Emerging Topics in Statistics and Biostatistics

Mohammad Arashi
Mina Norouzirad *Editors*

# Advances in Shrinkage and Penalized Estimation Strategies

## Honoring the Contributions of Professor A. K. Md. Ehsanes Saleh

Springer

# Emerging Topics in Statistics and Biostatistics

**Editor-in-Chief**

Ding-Geng Chen, College of Health Solutions, Arizona State University, Phoenix, USA

**Series Editor**

Hon Keung Tony Ng, Bentley University, Waltham, USA

**Editorial Board Members**

Andriëtte Bekker, Department of Statistics, University of Pretoria, Pretoria, South Africa

Carlos A. Coelho, NOVA University of Lisbon, Lisbon, Portugal

Maxim Finkelstein, Mathematical Statistics, University of the Free State, Bloemfontein, South Africa

Jeffrey R. Wilson, Department of Economics, W.P. Carey School, Arizona State University, Tempe, USA

Yuhlong Lio, Mathematical Sciences, University of South Dakota, Vermillion, USA

Emerging Topics in Statistics and Biostatistics (ETSB) showcases research and applications in statistics and biostatistics. It publishes books in statistical and biostatistical theory, applications, and education. Books may be monographs, edited volumes, textbooks and proceedings.

Mohammad Arashi • Mina Norouzirad
Editors

# Advances in Shrinkage and Penalized Estimation Strategies

## Honoring the Contributions of Professor A. K. Md. Ehsanes Saleh

Springer

*Editors*
Mohammad Arashi
Department of Statistics
Faculty of Mathematical Sciences
Ferdowsi University of Mashhad
Mashhad, Iran

Mina Norouzirad
Department of Mathematics and Center for Mathematics and Applications
NOVA School of Science and Technology
Caparica, Portugal

ISSN 2524-7735 ISSN 2524-7743 (electronic)
Emerging Topics in Statistics and Biostatistics
ISBN 978-3-031-94049-1 ISBN 978-3-031-94050-7 (eBook)
https://doi.org/10.1007/978-3-031-94050-7

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

*In memory of*



Professor A. K. Md. Ehsanes Saleh

# Preface

It is a great honor for us to edit this volume in recognition of Professor A. K. Md. Ehsanes Saleh, who serves as Distinguished Research Professor and also holds the title of Professor Emeritus at the School of Mathematics and Statistics, Carleton University, Ottawa, Canada.

In the recent years of his life, we had the privilege of working closely with him and benefiting from his vast knowledge, insight, and dedication. Professor Saleh was not only a remarkable academic but also a passionate mentor and collaborator. We had hoped to complete this book in his honor while he was still with us, but time passed too quickly.

This volume is a heartfelt tribute to Professor Saleh's remarkable contributions to statistics and mathematics. It brings together 25 papers from his colleagues, former students, and esteemed experts who generously contributed to this project. These works reflect his significant influence on the academic community and highlight his pioneering research in shrinkage estimation, robust estimation, nonparametric methods, penalized strategies, and variable selection, inspiring many scholars worldwide.

This book is organized into three parts based on Professor Saleh's influential work. Before these parts, we include a paper (Chap. 1) dedicated to an overview of Professor Saleh's academic life. What follows is a description of each part.

**Part I: Shrinkage Estimation Strategies** This part includes 10 chapters focusing on shrinkage estimators in various models. Topics covered include Liu-type estimators in linear regression (Chap. 2), shrinkage estimators in Bell regression models (Chap. 3), high-dimensional data (Chap. 4), tensor regression (Chap. 5), size-biased sampling (Chap. 6), balanced loss functions (Chap. 7), modified balanced loss functions (Chap. 8), super-efficiency (Chap. 9), wavelet denoising (Chap. 10), and minimaxity under concave loss (Chap. 11).

**Part II: Penalized Estimation and Variable Selection** This part contains 7 chapters focused on penalized estimation and shrinkage applications in various models. Topics include improved LASSO in semiparametric measurement error models (Chap. 12), weighted-average least squares estimation in panel data models

(Chap. 13), a comparison of penalized estimations in multiple regression (Chap. 14), limitations of model selection via t-tests in collinear regression (Chap. 15), ridge-based biased prediction in linear mixed models (Chap. 16), penalized L-estimation (Chap. 17), and variable selection in regression with dependent and asymmetrically distributed errors (Chap. 18).

**Part III: Robust Estimation and Nonparametric Methods** This part includes 7 papers focusing on robust estimation and nonparametric methods. Topics include shrinkage estimation in spatial autoregressive models with endogenous covariates (Chap. 19), robust neural networks (Chap. 20), mixture model estimation using component self-paced learning (Chap. 21), shrinkage estimation in generalized CIR processes with change points (Chap. 22), additive censored models (Chap. 23), confidence intervals in univariate normal mean estimation using preliminary test estimators (Chap. 24), and shrinkage estimators in energy supply interruption prediction (Chap. 25).

We extend our heartfelt gratitude to all the authors and reviewers whose invaluable contributions have made this volume possible. Each individual has played a significant role, and we hold them in the highest regard. The ordering of the papers does not imply any priority; rather, they are arranged alphabetically by the last name of the first author within each part.

Among these contributions, we wish to highlight two special papers: the first is the final collaborative work of the late Professor Saleh, co-authored with Professor Bashir Khan, whose ideas and discussions were in the last month of his life. We are deeply thankful to Professor Khan for granting us the privilege of including this work. The second is a paper by Professor William Edward Strawderman, who sadly passed away on October 1, 2024, during the preparation of this volume. We are profoundly grateful to his co-author, Professor Stavros Zinonos, for kindly seeing this work to completion. May both of them rest in peace, and may their invaluable contributions to the field of statistics continue to inspire future generations of researchers. Their dedication, insight, and passion for statistical science leave a lasting legacy, and we are honored to include their work in this volume.

We hope that this book not only honors the memory of Professor Saleh but also serves as lasting evidence of his passion for advancing knowledge and mentoring generations of scholars.

Mohammad Arashi's research is based upon research funded by Iran National Science Foundation, grant no. 4015320 and the National Research Foundation (NRF) of South Africa, Reference: RA211204653274, grant no. 151035, DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS), South Africa, and STATOMET at the Department of Statistics at the University of Pretoria. The work of Mina Norouzirad is funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UID/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications—NOVA Math).

Although there is great sadness for all of us on his passing, we can be joyous about his amazing legacy and a well-lived life! May he Rest in Peace.

Mashhad, Iran — Mohammad Arashi
Caparica, Portugal — Mina Norouzirad

# Contents


**1 A Legacy in Statistics: The Life and Contributions of Professor A. K. Md. Ehsanes Saleh** ..... 1
Mina Norouzirad and Mohammad Arashi
1.1 Personal Life, Family, and Passions ..... 1
1.2 Academic Background and Career ..... 2
1.3 Honors and Professional Recognition ..... 4
1.3.1 Keynote Speaker ..... 6
1.3.2 Invited Speaker ..... 6
1.3.3 Contributions to Learned Societies ..... 9
1.4 Mentorship and PhD Supervision ..... 10
1.5 Scholarly Publications ..... 11
1.5.1 Books ..... 11
1.5.2 Edited Volumes and Special Issues ..... 12
1.5.3 Peer-Reviewed Articles ..... 12
1.6 Specialized Fields ..... 14
1.6.1 Order Statistics ..... 16
1.6.2 Shrinkage Estimation ..... 18
1.6.3 Nonparametric Methods ..... 21
1.6.4 Decision Theory ..... 22
1.6.5 Time Series Models ..... 23
1.6.6 Robust Statistics ..... 24
1.6.7 Models with Measurement Error ..... 25
1.6.8 Elliptically Countoured Models ..... 26
1.6.9 Empirical Bayes Estimation ..... 27
1.6.10 Penalized Estimators ..... 28
1.6.11 Additional Contributions ..... 29
1.7 Final Reflections ..... 34
References ..... 35

**Part I Shrinkage Estimation Strategies**

**2 Restricted Liu-Type Regression Estimators in Linear Regression Model** ..... 51
Fikri Akdeniz and Esra Akdeniz
2.1 Introduction and Preliminaries ..... 51
2.2 Proposing Estimator ..... 55
2.3 Properties of $b_{rd}$ ..... 57
2.4 Comparisons Between the Restricted Liu-Type Estimator and the Restricted Ordinary Least Squares Estimator ..... 60
2.5 Experimental Results ..... 62
2.6 Concluding Results ..... 65
References ..... 65

**3 Shrinkage Strategies for Right-Censored Bell Regression Model with Application** ..... 67
Hossein Bevrani and Zahra Zandi
3.1 Introduction ..... 67
3.2 Right-Censored Bell Model and Unrestricted Estimator ..... 68
3.2.1 The Bell Model ..... 68
3.2.2 The Right-Censored Bell Model ..... 69
3.2.3 Unrestricted Estimator ..... 70
3.3 Restricted and Shrinkage Estimators ..... 71
3.3.1 Restricted Estimators ..... 71
3.3.2 Shrinkage Estimators ..... 71
3.4 Bias and Risk Analysis ..... 72
3.5 Simulation Study ..... 75
3.6 Doctor Visits Data ..... 78
3.7 Conclusion ..... 82
Appendix ..... 83
References ..... 86

**4 On a Class of Shrinkage Estimators of Normal Mean in High-dimensional Data with Unknown Covariance** ..... 87
Arash A. Foroushani and Sévérien Nkurunziza
4.1 Introduction ..... 87
4.2 Statistical Model and Some Fundamental Results ..... 88
4.3 Main Result ..... 90
4.3.1 Theorem 2 of Chételat and Wells (2012) Revisited ..... 91
4.3.2 Counterexample ..... 96
4.3.3 The Nonexistence Case of the Expectation of $F^{-1}$ ..... 98
4.4 Conclusion ..... 99
References ..... 100

**5 Some Stein-Rule Methods in Tensor Regression Model with High-Dimensional Data** ..... 101
Mai Ghannam and Sévérien Nkurunziza
5.1 Introduction ..... 101
5.2 Statistical Model and Notations ..... 102
5.2.1 CP Decomposition ..... 103
5.2.2 The Block Relaxation Algorithm ..... 103
5.3 Unrestricted and Restricted Estimators ..... 105
5.3.1 Asymptotic Properties of UE and RE ..... 106
5.4 Shrinkage Estimators and Asymptotic Results ..... 108
5.4.1 Shrinkage Estimators ..... 109
5.4.2 Asymptotic Distributional Risk (ADR) ..... 110
5.5 Application and Numerical Results ..... 112
5.5.1 Simulation Study ..... 112
5.5.2 Real Data Set ..... 116
5.6 Conclusion ..... 118
Appendix ..... 118
References ..... 128

**6 Some Implications of Preliminary-Test Estimation in the Context of Size-Biased Sampling** ..... 131
David E. Giles
6.1 Introduction ..... 131
6.2 Preliminary-Test Estimation ..... 132
6.2.1 The Preliminary-Testing Problem ..... 132
6.2.2 Properties of Preliminary-Test Estimators ..... 133
6.2.3 Preliminary Testing for Size-Biased Sampling ..... 133
6.3 The Generalized Gamma Distribution and Size-Biased Sampling ..... 135
6.3.1 Stacy's Distribution ..... 135
6.3.2 Some Single-Parameter Distributions ..... 136
6.4 A Monte Carlo Simulation Experiment ..... 140
6.4.1 The Test Statistic ..... 140
6.4.2 Bias and Risk Evaluations ..... 141
6.4.3 Optimal Critical Values ..... 142
6.5 Results ..... 144
6.5.1 Relative Biases ..... 144
6.5.2 Relative Risks ..... 147
6.5.3 Optimal Critical Values ..... 149
6.6 Concluding Remarks ..... 152
References ..... 155

**7 Study the Performance of New Shrinkage Estimators Under the Balanced Loss Function** ..... 159
Abdenour Hamdaoui, Abdelkader Benkhaled, and Amani Alahmadi
7.1 Introduction ..... 159
7.2 Preliminaries ..... 161
7.2.1 The Minimax Shrinkage Estimators ..... 162
7.2.2 Shrinkage Estimators Dominate the James-Stein Estimator ..... 165
7.3 Main Results ..... 170
7.4 Numerical Study ..... 173
7.5 Conclusion ..... 179
References ..... 180

**8 Shrinkage Estimators of the Location Parameter Under Modified Balanced Loss Functions** ..... 183
Lahoucine Hobbad, Idir Ouassou, and Mohamed Alahiane
8.1 Introduction and Preliminaries ..... 183
8.2 Risk Analysis for Loss $\omega\rho(\|\delta-\delta_0\|^2)+(1-\omega)\rho(\|\delta-\theta\|^2)$ ..... 185
8.3 Risk Analysis for Loss $\ell(\omega\|\delta-\delta_0\|^2+(1-\omega)\|\delta-\theta\|^2)$ ..... 191
8.4 Numerical Illustration ..... 196
8.4.1 Example 1 ..... 196
8.4.2 Example 2 ..... 197
8.5 Concluding Remarks ..... 198
Appendix ..... 199
References ..... 202

**9 Shrinkage Strategies and Superefficiency** ..... 205
Paul Kabaila
9.1 Introduction ..... 205
9.2 The Two Nested Linear Regression Models ..... 207
9.3 The Frequentist Model Averaged Estimator ..... 207
9.4 Convenient Expression for the Model Averaged Estimator ..... 208
9.5 Large Sample Framework ..... 209
9.6 Superefficiency of the Model Averaged Estimator ..... 209
9.6.1 A Theorem that Appears to Suggest that the Model Averaged Estimator is Better than the Least Squares Estimator ..... 210
9.6.2 The Model Averaged Estimator Is Actually Worse than the Least Squares Estimator ..... 212
9.7 Superefficiency of a Confidence Interval Centered on the Model Averaged Estimator ..... 213
9.7.1 A Theorem that Appears to Suggest that the Model Averaged Confidence Interval Is Better than the Usual Confidence Interval ..... 215

9.7.2 The Model Averaged Confidence Interval Is Actually Much Worse than the Usual Confidence Interval ..... 215
9.8 Discussion ..... 215
Appendix ..... 216
References ..... 226

**10 Shrinkage Estimation of Restricted Mean Vector Under Balanced Loss with Application in Wavelet Denoising** ..... 229
Hamid Karamikabir
10.1 Introduction ..... 229
10.2 Domination for the Positive Orthant Cones ..... 231
10.2.1 Domination for General Cones ..... 235
10.2.2 A Note for Nested Cones ..... 237
10.2.3 The Restricted Shrinkage Generalized Bayes Estimator ..... 240
10.2.4 The Restricted Soft Threshold Wavelet Shrinkage Estimator ..... 243
10.3 Simulation and Application ..... 248
10.3.1 Parkinson Speech Dataset ..... 248
10.4 Conclusion ..... 254
References ..... 254

**11 On Minimaxity of Shrinkage Estimators Under Concave Loss** ..... 257
Stavros Zinonos and William E. Strawderman
11.1 Introduction ..... 257
11.2 Normal Theory $\sigma^2$ Known ..... 259
11.3 Extensions ..... 261
11.3.1 Normal with Scale Unknown ..... 262
11.3.2 Scale Mixture of Normals with Unknown Scale ..... 263
11.4 Simulation Results ..... 263
11.5 Conclusion ..... 266
References ..... 266

**Part II Penalized Estimation and Variable Selection**

**12 Improved LASSO Estimator in Semiparametric Linear Measurement Error Models** ..... 271
Hadi Emami and Kourosh Dadkhah
12.1 Introduction ..... 271
12.2 Background and Definition ..... 273
12.2.1 Test Statistics ..... 277
12.2.2 Proposed LASSO Estimators and Their Asymptotic Distributions ..... 277
12.3 Asymptotic Distributional Risk Properties ..... 280
12.4 Algorithm ..... 282

12.5 Simulation ..... 284
12.5.1 Example 1 ..... 284
12.5.2 Example 2 ..... 288
12.6 Real Data ..... 290
12.7 Discussion ..... 294
References ..... 296

**13 Weighted-Average Least Squares Estimation of Panel Data Models** ..... 299
Giuseppe De Luca and Jan R. Magnus
13.1 Introduction ..... 299
13.2 The WALS Approach ..... 301
13.3 Nonspherical Errors ..... 306
13.4 Fixed and Random Effects ..... 306
13.5 WALS Estimation of the Fixed-Effects Model with i.i.d. Errors ..... 309
13.6 WALS Estimation of the Random-Effects Model with i.i.d. Errors ..... 310
13.7 The Baltagi–Wu Transformation ..... 312
13.8 The Impact of Legalized Abortion on Crime ..... 313
13.8.1 Preliminary Analysis of the Baseline Specifications ..... 314
13.8.2 Expanding the Set of Control Variables ..... 317
13.9 Concluding Remarks ..... 320
Appendix ..... 321
References ..... 324

**14 Performance of Some Test Statistics for Testing the Regression Coefficients for the One- and Two-Parameter Multicollinear Gaussian Multiple Linear Regression Models: An Empirical Comparison** ..... 327
Md Ariful Hoque, Zoran Bursac, and B. M. Golam Kibria
14.1 Introduction ..... 327
14.2 Statistical Methodology ..... 328
14.2.1 Model and Some Existing Estimators ..... 329
14.2.2 One-Parameter Estimators ..... 329
14.2.3 Stein-Type Estimators ..... 333
14.2.4 Two-Parameter Estimators ..... 336
14.3 Simulation Study ..... 339
14.3.1 Empirical Type I Error ..... 339
14.3.2 Empirical Power of the Test ..... 347
14.4 Real-Life Example ..... 352
14.5 Some Concluding Remarks ..... 353
References ..... 355

**15 Ineffectiveness of Model Selection via t-Test in Regression with Collinearity** ..... 359
Takeaki Kariya
15.1 Introduction ..... 359
15.1.1 Empirically Effective Modeling ..... 361
15.1.2 A Brief Review on Literature ..... 364
15.2 Effect of Collinearity on VSP-T ..... 365
15.2.1 Basic Setting of $t$-Testing ..... 366
15.3 Comprehensive Framework for Analyzing Collinearity Effect on Testing ..... 369
15.3.1 Simple Regression Models for Individual OLSEs ..... 369
15.3.2 Collinearity Effect on $t$-Test ..... 372
15.4 Test for Comparability of $t$-Values and Correlation Structure of $t$-Statistics ..... 373
15.4.1 Distribution of $(t_k, t_j)$ ..... 373
15.4.2 Testing the Equality of the Means of $t_k$ and $t_j$ ..... 377
15.5 Common Error Term Problem on MSP via VSP-T in View of NPTT ..... 378
15.5.1 MSP via VSP-T Within the Framework of NPTT ..... 380
15.5.2 A Remark on Indeterminacy of Null Distribution ..... 382
Appendix ..... 383
References ..... 385

**16 A New Ridge-Based Biased Prediction Technique in Linear Mixed Models** ..... 387
Özge Kuran and M. Revan Özkale
16.1 Introduction ..... 387
16.2 A New Ridge-Based Biased Prediction of Fixed and Random Effects ..... 390
16.3 Examination of Mean Square Error Performance ..... 394
16.3.1 The $k - d$ Estimator vs. the BLUE ..... 396
16.3.2 The $k - d$ Estimator vs. the Ridge Estimator ..... 397
16.3.3 The $k - d$ Estimator vs. the Liu Estimator ..... 397
16.4 Selection of the Biasing Parameters ..... 398
16.5 Real Data Analysis of Greenhouse Gases ..... 399
16.6 Conclusion ..... 402
References ..... 402

**17 L-Estimation of Location: Shrinkage and Selection** ..... 405
A. K. Md. Ehsanes Saleh and Bashir U. Khan
17.1 Introduction ..... 405
17.2 Asymptotic Distributional Biases and $L_2$-Risk of the Proposed L-Estimators ..... 409
17.2.1 Shrinkage L-Estimator of $\theta$ ..... 410
17.2.2 Ridge-Type L-Estimator of $\theta$ ..... 410
17.2.3 LASSO L-Estimator of $\theta$ ..... 412

17.2.4 Preliminary Test L-Estimator (PTLE) of $\theta$ ..... 413
17.2.5 Saleh-Type L-Estimator of $\theta$ ..... 413
17.2.6 Elastic Net L-Estimator of $\theta$ ..... 414
17.3 Risk Comparisons of the Different L-Estimators of $\theta$ ..... 416
17.3.1 Comparison of Restricted L-Estimator (RLE) and ULE ..... 417
17.3.2 Comparison of Preliminary Test L-Estimator (PTLE) and ULE ..... 417
17.3.3 Comparison of the Saleh-Type L-Estimator and ULE ..... 417
17.3.4 Risk Comparison of Ridge-Type L-Estimator and ULE ..... 418
17.3.5 Comparison of LASSO L-Estimators and ULE ..... 418
17.3.6 Comparison of LASSO and Ridge-Type L-Estimators ..... 419
17.3.7 Comparison of LASSO and Saleh-Type L-Estimators ..... 419
17.3.8 Comparison of LASSO L-Estimator and PTLE ..... 420
17.3.9 Comparison of PTLE and Ridge-Type L-Estimator ..... 420
17.3.10 Comparisons of Ridge-Type and Saleh-Type L-Estimators ..... 421
17.4 Application Using Logistic Regression ..... 421
17.5 Conclusion ..... 422
References ..... 424

**18 Variable Selection in Regression Models with Dependent and Asymmetrically Distributed Error Term** ..... 425
Yetkin Tuaç and Olcay Arslan
18.1 Introduction ..... 425
18.2 Regression Model ..... 427
18.3 Penalized Maximum Likelihood Estimation with Skew Distributions ..... 429
18.3.1 Skew Normal Distribution ..... 429
18.3.2 Steps of the ECM Algorithm ..... 435
18.3.3 Skew $t$ Distribution ..... 437
18.4 Numerical Examples ..... 447
18.4.1 Simulation ..... 448
18.4.2 Real Data ..... 457
18.5 Conclusion ..... 458
References ..... 459

**Part III Robust Estimation and Nonparametrics Methods**

**19 Shrinkage Estimator for Spatial Auto-regressive Model with Endogenous Covariates** ..... 463
Anoop Chaturvedi, Christian Heumann, and Shalabh
19.1 Introduction ..... 463
19.2 SAR Model with Endogenous and Exogeneous Regressors ..... 465

19.3 Combined Shrinkage Estimator and Asymptotic Distribution ..... 468
19.4 Comparison of Asymptotic Risks ..... 470
19.5 Monte Carlo Simulation ..... 472
19.6 Conclusions ..... 474
References ..... 476

**20 Regularization of Robust Neural Networks: Bayesian Connections and Outlier Detection** ..... 477
Jan Kalina
20.1 Introduction ..... 477
20.2 Bayesian Interpretation of Regularized Multilayer Perceptrons ... 478
20.2.1 The Bayes Theorem ..... 478
20.2.2 Multilayer Perceptrons ..... 479
20.2.3 $L_1$-Regularization ..... 480
20.2.4 $L_2$-Regularization ..... 481
20.2.5 General Remarks ..... 482
20.3 Bayesian Estimation in Linear Regression ..... 483
20.3.1 Bayesian Linear Regression: Basic Model ..... 483
20.3.2 Bayesian Linear Regression: Generalized Model ..... 485
20.4 Robust Regularized Multilayer Perceptrons ..... 485
20.5 Outlier Detection for Robust Regularized Multilayer Perceptrons ..... 489
20.5.1 Experiment: Artificial Data with $p = 1$ ..... 491
20.5.2 Experiments: A Real Dataset ..... 492
20.6 Conclusion ..... 494
References ..... 495

**21 Estimating Finite Mixture Models Using Component Self-Paced Learning** ..... 499
Andre R. Kleynhans, Frans H. J. Kanfer, and Salomon M. Millard
21.1 Introduction ..... 499
21.2 Finite Mixture Models ..... 500
21.2.1 Definition of the Finite Mixture Model ..... 501
21.2.2 Expectation Maximization Algorithm ..... 501
21.3 Self-Paced Learning ..... 503
21.4 Self-Paced Component Leaning ..... 505
21.5 Simulation Study ..... 507
21.5.1 Data Generation ..... 507
21.5.2 Model Fitting ..... 509
21.5.3 Simulation Study Results ..... 510
21.6 Real Data Analysis ..... 513
21.6.1 Iris Data Set ..... 513
21.6.2 Wine Data ..... 514
21.7 Discussion and Conclusion ..... 515
References ..... 515

**22 Shrinkage Estimation in Generalized CIR Processes with Change-point** ..... 517
Yunhong Lyu and Sévérien Nkurunziza
22.1 Introduction ..... 517
22.2 Statistical Model and Preliminary Results ..... 518
22.3 Estimation in the Case of One Known Change-Point ..... 527
22.3.1 The UMLE $\hat{\theta}_T$ and the RMLE $\tilde{\theta}_T$ ..... 527
22.3.2 Auxiliary Stationary and Ergodic Process ..... 528
22.3.3 Joint Asymptotic Normality of the UMLE and RMLE ..... 532
22.4 Shrinkage Estimators in the Case of Unknown Change-Point ..... 536
22.5 Numerical Evaluations and Real Data Analysis ..... 538
22.5.1 Parameter Estimation ..... 538
22.5.2 Relative Mean Squared Efficiency ..... 540
22.5.3 Real Data Analysis ..... 541
22.6 Conclusion ..... 544
Appendix ..... 545
References ..... 563

**23 Estimating and Pretesting in Additive Censored Models** ..... 565
Sévérien Nkurunziza
23.1 Introduction ..... 565
23.2 Statistical Model and Main Assumptions ..... 566
23.2.1 Statistical Model and Notation ..... 566
23.2.2 The Hypothesized Restriction and Main Assumptions ..... 568
23.2.3 Preliminary Results: The Unrestricted and Restricted Estimators ..... 569
23.3 The Main Result ..... 570
23.3.1 Joint Asymptotic Distribution of the Unrestricted and Restricted Estimators ..... 570
23.3.2 Shrinkage Preliminary Test Estimator ..... 573
23.3.3 Integrated Asymptotic Distributional Risk (IADR) ..... 575
23.4 Conclusion ..... 578
Appendix ..... 579
References ..... 583

**24 Confidence Interval for a Univariate Normal Mean Based on a Pretest Estimator** ..... 585
Nanami Taketomi, Jia-Han Shih, and Takeshi Emura
24.1 Introduction ..... 585
24.2 Background ..... 586
24.2.1 The CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ ..... 587
24.2.2 Asymptotic Properties of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ ..... 589

24.3 Proposed Confidence Interval ..... 593
24.3.1 Coverage Probabilities ..... 594
24.3.2 Asymptotic Property of the CP of the Pivot CI ..... 596
24.4 Simulation Studies ..... 598
24.5 Data Analysis ..... 599
24.6 Conclusion and Discussion ..... 602
Appendix ..... 603
References ..... 624

**25 Prediction of Interruptions in Energy Supply: A Machine Learning Study with Post-Shrinkage Modeling** ..... 627
Minjie Yu and Ravindra Khattree
25.1 Introduction ..... 627
25.2 Description of the Problem, Variables, and Basic Model ..... 628
25.3 The Statistical Model ..... 629
25.3.1 The Full Model Estimator ..... 629
25.3.2 Submodel Selection and Penalized Methods ..... 634
25.3.3 Post-Shrinkage Estimators ..... 635
25.4 The Analysis of the Data ..... 637
25.4.1 Full Model Estimation ..... 638
25.4.2 Submodel Estimation ..... 638
25.4.3 Development of Refined Estimators ..... 640
25.4.4 Future Predictions ..... 645
25.5 Copula Transformation-Based Modeling ..... 648
25.5.1 Full Model and Submodels for the Transformed Data ..... 649
25.5.2 Computation of Average Prediction Errors ..... 649
25.5.3 Future Predictions Using Copula-Transformed Approach ..... 653
25.6 The Analysis of Entire Data ..... 654
25.7 Conclusions and Closing Remarks ..... 656
References ..... 657

# Contributors

**Esra Akdeniz** Department of Medical Education, Marmara University, Istanbul, Türkiye

**Fikri Akdeniz** Department of Statistics, Çukurova University, Adana, Türkiye

**Mohamed Alahiane** Ecole Nationale de Commerce et de Gestion d'El Jadida, Jadida, Morocco

**Amani Alahmadi** Department of Mathematics, College of Science and Humanities in Ad Dawadmi, Shaqra University, Shaqra, Saudi Arabia

**Olcay Arslan** Department of Statistics, Faculty of Science, Ankara University, Ankara, Türkiye

**Abdelkader Benkhaled** Department of Biology, University of Mascara, Laboratory of Stochastic Models, Statistics and Applications, University Tahar Moulay of Saïda, Mascara, Algeria

**Hossein Bevrani** Department of Statistics, Faculty of Basic Sciences, University of Kurdistan, Sanandaj, Iran

**Zoran Bursac** Department of Biostatistics, Florida International University, Miami, FL, USA

**Anoop Chaturvedi** Department of Statistics, University of Allahabad, Allahabad, India

**Kourosh Dadkhah** Department of Statistics, Faculty of Basic Sciences, University of Kurdistan, Sanandaj, Iran

**Giuseppe De Luca** Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

**Hadi Emami** Department of Statistics, Faculty of Basic Sciences, University of Kurdistan, Sanandaj, Iran

**Takeshi Emura** Department of Fundamental Statistical Mathematics, The Institute of Statistical Mathematics, and Biostatistics Center, Kurume University, Fukuoka, Japan

**Arash A. Foroushani** Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada

**Mai Ghannam** Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada

**David E. Giles** University of Victoria, Victoria, BC, Canada

**Abdenour Hamdaoui** Department of Mathematics, University of Science and Technology, Mohamed Boudiaf, Oran, Algeria
Laboratory of Statistics and Random Modelisations of Tlemcen University, El Mnaouar, Oran, Algeria

**Christian Heumann** Department of Statistics, Ludwig Maximilians University, Munich, Germany

**Lahoucine Hobbad** National School of Applied Sciences, Cadi Ayyad University, Marrakesh, Morocco

**Md. Ariful Hoque** Department of Mathematics and Statistics, Florida International University, Miami, FL, USA

**Paul Kabaila** Department of Mathematical and Physical Sciences, La Trobe University, Melbourne, VIC, Australia

**Jan Kalina** The Czech Academy of Sciences, Institute of Computer Science, Prague, Czech Republic
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

**Frans H. J. Kanfer** Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa

**Hamid Karamikabir** Department of Statistics, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr, Iran

**Takeaki Kariya** Institute of Economic Research, Hitotsubashi University, Tokyo, Japan

**Bashir U. Khan** Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS, Canada

**Ravindra Khattree** Department of Mathematics and Statistics, Center for Data Science and Big Data Analytics, Oakland University, Rochester, MI, USA

**B. M. Golam Kibria** Department of Mathematics and Statistics, Florida International University, Miami, FL, USA

**Andre R. Kleynhans** Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa

**Özge Kuran** Department of Statistics, Faculty of Science, Dicle University, Diyarbakır, Türkiye

**Yunhong Lyu** Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada

**Jan R. Magnus** Department of Econometrics & Data Science, Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands

**Salomon M. Millard** Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa

**Sévérien Nkurunziza** Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada

**Idir Ouassou** National School of Applied Sciences, Cadi Ayyad University, Marrakesh, Morocco

**M. Revan Özkale** Department of Statistics, Faculty of Science and Letters, Çukurova University, Adana, Türkiye

**A. K. Md. Ehsanes Saleh** School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

**Shalabh** Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, India

**Jia-Han Shih** Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung, Taiwan

**William E. Strawderman** Department of Statistics, Rutgers University, New Jersey, USA

**Nanami Taketomi** Clinical Resarch Center in Hiroshima, Hiroshima University Hospital, Hiroshima, Japan

**Yetkin Tuac** Department of Statistics, Faculty of Science, Ankara University, Ankara, Türkiye

**Minjie Yu** DTE Energy, Distribution Operation Performance Center, Detroit, MI, USA

**Zahra Zandi** Department of Statistics, University of Kurdistan, Faculty of Basic Sciences, Sanandaj, Iran

**Stavros Zinonos** Global Biometrics and Data Managment, Pfizer Inc., New York, NY, USA

# List of Reviewers

**Marwan Al-Momani** Department of Mathematics, College of Sciences, The University of Sharjah, Sharjah, United Arab Emirates malmomani@sharjah.ac.ae

**Reza Arabi Belaghi** Unit of Applied Statistics and Mathematics, Department of Energy and Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden reza.belaghi@slu.se

**Mohammad Arashi** Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran arashi@um.ac.ir

**Yasin Asar** Department of Mathematics and Computer Sciences, Necmettin Erbakan University, Konya, Türkiye yasar@erbakan.edu.tr

**Anoob Chaturvedi** Department of Statistics, University of Allahabad, Prayagraj, India anoopchaturv@gmail.com

**Abdelkader Benkhaled** Department of Biology, University Mustapha Stambouli of Mascara, Mascara, Algeria benkhaled08@yahoo.fr

**N. Balakrishna** Department of Statistics, Cochin University of Science & Technology, Kochi, India nb@cusat.ac.in

**Kouider Djerfi** Department of Mathematics, University of Saida, Saida, Algeria djorfik@gmail.com

**Miguel Fonseca** Department of Mathematics and Center for Mathematics and Applications, NOVA School of Science and Technology, NOVA University Lisbon, Caparica, Portugal fmig@fct.unl.pt

**David E. Giles** Department of Economics, University of Victoria, Victoria, BC, Canada dgiles@uvic.ca

**Yeşim Güney** Department of Statistics, Faculty of Science, Ankara University, Ankara, Türkiye ydone@ankara.edu.tr

**Özlem Gürünlü Alma** Department of Statistics, Faculty of Sciences, Muğla Sıtkı Kocman University, Muğla, Türkiye ozlem.gurunlu@gmail.com

**Abdenour Hamdaoui** Department of Mathematics, University of Sciences and Technologies Mohamed Boudiaf, -USTOMB-, El Mnaouar, Bir El Djir, Oran, Algeria abdenour.hamdaoui@univ-usto.dz

**Shakhawat Hossain** Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, Canada sh.hossain@uwinnipeg.ca

**Hamid Karamikabir** Department of Statistics, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr, Iran h_karamikabir@pgu.ac.ir

**Yoshihiko Konno** Department of Mathematics, Osaka Metropolitan University, Osaka, Japan konno@fc.jwu.ac.jp

**Özge Kuran** Department of Statistics, Faculty of Science, Dicle University, Diyarbakır, Türkiye ozge-kuran@hotmail.com

**Éric Marchand** Département de Mathématiques, Université de Sherbrooke, Sherbrooke, Canada eric.marchand@usherbrooke.ca

**Mahsa Nadifar** Department of Statistics, University of Pretoria, Pretoria, South Africa mahsa.nst@gmail.com

**Sévérien Nkurunziza** Department of Mathematics and Statistics, University of Windsor, Windsor, Canada severien@uwindsor.ca

**Mina Norouzirad** Department of Mathematics and Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, NOVA University Lisbon, Caparica, Portugal mina.norouzirad@gmail.com; m.norouzirad@fct.unl.pt

**Idir Ouassou** Complex Systems Modeling Laboratory, National School of Applied Sciences, Cadi Ayyad University, Marrakesh, Morocco i.ouassou@uca.ac.ma

**Şenay Özdemir** Department of Statistics, Faculty of Arts and Sciences, Afyon Kocatepe University, Afyonkarahisar, Türkiye senayozdemir@aku.edu.tr

**Nestor Parolya** Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands n.parolya@tudelft.nl

**Jingfu Peng** Yau Mathematical Sciences Center, Tsinghua University, Beijing, China aypjf@mail.tsinghua.edu.cn

**Franco Peracchi** Institute for Economics and Finance (EIEF), University of Rome, Tor Vergata, Rome, Italy franco.peracchi@uniroma2.it

**Serge Provost** Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Canada provost@uwo.ca

**Resve Saleh** Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada ressaleh@shaw.ca

**Shalabh** Department of Mathematics & Statistics, Indian Institute of Technology, Kanpur, India shalab@iitk.ac.in

**Kadri Ulas Akay** Department of Mathematics, Istanbul University, Istanbul, Türkiye kulas@istanbul.edu.tr

**Gülesen Üstündağ Şiray** Department of Statistics, Faculty of Science and Letters, Çukurova University, Adana, Türkiye gustundag@cu.edu.tr

**Bahadır Yüzbaşı** Department of Econometrics, Inönü University, Malatya, Türkiye bahadir.yuzbasi@inonu.edu.tr

**Xinyu Zhang** Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China xinyu@amss.ac.cn

# Chapter 1
# A Legacy in Statistics: The Life and Contributions of Professor A. K. Md. Ehsanes Saleh




**Mina Norouzirad and Mohammad Arashi**

## 1.1 Personal Life, Family, and Passions

Professor Abul Khair Mohammed Ehsanes Saleh, widely published under the name A. K. Md. Ehsanes Saleh, was born on January 1, 1932, in Bikrampur, a village in the Munshiganj district of Dhaka, Bangladesh. He passed away peacefully at the age of 91 on September 3rd, 2023, in Ottawa, Canada, surrounded by his loving family. He is outlived by his devoted wife, Shahidara; his beloved daughter and sons, Isme, Resve, and Raihan; his loving grandchildren, Jasmine, Sarah, Malique, and Miguel; his sweet great-granddaughters, Anaya, Layla, and Samara; and a wider circle of Saleh and extended family members.

In his leisure time, Professor Saleh enjoyed playing the flute, cooking for his loved ones, and writing poetry. He was beloved by all who had the privilege of meeting him and remembered by friends and family as a pious, humble, and selfless man with unconditional love and faith in his Creator. He was one of the early members of the Ottawa Muslim Association. A devoted husband to his wife, Shahidara Saleh, he shared a sacred and loving bond with her for over seven decades. Truly, he was a role model, mentor, and an inspiration to his children, grandchildren, great-grandchildren, and all extended family members.

Professor Saleh was also dedicated to and delighted in the success of others, mentoring numerous graduate and doctoral students throughout his retirement, even

M. Norouzirad (✉)
Department of Mathematics and Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal
e-mail: m.norouzirad@fct.unl.pt

M. Arashi
Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: arashi@um.ac.ir

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_1

into his nineties. He was the first statistician of Bangladeshi descent who has published more than 200 research articles in many top journals. Never forgetting his roots and to further support the development of Bangladeshi students in the field, he established the Shahidara Saleh Trust Fund at the Institute of Statistical Research and Training (ISRT) at the University of Dhaka. He also provided scholarships by running a "star search" program that identified talented Bangladeshi students.

Professor Saleh was a pioneer in the early 1960s, becoming the first Bangladeshi to earn a doctorate in statistics from the newly established graduate program at the University of Western Ontario. He then joined Carleton University's Department of Mathematics and Statistics, where he founded the graduate program in statistics and spent the remainder of his career, with the exception of one year at Simon Fraser University in Vancouver, British Columbia, Canada. As a Professor Emeritus and Distinguished Research Professor at Carleton University, he devoted himself passionately to groundbreaking statistical research, publishing numerous journal articles and books over a remarkable 60-year career—continuing to contribute to his field until the final weeks of his life.

The outline of this chapter, devoted to the dedication of Professor Saleh, is as follows: Sect. 1.2 provides a comprehensive view of his academic background and career, covering both his degrees and professional appointments. In Sect. 1.3, we highlight the honors and recognitions Professor Saleh received, along with his active involvement in academic conferences and events. Section 1.4 discusses his role in mentoring graduate students and shaping future generations of statisticians. In Sect. 1.5, we review his vast scholarly output—including books, journal articles, and edited volumes—and his collaborations with international scholars. Section 1.6 provides an in-depth exploration of the specialized areas of statistics to which he made substantial contributions. Finally, Sect. 1.7 offers a reflective conclusion on Professor Saleh's enduring influence on the field of statistics, underscoring his groundbreaking work, mentorship, and lasting legacy in statistical theory and practice.

## 1.2 Academic Background and Career

Professor Saleh earned the following academic degrees:

**BSc** in Statistics, awarded in 1951 from Dhaka University, Bangladesh (stood sixth in order of merit with distinction)
**MS** in Statistics, awarded in 1953 from Dhaka University, Bangladesh
**MA** in Mathematical Statistics, awarded in 1962 from the University of Western Ontario, Canada
**PhD** in Mathematical Statistics, awarded in 1965 from the University of Western Ontario, Canada
**Title of thesis:** On the estimation of the parameters of the exponential distribution based on optimum order statistics in censored samples

It is notable that Professor Saleh is the first MA and PhD degree holder in Statistics from the University of Western Ontario (Kibria, 2009).

Professor Saleh held various academic and professional positions throughout his career, including:

**Distinguished Research Professor and Professor Emeritus**, School of Mathematics and Statistics, Carleton University (1997–2023)
**Director** of the Ottawa-Carleton Institute of Graduate Studies and Research (1991–1993)
**Full Professor**, Department of Mathematics and Statistics, Carleton University (1978–1997)
**Associate Professor**, Department of Mathematics and Statistics, Carleton University (1968–1978)
**Associate Professor**, Department of Mathematics, Simon Fraser University (1967–1968)
**Assistant Professor**, Department of Mathematics, Carleton University (1966–1967)
**Lecturer**, Department of Mathematics, University of Western Ontario (1964–1966)
**Statistical Officer**, Central Statistical Office, Karachi, Pakistan (1954–1961)

After retiring in 1997, he was appointed as a Distinguished Research Professor at Carleton University. Recognized for his outstanding contributions, he is regarded as one of the four elites in statistics and probability within the School of Mathematics and Statistics at Carleton University (Kibria, 2009).

Professor Saleh also served as a consultant for researchers at the following universities and institutes:

**Adjunct Professor**, Department of Biometry, Kansas University Medical Centre, Kansas, United States (1982–2023)—He provided consultation on survival analysis, multivariate data analysis, and nonparametric methods, focusing on developing new methodologies.
**Adjunct Professor**, Department of Statistics, Lahore University, Lahore, Pakistan (2000–2023).
**Statistical Advisor**, Statistics Canada (summer of 1975) in the Labour Force Survey Division.
**Member of the Statistical Consulting Services, Carleton University**, (1975–1985). Worked on the NRC IRAP project in collaboration with Dr. R. Thomas.
**Statistical Consultant**, the Sunnybrook Hospital, Toronto, Canada (1971–1972).
**Consultant**, Department of Social Welfare (1969–1971).
**Consultant**, Rio Algom Mines, Canada (1964–1965).

In his illustrious career, Professor Saleh held the position of Visiting Professor at several prestigious academic institutions, including:

- University of Toronto, Toronto, Ontario, Canada
- Stanford University, Standford, California, United States
- Massachusetts Institute of Technology, Cambridge, Massachusetts, United States
- University of California, Berkeley, California, United States

- Michigan State University, Michigan, United States
- Bowling Green State University (as a Eugene Lukacs Distinguished Professor), Ohio, United States
- Charles University, Prague, Czechia
- Kuwait University, Kuwait City, Kuwait
- Banach Center, Warszawa, Poland
- Department of Biometry at Kansas University Medical Center (as an Adjunct Professor), Kansas City, Kansas
- University of Southern Queensland, Toowoomba, Queensland, Australia
- Monash University, Melbourne, Victoria, Australia

Professor Saleh has held several prominent editorial roles in leading statistical journals, including the following:

**Editor-in-Chief**, Journal of Statistical Research—served in this role for over two decades

**Founding Editor-in-Chief**, Journal of Nonparametric Statistics—served from the journal's inception until 2003

**Editorial Board Member**, Journal of Statistical Studies—served from the journal's inception until 2017

## 1.3 Honors and Professional Recognition

Over the course of Professor Saleh's distinguished career, he received numerous awards and honors, including:

**Ogawa Award** (1991), jointly with Professor Tatsuya Kubokawa—For the best paper published in the Journal of the Japanese Statistical Society

**Q. M. Hossain Gold Medal**, awarded by the Bangladesh Statistical Association (1992)—In recognition of his exceptional research quality, fundamental and pioneering contributions to various areas of statistical science, prolific publication record, and dedicated service to the profession

**Eugene Lukacs Distinguished Visiting Professor**, Bowling Green State University, Ohio, United States (1992–1993)

**Award of Excellence and Plaque**, presented by the Canada Bangladesh Muslim Community, Ottawa, Canada (1999)—In recognition of his outstanding contributions to mathematical statistics

**Gold Medal and Certificate**, Islamic Society of Statistical Sciences (ISOSS) and Islamic Educational, Scientific and Cultural Organization (ISESCO), 1999—In recognition of outstanding research in statistics among the Organization of Islamic Conference member states

**"Pride of Bangladesh" Plaque**, Federation of Bangladesh Associations in North America (FOBANA), 14th Annual Meeting, Toronto, September 1–4, 2000—In

recognition of his distinction as an outstanding statistician and academic from Bangladesh

**Certificate of Recognition**, Natural Sciences and Engineering Research Council of Canada (NSERC), 2004—For important research achievements that have contributed to the sum total of human knowledge and to the economic and social well-being of Canadians over NSERC's 25 years of existence

**Gold Medal**, Islamic Society of Statistical Sciences (ISOSS), 2005—For outstanding contributions to statistical sciences and the impact of the Wiley book "An Introduction to Probability and Statistics"

**Honorary Member**, Statistical Society of Canada (SSC), 2005—For landmark contributions to nonparametric methods for preliminary test and Stein-type estimators, autoregression quantiles, and order statistics, founding the graduate program in Statistics at Carleton University, mentoring PhD students and postdoctoral researchers, and service to the profession, especially through the organization of symposium and edited volumes

**Gold Medal**, Institute of Statistical Research and Training (ISRT), University of Dhaka, Bangladesh, 2006—For lifetime achievements and contributions to statistical sciences and the development of the Journal of Statistical Research

**Gold Medal**, Statistical Society of Canada (SSC), 2007—For exceptional research and service to the statistical community

Colleagues and peers of Professor Saleh from around the world have acknowledged and celebrated his outstanding contributions in numerous ways. These honors include:

**Fellow**, Institute of Mathematical Statistics (IMS), 1991—In recognition of his contributions to the development, dissemination, and application of mathematical statistics

**Fellow**, American Statistical Association (ASA), 1995—For outstanding research on the development of improved estimation procedures based on preliminary test and shrinkage methods, contributions to nonparametric estimation, and dedicated service to the profession, particularly in organizing conferences and edited volumes

**Fellow**, Royal Statistical Society (RSS), 1967—Elected by the members of the Society

**Elected Fellow**, Bangladesh Academy of Sciences, 2003—For outstanding contributions to the development of statistical sciences in Bangladesh through impactful research

**Elected Member**, International Statistical Institute (ISI), The Hague, 1982

**Life Member**, Bangladesh Statistical Association

**Life Member**, Islamic Countries Society of Statistical Sciences (ISOSS)

**Expert Member**, University of Karachi, Pakistan (2000–2023) and University of Dhaka, Bangladesh (1992–1996, 2000–2023)

Professor Saleh retired in 1997. The following is a compilation of his presentations during his career as a keynote speaker, invited speaker, and contributor to

learned societies, based on the available records. It is important to note that there may be additional presentations that are not included on this list.

### *1.3.1 Keynote Speaker*

Professor Saleh delivered several keynote addresses at conferences and symposia throughout his career, sharing his expertise in statistical theory and its applications. Below is a list of his presentations:

"**Remarks on Regression Quantiles and its Applications**." The 3rd Eugene Lukacs Symposium in Statistics was held March 25–27, 1993.

"**Recent Development in Statistical Theory and Applications**," Q. M. Hossain Gold Medal Lecture, July 26, 1993, University of Dhaka, Bangladesh.

"**Recent Development of Preliminary Test Approach to Estimation**," The 217th Institute of Mathematical Statistics meeting. March 24–27, 1991, Houston, Texas, United States.

"**Recent Development of Preliminary Test Approach to Shrinkage Estimation, Inference under Uncertain Prior Information,"** University of Regina, April 6, 1990, Regina, Canada.

"**Shrinkage Least-Squares Estimation in General Multivariate Linear Models**," The 5th Pannonium Symposium on Mathematical Statistics, July 25–28, 1985, Visegrad, Hungary.

### *1.3.2 Invited Speaker*

Professor Saleh was invited to speak at various conferences and universities, where he shared his expertise in statistical theory and its applications. Below is a compilation of his invited talks:

"**Estimation of Location and Scale Parameters in Randomly Censored Samples**," University of Dhaka, July 26, 1993, Dhaka, Bangladesh

"**Logistic Distribution and Regression Estimators: Continuous and Binary Data Case**," Institute of Statistical Research and Training, University of Dhaka, July 24, 1993, Dhaka, Bangladesh

"**Various Regression Estimators and Their Properties**," Jahangirnagar University, July 22, 1993, Dhaka, Bangladesh

"**Estimation of Auto-Regressive Parameters**," Singapore National University, July 12, 1993, Singapore

"**L-Estimation of Regression Parameters Based on Selected Regression Quantiles**," Bowling Green State University, February 1993, Ohio, United States

Weekly seminar on "**Preliminary Test Approach to Shrinkage Estimation**," as Eugene Lukacs Distinguished (visiting) Professor, February–May 1993

"**Preliminary Test and Shrinkage Estimators of Regression Parameters**," Michigan State University, September 14, 1992, Michigan, United States

"**On Shrinkage Estimation via Preliminary Test**," Jahangirnagar University, December 22–29, 1991, Dhaka, Bangladesh

"**Preliminary Test Approach to Estimation**," Department of Statistics, University of Dhaka, December 18–20, 1991, Dhaka, Bangladesh

"**Estimation of Regression Coefficients When Errors are Autocorrelated**," Pacific Area Conference, December 10–15, 1991, Tokoyo, Japan

"**Asymptotic Theory for Simultaneous Estimation of Binomial Means**," Yokohama City University, December 8–10, 1991, Yokohama, Kanagawa Prefecture, Japan.

"**Estimation of Regression Coefficients When Errors are Autocorrelated**," The 3rd International Conference on Environmetrics, October 7–10, 1991, Madison, Wisconsin, United States

"**Estimation of Regression Coefficients in a Proportional Hazard Model**," International Conference on Order Statistics and Nonparametric Statistics, September 18–20, 1991, Alexandria, Egypt

"**Preliminary Test and Shrinkage Estimation of Regression Parameters—A Review**," Conference on Applied Statistics to Honor the 71st Birthday of Professor C. R. Rao, May 21–23, 1991

"**Pitman-Closeness of Some Preliminary Test and Shrinkage Estimators**," Conference on Pitman-Closeness Criterion, University of Texas, June 14–15, 1991, San Antonio, United States

"**Asymptotic Theory for Simultaneous Estimation of Binomial Means**," First Eugene Lukacs Symposium on Mathematical Statistics, Bowling Green State University, April 27–28, 1991, Bowling Green, Ohio, United States

"**On Shrinkage Estimation of Auto-Regressive Parameters in Gaussian Process**," American Mathematical Society, Special Sessions on Probability and Statistics, October 27–28, 1989, Ball State University, Muncie, Indiana, United States

"**How to Estimate Missing Cells in a Contingency Table**," Department of Statistics, Charles University, May 22–28, 1989, Prague, Czechoslovakia

"**Some Developments on Preliminary Test and Shrinkage Estimation**," Banach Mathematical Centre, May 10–22, 1989, Warsaw, Poland

"**Contingency Table and Estimation**," Department of Mathematics and Statistics, University of Windsor, March 10, 1989, Windsor, Ontario, Canada

"**Regression Quantiles and Improved L-Estimation in a Linear Model**," R. C. Bose Symposium, December 24–27, 1988, Delhi, India

"**Preliminary Test Approach to Estimation**," Department of Statistics and Actuarial Sciences, University of Western Ontario, March 10, 1988, London, Canada

"**Quantile Estimation Using Optimum Order Statistics**," Department of Statistics, University of Manitoba, April 12, 1987, Winnipeg, Canada

"**On Shrinkage Estimation for Parallelism Problem**," Department of Mathematics and Statistics, Bowling Green State University, March 15, 1986, Ohio, United States

"**On Improved Estimation of Cell-Proportions in a Contingency Table**," Institute of Mathematics, University of Tsukuba, December 15–16, 1986, Tsukuba, Japan

"**On Preliminary Test Version and Stein-Rule Estimation of Common Mean of Several Multinormal Populations**," Pacific Area Conference, December 10–14, 1986, Tokyo, Japan

"**Estimation of Regression Parameters Under Uncertain Restriction in a Subspace**" (with M. M. Ali), Department of Statistics and Actuarial Sciences, University of Western Ontario, May 15, 1986, London, Canada

"**On Improved Estimation of Location**," Department of Statistics, Purdue University, March 20, 1986, West Lafayette, Indiana, United States

"**R-Estimation of Regression Parameters**," Statistical Conference in the Basque Country, August 9–10, 1985, Bilbao, Spain

"**Shrinkage Least Squares Estimation in a General Multivariate Linear Model**," The 5th Pannonium Symposium on Mathematical Statistics, July 24–29, 1985, Visegrad, Hungary

"**Nonparametric Shrinkage Estimation of Locations**," Department of Industrial Engineering and Operation Research, Cornell University, March 20, 1985, New York, United States

"**Estimation of Location Parameters**," International Centre for Diarrhoeal Disease Research (ICDDR), Mohakhali Dhaka, February 20, 1985, Mohakhali, Dhaka, Bangladesh

"**Statistical Education in Canada**," Institute of Statistical Research and Training, University of Dhaka, January 18–22, 1985, Dhaka, Bangladesh

"**Preliminary Test Estimation**," Department of Statistics, University of Manitoba, October 15, 1983, Winnipeg, Canada

"**Nonparametric Test of Location Parameter After a Preliminary Test on Regression**," Meeting of European Statisticians, September 15, 1983, Palermo, Sicily, Italy

"**Preliminary Test Prediction in General Multivariate Linear Model**," Pacific Area Conference, December 14–17, 1982, Tokyo, Japan

"**On Quantile Estimation Using Order Statistics**," Department of Biometry, Kansas University Medical Centre, October 22, 1982, Kansas, United States

"**Preliminary Test Estimation of Locations**," Department of Statistics, University Missouri, October 20, 1982, Columbia, United States

"**A Review of Quantile Estimation Using Optimum Order Statistics**," Department of Mathematical Sciences, Ball State University, October 10, 1982, Muncie, Indiana, United States

"**A Course on Order Statistics and Estimation**," Department of Mathematics, Massachusetts Institute of Technology, February 15 to March 15, 1981, Cambridge, United States

"**Data Analysis**," Department of Statistics, Jahangirnagar University, December 14, 1978, Dhaka, Bangladesh

"**Order Statistics and Its Properties**," Department of Statistics, University of Dhaka, November 22, 1978, Dhaka, Bangladesh

"**Robust Estimation of Location**," Department of Statistics, University of Dhaka, November 20, 1978, Dhaka, Bangladesh
"**A Course on 'Estimation Theory'**," Department of Statistics, University of Dhaka, December 2–10, 1978, Dhaka, Bangladesh
"**A Course on 'Statistics for Social Sciences'**," Department of Statistics, Stanford University, June 15 to August 15, 1978, Standford, United States
"**A Course on 'Engineering Statistics'**," Department of Statistics, Stanford University, June 15 to August 15, 1978, Standford, United States
"**A Course on 'Elements of Nonparametric Statistics'**," Institute of Statistical Research and Training, University of Dhaka, August 2–6, 1976, Dhaka, Bangladesh
"**Characterization of Distributions Using Order Statistics**," Institute of Statistical Research and Training, University of Dhaka, August 23, 1976, Dhaka, Bangladesh
"**Hodges-Lehmann Estimate in Censored Samples**," Statistics Day Speaker, Ball State University, April 9–10, 1976, Muncie, Indiana, United States
"**Characterization of Distributions Using Spacings of Order Statistics**," Colloquium Speaker, Department of Mathematics and Statistics, University of Guelph, November 15, 1972, Guelph, Ontario, Canada

### *1.3.3 Contributions to Learned Societies*

Professor Saleh has contributed to various learned societies through his presentations. Below is a selection highlighting his involvement:

"**Estimating a Covariance Matrix of a Normal Distribution with Unknown Mean**," The 5th Purdue International Symposium on Statistical Decision Theory and Selected Topics, June 14–20, 1992, West Lafayette, Indiana, United States
"**Estimation of Means and Treatment Effects in a One-Way ANOVA Model**," Joint Meeting of Institute of Mathematical Statistics (IMS) and World Congress of the Bernoulli Society, August 13–18, 1990, Uppsala, Sweden
"**Estimation of Functions of Location and Scale Parameters Using Optimum Order Statistics**," Joint Meeting of IMS and World Congress of the Bernoulli Society, August 13–18, 1990, Uppsala, Sweden
"**Asymptotic Theory of Baseball Data Analysis**," The 47th Session of the International Statistical Institute, September 5, 1989, Paris, France
"**Preliminary Test Approach to Shrinkage Estimation**," Conference on Bayesian Statistics, June 15–16, 1989, Sherbrooke, Quebec, Canada
"**Shrinkage M-Estimation of Location Parameter**," Centennial Session of the International Statistical Institute, 1987, Amsterdam, Netherlands
"**Estimation of Quantiles of Exponential Distribution**," National Science Foundation (NSF) Meeting on Quantiles Processes, June 15–19, 1981, College Station, Texas, United States

"**On a Further Generalization of the Savage Test**," Joint Meeting of IMS and American Mathematical Society (AMS), August 15–18, 1977, Seattle, Washington, United States

"**Estimation of the Mean of a Multivariate Normal Distribution Using Preliminary Test**," Symposium on Statistical Decision Theory and Related Topics, May 17–19, 1976, Lafayette, Indiana, United States

"**Linear Order Estimate of Location Parameter with Pitman's Efficiency**," Symposium on Statistics and Related Topics, October 24–26, 1974, Ottawa, Canada

"**On the Behrens–Fisher Problem for the Exponential Distribution in Censored Samples**," Joint Meeting of European Statisticians and the 7th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, August 18–23, 1974, Prague, Czechoslovakia

"**Decomposition of Finite Mixtures**," Ontario Mathematical Meeting, March 23, 1974, Waterloo, Canada

"**Robust Estimation**," Department of Mathematics, March 23, 1972, Toronto, Canada

"**Characterization of Distributions Based on Expected Spacings between Consecutive Order Statistics**," Ontario Mathematical Meeting, March 10, 1971, London, Canada

"**Estimation of Regression Coefficients in a Polynomial Regression Model**," Summer Research Institute, June 22, 1971, Kingston, Canada

"**Robustness of Dixon's Statistics**," Ontario Mathematical Meeting, February 22, 1970, Ottawa, Canada

"**Estimation of Location After a Preliminary Test on Regression**," European Meeting of Statisticians, July 22–24, 1970, Hannover, Germany

"**Determination of Exact Optimum Order Statistics for Estimating the Parameters of Exponential Distribution in Censored Samples**," Summer Research Institute, June 22, 1967, Kingston, Canada

"**Optimum Quantiles for Exponential Distribution**," Institute of Mathematical Statistics Meeting, June 15, 1964, Chicago, United States

## 1.4 Mentorship and PhD Supervision

Professor Saleh founded the graduate program in Statistics at Carleton University and supervised his first PhD student, who graduated in 1971. Over the next 37 years, he mentored numerous master's and doctoral students, many of whom conducted research under his guidance in parametric theory—particularly on preliminary test and shrinkage estimation (Khan, 2009). Below is a selection of his PhD students and the thesis that stemmed from his supervision:

**Christine Scott** 1971, "On the Robustness of the Size of Student's $t$-Test"

**Jean-Paul Dionne** 1978, "Some Contributions to the Two-Sample Nonparametric Tests for Location and Scale Under Multiple Censoring"
**Syed Ejaz Ahmed** 1987, "Estimation Strategies Under Uncertain Prior Information"
**Abdunnabi M. Ali** 1990, "Interface of Preliminary Test Approach and Empirical Bayes Approach to Shrinkage Estimation"
**Soroush Alimoradi** 1995, "Some Contributions to L-Estimation in Linear Regression and Autoregression Models"
**Seyed Mohammad Mahdi Tabatabaey** 1995, "Preliminary Test Approach to Estimation: Regression Model with Spherically Symmetric Errors"
**Mohammad Abdul Matin** 1996, "Some Aspects of Statistical Inference of Logistic Regression Parameters" (Uppsala University; co-supervised)
**Javed Chowdhury** 1997, "Preliminary Test Approach to Shrinkage Estimation of the Parameters in a Variety of Statistical Models" (Co-supervised with Professor Vijay K. Rohatgi)
**Hyang Mi Kim** 2003, "On Some Contributions to Measurement Error Models"
**Mohammad Arashi** 2008, "Estimation of Parameters in Elliptically Contoured Distributions" (Ferdowsi University of Mashhad; co-supervised with Professor Seyed Mohammad Mahdi Tabatabaey)

## 1.5 Scholarly Publications

Professor Saleh authored numerous influential works in the field of statistics, including authored books, edited volumes, special issues, and peer-reviewed journal articles that have contributed to both theoretical development and applied methodologies.

### *1.5.1 Books*

His major research contributions are documented in his recent books, including:

- *An Introduction to Probability and Statistics* (Third Edition) by Rohatgi and Saleh (2015)
- *Theory of Preliminary Test and Stein-Type Estimation with Applications* by Saleh (2006)
- *Statistical Inference for Models with Multivariate t-Distributed Errors* by Saleh et al. (2014)
- *Theory of Ridge Regression Estimation with Applications* by Saleh et al. (2019)
- *Rank-Based Methods for Shrinkage and Selection: With Applications to Machine Learning* by Saleh et al. (2022)
- *Fundamentals of Robust Machine Learning: Handling Outliers and Anomalies in Data Science* by Saleh et al. (2025) —Completed by Professor Saleh just two weeks before his passing

### 1.5.2 Edited Volumes and Special Issues

In addition to his authored books, based on the details reported in Kibria (2009), Professor Saleh also contributed as an editor of several important volumes and special issues. Below are some of the edited volumes:

- *Nonparametric Statistics and Related Topics* by Saleh (1992b).
- *Data Analysis from Statistical Foundations: A Festschrift in Honour of the 75th Birthday of D. A. S. Fraser* by Saleh (2002a).
- A special volume of *Journal of Statistical Research* in honor of the 80th birthday of Professor C. R. Rao.
- A special volume of *Journal of Nonparametric Statistics* on "Regression Quantile" that was published in 1994.
- *Proceedings of the International Symposium on Statistics and Related Topics* by Csorgo et al. (1981).
- *Proceedings of the 3-Day Symposium on Statistics and Related Topics* (Saleh, 1974), Carleton Mathematics Series, 1974. This included the volume titled "An International Symposium on Nonparametric Statistics and Related Topics," which was held on September 15–17, 2006, at Carleton University in celebration of Professor Saleh's 75th birthday.

### 1.5.3 Peer-Reviewed Articles

We have compiled a comprehensive collection of 225 papers spanning various journals and conference proceedings, all of which were authored or co-authored by Professor Saleh throughout his career. Through an in-depth analysis of his publications, we found that he collaborated with a total of 80 different individuals, demonstrating his strong academic network and influence.

In the following, we present several plots based on the bibliographic information extracted from a BibTeX file containing data on Professor Saleh's publications, which was analyzed using Python 3.11.

Figure 1.1 visualizes these collaborations in the form of a network, where each node represents a co-author. The size of each node corresponds to the number of joint publications between Professor Saleh and the respective co-author, providing a clear visual representation of the depth and frequency of his academic partnerships over the years. This network serves not only as a testament to his wide-ranging influence but also as a glimpse into the academic relationships that have been shaped by his work.

Figure 1.2 highlights the diverse academic teams that Professor Saleh guided over the years, illustrating his role as a central figure in the academic community. Much like the sun at the center of a solar system, he radiated knowledge, providing guidance and inspiration to his students and collaborators. This co-authorship network visualizes the breadth of his influence, with each connection representing

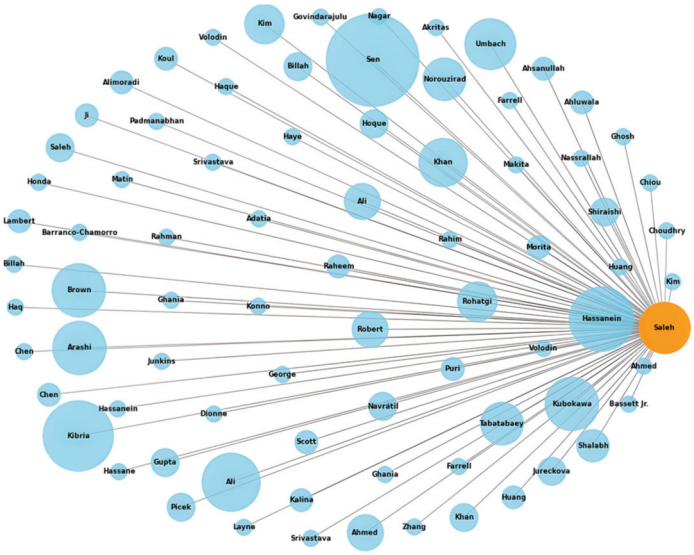



**Fig. 1.1** Professor Saleh's network of academic collaborations

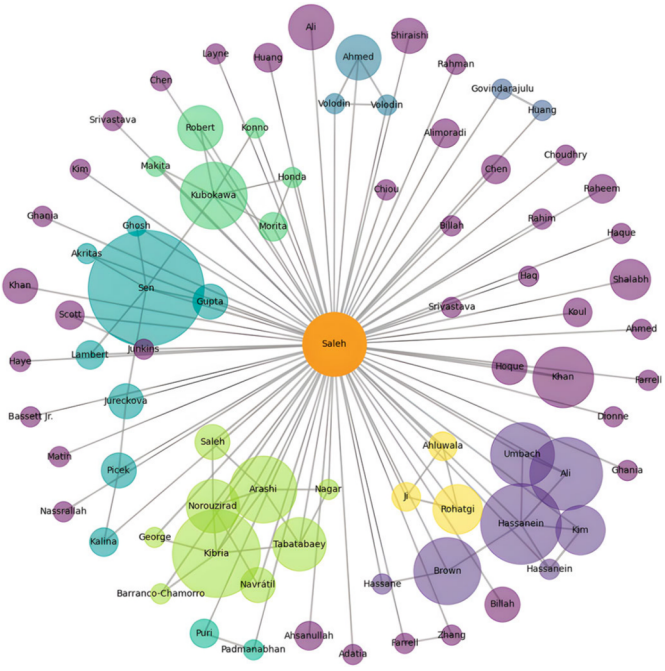



**Fig. 1.2** Professor Saleh's co-authorship network

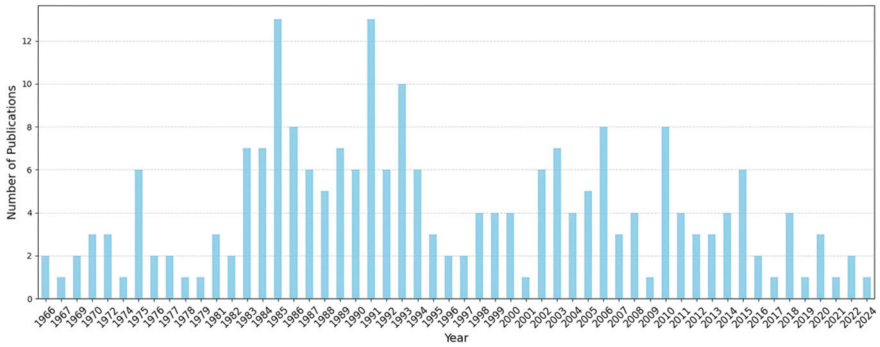



**Fig. 1.3** Yearly distribution of Professor Saleh's publications

a collaborative effort that enriched the field of statistics. The image serves as a testament to how his mentorship and collaborative spirit were key to the success of many researchers, shaping the academic journeys of countless individuals in the field.

Additionally, Fig. 1.3 presents the yearly distribution of his publications, offering a glimpse into the consistent productivity and research output that defined his career. The chart underscores not only the longevity of his academic contributions but also the sustained impact he had on statistical research over decades. Together, these figures provide a deeper understanding of Professor Saleh's far-reaching influence on both his immediate academic community and the broader field of statistics.

Furthermore, Fig. 1.4 displays the top 10 journals in which Professor Saleh's research was published. Notably, he contributed to the *Journal of Statistical Research*, an academic journal published in Bangladesh, publishing over 20 papers. His output in this journal greatly enhanced its academic visibility, helping establish it as a platform for statistical research.

## 1.6 Specialized Fields

Our analysis of Professor Saleh's published works reveals a rich and diverse body of research. Figure 1.5 presents a word cloud based on the titles of his papers, visually showcasing the breadth of his scholarly contributions. The cloud highlights the key themes and areas of focus in his work, reflecting his impact across multiple domains of statistics. Professor Saleh's pioneering research spanned various areas, including order statistics, shrinkage estimation, nonparametric methods, and robust statistics. He was particularly influential in bridging the fields of robust statistics and shrinkage estimation, making seminal contributions that advanced both areas significantly. Moreover, his groundbreaking work on autoregressive quantiles has

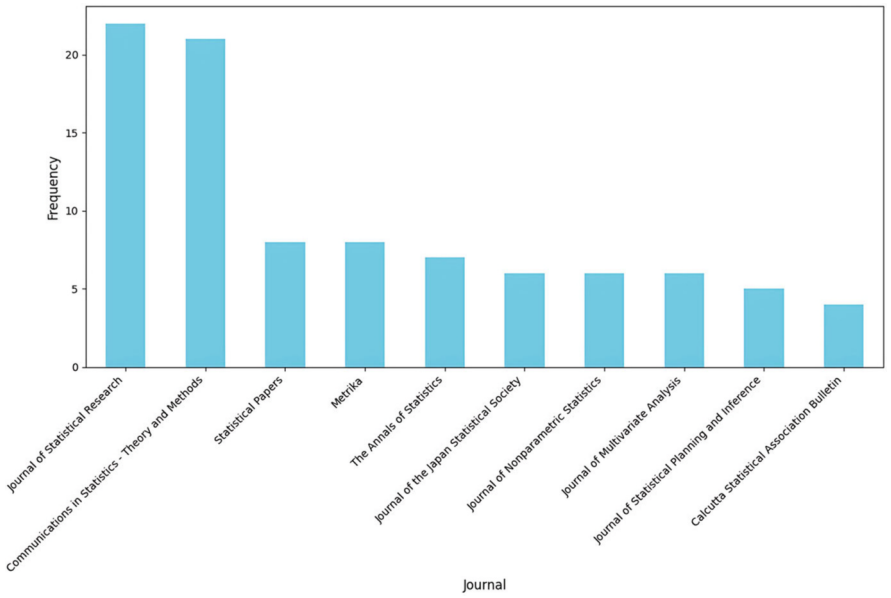



**Fig. 1.4** The top 10 most frequent journals of Professor Saleh's papers




**Fig. 1.5** The word cloud of titles of Professor Saleh's papers

had a profound effect on time series analysis, laying the foundation for many subsequent research topics.

In the following, we provide a comprehensive overview of Professor Saleh's remarkable body of work. His research not only advanced fundamental statistical theories but also had a lasting impact on applied methodologies.

### 1.6.1 Order Statistics

The earliest papers we found from Professor Saleh date back to the 1960s, with the first notable one published in 1966. His initial significant contributions were in the area of order statistics. The paper "Asymptotic optimum quantiles for the estimation of parameters of the negative exponential distribution" written by Saleh and Ali (1966) was part of his master's thesis and addressed a problem posed by Professor H. Leon Harter in the journal *Annals of Mathematical Statistics* in 1961 (Harter, 1960), which challenged the mathematical proof for the uniqueness of optimal quantiles when estimating the scale parameter of the exponential distribution based on a few quantiles. The second paper entitled "Estimation of the Parameters of the Exponential Distribution Based on Optimum Order Statistics in Censored Samples" (Saleh, 1966) came from his PhD thesis and was a response to a referee's comment on his earlier work, dealing with the issue of censored samples (Khan, 2009).

Given a random sample $X_1, \ldots, X_n$ which are assumed to be independently and identically distributed random variables with common cumulative distribution function $F$, the order statistics are the sorted values of the sample in ascending order $X_{(1)} \leq \cdots \leq X_{(n)}$ where $X_{(i)}$ represents the $i$th smallest value in the sample.

Censoring occurs when certain parts of the data are unavailable or missing, either intentionally due to the design of an experiment or unintentionally due to external factors. To illustrate this concept, consider a technical system composed of $n$ components that remains operational as long as at least $k$ components are functioning. The $(n - k + 1)$th order statistic, denoted as $X_{(n-k+1)}$, represents the lifetime of the system. If all component failures are fully observed, we obtain a complete sample of lifetimes. If failures can be observed only until the system breaks down, the data is right-censored beyond this point. So, the observed data consists of $X_{(1)} \leq \cdots \leq X_{(n-k+1)}$. The upper failure times $X_{(n-k+2)}, \ldots$, and $X_{(n)}$ are missing. In some cases, data is censored on both ends (left and right). This leads to the observed dataset as $X_{(r)} \leq X_{(r+1)} \leq \cdots \leq X_{(s)}$ where $1 \leq r < s \leq n$ that it is called double censoring. In middle censoring, some observations in the middle portion of the data are missing, i.e., the values between $X_{(r)}$ and $X_{(s)}$ are missing. Thus, the observed data consists of $X_{(1)}, \ldots, X_{(r+1)}, \ldots, X_{(n)}$. This type of censoring was studied and introduced by Saleh (1967) and Saleh and Ahsanullah (1970), who referred to it as censoring in the middle. Later, Saleh et al. (1983b) presented a refined and simplified version of the estimation method originally proposed by Saleh (1967).

Saleh (1981b) paved the way for the asymptotic theory of best linear unbiased estimator (BLUE) based on a few order statistics in estimating quantile functions.

Between 1982 and 1986, Professor Saleh, in collaboration with Professors Mir Masoom Ali and Dale Umbach, concentrated on quantile estimation and significance testing using order statistics. Their research applied optimal spacing and selected order statistics methods for quantile estimation across various distributions. Key contributions include:

- **Exponential distribution**: Quantile estimation using optimal spacing methods (Ali et al., 1982) and significance testing based on selected quantiles (Ali et al., 1985)
- **Pareto distribution**: Large sample estimation of Pareto quantiles (Saleh et al., 1985)
- **Double exponential distribution**: Estimation of quantiles and hypothesis testing using optimal spacing (Umbach et al., 1986), with additional work on significance testing based on selected quantiles (Umbach et al., 1983)
- **Location-scale family of distributions**: Quantile estimation using selected order statistics (Saleh et al., 1983a)
- **General methodology**: A broader framework for quantile estimation based on selected order statistics (Ali et al., 1983) and tests of significance (Saleh et al., 1984)

Additionally, they investigated simplified estimation of the parameters of the exponential distribution from censored data (specifically middle censoring) (Saleh et al., 1983b). Finally, the use of selected order statistics for estimating the life function of a chi distribution was explored in Ali et al. (1992).

Professor Saleh has been an adjunct professor in the Department of Biometry at the University of Kansas Medical Center since 1982. He collaborated with Professors Khatab M. Hassenein and Edward F. Brown, on order statistics, particularly in determining quantiles for various distributions and studying extreme-value distributions. Their work resulted in several publications, including:

- **Quantile estimation based on order statistics**: the extreme-value and Weibull distributions (Hassanein et al., 1984, 1986b, 1994), logistic distribution (Hassanein et al., 1985; Saleh et al., 1988b; Adatia & Saleh, 2012), normal distribution (Hassanein et al., 1986a,c), chi-distribution (Hassanein et al., 1990), uniform distribution (Adatia & Saleh, 2009)
- **Testing of hypotheses and significance testing**: Cauchy distribution (Saleh et al., 1985), location-scale distribution (Saleh & Hassanein, 1986b), normal distribution (Saleh et al., 1988a), simple linear model with extreme value errors (Saleh et al., 1994b), and with cauchy errors (Saleh et al., 1994a)
- **Survival analysis and distribution functions**: estimating survivor function using optimally selected order statistics (Hassanein et al., 1991) and for the extreme-value distribution based on selected order statistics (Hassanein et al., 1994)
- **Estimation in regression models**: parallelism problem (Saleh & Hassanein, 1986a) and quasi-empirical Bayes methodology for improving meta-analysis (Saleh et al., 2006)

Other significant contributions by Professor Saleh on order statistics are:

- Expected value of the spacings between order statistics (Govindarajulu et al., 1975)

- Characterization of distributions using spacings between order statistics (Saleh, 1975a) and expected spacing between consecutive order statistics (Saleh, 1976a)
- Asymptotic relative efficiency of some joint tests of location and scale parameters based on selected order statistics (Saleh & Sen, 1985a)
- A class of distributions connected to order statistics with nonintegral sample size (Rohatgi & Saleh, 1988)
- Estimation of regression coefficients in an exponential regression model with censored observations (Ahmed & Saleh, 1999b)

In a joint work with Professor Gui-Jing Chen (Saleh & Chen, 2000), Professor Saleh established the strong consistency and asymptotic normality of regression quantiles in linear models.

### *1.6.2 Shrinkage Estimation*

The family of shrinkage estimators for an arbitrary parameter $\beta$ (which can represent a mean, location, scale, or regression parameter) is defined as

$$\hat{\beta}^{\text{Shrinkage}} = g(\mathcal{T})\hat{\beta} + (1 - g(\mathcal{T})\tilde{\beta} = \tilde{\beta} - g(\mathcal{T})(\tilde{\beta} - \hat{\beta})$$

where $\tilde{\beta}$ denotes an estimator obtained using standard methods such as the maximum likelihood estimator (MLE), least squares estimator (LSE), or Bayes estimator, without imposing any restrictions on the parameter space; $\hat{\beta}$ represents an estimator derived under a restricted parameter space, which may incorporate prior information or expert knowledge; $\mathcal{T}$ is the test statistic used to evaluate the restriction correctness, commonly in a form of null hypothesis; and $g(\mathcal{T})$ is a nondecreasing function that determines the shrinkage behavior.

The function $g(\mathcal{T})$ can take various forms, leading to different types of shrinkage estimators:

- If $g(\mathcal{T}) = 0$, then $\hat{\beta}^{\text{Shrinkage}} = \tilde{\beta}$.
- If $g(\mathcal{T}) = 1$, then $\hat{\beta}^{\text{Shrinkage}} = \hat{\beta}$.
- If $g(\mathcal{T}) = I(\mathcal{T} \leq \mathcal{T}_\alpha)$, where $I(\cdot)$ is the indicator function and $\mathcal{T}_\alpha$ is the critical value of the test statistic at significance level $\alpha$, then $\hat{\beta}^{\text{Shrinkage}} = \hat{\beta}^{\text{PT}}(\alpha)$, the preliminary test estimator.
- If $g(\mathcal{T}) = c\mathcal{T}^{-1}$, where $c$ is a shrinkage parameter, then $\hat{\beta}^{\text{Shrinkage}} = \hat{\beta}^{\text{S}}$, the Stein-type estimator.
- If $g(\mathcal{T}) = c\mathcal{T}^{-1} + (1 - c\mathcal{T}^{-1})I(\mathcal{T} \leq c)$, then $\hat{\beta}^{\text{Shrinkage}} = \hat{\beta}^{\text{S+}}$, the positive-part (or rule) Stein-type estimator.

In 1978, Professor Bradley Efron invited Professor Saleh to Stanford University for the Summer and Fall terms. During his stay, Professor Saleh met Professor Charles Stein, and they discussed the nonparametric approach to preliminary test

estimation. Stein encouraged him to further explore Stein-type shrinkage estimation from a nonparametric perspective. This interaction marked a significant milestone in Professor Saleh's research trajectory.

Throughout his academic career, Professor Saleh investigated shrinkage and preliminary test estimators across a wide range of models, not limited to the nonparametric case. Covering the entirety of his contributions is a challenging task, but in the following sections, we highlight some key examples of his work and provide a more detailed discussion in subsequent sections.

Preliminary test estimation has been applied to various models, including but not limited to:

- Preliminary test estimator for the intercept after conducting a test on the slope in a simple regression model (Saleh & Ahsanullah, 1972)
- Preliminary test prediction in the general multivariate linear model (Saleh & Sen, 1985d).
- Application of preliminary test estimation in homogeneity of intercept in parallelism of regression models (Lambert et al., 1985)
- Relative performance of Stein-rule and preliminary test estimates in linear models under least squares theory (Saleh & Sen, 1987c).
- Preliminary test estimators of the mean for sampling from Student-t populations (Saleh & Khan, 1995).
- Preliminary test prediction of population totals under regression models with measurement error (Saleh & Kim, 2002).
- Preliminary test confidence sets for the mean of a multivariate normal distribution (Chiou & Saleh, 2002)
- Comparison between Han-Bancroft and Brook methods to determine optimal significance level for preliminary test estimator (Saleh & Kibria, 2005)
- Optimum critical value for preliminary test estimator (Kibria & Saleh, 2006)
- Preliminary test estimator for multivariate mean under different tests (Saleh & Kibria, 2006)
- Preliminary test approaches in high-dimensional multivariate regression models (Arashi et al., 2014).

Professor Saleh's work spans a wide array of models and applications for shrinkage estimation. Some notable contributions include:

- Shrinkage least squares estimation in multiple linear regression (Saleh & Sen, 1985e).
- Shrinkage estimators of multivariate location parameters (Sen & Sen, 1985).
- Shrinkage least squares estimation in parallelism problems (Saleh & Sen, 1986a).
- Shrinkage estimators for the parameters of location-scale family distributions in both complete and type II censored samples (Saleh & Rohatgi, 1988)
- Shrinkage estimators for a multivariate mean vector in the context of the two-sample problem (Ahmed & Saleh, 1989)

- Shrinkage estimator for contingency tables (Saleh et al., 1989)
- Shrinkage estimation for the coefficient matrix in multivariate regression models (Kubokawa & Saleh, 1990).
- Shrinkage estimation for the intercept vector in a simple linear multivariate normal regression model (Ahmed & Saleh, 1990)
- Shrinkage estimators of the common mean of several multinormal populations (Saleh & Sen, 1991)
- Shrinkage estimation of coefficient matrix in a growth curve model (Kubokawa et al., 1992)
- Shrinkage estimation of the parameters of an autoregressive Gaussian process (Saleh, 1993a)
- Shrinkage estimators of variance components (Saleh et al., 1993c)
- Shrinkage estimations for the component mean vector (Ahmed & Saleh, 1993)
- Shrinkage estimators of the intercept parameter in regression models with multivariate Student-t error (Khan & Saleh, 1997).
- Shrinkage estimators in mixed Multivariate Analysis of Variance (MANOVA) and Generalized MANOVA (GMANOVA) models (Konno et al., 1997).
- Shrinkage estimators for the univariate normal mean (Khan & Saleh, 2001).
- Shrinkage estimation of the intercept and slope parameter for linear regression model with uncertain prior information (Khan et al., 2005; Saleh et al., 2002)
- Shrinkage estimation of the mean vector of a multivariate normal distribution under various test statistics (Saleh & Kibria, 2003b)
- Shrinkage estimation of the parameters in the ill-conditioned Gaussian regression model (Kibria & Saleh, 2004)
- Shrinkage estimation of the mean vector of a multivariate normal distribution under subspace hypothesis (Srivastava & Saleh, 2005)
- Shrinkage estimation of regression parameters with a two-sample problem (Saleh & Khan, 2006)
- Shrinkage estimators in logistic regression model with skew-normally distributed explanatory variables (Matin & Saleh, 2006)
- Shrinkage estimation of parameters when the regression lines are parallel (Saleh & Khan, 2007a,b)
- Shrinkage estimation in ANOVA models (Saleh et al., 2017).
- Shrinkage estimation in non-orthogonal regression models (Saleh & Norouzirad, 2018).
- Shrinkage simultaneous estimation of several cumulative distribution functions under homogeneity constraint (Saleh et al., 2018)

One of the most renowned shrinkage estimators is the James–Stein estimator (James & Stein, 1961), which significantly improved the ordinary estimation of the mean vector for a random multivariate variable. In a joint work in 2007, Professor Saleh and his collaborators derived the asymptotic expansion of the coverage probability of James–Stein estimators (Ahmed et al., 2007).

### *1.6.3 Nonparametric Methods*

Nonparametric statistics refer to statistical methods that do not assume data comes from predefined models determined by a small number of parameters.

In data analysis, it is often necessary to check if two samples have the same distribution. If the data is normal, a t-test works well. But real-world data is not always normal, and it is hard to know its exact distribution. In these situations, rank tests offer a practical alternative. As nonparametric methods, rank tests do not rely on assumptions about the data's distribution, making them particularly valuable for analyzing datasets with unknown or non-normal distributions. Professor Saleh began his research in nonparametric statistics with the work of "Asymptotically most powerful rank tests of regression parameters from grouped data" (Saleh, 1969a).

He then collaborated with Professor Muni Shanker Srivastava on Saleh and Srivastava (1970) while both were research fellows at the Summer Research Institute at Queen's University in Kingston, Canada (Khan, 2009). The extension of Hodges-Lehmann estimate (Hodges & Lehmann, 1963) in censored samples was studied in Saleh (1976b). In the following years, Professor Saleh built upon the findings of Han and Bancroft (1968), which laid the foundation for studies on the estimation of location parameters. Consider $X_1, \ldots, X_n$ as independent random variables with absolutely continuous distribution functions $F_{X_i}(x) = F(x-\theta-\beta t_i)$, for $i = 1, \ldots, n$ and $x \in \mathbb{R}$ where $\{t_i, i \geq 1\}$ is a sequence of known constants (not all equal), and $\theta$ and $\beta$ are unknown parameters. If $\beta = 0$, the random variables $X_i$ are identically distributed with a common location parameter $\theta$. The estimator of $\theta$ varies depending on whether $\beta$ is specified or unknown. Building on this framework, Saleh and Sen (1978, 1979) provided nonparametric estimation methods for the location parameter using preliminary test estimation, further advancing the field. These two papers marked the beginning of decades-long collaboration with Pranab Kumar Sen from the University of North Carolina (Khan, 2009).

According to Khan (2009), the introduction of preliminary test estimation and testing in a nonparametric framework paved the way for rich and impactful research, with these two papers (i.e., Saleh & Sen, 1978, 1979) serving as the foundation for many subsequent studies.

Professor Saleh produced a remarkable sequence of over 30 significant research articles in collaboration with Professor Pranab Kumar Sen.[1]

In 1991, Professor Sen invited Professor Saleh to contribute an invited paper to the first volume of the *Journal of Nonparametric Statistics*. In this short paper (Saleh, 1991), he discussed two key topics: (i) rank tests for grouped data and (ii) the preliminary test approach to rank estimation of multiple parameters, highlighting both past and recent developments.

[1] Professor Sen was born in November 7, 1937, and passed away on December 31, 2023, a few months after the passing of Professor Saleh.

The work of Professor Saleh on the nonparametric methods, mostly joint with Professor Sen includes:

- **Nonparametric estimation** rank order preliminary test estimation in multiple linear regression model (Saleh & Sen, 1984a), shrinkage nonparametric estimation of location vector (Ahmed & Saleh, 1999a), and structural errors in variable models (Saleh & Chen, 2006)
- **Nonparametric tests** for location parameter using a preliminary test estimator in a simple linear regression (Saleh & Sen, 1982), multiple regression (Saleh & Sen, 1983b, 1984c); nonparametric tests of equality against ordered alternatives in the case of skew data (Padmanabhan et al., 1981); for the two-sample problem when the combined sample is type II censored (Ji et al., 1990; Saleh et al., 1991); rank test for the two-sample problem with censored samples (Saleh & Puri, 1984)
- **Nonparametric inference** of a preliminary test estimator based on a broad class of rank order statistics (Saleh & Sen, 1984b)
- **Asymptotic characteristics of nonparametric tests** based on rank statistic test of hypothesis of subset of parameters using a preliminary test estimation in multiple linear regression (Saleh & Sen, 1983a)
- **Parallelism models** estimation of intercept using a preliminary test (Lambert et al., 1984; Saleh et al., 1985), nonparametric shrinkage estimation (Saleh & Sen, 1985b)
- **Selection procedures**: nonparametric randomized selection procedures for location parameters (Huang & Saleh, 1985)

The Savage test is a statistical test used for comparing two independent samples from continuous distributions, where both distributions belong to the same family indexed by the parameter $\theta$. Specifically, let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be two independent samples drawn from distributions $F(x)$ and $G(y)$, respectively. Let $W_1, \ldots, W_N$ (where $N = n_1 + n_2$) denote the ordered observations from the combined sample of the $n_1$ $X$-observations and the $n_2$ $Y$-observations. Additionally, let the vector $Z = (Z_1, \ldots, Z_N)$ represent the rank order statistics, where $Z_i = 1$ if $W_i$ is an $X$-observation, and $Z_i = 0$ if $W_i$ is a $Y$-observation. Savage (1956) proposed a test statistic based on linear combinations of the rank order statistics $Z_1, \ldots, Z_N$ to test the null hypothesis $\mathcal{H}_0 : F = G$ against the alternative hypothesis $\mathcal{H}_1 : F \neq G$. This test is widely known as the Savage test. In 1977, Saleh and Dionne (1977) extended this approach by proposing a generalized version of the Savage test, which has important applications in life testing and clinical trials, allowing for more flexibility in testing hypotheses in these fields.

### *1.6.4 Decision Theory*

In 1986, Professor Saleh was invited to the Pacific Area Statistical Conference in Tokyo. During his visit to Japan, Professor Masatoshi Sugiura from the University

of Tsukuba hosted him for a two-day visit to his department. Professor Sugiura introduced Professor Saleh to two promising PhD scholars: Tatsuya Kubokawa, who specialized in decision-theoretic statistics, and Taka-aki Shiraishi, who focused on nonparametric statistics. Recognizing their potential, Professor Sugiura encouraged Professor Saleh to bring them to Canada for further training and collaborative research. At the time, Professor Saleh was working on a problem related to Pitman nearness in preliminary test estimators with Professor Sen which end with Saleh and Sen (1991b, 1991c). He was intrigued by the work of Professor Kubokawa (Khan, 2009), which led to a fruitful collaboration. Their joint work resulted in the publication Sen et al. (1989). Similarly, Professor Saleh's collaboration with Professor Shiraishi produced several significant contributions, including Saleh and Shiraishi (1989, 1992, 1993).

This collaboration with researchers from Japan resulted in the publication of several joint research papers, with contributions from scholars such as Professors Tatsuya Kubokawa, Takashi Shiraishi, Toshio Honda, and Yoshihiko Konno, as well as S. Makita.

The publications include works like Kubokawa and Saleh (1990), Kubokawa et al. (1991), Kubokawa et al. (1992), Saleh et al. (1993a, 1993b, 1993c), Kubokawa and Saleh (1994), Konno et al. (1997), Kubokawa et al. (2000), Saleh and Shiraishi (1989), Saleh and Shiraishi (1992), Saleh and Shiraishi (1993).

Based on the information extracted from Khan (2009), Professor Christian Robert from Ceremade, Université Paris-Dauphine, became involved in the collaboration when he was a regular visitor to Professor Saleh. His first paper (Robert & Saleh, 1991) with Professor Saleh addressed the empirical Bayes solution to the parallelism problem. Around the same time, both Professors Kubokawa and Robert collaborated on various topics, including the empirical Bayes solution to the covariance matrix estimation under entropy loss, the estimation of noncentrality parameters, and robust estimation of common regression coefficients under spherical symmetry (Kubokawa et al., 1991, 1992; Saleh et al., 1993c; Saleh & Robert, 2002).

### *1.6.5 Time Series Models*

At the 1989 Mathematical Society Conference at Ball State University, United States, Professor Saleh presented his newly published work on Stein estimation for the autoregressive (AR) model with Gaussian errors. This presentation brought his research on time series models to the forefront and captured the attention of Professor Hira L. Koul, who became interested in shrinkage estimators. Their collaboration began with a joint paper on R-estimation for the AR model (Koul & Saleh, 1993), marking the start of a productive partnership.

In 1992, while serving as the Eugene Lukacs Professor at Bowling Green University, Professor Saleh organized a conference on regression quantiles to assess progress in the field. Later, when Professor Koul invited him to Michigan State

University, Professor Saleh proposed studying the regression quantile problem in the context of the AR model (Khan, 2009). This collaboration further advanced their work on time series analysis.

During this period, Professor Saleh also made significant contributions to shrinkage estimation for autoregressive processes. In Saleh (1993a), he proposed shrinkage estimation for the parameters of an autoregressive Gaussian process. Additionally, Chen and Saleh (1993) explored the estimation of regression parameters when errors are autocorrelated, further expanding the scope of his research.

The ongoing collaboration between Professors Saleh and Koul culminated in their influential paper (Koul & Saleh, 1995), which has had a lasting impact on time series analysis. This partnership also inspired the PhD thesis of Dr. Soroush Alimoradi, titled "Some Contributions to L-Estimation in Linear Regression and Autoregression Models" (1995). Dr. Alimoradi's thesis led to two important publications, Alimoradi and Saleh (1998) and Saleh and Alimoradi (1999), which further enriched the field of estimation theory.

In 2009, Professor Saleh, in collaboration with Professors Fei Tang and B. M. Golam Kibria, developed shrinkage estimation methods for the parameters of a simple linear regression model with autocorrelated errors (Saleh et al., 2009).

### *1.6.6 Robust Statistics*

Robust statistics[2] are methods designed to maintain their properties even when the underlying distributional assumptions are violated. These methods have been developed for a wide range of problems, including estimating location, scale, and regression parameters. One key motivation for robust statistics is to create methods that are not overly influenced by outliers. Another is to ensure good performance in the presence of small deviations from parametric assumptions.

Professor Saleh undoubtedly was among the group of top statisticians in the world for his significant contributions in the theory of robust statistics. Drawing from Khan (2009), in 1974, Professor Saleh was invited to the 7th European Statistics Meeting in Prague, where he presented a paper titled "Behrens-Fisher Problem for the Exponential Distribution." This work challenged the traditional reliance on the normal theory and aimed to engage with Professor Jaroslav Hájek, a prominent figure in robust statistics. Unfortunately, Professor Hájek passed away just days before the meeting.

Nine years later, in 1983, Professor Saleh was invited by Professor Jana Jurečková to attend the Meeting of European Statisticians in Palermo, Sicily, Italy. This marked the beginning of a long-standing collaboration. In 1984, Professor Jurečková visited Carleton University, where she and Professor Saleh worked on regression quantiles and Stein estimation. Their partnership led to several influential

[2] https://en.wikipedia.org/wiki/Robust_statistics

papers on robustified Stein-type estimation, regression quantiles, and enhanced L-estimation for linear models, including Jurečková and Saleh (1990) and Saleh et al. (1991).

Professor Saleh's contributions to robust statistics span a variety of estimators and methods:

- **M-estimators**: (Saleh & Sen, 1985c; Sen & Saleh, 1987)
- **R-estimators**: (Saleh & Sen, 1986b)
- **L-estimators**: (Saleh & Sen, 1987b)
- **Shrinkage and preliminary test M-estimators** (Saleh & Sen, 1988)
- **Nonparametric and robust shrinkage estimation in multivariate linear models** (Saleh & Sen, 1987a)
- **Robust estimation of the median of a survey population parameter** (Bassett & Saleh, 1994)

The work of Saleh and Shiraishi (1989) further advanced the field by focusing on R- and M-estimators for regression parameters under specific constraints. Additionally, Saleh (2002b) explored simultaneous R-estimation of multivariate locations, addressing subspace hypotheses and data analysis from statistical foundations.

In his later years, Professor Saleh collaborated with his son, Professor Resve A. Saleh, from the Department of Electrical and Computer Engineering at the University of British Columbia, Vancouver, Canada, on machine learning. Their joint research focused on advancing robust statistical methods within the field. Together, they published a paper available on arXiv (Saleh & Saleh, 2021, 2024) and authored two books. One of these books was published in 2022 (Saleh et al., 2022), while the other, titled "Fundamentals of Robust Machine Learning: Handling Outliers and Anomalies in Data Science," was published in 2025 (Saleh et al., 2025).

### *1.6.7 Models with Measurement Error*

In many applied fields, response variables are modeled using one or more covariates that are not directly observable. Instead, researchers use substitute variables that can be observed to represent the original covariates. However, since these substitute variables contain measurement errors, such models are referred to as measurement error models (Carriquiry, 2001).

Professor Saleh had a PhD student, H. M. Kim, with whom he published several papers on the application of preliminary tests and Stein-type estimators in these models including Saleh and Kim (2002), Kim and Saleh (2003a,b, 2005, 2008, 2011).

In 2010, Professor Saleh co-authored a paper with Professors Jurečová and Jan Picek on the application of rank tests in measurement error models (Jurečková et al., 2010). The extended version of this paper is then published in the work of Navrátil and Saleh (2016).

Additionally, in collaboration with Professor Jan Picek and also Professor Jan Kalina, he contributed to the development of nonparametric estimation methods for these models in Saleh et al. (2009, 2012)

The Theil–Sen estimator in a measurement error model was studied by Sen and Saleh (2010). Additionally, quasi-empirical Bayes approaches were applied to measurement error models in Saleh (2010).

Together with Professor Shalabh, he explored the use of ridge and Liu estimators in measurement error models, as documented in Saleh and Shalabh (2014a, 2014b) and Saleh and Shalabh (2020), respectively.

### *1.6.8 Elliptically Countoured Models*

As reported in Khan (2009), the conventional use of the normal model for errors/responses has been increasingly questioned, particularly when the population distribution is symmetric but exhibits heavier tails than the normal distribution. Additionally, the normal model fails to account for dependent but uncorrelated responses. In such situations, the multivariate Student-t distribution offers a more suitable model for the population. He was also an external expert examiner for Professor Shahjahan Khan's PhD thesis at the University of Western Ontario in 1992, which focused on "Prediction distributions of multivariate models with multivariate Student-t errors." This sparked Professor Saleh's interest in improved estimation under the Student-t model. In 1994, during his first visit to Australia, Professor Saleh spent two weeks at the University of Southern Queensland, where he initiated joint work with Professor Shahjahan Khan. Their first collaborative paper is Saleh and Khan (1995).

Professor Khan visited Professor Saleh at Carleton University in both 1996 and 2006 to further their collaboration, leading to the production of 10 papers on various preliminary test and Stein-type shrinkage estimations across different models, including simple, multiple, and multivariate regression models, as well as the parallelism model when the distribution of errors is Student-t including Saleh and Khan (1995) Khan and Saleh (1997, 1998, 2008) Saleh and Khan (2003). Professor Khan's PhD student, Zahirul Hoque has also continued to work in this area, drawing inspiration and guidance from Professor Saleh including Saleh et al. (2002); Khan et al. (2005).

Professor Saleh conducted extensive research utilizing the Student-t distribution. In collaboration with Dr. Baki Billah, he investigated the conflict between Preliminary test estimators induced by three large-sample tests in a regression model with Student-t errors, leading to a series of publications (Saleh & Billah, 1998; Saleh and Billah, 2000a, 2000b; Billah & Saleh, 2003). Later, Saleh and Kibria (2004) revisited this issue, focusing on its resolution within the ridge regression framework. Additionally, Professor Saleh coauthored a book titled Statistical Inference for Models with Multivariate t-Distributed Errors (Saleh et al., 2014).

The normal distribution has been generalized in two significant directions: first, as a special case of the exponential family, and second, as a spherically symmetric distribution. The class of spherically symmetric distributions can be defined and characterized in various equivalent ways; a comprehensive review of these is provided in Fang et al. (1990). Professor Saleh supervised his PhD student, S. M. M. Tabatabaey, in applying preliminary test estimation to regression models with spherically symmetric errors. Together with Professor B. M. Golam Kibria, they published two joint papers: Saleh et al. (2004) and Tabatabaey et al. (2004).

A spherical distribution is an extension of the standard multivariate normal distribution, i.e., $N(\mathbf{0}, \boldsymbol{I}_n)$, where $\mathbf{0} = (0, \ldots, 0)^\top$ is the mean vector and $\boldsymbol{I}_n$ is the identity variance–covariance matrix of dimension $n$. In contrast, an elliptical distribution generalizes $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is the mean vector and $\boldsymbol{\Sigma}$ is the variance–covariance matrix. Professor Saleh, in collaboration with Professor Kibria, investigated the estimation of the mean vector of a multivariate elliptically contoured distribution (Saleh & Kibria, 2010). Meanwhile, Professor Tabatabaey supervised a PhD student, Mohammad Arashi, and invited Professor Saleh to join the supervision team. Together, they extended Professor Tabatabaey's thesis to elliptically contoured distributions, leading to several significant contributions. They derived shrinkage estimators in a parallelism model with elliptically distributed errors (Arashi et al., 2010), studied the mathematical characteristics of improved estimators for the mean and variance components in elliptically contoured models (Arashi et al., 2011), developed shrinkage estimation methods for regression coefficients with elliptically contoured errors (Arashi et al., 2013), and explored Bayesian statistical inference for the Laplacian class of matrix-variate elliptically contoured models (Arashi et al., 2015).

### *1.6.9 Empirical Bayes Estimation*

Empirical Bayes methods are statistical inference procedures in which the prior probability distribution is estimated from the data. This approach contrasts with standard Bayesian methods, where the prior distribution is fixed before any data are observed. Professor Saleh initiated research in empirical Bayes estimation in collaboration with Professor Malay Ghosh from Florida University and Professor Sen, resulting in the publication Ghosh et al. (1989). He further applied this methodology with Dr. Abdunnabi M. Ali in Ali and Saleh (1991c). In the same year, his joint work with Professor Robert on point and interval estimation in the parallelism model appeared in Robert and Saleh (1991).

Professor Saleh's PhD student, Abdunnabi M. Ali, extensively explored the use of preliminary test and empirical Bayes estimators in various models, leading to a series of publications: Ali and Saleh (1990, 1991a, 1991b, 1991c), Saleh and Ali (1991).

During his visit to Bowling Green State University in 1992–1993, Professor Saleh collaborated with Professor Arjun K. Gupta on categorical data analysis. This collaboration yielded two significant publications: Gupta and Saleh (1996, 1997).

Expanding his Bayesian approach, Professor Saleh collaborated with Professor Mohamedou Ould Haye on Bayes estimation methods for autoregressive moving average (ARMA) models, incorporating vague prior information on the moving average component, which led to Haye and Saleh (2006). The quasi-empirical Bayes (QEB) methodology, introduced on page 157 of Saleh (2006), was applied to enhance meta-analysis. Additionally, in a joint work with Professor Patrick J. Farrell, Professor Saleh developed a quasi-empirical Bayes estimation for the parameters of an ARMA model with subspace restrictions (Farrell & Saleh, 2010).

### *1.6.10 Penalized Estimators*

Consider the multiple linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ design matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of error terms following multivariate normal distribution with mean vector $\mathbf{0}$ and variance–covariance matrix $\sigma^2 \boldsymbol{I}_n$.

The classical estimator, ordinary least squares (OLS) estimator, is defined as $\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$. This estimator is obtained by minimizing the squared loss function $L(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$. A penalized estimator is defined by minimizing a penalized loss function of the form $L(\boldsymbol{\beta}) + \lambda P_\lambda(\boldsymbol{\beta})$, where $\lambda$ is a tuning parameter and $P_\lambda(\boldsymbol{\beta})$ is a penalty function chosen based on the objective of the study. Ridge regression (Hoerl & Kennard, 1970) is a method of estimating the coefficients of multiple regression models in scenarios where the independent variables are highly correlated. It is obtained by considering the penalty function $P(\boldsymbol{\beta}) = \sum_{i=1}^{p} \beta_i^2$, which leads to the ridge estimator $\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

While Professor Kibria was a graduate student at Carleton University, Professor Saleh introduced him to ridge regression and suggested incorporating shrinkage estimation into the framework. This collaboration led to their first joint publication Saleh and Kibria (1993), which introduced the preliminary test ridge estimator. This work marked the beginning of a productive series of papers they co-authored in this field. Their collaborative work includes:

- Comparing different test statistics in preliminary test ridge estimators (Saleh & Kibria, 2003a)
- Preliminary test ridge regression estimators with Student-t errors (Saleh & Kibria, 2004)
- Shrinkage ridge estimator in nonparameteric models (Saleh & Kibria, 2004)
- Shrinkage ridge estimator in probit regression models (Kibria & Saleh, 2012)

- Shrinkage ridge estimator in probit regression models (Saleh & Kibria, 2013)

Professor Saleh's other significant contributions to ridge estimation include:

- Shrinkage ridge R-estimators in autoregressive models (Saleh, 2014; Saleh & Kibria, 2011)
- Shrinkage ridge estimators in measurement error models (Saleh & Shalabh, 2014a, 2014b)
- Shrinkage ridge estimators in autoregressive models (Saleh & Ghania, 2015)
- Shrinkage ridge estimators in non-orthogonal models (Saleh & Norouzirad, 2018)

In 2019, Professor Saleh, in collaboration with Professors Kibria and Arashi, published a comprehensive book (Saleh et al., 2019) on ridge regression and shrinkage methods. That same year, in a joint paper with Professors B. M. Golam Kibria and Florence George, he conducted a comparative study of LASSO, ridge regression, preliminary test, and Stein-type estimators for the sparse Gaussian regression model (Saleh et al., 2019).

Mina Norouzirad, a former PhD student of Professor Mohammad Arashi, had the opportunity to visit Professor Saleh at Carleton University during 2016–2017. During her visit, Professor Saleh shared his expertise and suggested applying shrinkage methods to LASSO estimation (Tibshirani, 1996), where the penalty function is defined as $P(\boldsymbol{\beta}) = \sum_{i=1}^{n} |\beta_i|$. This collaboration resulted in several publications, including:

- Shrinkage estimation through convex optimization (Norouzirad et al., 2015a)
- Restricted LASSO and Double Shrinking (Norouzirad et al., 2015b)
- Rank theory approach to ridge, LASSO, preliminary test, and Stein-type estimators: A comparative study (Saleh et al., 2018).

Professor Saleh, in collaboration with Dr. Enayetur Raheem, published a paper (Saleh & Raheem, 2015) where they proposed using the LASSO estimator in place of the OLS estimator within the classical shrinkage estimation framework. Their work focused on theoretical analysis, particularly in the orthogonal case. In another paper Raheem and Saleh (2015) compared the shrinkage estimator with several penalized estimators, including LASSO, adaptive LASSO (Zou, 2006), smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), and elastic net (Zou & Hastie, 2005), under subspace restrictions.

### *1.6.11 Additional Contributions*

In addition to his work in the core areas of statistical theory, Professor Saleh made numerous other contributions across a broad range of statistical topics. These contributions span a variety of specialized areas, reflecting his diverse expertise and

deep understanding of statistical theory and practice. Professor Saleh's work on distributions, transformations, and model selection, as well as his research into classical statistical problems such as the Behrens-Fisher problem, highlights his ability to tackle both fundamental and emerging challenges in statistics. Furthermore, his scholarly contributions extend beyond technical research, with notable work on the significance of statistical approaches by eminent statisticians and insightful book reviews. This subsection delves into these additional contributions.

#### 1.6.11.1 Distributions

During his master's and PhD studies, Professor Saleh worked on estimating parameters of the exponential distribution based on order statistics. Throughout his academic career, he expanded his research to include estimating parameters in several distributions under various scenarios and using different methods. Below is a list of his contributions in this area:

- Decomposition of a finite mixture of distributions by minimum chi-square (Saleh, 1981a)
- Estimation of the common scale parameter of two Pareto distributions in censored samples (Rohatgi & Saleh, 1987)
- Logistic distribution and its applications (Saleh, 1989)
- Comparison of location parameters of two exponential distributions when scale parameters are different and unknown (Nassrallah & Saleh, 1994)
- Statistical theory for the parameters with Student-t distribution (Saleh, 2008)
- Preliminary test estimation of the parameters of exponential and Pareto distributions for censored samples (Kibria & Saleh, 2010; Barranco-Chamorro et al., 2013)
- Methods of moments estimation in finite mixtures (Farrell et al., 2011)
- New estimators of distribution functions (Saleh & Ghania, 2016)
- Sample design and estimation of parameters of half logistic distribution using generalized ranked-set sampling (Adatia & Saleh, 2020)

Truncated Distributions

In 1971, Dr. Christine Scott defended her thesis, "On Robustness of Size of Student's t-Test," under the supervision of Professor Saleh. This research led to the publication of a paper (Scott & Saleh, 1975), which explored the effect of truncation on the size of the t-test, contributing to the understanding of robustness in statistical hypothesis testing.

Professor Saleh supervised M. A. Rahim's master's thesis in 1972, from which the paper "Distribution of the Sum of Variates from Truncated Discrete Populations" (Saleh & Rahim, 1972) was derived. However, his early work on truncated

distributions dates back to Saleh (1970), focusing on estimating parameters for the Poisson distribution. He also co-authored a paper on determining the moments of order statistics in the truncated exponential distribution with Professor Christine Scott and D. Bruce Junkins (Saleh et al., 1975).

#### 1.6.11.2 Transformations

The arcsine transformation[3] converts a binomial random variable into one that is approximately normal, with a variance that depends only weakly on the binomial parameter $p$. Specifically, if $X$ follows a binomial distribution with parameters $n$ and $p$, i.e., $X \sim \text{Bin}(n, p)$, then the arcsine transformation of $X$ is given by $Y = \sin^{-1}(X/n)$. The transformed variable $Y$ is approximately normally distributed with mean $\text{E}[Y] \approx \sin^{-1}(p)$ and variance $\text{var}(Y) \approx 1/(4n)$. Professor Saleh applied this transformation to binomial data and compared it with normal approximations in Saleh (1969b).

#### 1.6.11.3 Gini's Mean Difference

Let $X_1, \ldots, X_n$ be a set of samples from a distribution $F$ with sample mean $\bar{X}_n$. The statistic $G_n = (1/(n(n-1))) \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|$ is known as Gini's mean difference. Professor Saleh, in collaboration with D. Layne, investigated and derived the moments of $G_n$ for samples drawn from an exponential distribution (Saleh & Layne, 1972).

#### 1.6.11.4 Behrens–Fisher Problem

The Behrens–Fisher distribution, named after Sir Ronald Fisher and Walter-Ulrich Behrens, is a family of probability distributions that arises in the context of the Behrens–Fisher problem. This problem concerns statistical inference about the difference between the means of two normally distributed populations when the ratio of their variances is unknown, meaning it is not assumed that the variances are equal.

The Behrens–Fisher distribution describes the distribution of the random variable $d = T_1 \cos(\theta) - T_1 \sin(\theta)$ where $T_1$ and $T_2$ are independent random variables following Student-t distributions with degrees of freedom $\nu_1$ and $\nu_2$, respectively, and $\theta$ is a constant. The family of Behrens–Fisher distributions is thus parameterized by $\nu_1$, $\nu_2$, and $\theta$.

In 1974, Professor Saleh, in collaboration with Professor Mizan Rahman, proposed a probability density function for the D-statistic using an Appell function.

[3] https://demonstrations.wolfram.com/TheArcsineTransformationOfABinomialRandomVariable/

This approach facilitated the computation of tail probabilities for any given pair of degrees of freedom $(\nu_1, \nu_2)$, leading to their publication (Rahman & Saleh, 1974).

In the following year, Professor Saleh extended this work by publishing "Extended Tables of Critical Values of the Behrens–Fisher D-Statistic" (Saleh, 1975b). Additionally, in 1977, he presented the paper "Behrens–Fisher Problem for the Exponential Distribution in Censored and Uncensored Samples" at the Seventh Prague Conference, which was later published in Saleh (1977).

#### 1.6.11.5 Exponential Regression

"Exponential regression" is used to model situations in which growth begins slowly and then accelerates rapidly without bound or where decay begins rapidly and then slows down to get closer and closer to zero. In 1975, Professor Saleh, jointly with Himanshu Choudhry, published the paper "On fitting exponential regressions" (Saleh & Choudhry, 1975)

#### 1.6.11.6 Estimation of Correlation Coefficient

If the response variable $Y$ follows the structural model $Y = \theta E$, where $\theta$ is an element of a unitary group of transformations $G$ acting on the sample space $X$, and $E$ is an error variable with a density function $f$ relative to an invariant measure $m$ on $X$, Hag and Saleh (1983) derived the structural estimation for both the bivariate correlation coefficient and the intra-class correlation coefficient.

#### 1.6.11.7 Model Selection

In many experiments, an observation (response) can be classified as either a success or a nonsuccess. Such binomial processes or models frequently arise in fields such as clinical studies, biostatistics, engineering, and social sciences. The problem of selecting a binomial model associated with the largest probability of success has been formulated in two distinct ways: the indifference zone formulation (Sobel & Hyyett, 1957) and the subset selection approach (Gupta & Sobel, 1960). In a joint paper with Professor Wen-Tao Huang, Professor Saleh proposed a mixed-type randomized rule for selecting superior models from binomial models, demonstrating its superiority over classical selection rules (Huang & Saleh, 1987).

#### 1.6.11.8 Cox Model

Proportional hazard models are a type of survival model used in statistics. These models analyze the time until an event occurs and how it relates to one or more influencing factors (covariates). In 1992, Professor Saleh developed methods for estimation in these models under prior information (Saleh, 1992a).

#### 1.6.11.9 Confidence Interval

For the interval (or, in general, set) estimation of the multivariate normal mean vector, Saleh (2003) investigated preliminary test and Stein-type estimators based on U-statistics. He introduced re-centered confidence intervals using these estimators and derived asymptotic expressions for their coverage probabilities. In addition, Saleh and Robert (2002) provided a comprehensive review of re-centered confidence sets from statistical foundations.

#### 1.6.11.10 Scholarly Contributions on Eminent Statisticians

Professor Saleh made scholarly contributions by authoring scientific and biographical articles on several eminent statisticians, highlighting their groundbreaking research and lasting impact on the field of statistics. His works provide discussions on their contributions to statistical theory, methodology, and applications, offering a comprehensive understanding of their influence on modern statistical practices.

He wrote about several distinguished statisticians, including:

- Saleh (1993b) and Saleh (2012)
  Professor Qazi Motahar Hossain,[4] whose contributions to statistical education and research were instrumental in founding the Institute of Statistical Research and Training (ISRT) in 1964, where he served as its first director; he was born in Faridpur, Bangladesh, on July 30, 1897, and passed away on October 9, 1981.
- Saleh (1995)
  Professor Calyampudi Radhakrishna Rao (well known as C. R. Rao[5]), a pioneer in statistical estimation and information theory, made groundbreaking contributions over an 80-year research career, profoundly shaping modern statistics. Born in Karnataka, India, on September 10, 1920, he became one of the most influential statisticians of the twentieth century. His work laid the foundation for numerous statistical methods, including the Cramér–Rao bound and Rao–Blackwell theorem. He passed away on August 22, 2023, just two weeks before his 103rd birthday, leaving behind a remarkable legacy in statistical science.
- Saleh and Haq (1996) and Saleh (2002a)
  Professor Donald Alexander Stuart Fraser (well known as D. A. S. Fraser[6]) born on April 29, 1925, and passing away on December 21, 2020, at the age of 95, made extraordinary contributions to the theory and foundations of statistics, particularly in the area of likelihood inference. His exceptional work continues to shape the field of statistics, with a profound and lasting impact both in Canada and globally. Fraser's pioneering research and intellectual legacy have been

[4] https://isrt.ac.bd/people/dr-qazi-motahar-hussain/

[5] https://imstat.org/2023/09/30/obituary-c-r-rao-1920-2023/

[6] https://www.statistics.utoronto.ca/news/eulogy-donald-s-fraser-1925-2020

instrumental in advancing statistical theory, and his influence remains vital in statistical science worldwide.

- Saleh (1998)
  Professor Ivan Peter Fellegi (well known as I. P. Fellegi[7]), born on June 22, 1935, is widely recognized for his groundbreaking contributions to statistics and public service. Two of Professor Fellegi's most influential research papers, "A Theory for Record Linkage" (1969) and "A Systematic Approach to Automatic Edit and Imputation" (1973), earned the distinction of being included among the 19 landmark papers in survey sampling published between 1934 and 1990. These papers had a profound impact on the fields of data linkage and statistical imputation, laying the groundwork for advancements in data processing and analysis. Professor Fellegi's extensive contributions to statistical methodology, particularly in survey sampling and official statistics, have garnered international recognition.
- Saleh (1999)
  Bradley Efron[8] (born on May 24, 1938) is especially known for proposing the bootstrap resampling technique, which has revolutionized the field of statistics and significantly impacted virtually every area of statistical application. The bootstrap was one of the first computer-intensive statistical techniques, replacing traditional algebraic derivations with data-based computer simulations.

Through these articles, Professor Saleh not only documented the remarkable achievements of these statisticians but also provided valuable insights into their intellectual journeys, shaping the appreciation of their work within the statistical community.

#### 1.6.11.11 Book Review

We found a book review paper written by Professor Saleh for the book titled "Meta-Analysis Methods for Health and Experimental Studies," authored by Professor Shahjahan Khan and published by Springer in 2020. This review was later published in the *Journal of the Royal Statistical Society, Series A (Statistics in Society)* in 2022 (Saleh, 2022).

## 1.7 Final Reflections

Professor A. K. Md. Ehsanes Saleh has left an enduring legacy that continues to influence the field of statistics to this day. His groundbreaking work in areas such as shrinkage estimation, order statistics, robust statistics, and nonparametric

[7] https://en.wikipedia.org/wiki/Ivan_Fellegi

[8] https://en.wikipedia.org/wiki/Bradley_Efron

methods has left an indelible mark on both theoretical and applied statistical research. His innovative ability to bridge different statistical domains, particularly by combining robust statistics with shrinkage techniques, has inspired a new generation of statisticians and researchers. Through his influential papers, books, and mentorship, Professor Saleh's legacy continues to shape modern statistical theory and methodology. His work remains a cornerstone for researchers in statistical inference, time series analysis, and beyond, ensuring that his contributions will remain a vital reference for future generations of statisticians.

Professor Saleh was a giant in the field and widely recognized for his contributions to the field, receiving special honors at an international conference organized by North South University, Dhaka, Bangladesh and Carleton University, Ottawa, Canada in 2008. Several journals, including the *Journal of Statistical Research*, *Pakistan Journal of Statistics*, and *Journal of Statistical Studies*, published special volumes in his honor, reflecting his profound influence on statistical sciences.

**Acknowledgments** We want to express our heartfelt gratitude to Professor Resve A. Saleh, the son of Professor Saleh, for his unwavering support and invaluable contributions throughout the preparation of this paper. We also extend our sincere thanks to Professor Saleh's family for their continued support and for sharing his remarkable legacy with the world. Their commitment to preserving his intellectual and personal legacy has been truly inspiring.

Additionally, the work of Mina Norouzirad is funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UID/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications—NOVA Math). Also, Mohammad Arashi's work is based on the research supported in part by the Iran National Science Foundation (INSF) grant No. 4015320.

## References

Adatia, A., & Saleh, A. K. Md. E. (2009). Estimation of quantiles of uniform distribution using generalized ranked-set sampling. In M. Ahmad, M. Hanif, & H. A. Muttlak (Eds.), *Ranked Set Sampling* (pp. 145–160).

Adatia, A., & Saleh, A. K. Md. E. (2012). Optimum regression quantiles for the estimation and tests of hypothesis of the parameters of a simple regression model with logistic errors. *Aligarh Journal of Statistics, 32*, 1–9.

Adatia, A., & Saleh, A. K. Md. E. (2020). Sample design and estimation of parameters of half logistic distribution using generalized ranked-set sampling. *Journal of Statistical Theory and Applications, 19*(1), 109–117.

Ahmed, S. E., & Saleh, A. K. Md. E. (1989). Pooling multivariate data. *Journal of Statistical Computation and Simulation, 31*(3), 149–167.

Ahmed, S. E., & Saleh, A. K. Md. E. (1990). Estimation strategies for the intercept vector in a simple linear multivariate normal regression model. *Computational Statistics & Data Analysis, 10*(3), 193–206.

Ahmed, S. E., & Saleh, A. K. Md. E. (1993). Improved estimation for the component mean-vector. *Journal of the Japan Statistical Society, 23*(2), 145–159.

Ahmed, S. E., & Saleh, A. K. Md. E. (1999a). Improved nonparametric estimation of location vectors in multivariate regression models. *Journal of Nonparametric Statistics, 11*(1–3), 51–78.

Ahmed, S. E., & Saleh, A. K. Md. E. (1999b). Estimation of regression coefficients in an exponential regression model with censored observations. *Journal of the Japan Statistical Society, 29*(1), 55–64.

Ahmed, S. E., Saleh, A. K. Md. E., Volodin, A. I., & Volodin, I. N. (2007). Asymptotic expansion of the coverage probability of James–Stein estimators. *Theory of Probability & Its Applications, 51*(4), 683–695.

Ali, A. M., & Saleh, A. K. Md. E. (1990). Estimation of the mean vector of a multivariate normal distribution under symmetry. *Journal of Statistical Computation and Simulation, 35*(3–4), 209–226.

Ali, A. M., & Saleh, A. K. Md. E. (1991a). Asymptotic theory of simultaneous estimation for binomial means. *Statistica Sinica, 1*(1), 271–294.

Ali, A. M., & Saleh, A. K. Md. E. (1991b). Estimation of means and treatment effects in a one-way ANOVA model. *Soochow Journal of Mathematics, 17*(2), 287–309.

Ali, A. M., & Saleh, A. K. Md. E. (1991c). Preliminary test and empirical Bayes approach to shrinkage estimation of regression parameters. *Journal of the Japan Statistical Society, 21*(1), 22–30.

Ali, M. M., Umbach, D., & Saleh, A. K. Md. E. (1982). Small sample quantile estimation of exponential distribution using optimal spacings. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002), 44*(2), 135–142.

Ali, M. M., Umbach, D., & Saleh, A. K. Md. E. (1985). Tests of significance for the exponential distribution based on selected quantiles. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002), 47*(3), 310–318.

Ali, M. M., Umbach, D., & Saleh, A. K. Md. E. (1992). Estimating life functions of chi distribution using selected order statistics. *IIE Transactions, 24*(5), 88–98.

Ali, M. M., Umbach, D., Saleh, A. K. Md. E., & Hassanein, K. M. (1983). Estimating quantiles using optimally selected order statistics. *Communications in Statistics - Theory and Methods, 12*(19), 2261–2271.

Alimoradi, S., & Saleh, A. K. Md. E. (1998). On some L-estimation in linear regression models. In *Handbook of statistics: Order statistics: applications* (Vol. 17, pp. 237–280). Elsevier.

Arashi, M., Norouzirad, M., & Saleh, A. K. Md. E. (2014). Preliminary test approach in high-dimensional multivariate regression models. In *Proceedings of the 12th Iranian Statistical Conference*, Kermanshah, Iran.

Arashi, M., Saleh, A. K. Md. E., Nagar, D. K., & Tabatabaey, S. M. M. (2015). Bayesian statistical inference for Laplacian class of matrix variate elliptically contoured models. *Communications in Statistics - Theory and Methods, 44*(13), 2774–2787.

Arashi, M., Saleh, A. K. Md. E., & Tabatabaey, S. M. M. (2010). Estimation of parameters of parallelism model with elliptically distributed errors. *Metrika, 71*, 79–100.

Arashi, M., Saleh, A. K. Md. E., & Tabatabaey, S. M. M. (2011). On mathematical characteristics of some improved estimators of the mean and variance components in elliptically contoured models. *Journal of the Iranian Statistical Society, 10*(2), 237–266.

Arashi, M., Saleh, A. K. Md. E., & Tabatabaey, S. M. M. (2013). Regression model with elliptically contoured errors. *Statistics – A Journal of Theoretical and Applied Statistics, 47*(6), 1266–1284.

Barranco-Chamorro, I., Kibria, B. M. G., & Saleh, A. K. Md. E. (2013). Erratum to corrections and complements to: Preliminary test estimation of the parameters of exponential and Pareto distributions for censored samples. *Statistical Papers, 54*(1), 251–254.

Bassett, G. W. Jr., & Saleh, A. K. Md. Ehsanes (1994). L1 estimation of the median of a survey population. *Journal of Nonparametric Statistics, 3*(3–4), 277–283.

Billah, Md. B., & Saleh, A. K. Md. E. (2003). Performance of the PTE based on the conflicting W, LR, and LM tests in regression model. In N. Balakrishnan (Ed.), *Advances on methodological and applied aspects of probability and statistics*. CRC Press.

Carriquiry, A. L. (2001). Measurement error models. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 9435–9442). Pergamon.

Chen, E. J., & Saleh, A. K. Md. E. (1993). Estimation of regression parameters when errors are autocorrelated. In K. Matusita, M. L. Puri, & T. Hayakawa (Eds.), *Statistical Sciences and Data Analysis: Proceedings of the 3rd Pacific Area Statistical Conference* (pp. 61–76).

Chiou, P. C., & Saleh, A. K. Md. E. (2002). Preliminary test confidence sets for the mean of a multivariate normal distribution. *Journal of Propagations in Probability and Statistics, 2*(2), 177–189.

Csorgo, M., Dawson, D. A., Rao, J. N. K., & Saleh, A. K. Md. E. (1981). *Proceedings of the International Symposium on Statistics and Related Topics, 5–8 May 1980*. North-Holland Publishers.

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Springer.

Farrell, P. J., & Saleh, A. K. Md. E. (2010). Quasi-empirical Bayes estimation of the parameters of an ARMA(p1, p2) model: Subspace restrictions. *Pakistan Journal of Statistics, 26*(1), 135–149.

Farrell, P. J., Saleh, A. K. Md. E., & Zhang, Z. (2011). Methods of moments estimation in finite mixtures. *Sankhyā: The Indian Journal of Statistics, 73*(2), 218–230.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association, 96*, 1348–1360.

Ghosh, M., Saleh, A. K. Md. E., & Sen, P. K. (1989). Empirical Bayes subset estimation in regression models. *Statistics and Decisions, 7*, 15–35.

Govindarajulu, Z., Huang, J. S., & Saleh, A. K. Md. E. (1975). Expected value of the spacings between order statistics. In G. P. Patil, S. Kotz, & J. K. Ord (Eds.), *A modern course on statistical distributions in scientific work* (pp. 143–147). Springer.

Gupta, A. K., & Saleh, A. K. Md. E. (1996). Estimation of multinomial probabilities under a model constraint. *Journal of Multivariate Analysis, 58*(2), 151–161.

Gupta, A. K., & Saleh, A. K. Md. E. (1997). Estimating odds-ratio: Homogeneity constraints. *Journal of the Italian Statistical Society, 6*, 67–81.

Gupta, S. S., & Sobel, M. (1960). Selecting a subset containing the best of several Binomial populations. In I. Olkin (Ed.) *Contributions to probability and statistics* (pp. 224–248). Standford University Press.

Hag, M. S., & Saleh, A. K. Md. E. (1983). On the structural estimation of certain correlation coefficients. *Statistische Hefte, 25*(1), 211–218.

Han, C. P., & Bancroft, T. A. (1968). On Pooling Means when Variance is Unknown. *Journal of the American Statistical Association, 63*(324), 1333–1342.

Harter, H. L. (1960). Tables of Range and Studentized Range. *The Annals of Mathematical Statistics, 31*(4), 1122–1147.

Hassanein, K. M., Brown, E. F., & Saleh, A. K. Md. E. (1994). Estimation of the survivor function, hazard function, and quantile-function for the extreme-value distribution based on selected order statistics. *Pakistan Journal of Statistics, 10*(1), 55–66.

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1984). Quantile estimates in complete and censored samples from extreme-value and Weibull distributions. *IEEE Transactions on Reliability, R-33*(5), 370–373.

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1985). Best linear unbiased quantile estimators of the logistic distribution using order statistics. *Journal of Statistical Computation and Simulation, 23*(1–2), 123–131.

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1986a). Best linear unbiased estimators for normal distribution quantiles for small sizes up to 20. *IEEE Transactions on Reliability, 35*(3), 327–329.

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1986b). Estimation and testing of quantiles of the extreme-value distribution. *Journal of Statistical Planning and Inference, 14*(2–3), 389–400.

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1986c). Small sample quantile estimation of the normal distribution using optimal order statistics. *Journal of Information and Optimization Sciences, 7*(2), 209–225.

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1990). Large sample estimation of the location and scale parameters of the chi-distribution using optimal selected order statistics. In *Modeling and Simulation, Proceedings of the Annual Pittsburgh Conference* (Vol. 21, pp. 263–270).

Hassanein, K. M., Saleh, A. K. Md. E., & Brown, E. F. (1991). Estimating survivor function using optimally selected order statistics. *Metrika, 38*, 99–115.

Haye, M. O., & Saleh, A. K. Md. E. (2006). Quasi-empirical Bayes methods of estimation in ARMA(p, q) models with vague prior information on MA(q). *Calcutta Statistical Association Bulletin, 58*(1–2), 53–76.

Hodges Jr., J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests, *The Annals of Mathematical and Statistics, 34*(2), 598–611.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Huang, W.-T., & Saleh, A. K. Md. E. (1985). Nonparametric randomized selection procedures for location parameters. *Bulletin of Institute of Mathematics, Academic Sinica, 13*(4), 297–316.

Huang, W.-T., & Saleh, A. K. Md. E. (1987). On a mixed-type randomized rule for selecting superior binomial models. *Biometrical Journal, 29*(7), 857–863.

James, W., & Stein, C. (1961). Estimation with quadratic loss, In J. Neyman (Ed.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Vol. 4.1, pp. 361–379).

Ji, P., Ahluwala, R., Saleh, A. K. Md. E., & Rohatgi, V. K. (1990). Locally most powerful tests for the two-sample problem when the combined sample is Type II censored. *Communications in Statistics - Theory and Methods, 19*(6), 2337–2355.

Jurečková, J., Picek, J., & Saleh, A. K. Md. E. (2010). Rank tests and regression rank score tests in measurement error models. *Computational Statistics & Data Analysis, 54*(12), 3108–3120.

Jurečková, J., & Saleh, A. K. Md. E. (1990). Robustified version of Stein's multivariate location estimation. *Statistics & Probability Letters, 9*(4), 375–380.

Khan, S. (2009). Scientific and scholarly contributions of Distinguished Research and Emeritus Professor A K Md Ehsanes Saleh. *International Journal of Statistical Sciences, 9*, 7–31.

Khan, S., Hoque, Z., & Saleh, A. K. Md. E. (2005). Estimation of the intercept parameter for linear regression model with uncertain non-sample prior information. *Statistical Papers, 46*(3), 379–395.

Khan, S., & Saleh, A. K. Md. E. (1997). Shrinkage pre-test estimator of the intercept parameter for a regression model with multivariate Student-t errors. *Biometrical Journal, 39*(2), 131–147.

Khan, S., & Saleh, A. K. Md. E. (1998). Comparison of estimators of means based on p-samples from multivariate student-t populations. *Communications in Statistics - Theory & Methods, 27*(1), 193–210.

Khan, S., & Saleh, A. K. Md. E. (2001). On the comparison of the pre-test and shrinkage estimators for the univariate normal mean. *Statistical Papers, 42*, 451–473.

Kibria, B. M. G. (2009). A. K. Md. Ehsanes Saleh: A luminous professor of statistics, *International Journal of Statistical Sciences, 9*, 1–5.

Khan, S., & Saleh, A. K. Md. E. (2008). Estimation of slope for linear regression model with uncertain prior information and Student-t error. *Communications in Statistics - Theory and Methods, 37*(16), 2564–2581.

Kibria, B. M. G., & Saleh, A. K. Md. E. (2004). Performance of positive rule estimator in the ill-conditioned Gaussian regression model. *Calcutta Statistical Association Bulletin, 55*(3–4), 209–240.

Kibria, B. M. G., & Saleh, A. K. Md. E. (2006). Optimum critical value for pre-test estimator. *Communications in Statistics - Simulation and Computation, 35*(2), 309–319.

Kibria, B. M. G., & Saleh, A. K. Md. E. (2010). Preliminary test estimation of the parameters of exponential and Pareto distributions for censored samples. *Statistical Papers, 51*(4), 757–773.

Kibria, B. M. G., & Saleh, A. K. Md. E. (2012). Improving the estimators of the parameters of a probit regression model: A ridge regression approach. *Journal of Statistical Planning and Inference, 142*(6), 1421–1435.

Kim, H. M., & Saleh, A. K. Md. E. (2003a). Preliminary test estimators of the parameters of simple linear models with measurement error. *Metrika, 57*, 223–251.

Kim, H. M., & Saleh, A. K. Md. E. (2003b). Improved estimation of regression parameters in measurement error models. *Calcutta Statistical Association Bulletin, 54*(3–4), 143–166.

Kim, H. M., & Saleh, A. K. Md. E. (2005). Improved estimation of regression parameters in measurement error models. *Journal of Multivariate Analysis, 95*(2), 273–300.

Kim, H. M., & Saleh, A. K. Md. E. (2008). Prediction of finite population total in measurement error models. In S. Shalabh & C. Heumann (Eds.), *Recent advances in linear models and related areas* (pp. 79–93). Physica-Verlag HD.

Kim, H. M., & Saleh, A. K. Md. E. (2011). Several simple linear models with measurement errors: Parallelism hypothesis. *Journal of Statistical Research, 45*(2), 155–173.

Konno, Y., Kubokawa, T., & Saleh, A. K. Md. E. (1997). Shrinkage estimators in a mixed MANOVA and GMANOVA model. *Statistics and Risk Modeling, 15*(1), 37–49.

Koul, H. L., & Saleh, A. K. Md. E. (1993). R-estimation of the parameters of autoregressive [AR(p)] models. *The Annals of Statistics, 21*(1), 534–551.

Koul, H. L., & Saleh, A. K. Md. E. (1995). Autoregression quantiles and related rank-score process. *Annals of Statistics, 23*(2), 670–689.

Kubokawa, T., Robert, C. P., & Saleh, A. K. Md. E. (1991). Robust estimation of common regression coefficients under spherical symmetry. *Annals of the Institute of Statistical Mathematics, 43*(4), 677–688.

Kubokawa, T., Robert, C. P., & Saleh, A. K. Md. E. (1992). Empirical Bayes estimation of the covariance matrix of a normal distribution with unknown mean under an entropy loss. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002), 54*(3), 402–410.

Kubokawa, T., & Saleh, A. K. Md. E. (1990). Sequential shrinkage estimation for a coefficient matrix in a multivariate regression model. *Journal of the Japan Statistical Society, 20*, 33–42.

Kubokawa, T., & Saleh, A. K. Md. E. (1994). Two-stage point estimation with a shrinkage stopping rule. *Metrika, 41*(1), 293–306.

Kubokawa, T., Saleh, A. K. Md. E., & Konno, Y. (2000). Bayes, minimax, and nonnegative estimators of variance components under Kullback-Leibler loss. *Journal of Statistical Planning and Inference, 86*(1), 201–214.

Kubokawa, T., Saleh, A. K. Md. E., & Morita, K. (1992). Improving on MLE of coefficient matrix in a growth curve model. *Journal of Statistical Planning and Inference, 31*(3), 169–177.

Lambert, A., Saleh, A. K. Md. E., & Sen, P. K. (1984). On least squares estimation of intercept after a preliminary test on parallelism of regression lines. *Communications in Statistics - Theory and Methods, 14*(4), 793–807.

Lambert, A., Saleh, A. K. Md. E., & Sen, P. K. (1985). Tests of homogeneity of intercepts after a preliminary test. *Communications in Statistics - Theory and Methods, 14*(9), 2243–2259.

Matin, M. A., & Saleh, A. K. Md. E. (2006). Small-sample properties of some improved estimators in logistic regression model with skew-normally distributed explanatory variables. *Journal of Statistical Research, 40*(1), 1–21.

Nassrallah, B., & Saleh, A. K. Md. E. (1994). Comparison of location parameters of two exponential distributions when scale parameters are different and unknown. *Statistical Papers, 35*(1), 57–69.

Navrátil, R., & Saleh, A. K. Md. E. (2016). Aligned rank tests in measurement error model. *Applied Mathematics, 61*(1), 47–59.

Norouzirad, M., Arashi, M., & Saleh, A. K. Md. E. (2015a). Shrinkage estimation through convex optimization. In *8th International Conference of the Iranian Society of Operations Research*. Ferdowsi University of Mashhad.

Norouzirad, M., Arashi, M., & Saleh, A. K. Md. E. (2015b). Restricted LASSO and double shrinking. arXiv preprint arXiv:1505.02913.

Padmanabhan, A. R., Puri, M. L., & Saleh, A. K. Md. E. (1981). Non-parametric tests of equality against ordered alternatives in the case of skew data with a biomedical application. In M. Csorgo, D. A. Dawson, J. N. K. Rao, & A. K. Md. E. Saleh (Eds.), *Statistics and Related*

*Topics: Proceedings of the International Symposium on Statistics and Related Topics held in Ottawa, May 5–7, 1980* (pp. 237–283). North-Holland Publishers.

Raheem, E., & Saleh, A. K. Md. E. (2015). Penalty, shrinkage, and preliminary test estimators under full model hypothesis. arXiv preprint arXiv:1503.06910.

Rahman, M., & Saleh, A. K. Md. E. (1974). Explicit form of the distribution of the Behrens–Fisher D–Statistic. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(1), 54–60.

Robert, C. P., & Saleh, A. K. Md. E. (1991). Point estimation and confidence set estimation in a parallelism model: An empirical Bayes approach. *Annals of Economics and Statistics, 23*, 65–89.

Rohatgi, V. K., & Saleh, A. K. Md. E. (1987). Estimation of the common scale parameter of two Pareto distributions in censored samples. *Naval Research Logistics, 34*(2), 235–238.

Rohatgi, V. K., & Saleh, A. K. Md. E. (1988). A class of distributions connected to order statistics with nonintegral sample size. *Communications in Statistics - Theory and Methods, 17*(6), 2005–2012.

Rohatgi, V. K., & Saleh, A. K. Md. E. (2015). *An Introduction to probability and statistics* (3rd ed.). John Wiley & Sons

Saleh, A. K. Md. E. (1966). Estimation of the parameters of the exponential distribution based on optimum order statistics in censored samples. *The Annals of Mathematical Statistics, 37*(6), 1717–1735.

Saleh, A. K. Md. E. (1967). Determination of exact optimum order statistics for estimating the parameters of exponential distribution from censored samples. *Technometrics, 9*(2), 279–292.

Saleh, A. K. Md. E. (1969a). Asymptotically most powerful rank tests of regression parameters from grouped data. *Journal of Statistical Research, 3*(2), 1–17.

Saleh, A. K. Md. E. (1969b). Arc-sin transformations and normal approximations of binomial data. *Journal of Statistical Research, 2*, 36–39.

Saleh, A. K. Md. E. (1970). Distribution of sum of independent truncated Poisson variates. *Journal of Statistical Research, 4*(1), 23–25.

Saleh, A. K. Md. E. (1974). *Proceedings of the 3-day symposium on Statistics and Related Topics, 24–26 October 1974*. Carleton Mathematics Series, No. 12.

Saleh, A. K. Md. E. (1975a). Characterization of distributions using spacings between order statistics. In G. P. Patil (Ed.), *A modern course on statistical distribution in scientific work* (pp. 143–147).

Saleh, A. K. Md. E. (1975b). Extended tables of critical values of Behrens-Fisher D-statistic. *Journal of Statistical Research, 8*, 29–51.

Saleh, A. K. Md. E. (1976a). Characterization of distributions using expected spacing between consecutive order statistics. *Journal of Statistical Research, 10*(1), 1–13.

Saleh, A. K. Md. E. (1976b). Hodges-Lehmann estimate of the location parameter in censored samples. *Annals of the Institute of Statistical Mathematics, 28*(1), 235–247.

Saleh, A. K. Md. E. (1977). Behrens-Fisher problem for the exponential distribution in censored and uncensored samples. In *Transactions of the Seventh Prague Conference and 1974 European Meetings* (pp. 465–472). Czech Academy of Sciences.

Saleh, A. K. Md. E. (1981a). Decomposition of finite mixture of distributions by minimum chi square. *Aligarh Journal of Statistics, 1*(2), 86–97.

Saleh, A. K. Md. E. (1981b). Estimating quantiles of exponential distribution. In M. Csorgo, D. A. Dawson, J. N. K. Rao and A. K. Md. E. Saleh (Eds.) *Statistics and Related Topics: Proceedings of the International Symposium on Statistics and Related Topics held in Ottawa, May 5–7, 1980* (pp. 145–151). North-Holland Publishers.

Saleh, A. K. Md. E. (1989). Logistic distribution and its applications. *Journal of Statistical Studies, 9*, 1–26.

Saleh, A. K. Md. E. (1991). Comments. *Journal of Nonparametric Statistics, 1*(1–2), 61–64.

Saleh, A. K. Md. E. (1992a). Estimation under uncertain prior information on the Cox Proportional Hazard Model. In P. K. Sen, & I. A. Salama (Eds.), *Proceedings of Order and Nonparametrics: Theory and Application*. Elsevier.

Saleh, A. K. Md. E. (1992b). *Nonparametric Statistics and Related Topics: Proceedings of the International Symposium on Nonparametric Statistics and Related Topics, Ottawa, Canada, 5–8 May, 1991*. North-Holland Publishers.

Saleh, A. K. Md. E. (1993a). On shrinkage estimation of the parameters of an autoregressive Gaussian process. *Theory of Probability & Its Applications, 37*(2), 250–260.

Saleh, A. K. Md. E. (1993b). Q. M. Hosain (1897–1981). *Journal of Statistical Research, 28*(1 & 2), 223–236.

Saleh, A. K. Md. E. (1995). C. R. Rao: A living legend in statistics. *Journal of Statistical Research, 29*(2), 93–98.

Saleh, A. K. Md. E. (1998). Ivan P. Fellegi—a world class official statistician. *Journal of Statistical Research, 32*(2), 113–115.

Saleh, A. K. Md. E. (1999). Bradley Efron – A world-class statistician. *Journal of Statistical Research, 33*(2), 114–115.

Saleh, A. K. Md. E. (2002a). *Data analysis from statistical foundations (a Festschrift in Honour of D.A.S. Fraser)*. Nova Science Publishers.

Saleh, A. K. Md. E. (2002b). Simultaneous R-estimation of multivariate locations: Subspace hypothesis and data analysis from statistical foundations. In *Statistical foundations* (pp. 179–194). Nova Science Publishers.

Saleh, A. K. Md. E. (2003). Asymptotic properties of Stein-type confidence sets: A U-statistics approach. *Sankhyā: The Indian Journal of Statistics (2003–2007), 65*(2), 411–421

Saleh, A. K. Md. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. Wiley & Sons.

Saleh, A. K. Md. E. (2008). Statistical theory for the parameters with Student's t-distribution. *Journal of Statistical Research, 42*(2), 47–68.

Saleh, A. K. Md. E. (2010). Quasi-empirical Bayes modeling of measurement error models and R-estimation of the regression parameters. *Journal of Statistical Research, 44*(2), 293–313.

Saleh, A. K. Md. E. (2012). On the paper "A talk on scope of statistics and statisticians" in Pakistan, by Q. M. Hossain. In: *Collected volume of statistical works of Q. M. Hossain with commentaries* (pp. 107–109).

Saleh, A. K. Md. E. (2014). Ridge autoregression R-estimation: Subspace restriction. In S. Lahiri, A. Schick, A. SenGupta, & T. Sriram (Eds.), *Contemporary developments in statistical theory* (pp. 119–136). Springer.

Saleh, A. K. Md. E. (2022). Book reviews: Meta-analysis methods for health and experimental studies. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 185*(2), 737–738.

Saleh, A. K. Md. E., & Ahsanullah, M. (1970). Optimum allocation of quantities in disjoint intervals for the BLUE of the parameters of exponential distribution when the sample is censored in the middle. *Naval Research Logistics, 17*(3), 331–349.

Saleh, A. K. Md. E., & Ahsanullah, M. (1972). Estimation of intercept in linear regression model with one independent variable after a preliminary test on regression coefficient. *International Statistical Reviews, 40*(2), 139–145.

Saleh, A. K. Md. E., Akritas, M., & Sen, P. K. (1985). Nonparametric Estimation of Intercepts after a preliminary test on parallelism of several regression lines. In P. K. Sen (Ed.), *Biostatistics: Statistics in biomedical, public health and environmental sciences* (pp. 221–235). North-Holland Publishers.

Saleh, A. K. Md. E., & Ali, A. M. (1991). On preliminary test and empirical Bayes estimator of finite population means. *Journal of Statistical Studies, 11*, 1–27.

Saleh, A. K. Md. E., & Ali, M. M. (1966). Asymptotic optimum quantiles for the estimation of parameters of the negative exponential distribution. *The Annals of Mathematical Statistics, 37*(1), 143–151.

Saleh, A. K. Md. E., Ali, M. M., & Umbach, D. (1983a). Estimating the quantile function of a location-scale family of distributions based on few selected order statistics. *Journal of Statistical Planning and Inference, 8*(1), 75–86.

Saleh, A. K. Md. E., Ali, M. M., & Umbach, D. (1983b). Simplified estimation of the parameters of exponential distribution from samples censored in the middle. *Communications in Statistics - Simulation and Computation, 12*(5), 609–627.

Saleh, A. K. Md. E., Ali, M. M., & Umbach, D. (1984). Tests of significance using selected sample quantiles. *Statistics & Probability Letters, 2*(5), 295–297.

Saleh, A. K. Md. E., Ali, M. M., & Umbach, D. (1985). Large sample estimation of Pareto quantiles using selected order statistics. *Metrika, 32*, 49–56.

Saleh, A. K. Md. E., & Alimoradi, S. (1999). On some L-estimation in auto-regression models. *Journal of Statistical Research, 32*, 85–111.

Saleh, A. K. Md. E., Arashi, M., & Tabatabaey, S. M. M. (2014). *Statistical inference for models with multivariate t-distributed errors*. John Wiley & Sons.

Saleh, A. K. Md. E., Arashi, M., & Kibria, B. M. G. (2019) *Theory of ridge regression estimation with applications*. John Wiley & Sons.

Saleh, A. K. Md. E., Arashi, M., Norouzirad, M., & Kibria, B. M. G. (2017). On shrinkage and selection of treatments ANOVA model. *Journal of Statistical Research, 51*(2), 165–191.

Saleh, A. K. Md. E., Arashi, M., Saleh, R. A., & Norouzirad, M. (2022). *Rank-based methods for shrinkage and selection: With application to machine learning*. John Wiley and Sons.

Saleh, A. K. Md. E., & Billah, B. (1998). Conflict between pre-test estimators induced by three large sample tests under a regression model with student t error. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(4), 593–606.

Saleh, A. K. Md. E., & Billah, B. (2000a). Performance of three large sample tests in the actual formation of the pretest estimators for regression models in the t-error. *Journal of Applied Statistical Science, 9*(3), 237–250.

Saleh, A. K. Md. E., & Billah, B. (2000b). Performance of the PTE based on the conflicting W, LR, and LM tests in regression models. In N. Balakrishnan (Ed.), *Proceedings of the International Indian Statistical Association (IISA) Conference* (pp. 275–285). McMaster University. Held October 11–15, 1998.

Saleh, A. K. Md. E., & Choudhry, G. H. (1975). On fitting exponential regressions. *Statistical Papers, 16*, 213–222.

Saleh, A. K. Md. E., & Chen, G. (2000). Strong consistency and asymptotic normality of regression quantiles in linear models. *Journal of Anhui University Natural Sciences Edition, 24*, 1–17.

Saleh, A. K. Md. E., & Chen, G. (2006). Moment and rank estimation in structural errors in variable models. *Journal of Chinese Statistical Association, 43*, 105–126.

Saleh, A. K. Md. E., & Dionne, J. P. (1977). On a further generalization of Savage test. *Communications in Statistics, 6*(12), 1213–1221.

Saleh, A. K. Md. E., & Ghania, A. F. (2015). Ridge autoregression estimation: LS method. *Communications in Statistics - Theory and Methods, 44*(15), 3303–3320.

Saleh, A. K. Md. E., & Ghania, A. F. (2016). New estimators of distribution functions. *Communications in Statistics - Theory and Methods, 45*(11), 3145–3157.

Saleh, A. K. Md. E., Gupta, A. K., & Sen, P. K. (1989). Improved estimation in contingency table: Independence structure. *Journal of the American Statistical Association, 84*(406), 525–532.

Saleh, A. K. Md. E., & Haq, M. S. (1996). Donald A.S. Fraser. *Journal of Statistical Research, 30*(2), 121–123.

Saleh, A. K. Md. E., & Hassanein, K. M. (1986a). On various F-estimation in a parallelism problem. *Soochow Journal of Mathematics, 12*, 83–94.

Saleh, A. K. Md. E., & Hassanein, K. M. (1986b). Testing equality of location parameters and quantiles of ($s \geq 2$) location-scale distributions. *Bulletin of the Institute of Mathematics Academia Sinica, 13*(1), 39–49.

Saleh, A. K. Md. E., Hassanein, K. M., & Ali, M. M. (1988a). Estimation and testing of hypotheses about the quantile function of the normal distribution. *Journal of Information and Optimization Sciences, 9*(1), 85–98.

Saleh, A. K. Md. E., Hassanein, K. M., & Ali, M. M. (1988b). Linear estimation and tests of quantiles of logistic distribution based on selected order statistics. *Journal of Information and Optimization Sciences, 9*(3), 445–464.

Saleh, A. K. Md. E., Hassanein, K. M., & Brown, E. F. (1985). Optimum spacings for the joint estimation and tests of hypothesis of location and scale parameters of Cauchy distribution. *Communications in Statistics - Theory and Methods, 14*(1), 247–254.

Saleh, A. K. Md. E., Hassanein, K. M., & Brown, E. F. (1994a). Joint estimation and tests of hypothesis of regression and scale parameters in a simple linear model with Cauchy errors based on a few regression quantiles. *Journal of Nonparametric Statistics, 3*(3–4), 309–315.

Saleh, A. K. Md. E., Hassanein, K. M., & Brown, E. F. (1994b). Optimum spacings, ABLUE and tests of hypothesis in simple linear model with extreme value errors. *Journal of Statistical Research, 28*, 53–64.

Saleh, A. K. Md. E., Hassanein, K. M., Hassanein, R. S., & Kim, H. M. (2006). Quasi-empirical Bayes methodology for improving meta-analysis. *Journal of Biopharmaceutical Statistics, 16*(1), 77–90.

Saleh, A. K. Md. E., Jurečková, J., & Sen, P. K. (1991). Regression quantiles and improved L-estimation in linear models. In R. R. Bahadur (Ed.), *Probability, statistics and design of experiments* (pp. 405–418).

Saleh, A. K. Md. E., & Khan, B. U. (2006). Estimation of regression parameters; A two-samples problem. *Journal of Statistical Theory and Applications, 4*(1), 91–107.

Saleh, A. K. Md. E., & Khan, B. U. (2007a). Improved estimation of regression parameters when the regression lines are parallel. *Journal of Applied Probability and Statistics, 1*(1), 1–13.

Saleh, A. K. Md. E., & Khan, B. U. (2007b). MSE comparisons of classical versus modified estimators: Parallelism restrictions. *Journal of Indian Statistical Association, 44*(1), 43–60.

Saleh, A. K. Md. E., & Khan, S. (1995). Preliminary test estimators of the mean for sampling from some Student-t populations. *Journal of Statistical Research, 29*, 67–88.

Saleh, A. K. Md. E., & Khan, S. (2003). Stein-type estimators for the mean vector in two-sample problem of multivariate Student-t populations with common covariance matrix. *International Journal of Statistical Sciences, 1*, 1–19.

Saleh, A. K. Md. E., Khan, S., & Hoque, Z. (2002). Estimation of the slope parameter for linear regression model with uncertain prior information. *Journal of Statistical Research, 36*, 55–73.

Saleh, A. K. Md. E., & Kibria, B. M. G. (1993). Performance of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics - Theory and Methods, 22*(10), 2747–2764.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2003a). Effect of W, LR, and LM tests on the performance of preliminary test ridge regression estimators. *Journal of the Japan Statistical Society, 33*(1), 119–136.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2003b). Estimation of the mean vector of a multivariate normal distribution under various test statistics. *Journal of Probability and Statistical Science, 1*, 141–155.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2004). Preliminary test ridge regression estimators with Student t-errors and conflicting test-statistics. *Metrika, 59*(2), 105–124.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2005). Comparison between Han-Bancroft and Brook methods to determine optimal significance level for pre-test estimator. *Journal of Probability and Statistical Sciences, 3*, 293–303.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2006). Pooling multivariate data under W, LR and LM tests. *Statistical Papers, 47*(1), 49–68.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2010). Estimation of the mean vector of a multivariate elliptically contoured distribution. *Calcutta Statistical Association Bulletin, 62*(3–4), 159–182.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2011). On some ridge regression estimators: A nonparametric approach. *Journal of Nonparametric Statistics, 23*(3), 819–851.

Saleh, A. K. Md. E., & Kibria, B. M. G. (2013). Improved ridge regression estimators for the logistic regression model. *Computational Statistics, 28*, 2519–2558.

Saleh, A. K. Md. E., Kibria, B. M. G., & George, F. (2018). Simultaneous estimation of several CDF's: homogeneity constraint. *Communications in Statistics - Theory and Methods, 47*(12), 2813–2826.

Saleh, A. K. Md. E., Kibria, B. M. G., & George, F. (2019). Comparative study of LASSO, ridge regression, preliminary test and Stein-type estimators for the sparse Gaussian regression model. *Statistics, Optimization and Information Computing, 7*(4), 626–641.

Saleh, A. K. Md. E., & Kim, H. M. (2002). Preliminary test prediction of population total under regression models with measurement error. *Pakistan Journal of Statistics, 18*(3), 335–357.

Saleh, A. K. Md. E., Kubokawa, T., Honda, T., & Morita, K. (1993a). Estimating a covariance matrix of a normal distribution with unknown mean. *Journal of Japan Statistical Society, 24*, 24–34.

Saleh, A. K. Md. E., Kubokawa, T., & Makita, S. (1993b). On improved positive estimators of variance components. *Statistics & Decisions, 24*(3), 3–16.

Saleh, A. K. Md. E., Kubokawa, T., & Robert, C. P. (1993c). Estimation of noncentrality parameters. *Canadian Journal of Statistics, 21*(1), 45–57.

Saleh, A. K. Md. E., & Layne, D. (1972). On moments of Gini's mean difference for samples from exponential distribution. *Journal of Statistical Research, 6*(2), 35–42.

Saleh, A. K. Md. E., Navrátil, R., & Norouzirad, M. (2018). Rank theory approach to ridge, LASSO, preliminary test and Stein-type estimators: A comparative study. *Canadian Journal of Statistics, 46*, 690–704.

Saleh, A. K. Md. E., & Norouzirad, M. (2018). On shrinkage estimation: non-orthogonal case. *SORT, Statistics, Optimization and Information Computing, 6*, 427–451.

Saleh, A. K. Md. E., Picek, J., & Kalina, J. (2009). Nonparametric estimation of regression parameters in measurement error models. *Metron, 67*(2), 177–200.

Saleh, A. K. Md. E., Picek, J., & Kalina, J. (2012). R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika, 75*(3), 311–328.

Saleh, A. K. Md. E., & Puri, M. L. (1984). Locally most powerful rank test for the two sample problem with censored samples. *Journal of Organizational Behaviour and Statistics, 1*(1), 189–196.

Saleh, A. K. Md. E., & Raheem, E. (2015). Improved LASSO. arXiv preprint arXiv:1503.05160.

Saleh, A. K. Md. E., & Rahim, M. A. (1972). Distribution of the sum of variates from truncated discrete populations. *Canadian Mathematical Bulletin, 15*(3), 395–398.

Saleh, A. K. Md. E., & Rohatgi, V. K. (1988). Estimation of parameters of location-scale family of distribution in complete and Type II censored samples. *Communications in Statistics - Theory and Methods, 17*(9), 2831–2858.

Saleh, A. K. Md. E., & Robert, C. P. (2002). Recentered confidence sets: A review of data analysis from statistical foundations. In A. K. Md. E. Saleh (Ed.) *Data analysis from statistical foundations: A Festschrift in Honour of the 75th birthday of D. A. S. Fraser* (pp. 295–308). Nova Science Publishers.

Saleh, A. K. Md. E., Rohatgi, V. K., Ahluwala, R., & Ji, P. (1991). Null distribution of locally most powerful tests for the two sample problem when the combined sample is type II censored. *Soochow Journal of Mathematics*, 10–25.

Saleh, R. A., & Saleh, A. K. Md. E. (2021). Solution to the non-monotonicity and crossing problems in quantile regression. arXiv preprint arXiv:2111.04805.

Saleh, R. A., & Saleh, A. K. Md. E. (2024). Statistical properties of the log-cosh loss function used in machine learning. arXiv preprint arXiv:2208.04564.

Saleh, A. K. Md. E., Scott, C., & Junkins, D. (1975). Exact first and second order moments of order statistics from the truncated exponential distribution. *Naval Research Logistics, 22*(1), 65–77.

Saleh, A. K. Md. E., & Sen, P. K. (1978). Nonparametric estimation of location parameter after a preliminary test on regression. *The Annals of Statistics, 6*(1), 154–168.

Saleh, A. K. Md. E., & Sen, P. K. (1979). Nonparametric estimation of location parameter after a preliminary test on regression in multivariate case. *Journal of Multivariate Analysis, 9*(2), 332–331.

Saleh, A. K. Md. E., & Sen, P. K. (1982). Nonparametric tests for location after a preliminary test on regression. *Communications in Statistics - Theory and Methods, 11*(6), 639–652.

Saleh, A. K. Md. E., & Sen, P. K. (1983a). Asymptotic properties of tests of hypothesis following preliminary test. *Statistics & Risk Modeling, 1*, 455–478.

Saleh, A. K. Md. E., & Sen, P. K. (1983b). Nonparametric tests of location parameter after a preliminary test on regression in the multivariate case. *Communications in Statistics - Theory and Methods, 12*(16), 1855–1872.

Saleh, A. K. Md. E., & Sen, P. K. (1984a). Least squares and rank order preliminary test estimation in general multivariate linear models. In J. K. Ghosh, & J. Roy (Eds.), *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference in Statistics: Application and new directions, Calcutta, December 16–19, 1981* (pp. 237–253). Statistical Publishing House.

Saleh, A. K. Md. E., & Sen, P. K. (1984b). Nonparametric preliminary test inference. In P. R. Krishnaiah, & P. K. Sen (Eds.), *Handbook of statistics* (Vol. 4, pp. 275–297).

Saleh, A. K. Md. E., & Sen, P. K. (1984c). Non-parametric shrinkage estimators of location in a multivariate simple linear model. In F. Konecny, J. Mogyorodi, & W. Wertz (Eds.), *Proceedings of the 4th Pannonian Symposium on Mathematics Statistics (PSMS)* (pp. 311–324). Bat Tatzmannsdorf.

Saleh, A. K. Md. E., & Sen, P. K. (1985a). Asymptotic relative efficiency of some joint tests of location and scale parameters based on selected order statistics. *Communications in Statistics - Theory and Methods, 14*(3), 621–633.

Saleh, A. K. Md. E., & Sen, P. K. (1985b). Nonparametric shrinkage estimation in a parallelism problem. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002), 47*, 156–165.

Saleh, A. K. Md. E., & Sen, P. K. (1985c). On shrinkage M-estimators of location parameters. *Communications in Statistics - Theory and Methods, 14*(10), 2313–2329.

Saleh, A. K. Md. E., & Sen, P. K. (1985d). Preliminary test Prediction in General multivariate linear model. In M. Matusita (Ed.), *Proceedings of the Pacific Area Statistical Conference* (pp. 619–638). North-Holland Publishers.

Saleh, A. K. Md. E., & Sen, P. K. (1985e). Shrinkage least squares estimation in a general multivariate linear models. In J. Mogyorodi, I. Vince, & W. Wertz (Eds.), *Proceedings of the 5th Pannonian Symposium on Mathematical Statistics* (pp. 307–325).

Saleh, A. K. Md. E., & Sen, P. K. (1986a). On shrinkage least squares estimation in a parallelism problem. *Communications in Statistics - Theory and Methods, 15*(5), 1451–1466.

Saleh, A. K. Md. E., & Sen, P. K. (1986b). On shrinkage R-estimation in a multiple regression model. *Communications in Statistics - Theory and Methods, 15*(7), 2229–2244.

Saleh, A. K. Md. E., & Sen, P. K. (1987a). Nonparametric and robust shrinkage estimation in some multivariate linear models. In J. K. Ghosh & S. D. Gupta (Eds.), *Proceedings on Some Multivariate Analysis* (pp. 403–423).

Saleh, A. K. Md. E., & Sen, P. K. (1987b). On the asymptotic distributional risk properties of pre-test and shrinkage L1-estimators. *Computational Statistics and Data Analysis, 5*(4), 289–299. North-Holland.

Saleh, A. K. Md. E., & Sen, P. K. (1987c). Relative performance of Stein-rule and preliminary test estimates in linear models: Least squares theory. *Communications in Statistics - Theory and Methods, 16*(2), 461–476.

Saleh, A. K. Md. E., & Sen, P. K. (1988). On shrinkage and preliminary test M-estimation in a parallelism problem. In I. B. MacNeill & G. J. Umphrey (Eds.), *Foundations of statistical inference* (pp. 105–119). D. Reidel Publishing Company.

Saleh, A. K. Md. E., & Sen, P. K. (1991). On preliminary test versions and Stein-rule estimators of the common mean of several multinormal populations. In M. Matusita (Ed.), *Statistics theory and data analysis II* (pp. 155–170). North-Holland.

Saleh, A. K. Md. E., & Sen, P. K. (1991b). Pitman-closeness of some preliminary test and shrinkage estimators. *Communications in Statistics - Theory and Methods, 20*(11) (pp. 3643–3657). Special issue on Pitman Measure of Closeness, Editors: C. R. Rao, P. K. Sen, J. Keating, & R. L. Mason.

Saleh, A. K. Md. E., & Sen, P. K. (1991c). On Pitman closeness of Pitman estimators. *Gujrat Statistical Reviews (C. G. Khatri Memorial Volume), 14*, 1–20.

Saleh, A. K. Md. E., & Shalabh, S. (2014a). A ridge regression estimation approach to the measurement error model. *Journal of Multivariate Analysis, 123*, 68–84.

Saleh, A. K. Md. E., & Shalabh, S. (2014b). Corrigendum to "Ridge regression estimation approach to measurement error model" [J. Multivariate Anal. 123 (2014) 68–84]. *Journal of Multivariate Analysis, 127*, 214.

Saleh, A. K. Md. E., & Shalabh, S. (2020). On Liu-type biased estimators in measurement error models. *Statistics, 54*(6), 1171–1213.

Saleh, A. K. Md. E., & Shiraishi, T. (1989). On some R-and M-estimators of regression parameters under uncertain restriction. *Journal of the Japan Statistical Society, 19*, 129–137.

Saleh, A. K. Md. E., & Shiraishi, T. (1992). On improved least squares estimators in multiple-design multivariate linear models under general restriction. In A. K. Md. E. Saleh (Ed.), *Proceedings of Nonparametric Statistics and Related Topics* (pp. 269–279). North-Holland.

Saleh, A. K. Md. Ehsanes, and Shiraishi, T. (1993). Robust estimation for the parameter of multiple-design multivariate linear model under general restriction. *Journal of Nonparametric Statistics, 2*, 295–305.

Saleh, A. K. Md. E., & Srivastava, M. S. (1970). On a class of nonparametric estimates of regression parameters. *Journal of Statistical Research, 4*(2), 133–139.

Saleh, A. K. Md. E., Tabatabaey, S. M. M., & Kibria, B. M. G. (2004). Simultaneous estimation of regression parameters with spherically symmetric errors under possible stochastic constraints. *International Journal of Statistical Science, 3*, 1–20.

Saleh, A. K. Md. E., Tang, F., & Kibria, B. M. G. (2009). Improved estimation of the parameters of simple linear regression model with autocorrelated errors. *Journal of Statistical Research, 43*(1), 89–107.

Saleh, R. A., Majzoub, S., Saleh, A. K. Md. E. (2025). Fundamentals of Robust Machine Learning: Handling Outliers and Anomalies in Data Science, Wiley, New Jersey.

Savage, I. R. (1956). Contributions to the theory of rank order statistics-the two-sample case. *The Annals of Mathematical Statistics, 27*(3), 590–615.

Scott, C., & Saleh, A. K. Md. E. (1975). The effect of truncation on the size of t-test. *Journal of Computation and Simulation, 3*(4), 345–368.

Sen, P. K., & Saleh, A. K. Md. E. (1985). On some shrinkage estimators of multivariate location. *The Annals of Statistics, 13*(1), 272–281.

Sen, P. K., & Saleh, A. K. Md. E. (1987). On preliminary test and shrinkage M-estimation in linear models. *The Annals of Statistics, 15*(4), 1580–1592.

Sen, P. K., Kubokawa, T., & Saleh, A. K. Md. E. (1989). The Stein paradox in the sense of the Pitman measure of closeness. *The Annals of Statistics, 17*(3), 1375–1386.

Sen, P. K., & Saleh, A. K. Md. E. (2010). The Theil–Sen estimator in a measurement error perspective. In *Nonparametrics and robustness in modern statistical inference and time series analysis: A Festschrift in honor of Professor Jana Jurečková* (pp. 224–234). Institute of Mathematical Statistics.

Sobel, M. and Hyyett, M. J. (1957). Selecting the best one of several Binomial populations, *The Bell System Technical Journal, 36*(2), 537–576.

Srivastava, M. S., & Saleh, A. K. Md. E. (2005). Estimation of the mean vector of a multivariate normal distribution: Subspace hypothesis. *Journal of Multivariate Analysis, 96*(1), 55–72.

Tabatabaey, S. M. M., Saleh, A. K. Md. E., & Kibria, B. M. G. (2004). Estimation strategies for parameters of the linear regression models with spherically symmetric distributions. *Journal of Statistical Research, 38*(1), 13–31.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological), 58*(1), 267–288.

Umbach, D., Ali, M. M., & Saleh, A. K. Md. E. (1983). Hypothesis testing for double exponential distribution based on optimal spacing. *Soochow Journal of Statistics, 10*, 45–55.

Umbach, D., Ali, M. M., & Saleh, A. K. Md. E. (1986). Estimation of the parameters of the double exponential distribution using symmetric spacings. *Soochow Journal of Mathematics, 12*, 99–108.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 67*(2), 301–320.

# Part I
# Shrinkage Estimation Strategies

# Chapter 2
# Restricted Liu-Type Regression Estimators in Linear Regression Model




**Fikri Akdeniz and Esra Akdeniz**

## 2.1 Introduction and Preliminaries

Consider the linear regression model, denoted by

$$y = X\beta + \varepsilon \tag{2.1}$$

where $y$ is an $n \times 1$ observable random vector, $X$ is a $n \times p$ design matrix of rank $p$, $\beta$ is a $p \times 1$ vector denoting unknown coefficients, and $\varepsilon$ is the $n \times 1$ vector of disturbances. We assume that the elements of $\varepsilon$ are independently distributed with expectation $\mathrm{E}[\varepsilon] = 0$ and dispersion matrix (variance–covariance matrix) $D(\varepsilon) = \sigma^2 I_n$. Simultaneously, $\beta$ satisfies linear restrictions

$$R\beta = r \tag{2.2}$$

for a given $m \times p$ nonzero matrix $R$ with rank $m < p$ and $r$ as a $m \times 1$ given vector. In this chapter, we refer restricted linear model (RLM) to Eq. (2.1) with restriction defined by Eq. (2.2). We will call an estimator $\tilde{\beta}(y)$ for $\beta$ a restricted estimator with respect to $R\beta = r$, if it satisfies $R\tilde{\beta}(y) = r$ for all $n \times 1$ vectors $y$. The well-known restricted ordinary least squares estimator (ROLSE) is

$$\beta^* = \hat{\beta} + S^{-1}R^\top \left(RS^{-1}R^\top\right)^{-1} (r - R\hat{\beta}) \tag{2.3}$$

F. Akdeniz (✉)
Department of Statistics, Çukurova University, Adana, Turkey

E. Akdeniz
Department of Medical Education, Marmara University, Istanbul, Turkey

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_2

which is a restricted estimator with respect to $R\beta = r$ in the above sense, where $\hat{\beta}^{\mathrm{OLS}} = \hat{\beta} = S^{-1}X^{\top}y$ and $S = X^{\top}X$. It is easy to see that $R\beta^* = r$ and

$$\mathrm{E}[\beta^*] = \beta + S^{-1}R^{\top}[RS^{-1}R^{\top}]^{-1}\delta^* \tag{2.4}$$

where $\delta^* = r - R\beta$. It is clear that, unless $\delta^* = 0$, $\beta^*$ is always a biased estimator for $\beta$.

The ROLSE is widely applied as unbiased estimator. In practical regression analysis, researchers often encounter the problem of multicollinearity. That is, $X^{\top}X$ is always ill-conditioned due to the multicollinearity in $X$. To overcome this problem, many studies on the general linear model without restrictions have been made. However, for ill-conditioned $X$ matrices, Hoerl and Kennard (1970) proposed the ordinary ridge regression (ORR) estimator as

$$\hat{\beta}(k) = \left(X^{\top}X + kI\right)^{-1}X^{\top}y, \quad k > 0\,. \tag{2.5}$$

Putting

$$W_k = \left(I + kS^{-1}\right)^{-1}, \quad k \geq 0, \tag{2.6}$$

we may write

$$\hat{\beta}(k) \;=\; W_k\hat{\beta}\,. \tag{2.7}$$

The symmetric positive definite matrix $W_k$ can thus also be called the ridge-shrinking matrix. The ridge-shrinkage restricted least squares estimator (RSRLSE), $\beta^*(k)$ introduced by Sarkar (1992), is defined as

$$\beta^*(k) = W_k\beta^*\,. \tag{2.8}$$

For $k = 0$, obviously $\beta^*(k) = \beta^*(0) = \beta^*$ and $\hat{\beta}(k) = \hat{\beta}(0) = \hat{\beta}$. In view of $\mathrm{E}[\beta^*(k)] = W_k\beta + W_kS^{-1}R^{\top}\left(RS^{-1}R^{\top}\right)^{-1}\delta^*$. It is clear that, unless $k = 0$ and $\delta^* = 0$ the RSRLSE estimator is always a biased estimator for $\beta$. However, it is easy to see that $\beta^*(k)$ does not satisfy $R\beta^*(k) = r$ for every outcome of $y$ and therefore is not a restricted estimator with respect to $R\beta = r$ in our sense.

Consider the least squares for regression to minimizing the sum of squared deviations objective which is defined as

$$\hat{\beta}^{\mathrm{OLS}} = \underset{\beta}{\mathrm{argmin}}\{(y - X\beta)^{\top}(y - X\beta)\}\,. \tag{2.9}$$

Adding to least squares objective given by Eq. (2.9) a penalizing function of the squared norm $\|d\hat{\beta}-\beta\|^2$ for the vector of regression coefficients yields a conditional objective:

$$L^* = \left\{(y - X\beta)^\top (y - X\beta) + (d\hat{\beta} - \beta)^\top (d\hat{\beta} - \beta)\right\} \tag{2.10}$$

where $d$ is a biasing parameter (see Lipovetsky & Conklin, 2005). The first-order condition of objective given by Eq. (2.10) minimized by the vector $\beta$ is

$$\frac{\partial L^*}{\partial \beta} = -2X^\top y + 2X^\top X\beta + 2\beta - 2d\hat{\beta} = 0\,. \tag{2.11}$$

From condition given by Eq. (2.11) of minimizing objective function defined in Eq. (2.10) by vector $\beta$, we obtain the system of equations

$$\left(X^\top X + I\right)\beta = X^\top y + d\hat{\beta}. \tag{2.12}$$

The solution of this system is

$$\begin{aligned} b_d &= \left(X^\top X + I\right)^{-1}\left(X^\top y + d\hat{\beta}\right) \\ &= (S + I)^{-1}(S + dI)\hat{\beta}, \quad 0 < d < 1. \end{aligned} \tag{2.13}$$

The estimator $b_d$ is due to Liu (1993). This combined Ridge–Stein estimator is called the Liu (**l**inear **u**nified) estimator by Akdeniz and Kaçıranlar (1995), Gruber (1998), Gross (2003), and Wang et al. (2004). The proposal of $b_d$ is motivated by augmenting $d\hat{\beta} = \beta + \varepsilon^0$ to Eq. (2.1). The derivation of $b_d$ is shown here. By artificially augmenting $d\hat{\beta} = \beta + \varepsilon^0$ to the equation $y = X\beta + \varepsilon$, we get

$$\begin{pmatrix} y \\ d\hat{\beta} \end{pmatrix} = \begin{pmatrix} X \\ I \end{pmatrix}\beta + \begin{pmatrix} \varepsilon \\ \varepsilon^0 \end{pmatrix}.$$

Define a vector $\tilde{\varepsilon} = \binom{\varepsilon}{\varepsilon^0}$, and then we have $\tilde{\varepsilon}^\top\tilde{\varepsilon} = (y - X\beta)^\top(y - X\beta) + (d\hat{\beta} - \beta)^\top(d\hat{\beta} - \beta)$.

Letting

$$F_d = (S + I)^{-1}(S + dI) \tag{2.14}$$

the Liu-type estimator $b_d$ can be expressed as

$$b_d = F_d\hat{\beta}.$$

Consequently, $\mathrm{E}[b_d] = F_d\beta$ and $D(b_d) = \sigma^2 F_d S^{-1} F_d^\top$. For $d = 1$, we have $F_{d=1} = I$ and evidently $b_{d=1} = \hat{\beta}$. Since

$$\mathrm{MSE}(b_d) = \mathrm{tr}(D(b_d)) + \|\mathrm{E}[b_d] - \beta\|^2, \tag{2.15}$$

we further obtain that

$$\mathrm{MSE}(b_d) = \sigma^2 \,\mathrm{tr}\left(F_d S^{-1} F_d^\top\right) + (1-d)^2 \beta^\top (S+I)^{-2}\beta. \tag{2.16}$$

Here $\|\cdot\|$ indicates the usual Euclidean norm (see Kaçıranlar et al., 1999). Kaçıranlar et al. (1999) introduced an alternative estimator of $\beta$, for each parameter $d \in (-\infty, +\infty)$ according to

$$\hat{\beta}_{rd} = F_d \beta^*. \tag{2.17}$$

Since this new estimator is obtained simply by replacing, in the Liu-type estimator (LTE), the OLS estimator by the restricted least squares estimator, we will call $\hat{\beta}_{rd}$ restricted Liu-type estimator (RLTE) of $\beta$. From this definition, it directly follows that the RLTE reduces to the restricted least squares estimator $\beta^*$ when $d = 1$. Akdeniz and Kaçıranlar (2001) proved some results giving necessary and sufficient conditions for the superiority of the RLTE over restricted ordinary least squares estimator (ROLSE) in the sense of the dispersion matrix of one dominating that of the order and mean squared error (MSE) matrix of one dominating that of other.

Torigoe and Ujiie (2006) compared two estimators, RLTE and ROLSE, using the generalized inverse of coefficient matrix of restriction and found some equivalency conditions for the superiority of RLTE with respect to MSE criterion.

Hubert and Wijekoon (2006) introduced another new Liu-type biased estimator called stochastic restricted Liu estimator $\hat{\beta}_{\mathrm{srd}}$ for $\beta$ and discuss its efficiency. Yang and Xu (2009) introduced an alternative improved stochastic restricted Liu estimator to the estimator proposed by Hubert and Wijekoon (2006). Özkale (2014) introduced the relative efficiency of the restricted estimators in linear regression models. Arashi et al. (2014) improved preliminary test and Stein-rule Liu estimators for the ill-conditioned elliptical linear regression model. Wu and Asar (2017) compared the restricted Liu estimator with MLE, RMLE, and Liu estimator in the mean squared error sense and they also present a method to choose a biasing parameter. Månsson and Kibria (2021) considered both unrestricted and restricted Liu estimators in the presence of multicollinearity for the Poisson regression model. Yüzbaşı et al. (2022) considered Liu-type shrinkage estimations in linear models.

It is seen that $\hat{\beta}_{rd}$ in Eq. (2.17) does not satisfy $R\hat{\beta}_{rd} = r$ unless $d = 1$; therefore, $\hat{\beta}_{rd}$ is not a restricted estimator with respect to $R\beta = r$ in our sense.

This chapter is organized as follows: In Sect. 2.2, the RLTE is proposed. In Sect. 2.3 disturbance and bias properties of the proposed estimator are presented. Comparisons between the restricted Liu estimator and restricted ordinary least squares estimator are given in Sect. 2.4. In Sect. 2.5, to illustrate the theoretical results a dataset on Portland cement is discussed.

## 2.2 Proposing Estimator

In this chapter, we suggest to minimize the sum of squared residuals under a linear restriction. So the restricted linear regression is transformed into an optimal problem with the restriction $R\beta = r$ as

$$\min_{\beta} \left\{ (y - X\beta)^{\top}(y - X\beta) + (d\hat{\beta} - \beta)^{\top}(d\hat{\beta} - \beta) \right\}. \tag{2.18}$$

For this problem, we can apply the Lagrange method. A Lagrange function is introduced as

$$\tilde{L} = (y - X\beta)^{\top}(y - X\beta) + (d\hat{\beta} - \beta)^{\top}(d\hat{\beta} - \beta) + 2\lambda\,(R\beta - r). \tag{2.19}$$

$\tilde{L}$ is a real-valued and differentiable function. The first-order condition gives

$$\frac{\partial \tilde{L}}{\partial \beta} = -2X^{\top}y + 2X^{\top}X\beta + 2R^{\top}\lambda + 2\beta - 2d\hat{\beta} = 0$$

$$\frac{\partial \tilde{L}}{\partial \lambda} = R\beta - r = 0.$$

Denoting the solution to this problem by $b_{rd}$, we get from Eq. (2.20),

$$\begin{aligned} b\hat{\beta}_{rd} &= \left(X^{\top}X + I\right)^{-1}\left(X^{\top}y + d\hat{\beta}\right) - \left(X^{\top}X + I\right)^{-1}R^{\top}\lambda, \\ &= b_d - (S + I)^{-1}R^{\top}\lambda. \end{aligned} \tag{2.20}$$

Including the restriction $Rb_{rd} - r = 0$ yields

$$Rb_{rd} = r = Rb_d - R(S + I)^{-1}R^{\top}\lambda, \tag{2.21}$$

and using $R(S + I)^{-1}R^{\top} > 0$, the optimal $\lambda$ is derived as

$$\hat{\lambda} = \left[R(S + I)^{-1}R^{\top}\right]^{-1}(r - Rb_d)\ . \tag{2.22}$$

Inserting $\hat{\lambda}$ in Eq. (2.21) and using the abbreviation $S_1 = S + I$, we get

$$b_{rd} = b_d + S_1^{-1}R^{\top}\left(RS_1^{-1}R^{\top}\right)^{-1}(r - Rb_d) \tag{2.23}$$

where $b_{rd}$ is the equality restricted Liu-type estimator. It is seen that $b_{rd}$ satisfies $Rb_{rd} = r$. Comparing Eq. (2.23) with ROLSE defined in Eq. (2.3), we can easily

find the form of our method where $b_d = S_1^{-1}\left(X^\top y + d\hat{\beta}\right)$ and $S_1 = S + I$, taking the place of $\hat{\beta}$ and $S = X^\top X$ of ROLSE respectively is completely consistent with that of ROLSE.

Furthermore, we can give another explanation to the RLTE by transforming the defined Eq. (2.23) into

$$b_{rd} = \left(I - S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}R\right)b_d + S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}r. \tag{2.24}$$

Since $R\left[S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}\right]R = R$ and the definition of generalized inverse, it is easy to see that $\left[S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}\right]$ is a generalized inverse of $R$. Defining $R^- = \left[S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}\right]$, Eq. (2.24) is equivalent to

$$b_{rd} = \left(I - R^-R\right)b_d + R^-r. \tag{2.25}$$

It is well known that the general solution of linear equation $R\beta = r$ with respect to $\beta$ is

$$R^-r + \left(I - R^-R\right)h$$

where $h$ is an arbitrary $p \times 1$ vector, and $R^- = R^g$ is a normalized generalized inverse of $R$ (Pringle & Rayner, 1971, see p. 15). Obviously, Eq. (2.25) satisfies the solution.

**Lemma 2.1 (Gross, 2003; Zhong & Yang, 2007)** *Consider that* $M_1 = S_1^{-1} - S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}RS_1^{-1}$, *and* $S_1 = X^\top X + I$.

*(i) The matrix* $M_1$ *can be written as*

$$M_1 = (QS_1Q)^+ = V\begin{pmatrix}(\Lambda + I)^{-1} & 0\\ 0 & 0\end{pmatrix}V^\top$$

*where* $M_1$ *is the Moore–Penrose inverse of the matrix* $QS_1Q$, $Q = I - R^\top\left(RR^\top\right)^{-1}R$, $\Lambda$ *is a* $(p-m)\times(p-m)$ *diagonal matrix with nonnegative main diagonal elements,* $V$ *is orthogonal matrix with* $V = (U, U^*)$,

$$Q = \left(U, U^*\right)\begin{pmatrix}I_{p-m} & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}U^\top\\ U^{*\top}\end{pmatrix} = UU^\top,$$

*and*

$$QX^\top XQ = V\begin{pmatrix}\Lambda & 0\\ 0 & 0\end{pmatrix}V^\top,$$

*as well as* $U^\top U = I_{p-m}$ *and* $U^\top X^\top XU = \Lambda$.

*(ii)* $M_1S_1M_1 = M_1$ *and* $M_0SM_0 = M_0$, *where* $S = X^\top X$ *and* $M_0 = S^{-1} - S^{-1}R^\top\left(RS^{-1}R^\top\right)^{-1}RS^{-1}$.

## 2.3 Properties of $b_{rd}$

In order to compare RLTE and ROLSE defined in Eqs. (2.23) and (2.3), respectively, we introduce the matrix-valued mean squared error (MMSE) which is called mean dispersion error (MDE).

The MDE of an estimator $b$ for $\beta$ is defined as

$$M(b,\beta) = \mathrm{E}\left[(b-\beta)(b-\beta)^\top\right]. \tag{2.26}$$

The following decomposition of the MDE often proves to be useful:

$$M(b,\beta) = \mathrm{cov}(b) + \mathrm{bias}(b)(\mathrm{bias}(b))^\top. \tag{2.27}$$

Now, we can give three theorems about the estimator $b_{rd}$ proposed in Sect. 2.2.

**Theorem 2.1** *If* $\beta$ *satisfies the linear restriction defined in Eq.* (2.3)*, then*

$$\mathrm{bias}\,(b_{rd}) = \mathrm{E}[b_{rd}] - \beta = -(1-d)M_1\beta. \tag{2.28}$$

*where* $M_1 = S_1^{-1} - S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}RS_1^{-1}$.

***Proof*** The bias of the Liu-type estimator is calculated as

$$\begin{aligned}\mathrm{E}[b_d] &= \mathrm{E}\left[(S+I)^{-1}(S+dI)\hat\beta\right]\\ &= \left[I-(1-d)S_1^{-1}\right]\beta.\end{aligned} \tag{2.29}$$

Let $\beta_0 = R^\top\left(RR^\top\right)^{-1}r = R^-r$, clearly seen that $\beta_0$ satisfies the linear restriction $R\beta = r$. So, we have

$$\begin{aligned}b_{rd} &= b_d + S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}(r - Rb_d)\\ &= S_1^{-1}S_1b_d + S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}r - S_1^{-1}R^\top\left(RS_1^{-1}R^\top\right)^{-1}RS_1^{-1}S_1b_d\end{aligned}$$

$$
\begin{aligned}
&= M_1 S_1 b_d + S_1^{-1} R^\top \left( R S_1^{-1} R^\top \right)^{-1} r \\
&= M_1 S_1 b_d + S_1^{-1} R^\top \left( R S_1^{-1} R^\top \right)^{-1} R\beta_0 \\
&= M_1 S_1 b_d + S_1^{-1} R^\top \left( R S_1^{-1} R^\top \right)^{-1} R S_1^{-1} S_1 \beta_0 + \beta_0 - S_1^{-1} S_1 \beta_0 \\
&= M_1 S_1 b_d + \beta_0 - M_1 S_1 \beta_0. \qquad (2.30)
\end{aligned}
$$

From Eqs. (2.29) and (2.30), we have

$$
\begin{aligned}
\mathrm{E}[b_{rd}] &= M_1 S_1 \left[ I - (1-d) S_1^{-1} \right] \beta + \beta_0 - M_1 S_1 \beta_0 \\
&= M_1 S_1 (\beta - \beta_0) + \beta_0 - (1-d) M_1 \beta. \qquad (2.31)
\end{aligned}
$$

Since the restrictions $R\beta = r$ are assumed to be true and $R\beta_0 = r$, we can easily get

$$
\begin{aligned}
M_1 S_1 (\beta - \beta_0) &= \left( I - R^- R \right) (\beta - \beta_0) \\
&= (\beta - \beta_0) - R^- (R\beta - R\beta_0) = \beta - \beta_0. \qquad (2.32)
\end{aligned}
$$

Thus, we obtain

$$
\begin{aligned}
\mathrm{E}[b_{rd}] &= \beta - \beta_0 - (1-d) M_1 \beta + \beta_0, \\
&= \beta - (1-d) M_1 \beta. \qquad (2.33)
\end{aligned}
$$

Since $M_1$ is a nonzero matrix, $b_{\mathrm{rd}}$ can be an unbiased estimator if and only if $d = 1$. Thus, the proof is completed. □

**Theorem 2.2** *Let us be given the estimator $b_{rd}$ under the linear regression model with true restrictions $R\beta = r$. Then*

$$
\begin{aligned}
\mathrm{MMSE}\,(b_{rd}, \beta) = M\,(b_{rd}, \beta) &= \sigma^2 M_1 (S + dI) S^{-1} (S + dI) M_1 \\
&+ (1-d)^2 M_1 \beta \beta^\top M_1. \qquad (2.34)
\end{aligned}
$$

***Proof*** From the representation in Eq. (2.30), we have

$$
\begin{aligned}
b_{rd} &= M_1 S_1 b_d - M_1 S_1 \beta_0 + \beta_0, \\
&= M_1 S_1 \left[ S_1^{-1} \left( X^\top y + d\hat{\beta} \right) \right] - M_1 S_1 \beta_0 + \beta_0, \\
&= \left( M_1 + d M_1 S^{-1} \right) X^\top y + (I - M_1 S_1)\, \beta_0\,. \qquad (2.35)
\end{aligned}
$$

Then, it follows that

$$\operatorname{cov}(b_{rd}) = \operatorname{cov}(Ay + \eta) = \operatorname{cov}(Ay) = \sigma^2 AA^\top$$

and

$$\operatorname{bias}(b_{rd}) = -(1-d)M_1\beta,$$

where $A = \left(M_1 + dM_1S^{-1}\right)X^\top = M_1(S+dI)S^{-1}X^\top$ and $\eta = (I - M_1S_1)\beta_0$. Since

$$M(b_{rd}, \beta) = \operatorname{cov}(b_{rd}) + \operatorname{bias}(b_{rd})\left(\operatorname{bias}(b_{rd})\right)^\top,$$

then, we obtain

$$\begin{aligned}\operatorname{MMSE}(b_{rd}, \beta) = M(b_{rd}, \beta) &= \sigma^2 M_1(S+dI)S^{-1}(S+dI)M_1 \\ &+ (1-d)^2 M_1\beta\beta^\top M_1. \end{aligned} \tag{2.36}$$

□

**Theorem 2.3** *Let us be given the estimators $\beta^*$ and $b_{rd}$ under the linear regression model with true restrictions $R\beta = r$. Then, the difference $D = \operatorname{cov}(\beta^*) - \operatorname{cov}(b_{rd})$ is always a nonnegative definite matrix. Moreover, $D$ is monotonously decreased with respect to $d$ in the sense of Löwner partial ordering.*

***Proof*** The well-known matrix risk of the ordinary restricted least squares estimator $\beta^*$ for the restrictions $R\beta = r$ can be written as (Rao & Toutenburg, 1999, p. 139)

$$\operatorname{cov}\left(\beta^*\right) = \sigma^2 S^{-1} - \sigma^2 S^{-1}R^\top\left(RS^{-1}R^\top\right)^{-1}RS^{-1} = \sigma^2 M_0. \tag{2.37}$$

From Theorem 2.2 we have

$$\operatorname{cov}(b_{rd}) = \sigma^2 M_1(S+dI)S^{-1}(S+dI)M_1.$$

Then, we have $D = \operatorname{cov}(\beta^*) - \operatorname{cov}(b_{rd}) = \sigma^2 M_0 - \sigma^2 M_1(S+dI)S^{-1}(S+dI)M_1$. By applying Lemma 2.1, we obtain

$$D = \sigma^2\left[M_0 - M_1\left(S + 2dI + d^2S^{-1}\right)M_1\right]$$

or

$$D = \sigma^2\left[M_0 - M_1 + (1-2d)M_1^2 - d^2M_1S^{-1}M_1\right].$$

Then, we have

$$M_0 = V \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^\top \quad \text{and} \quad M_1 = V \begin{pmatrix} (\Lambda + I)^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^\top.$$

Thus, we write

$$D = \sigma^2 V \begin{pmatrix} \Lambda^{-1} - (\Lambda + I)^{-1} + (1-2d)(\Lambda + I)^{-2} - d^2 \Lambda^{-1}(\Lambda + I)^{-2} & 0 \\ 0 & 0 \end{pmatrix} V^\top$$

$$= \begin{pmatrix} \Gamma & 0 \\ 0 & 0 \end{pmatrix}$$

where $\Gamma = \Lambda^{-1} - (\Lambda + I)^{-1} + (1-2d)(\Lambda + I)^{-2} - d^2 \Lambda^{-1}(\Lambda + I)^{-2}$. It is easy to check that the diagonal elements $\gamma_i$ of $\Gamma$ are given as

$$\gamma_i = \frac{1}{\lambda_i} - \frac{1}{1+\lambda_i} + \frac{1-2d}{(1+\lambda_i)^2} - \frac{d^2}{\lambda_i (1+\lambda_i)^2}.$$

Since $0 < d < 1$ and $\lambda_i > 0$ we have

$$\gamma_i = \frac{(1+\lambda_i)^2 - (\lambda_i + d)^2}{\lambda_i (1+\lambda_i)^2} = \frac{(1-d)(1+2\lambda_i + d)}{\lambda_i (1+\lambda_i)^2} > 0.$$

Thus, we obtain $\Gamma > 0$ and $D = (\operatorname{cov}(\beta^*) - \operatorname{cov}(b_{rd})) \geq 0$ in the sense of Löwner partial ordering. Since $\lambda_i > 0$ and $0 < d < 1$, then

$$\frac{\mathrm{d}\gamma_i}{\mathrm{d}d} = -\frac{2(\lambda_i + d)}{\lambda_i (1+\lambda_i)^2} < 0 \quad .$$

$D$ is monotonously decreasing with respect to $d$. □

## 2.4 Comparisons Between the Restricted Liu-Type Estimator and the Restricted Ordinary Least Squares Estimator

It is of interest to know under which conditions the matrix risk of the new estimator $b_{rd}$ is smaller than the matrix risk of $\beta^*$. For this we investigate the difference:

$$\Delta = M(\beta^*, \beta) - M(b_{rd}, \beta). \tag{2.38}$$

The well-known matrix risk of the restricted ordinary least squares estimator $\beta^*$ for the restrictions $R\beta = r$ can be written as

$$\operatorname{cov}(\beta^*) = \sigma^2 S^{-1} - \sigma^2 S^{-1} R^\top \left( R S^{-1} R^\top \right)^{-1} R S^{-1}. \tag{2.39}$$

Under the linear regression model defined in Eq. (2.1), the mean squared error matrix of $\beta^*$ is given by

$$M\left(\beta^*, \beta\right) = \sigma^2 M_0 + S^{-1}R^\top \left(RS^{-1}R^\top\right)^{-1} \delta\delta^\top \left(RS^{-1}R^\top\right)^{-1} RS^{-1} \tag{2.40}$$

where $M_0 = S^{-1} - S^{-1}R^\top \left(RS^{-1}R^\top\right)^{-1} RS^{-1}$, $\delta = R\beta - r$, and

$$\text{bias}\left(\beta^*\right) = -S^{-1}R^\top \left(RS^{-1}R^\top\right)^{-1} \delta.$$

From Theorem 2.2, the difference $\Delta$ is given by

$$\Delta = \sigma^2 M_0 - \sigma^2 M_1 (S + dI) S^{-1} (S + dI) M_1 - (1-d)^2 M_1 \beta\beta^\top M_1 \tag{2.41}$$

under $R\beta = r$ which is assumed to be true.

**Lemma 2.2 (Baksalary & Kala, 1983)** *Let $A$ be a nonnegative definite matrix and let $\boldsymbol{a}$ be a column vector. Then $A - \boldsymbol{a}\boldsymbol{a}^\top \geq 0$ if and only if $\boldsymbol{a}$ belongs to the range of $A$ and $\boldsymbol{a}^\top A^- \boldsymbol{a} \leq 1$, where $A^-$ is any g-inverse of $A$, that is, $AA^-A = A$.*

**Theorem 2.4** *Let us be given the estimators $\beta^*$ and $b_{rd}$. If $0 < d < 1$, then the matrix risk difference $\Delta = \sigma^2 D - (1-d)^2 M_1 \beta\beta^\top M_1$ is nonnegative definite if and only if*

$$\beta^\top \left[\frac{2}{1+d} Q + \left(QX^\top XQ\right)^+\right]^+ \beta \leq \alpha\sigma^2 \tag{2.42}$$

*where $D = \sigma^2 M_0 - \sigma^2 M_1 (S + dI) S^{-1} (S + dI) M_1$, $\alpha = (1+d)/(1-d)$ and $Q = I_p - R^\top \left(RR^\top\right)^{-1} R^\top$.*

***Proof*** Putting $\theta = (1-d) M_1 \beta$, the difference $\Delta$ becomes $\Delta = \sigma^2 D - \theta\theta^\top$, where $D$ is symmetric and nonnegative definite and $\theta$ belongs to the range of $D$. By applying Lemma 2.2, $\Delta$ is nonnegative definite if and only if $\theta^\top D^+ \theta \leq \sigma^2$. By applying Lemma 2.1, we obtain

$$\theta = (1-d) M_1 \beta = V \begin{pmatrix} (1-d)\left(\Lambda + I_{p-m}\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^\top \beta. \tag{2.43}$$

Since

$$D^+ = V \begin{pmatrix} \Gamma^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^\top,$$

then, we have

$$\theta^{\top} D^{+} \theta = \beta^{\top} V \begin{pmatrix} (1-d)^2 \left(\Lambda + I_{p-m}\right)^{-1} \Gamma^{-1} \left(\Lambda + I_{p-m}\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^{\top} \beta$$

$$= \beta^{\top} V \begin{pmatrix} \Omega & 0 \\ 0 & 0 \end{pmatrix} V^{\top} \beta, \tag{2.44}$$

where

$$\begin{aligned} \Omega &= (1-d)^2 \left(\Lambda + I_{p-m}\right)^{-1} \Gamma^{-1} \left(\Lambda + I_{p-m}\right)^{-1} \\ &= (1-d)^2 \left[\left(\Lambda + I_{p-m}\right) \Gamma \left(\Lambda + I_{p-m}\right)\right]^{-1} \\ &= (1-d)^2 \left[2(1-d) I_{p-m} + \left(1-d^2\right) \Lambda^{-1}\right]^{-1} \\ &= \frac{1-d}{1+d} \left(\frac{2}{1+d} I_{p-m} + \Lambda^{-1}\right)^{-1}, \end{aligned}$$

so that

$$\theta^{\top} D^{+} \theta = \frac{1}{\alpha} \beta^{\top} \left[ V \begin{pmatrix} \frac{2}{1+d} I_{p-m} + \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^{\top} \right]^{+} \beta$$

$$= \frac{1}{\alpha} \beta^{\top} \left[ V \begin{pmatrix} \frac{2}{1+d} I_{p-m} & 0 \\ 0 & 0 \end{pmatrix} V^{\top} + V \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} V^{\top} \right]^{+} \beta$$

where $D^{+}$ denotes the Moore–Penrose inverse of $D$. By applying Lemma 2.1 again, we have

$$\theta^{\top} D^{+} \theta = \frac{1}{\alpha} \beta^{\top} \left[ \frac{2}{1+d} Q + \left( Q X^{\top} X Q \right)^{+} \right]^{+} \beta \leq \sigma^2. \tag{2.45}$$

This result completes the proof of Theorem 2.4. □

## 2.5 Experimental Results

To illustrate the given results we now consider in this section a dataset on Portland cement originally due to Woods et al. (1932). The dataset has been widely analyzed by Kaçıranlar et al. (1999) and Mishra (2004).

These data come from an experimental investigation of the heat evolved during the setting and hardening of Portland cements of varied composition and the

dependence of this heat on the percentages of four compounds in the clinkers from which the cement was produced. In this example, the dependent variable $y$ is defined as heat evolved in calories per gram of cement. The four compounds considered by Woods et al. (1932) are tricalcium aluminate ($X_1$), tricalcium silicate ($X_2$), tetracalcium alumino ferrite ($X_3$), and $\beta$-dicalcium silicate ($X_4$). The four columns of the $13 \times 4$ matrix $X$ comprise the data on $x_1$, $x_2$, $x_3$, and $x_4$, respectively.

We assemble our data as

$$X = (x_1, x_2, x_3, x_4) = \begin{bmatrix} 7 & 26 & 6 & 60 \\ 1 & 29 & 15 & 52 \\ 11 & 56 & 8 & 20 \\ 11 & 31 & 8 & 47 \\ 7 & 52 & 6 & 33 \\ 11 & 55 & 9 & 22 \\ 3 & 71 & 17 & 6 \\ 1 & 31 & 22 & 44 \\ 2 & 54 & 18 & 22 \\ 21 & 47 & 4 & 26 \\ 1 & 40 & 23 & 34 \\ 11 & 66 & 9 & 12 \\ 10 & 68 & 8 & 12 \end{bmatrix}, \quad \sum_{j=1}^{4} x_j = \begin{bmatrix} 99 \\ 97 \\ 95 \\ 97 \\ 98 \\ 97 \\ 97 \\ 98 \\ 96 \\ 98 \\ 98 \\ 98 \\ 98 \end{bmatrix}, \quad y = \begin{bmatrix} 78.5 \\ 74.3 \\ 104.3 \\ 87.6 \\ 95.9 \\ 109.2 \\ 102.7 \\ 72.5 \\ 93.1 \\ 115.9 \\ 83.8 \\ 113.3 \\ 109.4 \end{bmatrix}.$$

Consider the following linear regression model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u. \tag{2.46}$$

The ordinary least squares estimate of $\beta$ is $\tilde{\beta}^{\mathrm{OLS}} = (2.1930, 1.1533, 0.7585, 0.4863)^{\top}$ and $\tilde{\sigma}^2_{\mathrm{OLS}} = 5.8455$ as well as $\hat{\mathrm{mse}}(\tilde{\beta}^{\mathrm{OLS}}) = 0.0638$. The condition number of $X$ is $\kappa(X) = 20.5846$ so $X$ may be considered "well-conditioned." We now, however, following Hald (1952, pp. 648–649) and Gorman and Toman (1966, pp.35–36), Daniel and Wood (1980, pp. 89), and Kaçıranlar et al. (1999) add a column of ones to the data matrix $X$ and fit an inhomogeneous linear regression model (regression model with an intercept). Now, consider the following regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u. \tag{2.47}$$

It is seen that the row totals of the $13 \times 4$ matrix $X$ (without the column of ones) are all approximately equal to 100 (the compounds are all measured in percentages). We still have $n = 13$ observations but there are now $p = 5$ unknown regression coefficients. The condition number of $X$ is now $\kappa(X) = 6056.3744$, which is exceptionally large and so $X$ may be considered as being quite "ill-conditioned." With an intercept, the associated inhomogeneous linear model certainly possesses "high multicollinearity" with possible effects on the OLS estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$,

**Table 2.1** OLS, RLT, and ROLS estimators for the cement data

| Parameter | OLSE | ROLSE | RLTE |
|---|---|---|---|
| $\beta_0$ | 62.405 | 61.454 | 24.831 |
| $\beta_1$ | 1.551 | 1.560 | 1.711 |
| $\beta_2$ | 0.510 | 0.520 | 0.784 |
| $\beta_3$ | 0.102 | 0.111 | 0.416 |
| $\beta_4$ | −0.144 | −0.134 | 0.188 |
| $\widehat{\text{sm}}\text{se}$ | 5.9830 | 5.3183 | 4.4356 |

$\hat{\beta}_4$. The ordinary least squares estimate of $\beta$ is presented in the first column of Table 2.1.

Thus, the estimated regression function is

$$\hat{y} = 62.4054 + 1.551\, x_1 + 0.5102\, x_2 + 0.1019\, x_3 - 0.1441\, x_4,$$

and $\hat{\sigma}^2_{\text{ols}} = 5.9830$ as well as $R^2 = 0.9824$, which implies that 98.24% of the total variation has been explained by the regressors.

Following Kaçıranlar et al. (1999, pp. 459), we may improve our estimates by adding to the inhomogeneous model the linear restriction $\beta_3 = \beta_2 - \beta_1$ or $R\beta = r = 0$ where $R = (0, 1, -1, 1, 0)$. The second column of Table 2.1 shows the restricted OLS estimator.

We test the linear hypothesis $H_0 : R\beta = r = 0$ in the unrestricted linear model defined in Eq. (2.47). The test statistic for testing the null hypothesis $H_0$ against $H_1 : R\beta \neq r = 0$, given our observations, is

$$F = \frac{\left(R\hat{\beta}^{\text{OLS}} - r\right)^\top \left(RS^{-1}R^\top\right)^{-1} \left(R\hat{\beta}^{\text{OLS}} - r\right)}{\hat{\sigma}^2} = 1.92\ .$$

Since $F_{0.95,4,8} = 5.0526$, $H_0$ is not rejected at the 5% significance level. We have the following estimates for $\sigma^2$:

(a) Unrestricted inhomogeneous model: $\hat{\sigma}^2_{\text{OLS}} = 5.9830$
(b) Restricted inhomogeneous model with the restriction $\beta_3 = \beta_2 - \beta_1$: $\hat{\sigma}^2_{\text{R}} = 6.5952$

Since $R = (0, 1, -1, 1, 0)^\top$, we have $R^- = \left(R^\top R\right)^{-1} R^\top = \frac{1}{3} R^\top$. Then, we obtain the restricted Liu-type estimator $b_{rd} = \left(I - R^- R\right) b_d + R^- r$ from Eq. (2.25). This estimator can be found in the third column of Table 2.1.

Comparing $\widehat{\text{mse}}(b_{rd}) = 4.4356$ with $\text{mse}(\beta^*) = 5.3183$ and $\text{MSE}(\hat{\beta}) = 5.9830$ shows that $b_{rd}$ has indeed a smaller estimated (scalar) mse than the other two estimators; this agrees with our theoretical findings in Theorem 2.4.

We obtain the restricted Liu-type estimator using Eq. (2.24). Using the estimated mse values, it will be possible to compare them. We found the restricted Liu-type estimator $b_{rd} = \left(I - R^- R\right) b_d + R^- r$ with Eq. (2.25).

Here $r = 0$ and $b_d$ was in the table. We used the generalized inverse of $R$ as

$$R^{-} = \frac{1}{3}\begin{bmatrix} 0 \\ 1 \\ -1 \\ 1 \\ 0 \end{bmatrix}$$

where $R^{-} = \left[S_1^{-1}R^{\top}\left(RS_1^{-1}R^{\top}\right)^{-1}\right]$. Also,

$$I - R^{-}R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2/3 & 1/3 & -1/3 & 0 \\ 0 & 1/3 & 2/3 & 1/3 & 0 \\ 0 & -1/3 & 1/3 & 2/3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

## 2.6 Concluding Results

In this chapter, we have compared the performance of the restricted Liu-type estimator $b_{rd}$ with the well-known restricted ordinary least squares estimator $\beta^*$ in terms of the matrix mean squared error. We can also conclude that $\mathrm{cov}\,(\beta^*) - \mathrm{cov}\,(b_{rd})$ is always a nonnegative definite. The estimator $b_{rd}$ appears to be a noteworthy competitor of $\beta^*$.

## References

Akdeniz, F., & Kaçıranlar, S. (1995). On the Almost Unbiased Generalized Liu Estimator and Unbiased Estimation of the Bias and MSE, *Communications in Statistics, Theory and Methods, 24*(7), 1789–1797.

Akdeniz, F., & Kaçıranlar, S. (2001). More on the new biased estimator in linear regression. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002), 63*, 321–325.

Arashi, M., Kibria, B. G., Norouzirad, M., & Nadarajah, S. (2014). Improved preliminary test and Stein-rule Liu estimators for the ill-conditioned elliptical linear regression model. *Journal of Multivariate Analysis, 126*, 53–74.

Baksalary, J. K., & Kala, R. (1983). Partial orderings between matrices one of which is of rank one, *Bulletin of the Polish Academy of Science, Mathematics, 31*, 5–7.

Daniel, C., & Wood, F. S. (1980). *Fitting equations to data: Computer analysis of multifactor data* (2nd ed.), with the assistance of John W. Gorman. Wiley.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics, 12*, 69–82.

Hubert, M. H., & Wijekoon, P. (2006). Improvement of the Liu estimation in linear regression model. *Statistical Papers, 47*(3), 471–479.

Gorman, J. W., & and Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics, 8*, 27–51.

Gross, J. (2003). *Linear regression*. Springer.

Gruber, M. H. J. (1998). *Improving efficiency by shrinkage: The James-Stein and ridge regression estimators*. Marcel Dekker.

Hald, A. (1952). *Statistical theory with engineering applications*. Wiley.

Kaçıranlar, S., Sakallığlu, S., Akdeniz, F., Styan, G. P. H., & Werner, H. J. (1999). A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland cement. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002), 61*, 443–459.

Lipovetsky, S., & Conklin, W. M. (2005). Ridge regression in two-parameter solution, *Applied Stochastic Models in Business and Industry, 21*, 525–540.

Liu, K. J. (1993). A new class of biased estimate in linear regression. *Communications in Statistics - Theory and Methods, 22*, 393–402.

Månsson, K., & Kibria, G. M. G. (2021). Estimating the unrestricted and restricted Liu estimators for the poisson regression model: Method and application. *Computational Economics, 58*(1), 311–326.

Mishra, S. K. (2004). Estimation under multicollinearity: Application of restricted Liu and maximum entropy estimators to the Portland cement dataset, MPRA paper 1809, University Library of Munich, Germany.

Özkale, M. R. (2014). The relative efficiency of the restricted estimators in linear regression models. *Journal of Applied Statistics, 41*(5), 998–1027.

Pringle, R. M., & Rayner, A. A. (1971). *Generalized inverse matrices with applications to statistics*. Hafner Publishing Company.

Rao, C. R., & Toutenburg, H. (1999). *Linear models-least squares and alternatives* (2nd ed.). Springer.

Sarkar, N. (1992). A new estimator combining the ridge regression and restricted least squares methods of estimation. *Communications in Statistics - Theory and Methods, 21*, 1987–2000.

Torigoe, N., & Ujiie, K. (2006). On the restricted Liu estimator in Gauss-Markov model. *Communications in Statistics - Theory and Methods, 35*(9), 1713–1722.

Wang, T., Zhang, Y. S., & Zhang, Y. (2004). Combined Ridge-Stein estimator in exterior orientation for linear pushbroom imagery, Commission IV, WG IV/7 (pp. 727–731); XXth ISPRS Congress, 12–23 July 2004, Istanbul, Turkey.

Woods, H., Steinour, H., & Starke, H. R. (1932). Effect of composition of Portland cement on heat evolved during hardening, *Industrial & Engineering Chemistry, 24*, 1207–1241.

Wu, J., & Asar, Y. (2017). More on the restricted Liu estimator in the logistic regression model. *Communications in Statistics - Simulation and Computation, 46*(5), 3680–3689.

Yang, H., & Xu, J. (2009). An alternative stochastic restricted Liu estimator in linear regression. *Statistical Papers, 50*, 639–647.

Yüzbaşı, B., Asar, Y., & Ahmed, S. E. (2022). Liu-type shrinkage estimations in linear models, *Statistics, 56*(2), 396–420.

Zhong, Z., & Yang, H. (2007). Ridge estimation to the restricted linear model. *Communications in Statistics - Theory and Methods, 36*, 2099–2115.

# Chapter 3
# Shrinkage Strategies for Right-Censored Bell Regression Model with Application




**Hossein Bevrani and Zahra Zandi**

## 3.1 Introduction

In various practical problems, the observations may be in the form of count data. In modeling such data, the Poisson regression model is widely applied when its mean and variance are equal. However, it often occurs that the count data exhibit over-dispersion (i.e., variance greater than the mean). In such case, the Bell regression (BR) model is an appropriate choice to describe the count data. The Bell distribution initially introduced by Castellares et al. (2018) is a one-parameter distribution that belongs to the exponential family of distributions for which the variance is larger than its mean.

The BR model has been applied to describe complete count data. However, count data can be subjected to right-censoring. It means that when only a lower bound of the count of interest is observed, or when the accurate value of the count data is higher than the observed value. Right-censoring has been used in many count models such as zero-inflated Poisson regression (Zandi et al., 2024), Poisson (Terza, 1985), generalized Poisson (Famoye & Wang, 2004) models, and Poisson mixtures model (Karlis et al., 2016). So far, no research has been done on the BR model with right-censored data. In the current chapter, we consider the right-censoring Bell (RCB) model and improve parameter estimation using shrinkage strategies when there exists some prior information about a subset of predictors which have no significant effect (known as inactive predictors) on the response variable. The prior information appears on the parameter vector as

$$Q\boldsymbol{\gamma} = \boldsymbol{q} \tag{3.1}$$

H. Bevrani (✉) · Z. Zandi
Department of Statistics, University of Kurdistan, Sanandaj, Iran
e-mail: bevrani@gmail.com; zahrazandi163@yahoo.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_3

where $\boldsymbol{\gamma}$ is the unknown regression parameter vector in the RCB model comprising the $p = k_1 + k_2$ parameter, where $k_1$ and $k_2$ are the number of parameters related to the active and inactive predictors, respectively, $\boldsymbol{Q}$ is a known matrix of order $k_2 \times p$, and $\boldsymbol{q}$ is a $k_2 \times 1$ vector with constant elements. Shrinkage estimators are derived by comparing parameters estimated from a full model containing all predictors to those estimated from a submodel including only active predictors. The parameters estimated from the full model are known as unrestricted estimators (UEs), while those from the submodel are referred to as restricted estimators (REs). These estimators contain linear shrinkage, preliminary test, shrinkage pretest, Stein, and positive Stein estimators initially introduced by Thompson (1968), Bancroft (1944), Ahmed (1992), and Stein (1956), respectively. The shrinkage estimators are defined as a linear combination of UE and RE in order to improve parameter estimation (Saleh, 2006). The shrinkage estimators have been introduced in various regression models, especially in count models, for example, in the BR model (Seifollahi et al., 2024), the zero-inflated Poisson model with right-censored data (Zandi et al., 2024), and the zero-inflated negative binomial model (Zandi et al., 2021), among several others.

In this chapter, our main goal is to enhance the accuracy of estimates in the RCB model by utilizing shrinkage strategies. We also aim to evaluate their performance with the UE through a Monte Carlo simulation and illustrate using a real dataset.

The rest of this chapter is organized as follows: the RCB model and UE are introduced in Sect. 3.2. The RE and shrinkage estimators are introduced in Sect. 3.3. The asymptotic distributional bias (ADB) and asymptotic distributional risk (ADR) of the proposed estimators are developed in Sect. 3.4. In Sect. 3.5, the Monte Carlo simulation study is presented with details to compare the performance of the proposed estimators. A real data application to illustrate our findings is analyzed in Sect. 3.6. Finally, concluding remarks are presented in Sect. 3.7.

## 3.2 Right-Censored Bell Model and Unrestricted Estimator

In this section, we will first introduce the Bell distribution. Then, we will discuss the RCB model and the estimation of its parameters using the maximum likelihood estimator (MLE), which is the UE.

### *3.2.1 The Bell Model*

Suppose that $Y$ is a random variable with the count observation which follows a Bell distribution with the probability mass function (p.m.f) as

$$\mathrm{P}(Y = y) = \frac{\delta^y \exp\{-\exp\{\delta + 1\}\}\mathcal{B}_y}{y!}, \qquad y = 0, 1, 2, \ldots, \tag{3.2}$$

where $\delta > 0$ and $\mathcal{B}_y$ is the Bell number defined as follows:

$$\mathcal{B}_y = \frac{1}{\exp\{1\}} \sum_{t=0}^{\infty} \frac{t^y}{t!}. \tag{3.3}$$

Based on Eq. 3.3, some Bell numbers for $y = 0, 1, \ldots, 5$ are as $\mathcal{B}_0 = \mathcal{B}_1 = 1$, $\mathcal{B}_2 = 2$, $\mathcal{B}_3 = 5$, $\mathcal{B}_4 = 15$, and $\mathcal{B}_5 = 52$ (Castellares et al., 2018). As mentioned, the variance is greater than the mean in Bell distribution which are given as follows:

$$\mathrm{E}[Y] = \delta \exp\{\delta\}, \qquad \mathrm{var}(Y) = \delta(1+\delta)\exp\{\delta\}. \tag{3.4}$$

The p.m.f. of the Bell distribution can be re-parameterized by considering $\theta = \delta \exp\{\delta\}$ and $\delta = \mathcal{W}_0(\theta)$, where $\mathcal{W}_0(\cdot)$ represents the Lambert function. Hence, the p.m.f. can be written as

$$\mathrm{P}(Y = y) = \exp\{1 - \exp\{\mathcal{W}_0(\theta)\}\} \frac{\big(\mathcal{W}_0(\theta)\big)^y \mathcal{B}_y}{y!}, \qquad y = 0, 1, \ldots, \quad \theta > 0, \tag{3.5}$$

where $\theta > 0$ and $\mathcal{W}_0(\theta) > 0$. It is clear that $\mathrm{E}(Y) = \theta$ and $\mathrm{var}(Y) = \theta\big(1 + \mathcal{W}_0(\theta)\big)$. The Bell distribution is denoted by Bell$(\mathcal{W}_0(\theta))$.

The BR model is applied when the count response variable, $Y_i$, for $i = 1, 2, \ldots, n$ distributed as Bell$(\mathcal{W}_0(\theta_i))$ such that $\ln(\theta_i) = (\boldsymbol{x}_i^\top \boldsymbol{\gamma}) \exp\{\boldsymbol{x}_i^\top \boldsymbol{\gamma}\}$ where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the $i$th row of design matrix $\mathcal{X}_{n\times p}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ is the regression coefficient vector where $p$ is the number of predictor variables.

### 3.2.2 The Right-Censored Bell Model

Suppose that the response variable $Y_i$ can be right-censored, i.e., a lower bound on $Y_i$ is observable for some individuals. Hence, we observe $(Y_i^*, \tau_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$, where $Y_i^* = \min(y_i, T_i)$, so that $T_i$ is a censoring random variable and $\tau_i = I(Y_i < T_i)$ is an indicator function with 1 and 0 values for $Y_i < T_i$ and $Y_i > T_i$, respectively. When $Y_i = T_i$, we set $Y_i^* = T_i$ and $\tau_i = 0$. The likelihood function in the RCB model denoted by $L_{\mathrm{RCB}}(\boldsymbol{\gamma})$ is given as

$$L_{\mathrm{RCB}}(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \Big(\mathbb{P}(Y_i = Y_i^*)\Big)^{\tau_i} \Big(\mathbb{P}(Y_i \geq Y_i^*)\Big)^{1-\tau_i}$$

$$= \prod_{i=1}^{n} \left(g_i(\boldsymbol{\gamma}) \frac{\big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\big)^{Y_i^*} \mathcal{B}_{Y_i^*}}{Y_i^*!}\right)^{\tau_i}$$

$$\left(1-\sum_{r=0}^{Y_i^*-1} g_i(\boldsymbol{\gamma})\Big[\frac{\big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\big)^r \mathcal{B}_r}{r!}\Big]\right)^{(1-\tau_i)},$$

where $e_i(\boldsymbol{\gamma}) = u_i(\boldsymbol{\gamma})\exp\{u_i(\boldsymbol{\gamma})\}$, $u_i(\boldsymbol{\gamma}) = \exp\{(\boldsymbol{x}_i^\top\boldsymbol{\gamma})\}$, and $g_i(\boldsymbol{\gamma}) = \exp\{1 - \mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\}$. We denote the log-likelihood function of the RCB model as $\mathcal{L}_{\text{RCB}}(\boldsymbol{\gamma})$ which is given as

$$\mathcal{L}_{\text{RCB}}(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\tau_i\left\{1-\exp\{\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\} + Y_i^*\ln\Big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\Big)\right\}$$

$$+\sum_{i=1}^{n}(1-\tau_i)\left\{\ln\left(1-\sum_{r=0}^{Y_i^*-1}\exp\{\big(1-\exp\{\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\}\big)\}\frac{\Big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\Big)^r\mathcal{B}_r}{r!}\right)\right\}. \tag{3.6}$$

### 3.2.3 *Unrestricted Estimator*

As mentioned, the estimation of parameters in the full model is called the UE. Hence, the UE (or MLE) of unknown regression parameter vector $\boldsymbol{\gamma}$ in the RCB model denoted by $\hat{\boldsymbol{\gamma}}^{\text{UE}}$ can be calculated by solving the following score equation using numerical methods:

$$\frac{\partial\,\mathcal{L}_{\text{RCB}}(\boldsymbol{\gamma})}{\partial\,\gamma_j} = \sum_{i=1}^{n}\tau_i\,(x_{ij})\left\{\frac{u_i(\boldsymbol{\gamma})\big(Y_i^*-e_i(\boldsymbol{\gamma})\big)}{\big(\exp\{u_i(\boldsymbol{\gamma})\}+e_i(\boldsymbol{\gamma})\big)\big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\big)}\right\}$$

$$+\sum_{i=1}^{n}(1-\tau_i)\,(x_{ij})\sum_{r=0}^{Y_i^*-1}\frac{\mathcal{B}_r}{r!}\left\{\Big(\exp\{1-\exp\{\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\}\}\Big)\right.$$

$$\left.\times\frac{\Big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\Big)^{r-1}\Big(u_i(\boldsymbol{\gamma})\,e_i(\boldsymbol{\gamma})+r\,u_i(\boldsymbol{\gamma})\Big)}{\big(\exp\{u_i(\boldsymbol{\gamma})\}+e_i(\boldsymbol{\gamma})\big)\,h(Y_i^*)}\right\}=0, \tag{3.7}$$

for $j = 1,\ldots,p$ where $h(s) = \mathbb{P}\big(\text{Bell}(\mathcal{W}_0(\theta_i)) \geq s\big)$ is the survival function of the Bell distribution. Under some regularity conditions of the MLE, when $n\to\infty$, $\sqrt{n}(\hat{\boldsymbol{\gamma}}^{\text{UE}}-\boldsymbol{\gamma}) \sim \mathcal{N}_p\big(\mathbf{0},\mathcal{I}^{-1}\big)$ with $\mathcal{I} = \lim_{n\to\infty}\frac{1}{n}\boldsymbol{I}(\boldsymbol{\gamma})$ is a nonsingular matrix (Hossain et al., 2015) and $\boldsymbol{I}(\boldsymbol{\gamma})$ is Fisher information matrix in the RCB model given in Appendix "The Fisher Information Matrix".

## 3.3 Restricted and Shrinkage Estimators

In the current section, we suggest the RE and shrinkage estimators to improve the estimation process in the RCB model.

### *3.3.1 Restricted Estimators*

The RE of parameter $\boldsymbol{\gamma}$ denoted by $\boldsymbol{\gamma}^{\mathrm{RE}}$ is calculated by maximizing Eq. (3.6) under the linear restriction given by Eq. (3.1). The general form of the RE is as follows:

$$\hat{\boldsymbol{\gamma}}^{\mathrm{RE}} = \hat{\boldsymbol{\gamma}}^{\mathrm{UE}} - \boldsymbol{I}^{-1}(\boldsymbol{\gamma})Q^{\top}\left[Q\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma})Q^{\top}\right]^{-1}\left(Q\hat{\boldsymbol{\gamma}}^{\mathrm{UE}} - \boldsymbol{q}\right). \tag{3.8}$$

### *3.3.2 Shrinkage Estimators*

In the following, we introduce the types of shrinkage estimators:

#### 3.3.2.1 Linear Shrinkage Estimator

The linear shrinkage estimator (LSE) of $\boldsymbol{\gamma}$ denoted by $\hat{\boldsymbol{\gamma}}^{\mathrm{LSE}}$ is defined as a linear combination of the UE and RE given as

$$\hat{\boldsymbol{\gamma}}^{\mathrm{LSE}} = \varphi\,\hat{\boldsymbol{\gamma}}^{\mathrm{RE}} + (1-\varphi)\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}, \quad \varphi \in [0,1], \tag{3.9}$$

where $\varphi$ indicates the accuracy of the prior information. The LSE carries out better for the value of $\varphi$ close to one. Its optimum value can be calculated by minimizing the mean squared error of this estimator (Zandi et al., 2021). The LSE chooses $\hat{\boldsymbol{\gamma}}^{\mathrm{RE}}$ or $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ if $\varphi = 1$ or $\varphi = 0$, respectively.

#### 3.3.2.2 Preliminary Test Estimator

The preliminary test estimator (PTE) of $\boldsymbol{\gamma}$ denoted by $\hat{\boldsymbol{\gamma}}^{\mathrm{PTE}}$ is defined as

$$\hat{\boldsymbol{\gamma}}^{\mathrm{PTE}} = \hat{\boldsymbol{\gamma}}^{\mathrm{UE}} - (\hat{\boldsymbol{\gamma}}^{\mathrm{UE}} - \hat{\boldsymbol{\gamma}}^{\mathrm{RE}})\,I\left(\mathcal{R}_n \leq \chi^2_{1-\alpha}(k_2)\right), \tag{3.10}$$

where $I(\cdot)$ is an indicator function and $\mathcal{R}_n$ is the likelihood ratio test statistic for testing $H_0 : \mathcal{Q}\boldsymbol{\gamma} = \boldsymbol{q}$ vs. $H_1 : \mathcal{Q}\boldsymbol{\gamma} \neq \boldsymbol{q}$ that is defined as

$$\mathcal{R}_n = 2 \ln \left( \frac{\mathcal{L}_{\text{RCB}}(\hat{\boldsymbol{\gamma}}^{\text{UE}})}{\mathcal{L}_{\text{RCB}}(\hat{\boldsymbol{\gamma}}^{\text{RE}})} \right), \tag{3.11}$$

where $\mathcal{L}_{\text{RCB}}(\hat{\boldsymbol{\gamma}}^{\text{UE}})$ and $\mathcal{L}_{\text{RCB}}(\hat{\boldsymbol{\gamma}}^{\text{RE}})$ are values of the log-likelihood function at the UE and RE, respectively. Under $H_0 : \mathcal{Q}\boldsymbol{\gamma} = \boldsymbol{q}$, the test statistic $\mathcal{R}_n$ is asymptotically distributed as $\chi^2(k_2)$ (Aitchison & Silvey, 1958). The PTE chooses the UE or RE when $H_0$ is accepted or is rejected, respectively.

#### 3.3.2.3 Shrinkage Preliminary Test Estimator

The shrinkage preliminary test estimator (SPE) of $\boldsymbol{\gamma}$ denoted by $\hat{\boldsymbol{\gamma}}^{\text{SPE}}$ is defined as

$$\hat{\boldsymbol{\gamma}}^{\text{SPE}} = \hat{\boldsymbol{\gamma}}^{\text{UE}} - \varphi\,(\hat{\boldsymbol{\gamma}}^{\text{UE}} - \hat{\boldsymbol{\gamma}}^{\text{RE}})\, I\big(\mathcal{R}_n \leq \chi^2_{1-\alpha}(k_2)\big). \tag{3.12}$$

#### 3.3.2.4 Stein Estimators

The Stein estimator (SE) of $\boldsymbol{\gamma}$ denoted by $\hat{\boldsymbol{\gamma}}^{\text{SE}}$ is given as

$$\hat{\boldsymbol{\gamma}}^{\text{SE}} = \hat{\boldsymbol{\gamma}}^{\text{UE}} - \Big(\frac{k_2 - 2}{\mathcal{R}_n}\Big)(\hat{\boldsymbol{\gamma}}^{\text{UE}} - \hat{\boldsymbol{\gamma}}^{\text{RE}}), \qquad k_2 = 3, 4, \ldots. \tag{3.13}$$

The positive Stein estimator (PSE) of $\boldsymbol{\gamma}$ denoted by $\hat{\boldsymbol{\gamma}}^{\text{PSE}}$ is an adjusted version of the SE to control the over-shrinking problem in it that is defined as

$$\hat{\boldsymbol{\gamma}}^{\text{PSE}} = \hat{\boldsymbol{\gamma}}^{\text{UE}} - \max\Big(0, \frac{k_2 - 2}{\mathcal{R}_n}\Big)(\hat{\boldsymbol{\gamma}}^{\text{UE}} - \hat{\boldsymbol{\gamma}}^{\text{RE}}). \tag{3.14}$$

The advantages of the shrinkage are discussed in Ahmed (2014).

## 3.4 Bias and Risk Analysis

This section addresses the development of the ADB and ADR of all estimators proposed in Sect. 3.3 by considering the sequence of local alternatives (SLA) as

$$\mathcal{A}_{(n)} : \mathcal{Q}\boldsymbol{\gamma} = \boldsymbol{q} + (n)^{-\frac{1}{2}}\mathcal{J}, \quad n = 1, 2, \ldots \tag{3.15}$$

where $\mathcal{J}$ is the vector of order $k_2 \times 1$ with constant elements. Let $\hat{\boldsymbol{\gamma}}^{\Diamond}$ indicate any of the suggested estimators. The ADB of $\hat{\boldsymbol{\gamma}}^{\Diamond}$ denoted by $B(\hat{\boldsymbol{\gamma}}^{\Diamond})$ has the following form:

$$B(\hat{\boldsymbol{\gamma}}^{\Diamond}) = \lim_{n\to\infty} \mathrm{E}\left[n^{1/2}\left(\hat{\boldsymbol{\gamma}}^{\Diamond} - \boldsymbol{\gamma}\right)\right].$$

To define the ADR of $\hat{\boldsymbol{\gamma}}^{\Diamond}$, we use a weighted quadratic loss function given as

$$L(\hat{\boldsymbol{\gamma}}^{\Diamond}, \boldsymbol{\gamma}; \boldsymbol{U}) = \left(n^{1/2}(\hat{\boldsymbol{\gamma}}^{\Diamond} - \boldsymbol{\gamma})\right)^{\top} \boldsymbol{U} \left(n^{1/2}(\hat{\boldsymbol{\gamma}}^{\Diamond} - \boldsymbol{\gamma})\right), \tag{3.16}$$

where $\boldsymbol{U}$ indicates a weight matrix. In our simulation studies, we use the usual quadratic loss function which is a special case of $L(\hat{\boldsymbol{\gamma}}^{\Diamond}, \boldsymbol{\gamma}; \boldsymbol{U})$, when $\boldsymbol{U} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Now, the ADR of $\hat{\boldsymbol{\gamma}}^{\Diamond}$, denoted by $R(\hat{\boldsymbol{\gamma}}^{\Diamond})$, can be defined as

$$R(\hat{\boldsymbol{\gamma}}^{\Diamond}, \boldsymbol{U}) = \int \cdots \int \boldsymbol{x}^{\top} \boldsymbol{U}\, \boldsymbol{x}\, \mathrm{d}F(\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{U}\boldsymbol{V}(\hat{\boldsymbol{\gamma}}^{\Diamond})), \tag{3.17}$$

where $F(\boldsymbol{x})$ is the cumulative distribution function of $\hat{\boldsymbol{\gamma}}^{\Diamond}$ under the SLA that is defined as

$$F(\boldsymbol{x}) = \lim_{n\to\infty} \mathbb{P}\big(n^{1/2}(\hat{\boldsymbol{\gamma}}^{\Diamond} - \boldsymbol{\gamma}) \le \frac{\boldsymbol{x}}{\mathcal{A}_{(n)}}\big).$$

Also, $\boldsymbol{V}(\hat{\boldsymbol{\gamma}}^{\Diamond})$ is the asymptotic covariance of $\hat{\boldsymbol{\gamma}}^{\Diamond}$ which is given as

$$\boldsymbol{V}(\hat{\boldsymbol{\gamma}}^{\Diamond}) = \lim_{n\to\infty} \mathrm{E}\left[n^{1/2}(\hat{\boldsymbol{\gamma}}^{\Diamond} - \boldsymbol{\gamma})\, n^{1/2}(\hat{\boldsymbol{\gamma}}^{\Diamond} - \boldsymbol{\gamma})^{\top}\right].$$

In the following theorems, we present the ADB and ADR of our suggested estimators in the RCB model under the regularity conditions (RC) of the ML estimation which are presented in Appendix "The Regularity Conditions".

**Theorem 3.1** *Under the SLA and RC when $n \to \infty$, the ADBs of the proposed estimators are given as*

$$\begin{aligned}
B(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}) &= \mathbf{0},\\
B(\hat{\boldsymbol{\gamma}}^{\mathrm{RE}}) &= -\mathcal{H}\mathcal{J},\\
B(\hat{\boldsymbol{\gamma}}^{\mathrm{LSE}}) &= -\varphi\,\mathcal{H}\mathcal{J},\\
B(\hat{\boldsymbol{\gamma}}^{\mathrm{PTE}}) &= -\mathcal{H}\mathcal{J}\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha}; \Delta^*)\\
B(\hat{\boldsymbol{\gamma}}^{\mathrm{SPE}}) &= -\varphi\,\mathcal{H}\mathcal{J}\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha}; \Delta^*),
\end{aligned}$$

$$B(\hat{\boldsymbol{\gamma}}^{\mathrm{SE}}) = -(k_2-2)\,\mathcal{H}\mathcal{J}\,\mathrm{E}\left[\frac{1}{\chi^2_{k_2+2}(\Delta^*)}\right],$$

$$B(\hat{\boldsymbol{\gamma}}^{\mathrm{PSE}}) = B(\hat{\boldsymbol{\gamma}}^{\mathrm{SE}}) - \mathcal{H}\mathcal{J}\,\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha};\,\Delta^*) + (k_2-2)\,\mathcal{H}\mathcal{J}\,\mathrm{E}\left[\frac{I(\chi^2_{k_2+2}(\Delta^*) < k_2-2)}{\chi^2_{k_2+2}(\Delta^*)}\right],$$

*where* $\mathcal{H} = \mathcal{I}^{-1}\boldsymbol{Q}^\top\left(\boldsymbol{Q}\,\mathcal{I}^{-1}\boldsymbol{Q}^\top\right)^{-1}$ *and* $\boldsymbol{\Psi}_v(\cdot;\,\Delta^*)$ *is the cumulative distribution function of the* $\chi^2_v(\Delta^*)$ *distribution.*

***Proof*** See Zandi et al. (2024) for a thorough explanation of the proof. □

**Theorem 3.2** *Under the SLA and RC when* $n \to \infty$*, the ADRs of the proposed estimators are given as*

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}};\boldsymbol{U}) = \mathrm{tr}\left(\boldsymbol{U}\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma})\right),$$

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{RE}};\boldsymbol{U}) = R(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}};\boldsymbol{U}) - \mathrm{tr}(\boldsymbol{U}\,\mathcal{H}\boldsymbol{Q}\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma})) + \Delta^*_U,$$

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{LSE}};\boldsymbol{U}) = R(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}};\boldsymbol{U}) - \varphi\,(2-\varphi)\;\mathrm{tr}(\boldsymbol{U}\,\mathcal{H}\boldsymbol{Q}\mathbf{I}^{-1}) + \varphi^2\Delta^*_U,$$

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{PTE}};\boldsymbol{U}) = R(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}};\boldsymbol{U}) - \mathrm{tr}(\boldsymbol{U}\,\mathcal{H}\boldsymbol{Q}\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma}))\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha};\,\Delta^*) + \Delta^*_U\,[2\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha};\,\Delta^*) - \boldsymbol{\Psi}_{k_2+4}(\chi^2_{k_2,\alpha};\,\Delta^*)],$$

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{SPE}};\boldsymbol{U}) = R(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}};\boldsymbol{U}) - \varphi\,(2-\varphi)\;\mathrm{tr}(\boldsymbol{U}\,\mathcal{H}\boldsymbol{Q}\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma}))\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha};\,\Delta^*) + \Delta^*_U\,[2\,\varphi\boldsymbol{\Psi}_{k_2+2}(\chi^2_{k_2,\alpha};\,\Delta^*) - \varphi\,(2-\varphi)\boldsymbol{\Psi}_{k_2+4}(\chi^2_{k_2,\alpha};\,\Delta^*)],$$

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{SE}};\boldsymbol{U}) = R(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}};\boldsymbol{U}) - 2\,(k_2-2)\;\mathrm{tr}(\boldsymbol{U}\,\mathcal{H}\boldsymbol{Q}\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma}))\left\{\mathrm{E}\left[\frac{1}{\chi^2_{k_2+2}(\Delta^*)}\right] - (k_2-2)\;\mathrm{E}\left[\frac{1}{(\chi^2_{k_2+2}(\Delta^*))^2}\right]\right\} + (k_2-2)\,\Delta^*_U\left\{2\;\mathrm{E}\left[\frac{1}{\chi^2_{k_2+2}(\Delta^*)}\right] - 2\;\mathrm{E}\left[\frac{1}{(\chi^2_{k_2+2}(\Delta^*))^2}\right] + (k_2-2)\;\mathrm{E}\left[\frac{1}{(\chi^2_{k_2+4}(\Delta^*))^2}\right]\right\},$$

$$R(\hat{\boldsymbol{\gamma}}^{\mathrm{PSE}};\boldsymbol{U}) = R(\hat{\boldsymbol{\gamma}}^{\mathrm{SE}};\boldsymbol{U}) - \Delta^*_U\;\mathrm{E}\left[\left(1-\frac{k_2-2}{\chi^2_{k_2+4}(\Delta^*)}\right)^2\right.$$

$$\times I(\chi^2_{k_2+4}(\Delta^*) < (k_2-2))\Bigg]$$

$$-\operatorname{tr}(\boldsymbol{U}\,\mathcal{H}\mathcal{Q}\,\boldsymbol{I}^{-1}(\boldsymbol{\gamma}))\ \mathrm{E}\left[\left(1-\frac{k_2-2}{\chi^2_{k_2+2}(\Delta^*)}\right)^2 I(\chi^2_{k_2+4}(\Delta^*) < (k_2-2))\right]$$

$$+2\,\Delta^*_U\ \mathrm{E}\left[\left(1-\frac{p_2-2}{\chi^2_{k_2+4}(\Delta^*)}\right) I(\chi^2_{k_2+4}(\Delta^*) < k_2-2)\right],$$

*where* $\Delta^*_U = \mathcal{J}^\top(\mathcal{H}^\top \boldsymbol{U}\,\mathcal{H})\,\mathcal{J}$ *and* $I(\cdot)$ *is an indicator function.*

***Proof*** See Zandi et al. (2024) for a thorough explanation of the proof. □

## 3.5 Simulation Study

In this section, we explore the performance of the RE and shrinkage estimators in the RCB model on simulated data by conducting an extensive Monte Carlo simulation experiment. We compare the performance of the suggested estimators with respect to the UE using simulated relative efficiency (SRE) criterion which is defined as

$$\mathrm{SRE}(\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}, \hat{\boldsymbol{\gamma}}^{\diamond}) = \frac{\mathrm{SMSE}(\hat{\boldsymbol{\gamma}})}{\mathrm{SMSE}(\hat{\boldsymbol{\gamma}}^{\diamond})},$$

where $\hat{\boldsymbol{\gamma}}^{\diamond}$ can be any one of the RE, LSE, PTS, SPE, SE, and PSE proposed in Sect. 3.3 and $\mathrm{SMSE}(\hat{\boldsymbol{\gamma}}^{\diamond})$ is defined as the simulated mean squared error of $\hat{\boldsymbol{\gamma}}^{\diamond}$:

$$\mathrm{SMSE}(\hat{\boldsymbol{\gamma}}^{\diamond}) = \frac{1}{1000}\sum_{s=1}^{1000}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}^{\diamond})^\top_s\,(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}^{\diamond})_s.$$

The value of SRE larger than one indicates that $\hat{\boldsymbol{\gamma}}^{\diamond}$ performs better than the UE. We use R statistical software to conduct the simulation experiments. We consider two different sample sizes $n = 150, 200$ and distinct censoring values $T = 4, 5, 6$, to generate the RCB model as

$$\ln(\theta_i) = (\boldsymbol{x}_i^\top \boldsymbol{\gamma})\exp\{\boldsymbol{x}_i^\top \boldsymbol{\gamma}\}, \quad i = 1, \ldots, n,$$

where the $i$th predictor variable $\boldsymbol{x}_i^\top$ (for $i = 1, \ldots, n$) is generated from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. We maximize the log-likelihood function given by Eq. (3.6) using the *optim* function to calculate $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$. We suppose the RCB model is sparse. Thus, we partition the parameter vector $\boldsymbol{\gamma}$ into two sub-vectors $\boldsymbol{\gamma}_1$ and

$\boldsymbol{\gamma}_2$ as $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top$, where $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ comprise active and inactive parameters, respectively, with the elements

$$\boldsymbol{\gamma}_1 = (\gamma_1, \gamma_2, \ldots, \gamma_{k_1})^\top, \qquad \boldsymbol{\gamma}_2 = (\underbrace{\gamma_{k_1+1}, \gamma_{k_1+2}, \ldots, \gamma_p}_{k_2})^\top.$$

We set $k_1 = 4$, $k_2 = 3, 5, 7$, and $p = k_1 + k_2 = 7, 9, 11$ in our simulations. Hence, the linear restriction given by Eq. (3.1) for a special case when $k_2 = 7$ is given as

$$\begin{pmatrix} 0&0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&1 \end{pmatrix} \times \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \\ \gamma_6 \\ \gamma_7 \\ \gamma_8 \\ \gamma_9 \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Therefore, the linear restriction given by Eq. (3.1) reduces to $\boldsymbol{\gamma}_2 = \mathbf{0}$ and we test $H_0 : \boldsymbol{\gamma}_2 = \mathbf{0}$ vs. $H_1 : \boldsymbol{\gamma}_2 \neq \mathbf{0}$. Under $H_0$, we define the true parameter vector as $\boldsymbol{\gamma}^{(0)} = (\boldsymbol{\gamma}_1^{(0)}, \boldsymbol{\gamma}_2^{(0)})^\top$, where $\boldsymbol{\gamma}_1^{(0)} = (0.07, 0.07, 0.05, 0.05)^\top$ and $\boldsymbol{\gamma}_2^{(0)} = \mathbf{0}_{k_2 \times 1}^\top$. We define the distance between the simulated parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^{(0)}$ in order to compare the performance of the suggested estimators when $H_0$ is true and otherwise as

$$\Delta = \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)}\|^2$$

where $\|\cdot\|$ is the Euclidean norm. We carry out the simulations for different values of $\Delta = 0, 0.2, 0.4, 0.6, 1$. As well, we select $\alpha = 0.05, 0.5$ and $\varphi = 0.5, 0.75$. The number of replications is set to 1, 000 for all cases.

We report the SREs of the suggested estimators in Tables 3.1, 3.2, and 3.3 and Figs. 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6.

**Finding Results**

Our results are summarized as the following items:

- The SRE of all suggested estimators is larger for smaller sample size.
- The RE and shrinkage estimators perform better for greater value of the censoring constant, $T$.
- Increasing the number of inactive predictors, $k_2$, leads to an increase in the SRE of all suggested estimators.
- The SREs of all estimators decrease as the value of $\Delta$ increases.

**Table 3.1** The SRE values of the suggested estimators for $T = 4$

| | | | | LSE | | PTE | | SPE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\varphi$ | | $\alpha$ | | $\alpha$ | | | |
| $n$ | $k_2$ | $\Delta$ | RE | 0.5 | 0.75 | 0.05 | 0.5 | 0.05 | 0.5 | SE | PSE |
| 150 | 3 | 0.0 | 1.586 | 1.394 | 1.540 | 1.518 | 1.130 | 1.353 | 1.123 | 1.078 | 1.258 |
| | | 0.2 | 0.947 | 1.206 | 1.120 | 0.916 | 0.987 | 1.090 | 0.993 | 1.058 | 1.084 |
| | | 0.4 | 0.428 | 0.866 | 0.618 | 0.794 | 0.984 | 0.925 | 0.989 | 1.002 | 1.007 |
| | | 0.6 | 0.224 | 0.591 | 0.354 | 0.956 | 1.000 | 0.983 | 1.000 | 1.002 | 1.002 |
| | | 1.0 | 0.088 | 0.293 | 0.150 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| | 5 | 0.0 | 1.990 | 1.610 | 1.890 | 1.941 | 1.259 | 1.585 | 1.243 | 1.426 | 1.702 |
| | | 0.2 | 1.179 | 1.412 | 1.381 | 1.109 | 1.006 | 1.271 | 1.027 | 1.210 | 1.284 |
| | | 0.4 | 0.533 | 1.035 | 0.767 | 0.789 | 0.984 | 0.948 | 0.990 | 1.068 | 1.071 |
| | | 0.6 | 0.279 | 0.717 | 0.441 | 0.966 | 1.000 | 0.988 | 1.000 | 1.033 | 1.033 |
| | | 1.0 | 0.110 | 0.362 | 0.187 | 1.000 | 1.000 | 1.000 | 1.000 | 1.013 | 1.013 |
| | 7 | 0.0 | 2.426 | 1.805 | 2.247 | 2.308 | 1.350 | 1.755 | 1.324 | 1.622 | 2.085 |
| | | 0.2 | 1.461 | 1.605 | 1.670 | 1.298 | 1.057 | 1.401 | 1.079 | 1.345 | 1.502 |
| | | 0.4 | 0.665 | 1.206 | 0.943 | 0.812 | 0.978 | 0.992 | 0.988 | 1.145 | 1.151 |
| | | 0.6 | 0.348 | 0.853 | 0.546 | 0.936 | 0.999 | 0.981 | 0.999 | 1.072 | 1.072 |
| | | 1.0 | 0.138 | 0.441 | 0.233 | 1.000 | 1.000 | 1.000 | 1.000 | 1.030 | 1.030 |
| 200 | 3 | 0.0 | 1.533 | 1.356 | 1.487 | 1.493 | 1.122 | 1.332 | 1.114 | 1.181 | 1.244 |
| | | 0.2 | 0.842 | 1.144 | 1.024 | 0.850 | 0.966 | 1.025 | 0.981 | 1.032 | 1.051 |
| | | 0.4 | 0.359 | 0.783 | 0.532 | 0.885 | 0.998 | 0.956 | 0.999 | 1.008 | 1.008 |
| | | 0.6 | 0.183 | 0.514 | 0.296 | 0.992 | 0.998 | 0.997 | 0.998 | 1.002 | 1.002 |
| | | 1.0 | 0.071 | 0.245 | 0.122 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| | 5 | 0.0 | 1.850 | 1.534 | 1.765 | 1.772 | 1.273 | 1.492 | 1.254 | 1.361 | 1.612 |
| | | 0.2 | 1.013 | 1.310 | 1.223 | 0.960 | 0.976 | 1.151 | 0.998 | 1.141 | 1.191 |
| | | 0.4 | 0.430 | 0.911 | 0.637 | 0.848 | 0.994 | 0.952 | 0.996 | 1.045 | 1.045 |
| | | 0.6 | 0.220 | 0.604 | 0.354 | 0.997 | 1.000 | 0.999 | 1.000 | 1.022 | 1.022 |
| | | 1.0 | 0.086 | 0.291 | 0.146 | 1.000 | 1.000 | 1.000 | 1.000 | 1.009 | 1.009 |
| | 7 | 0.0 | 2.197 | 1.707 | 2.062 | 2.089 | 1.368 | 1.657 | 1.342 | 1.459 | 1.962 |
| | | 0.2 | 1.210 | 1.480 | 1.446 | 1.112 | 1.004 | 1.290 | 1.030 | 1.244 | 1.352 |
| | | 0.4 | 0.515 | 1.050 | 0.759 | 0.822 | 0.989 | 0.961 | 0.993 | 1.094 | 1.096 |
| | | 0.6 | 0.263 | 0.707 | 0.423 | 0.987 | 1.000 | 0.996 | 1.000 | 1.046 | 1.046 |
| | | 1.0 | 0.103 | 0.345 | 0.175 | 1.000 | 1.000 | 1.000 | 1.000 | 1.019 | 1.019 |

- The RE performs best for all inactive predictors and censoring constants when $\Delta$ is equal to 0.
- The SRE of the LSE varies based on $\varphi$. A higher $\varphi$ results in an increased SRE for the LSE.
- The PTE performs better for smaller $\alpha$.
- The PSE estimator outperforms the SE when $\Delta$ is equal to 0 and around zero.

**Table 3.2** The SRE values of the suggested estimators for $T = 5$

| | | | | LSE | | PTE | | SPE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\varphi$ | | $\alpha$ | | $\alpha$ | | | |
| $n$ | $k_2$ | $\Delta$ | RE | 0.5 | 0.75 | 0.05 | 0.5 | 0.05 | 0.5 | SE | PSE |
| 150 | 3 | 0.0 | 1.615 | 1.414 | 1.568 | 1.387 | 1.052 | 1.272 | 1.049 | 1.096 | 1.181 |
| | | 0.2 | 1.029 | 1.243 | 1.187 | 0.965 | 0.990 | 1.048 | 0.997 | 1.061 | 1.072 |
| | | 0.4 | 0.493 | 0.928 | 0.693 | 0.895 | 0.993 | 0.961 | 0.695 | 1.010 | 1.010 |
| | | 0.6 | 0.264 | 0.655 | 0.410 | 0.989 | 1.000 | 0.996 | 1.000 | 1.003 | 1.003 |
| | | 1.0 | 0.106 | 0.339 | 0.178 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| | 5 | 0.0 | 2.005 | 1.620 | 1.905 | 1.566 | 1.097 | 1.382 | 1.091 | 1.417 | 1.476 |
| | | 0.2 | 1.268 | 1.452 | 1.456 | 1.059 | 1.006 | 1.126 | 1.014 | 1.230 | 1.242 |
| | | 0.4 | 0.607 | 1.108 | 0.855 | 0.896 | 0.995 | 0.973 | 0.997 | 1.077 | 1.079 |
| | | 0.6 | 0.325 | 0.795 | 0.507 | 0.989 | 1.000 | 0.996 | 1.000 | 1.037 | 1.037 |
| | | 1.0 | 0.131 | 0.417 | 0.220 | 1.000 | 1.000 | 1.000 | 1.000 | 1.015 | 1.015 |
| | 7 | 0.0 | 2.403 | 1.804 | 2.238 | 1.757 | 1.101 | 1.490 | 1.096 | 1.698 | 1.757 |
| | | 0.2 | 1.551 | 1.633 | 1.736 | 1.173 | 1.023 | 1.206 | 1.029 | 1.378 | 1.428 |
| | | 0.4 | 0.749 | 1.273 | 1.036 | 0.915 | 0.994 | 0.993 | 0.997 | 1.158 | 1.161 |
| | | 0.6 | 0.402 | 0.930 | 0.619 | 0.989 | 1.000 | 0.996 | 1.000 | 1.078 | 1.078 |
| | | 1.0 | 0.162 | 0.500 | 0.271 | 1.000 | 1.000 | 1.000 | 1.000 | 1.031 | 1.031 |
| 200 | 3 | 0.0 | 1.549 | 1.369 | 1.504 | 1.345 | 1.047 | 1.242 | 1.044 | 1.141 | 1.165 |
| | | 0.2 | 0.917 | 1.180 | 1.087 | 0.909 | 0.983 | 1.009 | 0.990 | 1.038 | 1.047 |
| | | 0.4 | 0.414 | 0.844 | 0.599 | 0.946 | 0.999 | 0.980 | 0.999 | 1.009 | 1.009 |
| | | 0.6 | 0.216 | 0.574 | 0.343 | 1.000 | 1.000 | 1.000 | 1.000 | 1.003 | 1.003 |
| | | 1.0 | 0.086 | 0.285 | 0.145 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| | 5 | 0.0 | 1.927 | 1.578 | 1.835 | 1.519 | 1.083 | 1.352 | 1.078 | 1.388 | 1.437 |
| | | 0.2 | 1.134 | 1.380 | 1.335 | 0.998 | 0.998 | 1.083 | 1.004 | 1.177 | 1.193 |
| | | 0.4 | 0.509 | 1.004 | 0.736 | 0.945 | 0.995 | 0.982 | 0.997 | 1.054 | 1.055 |
| | | 0.6 | 0.266 | 0.691 | 0.422 | 0.993 | 0.999 | 0.997 | 0.999 | 1.023 | 1.023 |
| | | 1.0 | 0.105 | 0.346 | 0.178 | 1.000 | 1.000 | 1.000 | 1.000 | 1.009 | 1.009 |
| | 7 | 0.0 | 2.277 | 1.745 | 2.130 | 1.756 | 1.108 | 1.487 | 1.102 | 1.638 | 1.716 |
| | | 0.2 | 1.341 | 1.546 | 1.563 | 1.074 | 1.004 | 1.139 | 1.012 | 1.326 | 1.347 |
| | | 0.4 | 0.599 | 1.143 | 0.864 | 0.931 | 0.998 | 0.983 | 0.999 | 1.122 | 1.123 |
| | | 0.6 | 0.312 | 0.796 | 0.495 | 0.994 | 0.999 | 0.998 | 0.999 | 1.058 | 1.058 |
| | | 1.0 | 0.123 | 0.403 | 0.209 | 1.000 | 1.000 | 1.000 | 1.000 | 1.023 | 1.023 |

## 3.6 Doctor Visits Data

We apply doctor visits data originating from the 1977–1978 Australian Health Survey to illustrate the performance of the suggested estimator in RCB model. The data include $n = 5,190$ observations. The response variable is the number of doctor visits of illnesses in the past two weeks and predictor variables are defined as $x_1$, gender; $x_2$, age in years divided by 100; $x_3$, annual income in tens of thousands of dollars; $x_4$, number of illnesses in the past two weeks; $x_5$, number of days of

**Table 3.3** The SRE values of the suggested estimators for $T = 6$

| | | | | LSE | | PTE | | SPE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\varphi$ | | $\alpha$ | | $\alpha$ | | | |
| $n$ | $k_2$ | $\Delta$ | RE | 0.5 | 0.75 | 0.05 | 0.5 | 0.05 | 0.5 | SE | PSE |
| 150 | 3 | 0.0 | 1.610 | 1.412 | 1.564 | 1.244 | 1.027 | 1.178 | 1.026 | 1.127 | 1.135 |
| | | 0.2 | 1.081 | 1.267 | 1.228 | 0.985 | 0.998 | 1.035 | 1.001 | 1.050 | 1.063 |
| | | 0.4 | 0.544 | 0.978 | 0.752 | 0.924 | 0.996 | 0.972 | 0.997 | 1.012 | 1.012 |
| | | 0.6 | 0.298 | 0.711 | 0.457 | 0.996 | 1.000 | 0.998 | 1.000 | 1.004 | 1.004 |
| | | 1.0 | 0.122 | 0.380 | 0.203 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| | 5 | 0.0 | 1.975 | 1.611 | 1.885 | 1.337 | 1.034 | 1.240 | 1.032 | 1.336 | 1.352 |
| | | 0.2 | 1.334 | 1.469 | 1.499 | 1.058 | 1.004 | 1.084 | 1.007 | 1.205 | 1.210 |
| | | 0.4 | 0.677 | 1.160 | 0.930 | 0.954 | 0.996 | 0.990 | 0.997 | 1.077 | 1.077 |
| | | 0.6 | 0.372 | 0.859 | 0.569 | 0.992 | 1.000 | 0.997 | 1.000 | 1.036 | 1.036 |
| | | 1.0 | 0.152 | 0.469 | 0.254 | 1.000 | 1.000 | 1.000 | 1.000 | 1.014 | 1.014 |
| | 7 | 0.0 | 2.343 | 1.785 | 2.196 | 1.387 | 1.038 | 1.273 | 1.036 | 1.527 | 1.546 |
| | | 0.2 | 1.602 | 1.637 | 1.763 | 1.117 | 1.011 | 1.122 | 1.012 | 1.350 | 1.357 |
| | | 0.4 | 0.819 | 1.314 | 1.107 | 0.966 | 0.999 | 0.999 | 0.999 | 1.154 | 1.154 |
| | | 0.6 | 0.451 | 0.989 | 0.683 | 0.999 | 1.000 | 1.000 | 1.000 | 1.076 | 1.076 |
| | | 1.0 | 0.185 | 0.553 | 0.307 | 1.000 | 1.000 | 1.000 | 1.000 | 1.028 | 1.028 |
| 200 | 3 | 0.0 | 1.574 | 1.385 | 1.527 | 1.222 | 1.021 | 1.161 | 1.020 | 1.108 | 1.123 |
| | | 0.2 | 0.992 | 1.219 | 1.152 | 0.956 | 0.993 | 1.009 | 0.997 | 1.041 | 1.046 |
| | | 0.4 | 0.470 | 0.905 | 0.666 | 0.972 | 0.999 | 0.989 | 0.999 | 1.008 | 1.009 |
| | | 0.6 | 0.251 | 0.635 | 0.392 | 1.000 | 1.000 | 1.000 | 1.000 | 1.003 | 1.003 |
| | | 1.0 | 0.101 | 0.325 | 0.169 | 1.000 | 1.000 | 1.000 | 1.000 | 1.001 | 1.001 |
| | 5 | 0.0 | 1.933 | 1.583 | 1.842 | 1.307 | 1.036 | 1.219 | 1.034 | 1.313 | 1.331 |
| | | 0.2 | 1.209 | 1.407 | 1.392 | 1.008 | 1.000 | 1.044 | 1.003 | 1.162 | 1.169 |
| | | 0.4 | 0.570 | 1.059 | 0.806 | 0.970 | 0.998 | 0.990 | 0.999 | 1.055 | 1.055 |
| | | 0.6 | 0.303 | 0.751 | 0.474 | 0.997 | 1.000 | 0.999 | 1.000 | 1.025 | 1.025 |
| | | 1.0 | 0.121 | 0.389 | 0.204 | 1.000 | 1.000 | 1.000 | 1.000 | 1.009 | 1.009 |
| | 7 | 0.0 | 2.313 | 1.763 | 2.162 | 1.362 | 1.039 | 1.254 | 1.037 | 1.500 | 1.521 |
| | | 0.2 | 1.440 | 1.583 | 1.642 | 1.056 | 1.004 | 1.084 | 1.006 | 1.296 | 1.301 |
| | | 0.4 | 0.675 | 1.209 | 0.952 | 0.976 | 1.001 | 0.994 | 1.000 | 1.120 | 1.121 |
| | | 0.6 | 0.358 | 0.867 | 0.560 | 1.000 | 1.000 | 1.000 | 1.000 | 1.059 | 1.059 |
| | | 1.0 | 0.143 | 0.455 | 0.241 | 1.000 | 1.000 | 1.000 | 1.000 | 1.023 | 1.023 |

reduced activity in the past two weeks due to illness or injury; $x_6$, general health questionnaire score using Goldberg's method; $x_7$, does the individual have private health insurance?; $x_8$, does the individual have free government health insurance due to low income?; $x_9$, does the individual have free government health insurance due to old age, disability, or veteran status?; $x_{10}$, is there a chronic condition not limiting activity?; and $x_{11}$, is there a chronic condition limiting activity?

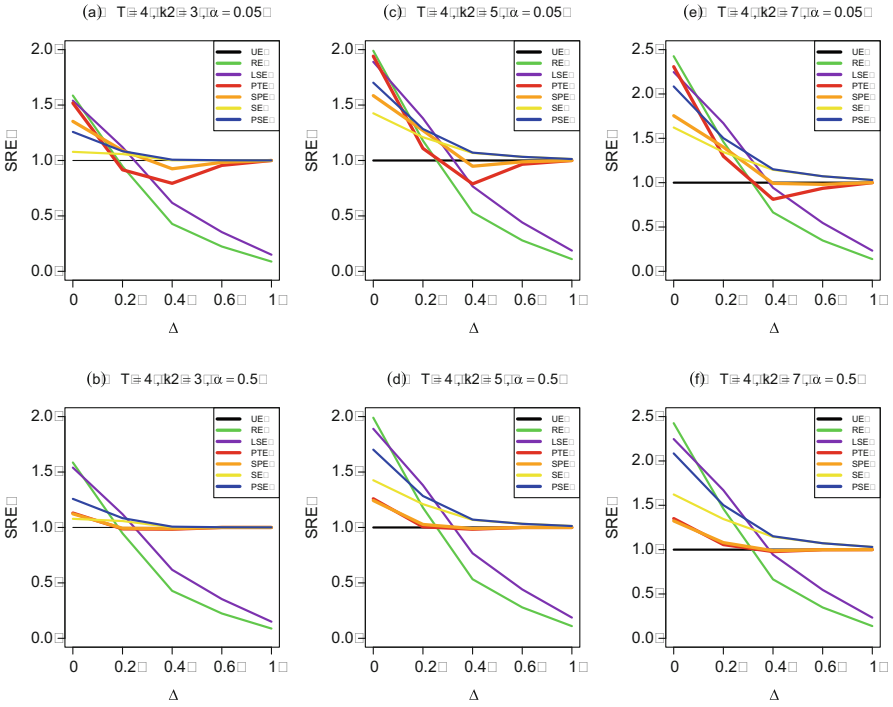



**Fig. 3.1** SREs of the proposed estimators with respect to the $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ for $n = 150$

The mean and variance of the response variable are 0.3017341 and 0.6370176, respectively, which show that the data is over-dispersed. By applying AIC to specify the significant predictor variables, the submodel contains $x_1$, $x_2$, $x_6$, and $x_7$ with the smallest value of AIC $=$ 17301.28 chosen as the true submodel. The true parameter is calculated by fitting an RCB model to data as $\boldsymbol{\gamma}^{(0)} = (0.475, -0.628, 0, 0, 0, 0.514, -0.672, 0, 0, 0, 0)^{\top}$. By applying bootstrap sampling, we choose 200 observations with replacement at 1, 000 times from the data to estimate the active parameters $\gamma_1$, $\gamma_2$, $\gamma_6$, and $\gamma_7$, related to the significant predictors for $\Delta = 0$, $\alpha = 0.01$, $T = 2$, and $\varphi = 0.75$. The results are presented in Table 3.4. The results reveal that all proposed estimators perform better than the UE except the SE which is inefficient for this special data with a lower value of SRE.

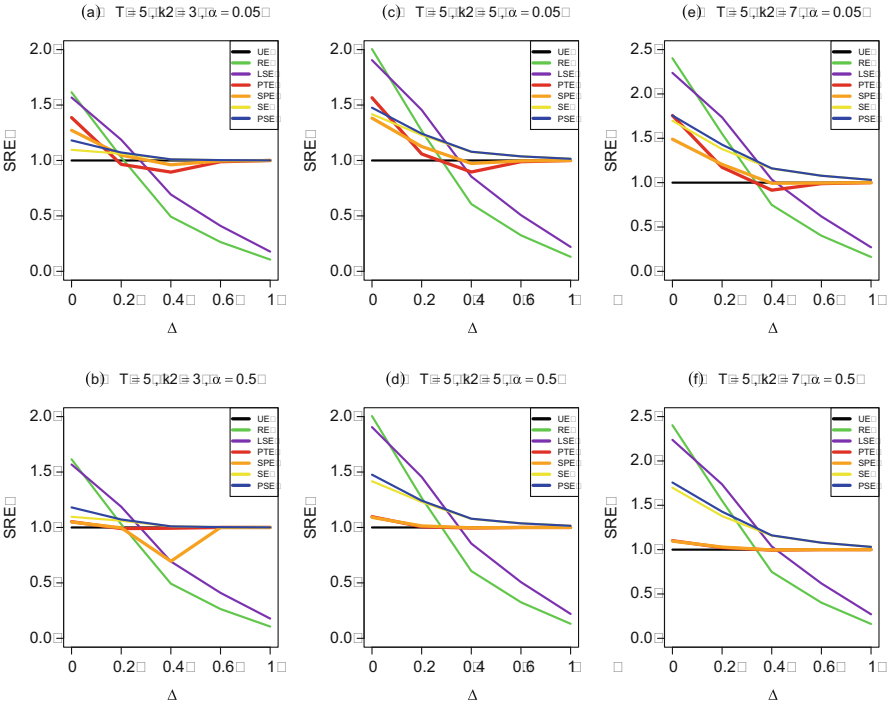



**Fig. 3.2** SREs of the proposed estimators with respect to the $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ for $n = 150$

**Table 3.4** Estimation, standard error (in the brackets), and SRE of the proposed estimators in doctor visit data

| Parameters | UE | RE | LSE | PTE |
|---|---|---|---|---|
| $\hat{\gamma}_1$ | − 0.016(0.619) | − 0.275(0.451) | − 0.210(0.470) | − 0.176(0.626) |
| $\hat{\gamma}_2$ | 0.675(1.016) | 0.225(0.949) | 0.338(0.875) | 0.404(0.945) |
| $\hat{\gamma}_6$ | 0.013(0.128) | 0.075(0.078) | 0.060(0.085) | 0.066(0.097) |
| $\hat{\gamma}_7$ | − 0.791(0.589) | − 0.911(0.404) | − 0.881(0.416) | − 0.875(0.430) |
| SRE | 1.000 | 2.257 | 2.168 | 1.584 |

| Parameters | SPE | SE | PSE |
|---|---|---|---|
| $\hat{\gamma}_1$ | − 0.136(0.611) | − 0.125(0.650) | − 0.122(0.614) |
| $\hat{\gamma}_2$ | 0.472(0.907) | 0.404(1.156) | 0.456(1.032) |
| $\hat{\gamma}_6$ | 0.052(0.100) | 0.047(0.130) | 0.059(0.109) |
| $\hat{\gamma}_7$ | − 0.854(0.440) | − 0.918(0.686) | − 0.882(0.562) |
| SRE | 1.554 | 0.914 | 1.127 |

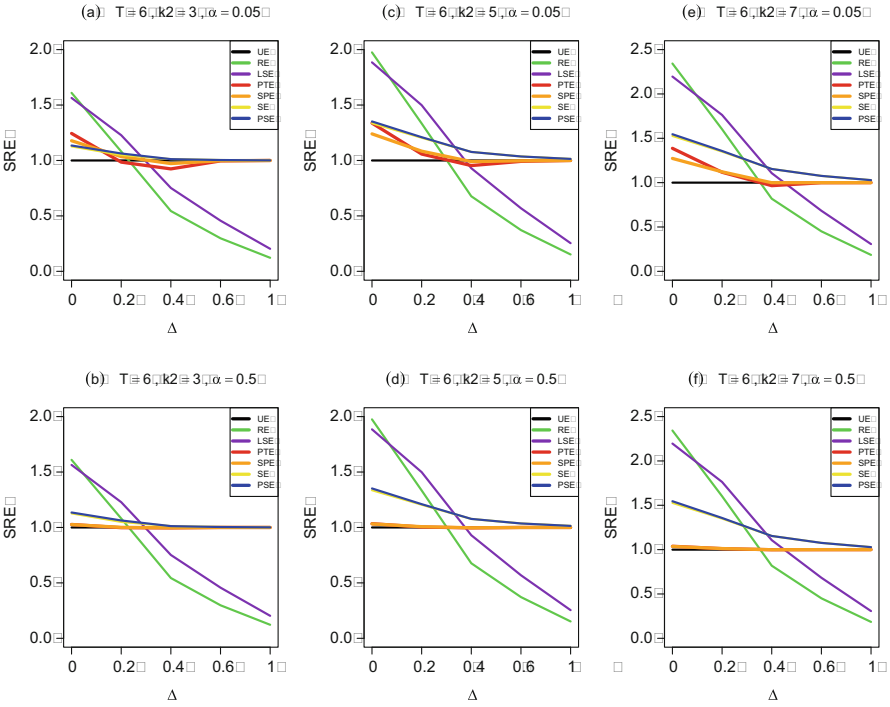



**Fig. 3.3** SREs of the proposed estimators with respect to the $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ for $n = 150$

## 3.7 Conclusion

In this chapter, we use the shrinkage estimators including linear shrinkage, preliminary test, shrinkage pretest, Stein, and positive Stein estimators to improve the efficiency of parameter estimation in the right-censored Bell model that contains some nonsignificant predictors. We develop the ADB and ADR of the suggested estimators. As well, we use a Monte Carlo simulation and a real dataset to compare the performance of the proposed estimator. The results demonstrate that by increasing censoring constant, all the proposed estimators perform better. When the submodel is chosen accurately ($\Delta = 0$), for all situations, all suggested estimators perform better than the UE, the RE has the best efficiency, the PTE performs better than the SPE, and the SRE of the PSE is greater than the SE. When $\Delta$ moves away from zero, the SRE of the RE sharply decreases and the Stein estimators have the same performance.

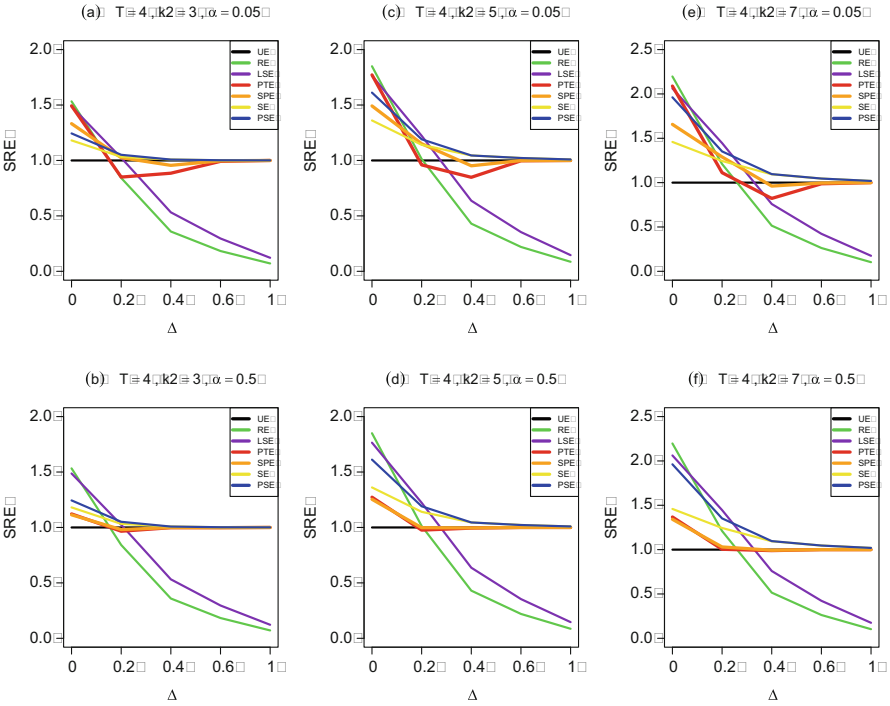



**Fig. 3.4** SREs of the proposed estimators with respect to the $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ for $n = 200$

# Appendix

## *The Fisher Information Matrix*

The Fisher information matrix in the RCB model is

$$
\begin{aligned}
\boldsymbol{I}(\boldsymbol{\gamma}) &= -\mathrm{E}\left[\frac{\partial^2 \mathcal{L}_{\mathrm{RCB}}(\boldsymbol{\gamma})}{\partial\, \gamma_j \partial\, \gamma_t}\right] \\
&= \sum_{i=1}^{n} \tau_i\,(x_{ij})\,(x_{it}) \left\{ \frac{a_i(\boldsymbol{\gamma})\, \mathcal{W}_0(e_i(\boldsymbol{\gamma}))C_1 - C_2\, C_3}{\left[\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\, a_i(\boldsymbol{\gamma})\right]^2} \right\} \\
&\quad + \sum_{i=1}^{n} (1-\tau_i)\,(x_{ij})\,(x_{it}) \\
&\quad \times \sum_{r=0}^{Y_i^*-1} \frac{\mathcal{B}_r}{r!} \left\{ \frac{A_1\big(A_2\, A_3 + A_4 + A_5\big) - A_6\big(A_7 + A_6\big)}{\left[a_i(\boldsymbol{\gamma})\, h(Y_i^*)\right]^2} \right\}
\end{aligned}
\tag{3.18}
$$

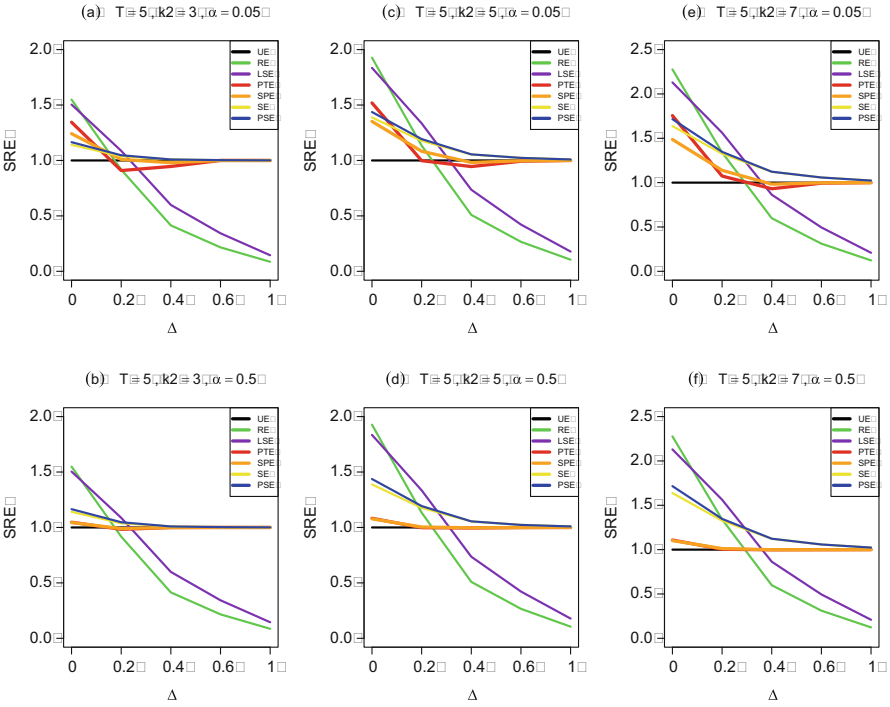



**Fig. 3.5** SREs of the proposed estimators with respect to the $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ for $n = 200$

where

$$C_1 = u_i(\boldsymbol{\gamma})\, Y_i^* - u_i^3(\boldsymbol{\gamma}) \exp\{u_i(\boldsymbol{\gamma})\}\,(u_i(\boldsymbol{\gamma}) + 2)$$

$$C_2 = u_i(\boldsymbol{\gamma})\, Y_i^* - u_i^2(\boldsymbol{\gamma}) \exp\{u_i(\boldsymbol{\gamma})\}$$

$$C_3 = e_i(\boldsymbol{\gamma})\,[2 + u_i(\boldsymbol{\gamma})]\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big) + u_i(\boldsymbol{\gamma})$$

$$A_1 = a_i(\boldsymbol{\gamma})\, h(Y_i^*)$$

$$A_2 = -\frac{u_i(\boldsymbol{\gamma})\, \exp\{\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\}\, d_i(\boldsymbol{\gamma})}{a_i(\boldsymbol{\gamma})}$$

$$A_3 = \big[\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\big]^{r-1}\big[u_i^2(\boldsymbol{\gamma})\, \exp\{u_i(\boldsymbol{\gamma})\} + r\, u_i(\boldsymbol{\gamma})\big]$$

$$A_4 = (r-1)\big[\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\big]^{r-2}\Big[\frac{u_i(\boldsymbol{\gamma})}{e_i(\boldsymbol{\gamma})}\Big]\, d_i(\boldsymbol{\gamma})\,\big[u_i^2(\boldsymbol{\gamma}) \exp\{u_i(\boldsymbol{\gamma})\} + r u_i(\boldsymbol{\gamma})\big]$$

$$A_5 = \big[2\, u_i(\boldsymbol{\gamma})\, e_i(\boldsymbol{\gamma}) + u_i^2(\boldsymbol{\gamma})\, e_i(\boldsymbol{\gamma}) + r\, u_i(\boldsymbol{\gamma})\big]\Big[d_i(\boldsymbol{\gamma})\big(\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\big)^{r-1}\Big]$$

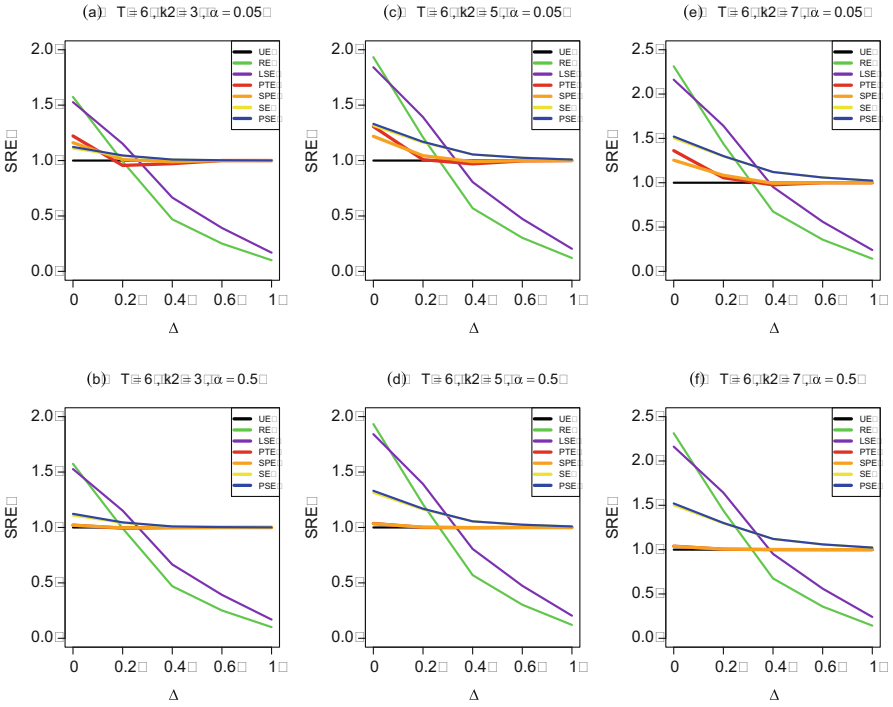



**Fig. 3.6** SREs of the proposed estimators with respect to the $\hat{\boldsymbol{\gamma}}^{\mathrm{UE}}$ for $n = 200$

$$A_6 = d_i(\boldsymbol{\gamma})\left[\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\right]^{r-1}\left[u_i^2(\boldsymbol{\gamma})\, e_i(\boldsymbol{\gamma}) + r\, u_i(\boldsymbol{\gamma})\right]$$

$$A_7 = h(Y_i^*)\left[u_i^2(\boldsymbol{\gamma})\, e_i(\boldsymbol{\gamma}) + r\, u_i(\boldsymbol{\gamma})\right]$$

$$a_i(\boldsymbol{\gamma}) = \exp\{u_i(\boldsymbol{\gamma})\} + e_i(\boldsymbol{\gamma})$$

$$d_i(\boldsymbol{\gamma}) = \exp\left\{1 - \exp\{\mathcal{W}_0\big(e_i(\boldsymbol{\gamma})\big)\}\right\}.$$

## *The Regularity Conditions*

The parameter vector $\boldsymbol{\gamma}_{p\times 1}$ is defined in an open subset $\mathcal{B}$ of $\Re^p$. The link function $\ln(\theta_i)$ has continuous partial derivative with respect to $\boldsymbol{\gamma}$ and $\theta_i$ up to the third order.

There exists a unique MLE (or unrestricted estimator) of $\boldsymbol{\gamma}$ for $\mathcal{L}_{\mathrm{RCB}}(\boldsymbol{\gamma})$. The moments of $\partial\, \mathcal{L}_{\mathrm{RCB}}(\boldsymbol{\gamma}) / \partial\, \boldsymbol{\gamma}$ exist at least up to the third order.

The information matrix $\boldsymbol{I}(\boldsymbol{\gamma})$ is finite and positive definite at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$, where $\boldsymbol{\gamma}^0$ is the true vector of parameters.

## References

Aitchison, J., & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics, 29*(3), 813–828.

Ahmed, S. E. (1992). Shrinkage preliminary test estimation in multivariate normal distributions. *Journal of Statistical Computation and Simulation, 43*(3–4), 177–195.

Ahmed, S. E. (2014). *Penalty, shrinkage and pretest strategies: Variable selection and estimation*. Springer.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics, 15*, 190–204.

Castellares, F., Ferrari, S. L., & Lemonte, A. J. (2018). On the Bell distribution and its associated regression model for count data. *Applied Mathematical Modelling, 56*, 172–185.

Famoye, F., & Wang, W. (2004). Censored generalized Poisson regression model. *Computational Statistics and Data Analysis, 46*(3), 547–560.

Hossain, S., Ahmed, S. E., & Doksum, K. A. (2015). Shrinkage, pretest, and penalty estimators in generalized linear models. *Journal of Statistical Methodology, 24*, 52–68.

Karlis, D., Papatla, P., & Roy, S. (2016). Finite mixtures of censored Poisson regression models. *Statistica Neerlandica, 70*(2), 100–122.

Saleh, A. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. John Wiley.

Seifollahi, S., Bevrani, H., & Yahya A. Z. (2024). Improved estimators in bell regression model with application, *Journal of Statistical Computation and Simulation, 94*(12), 2710–2726.

Stein, C. (1956). The admissibility of Hotelling's $T^2$-test. *Mathematical Statistics, 27*, 616–623.

Terza, J. V. (1985). A Tobit-type estimator for the censored Poisson regression model. *Economics Letters, 18*(4), 361–365.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association, 63*, 113–122.

Zandi, Z., Bevrani, H., & Arabi Belaghi, R. (2021). Improved shrinkage estimators in zero-inflated negative binomial regression model. *Hacettepe Journal of Mathematics and Statistics, 50*(6), 1855–1876.

Zandi, Z., Bevrani, H., & Arabi Belaghi, R. (2024). Shrinkage estimation in the zero-inflated Poisson regression model with right-censored data, *Communication in Statistics - Theory and Methods, 53* (13), 4898–4917.

Zandi, Z., Bevrani, H., & Arabi Belaghi, R. (2021). Using shrinkage strategies to estimate fixed effects in zero-inflated negative binomial mixed model, *Communications in Statistics - Simulation and Computation, 52*(7), 3201–3222.

# Chapter 4
# On a Class of Shrinkage Estimators of Normal Mean in High-Dimensional Data with Unknown Covariance




**Arash A. Foroushani and Sévérien Nkurunziza**

## 4.1 Introduction

In this chapter, we consider an estimation problem concerning the mean parameter vector of Gaussian samples in the context of unknown covariance matrix along with high-dimensional data. In particular, we revisit the important results in Chételat and Wells (2012) which deals with improved multivariate normal mean estimation with unknown covariance when the dimension of the parameter $p$ is greater than the sample $n$. More precisely, we establish a theorem which is useful in deriving the superiority of some shrinkage estimators, over the classical sample mean, in the context of high-dimensional data. The established result is a revised version of Theorem 2 of Chételat and Wells (2012). Namely, by the incorrect use of Cauchy-Schwarz inequality, at their page 3153, Chételat and Wells (2012) obtain an incorrect bound for the quantity $X^\top(T^+TA)^+(T^+TA)X$, and thus, there are some gaps to fill in the proof of the statistical methods of the quoted paper. Indeed, in contrast with the claim in Chételat and Wells (2012), $X^\top(T^+TA)^+(T^+TA)X \not\leq X^\top(T^+TA)^+(T^+TA)^+XX^\top(AT^+T)(T^+TA)X$. To highlight this last point, we give a counterexample which shows that the obtained bound is incorrect. We also provide an alternative proof that fills in the gaps left in the derivation of the quoted result. To this end, we derive a result which gives a sharp bound on the first moment of some random quantities which play a crucial role in the proof of Theorem 2 of Chételat and Wells (2012). Beyond such important statistical contributions, a salient feature of this chapter consists in the fact that the established results hold under a very realistic assumption about the rank of the estimator of the variance-covariance matrix. We point out that such an assumption is required for the method

A. A. Foroushani · S. Nkurunziza (✉)
Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada
e-mail: aghaeif@uwindsor.ca; severien@uwindsor.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_4

in Chételat and Wells (2012) to hold. To better highlight this last point, we give a counterexample which shows that without this additional assumption, the main result in the quoted paper does not hold.

The rest of this chapter is organized as follows. In Sect. 4.2, we present the statistical model and some fundamental results. Section 4.3 is about the main result of this chapter. This section is subdivided into three subsections. Specifically, Sect. 4.3.1 reformulates Theorem 2 in Chételat and Wells (2012) and we revise, in this subsection, the proof given in the quoted paper. Further, Sect. 4.3.2 gives a counterexample which shows that the inequality used in Chételat and Wells (2012) does not hold. Finally, in Sect. 4.3.3, we give an example which shows that without imposing an assumption about the rank of the estimator of the covariance-variance matrix, an important random quantity involved in the proof of Theorem 2 of Chételat and Wells (2012) has infinite expectation. Finally, in Sect. 4.4, we give some concluding remarks.

## 4.2 Statistical Model and Some Fundamental Results

In this section, we present the statistical model and set up some notations used in this chapter. To introduce some notations, let $\boldsymbol{I}_q$ denote the $q$-dimensional identity matrix, let $\boldsymbol{\Sigma}$ be a positive definite matrix, and let $\boldsymbol{\theta}$ be a $p$-column vector. For a given $p_1 \times p_2$-matrix $\boldsymbol{A}$, let $\mathrm{Vec}(\boldsymbol{A})$ stand for a vectorization operator that transforms the matrix into a column vector by vertically stacking the columns of the matrix $\boldsymbol{A}$. Further, let $\boldsymbol{A} \otimes \boldsymbol{B}$ denote the Kronecker product of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, let $\boldsymbol{U} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a $p$-column Gaussian random vector $\boldsymbol{U}$ with means $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and, for a random matrix $\boldsymbol{U}$, let $\boldsymbol{U} \sim \mathcal{N}_{p_1 \times p_2}(\boldsymbol{\mu}, \boldsymbol{A}_1 \otimes \boldsymbol{A}_2)$ stand for $\mathrm{Vec}(\boldsymbol{U}) \sim \mathcal{N}_{p_1 p_2}(\mathrm{Vec}(\boldsymbol{\mu}), \boldsymbol{A}_2 \otimes \boldsymbol{A}_1)$. We also denote $\boldsymbol{U} \sim \mathcal{W}_p(m, \boldsymbol{\Sigma})$ to stand for the $p \times p$-random matrix that follows Wishart distribution with degrees of freedom $m$ and scale parameter $\boldsymbol{\Sigma}$. For $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, let

$$\nabla_{\boldsymbol{A}} = \left(\frac{\partial}{\partial A_{ij}}\right)_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n},$$

$$\mathrm{div}_{\boldsymbol{A}} \boldsymbol{B} = \nabla_{\boldsymbol{A}} . \boldsymbol{B} = \mathrm{div}_{\mathrm{vec}(\boldsymbol{A})} \mathrm{vec}(\boldsymbol{B}) = \sum_{i,j} \frac{\partial B_{ij}}{\partial A_{ij}},$$

$$(\nabla_{\boldsymbol{A}} \boldsymbol{B})_{ij} = \sum_{\alpha} (\nabla_{\boldsymbol{A}})_{i\alpha} B_{\alpha j} = \sum_{\alpha} \frac{\partial B_{\alpha j}}{\partial A_{i\alpha}}.$$

In the sequel, we use the same notations/symbols as in Chételat and Wells (2012). In particular, we consider the scenario where we observe a $p$-random matrix $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ a positive definite matrix. In this chapter, we present a class of estimators for the mean vector $\boldsymbol{\theta}$ in the context of high-dimensional data with $\boldsymbol{\Sigma}$ unknown nuisance parameter. We suppose that the random matrix $\boldsymbol{S}$ is observed along with the matrix $\boldsymbol{X}$, and we suppose that $\boldsymbol{X}$ and $\boldsymbol{S}$ are independent in the context

where $\boldsymbol{S} \sim \mathcal{W}_p(n, \boldsymbol{\Sigma})$, with $p > n$. It should be noted that, as we are in the context of the Gaussian sampling plan, this assumption is without loss of generality. To better illustrate this point, let us consider the case where $N$ observations of $p$-column vectors $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N$ are independent and identically distributed as $\mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Xi})$. Under this sampling plan, the maximum likelihood estimators (MLEs) of $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are, respectively, given by

$$\boldsymbol{X} \equiv \bar{\boldsymbol{W}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{W}_i, \quad \boldsymbol{S} \equiv \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{W}_i - \bar{\boldsymbol{W}}\right)\left(\boldsymbol{W}_i - \bar{\boldsymbol{W}}\right)^\top. \tag{4.1}$$

Below, we rewrite $\boldsymbol{S}$ under the form of a simple quadratic form. To this end, let $\boldsymbol{W} = (\boldsymbol{W}_1 \vdots \boldsymbol{W}_2 \vdots \cdots \vdots \boldsymbol{W}_N)^\top$ and let $\boldsymbol{e}_N$ be an $N$-column vector with all components equal to 1. Together with the relation in Eq. (4.1), the matrix $\boldsymbol{S}$ can be rewritten as

$$\boldsymbol{S} = \frac{1}{N}\boldsymbol{W}^\top \left(\boldsymbol{I}_N - \frac{1}{N}\boldsymbol{e}_N\boldsymbol{e}_N^\top\right)\boldsymbol{W}. \tag{4.2}$$

From Eq. (4.2), one verifies that $\boldsymbol{S} \sim \mathcal{W}_p\left((N-1), \frac{1}{N}\boldsymbol{\Xi}\right)$. Further, since the matrix $\boldsymbol{I}_N - (1/N)\boldsymbol{e}_N\boldsymbol{e}_N^\top$ is idempotent, there exists an orthogonal matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{Q}^\top\left(\boldsymbol{I}_N - \frac{1}{N}\boldsymbol{e}_N\boldsymbol{e}_N^\top\right)\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{I}_{N-1} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{pmatrix}.$$

Let $\boldsymbol{V} = \boldsymbol{Q}^\top\boldsymbol{W}$. We have $\boldsymbol{V} \sim \mathcal{N}_{N\times p}\left(\boldsymbol{Q}^\top\left(\boldsymbol{e}_N \otimes \boldsymbol{\theta}^\top\right), \boldsymbol{I}_N \otimes \boldsymbol{\Xi}\right)$ and

$$N\,\boldsymbol{S} = \boldsymbol{V}^\top \begin{pmatrix} \boldsymbol{I}_{(N-1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{V} = \boldsymbol{V}_{(1)}^\top \boldsymbol{V}_{(1)}, \tag{4.3}$$

with

$$\boldsymbol{V}_{(1)} = \left[\boldsymbol{I}_{(N-1)} \vdots \boldsymbol{0}\right]\boldsymbol{V} \sim \mathcal{N}_{(N-1)\times p}\left(\boldsymbol{0}, \boldsymbol{I}_{(N-1)} \otimes \boldsymbol{\Xi}\right).$$

In the sequel, we let $n = N-1$, $\boldsymbol{\Sigma} = N^{-1}\boldsymbol{\Xi}$ and let $\boldsymbol{Y} = \boldsymbol{V}_{(1)}$. Thus, we can rewrite the sample covariance matrix $\boldsymbol{S}$ as $\boldsymbol{S} = \boldsymbol{Y}^\top\boldsymbol{Y}$.

Nevertheless, more research needs to be done for the cases of high- and ultrahigh-dimensional data, i.e., the case where the total number of features $p > n$. In this chapter, we solve the problem in the context of high-dimensional data. With Eqs. (4.1) and (4.3) in mind, to simplify the derivation of the main results of this chapter, we consider the statistical model that satisfies the following assumption.

**Assumption 4.1** *We observe a $p$-column random vector $\boldsymbol{X}$ independent of a $n \times p$-random matrix $\boldsymbol{Y}$ with $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} \sim \mathcal{N}_{n\times p}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a*

*$p \times p$-symmetric and positive definite matrix. We let $\boldsymbol{S} = \boldsymbol{Y}^\top \boldsymbol{Y}$ and $R = rank(\boldsymbol{S})$. We also consider that $P(R > 2) = 1$.*

It should be noticed that in contrast to the classical case where $n > p$, in high-dimensional data, the matrix $\boldsymbol{S}$ is singular with probability one. Because of that, the usual inverse $\boldsymbol{S}^{-1}$ does not exist, almost surely, while such an inverse is involved in classical James-Stein-type estimators. To overcome this problem, we adopt whenever needed the Moore-Penrose inverse. To introduce the notation, denote $\boldsymbol{A}^+$ as the Moore-Penrose inverse of the matrix $\boldsymbol{A}$, let

$$F = \boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{X}, \quad \boldsymbol{\delta}^0(\boldsymbol{X}) = \boldsymbol{X}, \tag{4.4}$$

and, for some real-valued function $r$ defined on $(0, \infty)$, let

$$\delta_r(\boldsymbol{X}, \boldsymbol{S}) = \left( \boldsymbol{I}_p - \frac{r(F)}{F} \boldsymbol{S}\boldsymbol{S}^+ \right) \boldsymbol{X}. \tag{4.5}$$

The estimator given in Eq. (4.5) is known as a Baranchik-type estimator (see Baranchik, 1970). For further details, the reader is referred to Fisher and Sun (2011), Fourdrinier et al. (2024), Haddouche et al. (2021), Karamikabir et al. (2018), and Lotfi et al. (2022).

In this chapter, we establish the optimality of the member of the class of estimators defined in Eq. (4.5). Note that the classical MLE is a member of the proposed class of estimators in Eq. (4.5) with $r = 0$; i.e., $\boldsymbol{\delta}^0(\boldsymbol{X}) = \boldsymbol{X} = \delta_0(\boldsymbol{X}, \boldsymbol{S})$. In this chapter, we derive the necessary and sufficient conditions for the estimator in Eq. (4.5) to have a finite quadratic risk function. As mentioned above, the novelty consists in the fact that we revise some main results in the above quoted paper.

## 4.3 Main Result

In this section, we give the main result of this chapter. In particular, we revise Theorem 2 of Chételat and Wells (2012). The established result is useful in deriving the superiority of the shrinkage estimators in Eq. (4.5) over the classical MLE. Nevertheless, without an assumption about the rank of the matrix $\boldsymbol{S}$, the quoted theorem is incorrect and, the fact is, due to some misuse of Cauchy-Schwarz inequality, some crucial steps in the proof given in Chételat and Wells (2012) are also incorrect. Thus, on top of revising this important result, we also correct its proof. We also give, in Sects. 4.3.2 and 4.3.3, some counterexamples which highlight the significance of the established result.

### 4.3.1 Theorem 2 of Chételat and Wells (2012) Revisited

In this subsection, we derive the revised version of Theorem 2 of Chételat and Wells (2012). As intermediate step, we first derive a result which plays a crucial role in the proof of the revised version of the quoted result. For context purposes, note that the major step in the proof of Theorem 2 of Chételat and Wells (2012) consists in proving that $\mathrm{E}[1/F] < \infty$. Nevertheless, as mentioned in Foroushani and Nkurunziza (2023), the proof given in the quoted paper relies on the incorrect use of Cauchy-Schwarz inequality and we give in Sect. 4.3.3 a counterexample which shows that if the rank of $\boldsymbol{S}$ is less than or equal to 2, $\mathrm{E}[1/F] = \infty$ (see also Foroushani & Nkurunziza, 2023). Below, we derive a theorem which proves that, in addition to the assumptions in Chételat and Wells (2012), if the rank of $S$ is greater than 2 with probability one, $\mathrm{E}[1/F] < \infty$.

**Theorem 4.1** *Suppose that Assumption 4.1 holds. Then,*

$$\mathrm{E}\left[\frac{1}{F}\right] \leqslant n\ \mathrm{tr}(\boldsymbol{\Sigma})\,\lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right).$$

***Proof*** Let $\boldsymbol{C}$ be $(R \times p)$-matrix of the form $\boldsymbol{C} = [\boldsymbol{I}_R \vdots \boldsymbol{0}_{R\times(p-R)}]$, consider $\boldsymbol{X}_{(1)} = \boldsymbol{C}\boldsymbol{X}$, let $\lambda^{\dagger}_{\min}(\boldsymbol{S}^{+})$ be the smallest strictly positive eigenvalues of $\boldsymbol{S}^{+}$, and let $\lambda^{\dagger}_{\max}(\boldsymbol{S}^{+})$ be the biggest nonzero eigenvalue of $\boldsymbol{S}^{+}$. We have

$$\lambda^{\dagger}_{\min}(\boldsymbol{S}^{+})\boldsymbol{X}^{\top}_{(1)}\boldsymbol{X}_{(1)} \leqslant F \leqslant \lambda^{\dagger}_{\max}(\boldsymbol{S}^{+})\boldsymbol{X}^{\top}_{(1)}\boldsymbol{X}_{(1)}. \tag{4.6}$$

We also have

$$\boldsymbol{X}_{(1)}|R \sim \mathcal{N}_R(\boldsymbol{C}\boldsymbol{\theta},\ \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\top}).$$

Note that, since $\boldsymbol{\Sigma}$ is a positive definite matrix, $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\top}$ is an $R \times R$-positive definite matrix, and then, we can set $\boldsymbol{A}^2 = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\top}$ where $\boldsymbol{A}$ is a positive definite matrix. Let $\boldsymbol{U} = \boldsymbol{A}^{-1}\boldsymbol{X}_{(1)}$. We have

$$\lambda_{\min}(\boldsymbol{A}^2)\boldsymbol{U}^{\top}\boldsymbol{U} \leqslant \boldsymbol{X}^{\top}_{(1)}\boldsymbol{X}_{(1)} = \boldsymbol{U}^{\top}\boldsymbol{A}^2\boldsymbol{U} \leqslant \lambda_{\max}(\boldsymbol{A}^2)\boldsymbol{U}^{\top}\boldsymbol{U}$$

Therefore, together with Eq. (4.6), we get

$$\begin{aligned}\frac{1}{\lambda^{\dagger}_{\max}(\boldsymbol{S}^{+})}\frac{1}{\lambda_{\max}(\boldsymbol{A}^2)}\frac{1}{\boldsymbol{U}^{\top}\boldsymbol{U}} \leqslant \frac{1}{F} &\leqslant \frac{1}{\lambda^{\dagger}_{\min}(\boldsymbol{S}^{+})}\frac{1}{\lambda^{\dagger}_{\min}(\boldsymbol{A}^2)}\frac{1}{\boldsymbol{U}^{\top}\boldsymbol{U}} \\ &= \lambda^{\dagger}_{\max}(\boldsymbol{S})\lambda^{\dagger}_{\max}(\boldsymbol{A}^{-2})\frac{1}{\boldsymbol{U}^{\top}\boldsymbol{U}},\end{aligned} \tag{4.7}$$

where $\lambda^{\dagger}_{\max}(\boldsymbol{S})$ and $\lambda^{\dagger}_{\max}(\boldsymbol{A}^{-2})$ are the biggest nonzero eigenvalues of $(\boldsymbol{S}^{+})^{+} = \boldsymbol{S}$ and $(\boldsymbol{A}^{2})^{+} = \boldsymbol{A}^{-2}$, respectively. We also have $\boldsymbol{U}|R \sim \mathcal{N}_R\left(\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{\theta},\ \boldsymbol{A}^{-1}\boldsymbol{A}^{2}\boldsymbol{A}^{-1}\right)$, i.e.,

$$\boldsymbol{U}|R \sim \mathcal{N}_R\left(\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{\theta},\ \boldsymbol{I}_R\right).$$

Further, note that the matrix $\boldsymbol{\Sigma}$ can be rewritten as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{A}^2 & \boldsymbol{\Sigma}_{21}^{\top} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\boldsymbol{A}^2 = [\boldsymbol{I}_R \vdots \boldsymbol{0}_{R\times(p-R)}]\boldsymbol{\Sigma}[\boldsymbol{I}_R \vdots \boldsymbol{0}_{R\times(p-R)}]^{\top}$. This shows that $\boldsymbol{A}^2$ is a principal sub-matrix of the matrix $\boldsymbol{\Sigma}$. Hence, by the inequality in Johnson and Robinson (1981), we get

$$\lambda_{\min}(\boldsymbol{A}^2) \geqslant \lambda_{\min}(\boldsymbol{\Sigma}) > 0 \tag{4.8}$$

and then

$$\frac{1}{F} \leqslant \frac{1}{\lambda^{\dagger}_{\min}(\boldsymbol{S}^{+})}\frac{1}{\lambda^{\dagger}_{\min}(\boldsymbol{\Sigma})}\frac{1}{\boldsymbol{U}^{\top}\boldsymbol{U}}. \tag{4.9}$$

Therefore,

$$\mathrm{E}\left[\frac{1}{F}\right] \leqslant \frac{1}{\lambda^{\dagger}_{\min}(\boldsymbol{\Sigma})}\,\mathrm{E}\left[\frac{1}{\lambda^{\dagger}_{\min}(\boldsymbol{S}^{+})}\frac{1}{\boldsymbol{U}^{\top}\boldsymbol{U}}\right].$$

Note that $\lambda_{\max}(\boldsymbol{S}) = 1\,/\,\lambda^{\dagger}_{\min}(\boldsymbol{S}^{+})$ depends on $\boldsymbol{S}$ only while $\boldsymbol{U}$ depends on $R$ and $\boldsymbol{X}$ only. Then, since $\boldsymbol{S}$ and $\boldsymbol{X}$ are independent, we get

$$\mathrm{E}\left[\frac{\lambda_{\max}(\boldsymbol{S})}{\boldsymbol{U}^{\top}\boldsymbol{U}}\right] = \mathrm{E}\left[\mathrm{E}\left[\frac{\lambda_{\max}(\boldsymbol{S})}{\boldsymbol{U}^{\top}\boldsymbol{U}}\Big|R\right]\right] = \mathrm{E}\left[\mathrm{E}\left[\lambda_{\max}(\boldsymbol{S})\Big|R\right]\mathrm{E}\left[\frac{1}{\boldsymbol{U}^{\top}\boldsymbol{U}}\Big|R\right]\right].$$

Further, since $\boldsymbol{U}\big|R \sim \mathcal{N}_R\left(\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{\theta},\ \boldsymbol{I}_R\right)$, we get

$$\boldsymbol{U}^{\top}\boldsymbol{U}\big|R \sim \chi^2_R(\delta) \quad \text{where} \quad \delta = (\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{\theta})^{\top}\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{\theta}.$$

Let $Z$ be a random variable such that $Z\big|R \sim \text{Poisson}(\delta/2)$. We have

$$\begin{aligned}\mathrm{E}\left[(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\big|R\right] &= \mathrm{E}\left[\mathrm{E}[(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\big|R, Z]\big|R\right] \\ &= \mathrm{E}\left[\mathrm{E}[(\chi^2_{R+2Z})^{-1}\big|R, Z]\big|R\right] \\ &= \mathrm{E}\left[\frac{2^{-1}\Gamma(\frac{R+2Z}{2}-1)}{\Gamma(\frac{R+2Z}{2})}\big|R\right].\end{aligned}$$

We have

$$\frac{2^{-1}\Gamma(\frac{R+2Z}{2}-1)}{\Gamma(\frac{R+2Z}{2})} = \frac{1}{R+2Z-2}.$$

Therefore, since $\mathrm{P}(R > 2) = \mathrm{P}(R \geqslant 3) = 1$ and $\mathrm{P}(Z \geqslant 0) = 1$, we conclude that $R + 2Z - 2 \geqslant 1$ with probability one, and then,

$$\mathrm{E}[(\boldsymbol{U}^\top\boldsymbol{U})^{-1}\big|R] = \mathrm{E}\left[\frac{1}{R+2Z-2}\big|R\right] \leqslant 1 \quad \text{almost surely.}$$

Hence,

$$\begin{aligned}\mathrm{E}\left[\frac{1}{F}\right] &\leqslant \lambda_{\max}(\boldsymbol{\Sigma}^{-1})\,\mathrm{E}\left[\frac{\lambda_{\max}(\boldsymbol{S})}{\boldsymbol{U}^\top\boldsymbol{U}}\right] \leqslant \lambda_{\max}(\boldsymbol{\Sigma}^{-1})\,\mathrm{E}\left[\mathrm{E}\left[\lambda_{\max}(\boldsymbol{S})\big|R\right]\right] \\ &= \lambda_{\max}(\boldsymbol{\Sigma}^{-1})\,\mathrm{E}\left[\lambda_{\max}(\boldsymbol{S})\right]. \qquad (4.10)\end{aligned}$$

Further, we have

$$\lambda_{\max}(\boldsymbol{S}) \leqslant \mathrm{tr}(\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{Y}^\top\boldsymbol{Y}) = \mathrm{tr}((\boldsymbol{Y})^\top\boldsymbol{Y}) = \mathrm{vec}^\top(\boldsymbol{Y})\mathrm{vec}(\boldsymbol{Y})$$

where $\mathrm{vec}(\boldsymbol{Y}) \sim \mathcal{N}_{np}(0, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Therefore,

$$\mathrm{E}[\lambda_{\max}(\boldsymbol{S})] \leqslant \mathrm{E}[\mathrm{tr}\left(\mathrm{vec}^\top(\boldsymbol{Y})\mathrm{vec}(\boldsymbol{Y})\right)] = \mathrm{tr}(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_n) = n\,\mathrm{tr}(\boldsymbol{\Sigma}).$$

Hence,

$$\mathrm{E}[\lambda_{\max}(\boldsymbol{S})] \leqslant n\,\mathrm{tr}(\boldsymbol{\Sigma}). \qquad (4.11)$$

Therefore, together with Eqs. (4.10) and (4.11), we get

$$\mathrm{E}\left[\frac{1}{F}\right] \leqslant \lambda_{\max}(\boldsymbol{\Sigma}^{-1})\,\mathrm{E}\left[\lambda_{\max}(\boldsymbol{S})\right] \leqslant \lambda_{\max}(\boldsymbol{\Sigma}^{-1}) n\,\mathrm{tr}(\boldsymbol{\Sigma}).$$

Hence, $\mathrm{E}\left[1/F\right] \leqslant n\,\mathrm{tr}(\boldsymbol{\Sigma})\lambda_{\max}(\boldsymbol{\Sigma}^{-1})$, this completes the proof. □

From Theorem 4.1, we establish the following result which shows that, under Assumptions 4.1 and 4.2, $\mathrm{E}\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right|\right] < \infty$. It should be noted that Chételat and Wells (2012) establish the same result without an assumption about the rank of $\boldsymbol{S}$ and there is a gap and a major error in their proof. Namely, in the next subsection, we give a counterexample which shows that if the rank of $\boldsymbol{S}$ is not greater than 2, it is possible to have $\mathrm{E}\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right|\right] = \infty$. Thus, the established result constitutes the revised version of Theorem 2 in Chételat and Wells (2012) and we provide a proof which fills in gaps left in the quoted paper.

**Theorem 4.2** *Let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{S}$, and $R$ be as in Assumption 4.1 and for $\boldsymbol{A}$ the symmetric positive definite square root of $\boldsymbol{\Sigma}$, $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{A}^{-1}$. Consider $r$ is any bounded differentiable nonnegative function $r : \mathbb{R} \to [0, C_1]$ with bounded derivative $|r'| \leq C_2$. Define*

$$\boldsymbol{G} = \frac{r^2(\boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{X})}{(\boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{X})^2} \boldsymbol{S}^+ \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{S}, \quad \textit{and} \quad \boldsymbol{H} = \boldsymbol{A}\boldsymbol{G}\boldsymbol{A}^{-1}.$$

*If $P(R > 2) = 1$, then,*

*(i)*

$$\mathrm{E}\left[\frac{1}{F}\left|r^2(F)(n + p - 2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^+) + 3)\right|\right] \leqslant C_1^2\,|n + p - 2\min(n, p) + 3|\, n\,\mathrm{tr}(\boldsymbol{\Sigma})\lambda_{\max}(\boldsymbol{\Sigma}^{-1})$$

*(ii)* $\mathrm{E}\left[\left|div_{vec(\tilde{\boldsymbol{Y}})} vec(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right|\right] < \infty.$

***Proof***

(i) We have

$$\frac{1}{F}\left|r^2(F)(n + p - 2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^+) + 3)\right| \leqslant \frac{1}{F} C_1^2 \left|n + p - 2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^+) + 3\right|.$$

Therefore, since $\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^{+}) = \min(n, p)$, we get

$$\mathrm{E}\left[\frac{1}{F}\left|r^2(F)(n+p-2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^{+})+3)\right|\right] \leqslant C_1^2\,|n+p-2\min(n,p)+3|\,\mathrm{E}\left[\frac{1}{F}\right].$$

Hence, since $\mathrm{P}\,(\mathrm{Rank}(\boldsymbol{S}) > 2) = 1$, from Theorem 4.1, we have

$$\mathrm{E}\left[\frac{1}{F}\right] < n\,\mathrm{tr}(\Sigma)\lambda_{\max}(\Sigma^{-1}).$$

This proves that

$$\mathrm{E}\left[\frac{1}{F}\left|r^2(F)(n+p-2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^{+})+3)\right|\right] \leqslant C_1^2\,|n+p-2\min(n,p)+3|\,n\,\mathrm{tr}(\Sigma)\lambda_{\max}(\Sigma^{-1});$$

this completes the proof of the statement in Part $(i)$.

(ii) One can verify that

$$\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H}) = \frac{r^2(F)}{F}(n+p-2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^{+})+3) - 4r(F)r'(F).$$

Then, by the triangle inequality, we get

$$\left|\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right| \leqslant \frac{1}{F}\left|r^2(F)(n+p-2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^{+})+3)\right| + \left|4r(F)r'(F)\right|,$$

and then, since $0 \leqslant r(F) \leqslant C_1$ and $0 \leqslant |r'(F)| \leqslant C_2$, this gives

$$\left|\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right| \leqslant \frac{1}{F}r^2(F)\left|n+p-2\,\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^{+})+3\right| + 4C_1C_2.$$

Therefore, from Part $(i)$, we get

$$\mathrm{E}\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right|\right] \leqslant C_1^2\,|n+p-2\min(n,p)+3|\,n\,\mathrm{tr}(\boldsymbol{\Sigma})\lambda_{\max}(\boldsymbol{\Sigma}^{-1}) + 4C_1C_2 < \infty.$$

This completes the proof. □

The result given in Part $(ii)$ may look similar to that stated in Theorem 2 of Chételat and Wells (2012). Nevertheless, here we require an additional but important assumption for which $\mathrm{P}(\mathrm{Rank}(\boldsymbol{S}) > 2) = 1$. Indeed, as pointed out in the counterexample given in Sect. 4.3.3, without this assumption, Theorem 4.2 does not necessarily hold. It should be noticed that the established result plays a crucial role in proving that the proposed shrinkage estimators in Eq. (4.5) dominate the classical MLE $\boldsymbol{X}$. For more details, we refer the proof of Theorem 1 of Chételat and Wells (2012). For the completeness, it should also be noticed that, to establish this result, one needs to also suppose that the function $r$ satisfies the following conditions.

**Assumption 4.2** *The real-valued function $r$ defined on $(0, \infty)$ is differentiable such that:*

*(i)* $0 \leqslant r(t) \leqslant \dfrac{[2(\min(n, p) - 2)]}{[|n - p| + 3]}$ *for all* $t > 0$.
*(ii)* $r$ *is nondecreasing.*
*(iii)* $r'$ *is bounded.*

### 4.3.2 Counterexample

In this subsection, we give an example which shows that the bound obtained at page 3153 of Chételat and Wells (2012) is not correct. To this end, we use the same notations as used in Chételat and Wells (2012) and let $T$ be a symmetric matrix and $A$ a positive definite matrix. Specifically, for a given $X$ a column vector, in contrast with the statement in Chételat and Wells (2012),

$$X^\top (T^+TA)^+(T^+TA)X \not\leqslant X^\top (T^+TA)^+(AT^+T)^+XX^\top (AT^+T)(T^+TA)X.$$

**Example 4.1** Let $\boldsymbol{A} = \boldsymbol{I}_4$, $\boldsymbol{X} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}^\top$, and

$$\boldsymbol{T} = \frac{1}{48}\begin{bmatrix}7 & 7 & 1 & 1\\ 7 & 7 & 1 & 1\\ 1 & 1 & 7 & 7\\ 1 & 1 & 7 & 7\end{bmatrix}.$$

Therefore,

$$\boldsymbol{T}^+ = \frac{1}{4}\begin{bmatrix}7 & 7 & -1 & -1\\ 7 & 7 & -1 & -1\\ -1 & -1 & 7 & 7\\ -1 & -1 & 7 & 7\end{bmatrix}$$

and then

$$\boldsymbol{T}^{+}\boldsymbol{T} = \frac{1}{192}\begin{bmatrix} 7 & 7 & -1 & -1 \\ 7 & 7 & -1 & -1 \\ -1 & -1 & 7 & 7 \\ -1 & -1 & 7 & 7 \end{bmatrix}\begin{bmatrix} 7 & 7 & 1 & 1 \\ 7 & 7 & 1 & 1 \\ 1 & 1 & 7 & 7 \\ 1 & 1 & 7 & 7 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \frac{1}{2}I_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Further, since $\boldsymbol{A} = \boldsymbol{I}_4$, we have $\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A} = \boldsymbol{T}^{+}\boldsymbol{T}$, and $\boldsymbol{A}\boldsymbol{T}^{+}\boldsymbol{T} = \boldsymbol{T}^{+}\boldsymbol{T}$. Note that

$$(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A})^{+} = \left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right]^{+} = \frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

and

$$(\boldsymbol{A}\boldsymbol{T}^{+}\boldsymbol{T})^{+} = \frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Therefore,

$$(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A})^{+}(\boldsymbol{A}\boldsymbol{T}^{+}\boldsymbol{T})^{+} = \left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right]\left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right] = \frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$(\boldsymbol{A}\boldsymbol{T}^{+}\boldsymbol{T})(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A}) = \left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right]\left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right] = \frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Thus,

$$\begin{aligned} \boldsymbol{X}^{\top}(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A})^{+}(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A})\boldsymbol{X} &= (1,0,0,0)\left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right](1,0,0,0)^{\top} \\ &= \left(\frac{1}{2},\frac{1}{2},0,0\right)(1,0,0,0)^{\top} = \frac{1}{2}, \end{aligned}$$

and

$$\begin{aligned} &\boldsymbol{X}^{\top}(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A})^{+}(\boldsymbol{A}\boldsymbol{T}^{+}\boldsymbol{T})^{+}\boldsymbol{X}\boldsymbol{X}^{\top}(\boldsymbol{A}\boldsymbol{T}^{+}\boldsymbol{T})(\boldsymbol{T}^{+}\boldsymbol{T}\boldsymbol{A})\boldsymbol{X} \\ &= (1,0,0,0)\left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right](1,0,0,0)^{\top} \\ &\quad \times (1,0,0,0)\left[\frac{1}{2}\boldsymbol{I}_2 \otimes \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right](1,0,0,0)^{\top} \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

Hence,

$$\frac{1}{2} = \boldsymbol{X}^\top (\boldsymbol{T}^+ \boldsymbol{T} \boldsymbol{A})^+ (\boldsymbol{T}^+ \boldsymbol{T} \boldsymbol{A}) \boldsymbol{X} \nleqslant \boldsymbol{X}^\top (\boldsymbol{T}^+ \boldsymbol{T} \boldsymbol{A})^+ (\boldsymbol{A} \boldsymbol{T}^+ \boldsymbol{T})^+ \boldsymbol{X} \boldsymbol{X}^\top \times (\boldsymbol{A} \boldsymbol{T}^+ \boldsymbol{T})(\boldsymbol{T}^+ \boldsymbol{T} \boldsymbol{A}) \boldsymbol{X} = \frac{1}{4},$$

This proves that the inequality used at page 3153 of Chételat and Wells (2012) is incorrect.

### 4.3.3 The Nonexistence Case of the Expectation of $F^{-1}$

In the following example, we highlight the importance of requiring the assumption about the $\text{rank}(\boldsymbol{S}) > 2$ with probability one. In particular, we show that, when $\text{P}(R \leqslant 2) > 0$, it is possible to have $\text{E}[1 / F] = \infty$.

**Example 4.2** Let $\boldsymbol{X} \sim \mathcal{N}_2\left((1, 1)^\top, \boldsymbol{I}_2\right)$ and $\boldsymbol{Y} = [U \vdots V]$ where $U$ and $V$ are independent random variables distributed as $\mathcal{N}(0, 1)$, i.e., $\boldsymbol{Y} \sim \mathcal{N}_{1\times 2}(0, 1 \otimes \boldsymbol{I}_2)$. Let $R$ be the rank of $\boldsymbol{S} = \boldsymbol{Y}^\top \boldsymbol{Y}$. Let $\boldsymbol{S}^+ = \boldsymbol{P} \boldsymbol{D}^+ \boldsymbol{P}^\top$ be the spectral decomposition of $\boldsymbol{S}^+$ where $\boldsymbol{D}^+ = \text{diag}(d_1, 0)$. We have

$$F = \boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{X} = \boldsymbol{X}^\top \boldsymbol{P} \boldsymbol{D}^+ \boldsymbol{P}^\top \boldsymbol{X} = (\boldsymbol{P}^\top \boldsymbol{X})^\top \boldsymbol{D}^+ \boldsymbol{P}^\top \boldsymbol{X}.$$

Therefore,

$$\frac{F}{d_1} = (\boldsymbol{P}^\top \boldsymbol{X})^\top \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{P}^\top \boldsymbol{X}.$$

Note that $d_1$ and $\boldsymbol{P}$ are functions of $(U, V)$ and note that

$$\boldsymbol{P}^\top \boldsymbol{X} \big| U, V \sim \mathcal{N}_2\left(\boldsymbol{P}^\top \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{I}_2\right).$$

Then,

$$\frac{\boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{X}}{d_1} \big| U, V \sim \chi^2_1(\delta_0)$$

where

$$\delta_0 = \begin{bmatrix} 1 & 1 \end{bmatrix} \boldsymbol{P} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{P}^\top \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Therefore,

$$\mathrm{E}\left[\frac{d_1}{F}\Big|U, V\right] = \mathrm{E}\left[\frac{d_1}{\boldsymbol{X}^\top \boldsymbol{S}^+ \boldsymbol{X}}\Big|U, V\right]$$
$$= \mathrm{E}\left[(\chi_1^2(\delta_0))^{-1}\Big|U, V\right] = +\infty, \quad \text{almost surely.}$$

Thus,

$$\mathrm{E}\left[\frac{1}{F}\Big|U, V\right] = \frac{1}{d_1}\mathrm{E}\left[\frac{d_1}{F}\Big|U, V\right] = +\infty,$$

with probability one. Hence, $\mathrm{E}[1/F] = +\infty$.

**Remark 4.1** By taking

$$r(x) = \frac{1}{(1+\exp\{-x\})}, \quad x \geqslant 0$$

along with $\boldsymbol{X}$ and $\boldsymbol{Y}$ in Example 4.2, one can prove that

$$\mathrm{E}\left[\left|\mathrm{div}_{\mathrm{vec}(\tilde{\boldsymbol{Y}})}\mathrm{vec}(\tilde{\boldsymbol{Y}}\boldsymbol{H})\right|\right] = \infty.$$

Details are omitted to save the space of this chapter. Nevertheless, such details are available from the authors upon request.

## 4.4 Conclusion

In this chapter, we derived an important result which plays a central role in studying the risk dominance of Baranchik-type estimator in the Gaussian sample case with high-dimensional data. In particular, under a very realistic additional assumption, we revise Theorem 2 of Chételat and Wells (2012). On top of revising such important result, we also provide an alternative proof which corrects the preexisting proof and fills in the gaps left in the derivation of some recent methods in literature. In our ongoing research, we expect to show that the additional assumption is not only sufficient but necessary. Further, in future research, we hope to be able to prove that the established result also holds for matrix-value random samples from elliptically contoured distribution in the context of high- and ultrahigh-dimensional data.

**Acknowledgments** The authors would like to thank the referees for helpful comments and useful insights. Further, Dr. S. Nkurunziza would like to acknowledge the financial support received from Discovery Grant (RGPIN-2019-04464) of the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics, 41*(2), 642–645.

Chételat, D., & Wells, M. T. (2012). Improved multivariate normal mean estimation with unknown covariance when $p$ is greater than $n$. *The Annals of Statistics, 40*(6), 3137–3160.

Fisher, T. J., & Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis, 55*(5), 1909–1918.

Foroushani, A. A., & Nkurunziza, S. (2023). A note on improved multivariate normal mean estimation with unknown covariance when $p$ is greater than $n$. arXiv:2311.13140.

Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2024). On priors which give Bayes minimax estimators of Baranchik's form. *Japanese Journal of Statistics and Data Science, 7*(1), 329–340.

Haddouche, A. M., Fourdrinier, D., & Mezoued, F. (2021). Scale matrix estimation of an elliptically symmetric distribution in high and low dimensions. *Journal of Multivariate Analysis, 181*, 104680.

Johnson, C. R., & Robinson, H. A. (1981). Eigenvalue inequalities for principal submatrices. *Linear Algebra and Its Applications, 37*, 11–22.

Karamikabir, H., Afshari, M., & Arashi, M. (2018). Shrinkage estimation of non-negative mean vector with unknown covariance under balance loss. *Journal of Inequalities and Applications, 2018*(1), 331.

Lotfi, R., Shahsavani, D., & Arashi, M. (2022). Classification in high dimension using the Ledoit–Wolf shrinkage method. *Mathematics, 10*(21), 4069.

# Chapter 5
# Some Stein-Rule Methods in Tensor Regression Model with High-Dimensional Data




**Mai Ghannam and Sévérien Nkurunziza**

## 5.1 Introduction

As technology in imaging, especially in the healthcare field, has developed, a new area of interest with respect to regression has emerged. We are now interested in regression involving tensor covariates and coefficients. A procedure that first comes to mind is to simply vectorize the tensor and apply classical regression on the vector. However, such a procedure results in a large number of parameters that require an inefficient, if not impossible, number of simulations in order to solve them. As such, it is of interest to find a more efficient method for regressing such tensors.

In this chapter, we propose a new method in estimating the tensor parameter. In addition, we show that the proposed tensor estimator is more efficient than the maximum likelihood estimator (MLE) proposed by Zhou et al. (2013) under some sufficient conditions. Furthermore, we reduce the number of parameters in the coefficient tensor by setting restrictions on some of the parameter components. We then develop a restricted estimator based upon these restrictions. To give some related references on tensor regression and their applications to neuroimaging data, we quote Fu et al. (2014), Guo et al. (2012), Li (2014), Tan et al. (2013), and Zhou et al. (2013) who proposed MLE or least squares estimators in the context of tensor regression. More recently, we quote Lee et al. (2023), Li and Zhang (2017), Liu et al. (2019), Lock (2018), and Li et al. (2018). Most of these papers have developed the estimators using dimension-reducing methods through tensor decomposition to approximate the coefficient tensor of interest.

M. Ghannam · S. Nkurunziza (✉)
Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada
e-mail: ghannamm@uwindsor.ca; severien@uwindsor.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_5

In this chapter, we extend the problem to the case where some uncertain prior information is available. We also use an iteratively obtained least squares estimator as in Zhou et al. (2013). We then propose shrinkage estimators by using a convex combination between the restricted and unrestricted estimators. For other references about the shrinkage method, we cite Judge and Bock (1978), Saleh (2006), and the references therein. To give some other recent references about shrinkage estimators, we quote Nkurunziza and Ahmed (2010) who proposed shrinkage estimators in the multifactor Vasicek model. Further, Chen and Nkurunziza (2015) established a class of shrinkage estimators in multiple regression involving structural changes and established the optimality of the shrinkage estimator. To the best of our knowledge, there is very little literature in which shrinkage estimation is applied in the context of high-dimensional tensor regression. As such, we establish shrinkage estimators in the context of the tensor regression model with high-dimensional data. We prove that shrinkage estimators are more efficient than the unrestricted estimator, and thus, we improve the results in Zhou et al. (2013).

The rest of this chapter is organized as follows. In Sect. 5.2, we present the statistical model and some notations. We also outline in this section the canonical decomposition or parallel factors (CP) decomposition and the block relaxation algorithm. In Sect. 5.3, we present the unrestricted estimator (UE) and restricted estimator (RE) and we present some conditions to ensure the identifiability of the statistical model. We then present the joint asymptotic normality of the UE and RE. In Sect. 5.4, we propose shrinkage estimators and we determine the asymptotic distributional risk (ADR) of the UE, RE, and shrinkage estimators. We also show that, under particular conditions, the shrinkage estimators dominate the UE. In Sect. 5.5, we present some simulations and analysis results of a real data set that illustrate the application of the proposed method. Finally, Sect. 5.6 gives some concluding remarks. We also present some additional results about the identifiability of the tensor parameter as well as proofs of the main results in the Appendix.

## 5.2 Statistical Model and Notations

In this section, we define the tensor regression model and present some notations used in this chapter. Let $\{Y_i, X_i, Z_i\}_{i=1}^n$ be independent and identically distributed (i.i.d.) data of size $n$. Here, for $i = 1, \ldots, n$, $Y_i$ is the response, while $X_i$, $Z_i$, $i = 1, \ldots, n$ represent the covariates with $Z_i \in \mathbb{R}^{p_0}$, $X_i \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$, for $i = 1, \ldots, n$. Let $A \circ B$ denote an outer product between vectors $A$ and $B$, and let $A \odot B$ denote the Khatri-Rao product of matrices $A$ and $B$. Let $\mu = \mathrm{E}[Y_i | X_i, Z_i]$ and let $g(\cdot)$ denote a strictly increasing link function. First, suppose that the distribution of the response $Y_i$ is a member of an exponential family where the link function is such that

$$g(\mu) = \alpha + \gamma^\top Z_i + \langle (B_D \odot \cdots \odot B_1)\, 1_R, \mathrm{vec}(X_i) \rangle ,$$

where $\alpha \in \mathbb{R}$, $\gamma \in \mathbb{R}^{p_0}$, $B \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$. In particular, if $g(\mu) = \mu$, one can predict the response $Y_i$ by using the following tensor regression model:

$$Y_i = \alpha + \gamma^\top Z_i + \langle (B_D \odot \cdots \odot B_1)\, \mathbf{1}_R, \mathrm{vec}(X_i) \rangle + \epsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random scalars, and $\mathbf{1}_n$ denotes an $n$-column vector with all entries equal to one.

### 5.2.1 CP Decomposition

In this section, we describe the CP decomposition and its use in the tensor regression model. First, note that, in estimating the tensor regression coefficients, we face working with an overwhelming number of covariates and parameters. One solution to avoid such a huge number of parameters is to decompose the array of interest. In this chapter, we follow the CP decomposition implemented in Zhou et al. (2013). We define the CP decomposition as follows. Let $A \circ B$ denote the Khatri-Rao product of matrices $A$ and $B$. An array $\mathscr{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ has a rank-$R$ decomposition if

$$\mathscr{B} = \sum_{r=1}^{R} \mathbf{B}_1^{(r)} \circ \mathbf{B}_2^{(r)} \circ \cdots \circ \mathbf{B}_D^{(r)} \tag{5.2}$$

where $\mathbf{B}_d^{(r)} \in \mathbb{R}^{p_d}$, for $d = 1, \ldots, D$, $r = 1, \ldots, R$ are column vectors and $\mathscr{B}$ cannot be written as a sum of less than $R$ outer products. This is known as the CP decomposition of tensor $\mathscr{B}$ and it can be expressed in a simpler manner as $\mathscr{B} = [\mathbf{B_1}, \ldots, \mathbf{B_D}]$ where $\mathbf{B_d} = [B_d^{(1)}, \ldots, B_d^{(R)}] \in \mathbb{R}^{p_d \times R}$. Once we have decomposed the array, we can use the resulting decomposition to replace the array in the link function. We can then proceed to estimate the coefficients by using the block relaxation algorithm (BRA).

### 5.2.2 The Block Relaxation Algorithm

Let $\ell(\alpha, \gamma, B_1, \ldots, B_D)$ denote the log-likelihood function of the observations $Y_1, \ldots, Y_n$ and, as mentioned above, consider a model with the following link function:

$$g(\mu) = \alpha + \gamma^\top Z + \langle \mathscr{B}, \mathrm{vec}X \rangle = \alpha + \gamma^\top Z + \langle \sum_{r=1}^{R} B_1^{(r)} \circ B_2^{(r)} \circ \cdots \circ B_D^{(r)}, X \rangle.$$

**Algorithm 5.1** Block relaxation algorithm for maximizing log-likelihood

**Initialize:** $(\alpha^{(0)}, \gamma^{(0)}) = \underset{\alpha,\gamma}{\text{argmax}}\,\ell(\alpha, \gamma, 0, \ldots, 0)$, $B_d^{(0)} \in p_d \times R$ **a random matrix for** $d = 1, \ldots, D$
**repeat**
**for** $d = 1$ **to** $D$ **do**
$B_d^{(t+1)} = \underset{B_d}{\text{argmax}}\,\ell(\alpha^{(t)}, \gamma^{(t)}, B_1^{(t+1)}, \ldots, B_{d-1}^{(t+1)}, B_d, B_{d+1}^{(t)}, \ldots, B_D^{(t)})$
**end for**
$(\alpha^{(t+1)}, \gamma^{(t+1)}) = \underset{\alpha,\gamma}{\text{argmax}}\,\ell(\alpha, \gamma, B_1^{(t+1)}, \ldots, B_D^{(t+1)})$
**until** $\ell(\theta^{(t+1)}) - \ell(\theta^{(t)}) < \epsilon$

This can be rewritten as

$$g(\mu) = \alpha + \gamma^\top Z + \langle (B_D \odot \cdots \odot B_1)\,\mathbf{1_R}, \text{vec}(X) \rangle, \tag{5.3}$$

where $X \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is a $D$-dimensional array variate, $\mathscr{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is a full coefficient array of the same size, $\gamma \in \mathbb{R}^{p_0}$, $B_d = \left[B_d^{(1)}, \cdots, B_d^{(R)}\right] \in \mathbb{R}^{p_d \times R}$, $B_D \odot \cdots \odot B_1 \in \mathbb{R}^{\prod_d p_d \times R}$ is the Khatri-Rao product, and $\mathbf{1_R}$ is the vector of R ones. Further, $\sum_{r=1}^{R} B_1^{(r)} \circ B_2^{(r)} \circ \cdots \circ B_D^{(r)}$ is the result of the CP decomposition of $\mathscr{B}$.

As mentioned above, we suppose that $\{y_i, X_i, z_i\}_{i=1}^n$ are i.i.d. data of size $n$ with the link function presented in the objective function defined in Eq. (5.3) and their log-likelihood is $\ell(\alpha, \gamma, B_1, \ldots, B_D)$. We assume that the distribution of $Y_i$, $i = 1, \ldots, n$, belongs to an exponential family. In contrast with the classical regression model where the sample size is bigger than the number of parameters, in this context, the number of parameters $1 + p_0 + \prod_{d=1}^{D} p_d$ is bigger than the sample size $n$. In particular, as in Zhou et al. (2013), in the context of neuroimaging data set, $1 + p_0 + \prod_{d=1}^{D} p_d$ is so extremely large that it is practically impossible to estimate simultaneously all of these parameters. The estimation of the coefficient tensor components then follows by iteratively solving for the $\gamma \in \mathbb{R}^{p_0}$ vector and the CP decomposition column vectors, $\mathbf{B}_d^{(r)} \in \mathbb{R}^{p_d}$, $d = 1, \ldots, D$, and $r = 1, \ldots, R$, using the block relaxation algorithm on the log-likelihood function (De Leeuw, 1994; Zhou et al., 2013). This is because the systematic part is not jointly linear for $(\mathbf{B_1}, \ldots, \mathbf{B_D})$ but is linear when looking at the individual columns separately and assuming all other columns constant at a given iteration. The algorithm described in Zhou et al. (2013) is presented in Algorithm 5.1. It can be seen how this method for regressing tensors is much more efficient than using the classical method as, without intercept, only $p_0 + R\sum_{d=1}^{D} p_d$ parameters are estimated using the algorithm proposed in Algorithm 5.1 versus $p_0 + \prod_{d=1}^{D} p_d$ parameters to be solved if the classical route were to be taken.

**Remark 5.1** In the applications, as in neuroimaging, which are the motivation of this chapter, $p_d < n$, $d = 1, \ldots, D$, but $\prod_{d=1}^{D} p_d > n$. The advantage of the block relaxation is to create the condition of estimating a large number of parameters with a small number of observations.

## 5.3 Unrestricted and Restricted Estimators

In this section, we determine the unrestricted least squares estimator (UE) and restricted least squares estimator (RE) of the parameter in the statistical model defined in Eq. (5.4). We then address some important factors that affect the identifiability of the statistical model and some conditions to ensure the existence of the unrestricted and restricted estimators. We also present the joint asymptotic normality of the UE and RE.

In the sequel, we focus on the case where $g(\mu) = \mu$. Hence, the above model can be rewritten as

$$Y = \alpha \mathbf{1}_n + \mathcal{Z}\gamma + \mathcal{X}\theta_0 + \epsilon, \tag{5.4}$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathcal{Z} = (Z_1, \ldots, Z_n)^\top, \quad \mathcal{X} = \begin{pmatrix} \text{vec}(X_1)^\top \\ \vdots \\ \text{vec}(X_n)^\top \end{pmatrix},$$

and

$$\theta_0 = (B_D \odot \cdots \odot B_1)\, 1_R, \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top.$$

From the above, we have $\mathcal{X} \in \mathbb{R}^{n \times \Pi_d p_d}$. Further, for the sake of simplicity, we focus on the model in Eq. (5.1) with $\alpha = 0$ and $\gamma = 0$, and thus, the statistical model under consideration is

$$Y_i = \langle (B_D \odot \cdots \odot B_1)\, 1_R, \text{Vec}(X_i) \rangle + \epsilon_i, \quad i = 1, \ldots, n,$$

where $Y_i$ and $\epsilon_i$ are random scalars, $(B_D \odot \cdots \odot B_1) \in \mathbb{R}^{\Pi p_d \times R}$, $1_R \in \mathbb{R}^R$ is a vector of ones, and $X_i \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$. Then, as it can be seen from Eq. (5.4), if $\alpha = 0$, $\gamma = 0$, the model in Eq. (5.1) can be rewritten as

$$Y = \mathcal{X}\theta_0 + \epsilon. \tag{5.5}$$

We also suppose that there exists uncertain prior information about the parameter of interest. In particular, we suppose that we are given $k_d \times p_d$-matrices $D$, $\zeta_d \in \mathbb{R}^{k_d \times p_d}$, with rank $k_d$, $k_d << p_d$, such that

$$\mathcal{H}_0 : \zeta_d B_d = r_d, \; d = 1, \ldots, D \tag{5.6}$$

for known $k_d$-column vectors $r_d$.

Using the BRA, the UE for each $B_d$, denoted by $\hat{B}_d$, is determined. Thus, the UE of $\theta_0$ is $\hat{\theta}_n = \hat{B}_D \odot \cdots \odot \hat{B}_1$. The matrices $\zeta_d$ represent some known constraints on

each dimension $d$ component of the parameter. This may be a linear combination of the parameters in the matrix $B_d$ or it may be simply that some components of $B_d$ are trivial. For example, in our numerical data analysis, we are interested in recovering a signal matrix of 0s and 1s. As such, we suspect that some of those parameters are 0 and we can add these constraints to our model through a restriction matrix of diagonal 1s corresponding to the parameter components and by setting $r = 0$. The corresponding RE for each $B_d$ is

$$\tilde{B}_d = \hat{B}_d + (\mathcal{X}_d^\top \mathcal{X}_d)^{-1} \zeta_d^\top [\zeta_d (\mathcal{X}_d^\top \mathcal{X}_d)^{-1} \zeta_d^\top]^{-1} (r_d - \zeta_d \hat{B}_d), \tag{5.7}$$

where $\mathcal{X}_d = \mathcal{X}_{(d)}(B_D \odot \cdots \odot B_{d+1} \odot B_{d-1} \odot \cdots \odot B_1)$. Here, $\mathcal{X}_{(d)}$ denotes the mode-$d$ matricization of $\mathcal{X}$. Since $B_d$ are unknown for $d = 1, \ldots, D$, they are replaced with their UEs. Thus, the RE for the target tensor regression model is

$$\tilde{\theta}_n = \tilde{B}_D \odot \cdots \odot \tilde{B}_1. \tag{5.8}$$

### 5.3.1 Asymptotic Properties of UE and RE

In this subsection, we derive some asymptotic properties of the UE and RE. To this end, as in Zhou et al. (2013), we also adopt the classical asymptotic setup with a fixed number of parameters and a large sample size $n$. At this point, current investigations do not allow us to derive asymptotic results in the context of high-dimensional data. Nevertheless, such an interesting problem is part of our ongoing research. Let $B_0 \in \mathbb{B}$ be a parameter point and assume there exists an open neighborhood of $\boldsymbol{B}_0$ in which the information matrix has a constant rank. Also, let $J_d \in \mathbb{R}^{\prod_{d=1}^{D} p_d \times p_d R}$ be the Jacobian of $B$ where $J_d = \Pi_d[(B_D \odot \cdots \odot B_{d+1} \odot B_{d-1} \odot \cdots \odot B_1) \otimes I_{p_d}]$, for permutation matrix $\Pi_d$ that reorders $\mathrm{Vec}(B_{(d)})$ to get $\mathrm{Vec}(B)$. Let $I_n(\boldsymbol{B}_0)$ be the Fisher information matrix of $\boldsymbol{B}_0$. As given in Proposition 5.6 in the Appendix, we have

$$I_n(\boldsymbol{B}_0) = [J_1 \cdots J_D]^\top \left[ \sum_{i=1}^{n} \frac{\mu^\top(\eta_i)^2}{\sigma_i^2} (\mathrm{Vec}(x_i))(\mathrm{Vec}(x_i))^\top \right] [J_1 \cdots J_D].$$

Below, we present first an important theorem regarding the consistency of the MLE that is given in Zhou et al. (2013). This theorem discusses the cases where the estimator $\hat{B}_n$ converges to the parameter of interest $\boldsymbol{B}_0$. To simplify some mathematical expressions, let

$$\boldsymbol{I}(\boldsymbol{B}_0) = \lim_{n \to \infty} \frac{1}{n} \boldsymbol{I}_n(\boldsymbol{B}_0), \tag{5.9}$$

where in the case that the covariates $X_i$, $i = 1, \ldots, n$, are random, this limit stands for the limit in probability, i.e., $(1/n)\boldsymbol{I}_n(\boldsymbol{B}_0)$ converges in probability to $\boldsymbol{I}(\boldsymbol{B}_0)$. Moreover, we make the following assumption.

**Assumption 5.1** *For a given $\boldsymbol{B}_0$, $\boldsymbol{I}(\boldsymbol{B}_0)$ is a nonrandom and positive definite matrix.*

**Theorem 5.1 (Consistency)** *Assume $B_0 = [B_{01}, \ldots, B_{0D}] \in \mathbb{B}$ is (globally) identifiable up to permutation and array covariates $X_i$ are i.i.d. from a bounded distribution, where $\mathbb{B}$ is the restricted parameter space defined to fix the scaling and permutation in order to eliminate non-identifiability of the parameter $B$, and $\hat{B}_n$ converges in probability to $\boldsymbol{B}_0$ in the following models:*

*(1) Normal tensor regression with a compact parameter space $B_0 \subset \mathbb{B}$*
*(2) Binary tensor regression*
*(3) Poisson tensor regression with a compact parameter space $B \subset \mathbb{B}$*

The proof of this proposition is given in Zhou et al. (2013). From this theorem, one establishes the following proposition that gives the asymptotic normality of $\hat{B}_n$.

**Proposition 5.1** *Let $B_0 = [B_{01}, \ldots, B_{0D}] \in \mathbb{B}$ be an interior point and suppose that the conditions of Theorem 5.1 hold along with Assumption 5.1. Then,*

$$\sqrt{n}[\mathrm{Vec}(\hat{B}_{n1}, \ldots, \hat{B}_{nD}) - \mathrm{Vec}(B_{01}, \ldots, B_{0D})] \xrightarrow[n\to\infty]{D} \alpha_1 \sim \mathcal{N}_p(0, \boldsymbol{I}^{-1}(\boldsymbol{B}_0)).$$

The proof of Proposition 5.1 is given in Zhou et al. (2013). From Proposition 5.1, we derive below the joint normality between the UE and the RE. To this end, we suppose that the following assumption holds.

**Assumption 5.2** *Suppose that*

$$\frac{\mathcal{X}^\top \mathcal{X}}{n} \xrightarrow[n\to\infty]{P} \boldsymbol{Q},$$

*where $\boldsymbol{Q}$ is a nonrandom and positive definite matrix.*

In the sequel, to avoid the asymptotic degeneracy, we also consider the following sequence of local alternative hypotheses:

$$\mathcal{H}_{1,n} : \boldsymbol{\eta}\boldsymbol{\theta}_0 = \boldsymbol{r} + \boldsymbol{\Delta}/\sqrt{n}, \quad n = 1, 2, \ldots \tag{5.10}$$

where $\boldsymbol{\Delta}$ is a fixed matrix such that $\|\Delta\| < \infty$. We consider the above sequence of hypotheses to compare the efficiency of the proposed estimators when the restriction deviates away by $\boldsymbol{\Delta}/\sqrt{n}$ from the neighborhood of the true fixed restriction, $\boldsymbol{r}$. To introduce some other notations, let $\tilde{\theta}_n$ and $\hat{\theta}_n$ denote the RE and UE of $\theta_0 = (B_{0D} \odot \cdots \odot B_{01})\,\mathbf{1}_R$, respectively. Further, let

$$\boldsymbol{M} = \boldsymbol{I}_p - \boldsymbol{Q}^{-1}\zeta^\top[\zeta\,\boldsymbol{Q}^{-1}\zeta^\top]^{-1}\zeta\,,$$

$$\boldsymbol{J} = \boldsymbol{Q}^{-1}\zeta^{\top}[\zeta\,\boldsymbol{Q}^{-1}\zeta^{\top}]^{-1}\zeta\,,$$

$$\Gamma_{11} = \boldsymbol{J}\boldsymbol{I}^{-1}(\boldsymbol{B}_0)\boldsymbol{J}^{\top}\,,$$

$$\Gamma_{12} = \boldsymbol{J}\boldsymbol{I}^{-1}(\boldsymbol{B}_0)\boldsymbol{M}^{\top}\,,$$

$$\Gamma_{21} = \boldsymbol{M}\boldsymbol{I}^{-1}(\boldsymbol{B}_0)\boldsymbol{J}^{\top}\,,$$

$$\Gamma_{22} = \boldsymbol{M}\boldsymbol{I}^{-1}(\boldsymbol{B}_0)\boldsymbol{M}^{\top}\,.$$

Also,

$$\Lambda_{11} = \boldsymbol{I}^{-1}(\boldsymbol{B}_0)\,,$$

$$\Lambda_{12} = \boldsymbol{I}^{-1}(\boldsymbol{B}_0)\boldsymbol{M}^{\top}\,,$$

$$\Lambda_{21} = \boldsymbol{M}\boldsymbol{I}^{-1}(\boldsymbol{B}_0)\,,$$

$$\Lambda_{22} = \boldsymbol{M}\boldsymbol{I}^{-1}(\boldsymbol{B}_0)\boldsymbol{M}^{\top}\,.$$

Further, let $\Lambda_{31} = \Lambda_{13}^{\top} = \Lambda_{11} - \Lambda_{21}$, $\Lambda_{32} = \Lambda_{23}^{\top} = \boldsymbol{J}\,\Lambda_{12}$, and $\Lambda_{33} = \Lambda_{11} - \Lambda_{21} - \Lambda_{21} + \Lambda_{22}$. Let $\epsilon_1 = \boldsymbol{J}\Delta$, $\epsilon_2 = -\epsilon_1$.

**Theorem 5.2** *Suppose that Assumptions 5.1–5.2 hold along with Eq. (5.10). Then,*

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ \sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \end{pmatrix} \xrightarrow[n\to\infty]{D} \begin{pmatrix} \alpha_1 \\ \alpha_3 \\ \alpha_5 \end{pmatrix} \sim \mathcal{N}_{3p}\left( \begin{pmatrix} 0 \\ \epsilon_2 \\ \epsilon_1 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} \end{pmatrix} \right).$$

The proof of this theorem is given in Appendix "Proof of Theorem 5.2".

**Remark 5.2** As mentioned above, the asymptotic theory in this section is established under the assumption that $\prod_{i=1}^{D} p_i$ is fixed while $n$ is large. Consequently, the asymptotic results are not established under high-dimensional settings. However, in this chapter, we address the high-dimensionality problem numerically by using the block relaxation algorithm proposed in Sect. 5.2.2.

## 5.4 Shrinkage Estimators and Asymptotic Results

In this section, we present shrinkage estimators (SEs) also known as Stein-type estimators. Such estimators combine the UE and the RE. The pioneer reference about the method is James and Stein (1961) who were the first to propose a family of

estimators that dominate the MLE of the mean of a $p$-dimensional normal random vector when $p \geq 3$. Geometrically, such estimators shrink all of the components of UE toward the origin, which can be very useful in reducing the number of parameters to be estimated. In the context of this chapter, the proposed estimators shrink toward the UE as the difference between UE and RE becomes larger, and they shrink toward the RE as difference between UE and RE becomes smaller.

The main contribution of this chapter consists in the fact that the estimation problem is studied in context of the tensor regression model defined in Eq. (5.5) with high-dimensional data setting. We also compare in this section the relative performance of the proposed estimators.

### 5.4.1 Shrinkage Estimators

For simplicity, let $I_n(\hat{B}_{0d})$ and $I(B_{0d})$ denote the $(d, d)$-th elements of $I_n(\hat{B}_0)$ and $I(B_0)$, respectively. The proposed shrinkage estimators are constructed as in Chen and Nkurunziza (2015) and Chen and Nkurunziza (2016). To this end, let

$$\hat{A}_d = n^{-1}\zeta_d^\top \left[\zeta_d I_n^{-1}(\hat{B}_{0d})\zeta_d^\top\right]^{-1} \zeta_d$$

and

$$\psi_d = n(\hat{B}_d - \tilde{B}_d)^\top \hat{A}_d(\hat{B}_d - \tilde{B}_d).$$

We consider the following set of shrinkage estimator components:

$$\hat{B}_d^s = \tilde{B}_d + \left(1 - \frac{k_d - 2}{\psi_d}\right)(\hat{B}_d - \tilde{B}_d), \tag{5.11}$$

where $k_d \geq 2$ is the rank of the constrained matrix $\zeta_d$ in Eq. (5.6). To avoid possible over-shrinking, we propose the positive part shrinkage estimator as

$$\hat{B}_d^{s+} = \tilde{B}_d + \max\left\{0, \left(1 - \frac{k_d - 2}{\psi_d}\right)\right\}(\hat{B}_d - \tilde{B}_d). \tag{5.12}$$

As mentioned above, the estimators $\hat{B}_d^s$ and $\hat{B}_d^{s+}$ shrink toward the $\hat{B}_d$ as the distance between $\hat{B}_d$ and $\tilde{B}_d$ becomes larger, and they shrink toward the $\tilde{B}_d$ as difference between $\hat{B}_d$ and $\tilde{B}_d$ becomes smaller. As a result, the shrinkage estimators for $\theta_0$ are $\hat{\theta}^s = \hat{B}_D^s \odot \cdots \odot \hat{B}_1^s$ and $\hat{\theta}^{s+} = \hat{B}_D^{s+} \odot \cdots \odot \hat{B}_1^{s+}$.

### 5.4.2 Asymptotic Distributional Risk (ADR)

In this subsection, we evaluate the relative performance of the UE, the RE, and SEs. To this end, as in Sect. 5.3.1, we adopt in this subsection the classical asymptotic setup with a fixed number of parameters and a large sample size $n$. As in Judge and Bock (1978), Nkurunziza and Ahmed (2010), Nkurunziza (2012), and Saleh (2006) among others, we use the weighted squared-error loss function to compare efficiencies of the proposed estimators. However, as discussed, for example, in Nkurunziza and Ahmed (2010), the finite sample distribution theory of the shrinkage estimators is difficult to obtain, even for normal data. Hence, it is difficult or rather impossible to obtain the finite sample risk of shrinkage estimators. To overcome this problem, we use the asymptotic distribution methods in order to evaluate the performance of the proposed estimators. In particular, we use the concept of asymptotic distributional risk (ADR) as defined in Ahmed et al. (2012), Hossain and Ahmed (2012), Saleh (2006), and Sen and Saleh (1985) among others. Namely, as given in the quoted papers, the ADR with an estimator $\hat{\theta}^*$ is defined as $\mathrm{ADR}(\hat{\theta}^*, \theta_0, W) = \mathrm{E}[\rho^\top W \rho]$ where $W$ is a weighting nonnegative definite matrix, and $\rho$ is a random vector such that $\sqrt{n}(\hat{\theta}^* - \theta_0) \xrightarrow[n\to\infty]{D} \rho$. To give some other references about the ADR, we also quote Hossain et al. (2016), Chen and Nkurunziza (2015), Chen and Nkurunziza (2016), and references therein. Below, we give the ADR of the proposed estimators. First, note that the UE, RE, and SEs are members of the following class of estimators:

$$\hat{\lambda}(h) = \tilde{\theta}_n + h(n(\hat{\theta}_n - \tilde{\theta}_n)^\top \hat{A}(\hat{\theta}_n - \tilde{\theta}_n))(\hat{\theta}_n - \tilde{\theta}_n), \tag{5.13}$$

where $h$ is a Borel measurable and real-valued integrable function. Indeed, by letting $k = \prod_{i=1}^{D} k_i$ and taking $h(x) \equiv 1$ and $h(x) \equiv 0$, we get the UE and RE, respectively. In the cases that $h(x) = 1 - (k-2)/x$, $x > 0$ and $h(x) = \max(0, 1 - (k-2)/x)$, $x > 0$, we get the SEs. Thereafter, to simplify the derivation of the ADR, we suppose that $h$ is a continuous real-valued and squared-integrable function. Note that the above class of estimators is similar as that in Chen and Nkurunziza (2015) and Chen and Nkurunziza (2016). The following proposition gives the ADR of the class of estimators given in Eq. (5.13).

**Proposition 5.2** *Suppose assumptions 5.1–5.2 hold. Also, assume that the sequence of local alternatives in Eq.* (5.10) *holds and suppose that* $W = A^{1/2} W^* A^{1/2}$ *with* $W^*$ *being a nonnegative definite matrix. Then, the* ADR *of the family of estimators in Eq.* (5.10) *is*

$$\begin{aligned}
\mathrm{ADR}(\hat{\lambda}(h), \theta_0, W) &= \mathrm{ADR}(\tilde{\theta}_n) - 2\,\mathrm{E}[h(\chi^2_{k+2}(\omega))]\epsilon_1^\top W \epsilon_1 \\
&\quad + \mathrm{E}[h^2(\chi^2_{k+4}(\omega))]\epsilon_1^\top W \epsilon_1 - 2\,\mathrm{E}[h(\chi^2_{k+2}(\omega))]\epsilon_1^\top A \Lambda_{31} W \epsilon_1 \\
&\quad + 2\,\mathrm{E}[h(\chi^2_{k+2}(\omega))]\,\mathrm{tr}(\Lambda_{31} W \Lambda_{33} A) - 2\,\mathrm{E}[h(\chi^2_{k+4}(\omega))]\epsilon_1^\top A \Lambda_{31} W \epsilon_1 \\
&\quad + \mathrm{E}[h^2(\chi^2_{k+2}(\omega))]\,\mathrm{tr}(W \Lambda_{33}).
\end{aligned}$$

From Proposition 5.2, we obtain the following proposition that gives the ADR of the UE and RE.

**Proposition 5.3** *Suppose that the conditions of Proposition 5.2 hold and assume that the sequence of local alternatives in Eq. (5.10) holds. Then,*

$$\mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = \mathrm{tr}(W I^{-1}(B_0))$$

*and*

$$\begin{aligned}\mathrm{ADR}(\tilde{\theta}_n, \theta_0, W) = \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) - 2\,\mathrm{tr}(W J I^{-1}(B_0)) \\ + \mathrm{tr}(W J I^{-1}(B_0) J^\top) + \epsilon_1^\top W \epsilon_1.\end{aligned}$$

Let $\mathbb{I}(A)$ be the indicator function of the event $A$. From Proposition 5.2, we also obtain the following proposition that gives the ADR of the proposed shrinkage estimators.

**Proposition 5.4** *Suppose that the conditions of Proposition 5.2 hold. Then,*

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) = \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) - 2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]\,\mathrm{tr}(W(\Lambda_{33} + \Lambda_{31}) \\ +(k^2-4)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top W \epsilon_1 \\ +(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W\Lambda_{33}) \\ +4(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A\Lambda_{31} W\epsilon_1,\end{aligned}$$

*and*

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) = \mathrm{ADR}(\hat{\theta}^s, \theta_0, W) \\ +2\,\mathrm{E}[\mathbb{I}(\chi_{k+2}^2(\omega) \le k-2) - (k-2)\chi_{k+2}^{-2}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega) \le k-2)]\epsilon_1^\top W\epsilon_1 \\ +2\,\mathrm{E}[\mathbb{I}(\chi_{k+2}^2(\omega) \le k-2) - (k-2)\chi_{k+2}^{-2}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega) \le k-2)]\epsilon_1^\top A\Lambda_{31} W\epsilon_1 \\ -2\,\mathrm{E}[\mathbb{I}(\chi_{k+2}^2(\omega) \le k-2) - (k-2)\chi_{k+2}^{-2}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega) \le k-2)]\,\mathrm{tr}(W\Lambda_{31}) \\ -2\,\mathrm{E}[\mathbb{I}(\chi_{k+4}^2(\omega) \le k-2) - (k-2)\chi_{k+4}^{-2}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega) \le (k-2))]\epsilon_1^\top A\Lambda_{31} W\epsilon_1 \\ -\mathrm{E}[\mathbb{I}(\chi_{k+2}^2(\omega) \le (k-2)) - 2(k-2)\chi_{k+2}^{-2}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega) \le (k-2)) \\ +(k-2)^2\chi_{k+2}^{-4}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega) \le (k-2))]\,\mathrm{tr}(W\Lambda_{33}) \\ -\mathrm{E}[\mathbb{I}(\chi_{k+4}^2(\omega) \le (k-2)) - 2(k-2)\chi_{k+4}^{-2}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega) \le (k-2)) \\ +(k-2)^2\chi_{k+4}^{-4}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega) \le (k-2))]\epsilon_1^\top W\epsilon_1.\end{aligned}$$

We show that the shrinkage estimators dominate the UE, i.e., we show that the ADR of SEs is less than the ADR of the UE.

As in Chen and Nkurunziza (2015, 2016), let $\Lambda_{\min}(\Pi)$ denote the largest eigenvalue of $\Pi$. Also, let

$$\Pi^* = \left(\Pi_1^* + \Pi_1^{*^\top}\right)/2$$

where

$$\Pi_1^* = A^{1/2}(\Lambda_{33} + 4\Lambda_{31}/(k+2))W\Lambda_{33}A^{1/2}.$$

Then, we derive the following proposition.

**Proposition 5.5** *Suppose that Assumptions 5.1–5.2 hold along with the sequence of alternatives in Eq. (5.10). Suppose $W$ is chosen such that:*

1. $\text{tr}(W\Lambda_{31}) \leq 0$.
2. $-\Lambda_{min}(W\Lambda_{33}) \leq \Lambda_{min}(W(\Lambda_{31} + \Lambda_{13})/2)$.
3. $\text{tr}(W(\Lambda_{33} + \Lambda_{31}) \geq \max(-\text{tr}(W\Lambda_{31}), (k+2)\Lambda_{max}(\Pi^*)/2)$.

*Then, for any non-centrality parameter $\omega \geq 0$,*

$$\text{ADR}(\hat{\theta}^{s+}, \theta_0, W) \leq \text{ADR}(\hat{\theta}^{s}, \theta_0, W) \leq \text{ADR}(\hat{\theta}_n, \theta_0, W). \tag{5.14}$$

*Further, the inequality is strict for some $\omega \geqslant 0$.*

The proof is very similar to that given in proof of Corollary 2.4.12 in Chen and Nkurunziza (2015, 2016). Further, for completeness, we also outline the proof in Appendix "Proof of Proposition 5.5".

## 5.5 Application and Numerical Results

In this section, we present some simulation results and numerical values from a neuroimaging data set. Thus, this section is divided into two subsections. Namely, Sect. 5.5.1 presents some simulation results, and in Sect. 5.5.2, we present an analysis of neuroimaging data.

### *5.5.1 Simulation Study*

In this subsection, we carry out a simulation study to analyze the application of the proposed method in small and moderate sample sizes. To this end, we use Matlab software to generate the data. The parameter matrix, $\beta$, is a $p_1 \times \cdots p_D$ matrix of 1s and 0s, for $D = 2$ and $D = 3$. To save the space, we only report here the case

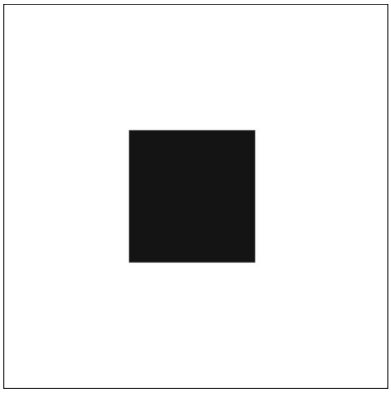


**Fig. 5.1** Image signal used for $\beta$

where $D = 2$. In particular, for the case where $p_1 = p_2 = p$, the parameter $\beta$ is a $p \times p$ matrix of 1s and 0s corresponding to a square image signal similar to that in Fig. 5.1. The value 1 is equivalent to a black signal (response) and 0 is equivalent to a white signal (nonresponse).

The covariate $X$ is a $p \times p \times n$ randomly generated tensor corresponding to $n$ responses. For each $\omega = 0, 1, 2, 3$, we generate 1000 samples of size $n$. At each replicate, the response $y$ is an $n$-column vector resulting from the sum of the tensor product between $X$ and $\beta$ and a randomly generated error value. The sample sizes are taken as $n = 100$, $n = 300$ or $n = 500$. We then make a rank-1 (taking $R = 1$) estimation of $\beta$ using the block relaxation algorithm as proposed in Zhou et al. (2013). This is done using the Matlab Tensor Regression Toolbox developed by Zhou et al. (2013). This gives two unrestricted estimators, $\hat{\beta}_1$ and $\hat{\beta}_2$, of the $p \times 1$ component vectors, $\beta_1$ and $\beta_3$, respectively. Recall that $\beta = \beta_2 \circ \beta_1$. To set up the $k \times p$ restriction matrix, $\zeta$, we choose $\zeta = (\zeta_{ij})$ where $\zeta_{ii} = 1$, for $i = 1, \ldots, k$, and $\zeta_{ij} = 0$, if $i \neq j$. In addition, we take $r$ to be a $k \times 1$ zero vector. This situation arises when one suspects that some components of the two parameter vectors are zero based on some prior knowledge. The restricted estimators, $\tilde{\beta}_1$ and $\tilde{\beta}_2$, are then found using Eq. (5.7). We then calculate the James-Stein shrinkage estimators, $\hat{\beta}_1^s$ and $\hat{\beta}_1^{s+}$, and the positive-rule shrinkage estimators, $\hat{\beta}_1^{s+}$ and $\hat{\beta}_2^{s+}$, by taking $f = k - 2$. We then obtain the matrix estimators for $\beta$, where $\hat{\beta}$, $\tilde{\beta}$, $\hat{\beta}^s$, and $\hat{\beta}^{s+}$ are the UE, RE, shrinkage estimator, and positive-rule shrinkage estimators, respectively. These matrix estimators are obtained using the following relations:

$$\hat{\beta} = \hat{\beta}_2 \circ \hat{\beta}_1,$$

$$\tilde{\beta} = \tilde{\beta}_2 \circ \tilde{\beta}_1,$$

$$\hat{\beta}^s = \hat{\beta}_2^s \circ \hat{\beta}_1^s,$$

$$\hat{\beta}^{s+} = \hat{\beta}_2^{s+} \circ \hat{\beta}_1^{s+}.$$

Further, we calculate the empirical ADR of the above estimators using the formula

$$\mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}^*) = \frac{1}{1000}\sum_{i=1}^{1000} n \ \mathrm{tr}\left((\hat{\beta}_i^* - \beta)^\top(\hat{\beta}_i^* - \beta)\right),$$

where $\hat{\beta}^* = \hat{\beta}, \tilde{\beta}, \hat{\beta}^s, \hat{\beta}^{s+}$ and $\hat{\beta}_i^* = \hat{\beta}_i, \tilde{\beta}_i, \hat{\beta}_i^s, \hat{\beta}_i^{s+}$ at the $i$th replicate. The mean square errors of the estimators relative to the unrestricted estimator (RMSE) are obtained from the average ADRs using

$$\mathrm{RMSE}(\hat{\beta}) = \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}) \,/\, \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}),$$

$$\mathrm{RMSE}(\tilde{\beta}) = \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}) \,/\, \mathrm{ADR}_{\mathrm{emp}}(\tilde{\beta}),$$

$$\mathrm{RMSE}(\hat{\beta}^s) = \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}) \,/\, \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}^s),$$

$$\mathrm{RMSE}(\hat{\beta}^{s+}) = \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}) \,/\, \mathrm{ADR}_{\mathrm{emp}}(\hat{\beta}^{s+}).$$

The resulting RMSEs are then plotted versus $\omega$. Figures 5.2, 5.3, 5.4, and 5.5 summarize the results for $\beta$ with $p = 15, 25, 48$, and 96, respectively. For the convenience of the reader, let us recall that $p = p_1 = p_2$, and accordingly, the number of parameters is $p \times p$. Because of that, it should be noted that, while the sample size is 500, for the case where $p = 25$, the number of parameters is 625 that corresponds to the case of high-dimensional data, and thus, classical methods are not applicable. Further, for the cases where $p = 48$ and $p = 96$, the number of parameters becomes 2304 and 9216, respectively. Thus, these two cases correspond to ultrahigh-dimensional data sets where classical methods are

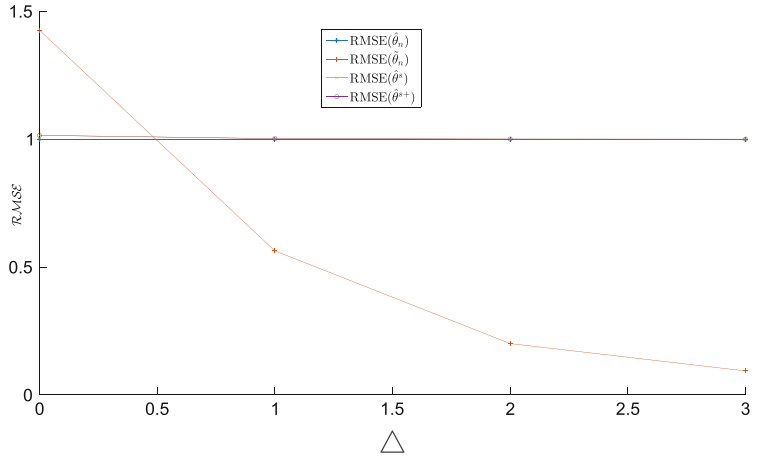



**Fig. 5.2** RMSE vs. $\omega$ for $p = 16, k = 5, n = 100$

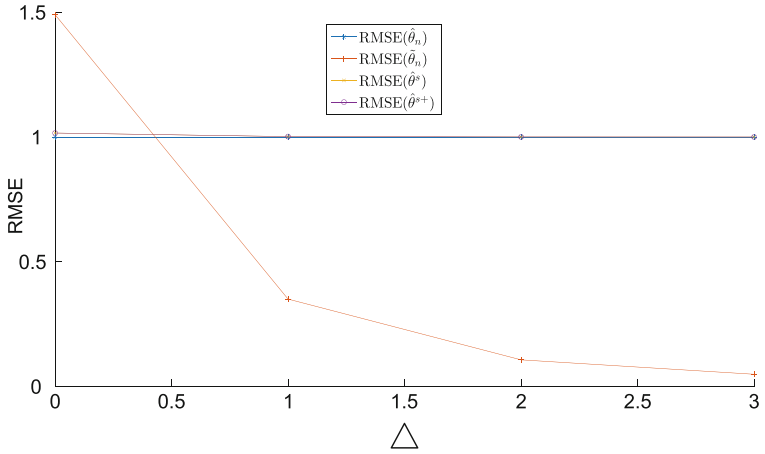



**Fig. 5.3** RMSE vs $\omega$ for $p = 25, k = 8, n = 300$

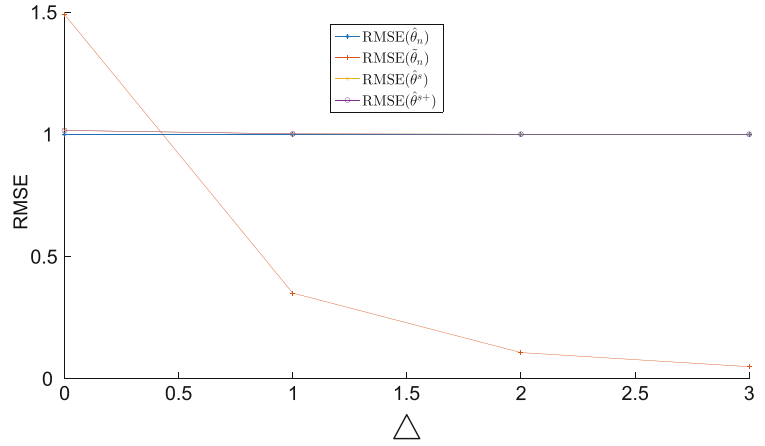



**Fig. 5.4** RMSE vs $\omega$ for $p = 48, k = 15, n = 500$

infeasible. As can be seen from the definition of the RMSE, RMSE $\geqslant 1$ if and only if $\text{ADR}_{\text{emp}}(\hat{\beta}) \geqslant \text{ADR}_{\text{emp}}(\hat{\beta}^*)$, i.e., $\hat{\beta}^*$ dominates $\hat{\beta}$ if and only if the RMSE of $\hat{\beta}^*$ is greater than 1. As can be seen in Figs. 5.2, 5.3, 5.4, and 5.5, the RMSE of the shrinkage estimators remains above the RMSE value of 1. Consequently, the ADR of the shrinkage estimators is always less than that of the unrestricted estimator, which corroborates the theoretical results in Eq. (5.14).

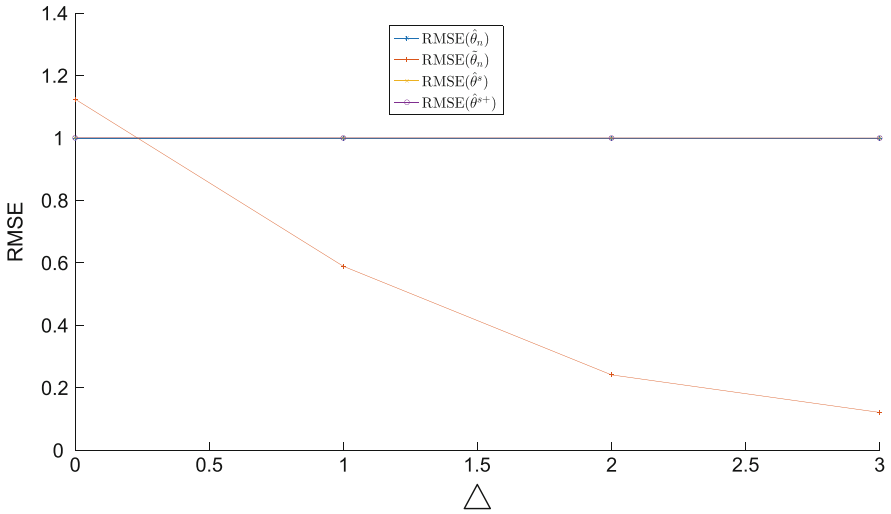



**Fig. 5.5** RMSE vs $\omega$ for $p = 96, k = 30, n = 500$

### *5.5.2 Real Data Set*

In this subsection, we perform an analysis on a real neuroimaging data set in order to illustrate the application of the proposed method. The attention deficit hyperactivity disorder (ADHD) data was acquired from the ADHD-200 Sample Initiative which consists of 776 subjects with 491 normal controls and 284 combined ADHD subjects and a test sample composed of 197 subjects. For the sake of simplicity, we obtained the preprocessed anatomical (MRI) data only for the test sample provided by the Athena pipeline. This test sample consisted of MRI data for 197 subjects distributed among 7 different locations. The MRI data for each subject is a $197 \times 233 \times 189$ tensor. To initialize our analysis, we follow a similar approach as in Zhou et al. (2013). In particular, we consider a GLM where $\eta = \langle B, X \rangle$, $B$ is a three-dimensional ball signal, and $X$ is the MRI anatomical tensor data obtained for each of the 197 subjects. The structure of $B$ admits a rank-1 decomposition with $B_1 \in \mathbb{R}^{197}$, $B_2 \in \mathbb{R}^{233}$, and $B_3 \in \mathbb{R}^{189}$. For each $B_i$, $i = 1, 2, 3$, the $(90 + j)$-th element is equal to $\sin(j\pi/14)$ for $j = 0, \ldots, 14$. All other elements are set to equal zero. The signal tensor is then obtained using $B = B_3 \circ B_2 \circ B_1$. The first difficulty of this data set is due to its high dimension.

Due to the vast size of the data and in the similar way as in Zhou et al. (2013), we have to downsize the data to about one-tenth of the original size using interpolation. The resulting dimensions of $B$ and each $X_i$, $i = 1, \ldots, 197$, are $20 \times 23 \times 19$. We then obtained the estimators for $B$ and the RMSE by performing 1000 replications. We bootstrapped residuals to produce the replicates as follows. We first obtain an initial UE from the data. We then find the fitted values based on the UE and we calculate the residuals. We resample the residuals with replacement

1000 times using the built-in bootstrap function in Matlab. This gives a sample of 1000 residuals. For each residual vector, we set the resampled response vector, $y^*$, to be the sum of the initial fitted values and the residuals. It is important to note that we performed the analysis on a computer with limited computational capabilities to analyze the vast size of the original MRI data. To overcome this issue, we were required to downsize the data to illustrate how the proposed methods may be used in applications. However, downsizing may result in under-fitting which can affect the robustness of the proposed model. In the future, we look to conduct the analysis on the original MRI data by using high-performance computing (HPC) and/or a graphics processing unit (GPU).

Using the block relaxation algorithm, we obtain $\hat{B}_1$, $\hat{B}_2$, and $\hat{B}_3$. The UE of $B$ is found using the relation $\hat{B} = \hat{B}_3 \circ \hat{B}_2 \circ \hat{B}_1$. To determine the restricted estimator, we set the restriction matrix of the first component vector $B_1$ to be $\zeta_1 = (\zeta_{1_{ij}})$, where $\zeta_{1_{ii}} = 1$ for $i = 1, \ldots, 7$, and $\zeta_{1_{ij}} = 0$ for $i \neq j$ with $i = 1, \ldots, 7$ and $j = 1, \ldots, 20$. We define the $8 \times 23$ restriction matrix of $B_2$ and the $6 \times 19$ restriction matrix of $B_3$ in the same manner. We then establish $\tilde{B}$, $\hat{B}^s$, and $\hat{B}^{s+}$ using the same relations in Sect. 5.5.1. We calculate the empirical ADR for the sample of 1000 using the following relations:

$$\mathrm{ADR}(\hat{B}) = \frac{1}{1000} \sum_{i=1}^{1000} n \ \mathrm{tr}\left((\hat{B}^{(i)} - \bar{\hat{B}})^\top (\hat{B}^{(i)} - \bar{\hat{B}})\right),$$

$$\mathrm{ADR}(\tilde{B}) = \frac{1}{1000} \sum_{i=1}^{1000} n \ \mathrm{tr}\left((\tilde{B}^{(i)} - \bar{\tilde{B}})^\top (\tilde{B}^{(i)} - \bar{\tilde{B}})\right),$$

$$\mathrm{ADR}(\hat{B}^s) = \frac{1}{1000} \sum_{i=1}^{1000} n \ \mathrm{tr}\left((\hat{B}^{s(i)} - \bar{\hat{B}}^s)^\top (\hat{B}^{s(i)} - \bar{\hat{B}}^s)\right),$$

$$\mathrm{ADR}(\hat{B}^{s+}) = \frac{1}{1000} \sum_{i=1}^{1000} n \ \mathrm{tr}\left((\hat{B}^{s+(i)} - \bar{\hat{B}}^{s+})^\top (\hat{B}^{s+(i)} - \bar{\hat{B}}^{s+})\right),$$

where $\hat{B}^{(i)}$, $\tilde{B}^{(i)}$, $\hat{B}^{s(i)}$, and $\hat{B}^{s+(i)}$ are the UE, RE, and SEs, respectively, at the $i$th replicate. We calculate the RMSE of the estimators using these relations. The numerical results are summarized in Table 5.1.

From Table 5.1, the RMSEs of the SEs are larger than 1, which shows that the shrinkage estimators are more efficient than the UE. This confirms that, even

**Table 5.1** RMSE of the proposed estimators for neuroimaging data

| RMSE | | | |
|---|---|---|---|
| $\hat{B}_n$ | $\tilde{B}_n$ | $\hat{B}^s$ | $\hat{B}^{s+}$ |
| 1 | $7.49 \times 10^{16}$ | $1 + 6.95 \times 10^{-8}$ | $1 + 1.82 \times 10^{-8}$ |

when working with real neuroimaging data, the SEs perform better than the UE. In addition, as the RMSE of the restricted matrix is very large, this may be an indication that the hypothesized restriction may hold.

## 5.6 Conclusion

In this chapter, we consider the estimation problem in a statistical model involving high-dimensional settings. In particular, we propose some efficient estimators in the context of a tensor regression model. This area of study continues to grow as information and technology progress. For example, there is now great interest in analyzing high-dimensional neuroimaging data such as MRI and fMRI. Classical regression is hardly applicable in such contexts as it requires an unrealistic sample size. As such, many researchers who study this field propose dimension-reducing decomposition of the tensor coefficient to make estimation more efficient. We follow suit by proposing a method based on tensor decomposition and the block relaxation algorithm as in Zhou et al. (2013). We also consider the case where some restrictions upon the tensor coefficient may further reduce the number of parameters to be estimated. Under these restrictions, we establish a RE. We investigate the joint asymptotic distribution of the UE and RE. We then propose SEs which are established by combining the UE and RE. In addition, we show that the SEs dominate the UE. To the best of our knowledge, we are the first to propose shrinkage estimators in the tensor regression context. To illustrate the application of our theoretical results, we perform some simulation studies and analyze a real neuroimaging data set.

In the future, we wish to extend the established results by including estimation with higher rank decomposition of the coefficient tensor. This is expected to provide even more efficient estimators for a tensor regression model involving more complex signals. We will also establish asymptotic theory in the context of the tensor regression model under high-dimensional settings.

**Acknowledgments** The authors would like to thank the referees for helpful comments and useful insights. Further, Dr. S. Nkurunziza would like to acknowledge the financial support received from Discovery Grant (RGPIN-2019-04464) of the Natural Sciences and Engineering Research Council of Canada (NSERC).

## Appendix

### *Identifiability Problem*

In this subsection, we outline three important factors that may affect the identifiability of the parameter of interest, $\mathscr{B}$: scaling, permutation, and uniqueness of

decomposition. As in Zhou et al. (2013), to overcome the identifiability problem, we restrict the parameter space so that we fix the scaling and permutation of $\mathscr{B}$. Namely, as in Zhou et al. (2013), we introduce a specific constrained parameterization: for $D > 2$, they suggest to put $(D-1)R$ restrictions on $B$. They do this by scaling $B_1, \ldots, B_{D-1}$ such that $\beta_{d1}^{(r)} = 1$, i.e., the first rows are ones. This sets the entries in the front row of $B_D$ and thus fixes scaling indeterminacy. For permutation indeterminacy, they assume the first row entries of $B_D$ are distinct and arranged in descending order. This gives rise to the restricted parameter space:

$$\mathbb{B} = \Big\{(B_1, \ldots, B_D) : \beta_{d1}^{(r)},\ \forall\, d = 1, \ldots, D,\ r = 1, \ldots, R, \quad \text{and} \quad \beta_{D1}^{(1)} > \cdots > \beta_{D1}^{(R)}\Big\}$$

which is open and convex. Namely, given a rank-$R$ decomposition of an array, let $B = [B_1, \ldots, B_D]$.

1. Scaling and permutation: these are defined as follows.
   a. Scaling: $B = [B_1\Lambda_1, \ldots, B_D\Lambda_D]$ for any diagonal matrices:
   $$\Lambda_d = \text{diag}(\lambda_{d1}, \ldots, \lambda_{dR})$$
   for $d = 1, \ldots, D$ such that $\Pi_d \lambda_{dr} = 1$ for $r = 1, \ldots, R$.
   b. Permutation: $B = [B_1\Pi, \ldots, B_D\Pi]$ for any $R$ by $R$ permutation matrix $\Pi$.
2. Uniqueness of decomposition of tensor: There may be also be a problem with the uniqueness of the decomposition of the tensor even after adjusting for scaling and permutation.

As explained in Zhou et al. (2013), these factors affect the determinacy of the true parameter and, as a result, the convergence of the unrestricted and the restricted estimators. The following set of propositions provide conditions that are useful for checking for uniqueness of decomposition for tensors.

**Proposition 5.6** *Suppose that a D-dimensional array $B \in \mathbb{R}^{p_1 \times \cdots \times p_2}$ has rank R.*

*Sufficiency: The decomposition is unique up to scaling and permutation if $\sum_{d=1}^{D} k_{B_d} \geqslant 2R + (D-1)$, where $k_A$ is the k-rank of a matrix A, that is, the maximum value k such that any k columns are linearly independent.*

*Necessity: If the decomposition is unique up to scaling and permutation, then $\min_{d=1,\ldots,D} rank(B_1 \odot \cdots \odot B_{d-1} \odot B_{d+1} \odot \cdots \odot B_D) = R$, which in turn implies that $\min_{d=1,\ldots,D}(\prod_{d' \neq d} rank(B_{d'})) \geqslant R$.*

**Remark 5.3** When $D = 3$, $R \leqslant p_3$ and $R(R-1) \leqslant p_1(p_1-1)p_2(p_2-1)/2$, the decomposition is unique for almost all such tensors except on a set of Lebesgue measure zero. When $D = 4$, $R \leqslant p_4$ and $R(R-1) \leqslant p_1p_2p_3(3p_1p_2p_3 - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, the decomposition is unique for almost all such tensors except on a set of Lebesgue measure zero.

For the detailed proof of this proposition, we refer the reader to Zhou et al. (2013). In the following proposition, we propose a sufficient and necessary condition for local identifiability.

**Proposition 5.7 (Identifiability)** *Given i.i.d. data points* $\{(y_i, x_i), i = 1, \ldots, n\}$ *from the tensor regression model, let* $B_0 \in \mathbb{B}$ *be a parameter point and assume there exists an open neighborhood of* $\boldsymbol{B}_0$ *in which the information matrix has a constant rank. Also, let* $J_d \in \mathbb{R}^{\prod_{d=1}^{D} p_d \times p_d R}$ *be the Jacobian of* $B$ *where* $J_d = \Pi_d[(B_D \odot \cdots \odot B_{d+1} \odot B_{d-1} \odot \cdots \odot B_1) \otimes I_{p_d}]$, *for permutation matrix* $\Pi_d$ *that reorders Vec* $B_{(d)}$ *to get Vec* $B$. *Then* $\boldsymbol{B}_0$ *is locally identifiable up to permutation if and only if the Fisher information matrix of* $\boldsymbol{B}_0$,

$$I_n(\boldsymbol{B}_0) = [J_1 \cdots J_D]^\top \left[ \sum_{i=1}^{n} \frac{\mu^\top(\eta_i)^2}{\sigma_i^2} (\mathrm{Vec}(x_i))(\mathrm{Vec}(x_i))^\top \right] [J_1 \cdots J_D],$$

*is nonsingular.*

The proof of this proposition is given in Zhou et al. (2013). Proposition 5.7 plays an important role in deriving asymptotic properties of the RE and UE.

## *Proof of Theorem 5.2*

We have

$$\sqrt{n}\big(\tilde{\theta}_n - \theta_0\big) = (I_p - A_n)\sqrt{n}(\hat{\theta}_n - \theta_0) + C_n, \tag{5.15}$$

where

$$\begin{aligned} A_n &:= (\mathcal{X}^\top \mathcal{X})^{-1} \zeta^\top [\zeta (\mathcal{X}^\top \mathcal{X})^{-1} \zeta']^{-1} \zeta, \\ C_n &:= -(\mathcal{X}^\top \mathcal{X})^{-1} \zeta^\top [\zeta (\mathcal{X}^\top \mathcal{X})^{-1} \zeta^\top]^{-1} \triangle. \end{aligned} \tag{5.16}$$

Similarly, we have

$$\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) = -C_n + A_n \sqrt{n}(\hat{\theta}_n - \theta_0). \tag{5.17}$$

Hence, we have

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ \sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \end{pmatrix} = \begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0) \\ C_n + (I_p - A_n)\sqrt{n}(\hat{\theta}_n - \theta_0) \\ -C_n + A_n\sqrt{n}(\hat{\theta}_n - \theta_0) \end{pmatrix}$$
$$= \begin{pmatrix} 0 \\ C_n \\ -C_n \end{pmatrix} + \begin{pmatrix} I_p \\ I_p - A_n \\ A_n \end{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta_0).$$

Then, by Assumption 5.2, we get

$$I_p - A_n \xrightarrow[n\to\infty]{P} M, \tag{5.18}$$

and

$$C_n \xrightarrow[n\to\infty]{P} \epsilon_2 = -\boldsymbol{J}\Delta. \tag{5.19}$$

Therefore, the result follows from relations (5.18) and (5.19) and Proposition 5.1 along with Slutsky's theorem. This completes the proof. □

## *Proof of Proposition 5.5*

We will first find the conditions for $\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) \le \mathrm{ADR}(\hat{\theta}_n, \theta_0, W)$ to hold. To do so, we show that $\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) \le 0$. Recall that

$$\begin{aligned} \mathrm{ADR}(\hat{\theta}^s, \theta_0, W) = \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) &+ 4(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1 \\ &-2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}) \\ &+(k^2-4)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top W\epsilon_1 \\ &+(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W\Gamma_{11}). \end{aligned}$$

We have

$$\begin{aligned} \mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}) \\ &+(k^2-4)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top W\epsilon_1 \\ &+(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W\Gamma_{11}) \\ &+4(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1. \end{aligned}$$

Then,

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))\\ &+(k-2)(k+2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top W\epsilon_1\\ &+4(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1\\ &+(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W\Gamma_{11}).\end{aligned}$$

So,

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))\\ &+(k-2)(k+2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A\Gamma_{11}W\epsilon_1\\ &+(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W\Gamma_{11})\\ &+(k-2)(k+2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A4\Gamma_{12}/(k+2)W\epsilon_1,\end{aligned}$$

which can be simplified to give

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12})\\ &+(k^2-4)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]\epsilon_1^\top A(\Gamma_{11}+4\Gamma_{12}\,/\,(k+2))W\epsilon_1\\ &+(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W\Gamma_{11}).\end{aligned}$$

Hence, we have

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)]C_1\\ &+(k^2-4)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]C_2\\ &+(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]C_3,\end{aligned}$$

where $C_1 = \mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))$, $C_2 = \epsilon_1^\top A(\Gamma_{11}+4\Gamma_{12}/(p+2))W\epsilon_1$, and $C_3 = \mathrm{tr}(W\Gamma_{11})$. Now, using the identity from Fu et al. (2014), we have

$$\mathrm{E}[\chi_{k+2}^{-2}(\omega)] = (k-2)\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)] + 2\omega\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)].$$

Therefore,

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-2(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]C_1 \\ &-4\omega(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]C_1 \\ &+(k^2-4)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)]C_2 \\ &(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]C_3.\end{aligned}$$

Then,

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)](2C_1 - C_3) \\ &-(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)](4\omega C_1 - (k+2)C_2).\end{aligned}$$

Since

$$\begin{aligned}2C_1 - C_3 &= 2\,\mathrm{tr}(W(\Gamma_{11} + \Gamma_{12})) - \mathrm{tr}(W\Gamma_{11}) \\ &= 2\,\mathrm{tr}(W\Gamma_{11}) + 2\,\mathrm{tr}(W\Gamma_{12}) - \mathrm{tr}(W\Gamma_{11}) \\ &= \mathrm{tr}(W(\Gamma_{11} + 2\Gamma_{12})),\end{aligned}$$

we can simplify the expression further to get

$$\begin{aligned}\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) = &-(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)]\,\mathrm{tr}(W(\Gamma_{11} + 2\Gamma_{12}) \\ &-(k-2)\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)](4\omega C_1 - (k+2)C_2).\end{aligned}$$

Thus, since $k \geq 2$, we conclude that for $\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) - \mathrm{ADR}(\hat{\theta}_n, \theta_0, W) \leq 0$, it suffices to have:

1. $\mathrm{tr}(W(\Gamma_{11} + 2\Gamma_{12})) \geq 0$.
2. $4\omega C_1 - (k+2)C_2 \geq 0$.

If $C_2 = 0$, then $4\omega C_1 - (k+2)C_2 \geq 0$ for all $\omega \geq 0$ since $C_1 = \mathrm{tr}(WJI^{-1}(B_0))$ and $JI^{-1}(B_0)$ is nonnegative definite. If $C_2 > 0$, then $4\omega C_1 - (k+2)C_2 \geq 0$ if $C_1 \geq (k+2)C_2/4\omega$. Now, recalling that $C_2 = \epsilon_1^\top A(\Gamma_{11} + 4\Gamma_{12}/(k+2))W\epsilon_1 = \epsilon_1^\top A(\Gamma_{11} + 4\Gamma_{12}/(k+2))W\Gamma_{11}A\epsilon_1$, we have

$$\frac{C_2}{\omega} = \frac{\epsilon_1^\top A(\Gamma_{11} + 4\Gamma_{12}/(k+2))W\Gamma_{11}A\epsilon_1}{\epsilon_1^\top A\epsilon_1}$$

$$= \frac{(A^{1/2}\epsilon_1)^\top A^{1/2}(\Gamma_{11} + 4\Gamma_{12}/(k+2))W\Gamma_{11}A^{1/2}A^{1/2}\epsilon_1}{(A^{1/2}\epsilon_1)^\top (A^{1/2}\epsilon_1)}.$$

Now, since $\Pi^*$ is symmetric by definition, we have

$$\frac{(A^{1/2}\epsilon_1)^\top \Pi^*(A^{1/2}\epsilon_1)}{(A^{1/2}\epsilon_1)^\top (A^{1/2}\epsilon_1)} \leq \Lambda_{\max}(\Pi^*).$$

Since $C_2 > 0$, then $\Pi_1^* > 0$, also, because $\epsilon_1 > 0$, $A > 0$, $\Gamma_{11} > 0$, and $W > 0$. For a matrix $X$, let $X > 0$ stand for $X$ is a positive definite matrix. Further, let $X \geq 0$ stand for $X$ is a nonnegative definite matrix. Hence, we have

$$\frac{\Pi_1^*}{2} \leq \frac{\Pi_1^* + \Pi_1^{*\top}}{2} = \Pi^*.$$

This implies that

$$\frac{(A^{1/2}\epsilon_1)^\top \frac{\Pi_1^*}{2}(A^{1/2}\epsilon_1)}{(A^{1/2}\epsilon_1)^\top (A^{1/2}\epsilon_1)} \leq \frac{(A^{1/2}\epsilon_1)^\top \Pi^*(A^{1/2}\epsilon_1)}{(A^{1/2}\epsilon_1)^\top (A^{1/2}\epsilon_1)} \leq \Lambda_{\max}(\Pi^*),$$

which results in

$$\frac{C_2}{2\omega} \leq \Lambda_{\max}(\Pi^*),$$

and this implies that

$$\frac{(k+2)C_2}{4\omega} \leq \frac{(k+2)\Lambda_{\max}(\Pi^*)}{2}.$$

Therefore, for $4\omega C_1 - (k+2)C_2 \geq 0$ to hold it suffices to have

$$C_1 \geq (k+2)\Lambda_{\max}(\Pi^*)/2. \tag{5.20}$$

In addition, since $\operatorname{tr}(W(\Gamma_{11} + 2\Gamma_{12})) \geq 0$, we have

$$\operatorname{tr}(W(\Gamma_{11} + \Gamma_{12})) + \operatorname{tr}(W\Gamma_{12}) \geq 0,$$

which means that $\operatorname{tr}(W(\Gamma_{11} + \Gamma_{12})) \geq -\operatorname{tr}(W\Gamma_{12})$. Hence, we also have

$$C_1 \geq -\operatorname{tr}(W\Gamma_{12}). \tag{5.21}$$

Combining Eqs. (5.20) and (5.21), we have $\mathrm{ADR}(\hat{\theta}^s, \theta_0, W) \leq \mathrm{ADR}(\hat{\theta}_n, \theta_0, W)$ provided that

$$C_1 \geq \max(-\operatorname{tr}(W\Gamma_{12}), (k+2)\Lambda_{\max}(\Pi^*)/2). \tag{5.22}$$

We will now determine the conditions for $\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) \leq \mathrm{ADR}(\hat{\theta}^{s}, \theta_0, W)$ to hold. To do so, we show that

$$\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) - \mathrm{ADR}(\hat{\theta}^{s}, \theta_0, W) \leq 0.$$

Recall that

$$\begin{aligned}
&\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) = \mathrm{ADR}(\hat{\theta}^{s}, \theta_0, W) \\
&+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\epsilon_1^\top W\epsilon_1 \\
&+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1 \\
&-2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{12}) \\
&-2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \leq k-2) - (k-2)\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \leq k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1 \\
&-\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) - 2(k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) \\
&\quad +(k-2)^2\chi^{-4}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{11}) \\
&-\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \leq k-2) - 2(k-2)\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \leq k-2) \\
&+(k-2)^2\chi^{-4}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \leq k-2)]\epsilon_1^\top W\epsilon_1.
\end{aligned}$$

We simplify the expression as

$$\begin{aligned}
&\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) = \mathrm{ADR}(\hat{\theta}^{s}, \theta_0, W) \\
&+\mathrm{E}[(k-2)^2\chi^{-4}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \leq k-2)]\epsilon_1^\top W\epsilon_1 \\
&+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq (k-2))]\epsilon_1^\top W\epsilon_1 \\
&+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1 \\
&-2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{12}) \\
&+2(k-2)\,\mathrm{E}[\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{12}) \\
&-\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{11}) \\
&+2(k-2)\,\mathrm{E}[\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{11}) \\
&-(k-2)^2\,\mathrm{E}[\chi^{-4}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \leq k-2)]\,\mathrm{tr}(W\Gamma_{11})
\end{aligned}$$

$$-2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2) - (k-2)\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1$$

$$-\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2) - 2(k-2)\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)].$$

Then,

$$\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) = \mathrm{ADR}(\hat{\theta}^{s}, \theta_0, W)$$

$$+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\epsilon_1^\top W\epsilon_1$$

$$+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1$$

$$-\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+2\Gamma_{12}))$$

$$+2(k-2)\,\mathrm{E}[\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))$$

$$-(k-2)^2\,\mathrm{E}[\chi^{-4}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\,\mathrm{tr}(W\Gamma_{11})$$

$$-2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2) - (k-2)\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1$$

$$-\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2) - 2(k-2)\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)$$

$$+(k-2)^2\chi^{-4}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)]\epsilon_1^\top W\epsilon_1.$$

Thus, we have

$$\mathrm{ADR}(\hat{\theta}^{s+}, \theta_0, W) = \mathrm{ADR}(\hat{\theta}^{s}, \theta_0, W)$$

$$+2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\epsilon_1^\top A\Gamma_{11}W\epsilon_1$$

$$-2\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2) - (k-2)\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1$$

$$-\,\mathrm{E}[\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+2\Gamma_{12}))$$

$$+2(k-2)\,\mathrm{E}[\chi^{-2}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))$$

$$-(k-2)^2\,\mathrm{E}[\chi^{-4}_{k+2}(\omega)\mathbb{I}(\chi^2_{k+2}(\omega) \le k-2)]\,\mathrm{tr}(W\Gamma_{11})$$

$$-\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)]\epsilon_1^\top A2\Gamma_{12}W\epsilon_1$$

$$+2(k-2)\,\mathrm{E}[\chi^{-2}_{k+4}(\omega)\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1$$

$$-\,\mathrm{E}[\mathbb{I}(\chi^2_{k+4}(\omega) \le k-2)]\epsilon_1 A\Gamma_{11}W\epsilon_1$$

$$+2(k-2)\,\mathrm{E}[\chi_{k+4}^{-2}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1 A\Gamma_{11}W\epsilon_1$$

$$-(k-2)^2\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1 A\Gamma_{11}W\epsilon_1.$$

This gives

$$\begin{aligned}
\mathrm{ADR}(\hat{\theta}^{s+},\theta_0,W) &= \mathrm{ADR}(\hat{\theta}^{s},\theta_0,W)\\
&+2\,\mathrm{E}[\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)-(k-2)\chi_{k+2}^{-2}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\\
&\times\epsilon_1^\top A(\Gamma_{11}+\Gamma_{12})W\epsilon_1\\
&-\mathrm{E}[\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+2\Gamma_{12}))\\
&+2(k-2)\,\mathrm{E}[\chi_{k+2}^{-2}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))\\
&-(k-2)^2\,\mathrm{E}[\chi_{k+2}^{-4}(\omega)\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\,\mathrm{tr}(W\Gamma_{11})\\
&-\mathrm{E}[\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1^\top A(\Gamma_{11}+\Gamma_{12})\\
&+2(k-2)\,\mathrm{E}[\chi_{k+4}^{-2}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1 A(\Gamma_{11}+\Gamma_{12})W\epsilon_1\\
&\quad-(k-2)^2\,\mathrm{E}[\chi_{k+4}^{-4}(\omega)\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1 A\Gamma_{11}W\epsilon_1.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathrm{ADR}(\hat{\theta}^{s+},\theta_0,W)-\mathrm{ADR}(\hat{\theta}^{s},\theta_0,W)\\
&=2\,\mathrm{E}[(1-(k-2)\chi_{k+2}^{-2}(\omega))\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\epsilon_1^\top A(\Gamma_{11}+\Gamma_{12})W\epsilon_1\\
&-\mathrm{E}[(1-(k-2)\chi_{k+2}^{-2}(\omega))^2\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\,\mathrm{tr}(W(\Gamma_{11}+\Gamma_{12}))\\
&-\mathrm{E}[(1-(k-2)^2\chi_{k+2}^{-4}(\omega))\mathbb{I}(\chi_{k+2}^2(\omega)\le k-2)]\,\mathrm{tr}(W\Gamma_{12})\\
&-\mathrm{E}[(1-(k-2)\chi_{k+4}^{-2}(\omega))^2\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1^\top A(\Gamma_{11}+\Gamma_{12})W\epsilon_1\\
&-\mathrm{E}[(1-(k-2)^2\chi_{k+4}^{-4}(\omega))\mathbb{I}(\chi_{k+4}^2(\omega)\le k-2)]\epsilon_1^\top A\Gamma_{12}W\epsilon_1.
\end{aligned}$$

Since

$$\mathrm{E}[(1-(k-2)\chi_{k+c}^{-2}(\omega))\mathbb{I}(\chi_{k+c}^2(\omega)\le k-2)]\le 0,$$

and

$$\mathrm{E}[(1-(k-2)\chi_{k+c}^{-2}(\omega))^2\mathbb{I}(\chi_{k+c}^2(\omega)\le k-2)]\ge 0,$$

for $c = 2, 4$, respectively, then for $\text{ADR}(\hat{\theta}^{s+}, \theta_0, W) - \text{ADR}(\hat{\theta}^{s}, \theta_0, W) \leq 0$, it suffices to have:

1. $\epsilon_1^\top A(\Gamma_{11} + \Gamma_{12})W\epsilon_1 \geq 0$.
2. $\text{tr}(W(\Gamma_{11} + \Gamma_{12})) \geq 0$.
3. $\text{tr}(W\Gamma_{12}) \leq 0$.
4. $\epsilon_1^\top A\Gamma_{12}W\epsilon_1 \leq 0$.

Therefore, using Eqs. (5.22), relation (5.14) holds if:

1. $-\Lambda_{\min}(W\Gamma_{11}) \leq \Lambda_{\min}(W(\Gamma_{12} + \Gamma_{21})/2)$.
2. $\text{tr}(W\Gamma_{12}) \leq 0$.
3. $\text{tr}(W(\Gamma_{11} + \Gamma 12)) \geq \max(-\text{tr}(W\Gamma_{12}), (k+2)\Lambda_{\max}(\Pi^*)/2)$.

This completes the proof. □

## References


Ahmed, S. E., Hossain, S., & Doksum, K. A. (2012). Lasso and shrinkage estimation in Weibull censored regression models. *Journal of Statistical Planning and Inference, 12*(6), 1273–1284.

Chen, F., & Nkurunziza, S. (2015). Optimal method in multiple regression with structural changes. *Bernoulli, 21*(4), 2217–2241.

Chen, F., & Nkurunziza, S. (2016). A class of Stein-rules in multivariate regression model with structural changes. *Scandinavian Journal of Statistics, 43*, 83–102.

De Leeuw, J. (1994). Block-relaxation algorithms in statistics. In *Information systems and data analysis* (pp. 308–324).

Fu, Y., Gao, J., Hong, X., & Tien, D. (2014). Tensor regression based on linked multiway parameter analysis. In *IEEE International Conference on Data Mining* (pp. 821–826).

Guo, W., Kotsia, I., & Patras, I. (2012). Tensor learning for regression. *Image Processing, 21*(2), 816–827.

Hossain, S., Ahmed, S. E., Yi, G. Y., & Chen, B. (2016). Shrinkage and pretest estimators for longitudinal data analysis under partially linear models. *Journal of Nonparametric Statistics, 28*(3), 531–549.

Hossain, S., & Ahmed, S. E.(2012). Shrinkage and penalty estimators of a poisson regression model. *Australian & New Zealand Journal of Statistics, 54*(3), 359–373.

James, W., & Stein, C. (1961). Estimation with quadratic loss, In *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol, 1, pp. 361–379).

Judge, G. G., & Bock, M. E. (1978) *The statistical implications of pre-test and Stein-rule estimators in econometrics*. North-Holland.

Lee, H. Y., Reisi Gahrooei, M., Liu, H., & Pacella, M. (2023). Robust tensor-on-tensor regression for multidimensional data modeling. *IISE Transactions, 56*(1), 43–53.

Li, X. (2014). *Tensor Based Statistical Models with Applications in Neuroimaging Data Analysis*. PhD Thesis, North Carolina State University.

Li, L., & Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association, 112*, 1131–1146.

Liu, J., Wu, Z., Xiao, L., Sun, J., & Yan, H. (2019). Generalized tensor regression for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 58*, 1244–1258.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics, 27*, 638–647.

Nkurunziza, S., & Ahmed, S. E. (2010). Shrinkage drift parameter estimation for multi-factor Ornstein-Uhlenbeck processes. *Applied Stochastic Models in Business and Industry, 26*(2), 103–124.

Nkurunziza, S. (2012). The risk of pretest and shrinkage estimators. *Statistics, 46*(3), 305–312.

Saleh, A. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. John Wiley & Sons.

Sen, P. K., & Saleh, A. K. Md. E. (1985). On some shrinkage estimators of multivariate location. *Annals of Statistic, 13*, 272–281.

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y. J., & Li, F. (2013). A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies, 28*, 12–27.

Li, X., Xu, D., & Zhou, H. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences, 10*, 520–545.

Zhou, h., Li, L., & Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association, 108*(502), 540–552.

# Chapter 6
# Some Implications of Preliminary-Test Estimation in the Context of Size-Biased Sampling




**David E. Giles**

## 6.1 Introduction

Size-biased sampling is a phenomenon that arises in a variety of fields, such as epidemiology, ecology, forestry, reliability studies, survival analysis, marketing, meteorology, genome mapping, and environmental and resource economics. Although it may not be directly observable, size-biased sampling occurs when the probability of a unit being selected into the sample is proportional to a predetermined weight function, with the latter depending on the observed value of the data for that unit. This phenomenon is discussed in the seminal contributions of Wicksell (1925), Fisher (1934), and Cox (1969). For example, if the probability of an observation being included in a sample is directly proportional to the size of that observation, we have so-called 'length-biased' sampling. Size-biased sampling is important because it affects the formulation of the appropriate data density and hence the likelihood function. If this is ignored, the likelihood function is misspecified and inferences based upon it will be invalid, *even asymptotically*.

If a random variable, $X$, has a density function $f(x;\theta)$, for some parameter (vector) $\theta$, then the corresponding 'weighted' density, to allow for size-biased sampling of order $c$, is defined as $f_c(x;\theta) = f(x;\theta)x^c/m'_c$, where $m'_r = \mathrm{E}[X^r]$ is the $r$th. moment of $X$. The case $c = 1$ corresponds to 'length-biased' sampling, while $c = 2$ corresponds to 'area-biased', etc. Size-biased sampling can arise with both discrete and continuous distributions. The empirical literature includes examples based on the discrete binomial, Poisson, and negative binomial distributions, as well as continuous ones such as the exponential, gamma, Weibull, half-normal, Nakagami, and Rayleigh distributions.

D. E. Giles (✉)
University of Victoria, Victoria, BC, Canada
e-mail: dgiles@uvic.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_6

An obvious preliminary-testing problem arises here, because in general the researcher will not know whether or not size bias is an issue in the sampling process. Interestingly, it appears that this particular preliminary-testing issue has not been analysed previously in the literature. Specifically, we might test the hypothesis, $\mathcal{H}_0$, 'simple random sampling was used', against the alternative hypothesis, $\mathcal{H}_1$, 'size-biased sampling was used'. If the null hypothesis is not rejected, we would use the maximum likelihood estimate (MLE) of $\theta$, basing the likelihood function on $f(x;\theta)$. On the other hand, if the null hypothesis is rejected, we would use the MLE of $\theta$, with the likelihood function constructed from $f_c(x;\theta) = f(x;\theta)\,x^c/m'_c$. The (sampling) properties of this preliminary-test estimator will differ from those of the two alternative MLEs and will depend on the value of $\theta$, the sample size, and the level of significance chosen for the test of $\mathcal{H}_0$.

Several tests of $\mathcal{H}_0$ against $\mathcal{H}_1$ have been proposed in the literature. For example, see Navarro et al. (2003), Akman et al. (2007), and Economou and Tzavelas (2013). In this study, we consider preliminary testing based on the simple distribution-free test introduced by Akman et al. (2007) and elaborated upon by Economou and Tzavelas (2013). Using this test and various distributions from the generalized gamma family, we investigate the relative risk of the preliminary-test estimator (and its two component estimators) under squared-error loss, as well as the relative biases of these estimators. Some of the results are obtained analytically, while others are based on the results of an extensive Monte Carlo simulation experiment.

The next section considers the preliminary-testing problem under study in more detail. The impact of size-biased sampling on the generalized gamma distribution and some of the members of this family are discussed in Sect. 6.3. The design of the Monte Carlo experiment is described in Sect. 6.4 and the associated results are presented in Sect. 6.5. Section 6.6 provides some concluding remarks.

## 6.2 Preliminary-Test Estimation

Preliminary-test estimators arise in a very wide range of situations, and their properties can be complicated. Often, researchers ignore the fact that such estimators have been used, and this can be a cause for serious concern. Such estimators have been studied extensively and deserve due attention.

### *6.2.1 The Preliminary-Testing Problem*

The preliminary-test estimation problem arises when the choice of estimator is conditioned on the application of some prior statistical test. This conditioning results in a 'preliminary-test' estimator (PTE) whose sampling properties depend upon the choice of test and significance level; the alternative estimators that are used, depending on the test's outcome; the values of all of the parameters in the problem;

and the sample size. Motivated by earlier work by Berkson (1942), preliminary-test estimation was first discussed by Bancroft (1944), in the context of estimating a regression parameter after a prior test of significance of one of the model's other parameters. Subsequently, a large literature on this topic (and that of 'preliminary-test testing') has emerged. This literature has been surveyed by a number of authors, including Bancroft and Han (1977), Judge and Bock (1978), Han et al. (1988), Giles and Giles (1993), and Saleh (2014).

### 6.2.2 Properties of Preliminary-Test Estimators

In general, preliminary-test estimators have desirable asymptotic (large sample) properties. This is the case in this study because the component estimators ($\tilde{\theta}$ and $\hat{\theta}$) are estimators. The distributions that are considered satisfy the usual regularity conditions, and so the MLEs are consistent—*if the underlying hypotheses are correct*. The preliminary-test estimator is a weighted combination of these estimators, with random weights whose properties depend on those of the sampling distribution of the test statistic, $\tilde{\lambda}$.

Provided that the test is itself consistent—ensuring that its power approaches one as the sample size grows indefinitely—then the preliminary-test estimator will be consistent. Economou and Tzavelas (2013) establish the asymptotic normality of the sampling distribution of $\tilde{\lambda}$. The consistency of the test is readily confirmed through Monte Carlo simulation, and this is discussed further in Sect. 6.4.

However, although the preliminary-test estimators under consideration are well-behaved asymptotically, their finite-sample properties require careful attention. First, it should be noted that PTEs are generally inadmissible (e.g. Cohen, 1965) due to the fact that they are a *discontinuous* weighted combination of the component estimators. Typically, they will also be biased (because the random weights generally are not independent of the test statistic's distribution), and the finite-sample risks of the estimators will be of interest. These are the properties on which the present study focuses.

### 6.2.3 Preliminary Testing for Size-Biased Sampling

Now let us consider the preliminary-test estimator of the parameter (vector) $\theta$ in the density $f(x;\theta)$, when the *potential level* of size bias is $c$. As noted in Sect. 6.1, in this study we investigate preliminary testing using the simple distribution-free test suggested by Akman et al. (2007) and extended by Economou and Tzavelas (2013). The test statistic is based on the quantity

$$\prod_{i=1}^{n}[f_c(x_i;\theta)/f(x_i;\theta)], \tag{6.1}$$

a large value of which will suggest that size-biased sampling has been used.

Using the definition of $f_c(x_i;\theta)$ and noting that $m'_c = \mathrm{E}[X^c]$ is typically an unobservable function of $\theta$, an equivalent test procedure is to reject in favour of size-biased sampling if the statistic

$$\tilde{\lambda} = \left[\prod_{i=1}^{n} x_i\right]^{1/n} /(\tilde{m}'_c))^{1/c} \tag{6.2}$$

is sufficiently large. In Eq. (6.2), $\tilde{m}'_c$ is the MLE of $m'_c$, in which $\theta$ is replaced with, $\tilde{\theta}$, its MLE based on the likelihood function formed using the 'base' (or 'unweighted') density, $f(x;\theta)$.

Let $\hat{\theta}$ denote the MLE based on the likelihood function formed using the 'weighted' density, $f_c(x;\theta)$. Then, the preliminary-test estimator, $\theta^*$, is equal to $\tilde{\theta}$ if $\tilde{\lambda} > k(\alpha)$, and it equals $\hat{\theta}$ if $\tilde{\lambda} \leq k(\alpha)$. Here, $k$ is a critical value that is determined by simulation and is dependent on the chosen significance level, $\alpha$. So, we can express the estimator as

$$\theta^* = \hat{\theta} I_{k(\alpha)} + \tilde{\theta}(1 - I_{k(\alpha)}) = \hat{\theta} + (\hat{\theta} - \tilde{\theta}) I_{k(\alpha)}, \tag{6.3}$$

where $I_{k(\alpha)}$ is an indicator function taking the value 1 if $\tilde{\lambda} > k(\alpha)$ and zero if $\tilde{\lambda} \leq k(\alpha)$.

We will be concerned with the relative bias and relative risk (both in percentage terms) of this preliminary-test estimator, together with the corresponding measures for $\tilde{\theta}$ and $\hat{\theta}$, under various scenarios. The %MSE provides the relative risk under quadratic loss, multiplied by 100. In this study, the biases and risks of $\tilde{\theta}$ and $\hat{\theta}$ can be obtained analytically for the three distributions that we consider. However, for all of the distributions that we explore, the biases and risks of the preliminary-test estimators will be obtained through Monte Carlo simulation, given the complexity of their structure.

We also consider the relative risks under both absolute error and LINEX loss functions, again expressed as percentages. Under the former loss function the percentage relative risk equals the mean relative absolute percentage error (%MAE). The LINEX loss function, proposed by Varian (1975) and elaborated upon by Zellner (1986), allows for an asymmetric loss structure. A simple form of this loss function is given by

$$L(\bar{\theta};\theta) = \exp\left\{w(\bar{\theta} - \theta)\right\} - w(\bar{\theta} - \theta) - 1; \quad w \neq 0 \tag{6.4}$$

where $\bar{\theta}$ is any one of the estimators of $\theta$ under consideration, and the parameter '$w$' is assigned to determine the degree of asymmetry. If $w > 0$, the overestimation of

$\theta$ is (exponentially) more costly than underestimation, while the converse is true if $w < 0$. For values of '$w$' close to zero, the loss function is approximately quadratic. The relative risk (in percentage terms) under LINEX loss is

$$\%\text{rel}R_L(\bar{\theta};\theta) = 100\,\text{E}[\exp\{w(\bar{\theta}-\theta)\} - w(\bar{\theta}-\theta) - 1]/\theta. \tag{6.5}$$

## 6.3 The Generalized Gamma Distribution and Size-Biased Sampling

Our focus in this chapter is on a distribution family that includes many specific distributions that are of practical importance in the presence of possible size-biased sampling. Some of these particular distributions are discussed in detail below.

### *6.3.1 Stacy's Distribution*

Stacy (1962) proposed a three-parameter generalization of the gamma probability distribution, which is very flexible in its form and contains a number of well-known distributions as special cases. The probability density function (p.d.f.) for a random variable, $Y$, that follows this (G-gamma) distribution is

$$f(y;a,d,p) = (p/a^d)y^{d-1}\exp\{-(y/a)^p\}/\Gamma(d/p); \quad y>0; a,d,p>0\,. \tag{6.6}$$

The $r$th raw moment of $Y$ is

$$\text{E}[Y^r] = a^r\Gamma((d+r)/p)/\Gamma(d/p). \tag{6.7}$$

So, under general size-biased sampling, the density of $Y$ becomes

$$f_c(y;a,d,p) = (p/a^{(d+c)})y^{(d+c)-1}\exp\{-(y/a)^p\}/\Gamma((d+c)/p); \quad y>0; a,d,p>0. \tag{6.8}$$

This is the G-gamma density with parameters $a$, $(d+c)$, and $p$, for $c = 1, 2, \ldots$ So, the G-gamma distribution has the property that call 'form invariance' under weighted sampling (Patil & Ord, 1976). See, also, Ducey and Gove (2015, p. 124).

Jabeen and Jan (2015) consider the G-gamma family as a whole when the sampling is size-biased, but they focus on the associated information measures. Several members of the G-gamma family of distributions are among those that have been applied in a wide variety of fields in the context of size-biased sampling. For example, Mir et al. (2013) and others discuss the exponential distribution ($d = p = 1$). The Weibull distribution ($d = p$) is investigated by several

authors, including Blumenthal (1967); Scheaffer (1972); Gove (2003a); Das and Roy (2011a); Giles (2021). Blumenthal (1967) and Scheaffer (1972) also consider the gamma distribution ($p = 1$) under size-biased sampling; and the Rayleigh distribution ($a = \sqrt{(2\sigma^2)}, d = p = 2$) is discussed by Das and Roy (2011b). The Nakagami distribution ($a = \sqrt{(\Omega/m)}, d = 2m, p = 2$) under length-biased sampling is investigated by Mudasir and Ahmad (2018), and Bashir and Rasul (2018) discuss the half-normal distribution ($a = \sqrt{(2\sigma^2)}, d = 1, p = 2$) in this context. (Here, $m$ and $\Omega$ are the shape and spread parameters of the Nakagami distribution, respectively, and $\sigma$ is the scale parameter for the half-normal and Rayleigh distributions.) Three of these distributions are discussed in the next subsection, and our attention will focus purely on length-biased sampling, rather than more general forms of size bias.

### 6.3.2 *Some Single-Parameter Distributions*

We consider three single-parameter members of the G-gamma family for which the MLEs of the parameters can be obtained analytically, and so many of the associated biases and risks can be expressed in closed form. For most members of the G-gamma family the MLEs must be obtained numerically, which precludes closed form expressions for the associated biases and risks.

**Exponential Distribution**

To elaborate on some of the general concepts introduced so far, consider the particular case where the random variable, $X$, follows an exponential distribution with a rate parameter, $\beta$. So, its p.d.f. is

$$f(x;\beta) = \beta \exp\{-\beta x\}, \qquad x > 0; \beta > 0 \tag{6.9}$$

and $\mathrm{E}[X] = 1/\beta$. This is a special case of the G-gamma distribution, with parameters $a = 1/\beta$, and $d = p = 1$. It is well known that the MLE of $\beta$ based on a sample of $n$ independent observations and Eq. (6.9) is given by $\tilde{\beta} = 1/\bar{x}$, where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$. Using the result that the sum of $n$ independent exponential random variables with a common rate parameter $\beta$ follows a gamma $(n, \beta)$ distribution, it follows immediately that $\mathrm{E}[\tilde{\beta}] = n\beta/(n-1)$ and $\mathrm{var}(\tilde{\beta}) = n^2\beta^2/[(n-1)(n-2)]$. So, if Eq. (6.9) is the correct density for $X$, the relative bias and relative mean squared error (MSE) of the MLE, in percentage terms, are

$$\%\mathrm{relBias}(\tilde{\beta}) = 100/(n-1) \tag{6.10}$$

and

$$\%\mathrm{relMSE}(\tilde{\beta}) = 100(n+2)/[(n-1)(n-2)], \tag{6.11}$$

respectively.

Under length-biased sampling, the p.d.f. for $X$ becomes

$$f_1(x;\beta) = xf(x;\beta)/\,\mathrm{E}[X] = x\beta^2\exp\{-\beta x\}; \qquad x > 0; \beta > 0 \tag{6.12}$$

which is the density for a gamma $(2, \beta)$-distributed random variable with a rate parameter, $\beta$.

In this case, the MLE for $\beta$ is $\hat{\beta} = 2/\bar{x}$, and the sum of the $n$ independent $x$s follows a gamma $(2n, \beta)$ distribution. It is easily shown that $\mathrm{E}[\hat{\beta}] = 2n\beta/(2n-1)$ and $\mathrm{var}(\hat{\beta}) = 2n^2\beta^2/[(n-1)(2n-1)^2]$. So, if Eq. (6.12) is the correct density for $X$, the relative bias and relative mean squared error (MSE) of the MLE are, in percentage terms,

$$\%\mathrm{relBias}(\hat{\beta}) = 100/(2n-1) \tag{6.13}$$

and

$$\%\mathrm{relMSE}(\hat{\beta}) = 100(n+1)/[(n-1)(2n-1)], \tag{6.14}$$

respectively.

Now, consider the situation where Eq. (6.9) is the correct density, but the rate parameter is estimated under the incorrect assumption that length-biased sampling has been used. In this case, it is easily shown that $\mathrm{E}[\hat{\beta}] = 2n\beta/(n-1)$ and $\mathrm{var}(\hat{\beta}) = 4n^2\beta^2/[(n-2)(n-1)^2]$, and so

$$\%\mathrm{relBias}(\hat{\beta}) = 100(n+1)/(n-1) \tag{6.15}$$

and

$$\%\mathrm{relMSE}(\hat{\beta}) = 100(n^3+4n^2-3n-2)/[(n-1)^2(n-2)]. \tag{6.16}$$

Conversely, if length-biased sampling has been used, so that Eq. (6.12) is appropriate, but this is ignored (or is unknown), we can show that that $\mathrm{E}[\tilde{\beta}] = n\beta/(2n-1)$ and $\mathrm{var}(\tilde{\beta}) = n^2\beta^2/[2(n-1)(2n-1)^2]$. It follows that

$$\%\mathrm{relBias}(\tilde{\beta}) = -100(n-1)\,/\,(2n-1) \tag{6.17}$$

and

$$\%\mathrm{relMSE}(\tilde{\beta}) = (100(2n^3-5n^2+6n-2))\,/\,[2(2n-1)^2(n-1)]. \tag{6.18}$$

Note that the expressions in Eqs. (6.10) and (6.11), as well as Eqs. (6.13)–(6.18), are independent of the true value of the rate parameter. It follows that the relative bias and relative MSE of the preliminary-test estimator of $\beta$ are also invariant to the true value of that parameter. However, the properties of the last of these estimators cannot easily be evaluated analytically, as the test statistic's distribution is unknown. These properties are explored *via* Monte Carlo simulations later in this chapter.

**Half-Normal Distribution**
Consider the half-normal distribution with a scale parameter, $\sigma$. Its p.d.f. is

$$f(x;\sigma) = (\sqrt{2/\pi}/\sigma)\exp\{-x^2/(2\sigma^2)\}; \quad x > 0; \sigma > 0. \tag{6.19}$$

In this case, $\mathrm{E}[X] = \sigma\sqrt{2/\pi}$, and so the weighted p.d.f. for length-biased sampling is

$$f_1(x;\sigma) = (x/\sigma^2)\ \exp\{-x^2/(2\sigma^2)\} \tag{6.20}$$

which is the G-gamma density with parameters $a = \sqrt{2\sigma^2}, d = p = 2$, or simply the Rayleigh density in Eq. (6.20) below. In this case, $\mathrm{E}[X] = \sigma\sqrt{\pi/2}$, and $\mathrm{E}[X^2] = 2\sigma^2$.

Based on an independent sample of $n$ values from Eq. (6.19), the MLE of the scale parameter is $\tilde{\sigma} = \sqrt{(1/n)\sum_{i=1}^n x_i^2}$. Using the results that for this distribution, $\sum_{i=1}^n x_i^2 = \sigma^2\chi^2_{(n)}$, and $\mathrm{E}[\chi_{(n)}] = \sqrt{2}\Gamma((n+1)/2)/\Gamma(n/2)$, it follows that the $\%\mathrm{relBias}(\tilde{\sigma}) = 100\sqrt{2/n}\Gamma((n+1)/2)/\Gamma(n/2) - 1$. Using the expression for $\mathrm{var}(\chi_{(n)})$ we can then derive $Var(\tilde{\sigma})$ and hence $\%MSE(\tilde{\sigma})$. On the other hand, if $\mathcal{H}_1$ is true and the random sample of data is generated by Eq. (6.20), the MLE of $\sigma$ is $\hat{\sigma} = \sqrt{(1/(2n))\sum_{i=1}^n x_i^2}$. The bias and MSE expressions for $\hat{\sigma}$ when the data are sampled from Eq. (6.19), and so $\mathcal{H}_0$ is true, are obtained in a similar manner. See the first column of Table 6.1.

On the other hand, under length-biased sampling (from Eq. (6.20)), it is easily shown that $\sum_{i=1}^n x_i^2$ is gamma$(n, 2\sigma^2)$-distributed, and its square root follows a Nakagami distribution. These results enable us to obtain the analytical expressions for the %Bias and %MSE for both $\tilde{\sigma}$ and $\hat{\sigma}$, respectively, in this case, as shown in Table 6.1.

**Rayleigh Distribution**
Finally, we can proceed in a similar manner when $X$ follows a Rayleigh distribution, with a scale parameter $\sigma$. Its p.d.f. is

$$f(x;\sigma) = (x/\sigma^2)\exp\{-x^2/(2\sigma^2)\}; \quad x > 0; \quad \sigma > 0 \tag{6.21}$$

and $\mathrm{E}[X] = \sigma\sqrt{\pi/2}$. Based on a random sample of size $n$, using Eq. (6.21), the MLE of $\sigma$ is $\tilde{\sigma} = \sqrt{(1/2n)\sum_{i=1}^n x_i^2}$.

The density in the case of length-biased sampling is

$$f_1(x;\sigma) = \sqrt{2}x^2\exp\left\{-x^2/(2\sigma^2)\right\}/[\sigma^3\sqrt{\pi}] \tag{6.22}$$

which is the G-gamma p.d.f. with parameters $a = \sqrt{2\sigma^2}, d = 3$, and $p = 2$. Based on a random sample of size $n$, using Eq. (6.22), the MLE of $\sigma$ is $\hat{\sigma} = \sqrt{(1/3n)\sum_{i=1}^n x_i^2}$.

**Table 6.1** Relative biases and relative MSEs of MLEs[a]

| MLEs | $\tilde{\beta}, \tilde{\sigma}$ | $\hat{\beta}, \hat{\sigma}$ |
|---|---|---|
| $H_0$ **True**[b] | | |
| Exponential | | |
| $(\beta)$ | | |
| Rel. bias | $1/(n-1)$ | $(n+1)/(n-1)$ |
| Rel. MSE | $(n+2)/[(n-1)(n-2)]$ | $[n^3+4n^2-3n-2]/[(n-1)^2(n-2)]$ |
| Half-normal | | |
| $(\sigma)$ | | |
| Rel. bias | $\sqrt{2/n}[\Gamma((n+1)/2)/\Gamma(n/2)]-1$ | $(1/\sqrt{n})[\Gamma((n+1)/2)/\Gamma(n/2)]-1$ |
| Rel. MSE | $2[1-\sqrt{2/n}\Gamma((n+1)/2)/\Gamma(n/2)]$ | $3/2-(2/\sqrt{n})\Gamma((n-1)/2)\Gamma(n/2)$ |
| Rayleigh | | |
| $(\sigma)$ | | |
| Rel. bias | $\Gamma(n+1/2)/(\sqrt{n}\Gamma(n))-1$ | $\sqrt{2/(3n)}\Gamma(n+1/2)/\Gamma(n)-1$ |
| Rel. MSE | $2[1-\Gamma(n+1/2)/(\sqrt{n}\Gamma(n))]$ | $5/3-2\sqrt{2/(3n)}\Gamma(n+1/2)/\Gamma(n)$ |
| $H_1$ **True**[b] | | |
| Exponential | | |
| $(\theta)$ | | |
| Rel. bias | $-(n-1)/(2n-1)$ | $1/(2n-1)$ |
| Rel. MSE | $[2n^3-5n^2+6n-2]/[2(2n-1)^2(n-1)]$ | $(n+1)/[(n-1)(2n-1)]$ |
| Half-normal | | |
| $(\sigma)$ | | |
| Rel. bias | $\sqrt{2/n}\Gamma(n+1/2)/\Gamma(n)-1$ | $\sqrt{1/n}\Gamma(n+1/2)/\Gamma(n)-1$ |
| Rel. MSE | $3-2\sqrt{2/n}\Gamma(n+1/2)/\Gamma(n)$ | $2[1-\sqrt{1/n}\Gamma(n+1/2)/\Gamma(n)]$ |
| Rayleigh[c] | | |
| $(\sigma)$ | | |
| Rel. bias | n.a. | n.a. |
| Rel. MSE | n.a. | n.a. |

[a] Multiply all results by 100 to obtain %relBias and %relMSE
[b] $H_0$ true, the population density is $f(x)$; $H_1$ true, the population density is $f_1(x)$
[c] Analytical expressions have not been established in this case

Using the result that the sum of $n$ squared Rayleigh-distributed random variables is gamma-distributed, the %Bias and %MSE of $\tilde{\sigma}$ are easily obtained when the true distribution is defined in Eq. (6.22). We can use the same result to derive the %Bias and %MSE of $\hat{\sigma}$ when Eq. (6.22) is the true distribution. These results are summarized in Table 6.1. However, the properties of $\tilde{\sigma}$ and $\hat{\sigma}$ when the true distribution is given by Eq. (6.22) are not easily obtained analytically. Table 6.2 provides an overview of the distributional relationships discussed above.

Once again, all of the relative bias and relative risk expressions are invariant to the value of the distribution's sole parameter, and the properties of the preliminary-test estimator are evaluated subsequently in a simulation experiment. Both the half-normal and Rayleigh distributions are special cases of the so-called generalized Rayleigh distribution. Xiao and Giles (2014) investigated the bias and MSE of the

**Table 6.2** Connections between the distributions

| Base density ($f(x)$) | $m'_c = \mathrm{E}[X^c]$ | Size-biased density ($f_c(x)$) |
|---|---|---|
| Exponential<br>(rate = $\beta$) | $\Gamma(c+1)/\beta^c$ | Gamma<br>(shape = $c+1$, rate = $\beta$) |
| Half-normal<br>(scale = $\sigma$) | $\sqrt{(2^c/\pi)}\sigma^c\Gamma((c+1)/2)$ | G-gamma<br>($a = \sqrt{2\sigma^2}, d = 1+c, p = 2$) |
| Rayleigh<br>(scale = $\sigma = a/\sqrt{2}$) | $2^{c/2-1}c\sigma^c\Gamma(c/2)$ | G-gamma<br>($a, d = c+2, p = 2$) |

MLEs of the parameters of the latter distribution under regular sampling, with and without analytical bias corrections.

## 6.4 A Monte Carlo Simulation Experiment

For each of the distributions discussed in the previous section, we have explored the relative risks and biases of the MLEs of their parameters under unweighted sampling, length-biased sampling, and preliminary testing. Recall that in the case of the exponential, Rayleigh, and half-normal distributions, most of these properties can be computed exactly. In all other situations, the properties of the various estimators are obtained via Monte Carlo simulation, with 100,000 replications in each case. All of the computations are undertaken using the R statistical software (R Core Team 2021), with random variables being generated using the 'ggamma' package (Saldanha and Suzuki 2022). The R code can be downloaded from https://github.com/DaveGiles1949/r-code. All of the results are invariant to the scale parameter of each of the distributions being considered, so this value (either $\sigma$ or $1/\beta$) is set to 1, without loss of generality.

### *6.4.1 The Test Statistic*

As noted in Sect. 6.2.2, the critical values for the test of $\mathcal{H}_0$ *versus* $\mathcal{H}_1$ depend on the distribution(s) in question, the sample size, and possibly the parameter value(s). The critical values are obtained within each part of the simulation study using further Monte Carlo replications. Specifically, for a particular distribution, parameter value, and sample size combination, 100,000 samples of size $n$ are generated from the null distribution. Then, for each sample the expression for the mean of the associated length-biased distribution is evaluated using the MLE(s) of the parameter(s) based on the null distribution. Equation (6.2) is then applied, thus generating 20,000 values of $\tilde{\lambda}$. For example, in the case of the exponential distribution, $m'_1 = (1/\beta)$, and replacing $\beta$ with $\tilde{\beta} = 1/\bar{x}$ yields $\tilde{\lambda} = [\prod_{i=1}^n x_i]^{1/n}/\bar{x}$. The construction of $\tilde{\lambda}$ for

**Table 6.3** Construction of the test statistic

| Base density ($f(x)$) | $m_1'$ | $\tilde{m}_1'$ | $\tilde{\lambda}$ |
|---|---|---|---|
| Exponential (rate = $\beta$) | $\frac{1}{\beta}$ | $\bar{x}$ | $\frac{[\prod_{i=1}^{n} x_i]^{1/n}}{\bar{x}}$ |
| Rayleigh (scale = $\sigma$) | $\sigma\sqrt{\frac{\pi}{2}}$ | $\frac{1}{2}\sqrt{\frac{\pi \sum_i^n x_i^2}{n}}$ | $\frac{2[\prod_{i=1}^{n} x_i]^{1/n}}{\sqrt{\frac{\pi \sum_i^n x_i^2}{n}}}$ |
| Half-normal (scale = $\sigma$) | $\sigma\sqrt{\frac{2}{\pi}}$ | $\sqrt{\frac{2 \sum_i^n x_i^2}{n\pi}}$ | $\frac{[\prod_{i=1}^{n} x_i]^{1/n}}{\sqrt{\frac{2 \sum_i^n x_i^2}{n\pi}}}$ |

the alternative hypothesis of length-biased sampling is summarized in Table 6.3 for each of the distributions under study.

The upper quantiles of the distribution of 100,000 $\tilde{\lambda}$ values provide the desired critical values for the test. A range of sample sizes is considered, and the impact of the choice of significance level for the preliminary tests is explored. The power of the $\tilde{\lambda}$ test is easily simulated for the distributions under consideration. In each case, 20,000 values of the test statistic are generated under the alternative hypothesis, $\mathcal{H}_1$, and the power is defined as the proportion of these values that exceed the associated critical value. These computations are performed for a range of sample sizes, and as shown in Fig. 6.1, the $\tilde{\lambda}$ test is consistent.

### 6.4.2 Bias and Risk Evaluations

In cases where we do not have an exact closed-form expression for the relative bias or risk of one of the estimators, these quantities are simulated as follows. If $\bar{\theta}_{(i)}$ is the $i$th simulated value of one of the estimators, $\bar{\theta}$ ( $= \tilde{\theta}, \hat{\theta}$, or $\theta^*$) of the parameter in question ($\theta$), then

$$\%\text{relBias}(\bar{\theta}; \theta) = 100\left[\frac{1}{\text{NREP}}\sum_{i=1}^{\text{NREP}} \bar{\theta}_{(i)} - \theta\right]/\theta \tag{6.23}$$

$$\%\text{relMSE}(\bar{\theta}; \theta) = 100\left[\frac{1}{\text{NREP}}\sum_{i=1}^{\text{NREP}} (\bar{\theta}_{(i)} - \theta)^2\right]/\theta^2 \tag{6.24}$$

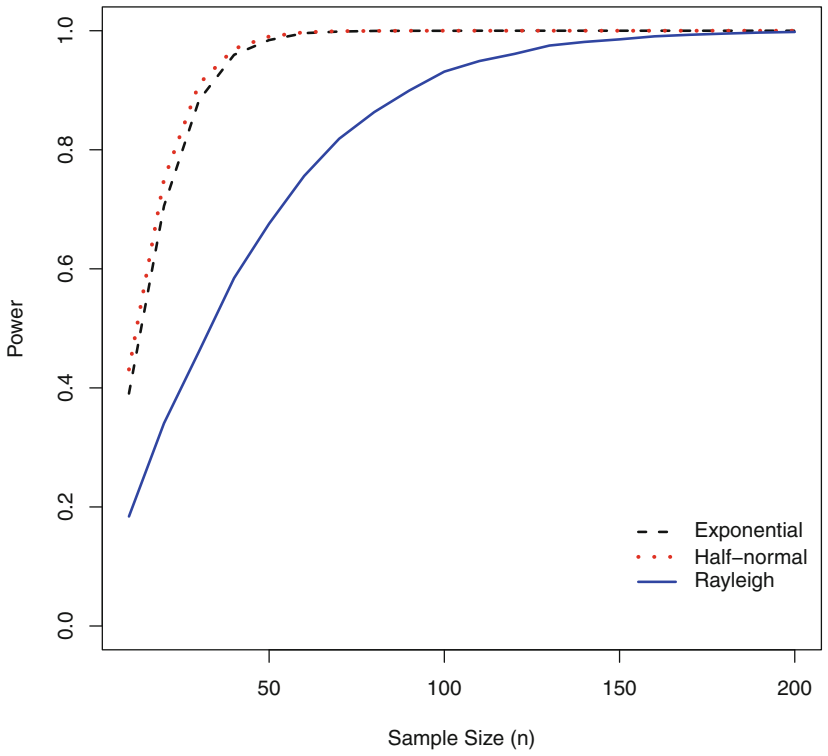



**Fig. 6.1** Powers of the $\tilde{\lambda}$ test when $\alpha = 5\%$

$$\%\text{relMAE}(\bar{\theta};\theta) = 100\left[\frac{1}{\text{NREP}}\sum_{i=1}^{\text{NREP}}|\bar{\theta}_{(i)}-\theta|\right]/\theta \tag{6.25}$$

$$\%\text{relR}_L(\bar{\theta};\theta) = 100\left[\frac{1}{\text{NREP}}\sum_{i=1}^{\text{NREP}}\exp(w(\bar{\theta}_{(i)}-\theta))-w(\bar{\theta}_{(i)}-\theta)-1\right]/\theta. \tag{6.26}$$

Recall that in this study, NREP = 20,000. This choice was justified by comparing the exact results provided in Sect. 6.3.2 with their simulated counterparts.

### *6.4.3 Optimal Critical Values*

In any preliminary-testing situation, the properties of the preliminary-test estimator depend on the underlying model and its parameter values as well as the sample size and values and the particular test that is employed to choose between the two

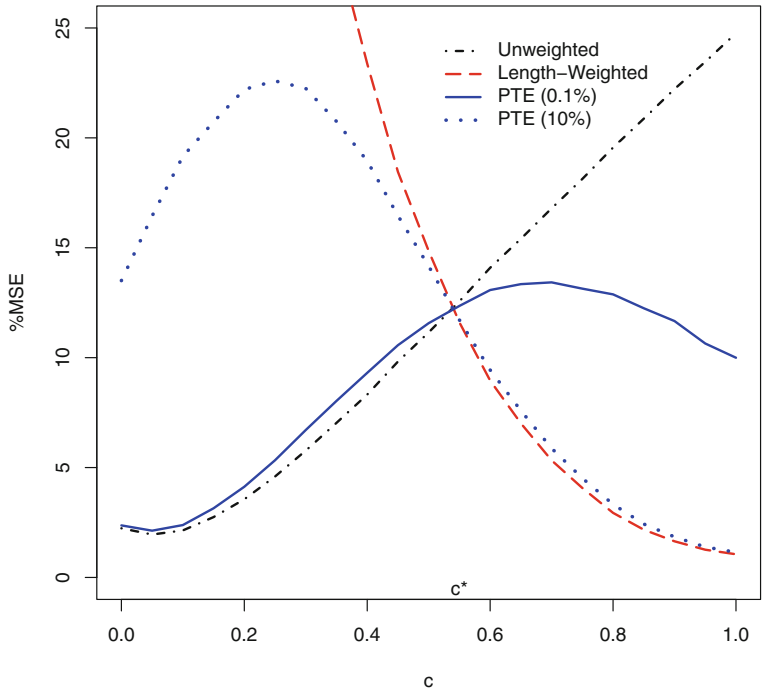



**Fig. 6.2** Exponential distribution: relative risks of the component and preliminary-test estimators under quadratic loss (%relMSE); $\alpha = 0.1\%$ and $\alpha = 10\%$ for the $\tilde{\lambda}$ test

'component' estimators. Not only are the properties of the latter test important, but so is the choice of significance level $\alpha$ (and hence critical value, $k$). This choice, together with the power of the test, determines the (random) weights that are assigned to the component estimators. In other words, it partially drives the random mixture that forms the preliminary-test estimator. The effect of changing the significance level is illustrated in Fig. 6.2, for the case of the exponential distribution and quadratic loss, when $n = 50$.

Figure 6.2 is constructed by allowing for the possibility of non-integer size-bias—that is, $c \in (0, 1)$. As $c$ varies from zero to one in value, the parameter ratio $(d/p)$ for the G-gamma distribution varies accordingly. In the case of the exponential distribution, for example, $c \in (0, 1)$ implies $(d/p) \in (1, 2)$. Examples of situations involving fractional biased sampling are discussed by Ducey (2009) and Ducey and Valentine (2009). The intersection of the 'unweighted' and 'length-biased' risk functions depends on the value of $n$ and occurs at $c = c^*$. The risk function for the preliminary-test estimator also passes through this point. This is because the latter estimator is constructed on the basis of null and alternative hypotheses that are 'point' hypotheses. We see that as $\alpha$ increases, the preliminary-test risks 'pivot' monotonically around this intersection point, rising to its left, and falling to its right.

This raises a natural question: 'What choice of significance level is best?' This question has been addressed by a number of authors for various other preliminary-testing problems. Several related criteria have been suggested for defining the 'optimal' choice of significance level or critical value. For example, see Sawa and Hiromatsu (1973), Toyoda and Wallace (1976), Brook (1976), Giles et al. (1992), and Kibria and Saleh (2006). We will consider the widely adopted criterion of 'mini-max regret', in the form suggested by Brook (1976).

For each value of $c$ in Fig. 6.2, consider the vertical 'distances' between the risk of the preliminary-test estimator and the *smaller* of the risks of the two component estimators. Here, the latter are the MLEs of $\beta$ for the exponential distribution, based on 'unweighted' and 'length-biased' sampling. We will label these distances $d_L(c;\alpha) = [\%\text{relMSE}(\beta^*) - \%\text{relMSE}(\tilde{\beta})]$, for $c < c^*$, and $d_U(c;\alpha) = [\%\text{relMSE}(\beta^*) - \%\text{relMSE}(\hat{\beta})]$, for $c \geq c^*$. These are termed the 'regret' of using the preliminary-test estimator instead of the component estimator that has lower (relative) risk for that value of $c$. These distances show the extent to which preliminary testing worsens the risk of the estimator, relative to the other two options.

The 'mini-max regret' criterion for obtaining an optimal value of $\alpha$, say $\alpha^*$, is to select the latter such that both $\max_{(c)}\{d_L(c;\alpha)\}$ and $\max_{(c)}\{d_U(c;\alpha)\}$ are minimized. Given the monotonic 'pivoting' of the preliminary-test risk function about the point $c^*$ noted above, this implies finding $\alpha^*$ (and hence $k^*$) such that $\max_{(c)}\{d_L(c;\alpha)\} = \max_{(c)}\{d_U(c;\alpha)\}$. Using $k^*$ as the critical value results in the best possible preliminary-testing strategy from the perspective of estimation risk.

## 6.5 Results

In this section, we present a representative set of results related to the sampling properties of the length-biased sampling preliminary-test estimator for each of the three distributions under study. These properties are compared, graphically, with those of the two component MLEs.

### *6.5.1 Relative Biases*

Although our primary focus is on the (relative) risks of the various estimators, under different loss functions, the relative biases of the estimators are also of interest. We have examined these in percentage terms, for various sample sizes, and we present the results for the illustrative case where a 5% significance level is chosen for the $\tilde{\lambda}$ preliminary test. These results appear in the top-left boxes of Figs. 6.3, 6.4, 6.5, 6.6, 6.7, and 6.8 and are ordered as follows.

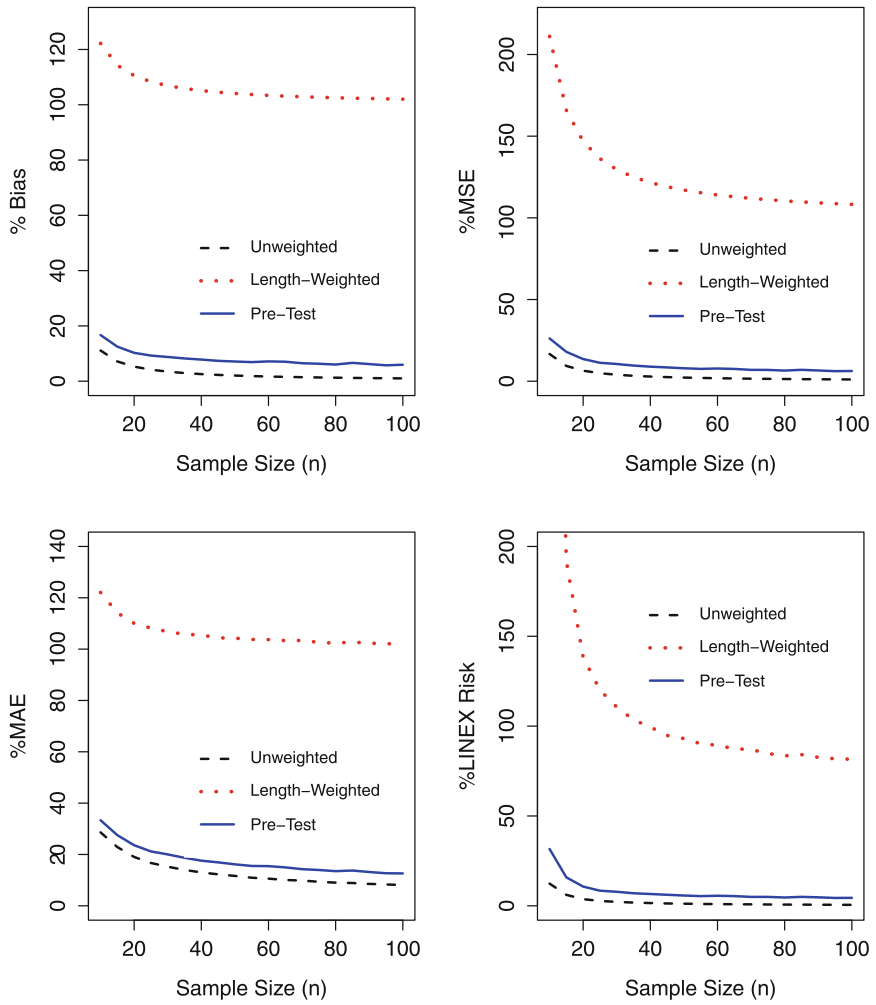



**Fig. 6.3** Exponential distribution: $\mathcal{H}_0$ true. Relative bias and relative risks of the component and preliminary-test estimators, in percentage terms. $\alpha = 5\%$. For the LINEX risk, $w = 2$ and $\beta = 1$

For the case of the exponential distribution, the exact relative biases of the component MLEs and the simulated relative bias of the PTE of the rate parameter, $\beta$, are shown as a function of the sample size, $n$ for the case where $\mathcal{H}_0$ is true in Fig. 6.3 and for the case where $\mathcal{H}_1$ is true in Fig. 6.8. The corresponding results for the half-normal and Rayleigh distributions appear in Figs. 6.5 and 6.6, as well as Figs. 6.7 and 6.8. (In the last of these figures, the biases of the component MLEs are also simulated—see the discussion associated with Table 6.1.)

In all of these figures, we see that the relative bias of the PTE (labelled 'pre-test' for brevity) always lies between the relative biases of the component MLEs.

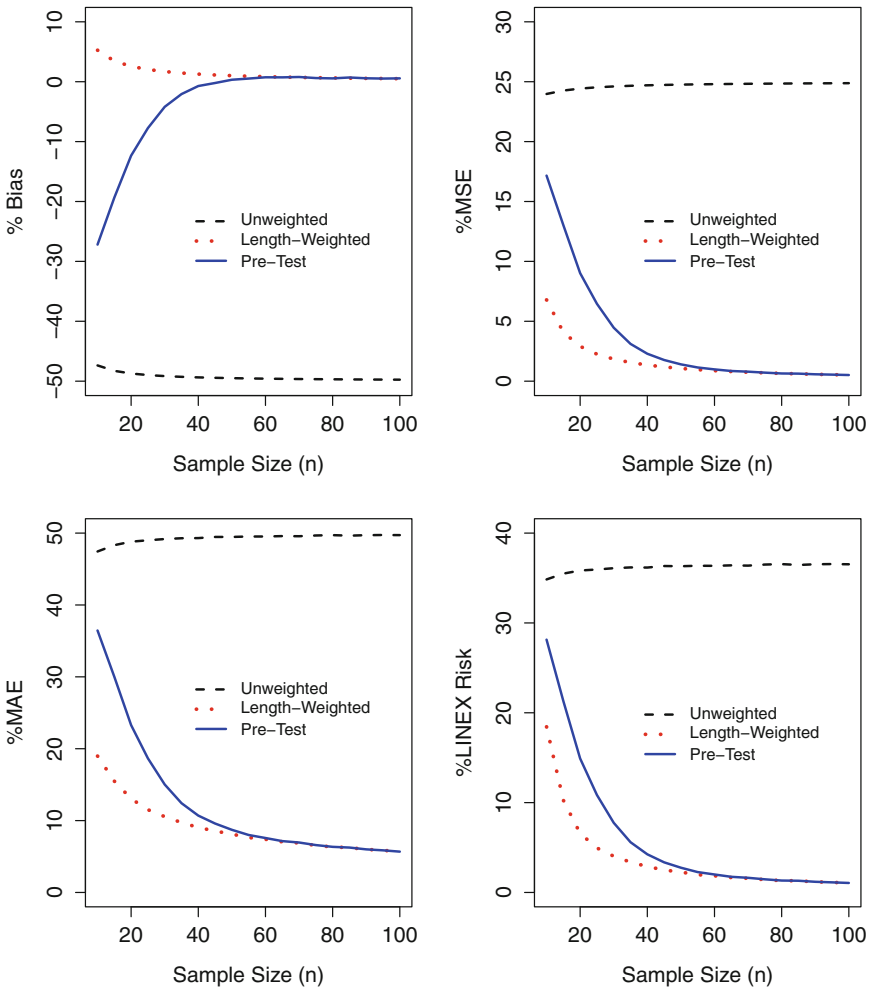



**Fig. 6.4** Exponential distribution: $\mathcal{H}_1$ true. Relative bias and relative risks of the component and preliminary-test estimators, in percentage terms. $\alpha = 5\%$. For the LINEX risk, $w = 2$ and $\beta = 1$

As expected, when $\mathcal{H}_0$ is true, the bias of the PTE is close to that of the MLE that assumes 'unweighted' sampling. Similarly, when $\mathcal{H}_1$ is true, it is close to the bias of the MLE that assumes 'length-weighted' sampling. The latter gap closes (and the bias approaches zero) as the sample size increases, even modestly. In addition, the absolute biases of the 'component' MLEs themselves can be as large as 50% in value (even when $n = 100$) if the hypothesis under which they are constructed is false. This reflects the inconsistency of these MLEs in such cases.

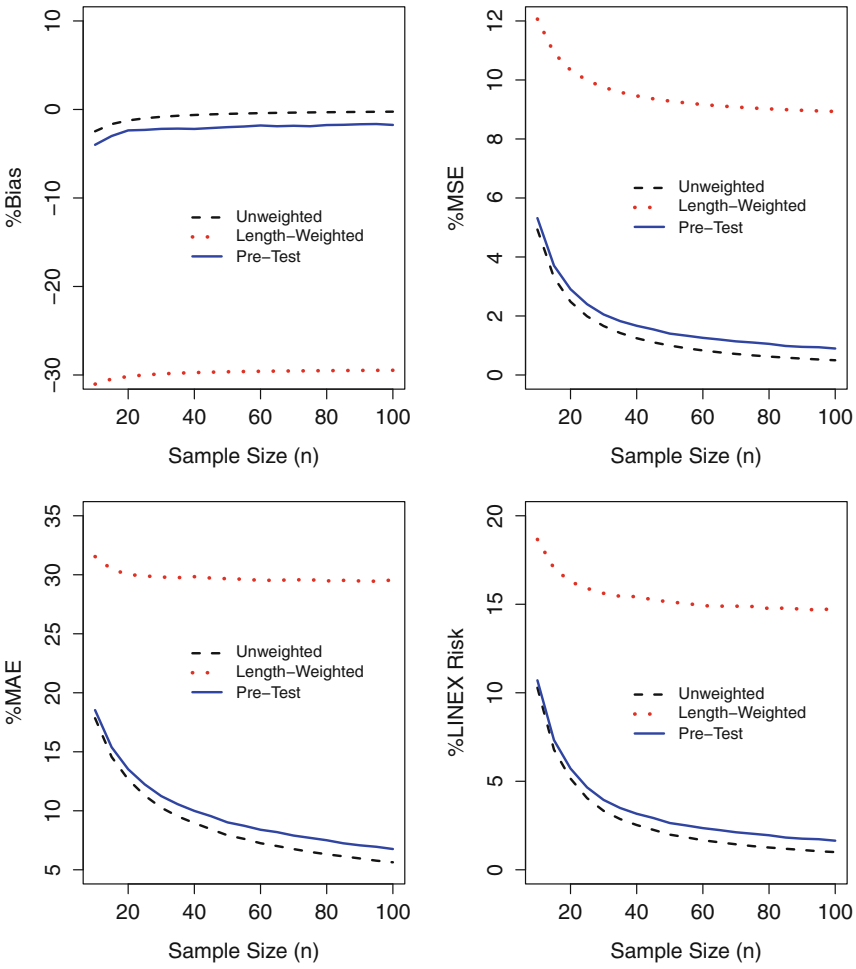



**Fig. 6.5** Half-normal distribution: $\mathcal{H}_0$ true. Relative bias and relative risks of the component and preliminary-test estimators, in percentage terms. $\alpha = 5\%$. For the LINEX risk, $w = 2$ and $\beta = 1$

### *6.5.2 Relative Risks*

The three (relative) risk functions for the 'unweighted' and 'size-biased' MLEs and for the associated PTEs also appear in Figs. 6.3, 6.4, 6.5, 6.6, 6.7, and 6.8, with the same ordering as for the relative biases. The relative LINEX risks of the PTEs depend on the value of the true parameter of each distribution, as well as on the choice of the asymmetry weight ($w$) for the LINEX loss function. The results for a single choice of each of these variables are illustrated in Figs. 6.2, 6.3, 6.4, 6.5, 6.6, and 6.7, but our overall simulation results show that these are representative choices.

We see that the risk function for the preliminary-test estimator lies between the risks of its two 'component estimators'. Interestingly, this implies that the

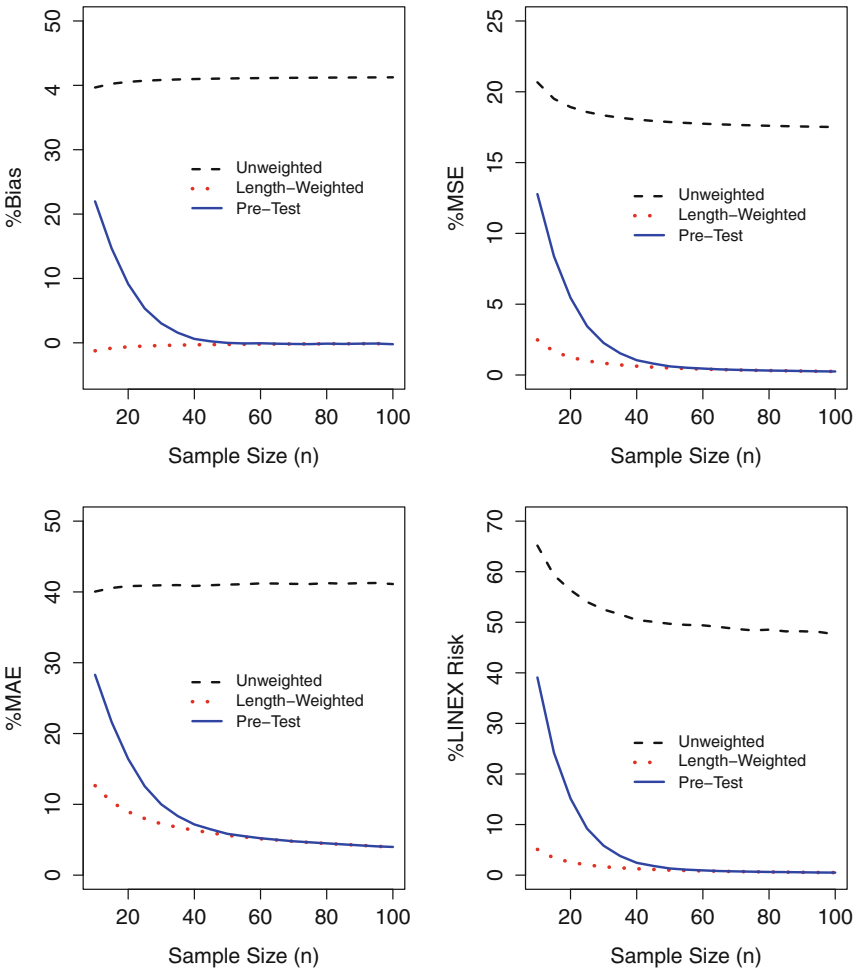



**Fig. 6.6** Half-normal distribution: $\mathcal{H}_1$ true. Relative bias and relative risks of the component and preliminary-test estimators, in percentage terms. $\alpha = 5\%$. For the LINEX risk, $w = 2$ and $\beta = 1$

preliminary-test estimator can never be the worst of the three estimators in terms of relative risk. This stands in contrast to many other results in the preliminary-testing literature. It arises because both the null and alternative hypotheses are point hypotheses. This result is discussed further in the next subsection. The pattern of behaviour of the risk of the PTE relative to the risks of the component MLEs follows that described for the biases in the last subsection. Again, the inconsistency of the component MLEs when the hypotheses on which they are based is false is clear from the %MSE results in these tables.

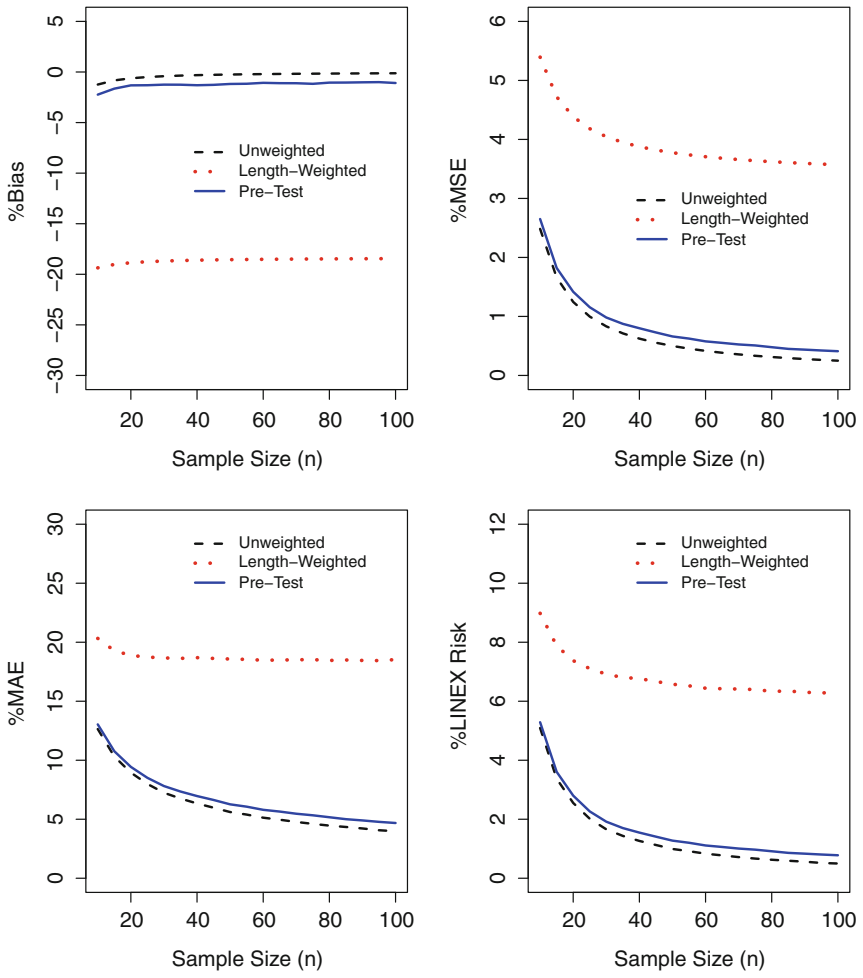



**Fig. 6.7** Rayleigh distribution: $\mathcal{H}_0$ true. Relative bias and relative risks of the component and preliminary-test estimators, in percentage terms. $\alpha = 5\%$. For the LINEX risk, $w = 2$ and $\beta = 1$

### *6.5.3 Optimal Critical Values*

As found by the various authors referenced in Sect. 6.4.3, for a given risk criterion and underlying distribution, the mini-max regret approach for determining $\alpha^*$ results in an almost constant optimal critical value, $k^*$, regardless of the sample size. Moreover, it is almost the same for the %relMSE and %relMAE risks. Of course, this implies that the optimal *significance level* varies greatly with the sample size, and it can be vastly different from the significance levels (such as 5% or 10%) that would commonly be used. This is shown in Table 6.4, for the three distributions under study and for the different risk functions being considered. This constancy

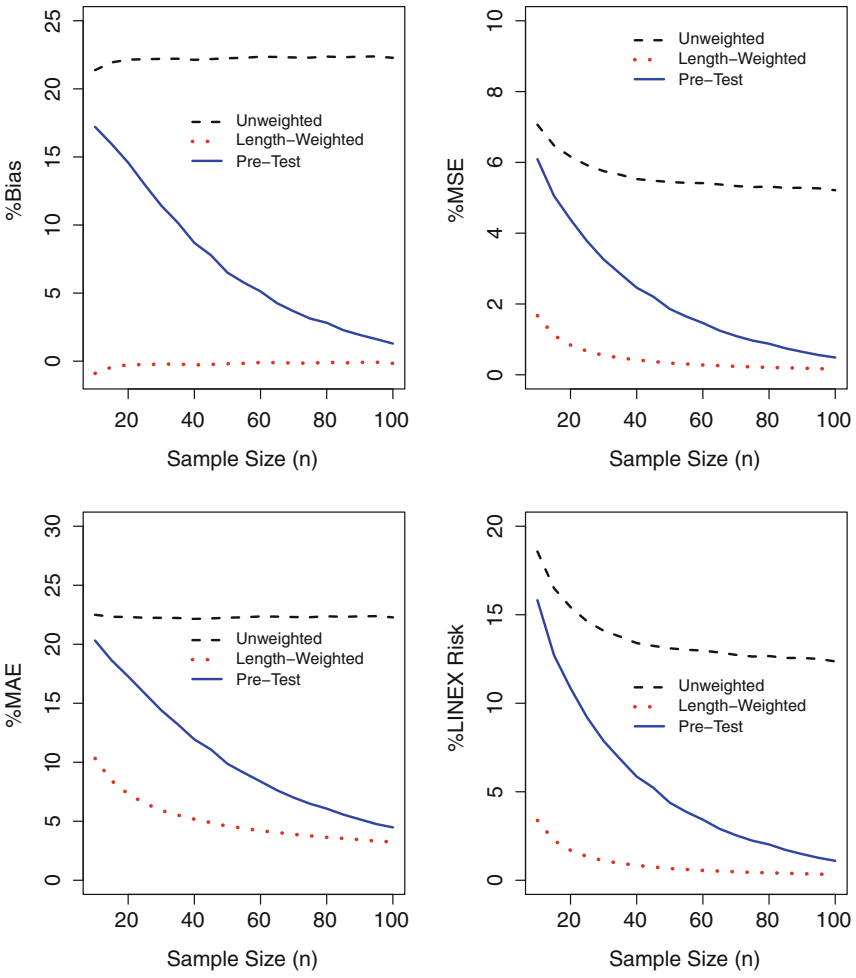



**Fig. 6.8** Rayleigh distribution: $\mathcal{H}_1$ true. Relative bias and relative risks of the component and preliminary-test estimators, in percentage terms. $\alpha = 5\%$. For the LINEX risk, $w = 2$ and $\beta = 1$

in the critical value across sample sizes is, of course, of considerable help to practitioners. For example, in the case of the half-normal distribution, choosing a critical value of approximately 0.8 to 0.85 will optimize the performance of the preliminary-test estimator for most sample sizes, regardless of the choice of risk function. A choice of $k^* = 0.9$ has the same effect for the Rayleigh distribution.

Figures 6.9, 6.10, and 6.11 illustrate the relative risks of the preliminary-test estimators for each distribution and each loss function, when the choice of critical value for the $\tilde{\lambda}$ preliminary test is optimized in terms of mini-max regret. For comparison, the corresponding risks when $\alpha = 5\%$ are also shown. The results shown in these figures are $n = 50$ or $n = 25$; and the LINEX risks are for $w = 2$, and $\beta = 1$ or $\sigma = 1$, for illustrative purposes. These choices are fully representative

**Table 6.4** Optimal critical values

| | Exponential | | Half-normal | | Rayleigh | |
|---|---|---|---|---|---|---|
| $n$ | $\alpha^*$ | $k^*$ | $\alpha^*$ | $k^*$ | $\alpha^*$ | $k^*$ |
| %relMSE-risk | | | | | | |
| 15 | 5.0% | 0.77 | 22.1% | 0.82 | 34.5% | 0.90 |
| 25 | 3.2% | 0.74 | 13.3% | 0.81 | 27.4% | 0.90 |
| 50 | 0.90% | 0.72 | 4.6% | 0.81 | 16.7% | 0.90 |
| 100 | 0.07% | 0.70 | 0.62% | 0.81 | 7.7% | 0.90 |
| 250 | 0.0002% | 0.69 | 0.006% | 0.80 | 1.02% | 0.89 |
| %relMAE-risk | | | | | | |
| 15 | 8.4% | 0.75 | 20.2% | 0.83 | 33.0% | 0.91 |
| 25 | 5.0% | 0.72 | 11.8% | 0.82 | 25.9% | 0.90 |
| 50 | 1.3% | 0.71 | 3.9% | 0.82 | 16.0% | 0.90 |
| 100 | 0.11% | 0.70 | 0.57% | 0.81 | 7.1% | 0.90 |
| 250 | 0.0002% | 0.69 | 0.006% | 0.81 | 1.1% | 0.89 |
| %relLINEX-risk ($w = 1$) | | | | | | |
| 15 | 2.4% | 0.80 | 26.5% | 0.80 | 38.5% | 0.89 |
| 25 | 2.0% | 0.76 | 16.2% | 0.80 | 30.2% | 0.89 |
| 50 | 0.56% | 0.72 | 5.8% | 0.80 | 18.9% | 0.89 |
| 100 | 0.04% | 0.71 | 0.95% | 0.80 | 8.5% | 0.89 |
| 250 | 0.00009% | 0.70 | 0.01% | 0.80 | 1.3% | 0.89 |
| %relLINEX-risk ($w = -1$) | | | | | | |
| 15 | 7.7% | 0.75 | 18.5% | 0.83 | 30.7% | 0.91 |
| 25 | 4.8% | 0.72 | 10.8% | 0.83 | 24.8% | 0.90 |
| 50 | 1.3% | 0.71 | 3.7% | 0.82 | 15.6% | 0.90 |
| 100 | 0.14% | 0.70 | 0.51% | 0.81 | 7.0% | 0.90 |
| 250 | 0.0002% | 0.69 | 0.003% | 0.81 | 0.89% | 0.90 |
| %relLINEX-risk ($w = 2$) | | | | | | |
| 15 | 1.1% | 0.83 | 31.7% | 0.77 | 42.6% | 0.89 |
| 25 | 1.1% | 0.77 | 19.3% | 0.79 | 33.2% | 0.89 |
| 50 | 0.35% | 0.73 | 7.3% | 0.79 | 20.3% | 0.89 |
| 100 | 0.023% | 0.72 | 1.2% | 0.79 | 9.3% | 0.89 |
| 250 | 0.00001% | 0.71 | 0.02% | 0.79 | 1.5% | 0.89 |
| %relLINEX-risk ($w = -2$) | | | | | | |
| 15 | 11.1% | 0.73 | 15.3% | 0.85 | 27.2% | 0.92 |
| 25 | 6.7% | 0.71 | 9.1% | 0.84 | 22.3% | 0.91 |
| 50 | 2.1% | 0.70 | 3.0% | 0.82 | 13.8% | 0.90 |
| 100 | 0.26% | 0.69 | 0.37% | 0.82 | 6.3% | 0.89 |
| 250 | 0.0008% | 0.68 | 0.002% | 0.82 | 0.78% | 0.90 |

of other situations that have been explored, and these figures clearly demonstrate the importance of using a (relatively constant) critical value, $k^*$, as provided in Table 6.4. Conventional choices of the significance level for the preliminary test can lead to a highly suboptimal risk for the PTE, regardless of the underlying distribution, the loss function, or the sample size.

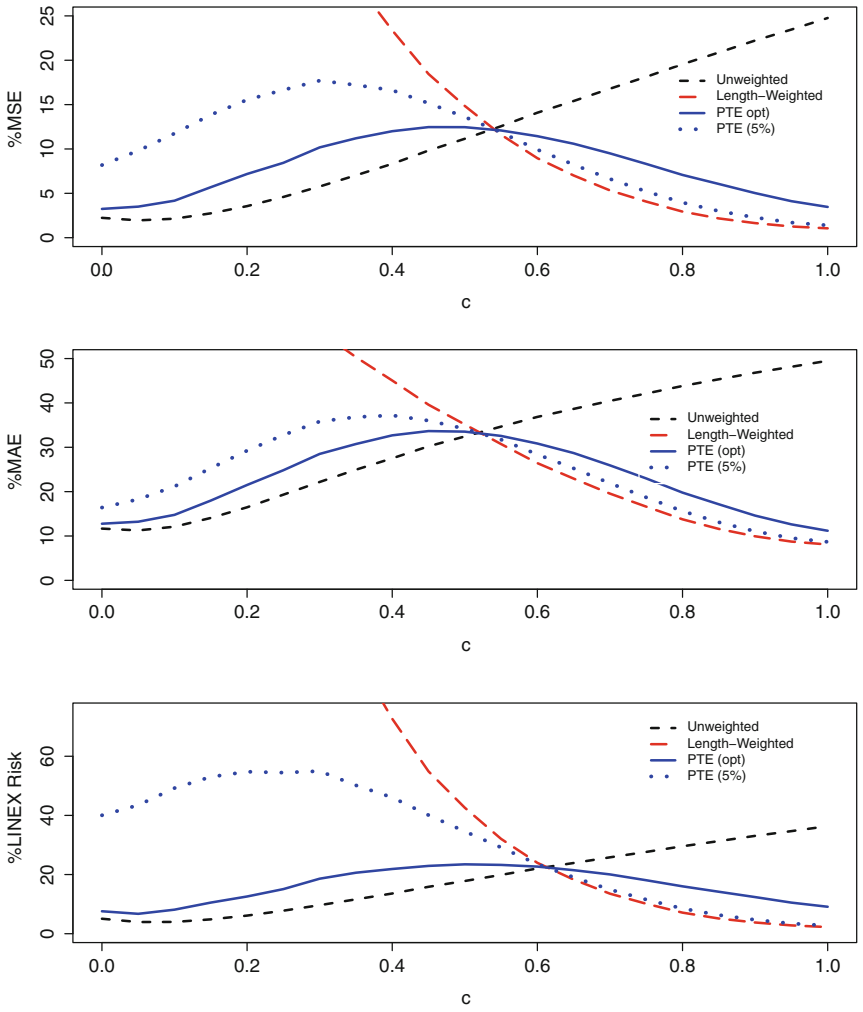



**Fig. 6.9** Exponential distribution: relative risks of the component and preliminary-test estimators, in percentage terms. $n = 50$; $\alpha = 5\%$ or $\alpha = \alpha^*$ (= opt). For the LINEX risk, $w = 2$ and $\beta = 1$

## 6.6 Concluding Remarks

Size-biased sampling is a statistical phenomenon that arises in a wide range of disciplines. If it has occurred, but this is ignored, then there can be detrimental consequences for the properties of estimators and tests based on the sampled data. One practical issue is that the researcher may not know whether such sampling has been used. If this is the case, then there is a small but important literature that provides tests for the hypothesis that simple random sampling has been used, *versus*

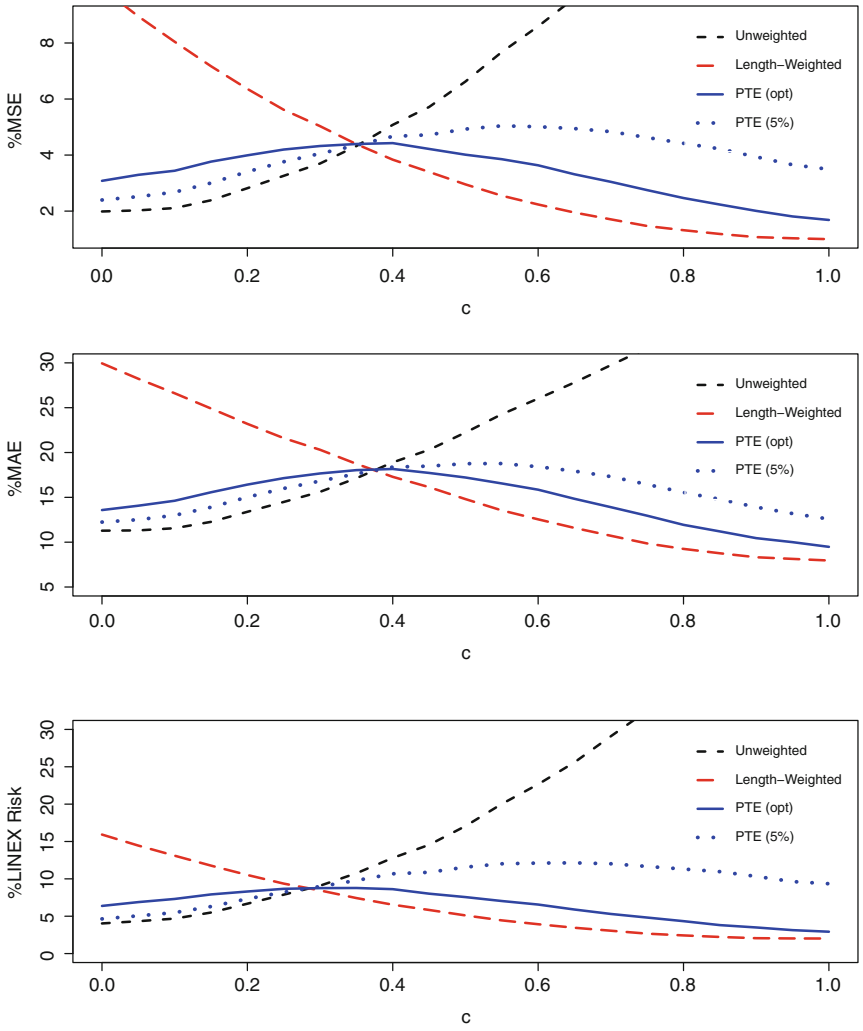



**Fig. 6.10** Half-normal distribution: relative risks of the component and preliminary-test estimators, in percentage terms. $n = 25$; $\alpha = 5\%$ or $\alpha = \alpha^*$ (= opt). For the LINEX risk, $w = 2$ and $\sigma = 1$

the alternative hypothesis that length-biased (or 'weighted') sampling of a fixed order is in place. These tests have asymptotic (large-sample) justification, but there is limited evidence about their performances in small samples.

If a researcher uses such a test to determine which distribution will form the basis of subsequent inferential procedures, then we have an example of 'preliminary testing'. Although the literature on preliminary-test inference (both estimation and testing) is substantial, it seems that to date it has not included the example of testing

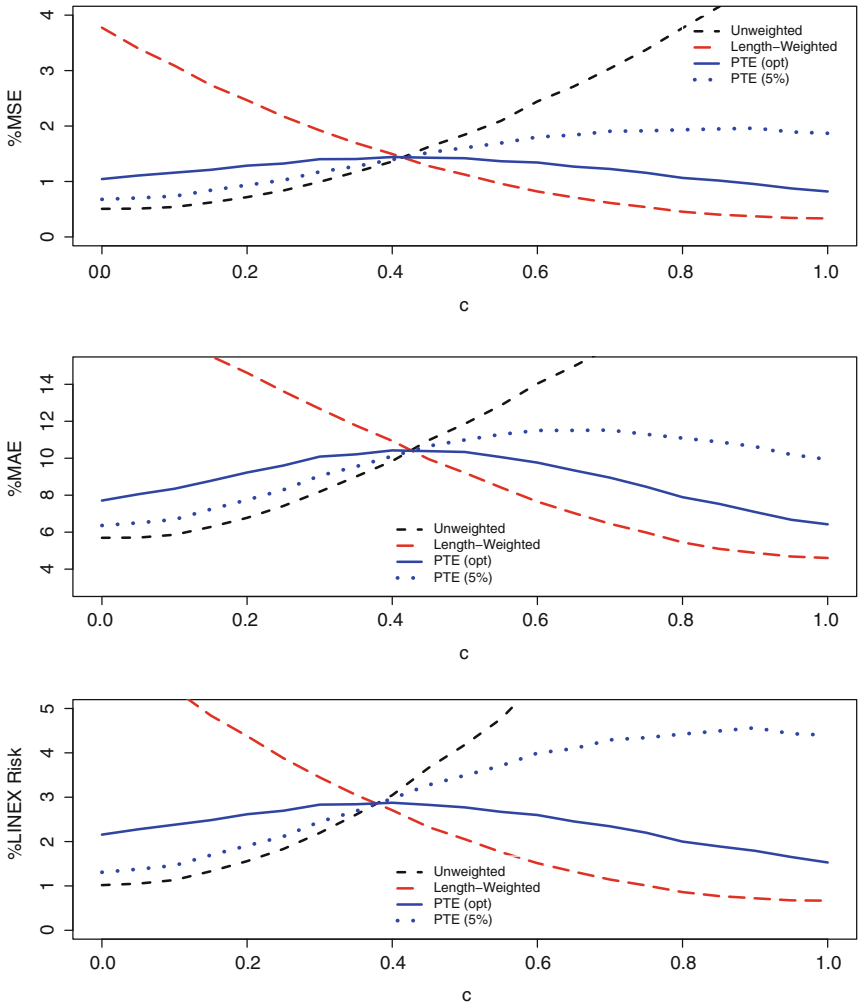



**Fig. 6.11** Rayleigh distribution: relative risks of the component and preliminary-test estimators, in percentage terms. $n = 50$; $\alpha = 5\%$ or $\alpha = \alpha^*$ (= opt). For the LINEX risk, $w = 2$ and $\sigma = 1$

for size-biased sampling. In this chapter, we have attempted to address this issue by exploring the sampling properties of preliminary-test estimators under various loss structures, for three simple distributions that have been used in the biased-sampling literature—the exponential, half-normal, and Rayleigh distributions.

Our results indicate that the relative bias and relative risks of these preliminary-test estimators lie between those of the two 'component' MLEs—those obtained from the 'unweighted' length-biased distributions. Accordingly, if we consider three possible strategies—always assume that $\mathcal{H}_0$ is true, always assume that $\mathcal{H}_1$ is true,

or apply a preliminary test using the $\tilde{\lambda}$ statistic—the preliminary-test strategy can never be the best of the three strategies, but it can also never be the worst.

An important finding is that when we seek to optimize the preliminary-testing strategy by choosing the critical value for the $\tilde{\lambda}$ test that satisfies the mini-max regret criterion, then this critical value should be constant across different sample sizes, for a particular distribution. It should not be varied to comply with a conventional significance level such as 5% or 10%. This finding is consistent with the corresponding preliminary-test literature for the linear regression model.

The distributions that have been studied in this chapter are specific members of the generalized gamma family, and there are several other members of this family that are commonly used in size-biased sampling. These include, for instance, the Weibull, gamma, and Nakagami distributions. Each of these distributions has more than one parameter, and the associated maximum likelihood estimators cannot be written in closed form, which complicates the evaluation of the corresponding preliminary-test estimators. However, exploring preliminary testing in the context of these distributions would be an important avenue for further research. Other extensions of the present study could incorporate additional loss functions. Each of those considered in this chapter is unbounded, and it would be interesting to undertake a similar investigation based on the 'reflected normal' loss function (Spiring, 1993; Giles, 2002), and the bounded LINEX (BLINEX) loss function (Wen & Levy, 2001), for example. Moreover, this chapter dealt solely with length-biased sampling. Higher-order size-biased sampling, such as area-biased , is important in several fields such as forestry (e.g. Gove, 2003b) and preliminary testing in that context would be worthy of study.

**Acknowledgments** I am grateful to the referees and editors for their helpful comments and suggestions that led to improvements in the chapter.

## References

Akman, O., Gamage, J., Jannot, J., Juliano, S., Thurman, A., & Whitman, D. (2007). A simple test for detection of length-biased sampling. *Journal of Biostatistics, 1*, 189–195.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics, 15*, 190–204.

Bancroft, T. A., & Han, C.-P. (1977). Inference based on conditional specification: A note and bibliography. *International Statistical Review, 45*, 117–127.

Bashir, S., & Rasul, M. (2018). Record values from size-biased half normal distribution: Properties and recurrence relations for the single and product moments. *American Journal of Computation, Communication and Control, 5*, 1–6.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325–335.

Blumenthal, S. (1967). Proportional sampling in life length studies. *Technometrics, 9*, 205–218.

Brook, R. J. (1976). On the use of a regret function to set significance points in prior tests of estimation. *Journal of the American Statistical Association, 71*, 126–131.

Cohen, A. (1965). Estimates of the linear combinations of parameters in the mean vector of a multivariate distribution. *Annals of Mathematical Statistics, 36*, 299–304.

Cox, D. R. (1969). Some sampling problems in technology. In N. L. Johnson & H. Smith (Eds.), *New Developments in Survey Sampling*, Wiley.

Das, K. K., & Roy, T. D. (2011a). On some length-biased weighted Weibull distributions. *Advances in Applied Science Research, 2*, 465–475.

Das, K. K., & Roy, T. D. (2011b). Applicability of length-biased weighted generalized Rayleigh distribution. *Advances in Applied Science Research, 2*, 320–327.

Ducey, M. J. (2009). Sampling trees with probability nearly proportional to biomass. *Forest Ecology and Management, 258*, 2110–2116.

Ducey, M. J., & Gove, J. H. (2015). Size-biased distributions in the generalized beta distribution family, with applications to forestry. *Forestry, 88*, 143–151.

Ducey, M. J., & Valentine, H. T. (2009). Direct sampling for stand density index. *Western Journal of Applied Forestry, 23*, 78–82.

Economou, P., & Tzavelas, G. (2013). Sample tests for detection of size-biased sampling mechanism. *Communications in Statistics—Theory and Methods, 42*, 3280–3295.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics, 6*, 13–25.

Giles, D. E. (2002). Preliminary test and Bayes estimation of a location parameter under 'reflected normal' loss. In A. Ullah, A. Wan, & A. Chaturvedi (Eds.), *Handbook of Applied Econometrics and Statistical Inference* (pp. 287–303). Marcel Dekker.

Giles, D. E. (2021). Improved maximum likelihood estimation for the Weibull distribution under length-biased sampling. *Journal of Quantitative Economics, 19*, 59–77.

Giles, J. A., & Giles, D. E. A. (1993). Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys, 7*, 145–197.

Giles, D. E. A., Lieberman, O., & Giles, J. A. (1992). The optimal size of a preliminary-test of linear restrictions in a mis-specified regression model. *Journal of the American Statistical Association, 87*, 1153–1157.

Gove, J. H. (2003a). Moment and maximum likelihood estimators for Weibull distributions under length- and area-biased sampling. *Environmental and Ecological Statistics, 10*, 455–467.

Gove, J. H. (2003b). Estimation and applications of size-biased distributions in forestry. In A. Amaro, D. Reed, & P. Soares (Eds.), *Modeling Forest Systems* (pp. 201–212). CABI Publishing.

Han, C.-P., Rao, C. V., & Ravichandran, J. (1988). Inference based on conditional specification: A second bibliography. *Communications in Statistics—Theory and Methods, 17*, 1945–1964.

Jabeen, S., & Jan, T. R. (2015). Information measures of size biased generalized gamma distribution. *International Journal of Scientific Engineering and Applied Science, 1*, 516–530.

Judge, G. G., & Bock, M. E. (1978). *The Statistical Implications of Pre-test and Stein Rule Estimators in Econometrics*. North-Holland.

Kibria, B. M. G., & Saleh, A. K. Md. E. (2006). Optimal critical value for pre-test estimator. *Communications in Statistics—Simulation and Computation, 35*, 309–319.

Mir, K. A., Ahmed, A., & Reshi, J. A. (2013). On size biased exponential distribution. *Journal of Modern Mathematics and Statistics, 7*, 21–25.

Mudasir, S., & Ahmad, S. P. (2018). Characterization and estimation of the length biased Nakagami distribution. *Pakistan Journal of Statistics and Operation Research, 3*, 697–715.

Navarro, J., Ruiz, J., & del Aguila, Y. (2003). How to detect biased samples? *Biometrics Journal, 45*, 91–112.

Patil, G. P., & Ord, J. K. (1976). On size-biased sampling and related form-invariant weighted distributions. *Sankhyā: The Indian Journal of Statistics, Series B, 38*, 48–61.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Saldanha, M. H. J., & Suzuki, A. K. (2022). Package 'ggamma'—The generalized gamma probability distribution. Retrieved from https://ftp.fau.de/cran/web/packages/ggamma/ggamma.pdf

Saleh, A. K. Md. E. (2014). *Theory of Preliminary Test and Stein-type Estimation with Applications*. Wiley.

Sawa, T., & Hiromatsu, T. (1973). Minimax regret significance points for a preliminary test in regression analysis. *Econometrica, 41*, 1093–1101.

Scheaffer, R. L. (1972). Size-biased sampling. *Technometrics, 14*, 635–644.

Spiring, F. A. (1993). The reflected normal loss function. *Canadian Journal of Statistics, 21*, 321–330.

Stacy, E. W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics, 33*, 1187–1192.

Toyoda, T., & Wallace, T. D. (1976). Optimal critical values for pre-testing in regression. *Econometrica, 44*, 365–375.

Varian, H. R. (1975). A Bayesian approach to real estate assessment. In L. J. Savage, S. E. Feinberg, & A. Zellner (Eds.), *Studies in Bayesian Econometrics and Statistics: In Honor of L. J. Savage* (pp. 195–208). North-Holland.

Wen, D., & Levy, M. S. (2001). BLINEX: A bounded asymmetric loss function with application to Bayesian estimation. *Communications in Statistics - Theory and Methods, 30*, 147–153.

Wicksell, S. D. (1925). The corpuscle problem: A mathematical study of a biometric problem. *Biometrika, 17*, 84–99.

Xiao, L., & Giles, D. E. (2014). Bias reduction for the maximum likelihood estimator of the generalized Rayleigh family of distributions. *Communications in Statistics—Theory and Methods, 43*, 1778–1792.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association, 81*, 446–451.

# Chapter 7
# Study the Performance of New Shrinkage Estimators Under the Balanced Loss Function




**Abdenour Hamdaoui, Abdelkader Benkhaled, and Amani Alahmadi**

## 7.1 Introduction

When dealing with natural phenomena, we encounter numerous variables, and to effectively analyze this type of data, we must utilize multivariate distributions. This method has become a very important technique for data modeling and offers valuable techniques for integrating data from different sources. This ability to represent relationships between variables is a fundamental part of various statistical analyses. In fact, estimating the mean vector of this distribution have a great deal of attention. The object of the present study is to focus mainly on the estimation of the mean parameter of a multivariate normal distribution using Stein-type shrinkage estimators.

Shrinkage estimators are a set of techniques designed to perform better by reducing the variance of an estimator. In standard estimation problems, we may use the sample mean, sample variance, or maximum likelihood estimates (MLE),

A. Hamdaoui (✉)
Department of Mathematics, University of Science and Technology, Mohamed Boudiaf, Oran, Algeria

University of Science and Technology, Mohamed Boudiaf, El Mnaouar, Oran, Algeria
e-mail: abdenour.hamdaoui@univ-usto.dz

A. Benkhaled
Department of Biology, University of Mascara, Laboratory of Stochastic Models, Statistics and Applications, University Tahar Moulay of Saïda, Mascara, Algeria
e-mail: benkhaled08@yahoo.fr

A. Alahmadi
Department of Mathematics, College of Science and Humanities in Ad Dawadmi, Shaqra University, Shaqra, Saudi Arabia
e-mail: aalahmadi@su.edu.sa

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_7

but these estimators may be suboptimal, especially in the case of small samples, high-dimensional data, or noisy settings. For high-dimensional problems, shrinkage can help avoid over-fitting by bringing extreme parameter estimates toward zero.

This field of research is started by Stein (1956) and James and Stein (1961), which showed in this time that the MLE of the population mean is inadmissible when the dimension of the parameter space is greater than or equal to three under mean square error loss. Given the importance of this research and its great relationship to natural phenomena, since then, many researchers have been immersed in this way; we cite, for example, Takahasi and Wakimoto (1968), Alam (1973), Strawderman (1973), Efron and Morris (1976), Arnold (1981), Stein (1981), Berger (1982), Sun (1995), Ahemd (2000), Hamdaoui and Benmansour (2015), and Nimet et al. (2017). Many details on parametric estimation using Stein-type estimators have been shown in Saleh (2006).

Recent studies in the context of shrinkage estimation include Norouzirad and Arashi (2019), Nimet and Selahattin (2019), Yüzbaşı et al. (2020), Benkhaled and Hamdaoui (2022), and Norouzirad et al. (2023). Hamdaoui et al. (2021b) considered the model $Z \sim N_q\left(\mu, I_q\right)$, and the authors are interested to estimate the unknown parameter $\mu$ by shrinkage estimators deduced from the MLE and the James-Stein estimator (JSE). They studied the minimaxity property of the estimators:

$$\delta_{a,r} = \left(1 - \frac{a}{\|Z\|^r}\right) Z,$$

where $2 \leq r < (q+2)/2$ and the real constant $a$ may depend on $q$. Then, they considered the estimators of the form

$$\delta_{b,r} = \delta_{\mathrm{JS}} + \frac{b\, Z}{\|Z\|^r}$$

with $2 \leq r < (q+2)/2$ and the real constant $b$ may depend on $q$. They showed that under the balanced loss function (BLF), these estimators dominate the JSE under some conditions on parameter $b$.

In this work, we deal with the model $Z \sim \mathcal{N}_q\left(\nu, I_q\right)$, and our goal is to estimate the unknown parameter $\nu$ by shrinkage estimators constructed from the MLE and the positive part of James-Stein estimator (PPJSE). To measure the quality of the suggested estimators, we use the risk function relatively to the BLF defined as, for any estimator $\delta$ of $\nu$,

$$\ell_\omega(\delta, \nu) = \omega\|\delta - \delta_0\|^2 + (1-\omega)\|\delta - \nu\|^2, \tag{7.1}$$

where $0 \leq \omega < 1$ and $\delta_0$ is the target estimator (in our study, we consider $\delta_0$ to be the MLE).

Two criteria are taken into account in the balance loss function: precision of estimation and goodness of fit. Therefore, the loss may be obtained considering the

arbitrary weight $\omega$ for each criterion. If we are interested in precision, the value of $\omega$ must be near to 0, and if we are interested in goodness of fit, the value of $\omega$ must be near to 1. When $\omega = 0$, we obtain the quadratic loss function. For the use of this loss, we can refer to Zellner (1994), Karamikabir et al. (2018), and Hamdaoui et al. (2021b).

The risk function associate to the BLF is defined by

$$\mathcal{R}_{\omega}(\delta, \nu) = \mathrm{E}[\ell_{\omega}(\delta, \nu)]. \tag{7.2}$$

We divide this manuscript as follows. In Sect. 7.2, we give preliminaries in which we recall some essential published results. In the first, we study the minimaxity of the general class of shrinkage estimator $\delta_g(Z) = Z + g(Z)$ with $g$ as a measurable differentiable function, namely, we give a sufficient condition on the function $g(\cdot)$ for that the estimator $\delta_g(Z)$ beat the MLE. Then, we treat the minimaxity of the Baranchik-type estimators, i.e., for $g(Z) = -\left(\phi(\|Z\|^2) / \|Z\|^2\right) Z$. Baranchik-type estimators are a specific class of shrinkage estimators that focus on enhancing the estimation of certain parameters in situations where classical estimators (such as sample mean or sample covariance) may not perform, by shrinking this estimators to a more stable value in order to moderate high variability or outlier effects. Afterward, we present the famous class of shrinkage estimators:

$$\delta_{h,\mathrm{JS}}(Z) = \delta_{\mathrm{JS}} + \frac{h(\|Z\|^2)}{\|Z\|^2} Z,$$

which do not only minimax but they are also dominate the JSE $:= \delta_{\mathrm{JS}}$ and give a simple condition on the function $h(\cdot)$ for that these estimators improve the JSE. We also show that the PPJSE is included in this class of estimators. In Sect. 7.3, we establish the main results in which we adopt a similar idea used in Hamdaoui et al. (2021b), but in this time, the aims is to improve the PPJSE, instead of the MLE and the JSE, and this leads us to more difficult calculations of risk functions. We construct a new family of estimators derived from the MLE and the PPJSE, and then we extract to this family the estimators that ameliorate the PPJSE under the BLF given in Eq. (7.1). Section 7.4 highlights a simulation study that compares the performance of the proposed approach with the PPJSE and JSE. The closing Sect. 7.5 crowns the whole work, wraps the conclusion, and provides new perspectives for future research.

## 7.2 Preliminaries

Let the model $Z \sim \mathcal{N}_q\left(\nu, I_q\right)$, and the aim is to estimate the unknown parameter $\nu = (\nu_1, \ldots, \nu_q) \in \mathbb{R}^q$ relatively to the BLF $\ell_{\omega}$ defined in Eq. (7.1). More precisely, every estimator $\delta = (\delta_1, \ldots, \delta_q)$ is evaluated through the risk function given in Eq. (7.2).

According to Hamdaoui et al. (2021a), the MLE of the unknown parameter $\nu$ is $\delta_0 = Z$, and its risk function is equal to $(1-\omega)q$. The MLE is minimax and inadmissible for $q \geq 3$; therefore, each estimator $\delta$ that dominates $\delta_0$ (i.e., $\mathcal{R}_\omega(\delta, \nu) \leq \mathcal{R}_\omega(Z, \nu)$ ) is minimax.

Next, we recall two famous classes of shrinkage estimators of the multivariate normal mean (MNM), and present the performance of these estimators under the BLF defined in Eq. (7.1).

### 7.2.1 The Minimax Shrinkage Estimators

Now, consider the large family of estimators derived from the MLE and defined by

$$\delta_g(Z) = Z + g(Z), \tag{7.3}$$

where the measurable function $g(Z)$ is differentiable.

**Proposition 7.1** *Relatively to the BLF $\ell_\omega$, the difference in risk between estimators $\delta_g(Z)$ and $\delta_0$ is expressed as*

$$\Delta_{\omega,g} = \mathrm{E}[\|g(Z)\|^2] + 2(1-\omega)\,\mathrm{E}[div\, g(Z)],$$

*where div $g(Z)$ denotes the divergence of the function $g(Z)$.*

***Proof*** Using the risk function associated with the BLF $\ell_\omega$, the difference in risk is equal to

$$\begin{aligned}
\Delta_{\omega,g} &= \mathcal{R}_\omega(\delta_g(Z), \nu) - \mathcal{R}_\omega(Z, \nu) \\
&= \omega\,\mathrm{E}\left[\|\delta_g(Z) - Z\|^2\right] + (1-\omega)\,\mathrm{E}\left[\|\delta_g(Z) - \nu\|^2\right] - (1-\omega)\,\mathrm{E}\left[\|Z - \nu\|^2\right] \\
&= \omega\,\mathrm{E}\left[\|g(Z)\|^2\right] + (1-\omega)\,\mathrm{E}\left[\|Z + g(Z) - \nu\|^2\right] - (1-\omega)\,\mathrm{E}\left[\|Z - \nu\|^2\right] \\
&= \omega\,\mathrm{E}\left[\|g(Z)\|^2\right] \\
&\quad +(1-\omega)\left(\mathrm{E}\left[\|g(Z)\|^2\right] + \mathrm{E}\left[\|Z - \nu\|^2\right] + 2\,\mathrm{E}\left[\,\langle Z - \nu, g(Z)\rangle\,\right]\right) \\
&\quad -(1-\omega)\,\mathrm{E}[\|Z - \nu\|^2] \\
&= \mathrm{E}\left[\|g(Z)\|^2\right] + 2(1-\omega)\,\mathrm{E}\left[\,\langle Z - \nu, g(Z)\rangle\,\right] \\
&= \mathrm{E}\left[\|g(Z)\|^2\right] + 2(1-\omega)\,\mathrm{E}\left[\sum_{i=1}^{q}(Z_i - \nu_i)g_i(Z)\right] \\
&= \mathrm{E}\left[\|g(Z)\|^2\right] + 2(1-\omega)\sum_{i=1}^{q}\mathrm{E}\left[(Z_i - \nu_i)g_i(Z)\right]
\end{aligned} \tag{7.4}$$

Lemma 1 of Stein (1981) leads to

$$\sum_{i=1}^{q} \mathrm{E}\left[(Z_i - \nu_i) g_i(Z)\right] = \sum_{i=1}^{q} \mathrm{E}\left[\frac{\partial}{\partial Z_i} g_i(Z)\right]$$

$$= \mathrm{E}\left[\sum_{i=1}^{q} \frac{\partial}{\partial Z_i} g_i(Z)\right]$$

$$= \mathrm{E}\left[\operatorname{div} g(Z)\right]. \tag{7.5}$$

Using both Eqs. (7.4) and (7.5), we achieve the request result. □

#### 7.2.1.1 Baranchik-Type and James-Stein Estimators

Rewrite the estimator defined in Eq. (7.3) by letting $g(Z) = -\left(\phi(\|Z\|^2) / \|Z\|^2\right)Z$, with the measurable function $\phi$ being differentiable and then $\delta_g(Z)$ becoming an estimator of the Baranchik type:

$$\delta_\phi(Z) = Z - \frac{\phi(\|Z\|^2)}{\|Z\|^2} Z\,. \tag{7.6}$$

Using Proposition 7.1 and the fact that

$$\operatorname{div}\left(-\frac{\phi(\|Z\|^2)}{\|Z\|^2}\right) = -(q-2)\frac{\phi(\|Z\|^2)}{\|Z\|^2} - 2\phi'(\|Z\|^2),$$

we deduce that the difference in risk between estimators $\delta_\phi(Z)$ and $\delta_0$ is given by

$$\Delta_{\omega,\phi} = \mathcal{R}_\omega(\delta_\phi(Z), \nu) - \mathcal{R}_\omega(Z, \nu)$$

$$= \mathrm{E}\left[\frac{\phi^2(\|Z\|^2)}{\|Z\|^2}\right] - 2(1-\omega)(q-2)\,\mathrm{E}\left[\frac{\phi(\|Z\|^2)}{\|Z\|^2}\right]$$

$$-4(1-\omega)\,\mathrm{E}\left[\phi'(\|Z\|^2)\right],$$

where $\phi'(\cdot)$ design the derivative function of $\phi(\cdot)$.

And the latter means that a sufficient condition for that $\delta_\phi(Z)$ defined in Eq. (7.6) is minimax is

$$\mathrm{E}\left[\frac{\phi(\|Z\|^2)}{\|Z\|^2}\left(\phi(\|Z\|^2 - 2(1-\omega)(q-2)\right) - 4(1-\omega)\phi'(\|Z\|^2)\right] \leq 0.$$

We note that this inequality is satisfied if the following classical conditions are also satisfied:

(i) The function $\phi(\cdot)$ is increasing,
(ii) $0 \leq \phi(\cdot) \leq 2(1-\omega)(q-2)$.

In the special case where $\phi(\|Z\|^2) = \gamma$ (i.e., $\delta_\gamma(Z) = Z - \gamma Z / \|Z\|^2$), the risk function of the estimator $\delta_\gamma(Z)$ is

$$\mathcal{R}_\omega(\delta_\gamma(Z), \nu) = (1-\omega)q + (\gamma^2 - 2(1-\omega)(q-2)\gamma)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\right].$$

It easy to check that a sufficient condition for that $\delta_\gamma(Z)$ dominates the MLE is

$$0 \leq \gamma \leq 2(1-\omega)(q-2),$$

and the optimal value of $\gamma$ that minimizes the risk function $\mathcal{R}_\omega(\delta_\gamma(Z), \nu)$ is

$$\hat{\gamma} = (1-\omega)(q-2).$$

If we replace $\gamma$ by $\hat{\gamma}$, we obtain the JSE as

$$\delta_{\mathrm{JS}} = \left(1 - \frac{\hat{\gamma}}{\|Z\|^2}\right) Z, \tag{7.7}$$

its risk function is equal to

$$\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) = (1-\omega)q - (1-\omega)^2(q-2)^2\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\right], \tag{7.8}$$

and the difference in risk between JSE and MLE is

$$\begin{aligned}\Delta_{\omega,\mathrm{JS}} &= \mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) - \mathcal{R}_\omega(Z, \nu) \\ &= -(1-\omega)^2(q-2)^2\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\right],\end{aligned}$$

and this explains the improvement to the JSE over the MLE.

### 7.2.2 Shrinkage Estimators Dominate the James-Stein Estimator

In this part, we consider a class of estimators that do not only dominate the MLE but also ameliorate the JSE. Let

$$\delta_{h,\mathrm{JS}}(Z) = \delta_{\mathrm{JS}} + h(Z) = \left(1 - \frac{\hat{\gamma}}{\|Z\|^2}\right) Z + h(Z), \tag{7.9}$$

where $\hat{\gamma} = (1-\omega)(q-2)$ and $h : \mathbb{R}^q \to \mathbb{R}^q$ is a measurable function.

**Proposition 7.2** *Under the BLF $\ell_\omega$, the difference in risk between estimators $\delta_{h,\mathrm{JS}}(Z)$ and $\delta_{\mathrm{JS}}$ is equal to*

$$\begin{aligned}\Delta_{\omega,h,\mathrm{JS}} = {} & \mathrm{E}\left[\|h(Z)\|^2\right] - 2(1-\omega)(q-2)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\,\langle Z, h(Z)\rangle\right] \\ & +2(1-\omega)\,\mathrm{E}\left[div\, h(Z)\right].\end{aligned} \tag{7.10}$$

***Proof*** Using the risk function associated with the BLF $\ell_\omega$, the difference in risk is equal to

$$\begin{aligned}\Delta_{\omega,h,\mathrm{JS}} &= \mathcal{R}_\omega(\delta_{h,\mathrm{JS}}(Z), \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) \\ &= \omega\,\mathrm{E}\left[\|\delta_{h,\mathrm{JS}}(Z) - Z\|^2\right] + (1-\omega)\,\mathrm{E}\left[\|\delta_{h,\mathrm{JS}}(Z) - \nu\|^2\right] - \mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) \\ &= \omega\,\mathrm{E}\left[\|\delta_{\mathrm{JS}} + h(Z) - Z\|^2\right] + (1-\omega)\,\mathrm{E}\left[\|\delta_{\mathrm{JS}} + h(Z) - \nu\|^2\right] \\ &\quad -\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) \\ &= \omega\,\mathrm{E}\left[\|\delta_{\mathrm{JS}} - Z\|^2 + \|h(Z)\|^2 + 2\left(\langle \delta_{\mathrm{JS}} - Z, h(Z)\rangle\right)\right] \\ &\quad +(1-\omega)\,\mathrm{E}\left[\|\delta_{\mathrm{JS}} - \nu\|^2 + \|h(Z)\|^2 + 2\left(\langle \delta_{\mathrm{JS}} - \nu, h(Z)\rangle\right)\right] \\ &\quad -\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) \\ &= \mathrm{E}\left[\|h(Z)\|^2\right] + 2\omega\,\mathrm{E}\left[\langle \delta_{\mathrm{JS}} - Z, h(Z)\rangle\right] \\ &\quad +2(1-\omega)\,\mathrm{E}\left[\langle \delta_{\mathrm{JS}} - \nu, h(Z)\rangle\right].\end{aligned} \tag{7.11}$$

The last equality according to the fact that

$$\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) = \omega\,\mathrm{E}\left[\|\delta_{\mathrm{JS}} - Z\|^2\right] + (1-\omega)\,\mathrm{E}\left[\|\delta_{\mathrm{JS}} - \nu\|^2\right].$$

As

$$\mathrm{E}\left[\,\langle \delta_{\mathrm{JS}} - Z, h(Z)\rangle\,\right] = -(1-\omega)(q-2)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\,\langle Z, h(Z)\rangle\,\right] \tag{7.12}$$

and

$$\begin{aligned}
\mathrm{E}\left[\,\langle \delta_{\mathrm{JS}} - \nu, h(Z)\rangle\,\right] &= \mathrm{E}\left[\,\langle \delta_{\mathrm{JS}} - Z + Z - \nu, h(Z)\rangle\,\right] \\
&= \mathrm{E}\left[\,\langle \delta_{\mathrm{JS}} - Z, h(Z)\rangle\,\right] + \mathrm{E}\left[\,\langle Z - \nu, h(Z)\rangle\,\right] \\
&= -(1-\omega)(q-2)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\,\langle Z, h(Z)\rangle\,\right] \\
&\quad + \mathrm{E}\left[\,\langle Z - \nu, h(Z)\rangle\,\right] \\
&= -(1-\omega)(q-2)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\,\langle Z, h(Z)\rangle\,\right] \\
&\quad + \mathrm{E}\left[\sum_{i=1}^{q}(Z_i - \nu_i)h_i(Z)\right],
\end{aligned}$$

and hence by applying Lemma 1 of Stein (1981), we have

$$\begin{aligned}
\mathrm{E}\left[\,\langle \delta_{\mathrm{JS}} - \nu, h(Z)\rangle\,\right] &= -(1-\omega)(q-2)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\,\langle Z, h(Z)\rangle\,\right] \\
&\quad + \mathrm{E}\left[\sum_{i=1}^{q}\frac{\partial}{\partial Z_i}h_i(Z)\right] \\
&= -(1-\omega)(q-2)\,\mathrm{E}\left[\frac{1}{\|Z\|^2}\,\langle Z, h(Z)\rangle\,\right] \\
&\quad + \mathrm{E}\left[\operatorname{div} h(Z)\right].
\end{aligned} \tag{7.13}$$

Using Eqs. (7.11), (7.12), and (7.13), we get the desired results. □

The result of the last proposition shows that a sufficient condition for that the estimator $\delta_{h,\mathrm{JS}}(Z)$ dominates $\delta_{\mathrm{JS}}$ is

$$\mathrm{E}\left[\|h(Z)\|^2\right] - 2(1-\omega)\,\mathrm{E}\left[\frac{q-2}{\|Z\|^2}\,\langle Z, h(Z)\rangle - \operatorname{div} h(Z)\right] \leq 0. \tag{7.14}$$

For $h(Z) = \left(\varphi(\|Z\|^2) \,/\, \|Z\|^2\right) Z$, namely, we consider the estimator

$$\delta_{\varphi,\mathrm{JS}}(Z) = \delta_{\mathrm{JS}} + \frac{\varphi(\|Z\|^2)}{\|Z\|^2} Z. \tag{7.15}$$

From Eq. (7.10), the difference in risk between estimators $\delta_{\varphi,\mathrm{JS}}(Z)$ and $\delta_{\mathrm{JS}}$ is expressed as

$$\begin{aligned}\Delta_{\omega,\varphi,\mathrm{JS}} &= \mathcal{R}_\omega(\delta_{\varphi,\mathrm{JS}}(Z), \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) \\ &= \mathrm{E}\left[\frac{\varphi^2(\|Z\|^2)}{\|Z\|^2}\right] + 4(1-\omega)\,\mathrm{E}\left[\varphi'(\|Z\|^2)\right].\end{aligned} \tag{7.16}$$

Equation (7.16) establishes that a sufficient condition for that the estimator $\delta_{\varphi,\mathrm{JS}}(Z)$ improves $\delta_{\mathrm{JS}}$ is

$$\Delta_{\omega,\varphi,\mathrm{JS}} \leq 0.$$

In the next result, we give a simple sufficient condition on the function $\varphi$ for that the estimator $\delta_{\varphi,\mathrm{JS}}(Z)$ dominates the JSE. Furthermore, this condition is independent of the derivative of the function $\varphi$ with respect to the variable $\|Z\|^2$. Consequently, we can assure the minimaxity of estimators of the form $\delta_{\varphi,\mathrm{JS}}(Z)$ without using the condition of the differentiability of the shrinkage function, which is a condition included in the majority of papers realized in this field because it appears immediately when the calculation of the risk function is based on Stein's identity given in Lemma 1 of Stein (1981).

**Theorem 7.1** *Under the BLF $\ell_\omega$, a sufficient condition for that $\delta_{\varphi,\mathrm{JS}}(Z)$ given in Eq. (7.15), dominating $\delta_{\mathrm{JS}}$, is*

$$0 \leq \varphi(\|Z\|^2) \leq 2\left(\hat{\gamma} - \|Z\|^2\right) \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \geq 1\right),$$

*where $\mathbb{I}(A)$ denotes the indicating function of set $A$.*

***Proof*** The risk function of the estimator $\delta_{\varphi,\mathrm{JS}}(Z)$, associated with the BLF $\ell_\omega$, is equal to

$$\begin{aligned}
\mathcal{R}_\omega(\delta_{\varphi,\mathrm{JS}}(Z),\nu) &= \omega\,\mathrm{E}\left[\|\delta_{\mathrm{JS}}+\frac{\varphi(\|Z\|^2)}{\|Z\|^2}Z-Z\|^2\right]\\
&\quad+(1-\omega)\,\mathrm{E}\left[\|\delta_{\mathrm{JS}}+\frac{\varphi(\|Z\|^2)}{\|Z\|^2}Z-\nu\|^2\right]\\
&= \omega\,\mathrm{E}\left[\|\delta_{\mathrm{JS}}-Z\|^2+\frac{\varphi^2(\|Z\|^2)}{\|Z\|^2}\right]\\
&\quad+2\omega\,\mathrm{E}\left[\left\langle\delta_{\mathrm{JS}}-Z,\frac{\varphi(\|Z\|^2)}{\|Z\|^2}Z\right\rangle\right]\\
&\quad+(1-\omega)\,\mathrm{E}\left[\|\delta_{\mathrm{JS}}-\nu\|^2+\frac{\varphi^2(\|Z\|^2)}{\|Z\|^2}\right]\\
&\quad+2(1-\omega)\,\mathrm{E}\left[\left\langle\delta_{\mathrm{JS}}-\nu,\frac{\varphi(\|Z\|^2)}{\|Z\|^2}Z\right\rangle\right]\\
&= \mathcal{R}_\omega(\delta_{\mathrm{JS}},\nu)+\omega\,\mathrm{E}\left[\frac{\varphi(\|Z\|^2)}{\|Z\|^2}\left(\varphi(\|Z\|^2)-2\,\hat{\gamma}\right)\right]\\
&\quad+(1-\omega)\,\mathrm{E}\left[\frac{\varphi(\|Z\|^2)}{\|Z\|^2}\left(\varphi(\|Z\|^2)-2\hat{\gamma}\right)\right]\\
&\quad-2(1-\omega)\,\mathrm{E}\left[\langle\nu,Z\rangle\frac{\varphi(\|Z\|^2)}{\|Z\|^2}-\varphi(\|Z\|^2)\right]. \qquad (7.17)
\end{aligned}$$

Using the formula (2.7) given by Benmansour and Mourid (2007), we have

$$\begin{aligned}
\mathrm{E}\left[\langle\nu,Z\rangle\frac{\varphi(\|Z\|^2)}{\|Z\|^2}\right] &= \mathrm{E}\left[\sum_{i=1}^{q}\left(\frac{\varphi(\|Z\|^2)}{\|Z\|^2}Z_i\nu_i\right)\right]\\
&= \lambda\,\mathrm{E}\left[\frac{\varphi(\chi^2_{q+2}(\lambda))}{\chi^2_{q+2}(\lambda)}\right]\\
&= \lambda\,\mathrm{E}\left[\frac{\varphi(\chi^2_{q+2K})}{\chi^2_{q+2K}}\right], \qquad (7.18)
\end{aligned}$$

where $\chi^2_{q+2}(\lambda)$ is the non-central chi-square distribution of the non-centrality parameter $\lambda$ and $K \sim P(\lambda/2)$ is a Poisson random variable with parameter $\lambda = \|\nu\|^2/2$.

Then, from Eqs. (7.17) and (7.18), the sufficient conditions for that $\delta_{\varphi,\mathrm{JS}}(Z)$ dominates $\delta_{\mathrm{JS}}$ are

$$\mathrm{E}\left[\frac{\varphi(\|Z\|^2)}{\|Z\|^2}\left(\varphi(\|Z\|^2)-2\hat{\gamma}\right)\right]\leq 0 \tag{7.19}$$

and

$$\mathrm{E}\left[\frac{\varphi(\|Z\|^2)}{\|Z\|^2}\left(\varphi(\|Z\|^2)-2\hat{\gamma}+2\|Z\|^2\right)\right]-2\lambda\,\mathrm{E}\left[\varphi(\chi^2_{q+2K})\frac{1}{\chi^2_{q+2K}}\right]\leq 0. \tag{7.20}$$

From Eqs. (7.19) and (7.20), we can prove easily that $\delta_{\varphi,\mathrm{JS}}(Z)$ improves $\delta_{\mathrm{JS}}$, if

$$0\leq\varphi(\|Z\|^2)\leq 2\hat{\gamma}$$

and

$$0\leq\varphi(\|Z\|^2)\leq 2\left(\hat{\gamma}-\|Z\|^2\right)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2}\geq 1\right).$$

Thus, a sufficient condition for that $\delta_{\varphi,\mathrm{JS}}(Z)$ dominates $\delta_{\mathrm{JS}}$ is

$$0\leq\varphi(\|Z\|^2)\leq 2\left(\hat{\gamma}-\|Z\|^2\right)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2}\geq 1\right).$$

□

**Example 7.1** Let $\varphi_1(\|Z\|^2)=(\hat{\gamma}-\|Z\|^2)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2}\geq 1\right)$, and therefore

$$\begin{aligned}\delta_{\varphi_1,\mathrm{JS}}&=\delta_{\mathrm{JS}}+\frac{\varphi_1(\|Z\|^2)}{\|Z\|^2}Z\\&=\left(1-\frac{\hat{\gamma}}{\|Z\|^2}\right)Z+\left(\frac{\hat{\gamma}}{\|Z\|^2}-1\right)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2}\geq 1\right)Z\\&=\left(1-\frac{\hat{\gamma}}{\|Z\|^2}\right)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2}\leq 1\right)Z\\&=\left(1-\frac{\hat{\gamma}}{\|Z\|^2}\right)^{+}Z=\delta_{\mathrm{JS}^+},\end{aligned} \tag{7.21}$$

where $(1-\hat{\gamma}/\|Z\|^2)^+=\max(0,1-\hat{\gamma}/\|Z\|^2)$, $\hat{\gamma}=(1-\omega)(q-2)$ and the estimator $\delta_{\mathrm{JS}^+}$ is so-called the PPJSE. It is clear that the function $\varphi_1(\|Z\|^2)$ satisfies the condition of Theorem 7.1, and then the estimator $\delta_{\mathrm{JS}^+}$ ameliorates $\delta_{\mathrm{JS}}$.

Based on Lemma 2.1 of Shao and Strawderman (1994) and relatively to the BLF, the risk function of the PPJSE and the difference in risk between $\delta_{\mathrm{JS}^+}$ and $\delta_{\mathrm{JS}}$ are expressed respectively as

$$\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) = \mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) + \mathrm{E}\left[\left(\|Z\|^2 + \frac{\hat{\gamma}^2}{\|Z\|^2} - 2q(1-\omega)\right)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \geq 1\right)\right], \tag{7.22}$$

and

$$\begin{aligned}\Delta_{\omega,\mathrm{JS}+,\mathrm{JS}} &= \mathcal{R}_\omega(\delta_{\delta_{\mathrm{JS}^+}}(Z), \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) \\ &= \mathrm{E}\left[\left(\|Z\|^2 + \frac{\hat{\gamma}^2}{\|Z\|^2} - 2q(1-\omega)\right)\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \geq 1\right)\right]. \end{aligned} \tag{7.23}$$

## 7.3 Main Results

In this section, we take the same model $Z = (Z_1, \ldots, Z_q) \sim \mathcal{N}_q\left(\nu, I_q\right)$ with unknown parameter $\nu = (\nu_1, \ldots, \nu_q)$ and identity covariance matrix $I_q$. Our aim is to estimate the mean parameter $\nu$ under the BLF $\ell_\omega$ defined in Eq. (7.1). The main object is to build a new large class of estimators derived from the MLE and the PPJSE and introduce between them the estimators that improve the PPJSE.

Here, we consider the estimator

$$\delta_{\beta,r,\mathrm{JS}^+} = \delta_{\mathrm{JS}^+} + \beta\left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right) Z, \tag{7.24}$$

where $r > 1$, $\hat{\gamma} = (q-2)(1-\omega)$ and the real constant $\beta$ may depend on $q$.

**Proposition 7.3** *Under the BLF $\ell_\omega$, the difference in risk between estimators $\delta_{\beta,r,\mathrm{JS}^+}$ and $\delta_{\mathrm{JS}^+}$ is*

$$\begin{aligned}\Delta_{\omega,\beta,r,\mathrm{JS}^+} &= \mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) \\ &= \beta^2\,\mathrm{E}\left[\frac{1}{\|Z\|^{4r-2}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right] \\ &\quad -4(1-\omega)(r-1)\beta\;\mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right]. \end{aligned} \tag{7.25}$$

***Proof*** The difference in risk function is equal to

$$\begin{aligned}
\Delta_{\omega,\beta,r,\mathrm{JS}^+} &= \mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) \\
&= \omega \,\mathrm{E}\left[ \left\| \delta_{\beta,r,\mathrm{JS}^+} - Z \right\|^2 \right] + (1-\omega)\,\mathrm{E}\left[ \left\| \delta_{\beta,r,\mathrm{JS}^+}(Z) - \nu \right\|^2 \right] \\
&\quad -\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) \\
&= \omega \,\mathrm{E}\left[ \left\| \delta_{\mathrm{JS}^+} + \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z - Z \right\|^2 \right] \\
&\quad +(1-\omega)\,\mathrm{E}\left[ \left\| \delta_{\mathrm{JS}^+} + \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z - \nu \right\|^2 \right] \\
&\quad -\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) \\
&= \omega \left( \mathrm{E}\left[ \left\| \delta_{\mathrm{JS}^+} - Z \right\|^2 \right] + \mathrm{E}\left[ \left\| \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\|^2 \right] \right) \\
&\quad +2\omega \,\mathrm{E}\left[ \left\langle \delta_{\mathrm{JS}^+} - Z, \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\rangle \right] \\
&\quad +(1-\omega) \Bigg( \mathrm{E}\left[ \left\| \delta_{\mathrm{JS}^+} - \nu \right\|^2 \right] \\
&\qquad + \mathrm{E}\left[ \left\| \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\|^2 \right] \Bigg) \\
&\quad +2(1-\omega)\,\mathrm{E}\left[ \left\langle \delta_{\mathrm{JS}^+} - \nu, \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\rangle \right] \\
&\quad -\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}(Z), \nu).
\end{aligned}$$

As

$$\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) = \omega \,\mathrm{E}\left[ \|\delta_{\mathrm{JS}^+} - Z\|^2 \right] + (1-\omega)\,\mathrm{E}\left[ \|\delta_{\mathrm{JS}^+} - \nu\|^2 \right],$$

and

$$\begin{aligned}
&\mathrm{E}\left[ \left\langle \delta_{\mathrm{JS}^+} - \nu, \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\rangle \right] \\
&\quad = \mathrm{E}\left[ \left\langle \delta_{\mathrm{JS}^+} - Z, \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\rangle \right] \\
&\qquad + \mathrm{E}\left[ \left\langle Z - \nu, \beta \left( \frac{1}{\|Z\|^2} \right)^r \mathbb{I}\left( \frac{\hat{\gamma}}{\|Z\|^2} \le 1 \right) Z \right\rangle \right],
\end{aligned}$$

then

$$\Delta_{\omega,\beta,r,\mathrm{JS}^+} = \beta^2 \,\mathrm{E}\left[\left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z\right]^2 + 2\beta\,\mathrm{E}\left[\left\langle \delta_{\mathrm{JS}^+} - Z, \left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z\right\rangle\right] + 2(1-\omega)\beta\,\mathrm{E}\left[\left\langle Z - \nu, \left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z\right\rangle\right]. \quad (7.26)$$

Using the fact that

$$\begin{aligned}
&\mathrm{E}\left[\left\langle \delta_{\mathrm{JS}^+} - Z, \left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z\right\rangle\right] \\
&= \mathrm{E}\left[\left\langle -\hat{\gamma}\frac{1}{\|Z\|^2}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z, \left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z\right\rangle\right] \\
&= -\hat{\gamma}\,\mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right)\right], \qquad (7.27)
\end{aligned}$$

and from Lemma 2.1 of Shao and Strawderman (1994), we have

$$\mathrm{E}\left[\left\langle Z - \nu, \left(\frac{1}{\|Z\|^2}\right)^r \mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right) Z\right\rangle\right] = (q-2r)\,\mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right)\right]. \quad (7.28)$$

Then, we put Eqs. (7.26), (7.27), and (7.28) together to get the result. □

The following theorem gives a sufficient condition for that $\delta_{\beta,r,\mathrm{JS}^+}$ defined in Eq. (7.24) improves $\delta_{\mathrm{JS}^+}$.

**Theorem 7.2** *Relatively to the BLF $\ell_\omega$, a sufficient condition for that $\delta_{\beta,r,\mathrm{JS}^+}$ defined in Eq. (7.24), dominating $\delta_{\mathrm{JS}^+}$, is*

$$0 \le \beta \le 4(r-1)(1-\omega)^r (q-2)^{r-1}.$$

***Proof*** The difference between the risk functions can be expressed using the formula below:

$$\begin{aligned}
\Delta_{\omega,\beta,r,\mathrm{JS}^+} &= \mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) \\
&= \beta^2\,\mathrm{E}\left[\frac{1}{\|Z\|^{4r-2}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right)\right] \\
&\quad -4(1-\omega)(r-1)\beta\,\mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \le 1\right)\right],
\end{aligned}$$

and the increase of the function $f_\alpha(x) = x^\alpha, \alpha > 0$ in $(0, \infty)$ leads to

$$\left[\frac{1}{(\|Z\|^2)^{r-1}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right] \leq \frac{1}{[(q-2)(1-\omega)]^{r-1}},$$

and we can deduce that

$$\mathrm{E}\left[\frac{1}{\|Z\|^{4r-2}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right] = \mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\frac{1}{(\|Z\|^2)^{r-1}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right]$$

$$\leq \mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right]\frac{1}{[(q-2)(1-\omega)]^{r-1}}. \tag{7.29}$$

Thus, a sufficient condition for that $\delta_{\beta,r,\mathrm{JS}^+}$ dominating $\delta_{\mathrm{JS}^+}$ is

$$\beta\left[\beta(1-\omega)^{1-r}(q-2)^{1-r} - 4(1-\omega)(r-1)\right]\mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right] \leq 0,$$

which is equivalent to

$$0 \leq \beta \leq \frac{4(1-\omega)(r-1)}{(q-2)^{1-r}(1-\omega)^{1-r}} = 4(r-1)(1-\omega)^r(q-2)^{r-1}.$$

□

According to the convexity of the risk function $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu)$ with respect to the variable $\beta$, the optimal value of $\beta$ which minimizes $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu)$ is

$$\hat{\beta} = 2(1-\omega)(r-1)\frac{\mathrm{E}\left[\frac{1}{\|Z\|^{2r}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right]}{\mathrm{E}\left[\frac{1}{\|Z\|^{4r-2}}\mathbb{I}\left(\frac{\hat{\gamma}}{\|Z\|^2} \leq 1\right)\right]}. \tag{7.30}$$

## 7.4 Numerical Study

We recall the risk function associated with the BLF, of the JSE given by Eq. (7.8), the risk of the PPJSE defined in Eq. (7.22), and the difference in risk between estimators $\delta_{\beta,r,\mathrm{JS}^+}$ and $\delta_{\mathrm{JS}^+}$ mentioned in Eq. (7.25) with $\hat{\beta}$ defined in Eq. (7.30).

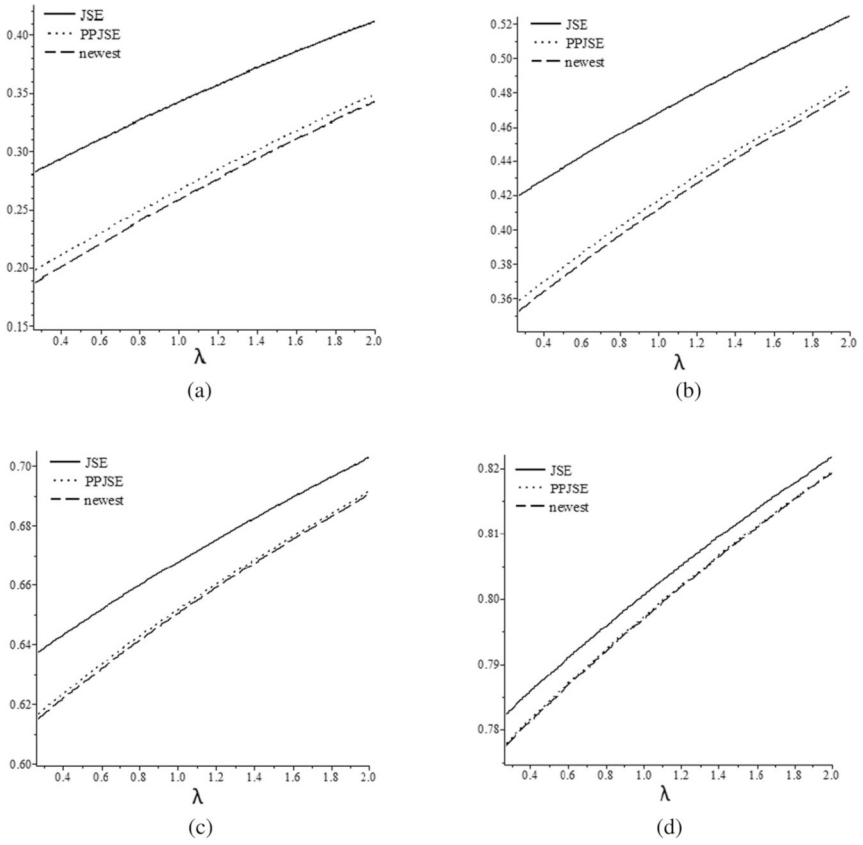



**Fig. 7.1** $\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) / \mathcal{R}_\omega(Z, \nu)$, $\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$, and $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$ as functions of $\lambda = \|\nu\|^2$ for $r = 1.5$, $q = 12$ and diverse values of $\omega$. (**a**) $q = 8$, $\omega = 0.01$, and $r = 1.5$. (**b**) $q = 8$, $\omega = 0.2$, and $r = 1.5$. (**c**) $q = 8$, $\omega = 0.5$, and $r = 1.5$. (**d**) $q = 8$, $\omega = 0.7$, and $r = 1.5$

First, we graph the risk ratios of JSE, PPJSE, and the new estimator i.e, newest $:= \delta_{\beta,r,\mathrm{JS}^+}$ (with $\beta := \hat{\beta}$) to the MLE denoted by $\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu)/\mathcal{R}_\omega(Z, \nu)$, $\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu)/\mathcal{R}_\omega(Z, \nu)$, and $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu)/\mathcal{R}_\omega(Z, \nu)$ as functions of $\lambda = \|\nu\|^2$ for diverse values of $q$, $\omega$ and $r$. One can check to see the dominance of one over the other to confirm our study.

Figures 7.1, 7.2, 7.3, and 7.4 show that for different values of $q$, $r$, and $\omega$, the risk ratios of JSE, PPJSE, and the newest $:= \delta_{\beta,r,\mathrm{JS}^+}$ to the MLE $Z$ are less than 1, and this means that the JSE, PPJSE, and $\delta_{\beta,r,\mathrm{JS}^+}$ dominate the MLE; therefore, they are minimax for the selected values of $q$, $r$, and $\omega$. We also observe that for the same values of $q$, $r$, and $\omega$, the PPJSE and $\delta_{\beta,r,\mathrm{JS}^+}$ dominate the JSE and the PPJSE, respectively, and we obtain a significant improvement of the PPJSE over the JSE

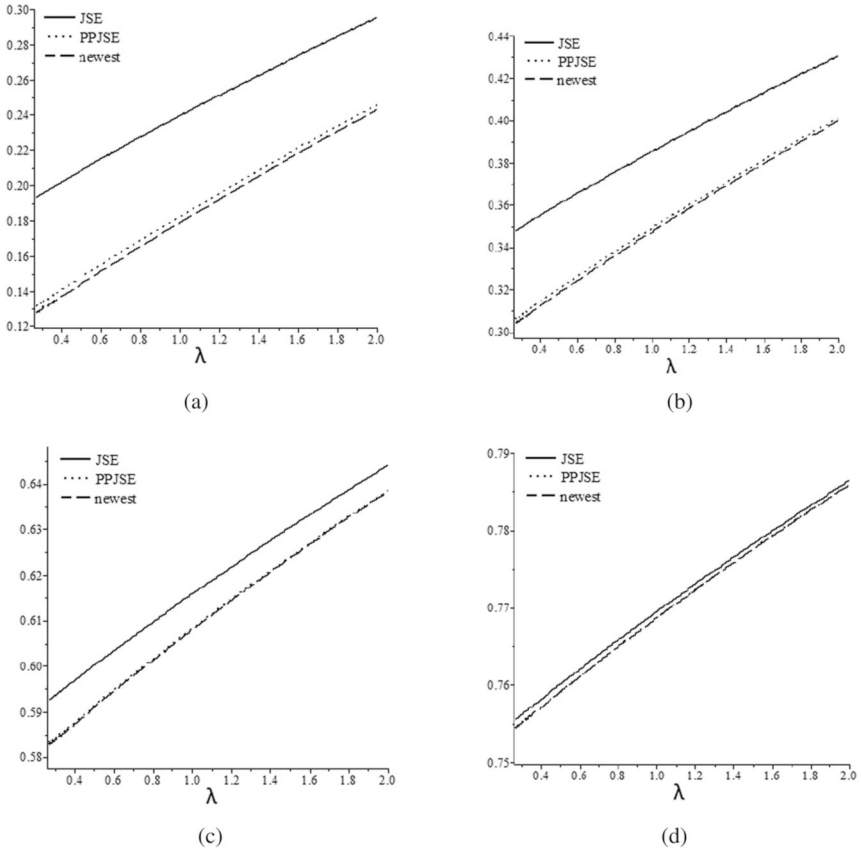



**Fig. 7.2** $\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) / \mathcal{R}_\omega(Z, \nu)$, $\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$, and $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$ as functions of $\lambda = \|\nu\|^2$ for $r = 1.5$, $q = 12$, and diverse values of $\omega$. (**a**) $q = 12$, $\omega = 0.01$, and $r = 1.5$. (**b**) $q = 12$, $\omega = 0.2$, and $r = 1.5$. (**c**) $q = 12$, $\omega = 0.5$, and $r = 1.5$. (**d**) $q = 12$, $\omega = 0.7$, and $r = 1.5$

and the $\delta_{\beta,r,\mathrm{JS}^+}$ over the PPJSE as the values of $\omega$ and $\lambda = \|\nu\|^2$ are near to zero, and this amelioration decreases as we increase the value of these parameters. The effect of the other parameters $q$ and $r$ on the performance of the estimators PPJSE over the JSE and the $\delta_{\beta,r,\mathrm{JS}^+}$ over the PPJSE is not clear and may be taken as a new problem to discuss about what is the optimized value of $r$ that maximizes the difference in risks between estimators $\delta_{\beta,r,\mathrm{JS}^+}$ and PPJSE.

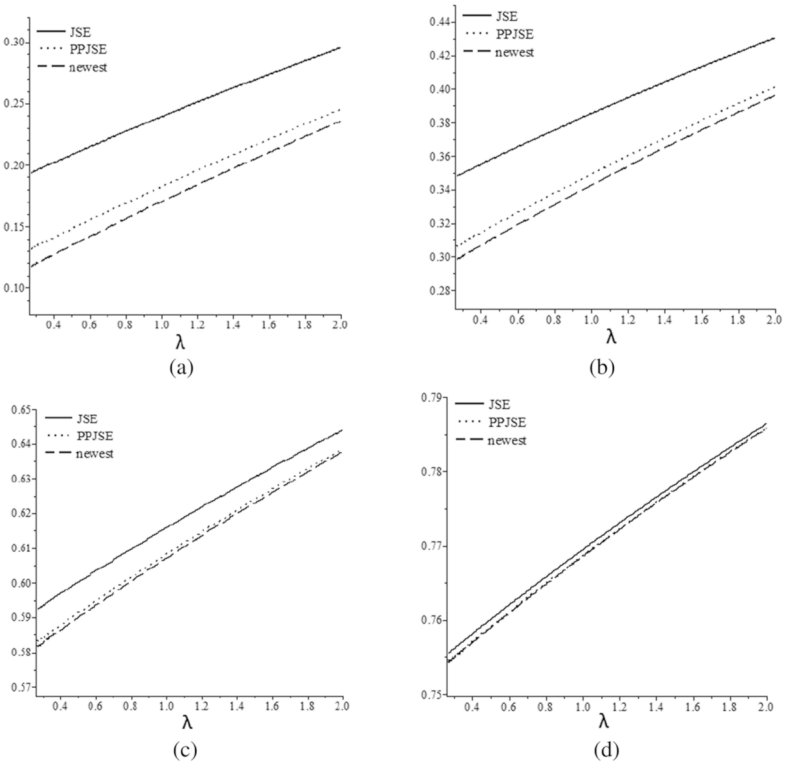



**Fig. 7.3** $\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) / \mathcal{R}_\omega(Z, \nu)$, $\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$, and $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$ as functions of $\lambda = \|\nu\|^2$ for $r = 2$, $q = 8$, and diverse values of $\omega$. (**a**) $q = 8$, $\omega = 0.01$, and $r = 2$. (**b**) $q = 8$, $\omega = 0.2$, and $r = 2$. (**c**) $q = 8$, $\omega = 0.5$, and $r = 2$. (**d**) $q = 8$, $\omega = 0.7$, and $r = 2$

To confirm our previous comments related to the improvement of the newest $:= \delta_{\beta,r,\mathrm{JS}^+}$ over the PPJSE, we give Figs. 7.5 and 7.6 that present the curve of the function of difference in risks between newest $:= \delta_{\beta,r,\mathrm{JS}^+}$ and PPJSE defined by Eq. (7.25) as functions of $\lambda = \|\nu\|^2$ for fixed values of $q$ ($q \in \{10, 14\}$), $\omega$ ($\omega \in \{0.01, 0.2, 0.5, 0.7\}$) and $r$ ($r \in \{3/2, 2, \sqrt{5}\}$).

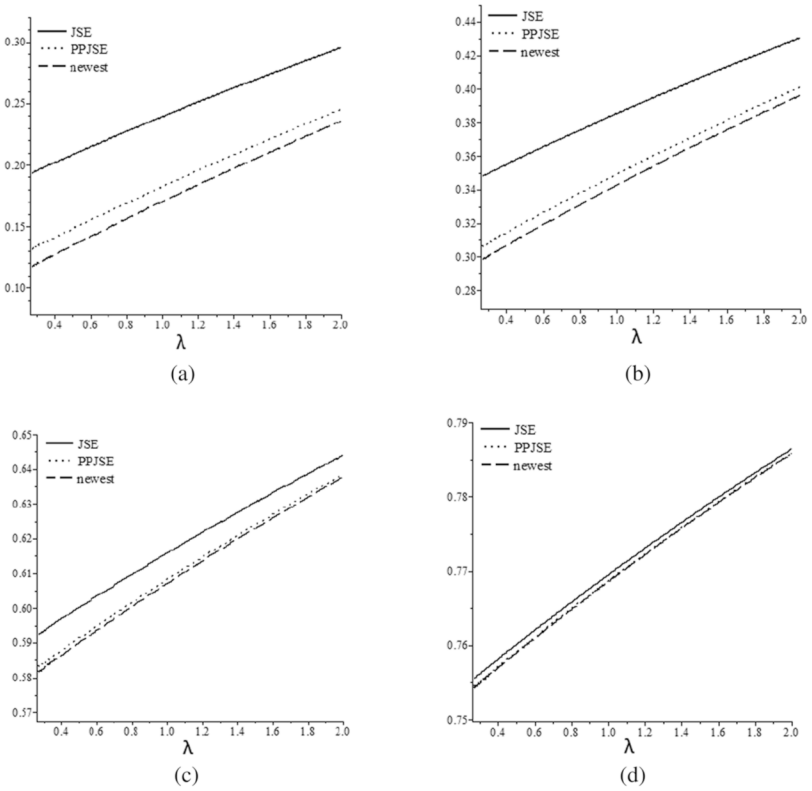



**Fig. 7.4** $\mathcal{R}_\omega(\delta_{\mathrm{JS}}, \nu) / \mathcal{R}_\omega(Z, \nu)$, $\mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$, and $\mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) / \mathcal{R}_\omega(Z, \nu)$ as functions of $\lambda = \|\nu\|^2$ for $r = 2$, $q = 12$, and diverse values of $\omega$. (**a**) $q = 12$, $\omega = 0.01$, and $r = 2$. (**b**) $q = 12$, $\omega = 0.2$, and $r = 2$. (**c**) $q = 12$, $\omega = 0.5$, and $r = 2$. (**d**) $q = 12$, $\omega = 0.7$, and $r = 2$

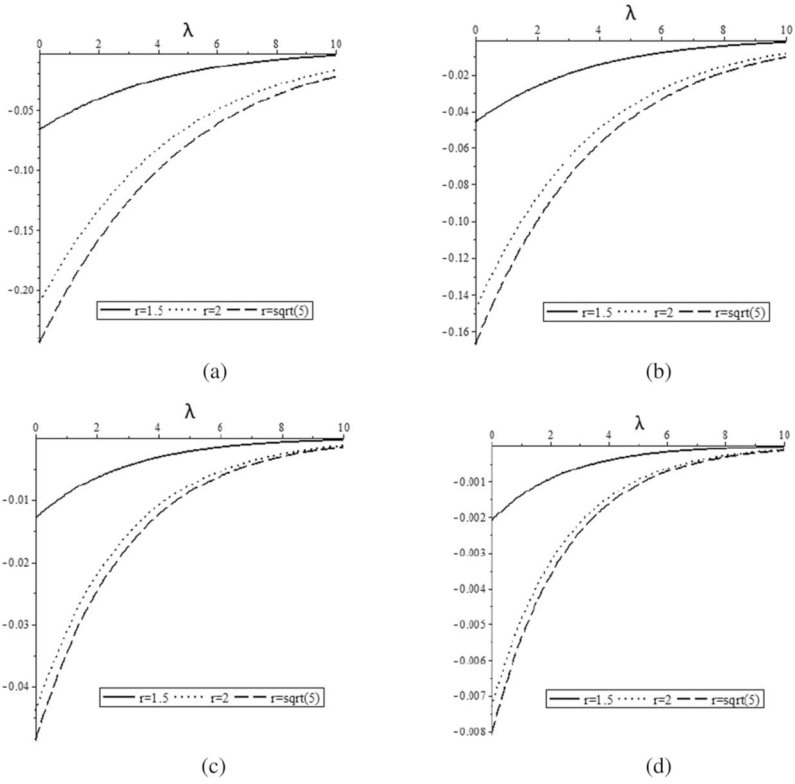



**Fig. 7.5** $\Delta_{\omega,\beta,r,\mathrm{JS}^+} = \mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) - \mathcal{R}_\omega(\delta_{\mathrm{JS}^+}, \nu)$ as functions of $\lambda = \|\nu\|^2$ for $q = 10$ and diverse values of $r$ and $\omega$. (**a**) $q = 10$ and $\omega = 0.01$. (**b**) $q = 10$ and $\omega = 0.2$. (**c**) $q = 10$ and $\omega = 0.5$. (**d**) $q = 10$ and $\omega = 0.7$

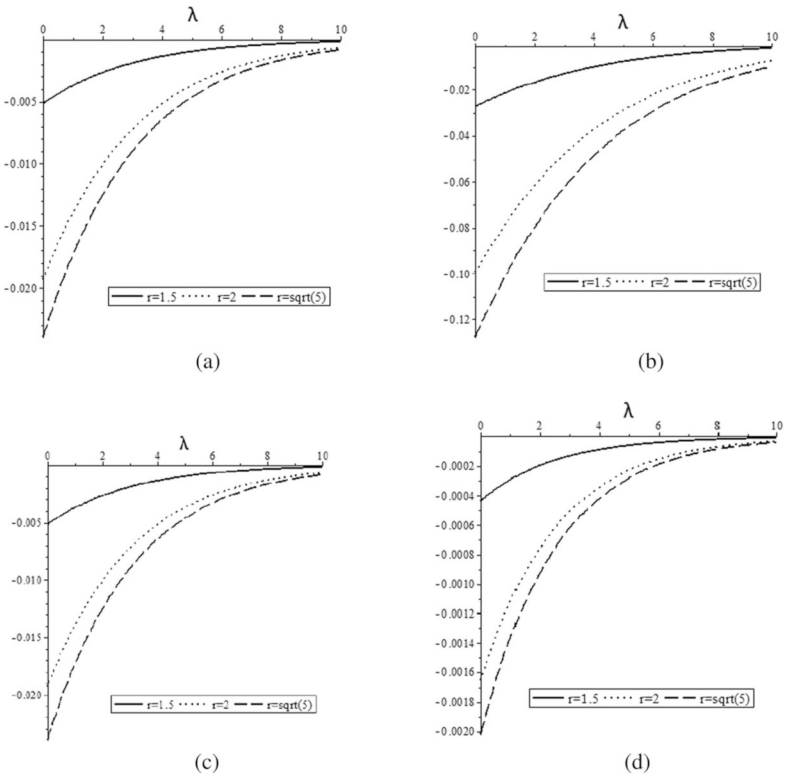



**Fig. 7.6** $\Delta_{\omega,\beta,r,\mathrm{JS}^+} = \mathcal{R}_\omega(\delta_{\beta,r,\mathrm{JS}^+}, \nu) - \mathcal{R}_\omega(\delta_{JS^+}, \nu)$ as functions of $\lambda = \|\nu\|^2$ for $q = 14$ and diverse values of $r$ and $\omega$. (**a**) $q = 14$ and $\omega = 0.01$. (**b**) $q = 14$ and $\omega = 0.2$. (**c**) $q = 14$ and $\omega = 0.5$. (**d**) $q = 14$ and $\omega = 0.7$

## 7.5 Conclusion

In this chapter, we are interested in estimation of the multivariate normal distribution $Z \sim \mathcal{N}_q\left(\nu, I_q\right)$ under the balanced loss function. We review some results regarding the sufficient conditions for the minimaxity of the estimators $\delta_g(Z) = Z + g(Z)$, where $g$ is a measurable differentiable function. Specifically, we discuss Baranchik-type estimators $\delta_h(Z) = (1 - (\phi(\|Z\|^2)/\|Z\|^2))Z$, where $h$ is also a measurable differentiable function. Additionally, we highlight that the well-known class of estimators $\delta_{h,\mathrm{JS}}(Z) = \delta_{\mathrm{JS}} + (h(\|Z\|^2)/\|Z\|^2)\,Z$ dominates the James-Stein estimator, concluding that the positive part of James-Stein estimator (PPJSE) is part of this class. Our main objective in this work was to construct the new class of estimators derived from the maximum likelihood estimator and the PPJSE as $\delta_{\beta,r,\mathrm{JS}^+} = \delta_{\mathrm{JS}^+} + \beta\left(1/\|Z\|^2\right)^r \mathbb{I}\left(\hat{\gamma}/\|Z\|^2 \leq 1\right) Z$, where $r$ is a real parameter that is not necessarily an integer, and to determine the interval of the values of the parameter $\beta$ for that $\delta_{\beta,r,\mathrm{JS}^+}$ improves PPJSE. All of simulation results supported

our theoretical main results. The underlying aim of using shrinkage and Baranchik-type estimators is to improve the quality and stability of the estimates and, in many cases, to obtain a more robust estimator with less risk.

As a future project to progress and enrich the ideas of this chapter, we propose to determine the optimal value of $r$ to obtain the best estimator in the class of estimators $\delta_{\beta,r,\mathrm{JS}^+}$, to generalize this study for a spherically symmetric distributions instead to the multivariate normal distribution, and to present a similar study in a Bayesian framework.

## References


Ahmed, S. E. (2000). Stein-type shrinkage quantile estimation. *Stochastic Analysis and Applications, 18*(3), 475–492.

Alam, K. (1973). A family of admissible minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics, 1*(3), 517–525.

Arnold, F. S. (1981). *The theory of linear models and multivariate analysis*. Wiley.

Berger, J. O. (1982). Selecting a minimax estimator of a multivariate mean. *The Annals of Statistics, 10*(1), 81–92.

Benkhaled, A., & Hamdaoui, A. (2022). General classes of shrinkage estimators for the multivariate normal mean with unknown variance: Minimaxity and limit of risk ratios. *Kragujevac Journal of Mathematics, 46*(2), 193–213.

Benmansour, D., & Mourid, T. (2007). Étude d'une classe d'estimateurs avec rétrécisseur de la moyenne d'une loi gaussienne. *Annales de l'I.S.U.P., 51*(1), 83–106.

Efron, B., & Morris, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics, 4*(1), 11–21.

Hamdaoui, A., & Benmansour, D. (2015). Asymptotic properties of risk ratios of shrinkage estimators. *Hacettepe Journal of Mathematics and Statistics, 44*(5), 1181–1195.

Hamdaoui, A., Almutiry, W., Terbeche, M., & Benkhaled, A. (2021a). Comparison of risk ratios of shrinkage estimators in high dimensions. *Mathematics, 52*(1), 1–14.

Hamdaoui, A., Terbeche, M., & Benkhaled, A. (2021b). On shrinkage estimators improving the James-Stein estimator under balanced loss function. *Pakistan Journal of Statistics and Operation Research, 17*(3), 711–727.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 361–379). University of California Press.

Karamikabir, H., Afshari, M., & Arashi, M. (2018). Shrinkage estimation of non-negative mean vector with unknown covariance under balanced loss. *Journal of Inequalities and Applications, 2018*(1), 1–11.

Nimet, O., Issam, D., & Selahattin, K. (2017). Feasible generalized Stein-rule restricted ridge regression estimators. *Journal of Applied Mathematics, Statistics and Informatics, 13*(2), 77–97.

Nimet, O., & Selahattin, K. (2019). Risk performance of some shrinkage estimators. *Communications in Statistics—Simulation and Computation, 50*(1), 1–20.

Norouzirad, M., & Arashi, M. (2019). Preliminary test and Stein-type shrinkage ridge estimators in robust regression. *Statistical Papers, 60*(1), 1–20.

Norouzirad, M., Arashi, M., Marques, F. J., & Mamode Khan, N. A. (2023). Feasible Stein-type and preliminary test estimations in the system regression model. *Statistics, Optimization and Information Computing, 11*(2), 258–275.

Saleh, A. K. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*, Wiley Series in Probability and Statistics.

Shao, P., & Strawderman, W. E. (1994). Improving on the James-Stein positive-part estimator. *The Annals of Statistics, 22*(3), 1517–1539.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 197–206). University of California Press.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics, 9*(6), 1135–1151.

Strawderman, W. E. (1973). Proper Bayes minimax estimators of the multivariate normal mean vector for the case of common unknown variances. *The Annals of Statistics, 1*(6), 1189–1194.

Sun, L. (1995). Risk ratio and minimaxity in estimating the multivariate normal mean with unknown variance. *Scandinavian Journal of Statistics, 22*(1), 105–120.

Takahasi, K., & Wakimoto, K. (1968). On the unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics, 20*(1), 1–31.

Yüzbaşı, B., Arashi, M., & Ahmed, S. E. (2020). Shrinkage estimation strategies in generalized ridge regression models: Low/high-dimension regime. *International Statistical Review, 88*(1), 1–23.

Zellner, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions. In J. O. Berger & S. S. Gupta (Eds.), *Statistical Decision Theory and Related Topics* (pp. 337–390). Springer.

# Chapter 8
# Shrinkage Estimators of the Location Parameter Under Modified Balanced Loss Functions




**Lahoucine Hobbad, Idir Ouassou, and Mohamed Alahiane**

## 8.1 Introduction and Preliminaries

Let the random observation have the form $(X, U)$, which is spherically symmetrically distributed around a vector of the form $(\theta, 0)$ with $\dim(X) = \dim(\theta) = p$ and $\dim(U) = \dim(0) = k$, denoted $(X, U) \sim SS_{(d,k)}(\theta, 0)$. This model with residual has been considered by Cellier et al. (1989), Brandwein and Strawderman (1991), Brandwein et al. (1993), Cellier and Fourdrinier (1995), and Fourdrinier and Strawderman (1996). These papers study classes of estimators that improve on the usual minimax estimator $X$. A particularly important class of such estimators, when a residual vector $U$ is available, is the class of robust James-Stein estimators $\delta_{\mathrm{JS}}(X, \|U\|^2) = \left(1 - a\|U\|^2/\|X\|^2\right) X$. The estimator $\delta_{\mathrm{JS}}(X, \|U\|^2)$ that improves on $X$ occurs when $0 < a < 2(d-2)/(k+2)$ under quadratic loss (see Fourdrinier & Strawderman, 1996). Brandwein and Strawderman (1991) showed that the estimator $\delta_{0,g}(X, \|U\|^2)$ that dominates the unbiased estimator $X$ under the concave loss function $L(\theta, \delta) = \ell(\|\theta - \delta\|^2)$ where the concave nondecreasing function $\ell$ also satisfies $t^{\alpha}\ell(t)$ is nondecreasing if $0 < a < (d - 2 - 2\alpha)/(d(k + 2))$. Fourdrinier

L. Hobbad · I. Ouassou (✉)
Cadi Ayyad University, National School of Applied Sciences, Av. Abdelkrim Khattabi, Marrakesh, Morocco
e-mail: la.hobbad@gmail.com; i.ouassou@uca.ac.ma

M. Alahiane
Chouaib Doukkali University, National School of Business and Management, LERSEM Laboratory, corner. Av. Ahmed Chaouki and Rue de Fès, El Jadida, Morocco
e-mail: alahianemed@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_8

et al. (2003) showed that the Baranchik-type estimator (Baranchik, 1970), defined as

$$\delta_{a,S}(X, \|U\|^2) = \left(1 - \frac{a\|U\|^2 S(\|X\|^2/\|U\|^2)}{\|X\|^2}\right) X \tag{8.1}$$

dominates the standard minimax estimator $X$ under the condition that the shrinkage parameter $a$ satisfies the inequality $0 < a < (2(d-2)/(k+2))$. Fourdrinier and Ouassou (2000); Fourdrinier et al. (2008) have studied the problem of estimating, under quadratic loss, the mean of a spherically symmetric distribution when its norm is supposed to be known and when a residual vector is available. Fourdrinier et al. (2003) have considered the problem of estimating a $d$-dimensional parameter $\theta = (\theta_1, \ldots, \theta_d)$ when the observation is a $(d+k)$-vector $(X, U)$. Two restrictions on the parameter $\theta$ are considered: first, they assume that $\theta_i \geq 0$ for $i = 1, \ldots, p$, and, second, thy suppose that only a subset of these $\theta_i$ are nonnegative.

Balanced loss functions (BLFs), introduced by Zellner (1994), are formulated to reflect two criteria, namely, goodness of fit and estimation precision. This loss function is defined as $\omega\|\delta - \delta_0\|^2 + (1-\omega)\|\delta - \theta\|^2$, with the first term measuring proximity of estimate $\delta$ to the target $\delta_0$ in comparison to the second term measuring proximity of $\delta$ to estimand $\theta$ weighted by $\omega$ and $1-\omega$, respectively, the special case where $\omega = 0$ corresponds to the scaled unbalanced loss function. In this case, the decisions are not influenced by the target $\delta_0$. Giles et al. (1996), Dey et al. (1999), Chaturvedi and Shalabh (2004), and Jafari Jozani et al. (2006) have studied the application of balanced loss function in Stein-rule estimation and some other areas of linear models. Marchand and Strawderman (2020) introduced a novel class of loss functions called modified balanced loss functions (MBLFs), denoted as

$$\omega\rho(\|\delta - \delta_0\|^2) + (1-\omega)\rho(\|\delta - \theta\|^2) \tag{i}$$

$$\ell(\omega\|\delta - \delta_0\|^2 + (1-\omega)\|\delta - \theta\|^2) \tag{ii}$$

where $\rho(\cdot) \geq 0$ and $\ell(\cdot) \geq 0$ are increasing and concave functions and satisfy a completely monotone property. Notably, these MBLFs can be considered as a natural extension of Zellner's balanced loss function, where the functions $\rho(\cdot)$ and $\ell(\cdot)$ take the simple form of $t$ for all nonnegative values of $t$ in the real numbers $\mathbb{R}_+$. Hobbad et al. (2021) generalized this class of the loss functions (i) and (ii), considering only that $\rho$ and $\ell$ are increasing and concave functions.

In this work, we are interested in the estimation of the unknown parameter $\theta$ under the modified balanced loss functions (MBLFs) (i) and (ii), and our objective is to find a sufficient condition on the constant $a$ and the function $g$ for which the robust shrinkage estimators

$$\delta_{\omega,g}(X, \|U\|^2) = X + a(1-\omega)\|U\|^2 g(X) \tag{8.2}$$

dominate the target estimator $\delta_0(X) = X$. In this context, we assume that there exists a nonpositive function $h(\cdot)$ such that $h(X)$ is subharmonic, $\mathrm{E}_\theta\left[\|h(X)\|\right] < \infty$, and $\mathrm{E}_{R,\theta}\left[R^2 h(W)\right]$ does not increase with $\mathrm{E}_{R,\theta}$, which is the expectation with respect to the uniform distribution $U_{R;\theta}$ on the sphere $S_{R;\theta} = \{y \in R^{p+k} : \|y - \theta\| = R\}$ of radius $R$ and centered at $\theta$. Moreover, the function $g(\cdot)$ is weakly differentiable and (a) $\mathrm{div}(g) \le h$ and (b) $\|g\|^2 + 2h \le 0$.

Notice that the James-Stein estimator and the Baranchik estimator can be viewed as special cases of a more general class of robust shrinkage estimators in Eq. (8.2). More specifically, the James-Stein estimator can be written in the form

$$\delta^{\mathrm{JS}}_{\omega,a}(X, \|U\|^2) = \left(1 - (1-\omega)a\frac{\|U\|^2}{\|X\|^2}\right)X.$$

Similarly, the Baranchik-type estimators are defined as

$$\delta_{\omega,a}(X, \|U\|^2) = \left(1 - a(1-\omega)\|U\|^2\frac{S(\|X\|^2)}{\|X\|^2}\right)X\,,$$

such that $a > 0$ and the function $S(\cdot)$ is twice differentiable and satisfies in the conditions $0 \le S(\cdot) \le 1$, $S'(\cdot) \ge 0$, and $S''(\cdot) \le 0$. Throughout this work, we deal with the expectations conditioned on the radius $R$ of a spherically symmetric distribution in $\mathbb{R}^d \times \mathbb{R}^k$ centered at $(\theta, 0)$ where $\theta \in \mathbb{R}^d$. These expectations reduce to integrals with respect to the uniform distribution $U_{R,\theta}$ on the sphere:

$$S_{R,\theta} = \left\{y = (x,u) \in \mathbb{R}^d \times \mathbb{R}^k \mid \sqrt{\|x\|^2 + \|u\|^2} = R\right\}.$$

If $\mathrm{E}_{R,\theta}[\psi]$ is the expectation of some function $\psi$ with respect to $U_{R,\theta}$, the expectation with respect to the entire distribution is given by $\mathrm{E}_\theta\left[\psi\right] = \mathrm{E}\left[\mathrm{E}_{R,\theta}[\psi]\right]$ where E is the expectation with respect to the distribution of the radius.

This chapter is organized as follows: In Sect. 8.2, we provide conditions on the constant $a$ and on the function $g$ such that the estimator $\delta_{\omega,g}(X, \|U\|^2)$ in Eq. (8.2) to dominate $\delta_0(X) = X$ under MBLFs (i) with $(X, U)$ has spherically symmetric distributions with residual in both cases: with density and without density. In Sect. 8.3, we study the same problem as the previous section under the loss function MBLFs (ii). In Sect. 8.4, we give some examples and simulations to illustrate our theoretical results. Section 8.5 devotes to concluding remarks.

## 8.2 Risk Analysis for Loss $\omega\rho(\|\delta - \delta_0\|^2) + (1-\omega)\rho(\|\delta - \theta\|^2)$

In this section, we will examine the minimax properties of the robust shrinkage estimators given by Eq. (8.2) for a location vector $\theta$ when the observed vector

$(X, U)$ has a spherically symmetric distribution with location parameter $(\theta, 0)$ under the MBLFs (i). We make the following assumption.

**Assumption 8.1** *The function $\rho$ satisfies the following conditions:*

*(i)* $\rho(0) = 0$.
*(ii)* $0 < \rho'(0) < +\infty$.
*(iii)* $\rho$ *is concave.*

The examples of functions $\rho$ satisfying Assumption 8.1 are:

$\rho(t) = t$,
$\rho(t) = 1 - \exp(-t/\alpha)$ with $\alpha > 0$
$\rho(t) = \log(1 + t)$
$\rho(t) = (1 + t/\gamma)^{\beta}$ with $\gamma > 0$ and $\beta \in (0, 1)$
$\rho(t) = r^2 t/(rt + 1)$ with $r > 0$
$\rho(t) = \arctan(t)$
$\rho(t) = \tanh(t)$

However, it is important to note that the last two examples do not exhibit complete monotonicity. In other words, $(-1)^k \rho^k(t) \geq 0$ for $t > 0$ and $k \in \mathbb{N}$ is not satisfied for this example.

If the vector $(X, U)$ follows the distribution $SS_{(d,k)}(\theta, 0)$ and has a density with respect to the Lebesgue measure, then its density is necessarily of the form $f(\|x - \theta\|^2 + \|u\|^2)$ and the radius $R$ has density $t \mapsto 2\pi^{p+k}\Gamma^{-1}((d + k)/2) f(t^2) t^{d+k-1}$.

**Theorem 8.1** *Suppose that $(X, U) \sim SS_{(d,k)}(\theta, 0)$, i.e., spherically symmetric distribution about $(\theta, 0)$. Then, with respect to the MBLFs (i), $\delta_{\omega,g}(X, U)$ has smaller risk than $\delta_0$ provided:*

1. $\|g\|^2/2 \leq -h \leq div(g)$, *where* $-h$ *is superharmonic*
2. $d > 4$ *and* $k > 2 + (4d / (d - 4))$
3. *For* $r^2 = \|X - \theta\|^2$,

$$0 < a \leq \frac{1}{k+2} \frac{\mathrm{E}[r^2 \rho'(r^2)]}{\omega \rho'(0)\, \mathrm{E}\left[r^2\right] + (1 - \omega)\, \mathrm{E}\left[r^2 \rho'(r^2)\right]}$$

***Proof*** First, we give an expression for the risk difference:

$$\begin{aligned}
&\Delta\mathcal{R}(\omega, \delta_0, \rho) \\
&\quad = \mathcal{R}(\omega, \rho, \delta) - \mathcal{R}(\theta, \rho, \delta_0) \\
&\quad = \mathrm{E}_\theta \left[ L_{\omega,\rho,\delta_0}(\theta, \delta) - L_{\omega,\rho,\delta_0}(\theta, \delta_0) \right] \\
&\quad = \mathrm{E}_\theta \left[ \omega\rho(\|\delta(X) - \delta_0(X)\|^2) + (1 - \omega)\rho(\|\delta(X) - \theta\|^2) \right. \\
&\qquad \left. -(1 - \omega)\rho(\|\delta_0(X) - \theta\|^2) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}_\theta \left[\omega\rho(\|X + a(1-\omega)\|U\|^2 g(X) - X\|^2) \right. \\
&\quad \left. +(1-\omega)\rho(\|X + a(1-\omega)\|U\|^2 g(X) - \theta\|^2)(1-\omega)\rho(\|X-\theta\|^2)\right] \\
&= \mathrm{E}_\theta \left[\omega\rho(\|a(1-\omega)\|U\|^2 g(X)\|^2) \right. \\
&\quad \left. +(1-\omega)\left\{\rho(\|X-\theta + a(1-\omega)\|U\|^2 g(X)\|^2) - \rho(\|X-\theta\|^2)\right\}\right] \\
&\le \mathrm{E}_\theta \left[\omega\rho'(0)a^2(1-\omega)^2\|U\|^4\|g(X)\|^2 + (1-\omega)\rho'(\|X-\theta\|^2)\right. \\
&\quad \left. \times\left(2a(1-\omega)\|U\|^2 g(X).(X-\theta) + a^2(1-\omega)^2||g(X)||^2\|U\|^4\right)\right].
\end{aligned}
$$

The last inequality comes from the fact that $\rho(t) \le \rho(0) + \rho'(0)t = \rho'(0)t$, which holds since $\rho$ is concave with $\rho(0) = 0$. Then, the difference in risk $\left(\Delta\mathcal{R}(\omega, \delta_0, \rho)\right)/a(1-\omega)^2$ is bounded by

$$
\begin{aligned}
\mathrm{E}_\theta \big[ \; & a\omega\rho'(0)\|U\|^4\|g(X)\|^2 + 2\rho'(\|X-\theta\|^2)\|U\|^2 g(X) \\
& \times(X-\theta) + a\,(1-\omega)\rho'(\|X-\theta\|^2)\,\|g(X)\|^2\|U\|^4\big].
\end{aligned} \tag{8.3}
$$

Using Lemma 8.5, we obtain

$$
\begin{aligned}
&\mathrm{E}_{R,\theta}\left[\rho'(R^2 - \|U\|^2)\|U\|^2 g(X).(X-\theta)\right] \\
&\le \mathrm{E}_{R,\theta}\left[\rho'(R^2 - \|U\|^2)\frac{\|U\|^2\frac{k}{2} + 1}{(k+2)(\|U\|^2)\frac{k}{2} - 1}\mathrm{div}(g(X))\right] \\
&= \mathrm{E}_{R,\theta}\left[\rho'(R^2 - \|U\|^2)\frac{\|U\|^4}{k+2}\mathrm{div}(g(X))\right].
\end{aligned}
$$

Hence, the difference in risk in Eq. (8.3) is bounded by

$$
\begin{aligned}
&\mathrm{E}_\theta\left[a\omega\rho'(0)\|U\|^4\|g(X)\|^2 + 2\rho'(R^2 - \|U\|^2)\frac{\|U\|^4}{k+2}\mathrm{div}(g(X))\right. \\
&\quad \left. +a(1-\omega)\rho'(\|X-\theta\|^2)\|g(X)\|^2\|U\|^4\right]. \\
&= \mathrm{E}_\theta\left[a\left(\omega\rho'(0) + (1-\omega)\rho'(\|X-\theta\|^2)\right)\|g(X)\|^2\|U\|^4\right. \\
&\quad \left. +\frac{2}{k+2}\rho'(R^2 - \|U\|^2)\|U\|^4\,\mathrm{div}(g(X))\right].
\end{aligned} \tag{8.4}
$$

Furthermore, suppose that the function $g(\cdot)$ is weakly differentiable and also satisfies $\operatorname{div}(g(X)) \leq h(X) \leq -1/2\|g(X)\|^2$; then the expectation in Eq. (8.4) is bounded by

$$\mathrm{E}\left[\left(a\omega\rho'(0) + a(1-\omega)\rho'(\|X-\theta\|^2) - \frac{\rho'(\|X-\theta\|^2)}{(k+2)}\right)(-2\|U\|^4 h(X))\right]. \tag{8.5}$$

Conditioning with respect to the radius $r = \|X-\theta\|$ and $R^2 = \|X-\theta\|^2 + \|U\|^2$, the expectation in Eq. (8.5) becomes

$$\begin{aligned}
&\mathrm{E}\Bigg[\left(a\omega\rho'(0) + a(1-\omega)\rho'(r^2) - \frac{\rho'(r^2)}{k+2}\right) \\
&\qquad \times (R^2-r^2)^2\,\mathrm{E}_{R,\theta}\left[-2h(X) | \|X-\theta\| = r,\, R\right]\Bigg] \\
&= \mathrm{E}\Bigg[\mathrm{E}\Big[\left(a\omega\rho'(0) + a(1-\omega)\rho'(\beta R^2)\right) \\
&\qquad \times (R^2-\beta R^2)^2\,\mathrm{E}_{R,\theta}\left[-2h(X) | \|X-\theta\|^2 = \beta R^2\right] | R\Big]\Bigg] \\
&\quad -\frac{1}{k+2}E\Big[\mathrm{E}\Big[\rho'(\beta R^2)(R^2-\beta R^2)^2\,\mathrm{E}_{R,\theta}\left[-2h(X) | \|X-\theta\|^2 = \beta R^2\right] | R\Big]\Big] \\
&= \mathrm{E}\Bigg[\mathrm{E}\Big[R^2\left(a\omega\rho'(0) + a(1-\omega)\rho'(\beta R^2)\right) \\
&\qquad \times \frac{(1-\beta)^2}{\beta}\,\mathrm{E}_{\theta}\left[-2\beta R^2 h(X) | \|X-\theta\|^2 = \beta R^2\right] | R\Big]\Bigg] \\
&\quad -\frac{1}{k+2}\mathrm{E}\Bigg[\mathrm{E}\Big[\beta R^2\left(-\rho'(\beta R^2)\right)\frac{(1-\beta)^2}{\beta}R^2 \\
&\qquad \times \mathrm{E}_{R,\theta}\left[-2h(X) | \|X-\theta\|^2 = \beta R^2\right] | R\Big]\Bigg].
\end{aligned} \tag{8.6}$$

We note that the function $(1-\beta)^2/\beta$ is decreasing and

$$\left(a\omega\rho'(0) + \rho'(\beta R^2)a(1-\omega)\right)\frac{(1-\beta)^2}{\beta}$$

is nondecreasing in $\beta$, since $\rho$ is concave. Also, for fixed $R$,

$$\mathrm{E}_\theta\left[-2\beta R^2 h(X)/\|X-\theta\| = \beta R^2\right]$$

is nonnegative and increasing in $\beta$. Using the covariance inequality and the fact that $\beta$ and $R$ are independent, the first expectation between the braces in Eq. 8.6 is bounded by

$$\begin{aligned}
&\mathrm{E}\Big[R^2\left(a\omega\rho'(0)+\rho'(\beta R^2)a(1-\omega)\right)\frac{(1-\beta)^2}{\beta} \\
&\qquad \mathrm{E}_{R,\theta}\left[-2\beta R^2 h(X)/\|X-\theta\|^2 = \beta R^2\right] \mid R\Big] \\
&\le \mathrm{E}\left[R^2\left(a\omega\rho'(0)+\rho'(\beta R^2)a(1-\omega)\right)\frac{(1-\beta)^2}{\beta}\right] \\
&\quad \times \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-2\beta R^2 h(X) \mid \|X-\theta\|^2 = \beta R^2\right] \mid R\right] \\
&\le \mathrm{E}\left[\beta R^2\left(a\omega\rho'(0)+\rho'(\beta R^2)a(1-\omega)\right)\right] \times \mathrm{E}\left[\frac{(1-\beta)^2}{\beta^2}\right] \\
&\quad \times \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-2\beta R^2 h(X) \mid \|X-\theta\|^2 = \beta R^2\right] \mid R\right] \\
&= \left(a\omega\rho'(0)\,\mathrm{E}\left[r^2\right] + a(1-\omega)\,\mathrm{E}\left[r^2\rho'(r^2)\right]\right) \times \mathrm{E}\left[\frac{(1-\beta)^2}{\beta^2}\right] \\
&\quad \times \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-2\beta R^2 h(X)/\|X-\theta\|^2 = \beta R^2\right] | R\right] \qquad (8.7)
\end{aligned}$$

In a similar way, from Lemma 8.2, for fixed $R$, $\mathrm{E}_\theta\left[-2h(X)/\|X-\theta\| = \beta R^2\right]$ is nonnegative and decreasing in $\beta$ and $-\beta R^2\rho'(\beta R^2)(1-\beta)^2$ is decreasing in $\beta$. It then follows from the covariance inequality and the independence of $\beta$ and $R$ that the expectation between the braces in the second expectation in Eq. (8.6) is bounded by

$$\begin{aligned}
&\frac{1}{k+2}\,\mathrm{E}\left[\beta R^2\left(-\rho'(\beta R^2)\right)\frac{(1-\beta)^2}{\beta}R^2\,\mathrm{E}_\theta\left[-2h(X)|\|X-\theta\|^2 = \beta R^2\right]|R\right] \\
&\le \frac{1}{k+2}\,\mathrm{E}\left[\beta R^2\left(-\rho'(\beta R^2)\right)\right] \\
&\quad \times \mathrm{E}\left[\frac{(1-\beta)^2}{\beta}R^2\,\mathrm{E}_{R,\theta}\left[-2h(X)|\|X-\theta\|^2 = \beta R^2\right]|R\right] \\
&= \frac{1}{k+2}\,\mathrm{E}\left[\beta R^2\left(-\rho'(\beta R^2)\right)\right]
\end{aligned}$$

$$\times \mathrm{E}\left[\frac{(1-\beta)^2}{\beta} R^2 \, \mathrm{E}_{R,\theta}\left[-2h(X)|\|X-\theta\|^2 = \beta R^2\right]|R\right]$$

$$\leq \frac{1}{k+2}\mathrm{E}\left[\beta R^2\left(-\rho'(\beta R^2)\right)\right]$$

$$\times \mathrm{E}\left[\frac{(1-\beta)^2}{\beta^2}\right] \times \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-2\beta R^2 h(X)|\|X-\theta\|^2 = \beta R^2\right]|R\right]$$

$$= -\frac{1}{k+2}\mathrm{E}\left[r^2\rho'(r^2)\right] \times \mathrm{E}\left[\frac{(1-\beta)^2}{\beta^2}\right] \times$$

$$\mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-2\beta R^2 h(X)|\|X-\theta\|^2 = \beta R^2\right]|R\right]. \tag{8.8}$$

By combining Eqs. (8.6)–(8.8), we obtain an upper bound on the risk difference:

$$\frac{\Delta\mathcal{R}(\theta)}{a(1-\omega)^2} \leq \left(a\omega\rho'(0)\,\mathrm{E}\left[r^2\right] + a(1-\omega)\,\mathrm{E}\left[r^2\rho'(r^2)\right] - \frac{1}{k+2}\mathrm{E}\left[r^2\rho'(r^2)\right]\right)$$
$$\times \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}_\theta\left[-2\beta R^2 h(X)/\|X-\theta\|^2 = \beta R^2\right]|R\right]\right]$$

Thus, a necessary condition for the risk difference to be negative is

$$a\omega\rho'(0)\,\mathrm{E}\left[r^2\right] + a(1-\omega)\,\mathrm{E}\left[r^2\rho'(r^2)\right] - \frac{1}{k+2}\mathrm{E}\left[r^2\rho'(r^2)\right] \leq 0$$

because $-h(\cdot)$ is a function that is positive. Hence, the proof is completed. □

The following theorem deals with the case where the vector $(X, U)$ has symmetric spherical distributions with density.

**Theorem 8.2** *Suppose that* $(X, U) \sim SS_{(d,k)}(\theta, 0)$ *with density of the form* $f(\|x-\theta\|^2 + \|u\|^2)$ *and* $\delta_{\omega,g}(X, U)$ *is defined by Eq.* (8.2)*. Then, with respect to the loss function (i),* $\delta_{\omega,g}(X, U)$ *has a smaller risk than* $\delta_0$ *provided the conditions of Theorem* 8.1 *hold and*

$$0 < a \leq \frac{K}{(k+2)(\omega\rho'(0) + K(1-\omega))},$$

*where* $K = \int_{\mathbb{R}^{d+k}} \rho'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2)\mathrm{d}(x, u)$.

***Proof*** As in the proof of Theorem 8.1, the risk difference $\Delta\mathcal{R}(\omega, \delta_0, \rho)/(a(1-\omega)^2)$ is bounded by

$$\mathrm{E}_\theta\left\{\omega\rho'(0)a^2(1-\omega)^2\|U\|^4\|g(X)\|^2 + (1-\omega)K\,\frac{\rho'(\|X-\theta\|^2)}{K}\right.$$

$$\times \left[2a(1-\omega)\|U\|^2 g(X).(X-\theta) + a^2(1-\omega)^2\|g(X)\|^2\|U\|^4\right]\Bigg\}$$

$$= \mathrm{E}_\theta\left[a\omega\rho'(0)\|U\|^4\|g(X)\|^2\right] + K\,\mathrm{E}^\star_\theta\left[2\|V\|^2 g(Y).(Y-\theta)\right]$$

$$+a(1-\omega)K\,\mathrm{E}^\star_\theta\left[\|g(Y)\|^2\|U\|^4\right] \tag{8.9}$$

where $\mathrm{E}^\star_\theta$ the expectation with respect to the density $f^\star$, where

$$f^\star(\|x-\theta\|^2 + \|u\|^2) = K^{-1}\rho'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2)$$

and $K = \int_{\mathbb{R}^{d+k}} \rho'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2) d(x,u)$.

In the second expectation of the right-hand side of the last equality, the function in this expectation depends on $\theta$. We can eliminate it by applying Lemma 8.5, as provided by Cellier and Fourdrinier (1995).

For a weakly differentiable function $g$, setting $q = 1$ in Eq. (8.13) in the Appendix yields

$$\mathrm{E}_\theta\left[\|V\|^2 g'(Y)(Y-\theta)\right] = \frac{1}{k+2}\,\mathrm{E}_\theta\left[\|V\|^4 \mathrm{div}(g(Y))\right].$$

Then, the quantity in Eq. (8.9) is bounded by

$$a\omega\rho'(0)^2\,\mathrm{E}_\theta\left[\|V\|^4\|g(Y)\|^2\right]$$

$$+K\,\mathrm{E}^\star_\theta\left[\left(\frac{2\|V\|^4}{k+2}\mathrm{div}(g(Y)) + a(1-\omega)\|g(Y)\|^2\|V\|^4\right)\right]$$

Since the function $g$ satisfies condition 1 of Theorem 8.1, then the risk difference $\Delta\mathcal{R}(\omega, \delta_0, \rho)/(a(1-\omega)^2)$ is bounded by

$$\mathrm{E}^\star_\theta\left[-2\|V\|^4 h(Y)\left(a(\omega\rho'(0) + K(1-\omega)) - \frac{K}{(k+2)}\right)\right];$$

hence, the proof is completed. □

## 8.3 Risk Analysis for Loss $\ell(\omega\|\delta-\delta_0\|^2 + (1-\omega)\|\delta-\theta\|^2)$

In this section, when the observation $(X, U) \sim SS_{(d,k)}(\theta, 0)$ (spherically symmetric distribution about $(\theta, 0)$), we present the result that states the dominance condition of estimators $\delta_{\omega,g}$ defined by Eq. (8.2) under the BLFs (ii). For the function $\ell$, we assume the following assumption.

**Assumption 8.2** *For the function $\ell$:*

*(i)* $\ell(\cdot) \geq 0$.
*(ii)* $\ell'(\cdot) > 0$.
*(iii) It is concave.*

Let $\Delta_{\omega,\ell,\delta_0}(\theta, \delta) = L_{\omega,\ell,\delta_0}(\theta, \delta) - L_{\omega,\ell,\delta_0}(\theta, \delta_0)$ be the difference in loss $L_{\omega,\ell,\delta_0}$ between estimates $\delta(X)$ and $\delta_0(X)$. We proceed with Lemma 8.1, which exploits the concavity of $\ell$ and relates the difference in losses $L_{\omega,\ell,\delta_0}$ between estimates $\delta(X) = \delta_0(X) + a(1-\omega)g(X)$ and $\delta_0(X)$, to the balanced squared-error loss difference $\Delta_{\omega,\ell,\delta_0}(\theta, \delta)$.

**Lemma 8.1** *Let $(X, U)$ have a spherically symmetric distribution with location parameter $(\theta, 0)$. For the problem of estimating $\theta$ under MBLFs (ii) with $\ell$ satisfying Assumption 8.2, we have*

$$\Delta_{\omega,\ell,\delta_0}(\theta, \delta) \leq (1-\omega)^2 \ell'((1-\omega)\|\delta_0 - \theta\|)\Delta_0(\theta, \delta).$$

***Proof*** The proof uses the fact that $\ell(a+b) - \ell(a) \leq b\ell'(a)$ since $\ell$ is concave, with $a = L_{\omega,\ell,\delta_0}(\theta, \delta_0)$ and $a + b = L_{\omega,\ell,\delta_0}(\theta, \delta_0 + (1-\omega)g)$. This yields

$$\begin{aligned}\Delta_{\omega,\ell,\delta_0}(\theta, \delta) &= L_{\omega,\ell,\delta_0}(\theta, \delta_0 + (1-\omega)g) - L_{\omega,\ell,\delta_0}(\theta, \delta_0) \\ &\leq (1-\omega)^2 \ell'((1-\omega)\|\delta_0 - \theta\|)\Delta_0(\theta, \delta)\end{aligned}$$

□

**Theorem 8.3** *Suppose that $(X, U) \sim SS_{(d,k)}(\theta, 0)$ with density of the form $f(\|x-\theta\|^2 + \|u\|^2)$ and $\delta_{\omega,g}(X, U)$ is defined by Eq. (8.2). Then, with respect to the MBLFs (ii), $\delta_{\omega,g}(X, U)$ has a smaller risk than $\delta_0$ provided condition 1 of Theorem 8.1 holds and $0 < a < ((k+2)(1-\omega))^{-1}$.*

***Proof*** Let $\mathrm{E}^*_\theta$ be the expectation with respect to the density $f^\star$, where

$$f^\star(\|x-\theta\|^2 + \|u\|^2) = L^{-1}\ell'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2)$$

and $L = \int_{\mathbb{R}^{d+k}} \ell'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2)\mathrm{d}(x, u)$.

Using the same approach as Brandwein and Strawderman (1980) and Marchand and Strawderman (2020), we calculate the difference between the risks of two estimators $\delta_0(X) = X$ and $\delta_{\omega,g}(X)$ with respect to the MBLFs (ii). Using Lemmas 8.5 and 8.1 as well as condition 1 of Theorem 8.1, respectively, we obtain

$$\begin{aligned}\Delta\mathcal{R}(\theta) &= \mathrm{E}_\theta\left[\Delta_{\theta,\delta_0,\ell}(\theta, \delta_{\omega,g})\right] \\ &\leq \mathrm{E}_\theta\left[(1-\omega)^2\ell'((1-\omega)\|\delta_0 - \theta\|^2)\Delta_0(\theta, \delta_{\omega,g})\right] \\ &= (1-\omega)^2\,\mathrm{E}_\theta\left[\ell'((1-\omega)\|X-\theta\|^2)(\|\delta_{\omega,g}(X) - \theta\|^2 - \|X-\theta\|^2)\right]\end{aligned}$$

$$
\begin{aligned}
&= (1-\omega)^2\,\mathrm{E}_\theta\Big[\ell'((1-\omega)\|X-\theta\|^2)\\
&\qquad\qquad \times\Big(\|X+a(1-\omega)\|U\|^2 g(X)-\theta\|^2-\|X-\theta\|^2\Big)\Big]\\
&= a(1-\omega)^3 L\left\{\mathrm{E}^\star_\theta\left[a(1-\omega)\|V\|^4\|g(Z)\|^2\right]+2\,\mathrm{E}^\star_\theta\left[\|V\|^2 g(Z).(Z-\theta)\right]\right\}\\
&= a(1-\omega)^3\ L\,\mathrm{E}^\star_\theta\left[a(1-\omega)\|V\|^4\|g(Z)\|^2+\frac{2\|V\|^4}{k+2}\mathrm{div}(g(Z))\right]\\
&\le a\,(1-\omega)^3 L\,\mathrm{E}^\star_\theta\left[-2a(1-\omega)\|V\|^4 h(Z)+\frac{2\|V\|^4}{k+2}h(Z)\right]\\
&= a(1-\omega)^3 L\,\mathrm{E}^\star_\theta\left[-2\|V\|^4 h(Z)(a(1-\omega))-\frac{1}{k+2}\right].
\end{aligned}
$$

The result follows immediately from the hypothesis $h(\cdot)<0$. □

**Theorem 8.4** *Suppose that $(X,U)$ have a spherically symmetric distribution with location parameter $(\theta,0)$ and $\delta_{\omega,g}(X)$ is defined by Eq. (8.2). Then, with respect the loss MBLFs (ii), $\delta_{\omega,g}(X)$ has smaller risk than $\delta_0(X)=X$ provided condition 1 of Theorem 8.1 holds and $0<a<((k+2)(1-\omega))^{-1}$.*

***Proof*** In the same way as in the proof of Theorem 8.3, using the fact that the function $\ell(\cdot)$ is concave, we show that the risk difference $\Delta\mathcal{R}(\theta)$ is bounded by

$$
\begin{aligned}
&a(1-\omega)^3\,\mathrm{E}_\theta\Big[\ell'((1-\omega)\|X-\theta\|^2)\\
&\quad\times\Big(2\|U\|^2 g'(X)(X-\theta)+a(1-\omega)\|U\|^4\|g(X)\|^2\Big)\Big]\\
&= a(1-\omega)^3\,\mathrm{E}_\theta\Big[\ell'((1-\omega)(R^2-\|U\|^2))\\
&\quad\times\Big(2\|U\|^2 g(X)'(X-\theta)+a(1-\omega)\|U\|^4\|g(X)\|^2\Big)\Big]\\
&= a(1-\omega)^3\Big(\mathrm{E}_\theta\Big[2\|U\|^2\ell'((1-\omega)(R^2-\|U\|^2))g(X)'(X-\theta)\Big] \quad (8.10)\\
&\quad+\mathrm{E}_\theta\Big[\ell'((1-\omega)(R^2-\|U\|^2))a(1-\omega)(\|U\|^2)^2\|g(X)\|^2\Big]\Big)
\end{aligned}
$$

Using Lemma 8.6, for the function $m(t)=\ell'((1-\omega)(R^2-t))t$ and its corresponding function $M(t)=-\frac{1}{2}\int_0^t m(t)t^{(k-2)/2}\,\mathrm{d}t$. By integration by parts, we have

$$
\begin{aligned}
M(t) &= -\frac{1}{2}\int_0^t \ell'((1-\omega)(R^2-t))t^{\frac{k}{2}}\mathrm{d}t\\
&= -\ell'((1-\omega)(R^2-t))\frac{t^{\frac{k+2}{2}}}{k+2}-(1-\omega)\int_0^t \ell''((1-\omega)(R^2-t))\frac{t^{\frac{k+2}{2}}}{k+2}\mathrm{d}t
\end{aligned}
$$

$$\geq -\ell'((1-\omega)(R^2-t))\frac{t^{\frac{k+2}{2}}}{k+2},$$

since the function $\ell(\cdot)$ is concave then (i.e., $\ell''(t) \leq 0$ for any $t \geq 0$).

Conditioning with respect to the radius $R$, the first expectation in Eq. (8.10) becomes

$$\begin{aligned}
&\mathrm{E}_{R,\theta}\left[\ell'((1-\omega)(R^2-\|U\|^2))\|U\|^2 g(X).(X-\theta)\right] \\
&= \mathrm{E}_{R,\theta}\left[m\left(\|U\|^2\right) g(X)\,(X-\theta)\right] \\
&= -\mathrm{E}_{R,\theta}\left[\frac{M(\|U\|^2)}{(\|U\|^2)^{\frac{k}{2}-1}}\mathrm{div}(g(X))\right] \\
&\leq \mathrm{E}_{R,\theta}\left[\ell'((1-\omega)(R^2-\|U\|^2))\frac{\|U\|^{k+2}}{(k+2)\|U\|^{k-2}}\mathrm{div}(g(X))\right] \\
&= \mathrm{E}_{R,\theta}\left[\ell'((1-\omega)(R^2-\|U\|^2))\frac{\|U\|^4}{k+2}\mathrm{div}(g(X))\right] \qquad (8.11)
\end{aligned}$$

Combining condition 1 of Theorem 8.1 and Eqs. (8.10)–(8.11), an upper bound on the difference between the risk $\Delta\mathcal{R}(\theta)/(a(1-\omega)^3)$ is given by

$$\begin{aligned}
&= \mathrm{E}_\theta\left[2\ell'((1-\omega)(R^2-\|U\|^2))\frac{\|U\|^4}{k+2}\mathrm{div}(g(X))\right. \\
&\qquad \left.+\ell'((1-\omega)(R^2-\|U\|^2))\|U\|^4(1-\omega)\|g(X)\|^2\right] \\
&\leq \mathrm{E}_\theta\left[2\ell'((1-\omega)(R^2-\|U\|^2))\|U\|^4\frac{1}{k+2}h(X)\right. \\
&\qquad \left.+\ell'((1-\omega)(R^2-\|U\|^2))\|U\|^4 a(1-\omega)(-2h(X))\right] \\
&= \mathrm{E}_\theta\left[-2\ell'((1-\omega)(R^2-\|U\|^2))\|U\|^4 h(X)\left(a(1-\omega)-\frac{1}{k+2}\right)\right].
\end{aligned}$$

The result follows immediately from the assumption that both functions $\ell'(\cdot)$ and $-h(\cdot)$ are positive. This implies the desired result. □

**Theorem 8.5** *Suppose that the vector $(X, U)$ has a spherically symmetric distribution with location parameter $(\theta, 0)$ and $\delta_{\omega,g}(X)$ is defined by 8.2. Assume that the concave and nondecreasing function $\ell(t)$ also satisfies the property that $t^\alpha \ell(t)$ is nondecreasing for $\alpha \geq 0$. Then, with respect the loss MBLFs (ii), $\delta_{\omega,g}(X)$ has smaller risk than $\delta_0(X) = X$ provided conditions 1 and 2 of Theorem 8.1 hold and $0 < a \leq (d-2-2\alpha)/(d(k+2)(1-\omega))$.*

***Proof*** Let $r^2 = \|X-\theta\|^2$ and $R^2 = \|X-\theta\|^2 + \|U\|^2$. From Eq. (8.10) and Lemma 8.7, we have

$$\begin{aligned}
\frac{\Delta\mathcal{R}(\theta)}{(1-\omega)^3} \le\\
&\mathrm{E}_\theta\left[\ell'((1-\omega)\|X-\theta\|^2)\left(\|X+a(1-\omega)\|U\|^2 g(X)-\theta\|^2-\|X-\theta\|^2\right)\right]\\
&= \mathrm{E}\Big[\mathrm{E}\left[\ell'((1-\omega)r^2)\left(2a^2(1-\omega)\|U\|^2/r^2 - 2a\|U\|^2/d\right)\right.\\
&\qquad\left.\times\,\mathrm{E}_\theta\left[-r^2h(W)\mid\|X-\theta\|^2=r^2\right]\mid R,r\right]\Big]\\
&\le \mathrm{E}\Big[\mathrm{E}\left[\ell'((1-\omega)\beta R^2)\left(2a^2(1-\omega)\frac{((1-\beta)R^2)^2}{\beta R^2} - \frac{2a(1-\beta)R^2}{d}\right)\right.\\
&\qquad\left.\times\,\mathrm{E}_\theta\left[-\beta R^2h(W)\mid\|X-\theta\|^2=\beta R^2\right]\mid R\right]\Big]\\
&= 2a\,\mathrm{E}\Big[\mathrm{E}\Big[R^2(\beta R^2)^\alpha\ell'((1-\omega)\beta R^2)\frac{1-\beta}{(\beta R^2)^\alpha}\\
&\quad\times\left(\frac{a(1-\omega)(1-\beta)}{\beta}-\frac{1}{d}\right)\mathrm{E}_\theta\left[-\beta R^2h(W)\mid\|X-\theta\|^2=\beta R^2\right]\mid R\Big]\Big]\\
&= 2a\,\mathrm{E}\left[\mathrm{E}\left[\left\{\frac{1-\beta}{\beta^\alpha}\left(\frac{a(1-\omega)(1-\beta)}{\beta}-\frac{1}{d}\right)\right\}\right.\right.\\
&\left.\left.\times\left\{R^{2-2\alpha}(\beta R^2)^\alpha\ell'((1-\omega)\beta R^2)\,\mathrm{E}\left[-\beta R^2h(W)\mid\|X-\theta\|^2=\beta R^2\right]\right\}\mid R\right]\right]
\end{aligned}$$

Next, for fixed $R$, both functions

$$\beta \to \frac{1-\beta}{\beta^\alpha}\left(\frac{a(1-\omega)(1-\beta)}{\beta}-\frac{1}{d}\right)$$

and

$$\beta \to (\beta R^2)^\alpha\ell'((1-\omega)\beta R^2)\,\mathrm{E}\left[-\beta R^2h(W)\mid\|X-\theta\|^2=\beta R^2\right]$$

are nonnegative and nondecreasing in $\beta$. Hence, it follows from the covariance inequality that the previous expectation is less than

$$\begin{aligned}
&2a\,\mathrm{E}\Big[\mathrm{E}\left(\mathrm{E}_{R,\theta}\left[-\beta R^2h(W)\mid\|X-\theta\|^2=\beta R^2\right]R^{2-2\alpha}\right.\\
&\qquad\left.(\beta R^2)^\alpha\ell'((1-\omega)\beta R^2)\mid R\right)\Big]\\
&\times\,\mathrm{E}\left[a(1-\omega)\beta^{-\alpha}(1-\beta)^2/\beta-\frac{\beta^{-\alpha}(1-\beta)}{d}\right].
\end{aligned}$$

Since the first expectation of this term is not negative, it is sufficient that the second expectation is negative for the last equation to be negative. Indeed, by Lemma 8.8, the random variable $\beta$ has the distribution Beta $(d/2, k/2)$, so the last expectation is negative if and only if

$$0 < a \leq \frac{1}{d(1-\omega)} \frac{\mathrm{E}[(\beta^{-\alpha}(1-\beta))]}{\mathrm{E}[(\beta^{-\alpha}(1-\beta)^2)/\beta]} = \frac{d-2-2\alpha}{(1-\omega)(k+2)d};$$

hence, the proof is completed. □

## 8.4 Numerical Illustration

In this section, we present some simulations of the theoretical results of Sect. 8.3. We will consider the Baranchik-type estimators and deal with two models, the first with density and the second without density. Before starting the simulation in Python 3.12, the Baranchik-type estimators defined in Eq. (8.1) dominate the estimator $\delta_0(X) = X$ if the parameter $a$ verifies

$$0 < a < 2\frac{d-2}{((k+2)(1-\omega))}$$

using

$$\mathrm{div}(g(X)) = -2aS'(\|X\|^2) - a(d-2)S(\|X\|^2)/\|X\|^2 \leq -a(d-2)S(\|X\|^2)/\|X\|^2$$

and $\|g(X)\|^2 = a^2S^2(\|X\|^2)/\|X\|^2 \leq a^2S(\|X\|^2)/\|X\|^2$. In this simulation, we consider $\ell(t) = t^q$, $q = 1/2$, $S(t) = t/(1+t)$, $\omega = 1/3$, $d = 20$, and $k = 10$. We get the cut-off point $a_0 = 2(d-2)/((k+2)(1-\omega)) = 4.5$.

### *8.4.1 Example 1*

In this example, we assume that $(X, U) \in \mathbb{R}^{d+k}$ follows the normal distribution with identity covariance matrix. Figure 8.1 gives, for different values of $a$, the risk of the estimator $\delta_0$ and that of the Baranchik-type estimators $\delta_{\omega,a}(X, \|U\|^2)$ as a function of the norm of $\theta$ under the MBLFs (ii) where $\ell(t) = t^{1/2}$. The estimator $\delta_{\omega,0}(X, \|U\|^2)$ is minimax at constant risk $\mathcal{R}(\theta, \ell, \delta_0) = \mathrm{E}_0((1-\omega)\|X\|^2) \approx 3.605$, which can be obtained using the fact that $\|X-\theta\|^2 \sim \chi^2_d$. We note that the Baranchik-type estimators dominate the $\delta_0$ estimator if the parameter $a$ is less than 4.5.

The simulation results indicate that for values of the parameter $a$ that are less than the $a_0$ (i.e., $a = 1, 2, 3, 3.5$), the risk curve of the Baranchik-type estimator

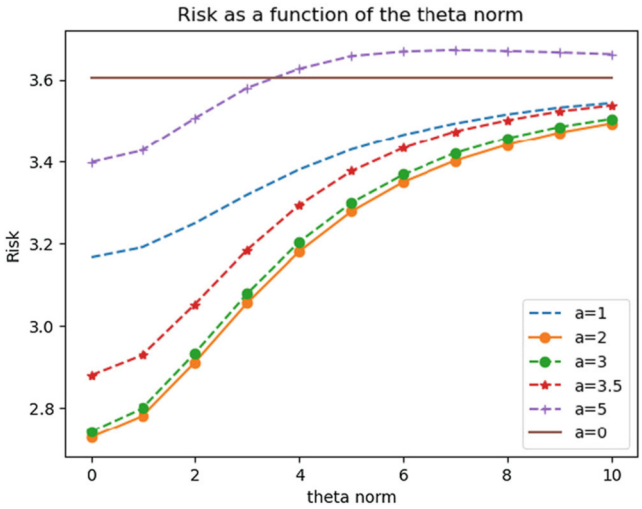



**Fig. 8.1** The risk $\mathcal{R}(\theta, \ell, \delta_{\omega,a}(X, \|U\|^2)$ as a function of the parameter $a$

$\delta_{\omega,a}(X, \|U\|^2)$ is lower than that of the usual estimator $\delta_{\omega,0}(X, \|U\|^2)$. However, once the value of $a$ ($a = 5$) surpasses the bound $a_0$, the risk curve of the Baranchik-type estimator rises above that of the usual estimator, showing the theoretical result of dominance.

### 8.4.2 *Example 2*

We will first state the following theorem, which characterizes random vectors having a spherically symmetric distribution. This characterization allows us to simulate the random vector $(X, U) \sim SS_{(d,k)}(\theta, 0)$.

**Theorem 8.6 (Fang et al. 1990)** *If the $n$-random vector $X$ has a spherically symmetric distribution around 0, partition $X$ into $m$ parts $X = ((X^{(1)})^\top, \ldots, (X^{(m)})^\top)^\top$ each with $n_1, \ldots, n_m$ components, respectively, then $((X^{(1)})^\top, \ldots, (X^{(m)})^\top)^\top$ possesses the same distribution as $(rd_1U_1^\top, \ldots, rd_m U_m^\top)^\top$ where $r$ has the same distribution as $\|X\|$, for $i = 1, \ldots, m$, $d_i \geq 0$, and $(d_1^2, \ldots, d_m^2)$ has the multivariate Dirichlet distribution $D_m(n_1/2, \ldots, n_m/2)$, $U_1, \cdots, U_m$ and $(d_1^2, \ldots, d_m^2)$ are independent and $U_j$ follows the uniform distribution on the unit sphere of dimension $n_j$, $j = 1, \ldots, m$.*

In this example, we consider a scenario in which a spherically symmetric vector does not necessarily have a probability density function. We start by considering a real variable $R$ that adheres to a Poisson distribution with parameter $\lambda = 10$. Then,

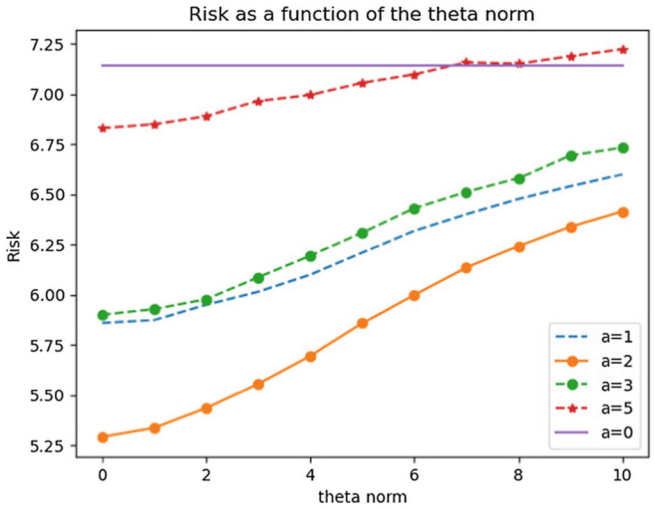



**Fig. 8.2** The risk $\mathcal{R}(\theta, \ell, \delta_{\omega,a}(X, \|U\|^2)$ as a function of the parameter $a$

we introduce a random vector $D = (D_1, D_2)$ that follows a Dirichlet distribution with parameters $d/2$ and $k/2$, i.e., $D(d/2,\ k/2)$.

Using the normal distribution, we simulate two random vectors $W_1$, which is uniformly distributed on the unit sphere in $\mathbb{R}^d$, and $W_2$, which is distributed on the unit sphere in $\mathbb{R}^k$. We then define the vectors $X$ and $U$ as follows: $X = R\sqrt{D_1}W_1+\theta$ and $U = R\sqrt{D_2}W_2$. By Theorem 8.6, the vector $(X, U)$ is characterized by a spherically symmetric distribution centered around $(\theta, 0)$. Figure 8.2 shows the variation of the risk of the estimator $\delta_{\omega,a}(X, \|U\|^2)$ as a function of the norm of $\theta$ for different values of the parameter $a$ under the MBLFs (ii), clearly showing that the estimator $\delta_{\omega,a}(X, \|U\|^2)$ dominates the usual estimator $\delta_{\omega,0}(X, \|U\|^2) = \delta_0(X) = X$ for values of below the cut-off point $a_0 = 4.5$.

Figure 8.2 is generated by simulating the data using Python 3.12, and we determined that the constant risk $\mathcal{R}(\theta, \ell, \delta_0) = \mathrm{E}_0((1-\omega)\|X\|^2) \approx 7.14$. As long as the parameter value $a$ is slightly below $a_0$ ($a = 1, 2, 3$), the risk curve of the Baranchik-type estimator stays below the risk curve of the benchmark estimator $\delta_{\omega,0}(X, \|U\|^2)$. However, if $a$ surpasses $a_0$ ($a = 5$), the risk curve of the Baranchik-type estimator exceeds that of the benchmark estimator, thus confirming the theoretical result obtained.

## 8.5 Concluding Remarks

For spherically symmetric distributions of the form $(X, U)$, we provide frequentist improvements of the risk on the reference estimator $\delta_0(X) = X$ associated with the balanced loss functions (i) and (ii). The results apply to a large class of Baranchk-

type estimators $\delta_{w,a}(X, \|U\|^2)$ with respect to the choices of $\rho$ in (i) and $\ell$ in (ii), as well as with respect to the weight $\omega$ in these losses. The results extend earlier results of Marchand and Strawderman (2020) that apply to mixtures of normal scales and completely monotone $\rho'$ or $\ell'$ and of Hobbad et al. (2021) that apply spherically symmetric distributions with density $f(\|X-\theta\|^2)$, while the results here apply to random observation of the form $(X, U)$ that is spherically symmetrically distributed around a vector of the form $(\theta, 0)$, as well as to choices of $\rho$ and $\ell$ that are monotone increasing and concave.

## Appendix

In this section, we review and establish several lemmas essential for this chapter.

**Lemma 8.2 (Du Plessis 1970)** *If $h$ is a superharmonic function on $\mathbb{R}^p$ and $Z$ is a random variable with a uniform distribution on the sphere centered at the origin and of radius $\tau$, then, for any $\theta \in \mathbb{R}^p$, the expectation $\mathrm{E}_\theta\,[h(\theta+Z)]$ is a nonincreasing function of $\tau$.*

**Lemma 8.3** *Give $X$ with a spherically symmetric distribution around $\theta$ and $\beta : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ a nondecreasing function. Then, $\mathrm{E}\left[R^2 \dfrac{\beta(\|X\|^2)}{\|X\|^2} \middle| \|X-\theta\| = R\right]$ is a nondecreasing function of $R > 0$.*

***Proof*** See Brandwein and Strawderman (1979, 1980), pp. 394–395, within the proof of their Theorem 3.3.1. □

**Lemma 8.4** *Let $(X, U)$ have a spherically symmetric distribution around $(\theta, 0)$ with the density of the form $f(\|x-\theta\|^2 + \|u\|^2)$, with $\theta \in \mathbb{R}^d$ and $\rho$ satisfying Assumption 8.1. Then, for $d \geq 4$,*

$$\mathrm{E}_\theta\left[\rho\left(-\|U\|^2 h(X)\right)\right] \leq \rho'(0)\,\mathrm{E}_\theta\left[-\|U\|^2 h(X)\right] \leq \rho'(0)\,\mathrm{E}_\theta\left[-\|V\|^2 h(Y)\right], \tag{8.12}$$

*where $(Y, V) \sim f^*(\|y-\theta\|^2 + \|u\|^2)$, with*

$$f^\star(x,u) = \left(\rho'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2)\right)/K$$

*where $K = \int_{\mathbb{R}^{d+k}} \rho'((1-\omega)\|x-\theta\|^2) f(\|x-\theta\|^2 + \|u\|^2)\mathrm{d}(x,u)$.*

***Proof*** The first inequality of Eq. (8.12) follows from $\rho(t) \leq \rho(0) + \rho'(0)t = \rho'(0)t$, which is true because $\rho$ satisfies Assumption 8.1. Let $r^2 = \|X-\theta\|^2$, $R^2 = \|X-\theta\|^2 + \|U\|^2$ and note that the conditional distribution of $X$ given $r$ and $R$ is $SS_d(\theta)$.

Then, the random variable $\beta = \|X-\theta\|^2/(\|X-\theta\|^2+\|U\|^2)$ has a Beta$(d/2, k/2)$ distribution, independent of $R^2 = \|X-\theta\|^2 + \|U\|^2$, and we have

$$\begin{aligned}\mathrm{E}_\theta\left[-\|U\|^2 h(X)\right] &= \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-\|U\|^2 h(X) \mid r, R\right]\right] \\ &= \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[-(R^2-r^2)h(X) \,\middle|\, r, R\right]\right]\end{aligned}$$

Note that $\beta R^2 = r^2$, so

$$\begin{aligned}&= \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[\frac{\rho'(\beta R^2)}{K}\,\mathrm{E}\left[-(1-\beta)R^2 h(X) \mid \|X-\theta\|^2 = \beta R^2\right]\frac{K}{\rho'(\beta R^2)}\right]\Big|R\right] \\ &= \mathrm{E}\left[\mathrm{E}_{R,\theta}\left[\frac{\rho'(\beta R^2)}{K}\,\mathrm{E}\left[-\|U\|^2 h(X) \mid \|X-\theta\|^2 = \beta R^2\right]\frac{K}{\rho'(\beta R^2)}\right] \mid R\right] \\ &= \mathrm{E}^*\left[\mathrm{E}^*\left[\mathrm{E}\left[-\|V\|^2 h(Y) \mid \|Y-\theta\|^2 = \beta R^2\right]\frac{K}{\rho'(\beta R^2)}\right] \mid R\right].\end{aligned}$$

For fixed $R$, from Lemma 8.3, the function $\beta \longrightarrow \mathrm{E}_\theta\left[-\|V\|^2 h(Y)/\|Y-\theta\| = \beta R^2\right]$ is decreasing and $\beta \longrightarrow 1/\rho'(\beta R^2)$ is increasing, so using the covariance inequality, we get

$$\begin{aligned}\mathrm{E}_\theta\left[-\|U\|^2 h(X)\right] &\le \mathrm{E}^\star\left[\mathrm{E}\left(-\|V\|^2 h(Y) \mid \|Y-\theta\|^2 = \beta R^2\right) \mid R\right] \\ &\quad \times \mathrm{E}^\star\left[\mathrm{E}^\star\left[\frac{K}{\rho'(\beta R^2)}\right] \mid R\right] \\ &= \mathrm{E}_\theta\left[-\|V\|^2 h(Y)\right].\end{aligned}$$

Hence, the proof is completed. □

**Lemma 8.5** *For every weakly differentiable function $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, for every integer $q$, and for every $\theta \in \mathbb{R}^d$, we have*

$$\mathrm{E}_\theta\left[(\|V\|^2)^q\, g'(Y)\,(Y-\theta)\right] = \frac{1}{k+2q}\,\mathrm{E}_\theta\left[(\|V\|^2)^{q+1}\mathrm{div}(g(Y))\right], \tag{8.13}$$

*provided these expectation exist.*

**Lemma 8.6 (Cellier and Fourdrinier, 1995)** *For every weakly differentiable function $g : \mathbb{R}^d \to \mathbb{R}^d$ and for every real valued function $m$, then, for any $\theta \in \mathbb{R}^d$,*

$$\mathrm{E}_{R,\theta}\left[m(\|U\|^2)g'(X)(X-\theta)\right] = -\,\mathrm{E}_{R,\theta}\left[\frac{M(\|U\|^2)}{(\|U\|^2)^{\frac{k}{2}-1}}\mathrm{div}(g(X))\right],$$

*provided these expectation exist, where $M$ is the indefinite integral, vanishing at* 0, *of the function* $t \longrightarrow -2^{-1}h(t)t^{\frac{k}{2}-1}$.

**Lemma 8.7** *For* $d \geq 4$, *let* $X \sim SS_d(\theta)$ *and also* $g(X)$ *map* $\mathbb{R}^d$ *into* $\mathbb{R}^d$ *be weakly differentiable and such that:*

1. $\|g(X)\|^2/2 \leq -h(X) \leq -\mathrm{div}(g(X))$
2. $-h(X)$ *is superharmonic and* $\mathrm{E}_\theta\left[r^2 h(W) \mid r\right]$ *is a non-decreasing function of* $r$,

*where* $W$ *has a uniform distribution on the sphere of radius* $r$ *centered at* $\theta$. *Then,*

$$\begin{aligned}&\mathrm{E}_\theta\left[\|X + ag(X) - \theta\|^2 - \|X - \theta\|^2\right] \\ &\quad \leq \mathrm{E}_\theta\left[\left(-2\frac{a^2}{r^2} + 2a\frac{1}{d}\right)\mathrm{E}_{r,\theta}\left[r^2 h(W) \mid \|X - \theta\|^2 = r^2\right]\right] \end{aligned} \tag{8.14}$$

***Proof*** We have

$$\begin{aligned}\mathrm{E}_\theta\left[\|X + ag(X) - \theta\|^2 - \|X - \theta\|^2\right] &= \mathrm{E}_\theta\left[2a^2\|g(X)\|^2/2 + 2a(X-\theta)^\top g(X)\right] \\ &\leq \mathrm{E}_\theta\left[-2a^2 h(X)\right] \\ &\quad +2a\left[\frac{1}{d}\int_{\mathbb{R}^+} r^2 \int_{B_{r,\theta}} \nabla g(x)\mathrm{d}V_{r,\theta}(x)\mathrm{d}\rho(r)\right] \\ &\leq \mathrm{E}_\theta\left[-2a^2 h(X)\right] \\ &\quad +2a\left[\frac{1}{d}\int_{\mathbb{R}^+} r^2 \int_{B_{r,\theta}} h(x)\mathrm{d}V_{r,\theta}(x)\mathrm{d}\rho(r)\right]\end{aligned}$$

By subharmonicity of $h$, we have

$$\int_{B_{r,\theta}} h(x)dV_{r,\theta}(x)\mathrm{d}\rho(r) \leq \int_{S_{r,\theta}} h(x)dU_{r,\theta}(x)\mathrm{d}\rho(r).$$

Hence, Eq. (8.14) is satisfied and the proof is completed. □

**Lemma 8.8 (Muirhead 1982)** *If* $(Z, U)$ *has a spherically symmetric distribution about* $(0, 0)$ *when* $\dim Z = d$ *and* $\dim(U) = k$, *then the distribution of* $Z^\top Z/\left(Z^\top Z + \|U\|^2\right)$ *is Beta* $(d/2,\ k/2)$ *and is independent of* $Z^\top Z + \|U\|^2$.

**Acknowledgments** The authors thank the editors and reviewers for their very helpful and constructive comments.

## References


Baranchik, A. J. (1970). A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Annals of Mathematical Statistics, 41*(2), 642–645.

Brandwein, A. & Strawderman, W. E. (1979). Minimax estimation of mean of spherically symmetric distributions under general quadratic loss. *Journal of Multivariate Analysis, 9*(2), 579–588.

Brandwein, A. & Strawderman, W. E. (1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *Annals of Statistics, 8*(2), 279–284.

Brandwein, A. & Strawderman, W. E. (1991). Generalizations of James-Stein estimators under spherical symmetry. *Annals of Statistics, 19*(3), 639.

Brandwein, A., Ralescu, S., & Strawderman, W. E. (1993). Shrinkage estimators of the location parameter for certain spherically symmetric distributions. *Annals of the Institute of Statistical Mathematics, 45*(3), 551–565.

Cellier, D., Fourdrinier, D., & Robert, C. (1989). Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. *Journal of Multivariate Analysis, 138*(12), 29–52.

Cellier, D., & Fourdrinier, D. (1995). Shrinkage estimators under spherical symmetry for the general linear model. *Journal of Multivariate Analysis, 52*(2), 338–351.

Chaturvedi, A. & Shalabh (2004). Risk and Pitman closeness properties of feasible generalized double k-class estimators in linear regression models with non-spherical disturbances under balanced loss function. *Journal of Multivariate Analysis, 90*(2), 229–256.

Dey, D., Ghosh, M., & Strawderman, W. E. (1999). On estimation with balanced loss functions. *Statistics Probability Letters, 45*(2), 97–101.

Du Plessis, N. (1970). *An introduction to potential theory*. Wiley-Interscience.

Fang, K.-T., Kotz, S., & Ng, K. W. (1990). *Symmetrie multivariate and related distributions*. Springer Science + Business Media, B.V.

Fourdrinier, D. & Ouassou, I. (2000). Estimation of the mean of a spherically symmetric distribution with constraints on the norm. *The Canadian Journal of Statistics, 28*(2), 399–415.

Fourdrinier, D. & Strawderman, W. E. (1996). A paradox concerning shrinkage estimators: should a known scale parameter be replaced by an estimated value in the shrinkage factor? *Journal of Multivariate Analysis, 59*(2), 109–140.

Fourdrinier, D., Ouassou, I., & Strawderman, W. E. (2003). Estimation of a parameter vector when some components are restricted. *Journal of Multivariate Analysis, 86*(1), 14–27.

Fourdrinier, D., Ouassou, I., & Strawderman, W. E. (2008). Estimation of a mean vector under quartic loss. *Journal of Statistical Planning and Inference, 138*(12), 3841–3841.

Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix, *Journal of Multivariate Analysis, 85*(1), 24–39.

Giles, J. A., Giles, D. E. A., & Ohtani, K. (1996). The exact risks of some pre-test and Stein-type regression estimates under balanced loss. *Communications in Statistics—Theory and Methods, 25*(12), 2901–2924.

Hobbad, L., Marchand, E., & Ouassou, I. (2021). On shrinkage estimation of a spherically symmetric distribution for balanced loss functions. *Journal of Multivariate Analysis, 186*, 104794.

James, W. & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 361–379.

Jafari Jozani, M., Marchand, E., & Parsian, A. (2006). On estimation with weighted balanced-type loss function. *Statistics & Probability Letters, 76*(8), 773–780.

Marchand, E. & Strawderman, W. E. (2020). On shrinkage estimation for balanced loss functions. *Journal of Multivariate Analysis, 175*, 104558.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley.

Zellner, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions. In: S. S. Gupta, J. O. Berger (Eds.), *Statistical Decision Theory and Related Topics* (pp. 371–390). Springer.

# Chapter 9
# Shrinkage Strategies and Superefficiency




**Paul Kabaila**

## 9.1 Introduction

Suppose that the probability distribution of the sample $y_1, \ldots, y_n$ is determined by the unknown $r$-dimensional parameter $\boldsymbol{\lambda}$. Let $\boldsymbol{I_n}(\boldsymbol{\lambda})$ denote the information matrix for this sample and suppose that $n^{-1}\boldsymbol{I_n}(\boldsymbol{\lambda})$ converges to the nonsingular matrix $\boldsymbol{I}(\boldsymbol{\lambda})$, as $n \to \infty$. Let $\hat{\boldsymbol{\lambda}}_n$ denote the maximum likelihood estimator, based on $y_1, \ldots, y_n$, of $\boldsymbol{\lambda}$. Under regularity conditions, for each $\boldsymbol{\lambda}$, $n^{1/2}\left(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}\right)$ converges in distribution to $N\left(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\lambda})\right)$, as $n \to \infty$.

Let $\tilde{\boldsymbol{\lambda}}_n$ denote an estimator, based on $y_1, \ldots, y_n$, of $\boldsymbol{\lambda}$. Suppose that, for each $\boldsymbol{\lambda}$, $n^{1/2}\left(\tilde{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}\right)$ converges in distribution to $N\left(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\lambda})\right)$, as $n \to \infty$. Beginning with R. A. Fisher, it was a long-held belief that

$$\boldsymbol{V}(\boldsymbol{\lambda}) - \boldsymbol{I}^{-1}(\boldsymbol{\lambda}) \text{ is positive semidefinite} \tag{9.1}$$

for all $\boldsymbol{\lambda}$. This belief was shattered by Hodges's example, first published by Le Cam (1953). In this example, $y_1, \ldots, y_n$ are independent and identically $N(\lambda, 1)$ distributed and Hodges's estimator of $\lambda$ does not satisfy in Eq. (9.1) at $\lambda = 0$. In the general case, the values of $\boldsymbol{\lambda}$ at which Eq. (9.1) is not satisfied are called points of superefficiency. As proved by Le Cam (1953) and Bahadur (1964), the set of points $\boldsymbol{\lambda}$ for which Eq. (9.1) fails is a set of Lebesgue measure zero. If $\boldsymbol{I}(\boldsymbol{\lambda})$ is continuous, as is commonly the case, then continuity of $\boldsymbol{V}(\boldsymbol{\lambda})$ guarantees that Eq. (9.1) is satisfied for all $\boldsymbol{\lambda}$. These results have obvious implications for the case that the parameter of interest is a scalar $\theta$ that is some specified smooth function of $\boldsymbol{\lambda}$.

P. Kabaila (✉)
Department of Mathematical and Physical Sciences, La Trobe University, Bundoora, VIC, Australia
e-mail: P.Kabaila@latrobe.edu.au

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_9

Hodges's superefficient estimator is a shrinkage estimator that, at first sight, appears to be better than the usual estimator, the sample mean, when $n$ is large. However, upon careful examination, this turns out to be an illusion, as is made clear by, e.g., Example 8.1 of van der Vaart (1998). In fact, the performance of this estimator is actually much worse than the sample mean when $n$ is large. Kabaila (1996) provided an illustration of the fact that estimators constructed after consistent preliminary model selection, such as BIC, are also superefficient. These estimators may be viewed as being obtained by shrinkage.

Kabaila (1995) described a confidence interval analogue of Hodges's superefficient estimator (see also Example 3 of Beran, 1992). The author also described a prediction interval analogue of this estimator. In summary, Kabaila (1995) described how to extend the concept of superefficiency to confidence intervals and prediction intervals. In addition, the author showed that confidence intervals and prediction intervals constructed after consistent preliminary model selection display a superefficiency phenomenon. At first sight, these intervals appear to be extremely good when the sample size is large. However, this turns out to be an illusion, since the performances of these intervals are actually much worse than the usual intervals when the sample size is large. The source of this illusion has been discussed in Section 6 of Kabaila (2009). These intervals may also be viewed as being obtained by shrinkage.

The lasso and other similar estimators, such as SCAD, are shrinkage estimators. Under conditions that make these estimators satisfy the sparsity property, Leeb and Pötscher (2008) have shown that, in the context of two nested linear regression models, these are superefficient. Leeb and Pötscher (2005) and Pötscher (2009) have provided examples of confidence sets based on the lasso and other similar estimators that satisfy the sparsity property and display a superefficiency phenomenon.

The frequentist model averaged estimators proposed by Buckland et al. (1997) are also shrinkage estimators. We consider the case of two nested linear regression models. Kabaila et al. (2020) derive convenient formulas for these estimators and the confidence intervals based on them in Sections 3.1 and 3.3, respectively. Starting with these finite-sample formulas, we rigorously prove that these estimators and confidence intervals display a superefficiency phenomenon and that, despite initial appearances, they are actually much poorer than the usual estimators and confidence intervals. By contrast, Kabaila et al. (2020) do not deal with superefficiency, and their large sample results, while plausible and numerically verified, are not proved in full rigor.

Sections 9.2–9.5 set the scene for the rest of this chapter. We consider the two nested linear regression models described in Sect. 9.2 and the frequentist model averaged estimator described in Sect. 9.3. We use the convenient expression for this estimator described in Sect. 9.4 and the large sample framework described in Sect. 9.5. The main results are presented in Sect. 9.6 and 9.7, which deal with the superefficiency of the model average estimator, using an approach described by Le Cam (1953), and a confidence interval centered on this estimator, respectively.

## 9.2 The Two Nested Linear Regression Models

Consider two nested linear regression models with unknown error variance, which we call the full model $\mathcal{M}_2$, and a simpler model $\mathcal{M}_1$. The full model $\mathcal{M}_2$ is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is an $n$-vector of random responses, $\boldsymbol{X}$ is a $n \times p$ matrix with known linearly independent columns, $\boldsymbol{\beta}$ is an $p$-vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an $n$-vector of random errors with a $N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ distribution in which $\sigma^2$ is an unknown positive parameter. We assume throughout that that $p$ is given ($p \geq 3$) and $n > p$. We will consider the case that $n \to \infty$. For conciseness, we do not make explicit in the notation any dependence on $n$, except when this is unavoidable.

Suppose that the parameter of interest is $\theta = \boldsymbol{a}^\top \boldsymbol{\beta}$, where $\boldsymbol{a}$ is a specified $p$-vector ($\boldsymbol{a} \neq \boldsymbol{0}$). To define the simpler model, we define another parameter $\tau = \boldsymbol{c}^\top \boldsymbol{\beta} - t$, where the vector $\boldsymbol{c}$ and the number $t$ are specified and $\boldsymbol{a}$ and $\boldsymbol{c}$ are linearly independent. The model $\mathcal{M}_1$ is $\mathcal{M}_2$ with $\tau = 0$. Let $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ denote the least squares estimator of $\boldsymbol{\beta}$ and $\hat{\sigma}^2 = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) / (n - p)$ the usual unbiased estimator of $\sigma^2$. Let $\hat{\theta} = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}}$ and $\hat{\tau} = \boldsymbol{c}^\top \hat{\boldsymbol{\beta}} - t$.

## 9.3 The Frequentist Model Averaged Estimator

Following Buckland et al. (1997, p. 604), consider the estimator of $\theta$, averaged over the models $\mathcal{M}_1$ and $\mathcal{M}_2$, that is, $\tilde{\theta} = w_1 \hat{\theta}_1 + w_2 \hat{\theta}_2$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the usual estimators of $\theta$ under the models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, and $w_1$ and $w_2$, satisfying $w_1 + w_2 = 1$, are the nonnegative data-based weights for the models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Now, $\hat{\theta}_2 = \hat{\theta}$ and

$$\hat{\theta}_1 = \hat{\theta} - \frac{\boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}}{\boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}} \hat{\tau} = \hat{\theta} - \frac{\mathrm{cov}(\hat{\theta}, \hat{\tau})}{\mathrm{var}(\hat{\tau})} \hat{\tau},$$

since $\mathrm{cov}(\hat{\theta}, \hat{\tau}) = \sigma^2 \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}$ and $\mathrm{var}(\hat{\tau}) = \sigma^2 \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}$. Therefore,

$$\tilde{\theta} = (1 - w_1)\hat{\theta} + w_1 \left( \hat{\theta} - \frac{\mathrm{cov}(\hat{\theta}, \hat{\tau})}{\mathrm{var}(\hat{\tau})} \hat{\tau} \right) = \hat{\theta} - \frac{\mathrm{cov}(\hat{\theta}, \hat{\tau})}{\mathrm{var}(\hat{\tau})} \hat{\tau}\, w_1.$$

Let $\mathrm{GIC}_k$ be the generalized information criterion for model $\mathcal{M}_k$, given by

$$\mathrm{GIC}_k = -2\, L_k + d_n(p + k - 1),$$

where $L_k$ is the maximized log-likelihood for model $\mathcal{M}_k$ ($k = 1, 2$), $d_n$ is nonnegative for every $n$, and $d_n = o(n)$. For AIC, $d_n = 2$ and for BIC, $d_n = \ln(n)$. Suppose that $w_1 = \exp\{-\text{GIC}_1 / 2\} / \big(\exp\{-\text{GIC}_1 / 2\} + \exp\{-\text{GIC}_2 / 2\}\big)$. Let $v_{\tau,n} = \text{var}\,(\hat{\tau})/\sigma^2 = \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}$, which is a known quantity, and $m_n = n - p$. Let $\gamma_n = \tau \big/ \big(\sigma\, v_{\tau,n}^{1/2}\big)$ and $\hat{\gamma}_n = \hat{\tau} \big/ \big(\hat{\sigma}\, v_{\tau,n}^{1/2}\big)$. As shown in Section 3 of Kabaila et al. (2020), $w_1 = w_1(\hat{\gamma}_n)$ where

$$w_1(x) = \frac{1}{1 + \left(1 + \dfrac{x^2}{m_n}\right)^{n/2} \exp\left\{-\dfrac{d_n}{2}\right\}}.$$

The model averaged estimator $\tilde{\theta} = w_1\, \hat{\theta}_1 + (1 - w_1)\, \hat{\theta}_2$ shrinks the estimator $\hat{\theta}_2$ toward the estimator $\hat{\theta}_1$ by an amount that increases as $\hat{\gamma}_n^2$, which measures the discord between the data and model $\mathcal{M}_1$, decreases. As noted in Example 7.7 of Chapter 4 and Section 6 of Chapter 5 of Lehmann and Casella (1998), shrinkage can be toward a target selected by the data rather than a specified point in the parameter space.

## 9.4 Convenient Expression for the Model Averaged Estimator

Define

$$\rho_n = \text{cor}\,\big(\hat{\theta}, \hat{\tau}\big) = \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c} \Big/ \big(\boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}\; \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{c}\big)^{1/2},$$

which is a known quantity. Let $W_n = \hat{\sigma}/\sigma$ and $\tilde{\gamma}_n = \hat{\tau} \big/ \big(\sigma\, v_{\tau,n}^{1/2}\big)$. Let $k_n(x) = x\, w_1(x)$. As shown in Section 3.1 of Kabaila et al. (2020), the model averaged estimator is

$$\tilde{\theta} = \hat{\theta} - \rho_n\, W_n \left(\text{var}(\hat{\theta})\right)^{1/2} k_n\big(\tilde{\gamma}_n / W_n\big). \tag{9.2}$$

This is a convenient expression for $\tilde{\theta}$ because it is determined by $\big(\hat{\theta}, \tilde{\gamma}_n\big)$, and $W_n$, which are independent, with $\big(\hat{\theta}, \tilde{\gamma}_n\big)$ having bivariate normal distribution and $W_n$ converging in probability to 1 as $n \to \infty$.

## 9.5 Large Sample Framework

Convergences of sequences of real numbers and convergences in probability and distribution of sequences of random variables will be as $n \to \infty$. For a sequence of real numbers $x_n$, we write $x_n \to \infty$ if, for every real number $b$, there exists $N_b < \infty$ such that $x_n > b$ for all $n > N_b$. For a sequence of real-valued random variables $X_n$, we write $X_n \xrightarrow{p} \infty$ if, for every real number $b$, $\mathrm{P}(X_n < b) \to 0$ (see, e.g., p. 431 of Lehmann & Romano, 2005).

Assume that $\boldsymbol{D} = \lim_{n\to\infty} \boldsymbol{X}^\top \boldsymbol{X}/n$ exists and is nonsingular. Therefore,

$$v_{\tau,n} = n^{-1}\boldsymbol{c}^\top \big(\boldsymbol{X}^\top \boldsymbol{X}/n\big)^{-1}\boldsymbol{c} = n^{-1}\, v + o(n^{-1}),$$

where $v = \boldsymbol{c}^\top \boldsymbol{D}^{-1}\boldsymbol{c} > 0$ and

$$\rho_n \to \rho = \boldsymbol{a}^\top \boldsymbol{D}^{-1}\boldsymbol{c} \big/ \big(\boldsymbol{a}^\top \boldsymbol{D}^{-1}\boldsymbol{a}\ \boldsymbol{c}^\top \boldsymbol{D}^{-1}\boldsymbol{c}\big)^{1/2}.$$

Henceforth, suppose that $\rho \neq 0$.

## 9.6 Superefficiency of the Model Averaged Estimator

Let $G_n = \big(\hat{\theta} - \theta\big)\big/\big(\mathrm{var}(\hat{\theta})\big)^{1/2}$, which has an $N(0, 1)$ distribution. We assess the efficiency of the model averaged estimator $\tilde{\theta}$ by comparing the concentration about zero of the distribution of $T_n = \big(\tilde{\theta} - \theta\big)\big/\big(\mathrm{var}(\hat{\theta})\big)^{1/2}$ with the concentration about zero of the $N(0, 1)$ distribution of $G_n$. It follows from Eq. (9.2) that

$$T_n = G_n - \rho_n\, W_n\, k_n\big(\hat{\gamma}_n\big), \tag{9.3}$$

where $\hat{\gamma}_n = \tilde{\gamma}_n \big/ W_n$. As noted by Kabaila et al. (2020), $\big(G_n, \tilde{\gamma}_n\big)$ and $W_n$ are independent:

$$\begin{bmatrix} G_n \\ \tilde{\gamma}_n \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \gamma_n \end{bmatrix}, \begin{bmatrix} 1 & \rho_n \\ \rho_n & 1 \end{bmatrix}\right)$$

and $W_n \xrightarrow{p} 1$. We consider two cases: (a) $d_n \to \infty$ and (b) $d_n = d$ for all $n$ ($d \geq 0$). We will show that in both cases, the averaged estimator $\tilde{\theta}$ is *superefficient*, using the following basic result.

**Theorem 9.1** *Suppose that $\gamma_n = \gamma$ for all $n$. Let*

$$\begin{bmatrix} G \\ H \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \gamma \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{9.4}$$

*If* $d_n \to \infty$, *then* $T_n \xrightarrow{d} G - \rho H$. *If* $d_n = d$ *for all* $n$ $(d \geq 0)$, *then* $T_n \xrightarrow{d} G - \rho\, k(H; d)$, *where*

$$k(h; d) = \frac{h}{1 + \exp\left\{\dfrac{h^2 - d}{2}\right\}}. \tag{9.5}$$

The proof is outlined in appendix "Proof of Theorem 9.1".

Consider the case that $d_n = d$ for all $n$ $(d \geq 0)$. Using Theorem 9.1, we compare the large sample efficiency of the model averaged estimator $\tilde{\theta}$ with that of $\hat{\theta}$ by comparing the concentration about 0 of the distribution of $G - \rho\, k(H; d)$ with that of $G$. Define $\tilde{e}(c; d, \rho, \gamma) = \mathrm{P}\big(|G - \rho\, k(H; d)| \leq c\big)$, where $(G, H)$ satisfies in Eq. (9.4). Considered as a function of $c > 0$, $\tilde{e}(c; d, \rho, \gamma)$ is a measure of the concentration of the distribution of $G - \rho\, k(H; d)$ about 0, for given $\rho$ and $\gamma$. Let $e(c) = \mathrm{P}(|G| \leq c)$, where $G \sim N(0, 1)$. We use $\tilde{e}(c; d, \rho, \gamma) - e(c)$, considered as a function of $c > 0$, to compare the concentrations about 0 of the distributions of $G - \rho\, k(H; d)$ and $G$, for given $\rho$ and $\gamma$.

Lemmas 9.1 and 9.2, stated and proved in appendix "Evenness of $\tilde{e}(c; d, \rho, \gamma)$ in $\rho$ and $\gamma$ and a Computationally Convenient Formula for $\tilde{e}(c; d, \rho, \gamma)$", show that $\tilde{e}(c; d, \rho, \gamma)$ is (a) an even function of $\rho$ for given $\gamma$ and (b) an even function of $\gamma$ for given $\rho$. Therefore, when using the criterion $\tilde{e}(c; d, \rho, \gamma) - e(c)$, it is sufficient to consider $\rho > 0$ (recall the assumption that $\rho \neq 0$) and $\gamma \geq 0$. When numerically evaluating this criterion, we use the computationally convenient formula for $\tilde{e}(c; d, \rho, \gamma)$ given in Lemma 9.3, which is stated and proved in appendix "Evenness of $\tilde{e}(c; d, \rho, \gamma)$ in $\rho$ and $\gamma$ and a Computationally Convenient Formula for $\tilde{e}(c; d, \rho, \gamma)$".

### 9.6.1 *A Theorem that Appears to Suggest that the Model Averaged Estimator is Better than the Least Squares Estimator*

Let $\psi = \tau/\sigma$. The following theorem considers $\psi$ fixed. In other words, it uses so-called fixed parameter asymptotics.

**Theorem 9.2** *Suppose that* $\psi$ *is fixed.*

*(a) Suppose that* $\psi = 0$ *and* $d_n \to \infty$. *Then,* $T_n \xrightarrow{d} N(0, 1 - \rho^2)$.

*(b) Suppose that* $\psi = 0$ *and* $d_n = d$ *for all* $n$ $(d \geq 0)$. *Then,* $T_n \xrightarrow{d} G - \rho\, k(H; d)$, *where* $(G, H)$ *satisfies in Eq.* (9.4) *with* $\gamma = 0$, *and* $k(h; d)$ *is given by Eq.* (9.5).

*(c) Suppose that* $\psi \neq 0$. *Then,* $T_n \xrightarrow{d} N(0, 1)$.

The proof is outlined in appendix "Proof of Theorem 9.2".

Consider the case that $d_n \to \infty$. Taken together, parts (a) and (c) of Theorem 9.2 imply that the model averaged estimator $\tilde{\theta}$ is superefficient. At first sight, this *appears* to suggest that $\tilde{\theta}$ is better than the estimator $\hat{\theta}$, when $n$ is large.

Now, consider the case that $d_n = d$ for all $n$ ($d \geq 0$). Suppose that $\psi = 0$, so that $\gamma_n = 0$ for all $n$. The following theorem is proved, using Theorem 9.2(b).Taken together, the following theorem and Theorem 9.2(c) imply that the model averaged estimator $\tilde{\theta}$ is superefficient. At first sight, this *appears* to suggest that $\tilde{\theta}$ is a better estimator than $\hat{\theta}$, when $n$ is large.

**Theorem 9.3** *Suppose that $\psi = 0$ and $d_n = d$ for all $n$ ($d \geq 0$). Also suppose that $c > 0$ and $\rho > 0$. Then,*

*(a) $\tilde{e}(c; d, \rho, 0) - e(c)$ is an increasing function of $d \geq 0$.*
*(b) $\tilde{e}(c; 0, \rho, 0) - e(c) > 0$.*

*Hence, $\tilde{e}(c; d, \rho, 0) - e(c) > 0$ for all $d \geq 0$.*

The proof is given in appendix "Proof of Theorem 9.3".

Figure 9.1 provides an illustration of the conclusion of this theorem for $\rho = 0.5$. In summary, for both cases (a) $d_n \to \infty$ and (b) $d_n = d$ for all $n$ ($d \geq 0$), Theorems 9.2 and 9.3 imply that the model averaged estimator $\tilde{\theta}$ is superefficient. Initially, this *appears* to suggest that $\tilde{\theta}$ is a better estimator than $\hat{\theta}$ (the usual estimator under the full model $\mathcal{M}_2$), when $n$ is large.

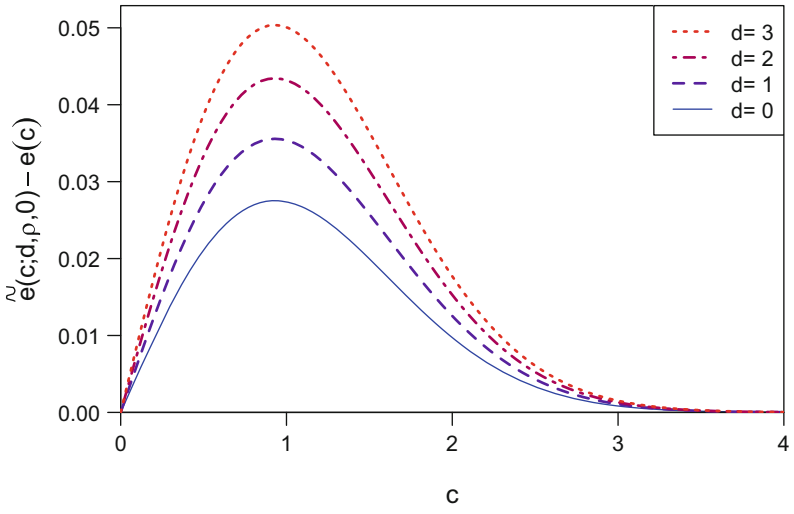



**Fig. 9.1** Illustration of the conclusion of Theorem 9.3 for $\rho = 0.5$

### 9.6.2 The Model Averaged Estimator Is Actually Worse than the Least Squares Estimator

Consider the case that $d_n \to \infty$. The following theorem establishes that the model averaged estimator $\tilde{\theta}$ is *actually much worse* than $\hat{\theta}$.

**Theorem 9.4** *Suppose that $d_n \to \infty$. Let $\gamma_n = d_n^{q/2}$, where $0 < q < 1$. Then, $|T_n| \xrightarrow{p} \infty$.*

The proof is given in appendix "Proof of Theorem 9.4".

Now, consider the case that $d_n = d$ for all $n$ ($d \geq 0$). Suppose that $\gamma_n = \gamma$ for all $n$. We assess the efficiency of $\hat{\theta}$ relative to the model averaged estimator $\tilde{\theta}$ by $\mathrm{E}[T_n^2] - \mathrm{E}[G^2] = \mathrm{E}[T_n^2] - 1$. By Theorem 9.1, $T_n \xrightarrow{d} G - \rho\, k(H; d)$, where the function $k$ is defined by Eq. (9.5). Proposition 2.4 of Shao (2003) implies that

$$\liminf_n \mathrm{E}[T_n^2] - 1 \geq a\, e(\hat{\theta}, \tilde{\theta}; d, \rho, \gamma),$$

where

$$a\, e(\hat{\theta}, \tilde{\theta}; d, \rho, \gamma) = \mathrm{E}\left[(G - \rho\, k(H; d))^2\right] - 1. \tag{9.6}$$

Theorem 9.5 provides a computationally convenient formula for $ae(\hat{\theta}, \tilde{\theta}; d, \rho, \gamma)$ in Eq. (9.6).

**Theorem 9.5** *Suppose that $d_n = d$ for all $n$ ($d \geq 0$). Also, suppose that $\gamma_n = \gamma$ for all $n$. Then,*

$$ae(\hat{\theta}, \tilde{\theta}; d, \rho, \gamma) = \rho^2\, g(d, \gamma),$$

*where*

$$g(d, \gamma) = \int_{-\infty}^{\infty} \left(k^2(y + \gamma; d) - 2\, k(y + \gamma; d)\, y\right) \phi(y)\, \mathrm{d}y.$$

*For every $d \geq 0$, $g(d, \gamma)$ is an even function of $\gamma$.*

The proof is given in appendix "Proof of Theorem 9.5".

Figure 9.2 is a graph of $\max_\gamma g(d, \gamma) = \max_{\gamma \geq 0} g(d, \gamma)$, as a function of $d \in [0, 20]$. This is an increasing function of $d$.

In summary, for both (a) $d_n \to \infty$ and (b) $d_n = d$ for all $n$ ($d \geq 0$), the model averaged estimator $\tilde{\theta}$ is *superefficient*. At first sight, this *appears* to suggest that $\tilde{\theta}$ is better than the estimator $\hat{\theta}$ (the usual estimator under the full model $\mathcal{M}_2$), when $n$ is large. However, as have shown in the previous subsection, $\tilde{\theta}$ is actually *much poorer* than $\hat{\theta}$.

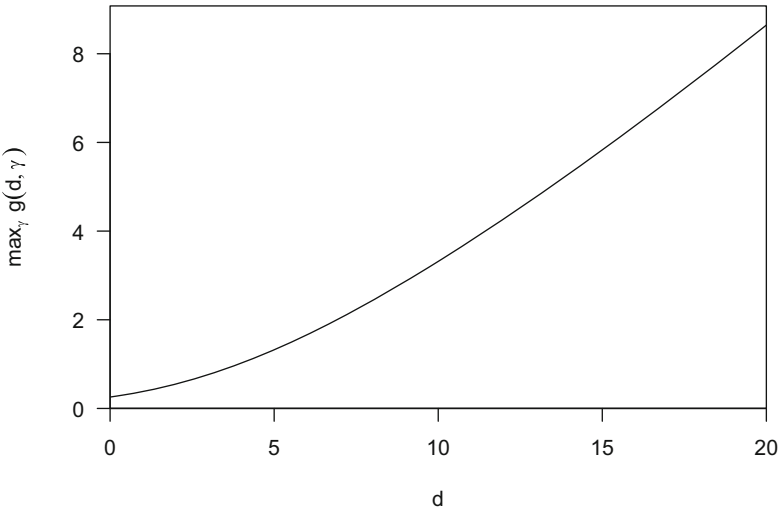



**Fig. 9.2** Plot of the graph of $\max_\gamma g(d, \gamma)$ as a function of $d \in [0, 20]$

## 9.7 Superefficiency of a Confidence Interval Centered on the Model Averaged Estimator

The concept of superefficiency was first formulated in the context of point parameter estimation. Kabaila (1995) described the extension of this concept to confidence intervals. In this section, we describe a confidence interval for $\theta$ proposed by Buckland et al. (1997), which is centered on $\tilde{\theta}$ and has nominal coverage $1 - \alpha$. Kabaila et al. (2020) call this confidence interval $J$ and assess its finite-sample coverage probability and expected length properties. In this section, we show that this confidence interval is *superefficient*.

We begin by describing the usual $1 - \alpha$ confidence interval for $\theta$, assuming that the model $\mathcal{M}_1$ is true. The variance of $\hat{\theta}_1$, the usual estimator of $\theta$ under the model $\mathcal{M}_1$, is $\sigma^2 v_{\theta,n} \left(1 - \rho_n^2\right)$, where $v_{\theta,n} = \text{var}\left(\hat{\theta}\right)/\sigma^2 = \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}$, which is a known quantity. The usual estimator of $\sigma^2 v_{\theta,n} \left(1 - \rho_n^2\right)$, assuming that the model $\mathcal{M}_1$ is true, is

$$\frac{m_n \hat{\sigma}^2 + \left(\hat{\tau}^2 / v_{\tau,n}\right)}{m_n + 1} v_{\theta,n} \left(1 - \rho_n^2\right) = \hat{\sigma}^2 v_{\theta,n} \frac{m_n + \hat{\gamma}_n^2}{m_n + 1} \left(1 - \rho_n^2\right).$$

Therefore, the usual $1 - \alpha$ confidence interval for $\theta$, assuming that the model $\mathcal{M}_1$ is true, is centered on $\hat{\theta}_1$ and has half-width

$$t_{n-p+1,1-\alpha/2} \, \hat{\sigma} \, v_{\theta,n}^{1/2} \left( \frac{m_n + \hat{\gamma}_n^2}{m_n + 1} \left(1 - \rho_n^2\right) \right)^{1/2}, \tag{9.7}$$

where $t_{n-p+1,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the $t_{n-p+1}$ distribution.

Now, consider the confidence interval $J$ for $\theta$ described in Section 3.3 of Kabaila et al. (2020). This confidence interval is centered on $\tilde{\theta}$ and has nominal coverage $1-\alpha$. Its half-width is based on the estimate put forward by Buckland et al. (1997) of their formula (9). As shown in Section 3.3 of Kabaila et al. (2020), this half-width is

$$t_{n-p,1-\alpha/2}\,\hat{\sigma}\,v_{\theta,n}^{1/2}\,r_n(\hat{\gamma}_n,\rho_n), \tag{9.8}$$

where

$$r_n(x,\rho_n) = w_1(x)\left(\frac{m_n+x^2}{m_n+1}\left(1-\rho_n^2\right)+\left(1-w_1(x)\right)^2\rho_n^2\,x^2\right)^{1/2}$$

$$+\left(1-w_1(x)\right)\left(1+w_1^2(x)\,\rho_n^2\,x^2\right)^{1/2}.$$

The usual $1-\alpha$ confidence interval for $\theta$, assuming that the model $\mathcal{M}_2$ is true, is centered on $\hat{\theta}_2$ and has half-width

$$t_{n-p,1-\alpha/2}\,\hat{\sigma}\,v_{\theta,n}^{1/2}\,. \tag{9.9}$$

Denote this confidence interval by $I$. We may view the half-width defined in Eq. (9.8) as being obtained by shrinking the half-width given by Eq. (9.9) toward the half-width in Eq. (9.7). To summarize, both the center and the half-width of the confidence interval $J$ are obtained by shrinkage toward the center and half-width, respectively, of the usual $1-\alpha$ confidence interval for $\theta$ obtained under the model $\mathcal{M}_1$.

Let CP denote the coverage probability of the confidence interval $J$. It may be shown that

$$\text{CP} = \text{P}\Big(-t_{n-p,1-\alpha/2}\,W_n\,r_n\big(\hat{\gamma}_n,\rho_n\big) \le T_n \le t_{n-p,1-\alpha/2}\,W_n\,r_n\big(\hat{\gamma}_n,\rho_n\big)\Big). \tag{9.10}$$

Also, let SL denote the scaled length of the confidence interval $J$, defined to be

$$\frac{\text{half-width of the confidence interval } J}{\text{half-width of the confidence interval } I} = \frac{\text{Eq. (9.8)}}{\text{Eq. (9.9)}} = r_n(\hat{\gamma}_n,\rho_n).$$

In the next two subsections, we assess the confidence interval $J$ using both CP and SL. Henceforth, we consider the case that $d_n \to \infty$. Recall the assumption made throughout that $d_n = o(n)$.

### 9.7.1 A Theorem that Appears to Suggest that the Model Averaged Confidence Interval Is Better than the Usual Confidence Interval

Let $\psi = \tau/\sigma$. Theorem 9.6 considers $\psi$ fixed. In other words, it uses so-called fixed parameter asymptotics.

**Theorem 9.6** *Suppose that $\psi$ is fixed and $d_n \to \infty$.*

*(a) Suppose that $\psi = 0$. Then,* $\text{CP} \to 1 - \alpha$ *and* $\text{SL} \xrightarrow{p} \left(1 - \rho^2\right)^{1/2}$.

*(b) Suppose that $\psi \neq 0$. Then,* $\text{CP} \to 1 - \alpha$ *and* $\text{SL} \xrightarrow{p} 1$.

The proof is outlined in appendix "Proof of Theorem 9.6".

Taken together, parts (a) and (b) of this theorem imply that the model averaged confidence interval $J$ is *superefficient*. At first sight, this *appears* to suggest that the confidence interval $J$ is better than the confidence interval $I$, the usual $1 - \alpha$ confidence interval for $\theta$ under the model $M_2$, when $n$ is large.

### 9.7.2 The Model Averaged Confidence Interval Is Actually Much Worse than the Usual Confidence Interval

Theorem 9.7 has the immediate consequence that the infimum over the parameter space of CP converges to 0, as $n \to \infty$. In other words, the model averaged confidence interval $J$ is not a valid confidence interval when $n$ is large. Consequently, $J$ is *actually much worse* than $I$ when $n$ is large.

**Theorem 9.7** *Suppose that $d_n \to \infty$. Let $\gamma_n = d_n^{q/2}$, where $0 < q < 1$. Then,* $\text{CP} \to 0$.

The proof is given in appendix "Proof of Theorem 9.7".

In summary, for $d_n \to \infty$, the model averaged confidence interval $J$ is *superefficient*. At first sight, this *appears* to suggest that $J$ is better than the usual $1 - \alpha$ confidence interval $I$. However, as shown in the previous subsection, $J$ is *actually much worse* than $I$.

## 9.8 Discussion

Stein-type estimators and confidence sets for the multivariate normal mean, which dominate the usual estimators and confidence sets, are obtained by shrinkage; see, e.g., Saleh (2006). Confidence intervals for the normal variance that dominate the usual confidence intervals are described by Maatta and Casella (1990) and Goutis

and Casella (1991). These intervals are also obtained by shrinkage. In other words, shrinkage can sometimes lead to remarkably good statistical procedures.

However, shrinkage can also lead to inferior statistical procedures. As we have demonstrated, frequentist model averaging of the type proposed by Buckland et al. (1997) is obtained by shrinkage and leads to estimators and confidence intervals that are superefficient. While superefficient estimators and confidence intervals may initially appear to be better than the usual estimators and confidence intervals, they are actually much worse.

## Appendix

### *Proof of Theorem 9.1*

Suppose that $\gamma_n = \gamma$ for all $n$, so that $\tilde{\gamma}_n \sim N(\gamma, 1)$. Since $\hat{\gamma}_n = \tilde{\gamma}_n / W_n$,

$$k_n(\hat{\gamma}_n) = \frac{\tilde{\gamma}_n / W_n}{1 + \left(1 + \dfrac{(\tilde{\gamma}_n^2 / W_n^2)/2}{m_n/2}\right)^{m_n/2} \left(1 + \dfrac{\tilde{\gamma}_n^2 / W_n^2}{m_n}\right)^{p/2} \exp\left\{-\dfrac{d_n}{2}\right\}}. \tag{9.11}$$

Recall that $W_n \xrightarrow{p} 1$. Since $\tilde{\gamma}_n^2 = O_p(1)$ and $m_n = O(n)$, $\left(\tilde{\gamma}_n^2 / W_n^2\right)/m_n \xrightarrow{p} 0$. Since $p$ (the dimension of the parameter vector $\boldsymbol{\beta}$) is fixed,

$$\left(1 + \frac{\tilde{\gamma}_n^2 / W_n^2}{m_n}\right)^{p/2} \xrightarrow{p} 1.$$

Also, since $\tilde{\gamma}_n^2 / W_n^2 - \tilde{\gamma}_n^2 \xrightarrow{p} 0$ and $\tilde{\gamma}_n^2$ has a noncentral $\chi^2$ distribution with 1 degree of freedom and noncentrality parameter $\gamma^2$,

$$\left(1 + \frac{\tilde{\gamma}_n^2/2}{m_n/2}\right)^{m_n/2} \xrightarrow{p} \exp\left(\tilde{\gamma}_n^2/2\right).$$

Finally, $\tilde{\gamma}_n / W_n - \tilde{\gamma}_n \xrightarrow{p} 0$. Therefore,

$$k_n(\hat{\gamma}_n) - \frac{\tilde{\gamma}_n}{1 + \exp\left\{\dfrac{\tilde{\gamma}_n^2 - d_n}{2}\right\}} \xrightarrow{p} 0.$$

We now consider the following two cases:

**Case 1:** In this case, $d_n \to \infty$

$$\exp\left\{\frac{\tilde{\gamma}_n^2 - d_n}{2}\right\} \xrightarrow{p} 0,$$

so that $k_n\big(\hat{\gamma}_n\big) - \tilde{\gamma}_n \xrightarrow{p} 0$. Hence, $T_n \xrightarrow{d} G - \rho\, H$.

**Case 2:** In this case, $d_n = d$ for all $n$ ($d \geq 0$). Thus, $T_n \xrightarrow{d} G - \rho\, k(H; d)$.

## *Evenness of $\tilde{e}(c; d, \rho, \gamma)$ in $\rho$ and $\gamma$ and a Computationally Convenient Formula for $\tilde{e}(c; d, \rho, \gamma)$*

The following lemma implies that, when assessing the properties of $\tilde{e}(c; d, \rho, \gamma)$ for $\rho \neq 0$, it is sufficient to consider $\rho > 0$.

**Lemma 9.1** *For all $c \geq 0$, $d \geq 0$, and $\gamma \in \mathbb{R}$, $\tilde{e}(c; d, \rho, \gamma)$ is an even function of $\rho$.*

***Proof*** By definition, $\tilde{e}(c; d, \rho, \gamma) = \mathrm{P}\big(|G - \rho\, k(H; d)| \leq c\big)$, where the distribution of $(G, H)$ is given by Eq. (9.4). Now,

$$\begin{bmatrix} -G \\ H \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \gamma \end{bmatrix}, \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\right).$$

Therefore, $\tilde{e}(c; d, -\rho, \gamma) = \mathrm{P}\big(|(-G) - (-\rho)\, k(H; d)| \leq c\big) = \tilde{e}(c; d, \rho, \gamma)$. □

The following lemma implies that, when assessing the properties of $\tilde{e}(c; d, \rho, \gamma)$, it is sufficient to consider $\gamma \geq 0$.

**Lemma 9.2** *For all $c \geq 0$, $d \geq 0$, and $\rho \in (-1, 1)$, $\tilde{e}(c; d, \rho, \gamma)$ is an even function of $\gamma$.*

***Proof*** By definition, $\tilde{e}(c; d, \rho, \gamma) = \mathrm{P}\big(|G - \rho\, k(H; d)| \leq c\big)$, where the distribution of $(G, H)$ is given by Eq. (9.4). Now,

$$\begin{bmatrix} -G \\ -H \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ -\gamma \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

Therefore, $\tilde{e}(c; d, \rho, -\gamma) = \mathrm{P}\big(|(-G) - \rho\, k(-H; d)| \leq c\big)$. Now, $k(-H; d) = -k(H; d)$. Hence, $\tilde{e}(c; d, \rho, -\gamma) = \mathrm{P}\big(|-G + \rho\, k(H; d)| \leq c\big) = \tilde{e}(c; d, \rho, \gamma)$. □

The following lemma provides a computationally convenient formula for $\tilde{e}(c; d, \rho, \gamma)$.

**Lemma 9.3** *Let* $\Psi(a, b; \mu, v) = P(a \leq Z \leq b)$, *where* $Z \sim N(\mu, v)$. *Then,*

$$\tilde{e}(c; d, \rho, \gamma) = \int_{-\infty}^{\infty} \Psi\big(-c, c; \rho\big(y - k(y + \gamma; d)\big), 1 - \rho^2\big)\, \phi(y)\, dy,$$

*where* $\phi$ *denotes the* $N(0, 1)$ *density function.*

***Proof*** We use

$$P\big(|G - \rho\, k(H; d)| \leq c\big) = \int_{-\infty}^{\infty} P\big(|G - \rho\, k(H; d)| \leq c \,\big|\, H = h\big)\, \phi(h - \gamma)\, dh. \tag{9.12}$$

It follows from Eq. (9.4) that, conditional on $H = h$, $G - \rho\, k(H; d)$ has a normal distribution with mean $\rho\,\big(h - \gamma - k(h; d)\big)$ and variance $1 - \rho^2$. Therefore,

$$P\big(|G - \rho\, k(H; d)| \leq c \,\big|\, H = h\big) = \Psi\big(-c, c; \rho\big(h - \gamma - k(h; d)\big), 1 - \rho^2\big).$$

It follows from Eq. (9.12) that

$$P\big(|G - \rho\, k(H; d)| \leq c\big) = \int_{-\infty}^{\infty} \Psi\Big(-c, c; \rho\big(h - \gamma - k(h; d)\big), 1 - \rho^2\Big)\, \phi(h - \gamma)\, dh.$$

The result is obtained by changing the variable of integration to $y = h - \gamma$. □

Lemmas 9.1 and 9.2 show that $\tilde{e}(c; d, \rho, \gamma)$ is (a) an even function of $\rho$ for given $\gamma$ and (b) an even function of $\gamma$ for given $\rho$. Therefore, when using the criterion $\tilde{e}(c; d, \rho, \gamma) - e(c)$, it is sufficient to consider $\rho > 0$ (recall the assumption that $\rho \neq 0$) and $\gamma \geq 0$. When numerically evaluating this criterion, we use the computationally convenient formula for $\tilde{e}(c; d, \rho, \gamma)$ given in Lemma 9.3.

## *Proof of Theorem 9.2*

To prove parts (a) and (b), suppose that $\psi = 0$, so that $\gamma_n = 0$ for all $n$. The following are consequences of Theorem 9.1. If $d_n \to \infty$, then $T_n \xrightarrow{d} G - \rho\, H$, so that $T_n \xrightarrow{d} N(0, 1 - \rho^2)$. If $d_n = d$ for all $n$ ($d \geq 0$), then $T_n \xrightarrow{d} G - \rho\, k(H; d)$, where $(G, H)$ satisfies in Eq. (9.4) with $\gamma = 0$, and $k(h; d)$ is given by Eq. (9.5).

To prove part (c), suppose that $\psi \neq 0$. Now,

$$\big|k_n(x)\big| < \frac{|x|}{1 + \left(1 + \dfrac{x^2/2}{m_n/2}\right)^{m_n/2} \exp\left\{-\dfrac{d_n}{2}\right\}} \tag{9.13}$$

for all $x \in \mathbb{R}$. By 4.2.36 on p. 70 of Abramowitz and Stegun (1964),

$$\left(1 + \frac{z}{y}\right)^{y} > \exp\left\{\frac{z}{1 + z\, y^{-1}}\right\} \qquad y > 0, z > 0.$$

Hence,

$$\left(1 + \frac{x^2/2}{m_n/2}\right)^{m_n/2} > \exp\left\{\frac{x^2/2}{1 + (x^2/m_n)}\right\}$$

for all $x \in \mathbb{R}$. Therefore, the right-hand side of Eq. (9.13) is less than

$$\frac{|x|}{1 + \exp\left\{\dfrac{x^2/2}{1 + (x^2/m_n)} - \dfrac{d_n}{2}\right\}}$$

for all $x \in \mathbb{R}$. Since $\hat{\gamma}_n = \tilde{\gamma}_n / W_n$,

$$\left|k_n(\hat{\gamma}_n)\right| < \frac{\left|\tilde{\gamma}_n / W_n\right|}{1 + \exp\left\{\dfrac{\tilde{\gamma}_n^2/(2\, W_n^2)}{1 + \tilde{\gamma}_n^2/(m_n\, W_n^2)} - \dfrac{d_n}{2}\right\}}. \tag{9.14}$$

Recall that $v_{\tau,n} = n^{-1}\, v + o(n^{-1})$, where $v = \boldsymbol{c}^{\top} \boldsymbol{D}^{-1} \boldsymbol{c} > 0$. Therefore,

$$\gamma_n = \psi / v_{\tau,n}^{1/2} = n^{1/2}\big((\psi/v^{1/2}) + o(1)\big).$$

Since $\tilde{\gamma}_n \sim N(\gamma_n, 1)$, $\tilde{\gamma}_n = n^{1/2}\big((\psi/v^{1/2}) + o_p(1)\big)$ and $\tilde{\gamma}_n^2 = n\big((\psi^2/v) + o_p(1)\big)$. Also, since $n/m_n \to 1$,

$$\tilde{\gamma}_n^2 / m_n = \big(\psi^2/v\big) + o_p(1).$$

By assumption, $d_n = o(n)$. Hence,

$$\exp\left\{\frac{\tilde{\gamma}_n^2/(2\, W_n^2)}{1 + \tilde{\gamma}_n^2/(m_n\, W_n^2)} - \frac{d_n}{2}\right\} = \exp\left\{n\left(\frac{\psi^2/v}{1 + (\psi^2/v)} + o_p(1)\right)\right\}.$$

Therefore, the right-hand side of Eq. (9.14) is

$$\frac{\left|n^{1/2}\big((\psi/v^{1/2}) + o_p(1)\big)\right|}{1 + \exp\left\{n\left(\dfrac{\psi^2/v}{1 + (\psi^2/v)} + o_p(1)\right)\right\}},$$

which converges in probability to 0. Hence, $k_n\left(\hat{\gamma}_n\right) \xrightarrow{p} 0$. Therefore, $T_n - G \xrightarrow{p} 0$. Consequently, $T_n \xrightarrow{d} N(0, 1)$.

## *Proof of Theorem 9.3*

To prove part (a), we use the following well-known result.

**Lemma 9.4** *Suppose that $Z \sim N(0, 1)$ and $c' > 0$. Then, $\mathrm{P}\left(-c' \leq Z + \mu' \leq c'\right)$ is a decreasing function of $\mu' \in [0, \infty)$ and an increasing function of $\mu' \in (-\infty, 0]$.*

**Proof of Theorem 9.3**
Suppose that $\psi = 0$ and $d_n = d$ for all $n$ ($d \geq 0$). Also suppose that $c > 0$ and $\rho > 0$. Throughout this proof, we restrict attention to $d \geq 0$. By Lemma 9.3,

$$\tilde{e}(c; d, \rho, 0) = \int_{-\infty}^{\infty} \Psi\left(-c, c; \rho\left(y - k(y; d)\right), 1 - \rho^2\right) \phi(y)\, \mathrm{d}y. \tag{9.15}$$

Here,

$$y - k(y; d) = \frac{y}{1 + \exp\{-y^2/2\}\, \exp\{d/2\}}. \tag{9.16}$$

Now, consider the dependence of the integrand in Eq. (9.15) on $d$, for given $y \neq 0$. Define $\mu = \rho(y - k(y; d))$, so that

$$\begin{aligned} \Psi\left(-c, c; \rho\left(y - k(y; d)\right), 1 - \rho^2\right) &= \Psi\left(-c, c; \mu, 1 - \rho^2\right) \\ &= \mathrm{P}\left(-c' \leq Z + \mu' \leq c'\right), \end{aligned} \tag{9.17}$$

where $Z \sim N(0, 1)$, $c' = c\big/(1 - \rho^2)^{1/2}$, and $\mu' = \mu\big/(1 - \rho^2)^{1/2}$. As $d$ increases, $\mu'$ approaches 0 monotonically. Hence, by Lemma 9.4, the integrand in Eq. (9.15) is an increasing function of $d$. Therefore, Eq. (9.15) is an increasing function of $d$.

To prove part (b), suppose that $d = 0$. It follows from Eq. (9.16) that

$$\Psi\left(-c, c; \rho\left(y - k(y; 0)\right), 1 - \rho^2\right) = \Psi\left(-c, c; \frac{\rho\, y}{1 + \exp\{-y^2/2\}}, 1 - \rho^2\right).$$

Now,

$$\begin{aligned} e(c) &= \mathrm{P}(|G| \le c) \\ &= \int_{-\infty}^{\infty} \mathrm{P}(|G| \le c \mid H = h)\, \phi(h-\gamma)\, \mathrm{d}h \\ &= \int_{-\infty}^{\infty} \Psi\left(-c, c; \rho(h-\gamma), 1-\rho^2\right) \phi(h-\gamma)\, \mathrm{d}h, \end{aligned}$$

since the distribution of $G$, conditional on $H = h$, is $N(\rho(h-\gamma), 1-\rho^2)$. Change the variable of integration to $y = h - \gamma$, so that

$$e(c) = \int_{-\infty}^{\infty} \Psi\left(-c, c; \rho y, 1-\rho^2\right) \phi(y)\, \mathrm{d}y.$$

Hence, $\tilde{e}(c; 0, \rho, 0) - e(c)$ is equal to

$$\int_{-\infty}^{\infty} \left( \Psi\left(-c, c; \frac{\rho\, y}{1+\exp\{-y^2/2\}}, 1-\rho^2\right) - \Psi\left(-c, c; \rho y, 1-\rho^2\right) \right) \phi(y)\, \mathrm{d}y.$$

Now,

$$\begin{aligned} &\Psi\left(-c, c; \frac{\rho\, y}{1+\exp\{-y^2/2\}}, 1-\rho^2\right) - \Psi\left(-c, c; \rho y, 1-\rho^2\right) \\ &\quad = \mathrm{P}\big(-c' \le Z + \mu' \le c'\big) - \mathrm{P}\big(-c' \le Z + \mu'' \le c'\big) \end{aligned} \tag{9.18}$$

where $Z \sim N(0,1)$, $c' = c\big/(1-\rho^2)^{1/2}$,

$$\mu' = \frac{\rho\, y}{(1+\exp\{-y^2/2\})(1-\rho^2)^{1/2}}, \quad \text{and} \quad \mu'' = \frac{\rho\, y}{(1-\rho^2)^{1/2}}.$$

If $y \neq 0$, then $|\mu'| < |\mu''|$ and, by Lemma 9.4, Eq. (9.18) is positive. Hence, $\tilde{e}(c; 0, \rho, 0) - e(c) > 0$.

## *Proof of Theorem 9.4*

We use the following two lemmas, which are also used in the proof of Theorem 9.7.

**Lemma 9.5**

$$\left(1+\frac{\hat{\gamma}_n^2}{m_n}\right)^{n/2} \exp\left\{-\frac{d_n}{2}\right\} < \exp\left\{\frac{\hat{\gamma}_n^2-d_n}{2}\right\}\left(1+\frac{\hat{\gamma}_n^2}{m_n}\right)^{p/2}. \tag{9.19}$$

***Proof*** By the result 4.2.36 on p. 70 of Abramowitz and Stegun (1964),

$$\left(1+\frac{x}{y}\right)^{y} < \exp\{x\} \qquad x>0, y>0.$$

Therefore,

$$\left(1+\frac{\hat{\gamma}_n^2}{m_n}\right)^{n/2} \exp\left\{-\frac{d_n}{2}\right\} = \left(1+\frac{\hat{\gamma}_n^2/2}{m_n/2}\right)^{m_n/2}\left(1+\frac{\hat{\gamma}_n^2}{m_n}\right)^{p/2} \exp\left\{-\frac{d_n}{2}\right\}$$

$$< \exp\left\{\frac{\hat{\gamma}_n^2}{2}\right\}\left(1+\frac{\hat{\gamma}_n^2}{m_n}\right)^{p/2} \exp\left\{-\frac{d_n}{2}\right\}.$$

□

**Lemma 9.6** *Suppose that* $d_n \to \infty$ *and* $\gamma_n = d_n^{q/2}$, *where* $0 < q < 1$. *Then,* $\hat{\gamma}_n^2 = O_p\left(d_n^q\right)$, *and* $w_1(\hat{\gamma}_n) \xrightarrow{p} 1$.

***Proof*** The distribution of $\tilde{\gamma}_n$ is the same as that of $d_n^{q/2} + Z$, where $Z \sim N(0,1)$. Therefore, $\tilde{\gamma}_n^2$ has the same distribution as $\left(d_n^{q/2}+Z\right)^2 = d_n^q + Z^2 + 2\,d_n^{q/2}Z$. Therefore, $\tilde{\gamma}_n^2 = O_p(d_n^q)$. Since $1/W_n^2 = 1 + o_p(1)$, $\hat{\gamma}_n^2 = \tilde{\gamma}_n^2/W_n^2 = O_p(d_n^q)$. It follows from Eq. (9.19) that

$$\left(1+\frac{\hat{\gamma}_n^2}{m_n}\right)^{n/2} \exp\left\{-\frac{d_n}{2}\right\} \xrightarrow{p} 0.$$

Therefore, $w_1(\hat{\gamma}_n) \xrightarrow{p} 1$. □

**Proof of Theorem 9.4**
Suppose that $d_n \to \infty$ and $\gamma_n = d_n^{q/2}$, where $0 < q < 1$. Since $\tilde{\gamma}_n \sim N(\gamma_n, 1)$ and $\gamma_n \to \infty$, $\tilde{\gamma}_n \xrightarrow{p} \infty$. Hence, $\hat{\gamma}_n \xrightarrow{p} \infty$. By Lemma 9.6, $w_1(\hat{\gamma}_n) \xrightarrow{p} 1$, so that $k_n\left(\hat{\gamma}_n\right) = \hat{\gamma}_n\, w_1(\hat{\gamma}_n) \xrightarrow{p} \infty$. We assume throughout that $\rho \neq 0$, so that the result follows from Eq. (9.3) for $T_n$.

### *Proof of Theorem 9.5*

Suppose that $d_n = d$ for all $n$ ($d \geq 0$). Also suppose that $\gamma_n = \gamma$ for all $n$. By definition,

$$\begin{aligned} a\, e(\hat{\theta}, \tilde{\theta}; d, \rho, \gamma) &= \mathrm{E}\left((G - \rho\, k(H; d))^2\right) - 1 \\ &= -2\,\rho\ \mathrm{E}\left(G\, k(H; d)\right) + \rho^2\ \mathrm{E}\left(k^2(H; d)\right). \end{aligned}$$

Now,

$$\mathrm{E}\left(k^2(H; d)\right) = \int_{-\infty}^{\infty} k^2(h; d)\, \phi(h - \gamma)\, \mathrm{d}h$$

and

$$\begin{aligned} \mathrm{E}\left(G\, k(H; d)\right) &= \int_{-\infty}^{\infty} \mathrm{E}\left(G\, k(H; d) \,\middle|\, H = h\right) \phi(h - \gamma)\, \mathrm{d}h \\ &= \int_{-\infty}^{\infty} \mathrm{E}\left(G \,\middle|\, H = h\right) k(h; d)\, \phi(h - \gamma)\, \mathrm{d}h \\ &= \rho \int_{-\infty}^{\infty} k(h; d)\, (h - \gamma)\, \phi(h - \gamma)\, \mathrm{d}h, \end{aligned}$$

since the distribution of $G$, conditional on $H = h$, is $N\left(\rho(h - \gamma), 1 - \rho^2\right)$, so that $\mathrm{E}\left(G \,\middle|\, H = h\right) = \rho(h - \gamma)$. Therefore, $a\, e(\hat{\theta}, \tilde{\theta}; d, \rho, \gamma) = \rho^2\, g(d, \gamma)$, where

$$g(d, \gamma) = \int_{-\infty}^{\infty} \left(k^2(h; d) - 2\, k(h; d)\, (h - \gamma)\right) \phi(h - \gamma)\, \mathrm{d}h.$$

Change the variable of integration to $y = h - \gamma$, so that

$$g(d, \gamma) = \int_{-\infty}^{\infty} \left(k^2(y + \gamma; d) - 2\, k(y + \gamma; d)\, y\right) \phi(y)\, \mathrm{d}y.$$

Hence,

$$g(d, -\gamma) = \int_{-\infty}^{\infty} \left(k^2(y - \gamma; d) - 2\, k(y - \gamma; d)\, y\right) \phi(y)\, \mathrm{d}y.$$

Change the variable of integration to $x = -y$, so that

$$g(d, -\gamma) = \int_{-\infty}^{\infty} \left(k^2(x + \gamma; d) - 2\, k(x + \gamma; d)\, x\right) \phi(x)\, \mathrm{d}x = g(d, \gamma),$$

since $\phi$ is an even function.

## *Proof of Theorem 9.6*

Suppose that $\psi$ is fixed and $d_n \to \infty$.

To prove part (a), suppose that $\psi = 0$, so that $\gamma_n = 0$ for all $n$. By Theorem 9.1, $T_n \xrightarrow{d} G - \rho H$, where $(G, H)$ has the probability distribution given by Eq. (9.4). It follows from Eq. (9.19) that

$$w_1(\hat{\gamma}_n) > \frac{1}{1 + \exp\left\{\dfrac{\hat{\gamma}_n^2 - d_n}{2}\right\}\left(1 + \dfrac{\hat{\gamma}_n^2}{m_n}\right)^{p/2}}. \tag{9.20}$$

Since $\hat{\gamma}_n = \tilde{\gamma}_n / W_n$, where $\tilde{\gamma}_n \sim N(0,1)$ and $W_n \xrightarrow{p} 1$, $\hat{\gamma}_n \xrightarrow{d} N(0,1)$. Therefore, the right-hand side of Eq. (9.20) converges in probability to 1. Since $w_1(x) \leq 1$ for all $x \in \mathbb{R}$, this implies that $w_1(\hat{\gamma}_n) \xrightarrow{p} 1$. Hence,

$$r_n(\hat{\gamma}_n, \rho_n) = w_1(\hat{\gamma}_n)\left(\frac{m_n + \hat{\gamma}_n^2}{m_n + 1}(1 - \rho_n^2) + (1 - w_1(\hat{\gamma}_n))^2 \rho_n^2 \hat{\gamma}_n^2\right)^{1/2}$$

$$+(1 - w_1(\hat{\gamma}_n))(1 + w_1^2(\hat{\gamma}_n)\rho_n^2 \hat{\gamma}_n^2)^{1/2}$$

converges in probability to $(1 - \rho^2)^{1/2}$. Therefore, SL $\xrightarrow{p}$ $(1 - \rho^2)^{1/2}$. Since $t_{n-p,1-\alpha/2} \to z_{1-\alpha/2}$ and $W_n \xrightarrow{p} 1$,

$$\text{CP} \to \text{P}\left(-z_{1-\alpha/2}(1 - \rho^2)^{1/2} \leq G - \rho H \leq z_{1-\alpha/2}(1 - \rho^2)^{1/2}\right) = 1 - \alpha.$$

To prove part (b), suppose that $\psi \neq 0$. It may be shown that CP is an even function of $\psi$. Also, SL $= r_n(\hat{\gamma}_n, \rho_n) = r_n(-\hat{\gamma}_n, \rho_n)$. Therefore, without loss of generality, suppose that $\psi > 0$.

Since $r_n(\hat{\gamma}_n, \rho_n)$ is equal to

$$\left((1 - \rho_n^2)\left(\frac{m_n}{m_n + 1} w_1^2(\hat{\gamma}_n) + \frac{(\hat{\gamma}_n w_1(\hat{\gamma}_n))^2}{m_n + 1}\right) + (1 - w_1(\hat{\gamma}_n))^2 \rho_n^2 (\hat{\gamma}_n w_1(\hat{\gamma}_n))^2\right)^{1/2}$$

$$+(1 - w_1(\hat{\gamma}_n))(1 + \rho_n^2 (\hat{\gamma}_n w_1(\hat{\gamma}_n))^2)^{1/2},$$

we prove that $r_n(\hat{\gamma}_n, \rho_n) \xrightarrow{p} 1$ by proving that $w_1(\hat{\gamma}_n) \xrightarrow{p} 0$ and $\hat{\gamma}_n w_1(\hat{\gamma}_n) \xrightarrow{p} 0$. Define the event $A_n = \{\gamma_n/2 < \hat{\gamma}_n < 3\gamma_n/2\}$. It may be shown that $\text{P}(A_n) \to 1$. Suppose that $\epsilon > 0$. Note that

$$\text{P}(|w_1(\hat{\gamma}_n)| > \epsilon) \leq \text{P}(\{|w_1(\hat{\gamma}_n)| > \epsilon\} \cap A_n) + \text{P}(A_n^c),$$

where $A_n^c$ denotes the complement of the event $A_n$. It may be shown that $\mathrm{P}\big(\{|w_1(\hat{\gamma}_n)| > \epsilon\} \cap A_n\big) \to 0$. It may similarly be shown that $\hat{\gamma}_n\, w_1(\hat{\gamma}_n) \xrightarrow{p} 0$. Therefore, $r_n(\hat{\gamma}_n, \rho_n) \xrightarrow{p} 1$. Hence, SL $\xrightarrow{p}$ 1 and, by Theorem 9.2(c), for $Z \sim N(0, 1)$,

$$\mathrm{CP} \to \mathrm{P}\big(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}\big) = 1 - \alpha.$$

## *Proof of Theorem 9.7*

To prove Theorem 9.7, we need Lemma 9.7.

**Lemma 9.7** *Suppose that $d_n \to \infty$ and $\gamma_n = d_n^{q/2}$, where $0 < q < 1$. Then, $r_n(\hat{\gamma}_n, \rho_n) \xrightarrow{p} (1 - \rho^2)^{1/2}$.*

***Proof*** It follows from

$$0 \le 1 - w_1(\hat{\gamma}_n) < \left(1 + \frac{\hat{\gamma}_n^2}{m_n}\right)^{n/2} \exp\left\{-\frac{d_n}{2}\right\}$$

and Eq. (9.19) inside Lemmas 9.5 and 9.6 that

$$0 \le |\hat{\gamma}_n|\big(1 - w_1(\hat{\gamma}_n)\big) < |\hat{\gamma}_n| \exp\left\{\frac{\hat{\gamma}_n^2 - d_n}{2}\right\} \left(1 + \frac{\hat{\gamma}_n^2}{m_n}\right)^{p/2},$$

where $|\hat{\gamma}_n| = O_p(d_n^{q/2})$ and $\hat{\gamma}_n^2 = O_p(d_n^q)$. Hence,

$$\hat{\gamma}_n\big(1 - w_1(\hat{\gamma}_n)\big) \xrightarrow{p} 0.$$

The result now follows from Lemma 9.6 and the expression

$$r_n(\hat{\gamma}_n, \rho_n) = w_1(\hat{\gamma}_n) \left(\frac{m_n + \hat{\gamma}_n^2}{m_n + 1}\left(1 - \rho_n^2\right) + \rho_n^2 \left(\hat{\gamma}_n\big(1 - w_1(\hat{\gamma}_n)\big)\right)^2\right)^{1/2}$$

$$+\left(\big(1 - w_1(\hat{\gamma}_n)\big)^2 + w_1^2(\hat{\gamma}_n)\, \rho_n^2 \left(\hat{\gamma}_n\big(1 - w_1(\hat{\gamma}_n)\big)\right)^2\right)^{1/2}.$$

□

**Proof of Theorem 9.7**
Suppose that $d_n \to \infty$ and $\gamma_n = d_n^{q/2}$, where $0 < q < 1$. By Lemma 9.7,

$$t_{n-p,1-\alpha/2}\, W_n\, r_n\big(\hat{\gamma}_n, \rho_n\big) \xrightarrow{p} \big(1-\rho^2\big)^{1/2} z_{1-\alpha/2},$$

since $t_{n-p,1-\alpha/2} \to z_{1-\alpha/2}$ and $W_n \xrightarrow{p} 1$. We use Eq. (9.10) for CP. Define the event $B_n = \big\{t_{n-p,1-\alpha/2}\, W_n\, r_n\big(\hat{\gamma}_n, \rho_n\big) < z_{1-\alpha/2}\big\}$ and note that $\mathrm{P}\big(B_n\big) \to 1$. It may be shown that

$$\mathrm{CP} \le \mathrm{P}\Big(-z_{1-\alpha/2} \le T_n \le z_{1-\alpha/2}\Big) + \mathrm{P}\big(B_n^c\big),$$

where $B_n^c$ denotes the complement of the event $B_n$. Since $\mathrm{P}\big(B_n^c\big) \to 0$, the result follows from Theorem 9.4.

**Acknowledgments** The author is grateful to the anonymous reviewers, whose comments and suggestions led to an improved paper.

## References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of Mathematical Functions*. Dover.
Bahadur, R. R. (1964). On Fisher's bound for asymptotic variances. *Annals of Mathematical Statistics, 35*, 1545–1552.
Beran, R. (1992). The radial process for confidence sets. In R. M. Dudley, M. G. Hahn, & J. Kuelbs (Eds.), *Probability in Banach spaces, 8: Proceedings of the eighth international conference* (pp. 479–496). Birkhauser.
Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics, 53*, 603–618.
Goutis, C., & Casella, G. (1991). Improved invariant confidence intervals for a normal variance. *The Annals of Statistics, 19*, 2015–2031.
Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory, 11*, 537–549.
Kabaila, P. (1996). The evaluation of model selection criteria: pointwise limits in the parameter space. In D. L. Dowe, K. B. Korb, & J. J. Oliver (Eds.), *Information, statistics and induction in science* (pp. 114–118). World Scientific.
Kabaila, P. (2009). The coverage properties of confidence regions after model selection. *International Statistical Review, 77*, 405–414.
Kabaila, P., Welsh, A. H., & Wijethunga, C. (2020). Finite sample properties of confidence intervals centered on a model averaged estimator. *Journal of Statistical Planning and Inference, 207*, 10–26.
Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics, 1*, 277–330.
Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory, 21*, 21–59.
Leeb, H., & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics, 142*, 201–211.
Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.
Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer.

Maatta, J. M., & Casella, G. (1990). Developments in decision-theoretic variance estimation. *Statistical Science, 28*, 151–156.
Pötscher, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhya A, 71*, 1–18.
Saleh, A. K. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. Wiley.
Shao, J. (2003). *Mathematical statistics* (2nd ed.). Springer.
van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

# Chapter 10
# Shrinkage Estimation of Restricted Mean Vector Under Balanced Loss with Application in Wavelet Denoising




**Hamid Karamikabir**

## 10.1 Introduction

In many practical scenarios, we encounter numerous variables when studying natural phenomena. Addressing the rise in random variables within parametric statistics poses distinctive difficulties. To effectively tackle or articulate such scenarios, we need to employ distributions like multivariate distributions and probability distributions involving pairs, triples, and so forth, based on the quantity of variables involved.

Comprehending multivariate distributions is essential in statistics due to their crucial role in modeling intricate real-world scenarios involving multiple variables. Unlike univariate distributions, which describe the distribution of a single variable, multivariate distributions capture the collective behavior of two or more variables simultaneously. This capability to represent relationships between variables is indispensable in various statistical analyses, including regression, hypothesis testing, and clustering.

Multivariate distributions offer valuable insights into the interdependencies and correlations among variables, enabling statisticians to evaluate the overall behavior of a system rather than focusing solely on individual variables. They facilitate the characterization of complex patterns and structures within datasets, enabling more nuanced interpretations and well-informed decision-making.

Furthermore, multivariate distributions form the basis for numerous statistical techniques and models, such as multivariate regression, principal component analysis, and factor analysis. By accurately representing the joint distribution of variables,

H. Karamikabir (✉)
Department of Statistics, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf Universit, Bushehr, Iran
e-mail: h_karamikabir@pgu.ac.ir

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_10

**Table 10.1** Some subclasses of $p$ dimensional $\mathcal{SS}$ distributions

| Type | p.d.f. |
|---|---|
| Kotz type | $\mathbb{N}(x^\top x)^{N-1}\exp(-r(x^\top x)^s),\ r,s>0,\ 2N+p>2$ |
| Multivariate normal | $\mathbb{N}\exp(-\frac{1}{2}x^\top x)$ |
| Scale mixture | $\mathbb{N}\int_0^\infty t^{-p/2}\exp(-x^\top x/2t)dG(t),\ G(t)\ a\ c.d.f$ |
| Multivariate $t$ | $\mathbb{N}(1+x^\top x/s)^{(p+m)/2}, s>0, m>0$ an integer |
| Logistic | $\mathbb{N}\exp(-x^\top x)/[1+\exp(-x^\top x)]^2$ |
| Multivariate Cauchy | $\mathbb{N}(1+x^\top x/s)^{(p+1)/2}, s>0$ |
| Pearson Type II | $\mathbb{N}(1-x^\top x)^m, m>0$ |
| Pearson Type VII | $\mathbb{N}(1+x^\top x/s)^{-N}, N>p/2, s>0$ |

statisticians can enhance parameter estimation, prediction accuracy, and inference of causal relationships.

In this chapter, we use the family of multivariate spherically symmetric ($\mathcal{SS}$) distribution. The concept of the spherical contour distribution holds significance across diverse domains, notably in multivariate analysis and spatial statistics. It depicts the spread of points or data observations across the surface of a hyper-sphere , where every point is equally distant from a central reference point. Known for its symmetry and uniformity, this distribution serves as a valuable instrument for modeling spatial patterns and investigating inter-variable relationships within a spherical framework.

The family of multivariate spherically symmetric distribution is defined as

**Definition 10.1** The random vector $X = (X_1, \ldots, X_p)^\top$ is said to have a $\mathcal{SS}$ distribution with mean vector parameters $\theta = (\theta_1, \ldots, \theta_p)^\top$ and covariance matrix $\Sigma$, if $X$ and $HX$ have the same distribution for all $p \times p$ orthogonal matrices $H$.

Assume $(X, U)$ is a $p+k$ random vector having a $\mathcal{SS}$ distribution around the $p+k$ vector $(\theta, 0)$ where dim $X$, $\theta$ is $p$ and dim $U$, $0$ is $k$. The probability density function (p.d.f.) of the $\mathcal{SS}$ distribution is of the form

$$\mathbb{N}\ f\left(\|x-\theta\|^2 + \|u\|^2\right), \tag{10.1}$$

where $\mathbb{N}$ is the normalizing constant.

Table 10.1 shows some types of $\mathcal{SS}$ distributions. For more information about the $\mathcal{SS}$ distribution, refer to Fang et al. (1990).

Evaluating the performance of an estimator, we need to set a measure. In this paper, we use the balanced loss function.

**Definition 10.2** Suppose that $X$ is a random vector with unknown mean vector parameter $\theta$. The balanced loss function ($\mathcal{BLF}$) is defined as

$$L(\theta, \delta) = \omega\|\delta - \delta^*\|^2 + (1-\omega)\|\delta - \theta\|^2, \quad 0 \le \omega < 1, \tag{10.2}$$

where $\delta$ is an estimator of $\theta$ and $\delta^*$ is a target estimator of $\theta$ obtained, for instance, using the criterion of maximum likelihood , least squares , unbiasedness, etc.

The special case of $\mathcal{BLF}$ is the weighted quadratic loss when $\omega = 0$. The associated risk function with respect to (w.r.t) loss function defined in Eq. (10.2) will be $\mathcal{R}(\theta, \delta) = \mathrm{E}_\theta [L(\theta, \delta)]$. $\mathcal{BLF}$ was introduced by Zellner (1994) to reflect two criteria: precision of estimation and goodness of fit. For more details about the use of this loss, we refer to Cao and He (2017), Zinodiny et al. (2017), Karamikabir et al. (2018) and Karamikabir and Afshari (2020, 2021, 2022), to mention a few.

The purpose of this paper is to estimate the mean vector parameter of $\mathcal{SS}$ distribution. For this purpose, the location vector is constrained by limitations such as positive orthant, simplicial, polyhedral, general convex cones, and nested cones, and an estimator has been introduced that has lower risk compared to the natural estimator. Also, by using the soft threshold wavelet shrinkage estimator, we introduce the two types of the restricted soft threshold wavelet shrinkage estimator by finding a specific threshold (SURE).

In recent years, the problem of estimating a mean vector parameter for multivariate distribution has made a lot of progress. For example, we can refer to Ouassou and Strawderman (2002), Maruyama (2003), Fourdrinier et al. (2003, 2008, 2014), Marchand and Strawderman (2020), Karamikabir and Afshari (2020, 2021, 2022); Karamikabir et al. (2023, 2024), and Batvandi et al. (2025). Additionally, in the field of estimating the mean matrix in matrix-variate distributions, references can be referred to Karamikabir et al. (2025), Batvandi et al. (2024, 2025), and Karamikabir (2024).

The paper is outlined as follows: In Sect. 10.2, we introduce the natural and shrinkage estimator of the mean vector parameter, and we show that the shrinkage estimator dominate the natural estimator in the restricted parameter space. In Sect. 10.2.1, we extend this results to simplicial cones and polyhedral and general pointed convex cones. In Sect. 10.2.2, we show the domination condition for nested cones and for orthogonal transformations. We find the restricted shrinkage generalized Bayes estimator in Sect. 10.2.3, and in Sect. 10.2.4, we find the restricted soft threshold wavelet shrinkage estimator for the $\mathcal{SS}$ distribution under $\mathcal{BLF}$. In Sect. 10.3, we show the numerical performance of the proposed estimators using a Monte Carlo simulation study. Finally, we use the Parkinson speech dataset to further demonstrate the superior performance of the restricted soft threshold wavelet shrinkage estimation in Sect. 10.3.1, and the chapter is concluded in Sect. 10.4.

## 10.2 Domination for the Positive Orthant Cones

Parameter estimation of the mean vector is important in multivariate distributions, especially when the vector parameter space is restricted. The concept of a restricted parameter space in statistics is essential for grasping the boundaries and limitations that govern the potential values of parameters within a statistical framework. In

different statistical investigations, specific parameters may encounter limitations imposed by theoretical principles, practical limitations, or prior understanding of the system being analyzed. These limitations manifest in diverse ways, including inequality constraints, linear combinations, or constraints on parameter ranges. Proficiently incorporating and modeling these restrictions is critical for precise inference and comprehension of statistical outcomes.

The positive orthant cones, commonly known as orthant cones or simply positive cones, represent geometric constructs within the realms of convex optimization and linear algebra. These cones are defined as subsets of vector spaces containing all vectors with nonnegative components along specific dimensions. Their significance in convex optimization stems from their natural occurrence in problems requiring nonnegative constraints. For instance, in linear programming, such constraints are prevalent in practical scenarios like resource allocation. Here, negative values are illogical (e.g., negative quantities of resources), making positive orthant cones indispensable for modeling and solving such problems.

In this section, we consider the problem of estimating a mean vector, $\theta = (\theta_1, \ldots, \theta_p)^\top$ by the estimator $\delta = (\delta_1, \ldots, \delta_p)^\top$ when the mean vector restricted to a closed convex cone, $C$. The observation consists of $(X, U)$ where $\dim(X) = p \geq 3$ and $\dim(U) = k \geq 1$. Also, we assume throughout that $\mathrm{E}_\theta(\|X-\theta\|^2+\|U\|^2) \leq \infty$ and hence $\mathrm{E}_\theta(X) = \theta$ and $\mathrm{E}_\theta(U) = 0$.

Under these conditions, the natural estimator proposed by Ouassou and Strawderman (2002) is as

$$\delta_0(X) = P_C X = X + \Gamma_C(X), \tag{10.3}$$

where $P_C X$ is the orthogonal projection of $X$ onto the cone $C$ and $\Gamma_C(X) = P_C X - X$. We consider the class of the shrinkage estimators of the form

$$\delta(X, U) = X + \Gamma_C(X) + \frac{\|U\|^2}{k+2} g(X) I_C(X), \tag{10.4}$$

where the vector function $\Gamma_C(X) = \big(\Gamma_C(X_1), \ldots, \Gamma_C(X_p)\big)^\top$ is defined as

$$\Gamma_C(X_i) = \begin{cases} -X_i & X_i < 0, \\ 0 & X_i \geq 0, \end{cases} \quad \text{or} \quad \Gamma_C(X) = 0 \quad \text{if} \quad X \in C,$$

with $I_C(X)$ as the indicator function and $g(\cdot)$ as the differentiable function.

For any function $\Gamma_C(X) \in \mathbb{R}^p$ and any set $C$ such that $\Gamma_C(X) = 0$ if $X \in C$, the difference in risk between the natural estimator $\delta_0(X)$ and the shrinkage estimator $\delta(X, U)$ is given by

$$\Delta\mathcal{R}(\theta) = \mathrm{E}_\theta \left[ \omega \|\delta - \delta^*\|^2 + (1-\omega)\|\delta - \theta\|^2 \right]$$

$$- \mathrm{E}_\theta \left[ \omega \|\delta_0 - \delta^*\|^2 - (1-\omega)\|\delta_0 - \theta\|^2 \right]. \tag{10.5}$$

Now, for $\mathcal{BLF}$ in Eq. (10.2), we suppose that the target estimator $\delta^*(X) = \delta_0(X)$, and again the difference in risk (10.5) is given by

$$\begin{aligned}
\Delta\mathcal{R}(\theta) = \mathrm{E}_\theta \Bigg[ & \omega \left\| \frac{\|U\|^2}{k+2} g(X) I_C(X) \right\|^2 \\
& + (1-\omega) \left\| X + \Gamma_C(X) + \frac{\|U\|^2}{k+2} g(X) I_C(X) - \theta \right\|^2 \Bigg] \\
& - \mathrm{E}_\theta \left[ (1-\omega) \|X + \Gamma_C(X) - \theta\|^2 \right] \\
= \mathrm{E}_\theta \Bigg[ & \frac{\|U\|^4}{(k+2)^2} \|g(X)\|^2 I_C(X) \\
& + 2(1-\omega) \frac{\|U\|^2}{k+2} g(X)^\top I_C(X)(X + \Gamma_C(X) - \theta) \Bigg].
\end{aligned} \tag{10.6}$$

Since $g(X)^\top \Gamma_C(X) I_C(X) = 0$, the difference in risk (10.6) is given by

$$\Delta\mathcal{R}(\theta) = \mathrm{E}_\theta \left[ \left( \frac{\|U\|^4}{(k+2)^2} \|g(X)\|^2 + 2(1-\omega) \frac{\|U\|^2}{k+2} g(X)^\top (X - \theta) \right) I_C(X) \right]. \tag{10.7}$$

The results for the positive orthant cone $C = \{\theta | \theta_i \geq 0, i = 1, \ldots, p\} = \mathbb{R}^p_+$ depend on the following lemma.

**Lemma 10.1 (Ouassou & Strawderman, 2002)** *Suppose that $(X, U)$ is a $(p+k)$ dimensional random vector having a $\mathcal{SS}$ distribution with p.d.f. in Eq. (10.1) and assume that the mean vector $\theta$ is restricted to the positive orthant $C$. Suppose $g(\cdot)$ is a differentiable function that satisfies*

*(i)* $\lim_{x_i \to 0} g(x_i) \int_{\|x-\theta\|^2 + \|u\|^2}^\infty f(s)\mathrm{d}s = 0,$

*(ii)* $\lim_{x_i \to \infty} g(x_i) \int_{\|x-\theta\|^2 + \|u\|^2}^\infty f(s)\mathrm{d}s = 0,$

*for each $i = 1, \ldots, p$ and all $\theta$, $u$ and $x_j > 0$ $(j \neq i)$. Then,*

$$\mathrm{E}_\theta \left[ \|g(X)\|^2 \|U\|^4 \prod_{i=1}^p I_{[0,\infty)}(X_i) \right] = \frac{\mathbb{N}(k+2)}{2} \mathrm{E}^*_\theta \left[ \|g(X)\|^2 \|U\|^2 \prod_{i=1}^p I_{[0,\infty)}(X_i) \right],$$

*and*

$$\mathrm{E}_\theta\left[(X-\theta)^\top g(X)\|U\|^2\prod_{i=1}^{p}I_{[0,\infty)}(X_i)\right]=\frac{\mathbb{N}}{2}\,\mathrm{E}^*_\theta\left[\|U\|^2\nabla_X\cdot g(X)\prod_{i=1}^{p}I_{[0,\infty)}(X_i)\right]$$

*provided the expectations are finite. Here, $\nabla_X$ denotes the gradient w.r.t. the variable $X$, and $\mathrm{E}^*_\theta$ denotes the expectation, w.r.t. the distribution $\mathbb{N}^{-1}F(\|X-\theta\|^2+\|U\|^2)$, where $F$ is defined as $F(t)=\int_t^\infty f(s)\mathrm{d}s$ (i.e., $-F'(t)=f(t)$) and $\mathbb{N}$ is the normalizing constant.*

The main result of this section is the following theorem.

**Theorem 10.1** *Assume that $(X,U)$ is a $(p+k)$ dimensional random vector having a $\mathcal{SS}$ distribution with p.d.f. in Eq. (10.1), where the set of points in $(0,\infty)$ for which $f$ is discontinuous has Lebesgue measure zero. Let $\theta$ be restricted to the positive orthant cone ($\theta\in C=\mathbb{R}^p_+$). If condition Lemma 10.1 holds and $\|g(X)\|^2+2(1-\omega)\nabla_X\cdot g(X)\le 0$, then the shrinkage estimator $\delta(X,U)$ in Eq. (10.4) dominates the natural estimator $\delta_0(X)$ under $\mathcal{BLF}$. We suppose that all expectations are finite.*

***Proof*** By using Lemma 10.1 for the difference in risk $\Delta\mathcal{R}(\theta)$ in Eq. (10.7), the difference in risk is equal to

$$\begin{aligned}
&\mathrm{E}_\theta\left[\left(\|g(X)\|^2\left(\frac{\|U\|^2}{k+2}\right)^2+2(1-\omega)\frac{\|U\|^2}{k+2}g(X)^\top(X-\theta)\right)I_C(X)\right]\\
&=\mathrm{E}_\theta\left[\left(\|g(X)\|^2\left(\frac{\|U\|^2}{k+2}\right)^2+2(1-\omega)\frac{\|U\|^2}{k+2}g(X)^\top(X-\theta)\right)\Pi_{i=1}^{p}I_{[0,\infty)}(X_i)\right]\\
&=\frac{\mathbb{N}}{2(k+2)}\,\mathrm{E}^*_\theta\left[\left(\|g(X)\|^2\|U\|^2+2(1-\omega)\|U\|^2\nabla_X\cdot g(X)\right)\Pi_{i=1}^{p}I_{[0,\infty)}(X_i)\right]\\
&=\frac{\mathbb{N}}{2(k+2)}\,\mathrm{E}^*_\theta\left[\|U\|^2\left(\|g(X)\|^2+2(1-\omega)\nabla_X\cdot g(X)\right)\Pi_{i=1}^{p}I_{[0,\infty)}(X_i)\right]\le 0.
\end{aligned}\tag{10.8}$$

A necessary condition for the risk difference in Eq. (10.8) to be negative is $\|g(X)\|^2+2(1-\omega)\nabla_X\cdot g(X)\le 0$. The proof is complete. □

In the case where $\omega=0$ in Theorem 1, the result of Theorem 2 by Ouassou and Strawderman (2002) is obtained. In fact, the difference in the loss functions in these two theorems has led to this new condition. In this chapter, the balanced

loss function is used, whereas in Ouassou and Strawderman (2002), the standard quadratic loss function is applied.

In the following section, we will extend the domination to simplicial, polyhedral, and general convex cones with non-empty interior.

### *10.2.1 Domination for General Cones*

In this subsection, we generalize the results of the preceding section to simplicial cones. Furthermore, we extend the dominance results to a polyhedral and general pointed convex cones. If the cone $C$ has a linearly independent system of generators, then it is a simplicial cone. Using the standard decomposition of a polyhedral cone into the union of simplicial cones with nonintersecting interiors, the results for polyhedral cones can be obtained. Similarly, using the approximation of a general cone by a sequence of polyhedral cones, the results can be obtained for general cones.

In this section, we suppose that $g(X) = h(\|X\|^2)X$ in the shrinkage estimator given by Eq. (10.4), although some results can be obtained for non-orthogonally invariant $g(\cdot)$. The difference in risk $\Delta\mathcal{R}(\theta)$ in Eq. (10.7) is equal to

$$\mathrm{E}_\theta\left[\left(\frac{\|U\|^2}{k+2}\right)^2 h^2(\|X\|^2)\|X\|^2 I_C(X) \right.$$
$$\left. +2(1-\omega)\frac{\|U\|^2}{k+2}h(\|X\|^2)X^\top(X-\theta)I_C(X)\right]. \tag{10.9}$$

By using Lemma 10.1 for $\Delta\mathcal{R}(\theta)$ in Eq. (10.9), we have the following equations:

$$\mathrm{E}_\theta\left[\left(\frac{\|U\|^2}{k+2}\right)^2 h^2(\|X\|^2)\|X\|^2 I_C(X)\right] =$$
$$\frac{\mathbb{N}}{2(k+2)}\,\mathrm{E}^*_\theta\left[\|U\|^2 h^2(\|X\|^2)\|X\|^2 I_C(X)\right], \tag{10.10}$$
$$\mathrm{E}_\theta\left[(1-\omega)\frac{\|U\|^2}{k+2}h(\|X\|^2)X^\top(X-\theta)I_C(X)\right] =$$
$$\frac{\mathbb{N}(1-\omega)}{2(k+2)}\,\mathrm{E}^*_\theta\left[p\,\|U\|^2 h(\|X\|^2)I_C(X) + 2\|U\|^2\|X\|^2 h'(\|X\|^2)I_C(X)\right]. \tag{10.11}$$

By replacing Eqs. (10.10) and (10.11) for $\Delta\mathcal{R}(\theta)$ in Eq. (10.9), we have the following difference in risk:

$$\Delta\mathcal{R}(\theta) = \frac{\mathbb{N}}{2(k+2)} \mathrm{E}_\theta^* \Big[ \Big( \|X\|^2 h^2(\|X\|^2) + 2(1-\omega) p h(\|X\|^2) + 4(1-\omega)\|X\|^2 h'(\|X\|^2) \Big) \|U\|^2 I_C(X) \Big]. \quad (10.12)$$

Similar to Theorem 10.1, we have the following theorem:

**Theorem 10.2** *Assume that $(X, U)$ is a $(p+k)$ dimensional random vector having a $\mathcal{SS}$ distribution with p.d.f. in Eq. (10.1). Suppose $\theta$ is restricted to the simplicial cone $C \subset \mathbb{R}^p$. Under $\mathcal{BLF}$, the shrinkage estimator $\delta(X, U)$ in Eq. (10.4) with $g(X) = h(\|X\|^2)X$ dominates the natural estimator $\delta_0(X)$ in Eq. (10.3) provided that (i) and (ii) of Lemma 10.1 and also $h(\cdot)$ are differentiable and satisfy*

$$t\, h^2(t) + 2(1-\omega) p h(t) + 4(1-\omega) t h'(t) \leq 0 \quad \forall, t > 0.$$

***Proof*** It is enough to put $g(X) = h(\|X\|^2)X$ in Theorem 10.1 and make the difference in risk given by Eq. (10.12) smaller than 0. □

The following Theorem extends Theorem 10.2 to a polyhedral cone $C$.

**Theorem 10.3** *The results of Theorem 10.2 hold if the cone $C$ is a polyhedral cone.*

***Proof*** For any polyhedral cone $C$, there exist simplicial cones $C_1, \ldots, C_l$ of dimension $p$ such that $C = \cup_{i=1}^l C_i$ and $C^I \cap C_j^I = \emptyset$ for all $i \neq j$ (see Bock, 1982). Hence, $I_C(X) = \sum_{i=1}^l I_{C_i}(X)$ almost everywhere. Hence,

$$\mathrm{E}_\theta \left[ \left( \|g(X)\|^2 \left( \frac{\|U\|^2}{k+2} \right)^2 + 2(1-\omega) \frac{\|U\|^2}{k+2} g(X)^\top (X-\theta) \right) I_C(X) \right]$$
$$= \sum_{i=1}^{l} \mathrm{E}_\theta \left[ \left( \|g(X)\|^2 \left( \frac{\|U\|^2}{k+2} \right)^2 + 2(1-\omega) \frac{\|U\|^2}{k+2} g(X)^\top (X-\theta) \right) I_{C_i}(X) \right]$$

and the result follows from Theorem 10.2. □

The final result extends the above to a general convex cone.

**Theorem 10.4** *Let $C$ be a convex cone with non-empty interior. The results of Theorems 10.2 and 10.3 remain valid for $C$.*

***Proof*** Let $C_i$ be a sequence of polyhedral cones with non-empty interior that approach $C$ from within. Hence, $I_{C_i}(X)$ approaches $I_C(X)$ from below. Also, the

difference in risk between the shrinkage estimator $\delta_i(X,U) = X + \Gamma_{C_i}(X) + \big(\|U\|^2/(k+2)\big)g(X)I_{C_i}(X)$ and the natural estimator $\delta_{C_i}(X) = X + \Gamma_{C_i}(X)$ is

$$\Delta_i\mathcal{R}(\theta) = \mathrm{E}_\theta\left[\left(\|g(X)\|^2\left(\frac{\|U\|^2}{k+2}\right)^2 + 2(1-\omega)\frac{\|U\|^2}{k+2}g(X)^\top(X-\theta)\right)I_{C_i}(X)\right] \le 0.$$

Since $I_{C_i}(X) \le I_C(X)$ and all risks are assumed finite, we have

$$\lim_{i\to\infty}\Delta_i\mathcal{R}(\theta) = \Delta\mathcal{R}(\theta) \le 0,$$

which completes the proof. □

### *10.2.2 A Note for Nested Cones*

In this subsection, similar to Fourdrinier et al. (2003), we assume there are two positive closed convex $C_1$ and $C_2$ such that $C_1 \subset C_2$ and such that if $\theta \in C_2 - C_1$, there is an orthogonal transformation $Q$ such that:

(*a*) $C_1 \subset QC_2$
(*b*) $Q\theta \in C_1$

Similar to the natural estimator defined in Eq. (10.3), when $\theta$ belongs to one of the cones $C_i$ $(i = 1, 2)$, a natural estimator is $\delta_{0i}(X) = X + \Gamma_{C_i}(X)$. Also, the shrinkage estimator in this case is as

$$\delta_i(X,U) = X + \Gamma_{C_i}(X) - \frac{\|U\|^2}{k+2}g(\|X\|^2)X, \qquad i = 1, 2,$$

where $g(\cdot)$ is a nonnegative function.

We want to show that when $\theta$ is restricted to $C_1$ and an alternative estimator $\delta_1(X,U)$ dominating $\delta_{01}(X)$ is available, we can derive a competitive estimator $\delta_2(X,U)$ dominating $\delta_{02}(X)$ when $\theta$ is restricted to $C_2$.

**Lemma 10.2 (Fourdrinier et al., 2003)** *If $C_1$ and $C_2$ are positive closed convex cones such that $C_1 \subset C_2$, then $X^\top P_{C_1}(X) \le X^\top P_{C_2}(X)$ for all $X \in \mathbb{R}^p$. Hence, also $X^\top\Gamma_{C_1}(X) \le X^\top\Gamma_{C_2}(X)$ where $\Gamma_{C_i}(X) = P_{C_i}(X) - X$ $(i = 1, 2)$.*

**Lemma 10.3 (Fourdrinier et al., 2003)** *If $C$ is a positive closed convex cone and $Q$ is an orthogonal transformation, then $P_{QC}(QX) = QP_C(X)$ and hence $P_{QC}(QX) - QX = Q(P_C(X) - X)$.*

**Theorem 10.5** *Assume conditions (a) and (b) hold and suppose that the target estimator $\delta^*(X) = \delta_{02}(X)$ under the $\mathcal{BLF}$. If the shrinkage estimator $\delta_1(X,U)$ dominates the natural estimator $\delta_{01}(X)$ when $\theta$ is restricted to $C_1$, then the*

*shrinkage estimator* $\delta_2(X, U)$ *dominates the natural estimator* $\delta_{02}(X)$ *when* $\theta$ *is restricted to* $C_2$.

***Proof*** Since $C_1 \subset C_2$, we will separate the two cases: $\theta \in C_1$ and $\theta \in C_2 - C_1$.

Suppose first that $\theta \in C_1$. Note that the risk difference $\Delta_2\mathcal{R}(\theta)$ between $\delta_2(X, U)$ and $\delta_{02}(X)$ is equal to

$$\begin{aligned}
&\mathrm{E}_\theta\left[\omega\left\|\delta_2-\delta_{02}\right\|^2+(1-\omega)\left\|\delta_2-\theta\right\|^2\right]\\
&\quad-\mathrm{E}_\theta\left[\omega\left\|\delta_{02}-\delta_{02}\right\|^2-(1-\omega)\left\|\delta_{02}-\theta\right\|^2\right]\\
&=\mathrm{E}_\theta\left[\omega\left\|-\frac{\|U\|^2}{k+2}g(\|X\|^2)X\right\|^2+(1-\omega)\left\|X+\Gamma_{C_2}(X)-\frac{\|U\|^2}{k+2}g(\|X\|^2)X-\theta\right\|^2\right]\\
&\quad-\mathrm{E}_\theta\left[(1-\omega)\left\|X+\Gamma_{C_2}(X)-\theta\right\|^2\right]\\
&=\mathrm{E}_\theta\left[\frac{\|U\|^4}{(k+2)^2}g^2(\|X\|^2)\|X\|^2-2(1-\omega)\frac{\|U\|^2}{k+2}g(\|X\|^2)X^\top(X+\Gamma_{C_2}(X)-\theta)\right]\\
&=\mathrm{E}_\theta\left[\frac{\|U\|^4}{(k+2)^2}g^2(\|X\|^2)\|X\|^2-2(1-\omega)\frac{\|U\|^2}{k+2}g(\|X\|^2)X^\top(X-\theta)\right.\\
&\quad\left.-2(1-\omega)\frac{\|U\|^2}{k+2}g(\|X\|^2)X^\top\Gamma_{C_2}(X)\right].
\end{aligned}$$

By Lemma 10.2, it is clear that $X^\top\Gamma_{C_2}(X) \geq X^\top\Gamma_{C_1}(X)$ for each $X \in \mathbb{R}^p$. Hence, since $g(\|X\|^2)$ is nonnegative and we assume that domination of $\delta_1(X, U)$ over $\delta_{01}(X)$ for $\theta \in C_1$, we have

$$\begin{aligned}
\Delta_2\mathcal{R}(\theta) &\leq \mathrm{E}_\theta\left[\frac{\|U\|^4}{(k+2)^2}g^2(\|X\|^2)\|X\|^2-2(1-\omega)\frac{\|U\|^2}{k+2}g(\|X\|^2)X^\top(X-\theta)\right.\\
&\qquad\left.-2(1-\omega)\frac{\|U\|^2}{k+2}g(\|X\|^2)X^\top\Gamma_{C_1}(X)\right]\\
&= \Delta_1\mathcal{R}(\theta) \leq 0. \qquad (10.13)
\end{aligned}$$

Suppose now that $\theta \in C_2 - C_1$ and $Q$ is an orthogonal transformation. Let $Y = QX$ and note that $(Y, V)$ is $\mathcal{SS}$ distribution about $(v, 0)$ where $v = Q\theta$. Consider the problem of estimating $v$ when $v$ is restricted to $v = QC_2$. Since $C_1 \subset QC_2$, we

have, by the result in the first case, that for $Q\theta = v \in C_1$. Similar to Eq. (10.13), for $(Y, V)$, we have the following equation:

$$\mathrm{E}_v \left[\omega\| - \frac{\|V\|^2}{k+2} g(\|Y\|^2)Y\|^2 + (1-\omega)\|Y + \Gamma_2^*(Y) - \frac{\|V\|^2}{k+2} g(\|Y\|^2)Y - v\|^2\right] - \mathrm{E}_v\left[(1-\omega)\|Y + \Gamma_2^*(Y) - v\|^2\right] \le 0.$$

Using the orthogonality property of $Q$, we know $Q^\top Q = I$. Substituting this, we get $(QV)^\top(Qv) = V^\top Q^\top QV = V^\top IV = V^\top V$ and $\|QV\|^2 = \|V\|^2$. Now, we have equivalently

$$\begin{aligned}\mathrm{E}_\theta \Bigg[&\omega \frac{\|QV\|^4}{(k+2)^2} g^2(\|QX\|^2)\|QX\|^2 \\ &+(1-\omega)\|QX + \Gamma_2^*(QX) - \frac{\|QV\|^2}{k+2} g(\|QX\|^2)QX - Q\theta\|^2\Bigg] \\ &-(1-\omega)\,\mathrm{E}_\theta\left[\|QX + \Gamma_2^*(QX) - Q\theta\|^2\right] \le 0.\end{aligned}$$

where $\Gamma_2^*(QX) = P_{QC_2}(QX) - QX = Q(P_{C_2}(X) - X) = Q\Gamma_{C_2}(X)$ as shown in Lemma 10.3. Hence,

$$\begin{aligned}&\mathrm{E}_\theta\left[\omega \frac{\|QV\|^4}{(k+2)^2} g^2(\|QX\|^2)\|QX\|^2 + (1-\omega)\|QX + Q\Gamma_{C_2}(X)\right. \\ &\left. - \frac{\|QV\|^2}{k+2} g(\|QX\|^2)QX - Q\theta\|^2\right] - (1-\omega)\,\mathrm{E}_\theta\left[\|QX + Q\Gamma_{C_2}(X) - Q\theta\|^2\right] \\ &= \mathrm{E}_\theta\left[\omega \frac{\|U\|^4}{(k+2)^2} g^2(\|X\|^2)\|X\|^2\right. \\ &\left. +(1-\omega)\left\|QX + Q\Gamma_{C_2}(X) - \frac{\|U\|^2}{k+2} g(\|X\|^2)QX - Q\theta\right\|^2\right] \\ &-(1-\omega)\,\mathrm{E}_\theta[\|QX + Q\Gamma_{C_2}(X) - Q\theta\|^2] \\ &= \mathrm{E}_\theta\left[\omega \frac{\|U\|^4}{(k+2)^2} g^2(\|X\|^2)\|X\|^2\right. \\ &\left. +(1-\omega)\|Q(X + \Gamma_{C_2}(X) - \frac{\|U\|^2}{k+2} g(\|X\|^2)X - \theta)\|^2\right] \\ &-(1-\omega)\,\mathrm{E}_\theta\left[\|Q(X + \Gamma_{C_2}(X) - \theta)\|^2\right]\end{aligned}$$

$$= \mathrm{E}_\theta \left[ \omega \frac{\|U\|^4}{(k+2)^2} g^2(\|X\|^2)\|X\|^2 \right.$$
$$\left. +(1-\omega)\left\| (X + \Gamma_{C_2}(X) - \frac{\|U\|^2}{k+2} g(\|X\|^2)X - \theta) \right\|^2 \right]$$
$$-(1-\omega)\,\mathrm{E}_\theta \left[ \|(X + \Gamma_{C_2}(X) - \theta)\|^2 \right]$$
$$= \Delta_2 \mathcal{R}(\theta).$$

Now, since in the first case of the theorem we showed that for $\theta \in C_1$, $\Delta_2\mathcal{R}(\theta) \leq \Delta_1\mathcal{R}(\theta) \leq 0$, using this inequality, we also have $\Delta_2\mathcal{R}(\theta) \leq 0$. □

### *10.2.3 The Restricted Shrinkage Generalized Bayes Estimator*

The generalized Bayes estimator marks a significant evolution in statistical inference, offering a flexible approach to decision-making that admits uncertainty. Expanding upon the core tenets of Bayesian statistics, this estimator broadens the horizon of inference by integrating additional information beyond the observed data. Central to its methodology is the amalgamation of prior knowledge , typically represented through prior distributions, with the likelihood function derived from observed data. This integration enables the estimator to render more informed and robust decisions, especially in scenarios with limited sample sizes or noisy data.

An inherent strength of the generalized Bayes estimator lies in its adaptability to diverse decision-making contexts, achieved through the specification of relevant prior distributions. By carefully selecting priors that encapsulate pertinent information about the parameters of interest, statisticians can tailor the estimator to address specific application domains and decision scenarios.

Furthermore, the generalized Bayes estimator offers versatility in incorporating various types of prior information, including subjective assessments, historical data, expert insights, and external sources. This adaptability renders it well-suited for tackling intricate inference challenges across a spectrum of disciplines, including engineering, economics, medicine, and the social sciences.

In this subsection, we consider generalized Bayes (GB) estimators of $\theta$ for priors of the form $\pi(\|\theta\|^2)$. Any shrinkage estimator $\delta(X, U)$ of $\theta$ is evaluated by its risk function associated with $\mathcal{BLF}$. We know that the Bayes estimator under quadratic loss function $\|\delta - \theta\|^2$ for the priors given by $\pi(\|\theta\|^2)$ is equal to the expected value of posterior distribution as

$$\mathrm{E}\,[\theta|(X,U)] = X + \frac{1}{m(X,U)} \int_{\mathbb{R}^p} (\theta - X) f(\|X-\theta\|^2 + \|U\|^2)\pi(\|\theta\|^2)\mathrm{d}\theta, \tag{10.14}$$

where $m(X, U)$ is the marginal density as

$$m(X, U) = \int_{\mathbb{R}^p} f(\|X - \theta\|^2 + \|U\|^2)\pi(\|\theta\|^2)\mathrm{d}\theta\,.$$

Now, define for any $(X, U) \in \mathbb{R}^p \times \mathbb{R}^k$,

$$M(X, U) = \int_{\mathbb{R}^p} F(\|X - \theta\|^2 + \|U\|^2)\pi(\|\theta\|^2)\mathrm{d}\theta\,.$$

By this definition, the expected value in (10.14) is

$$\mathrm{E}\,[\theta|(X, U)] = X + \frac{\nabla_X M(X, U)}{m(X, U)}. \tag{10.15}$$

The target estimator $\delta^*(X) = \delta_{02}(X)$ under $\mathcal{BLF}$. If the shrinkage estimator $\delta_1(X, U)$ dominates the natural estimator $\delta_{01}(X)$ when $\theta$ is restricted to $C_1$, then the shrinkage estimator $\delta_2(X, U)$ dominates the natural estimator $\delta_{02}(X)$ when $\theta$ is restricted to $C_2$.

**Theorem 10.6** *Suppose that $(X, U)$ is a $p + k$ random vector having a $\mathcal{SS}$ distribution, under the $\mathcal{BLF}$, the GB estimator for $\theta$ w.r.t. the prior $\pi(\|\theta\|^2)$ is*

$$\delta_\pi(X, U) = \omega\delta^*(X) + (1 - \omega)\left(X + \frac{\nabla_X M(X, U)}{m(X, U)}\right) I_C(X), \tag{10.16}$$

*where $\delta^*(X)$ is an arbitrary target estimator in $\mathcal{BLF}$ and $C$ is a closed convex cone.*

***Proof*** For the arbitrary prior distribution $\pi(\|\theta\|^2)$ and the posterior distribution $\pi(\theta|X)$, it is straightforward to check that the posterior expected loss of an arbitrary estimator $\delta(X)$ under the $\mathcal{BLF}$ in Eq. (10.2) is given by

$$\begin{aligned}\rho\left(\pi(\theta|X), \delta(X)\right) &= \int_\Theta L_{\omega,\delta_0}(\theta, \delta)\pi(\theta|X)\mathrm{d}\theta \\ &= \mathrm{E}\left[\omega\|\delta(X) - \delta^*(X)\|^2 + (1 - \omega)\|\delta(X) - \theta\|^2 \middle| X\right].\end{aligned}$$

A GB estimator is obtained by

$$\begin{aligned}\frac{\partial\rho\left(\pi(\theta|X), \delta(X)\right)}{\partial\delta(X)} &= 2\omega\big(\delta(X) - \delta^*(X)\big) + 2(1 - \omega)\big(\delta(X) - \mathrm{E}\,[\theta|X]\big) \\ &= 2\delta(X) - 2\omega\delta^*(X) - 2(1 - \omega)\,\mathrm{E}\,[\theta|X] = 0;\end{aligned}$$

then, by condition with respect to $(X, U)$, we have

$$\delta_\pi(X, U) = \omega\delta^*(X) + (1 - \omega)E\left(\theta|(X, U)\right)\,.$$

Then, by using the expected value in Eq. (10.15), the GB estimator $\delta_\pi(X, U)$ can be written as

$$\delta_\pi(X, U) = \omega\delta^*(X) + (1-\omega)\left(X + \frac{\nabla_X M(X, U)}{m(X, U)}\right).$$

□

If we suppose that the target estimator $\delta^*(x)$ is equal to the natural estimator $\delta_0(X)$ in Eq. (10.3) and restricted to a cone, then the restricted generalized Bayes (RGB) estimator $\delta_\pi(X, U)$ is as

$$\delta_{R,\pi}(X, U) = X + \omega\Gamma_C(X) + (1-\omega)\left(\frac{\nabla_X M(X, U)}{m(X, U)}\right).$$

Using Theorem 10.6, the general form of the RGB estimators can be considered as follows:

$$\delta_{R,\pi}(X, U) = X + \omega\Gamma_C(X) + (1-\omega)g(X, U). \tag{10.17}$$

Note that the function $m(X, U)$, which is the marginal density, must be positive and differentiable to allow the application of the relationship $\nabla_X M(X, U)/m(X, U)$ and the function $M(X, U)$ should be continuous and differentiable and numerically behave in a way that ensures the proper computation of the expected value $\mathrm{E}[\theta|(X, U)]$.

**Example 10.1** Suppose that $X|\theta \sim N_p(\theta, \sigma^2 I_p)$ and the prior distribution $\theta \sim N_p(\mu, \tau^2 I_p)$. The posterior distribution $\pi(\theta|X)$ is

$$\begin{aligned}
\pi(\theta|X) &\propto f(X|\theta)\,\pi(\theta)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left(x^\top x + \theta^\top\theta - 2\theta^\top x\right) - \frac{1}{2\tau^2}\theta^\top\theta\right\}\\
&\propto \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\theta^\top\theta + \frac{1}{\sigma^2}\theta^\top x\right\}\\
&= \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left(\theta^\top\theta - \frac{2\tau^2}{\tau^2+\sigma^2}\theta^\top x\right)\right\}\\
&\propto \exp\left\{-\frac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\left\|\theta - \frac{\tau^2}{\tau^2+\sigma^2}x\right\|^2\right\}\\
&\sim N_p\left(\frac{\tau^2}{\tau^2+\sigma^2}X, \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}I_p\right).
\end{aligned}$$

Now, assume the following domination conditions on $\tau$ and $\sigma$:

- When $\sigma^2 \geq \tau^2$, meaning the noise variance dominates the prior variance, which is typical when the observed data carries more information than the prior belief
- When $\tau^2 \geq \sigma^2$, meaning the prior variance dominates, often encountered when the prior knowledge is strong relative to the observed data

Then, similar to Eq. (10.16), the RGB estimator $\delta_\pi(X)$ is

$$\delta_\pi(X) = X + \omega\Gamma_C(X) + (1-\omega)\frac{\tau^2}{\tau^2+\sigma^2}X.$$

### *10.2.4 The Restricted Soft Threshold Wavelet Shrinkage Estimator*

The wavelet shrinkage method holds significant importance in signal and image processing, especially for denoising purposes. In today's data-driven world, where noise often contaminates signals and images, the capability to remove noise while preserving essential features is crucial. Wavelet shrinkage presents a potent solution to this challenge, utilizing the distinctive characteristics of wavelet transforms to efficiently reduce noise without compromising signal fidelity.

This technique works by breaking down the signal or image into its wavelet coefficients, which represent different frequency components at varying scales. By applying thresholding or shrinkage to these coefficients, the method selectively diminishes noise while retaining the signal components, resulting in an improved signal-to-noise ratio. Through adaptive adjustment of the thresholding parameters based on the data's statistical properties, wavelet shrinkage can outperform traditional denoising methods. The basic idea is to transform the noisy signal into the wavelet domain, shrink the wavelet coefficients, and then transform the signal back to the original domain.

The significance of wavelet shrinkage extends beyond denoising tasks, finding widespread application in fields such as medical imaging, telecommunications, and financial signal processing. Its capability to capture both local and global signal characteristics efficiently makes it suitable for diverse applications where preserving signal integrity is critical.

The soft threshold wavelet shrinkage (STWS) estimator of Donoho and Johnstone (1995) is given by

$$\delta_\lambda^{\text{soft}}(X_i) = (X_i - \text{sgn}(X_i)\lambda)I(|X_i| > \lambda).$$

The shrinkage form of $\delta_\lambda^{\text{soft}}(X)$ estimator is

$$\delta_\lambda^{\text{soft}}(X) = X + g(X), \tag{10.18}$$

where $g(X) = (g(X_1), \ldots, g(X_p))^\top$, and $g(X_i)$ is defined as

$$g(X_i) = \begin{cases} \lambda & X_i < -\lambda, \\ -X_i & |X_i| \leq \lambda, \\ -\lambda & X_i > \lambda. \end{cases} \tag{10.19}$$

So far, we have considered using different types of thresholds as shrinkage wavelets, such as the universal $\lambda^* = \sqrt{2\sigma^2 \log p}$ ($p$ is the dimension of the parameter vector and $\sigma^2$ is the variance), the Lorenz method (Lorentz 1905), the percentile method (Mallat, 1989), the false discovery rate method (Benjamini & Hochberg, 1995), cross-validation (CV) (Nason, 1996), and block thresholds (Cai & Silverman, 2001), as well as Stein's unbiased risk estimate (SURE) (Donoho & Johnstone, 1995). The results of this subsection depend on the following lemma.

**Lemma 10.4 (Stein, 1981)** *Suppose that $X \sim N_p(\theta, \sigma^2 I_p)$, with known $\sigma^2$. Then,*

$$\mathrm{E}_\theta[(X-\theta)^\top g(X)] = \sigma^2 \,\mathrm{E}[\nabla_X \cdot g(X)],$$

*where $g : \mathbb{R}^p \to \mathbb{R}^p$ is a function for which the two expectations $\mathrm{E}_\theta[(X-\theta)^\top g(X)]$ and $\mathrm{E}[\nabla_X \cdot g(X)]$ both exist.*

One of the methods of selecting the threshold is minimizing the SURE, which was introduced by Donoho and Johnstone (1995). Suppose that $\theta = (\theta_1, \ldots, \theta_p)^\top$ and $X_i \sim N(\theta_i, 1)$, for $i = 1, \ldots, p$. The function $g(X)$ in the shrinkage estimator representation $\delta_\lambda^{\text{soft}}(X)$ in Eq. (10.18) is weakly differentiable. Under the quadratic loss function $\|\delta(X) - \theta\|^2$, by Lemma 10.4, we have

$$\mathcal{R}(\theta, \delta(X)) = \mathrm{E}_\theta\left[\|\delta(X) - \theta\|^2\right] = p + \mathrm{E}_\theta\left[\|g(X)\|^2\right] + \mathrm{E}_\theta\left[2\nabla_X \cdot g(X)\right]. \tag{10.20}$$

Now, by using Lemma 10.4, we have from the risk given by Eq. (10.20) that the quantity

$$\mathrm{SURE}(X, \lambda) = p - 2\sum_{i=1}^{p} I(|X_i| \leq \lambda) + \sum_{i=1}^{p} (|X_i| \wedge \lambda)^2,$$

is an unbiased estimate of the risk; that is, $\mathrm{E}_\theta\left[\|\delta_\lambda^{\text{soft}}(X) - \theta\|^2\right] = \mathrm{E}_\theta\left[\mathrm{SURE}(X, \lambda)\right]$. Consider using this estimator to select a threshold as

$$\lambda^{\text{sure}} = \underset{0 \leq \lambda \leq \lambda^*}{\operatorname{argmin}} \left\{\mathrm{SURE}(X, \lambda)\right\}.$$

Now, we can achieve the SURE thresholds for two following shrinkage estimators $\delta(X, U)$ in Eq. (10.4) and $\delta_\pi(X, U)$ in Eq. (10.17). We suppose that $g(X, U) = g(X)$ in Eq. (10.19), and then the restricted soft threshold wavelet shrinkage (RSTWS) estimator can be the same as $\delta_\lambda^{\text{soft}}(X)$ in Eq. (10.18) as

$$\delta_\lambda^{\text{soft}}(X, U) = X + \Gamma_C(X) + \frac{\|U\|^2}{k+2} g(X) I_C(X), \tag{10.21}$$

$$\delta_{\pi,\lambda}^{\text{soft}}(X) = X + \omega\Gamma_C(X) + (1-\omega) g(X) I_C(X). \tag{10.22}$$

In this subsection, similar to the previous subsection, we suppose that the target estimator $\delta^*(X) = \delta_0(X)$ in Eq. (10.3). With a same approach to Donoho and Johnstone (1995), we have the following theorems.

**Theorem 10.7** *Suppose that $(X, U)$ is a $(p + k)$ random vector having a $\mathcal{SS}$ distribution. For the RSTWS estimator $\delta_\lambda^{\text{soft}}(X, U)$ in Eq. (10.21), the value of the threshold is given by*

$$\lambda^{\text{sure}} = \underset{0 \le \lambda \le \lambda^*}{\text{argmin}}\, \text{SURE}(\delta_\lambda^{\text{soft}}(X), \delta_0(X)),$$

*where* $\text{SURE}(\delta_\lambda^{\text{soft}}(X), \delta_0(X))$ *is as follows:*

$$\begin{aligned}
&\frac{\mathbb{N}\|U\|^2}{2(k+2)} \sum_{i=1}^{p} (|X_i| \wedge \lambda)^2 I_C(X) + p(1-\omega) + \sum_{i=1}^{p} X_i^2 I(X_i < 0) \\
&\quad - \frac{\mathbb{N}\|U\|^2}{k+2} (1-\omega) \sum_{i=1}^{p} I(|X_i| \le \lambda) I_C(X) + \mathbb{N}(1-\omega) \sum_{i=1}^{p} I(X_i < 0),
\end{aligned}$$

*where $\mathbb{N}$ is the normalizing constant and $\lambda^*$ is the universal threshold.*

***Proof*** By Lemma 10.1 and under $\mathcal{BLF}$ in Eq. (10.2), the risk $R(\theta, \delta_\lambda^{\text{soft}}(X, U))$ is given as follows:

$$\begin{aligned}
&\text{E}\left[\omega \|\delta_\lambda^{\text{soft}}(X, U) - \delta_0(X)\|^2 + (1-\omega)\|\delta_\lambda^{\text{soft}}(X, U) - \theta\|^2\right] \\
&= \text{E}\left[\omega \left\| \frac{\|U\|^2}{k+2} g(X) I_C(X) \right\|^2 + (1-\omega) \left\| X + \Gamma_C(X) + \frac{\|U\|^2}{k+2} g(X) I_C(X) - \theta \right\|^2\right] \\
&= \text{E}\left[ \frac{\|U\|^4}{(k+2)^2} \|g(X)\|^2 I_C(X) + (1-\omega)\|X - \theta\|^2 + (1-\omega)\|\Gamma_C(X)\|^2 \right. \\
&\qquad \left. + 2\frac{\|U\|^2}{k+2}(1-\omega)(X-\theta)^\top g(X) I_C(X) + 2(1-\omega)(X-\theta)^\top \Gamma_C(X) \right.
\end{aligned}$$

$$+2\frac{\|U\|^2}{k+2}(1-\omega)\Gamma_C^{\top}(X)g(X)I_C(X)\Big]$$

$$=\mathrm{E}\Big[\frac{\mathbb{N}\|U\|^2}{2(k+2)}\sum_{i=1}^{p}(|X_i|\wedge\lambda)^2 I_C(X)+p(1-\omega)+(1-\omega)\sum_{i=1}^{p}X_i^2 I(X_i<0)$$

$$+\frac{\mathbb{N}\|U\|^2}{k+2}(1-\omega)\nabla_X\cdot g(X)I_C(X)+\mathbb{N}(1-\omega)\nabla_X\cdot\Gamma_C(X)\Big].$$

Since $\nabla_X\cdot g(X)=-\sum_{i=1}^{p}I(|X_i|\le\lambda)$ and $\nabla_X\cdot\Gamma_C(X)=\sum_{i=1}^{p}I(X_i<0)$, then where the $\mathrm{SURE}(\delta_\lambda^{\mathrm{soft}}(X),\delta_0(X))$ is as

$$\frac{\mathbb{N}\|U\|^2}{2(k+2)}\sum_{i=1}^{p}(|X_i|\wedge\lambda)^2 I_C(X)+p(1-\omega)+\sum_{i=1}^{p}X_i^2 I(X_i<0)$$

$$-\frac{\mathbb{N}\|U\|^2}{k+2}(1-\omega)\sum_{i=1}^{p}I(|X_i|\le\lambda)I_C(X)+\mathbb{N}(1-\omega)\sum_{i=1}^{p}I(X_i<0),$$

and the SURE is an unbiased estimate of risk, that is,

$$\mathrm{E}\Big[L_{\omega,\delta_0}(\theta,\delta_\lambda^{\mathrm{soft}}(X))\Big]=\mathrm{E}\Big[\mathrm{SURE}(\delta_\lambda^{\mathrm{soft}}(X),\delta_0(X))\Big],$$

and $\lambda^{\mathrm{sure}}=\underset{0\le\lambda\le\lambda^*}{\mathrm{argmin}}\big\{\mathrm{SURE}_{(X,\lambda)}(\delta_\lambda^{\mathrm{soft}}(X),\delta_0(X))\big\}$. □

**Corollary 10.1** *Suppose that $X\sim N_p(\theta,\sigma^2 I_p)$ with known $\sigma^2$. For the RSTWS estimator $\delta_\lambda^{\mathrm{soft}}(X,U)$ in Eq. (10.21), the value of the threshold is given by*

$$\lambda^{\mathrm{sure}}=\underset{0\le\lambda\le\lambda^*}{\mathrm{argmin}}\big\{\mathrm{SURE}(\delta_\lambda^{\mathrm{soft}}(X),\delta_0(X))\big\},$$

*where* $\mathrm{SURE}(\delta_\lambda^{\mathrm{soft}}(X),\delta_0(X))$ *is given as*

$$\sum_{i=1}^{p}(|X_i|\wedge\lambda)^2 I_C(X)+p(1-\omega)+\sum_{i=1}^{p}X_i^2 I(X_i<0)$$
$$-(1-\omega)\sum_{i=1}^{p}I(|X_i|\le\lambda)I_C(X)+(1-\omega)\sum_{i=1}^{p}I(X_i<0).$$

***Proof*** In the proof of Theorem 10.7, use Lemma 10.4 instead of Lemma 10.1. □

**Theorem 10.8** *Suppose that* $X \sim N_p(\theta, \sigma^2 I_p)$. *For the RSTWS estimator* $\delta^{\mathrm{soft}}_{\pi,\lambda}(X)$ *in Eq.* (10.22)*, the value of the threshold is given by*

$$\lambda^{\mathrm{sure}} = \underset{0\le\lambda\le\lambda^*}{\operatorname{argmin}} \left\{ \mathrm{SURE}(\delta^{\mathrm{soft}}_{\pi,\lambda}(X), \delta_0(X)), \right\}$$

*where* $\mathrm{SURE}(\delta^{\mathrm{soft}}_{\pi,\lambda}(X), \delta_0(X))$ *is given as*

$$\begin{aligned}
&\omega(1-\omega)\sum_{i=1}^{p} X_i^2 I(X_i < 0) + (1-\omega)^2 \sum_{i=1}^{p} (|X_i| \wedge \lambda)^2 I_C(X) \\
&\quad -2(1-\omega)\sum_{i=1}^{p} I(|X_i| \le \lambda) I_C(X) + 2(1-\omega)\sum_{i=1}^{p} I(X_i < 0) + p(1-\omega).
\end{aligned}$$

***Proof*** By Lemma 10.4 and under $\mathcal{BLF}$ in Eq. (10.2), the risk $R(\theta, \delta^{\mathrm{soft}}_{\pi,\lambda}(X))$ is given as

$$\begin{aligned}
&\mathrm{E}\left[\omega\|\delta^{\mathrm{soft}}_{\pi,\lambda}(X) - \delta_0(X)\|^2 + (1-\omega)\|\delta^{\mathrm{soft}}_{\pi,\lambda}(X) - \theta\|^2\right] \\
&= \mathrm{E}\Big[\omega\|(\omega-1)\Gamma_C(X) + (1-\omega)g(X)I_C(X)\|^2 \\
&\qquad +(1-\omega)\|X + \omega\Gamma_C(X) + (1-\omega)g(X)I_C(X) - \theta\|^2\Big] \\
&= \mathrm{E}\Big[\omega(1-\omega)\|\Gamma_C(X)\|^2 + (1-\omega)^2\|g(X)\|^2 I_C(X) \\
&\qquad +2(1-\omega)(X-\theta)^\top g(X) I_C(X) + 2(1-\omega)(X-\theta)^\top \Gamma_C(X)\Big] + p(1-\omega) \\
&= \mathrm{E}\Big[\omega(1-\omega)\sum_{i=1}^{p} X_i^2 I(X_i < 0) + (1-\omega)^2 \sum_{i=1}^{p} (|X_i| \wedge \lambda)^2 I_C(X) \\
&\qquad +2(1-\omega)\nabla_X \cdot g(X) I_C(X) + 2(1-\omega)\nabla_X \cdot \Gamma_C(X)\Big] + p(1-\omega),
\end{aligned}$$

then, $\mathrm{SURE}(\delta^{\mathrm{soft}}_{\pi,\lambda}(X), \delta_0(X))$ is

$$\begin{aligned}
&\omega(1-\omega)\sum_{i=1}^{p} X_i^2 I(X_i < 0) + (1-\omega)^2 \sum_{i=1}^{p} (|X_i| \wedge \lambda)^2 I_C(X) \\
&\quad -2(1-\omega)\sum_{i=1}^{p} I(|X_i| \le \lambda) I_C(X) + 2(1-\omega)\sum_{i=1}^{p} I(X_i < 0) + p(1-\omega),
\end{aligned}$$

and the SURE is an unbiased estimate of risk, that is,

$$\mathrm{E}\left[L(\theta, \delta^{\mathrm{soft}}_{\pi,\lambda}(X))\right] = \mathrm{E}\left[\mathrm{SURE}(\delta^{\mathrm{soft}}_{\pi,\lambda}(X), \delta_0(X))\right],$$

and $\lambda^{\mathrm{sure}} = \underset{0\le\lambda\le\lambda^*}{\mathrm{argmin}} \left\{\mathrm{SURE}(\delta^{\mathrm{soft}}_{\pi,\lambda}(X), \delta_0(X))\right\}$. □

## 10.3 Simulation and Application

In this section, we validated theoretical results through simulation studies aimed at assessing the efficacy of the novel RSTWS estimators. All simulations were conducted using the mvtnorm and wavethresh packages in R-4.2.3.

To investigate the risk of estimators, a Monte Carlo simulation study was performed to compare the risk value of estimators for $N_8(\theta, \sigma^2 I_4)$ and $N_{16}(\theta, \sigma^2 I_8)$ where the selected value of $\sigma^2$ is equal to 0.5, 1, 2.

Table 10.1 presents the average of risk for the RSTWS estimators for $p$, $\omega$, and $\sigma^2$ by using the target estimator $\delta_0(X)$. To calculate the average of risk, we first calculate the risk values through 1, 000 iterations of Monte Carlo simulation and then calculate the average of these 1, 000 iterations. In this table, we suppose that $U \sim N_p(0, I_p)$ and $\theta$ is selected as $\left(\sqrt{j}, 0, \ldots, 0\right)$ and $j = 0, 0.1, 0.2, \ldots, 10$. In this case, $\|\theta\| = \theta^\top\theta = \sum_{i=1}^{p} \theta_i = j$.

As can be seen in Table 10.1, by increasing the values of $\sigma^2$, the average of risk increases. Also, the average of risk values obtained in the RSTWS estimator $\delta^{\mathrm{soft}}_{\pi,\lambda}(X)$ is lower than that of the other estimators for $\omega = 0.8$ (bold in table). In general, by decreasing the values of dim $p$, the values of the average of risk decreases. For the RSTWS estimators $\delta^{\mathrm{soft}}_{\lambda}(X, U)$ and $\delta^{\mathrm{soft}}_{\pi,\lambda}(X)$, by increasing the values of $\omega$, the values of the average of risk decrease, the opposite of the estimator $\delta^{\mathrm{soft}}_{\lambda}(X)$.

The risk values of the RSTWS estimators w.r.t. $\|\theta\|$ are plotted in Figs. 10.1, 10.2, 10.3 for selected values of $p = 8, 16$, $\omega = 0.2, 0.5, 0.8$, and $\sigma^2 = 0.5, 1, 2$. These figures are plotted using the simulated data in Table 10.2, with the difference that it shows the risk behavior for each value of $\|\theta\|$. In these figures by increasing the values of $\sigma^2$ and dimension $p$, the average of risk increases. In Fig. 10.1, when $\omega = 0.2$ and $\sigma^2 = 2$, the RSTWS estimator $\delta^{\mathrm{soft}}_{\lambda}(X)$ is better than other estimators. In all other cases, the RSTWS estimator $\delta^{\mathrm{soft}}_{\lambda}(X, U)$ on average is better than other estimators.

### *10.3.1 Parkinson Speech Dataset*

In this subsection, we assess the effectiveness of the RSTWS estimators on a real dataset, a task that poses a significant challenge in this field. Our objective is to

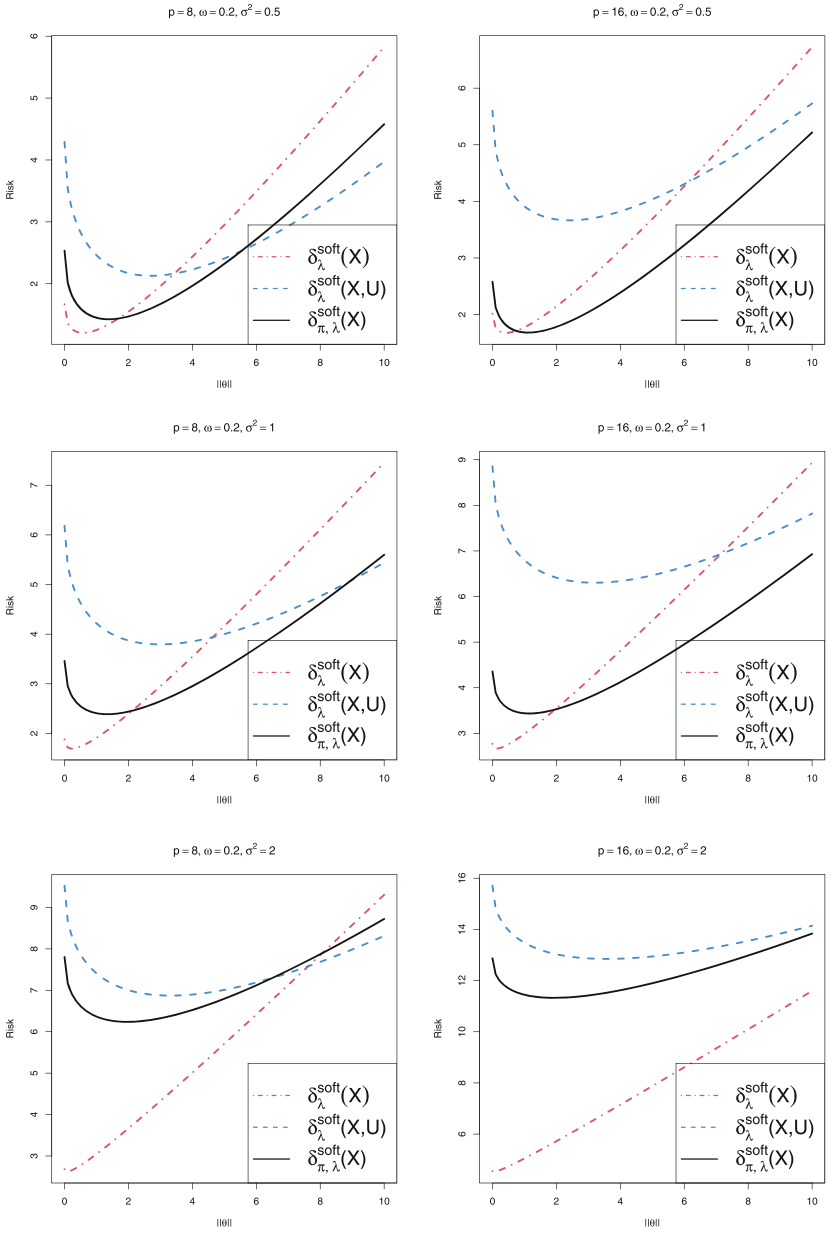



**Fig. 10.1** The risk values for the RSTWS estimators with $\omega = 0.2$ and selected values of $p$ and $\sigma^2$

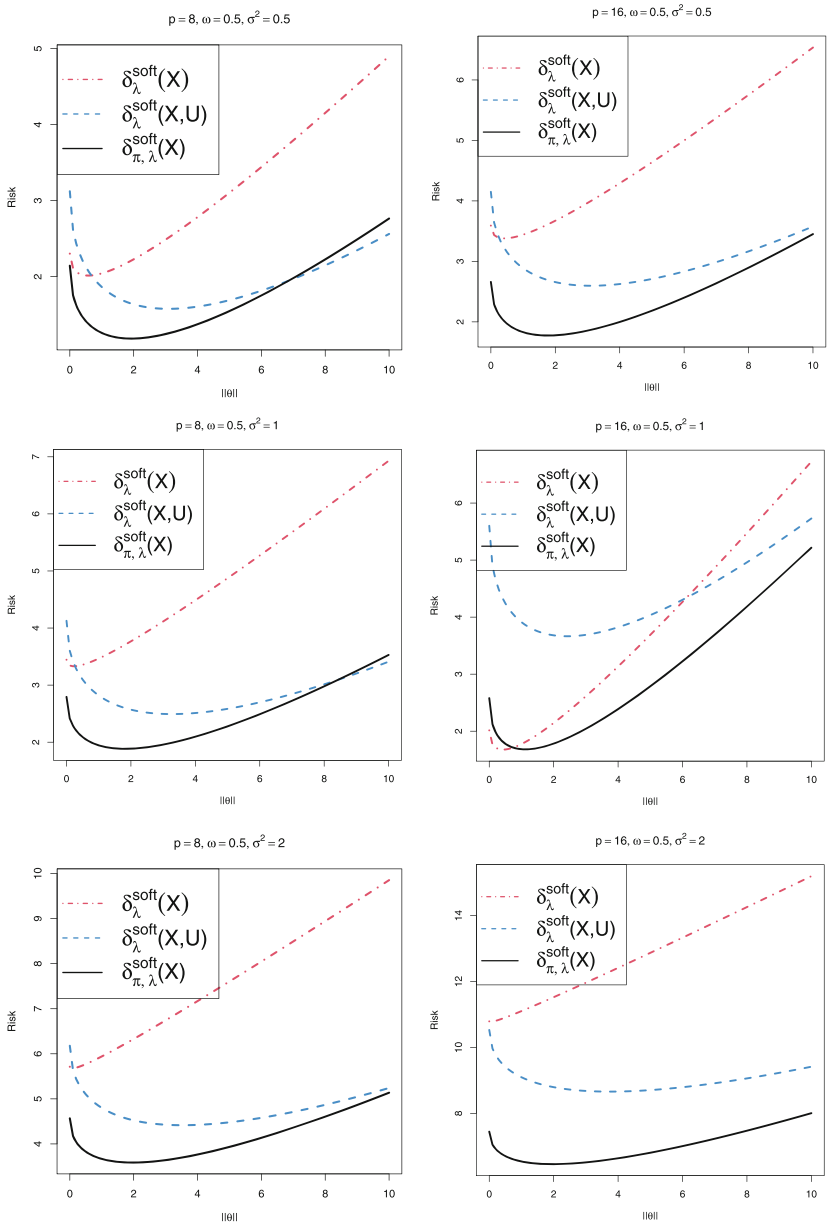



**Fig. 10.2** The risk values for the RSTWS estimators with $\omega = 0.5$ and selected values of $p$ and $\sigma^2$

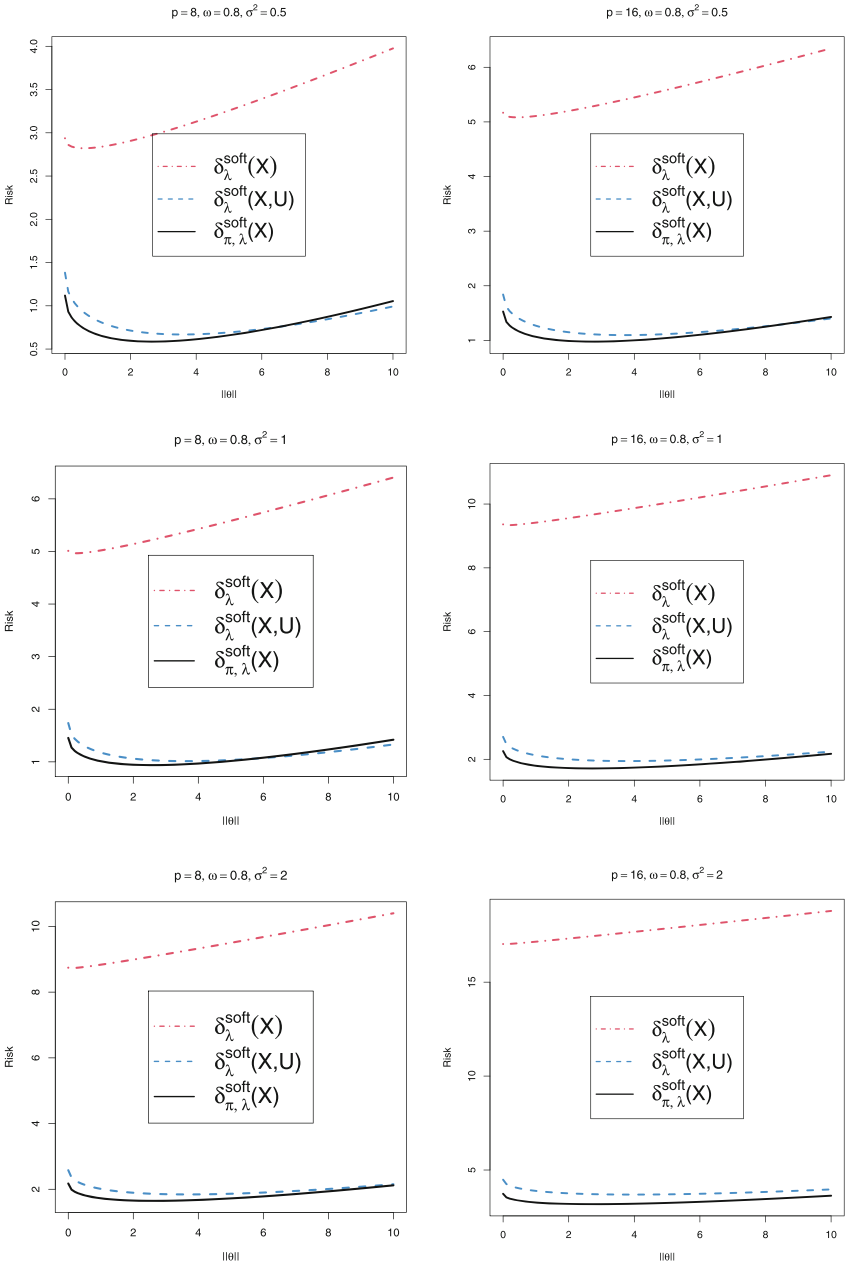



**Fig. 10.3** The risk values for the RSTWS estimators with $\omega = 0.8$ and selected values of $p$ and $\sigma^2$

**Table 10.2** The average of risk for the RSTWS estimators

| | | $\delta^{\text{soft}}_{\pi,\lambda^{\text{sure}}}(X)$ | | | $\delta^{\text{soft}}_{\lambda^{\text{sure}}}(X,U)$ | | | $\delta^{\text{soft}}_{\lambda^*}(X)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | $p$ | $\omega=0.2$ | $\omega=0.5$ | $\omega=0.8$ | $\omega=0.2$ | $\omega=0.5$ | $\omega=0.8$ | $\omega=0.2$ | $\omega=0.5$ | $\omega=0.8$ |
| 0.5 | 8 | 2.574 | 1.732 | **0.746** | 2.761 | 1.926 | 0.794 | 3.113 | 3.205 | 3.298 |
| | 16 | 3.015 | 2.362 | **1.133** | 4.368 | 2.945 | 1.219 | 3.828 | 4.725 | 5.621 |
| 1 | 8 | 3.563 | 2.458 | **1.102** | 4.371 | 2.826 | 1.137 | 4.266 | 4.937 | 5.608 |
| | 16 | 4.747 | 3.692 | **1.871** | 6.847 | 4.758 | 2.066 | 5.549 | 7.803 | 10.058 |
| 2 | 8 | 7.096 | 4.123 | **1.809** | 7.340 | 4.731 | 1.966 | 5.767 | 7.641 | 9.515 |
| | 16 | 12.193 | 6.994 | **3.337** | 13.346 | 8.964 | 3.811 | 7.916 | 12.894 | 17.872 |




**Fig. 10.4** Wavelet shrinkage method

denoise the dataset using the wavelet shrinkage method algorithm. To achieve this, we first introduce white noise to the dataset, resulting in noisy data. Subsequently, employing the denoising technique on the noisy data, we aim to extract the underlying main dataset.

To accomplish this, we utilize the SURE thresholds discussed in Sect. 10.2.4 to convert the noisy data into wavelet coefficients via discrete wavelet transformation (DWT). We then employ the STWS estimator to threshold (or shrink) these wavelet coefficients, a process known as shrinkage. Finally, by applying the inverse discrete wavelet transformation (Inv. DWT) to the denoised coefficients, we obtain an approximation of the main dataset, known as the wavelet shrinkage method. The steps involved in the wavelet shrinkage method are illustrated in Fig. 10.4.

To achieve this objective, we initially compute the mean squared error (MSE) value for each row of data. Subsequently, we calculate the average of these MSE values to obtain the average mean squared error (AMSE).

The Parkinson speech dataset comprises recordings from 20 individuals with Parkinson's disease (PD) and 20 healthy subjects, as shown by Sakar et al. (2013). Each subject contributed multiple types of sound recordings (26 in total). The PD database encompasses both training and test files. The training data originates from 20 individuals with Parkinson's disease (6 female, 14 male) and 20 healthy individuals (10 female, 10 male) who visited the Department of Neurology at Cerrahpasa Faculty of Medicine, Istanbul University. Various types of sound recordings were collected from each subject.

In this study, we utilized only 16 variables from the training dataset, comprising a total of 1040 observations. The dataset is "Parkinson speech dataset" that can be accessed via https://doi.org/10.24432/C5NC8M.

Table 10.3 presents the AMSE based on the RSTWS estimators, the soft universal threshold $\delta^{\text{soft}}_{\lambda^*}(X)$, the soft CV threshold, and the soft Bayes threshold estimator

**Table 10.3** The AMSE values for the RSTWS estimators, the soft universal threshold, the soft CV threshold, and the soft Bayes threshold estimator for the Parkinson speech dataset

| Filter | | $\delta^{\text{soft}}_{\pi,\lambda^{\text{sure}}}(X)$ | | | $\delta^{\text{soft}}_{\lambda^{\text{sure}}}(X,U)$ | | | | | Bayes |
|---|---|---|---|---|---|---|---|---|---|---|
| number | Family | $\omega=0.2$ | $\omega=0.5$ | $\omega=0.8$ | $\omega=0.2$ | $\omega=0.5$ | $\omega=0.8$ | $\delta^{\text{soft}}_{\lambda^*}(X)$ | CV | soft |
| 5 | DLA | **0.4945** | 0.5216 | 0.5901 | 0.5673 | 0.6716 | 0.8416 | 0.6364 | 0.5990 | 0.6812 |
| 10 | DLA | **0.4934** | 0.5180 | 0.5829 | 0.5667 | 0.6720 | 0.8419 | 0.6217 | 0.5613 | 0.6648 |
| 5 | DEP | **0.5060** | 0.5376 | 0.6111 | 0.5725 | 0.6755 | 0.8426 | 0.6556 | 0.6265 | 0.7026 |
| 10 | DEP | **0.5087** | 0.5421 | 0.6128 | 0.5729 | 0.6751 | 0.8420 | 0.6541 | 0.6268 | 0.6849 |

for the Parkinson speech dataset. Daubechies' extremal phase (DEP) wavelet and Daubechies' least-asymmetric (DLA) wavelet families are used in this table for selected value of $\omega$. As can be seen in the table, the AMSE values obtained in the RSTWS estimator $\delta^{\text{soft}}_{\pi,\lambda^{\text{sure}}}(X)$ with $\omega = 0.2$ are lower than that of the other methods for standardized data. Also, as the filter number increases, the accuracy of estimation increases (AMSE decreases), and in this dataset, the accuracy of the DLA wavelet is better than the DEP wavelet. Note that if the data were not standardized, the AMSE values for the three estimators, the soft CV threshold, the soft Bayes threshold, and $\delta^{\text{soft}}_{\lambda^*}(X)$ would be much higher than the two RSTWS estimators $\delta^{\text{soft}}_{\pi,\lambda^{\text{sure}}}(X)$ and $\delta^{\text{soft}}_{\lambda^{\text{sure}}}(X, U)$.

## 10.4 Conclusion

In this chapter, we study the estimation of a mean vector restricted to a convex cone with five types of restrictions: positive orthant, simplicial, polyhedral, general convex cones, and nested cones. We find estimators that improve on the natural estimator in the general case of a $\mathcal{SS}$ distribution, and we introduce the restricted generalized Bayes estimation. Also, we introduce two types of the restricted soft SURE threshold wavelet shrinkage estimator. To denoising, we present a simulation study to test the validity of the proposed estimator. In this regard, we investigate the behavior of the RSTWS estimator, and the Parkinson speech dataset is given to test the efficiency of this estimator in denoising. After denoising the real datasets, by computing the average mean square error, we find that the proposed estimator dominates other competing estimators. The results demonstrate that the target estimator is appropriate and that increasing the dimension $p$ decreases the accuracy of the new RSTWS estimator. Also, the average of risk values obtained in the RSTWS estimator $\delta^{\text{soft}}_{\pi,\lambda}(X)$ is lower than that of the other estimators for $\omega = 0.8$.

**Acknowledgments** The authors would like to thank the editors and reviewers for their valuable comments, which greatly improved the readability of this paper. Also, we would like to thank the research committee of Persian Gulf University.

## References

Batvandi, Z., Afshari, M., & Karamikabir, H. (2025). Two new Bayesian-wavelet thresholds estimations of elliptical distribution parameters under non-linear exponential balanced loss. *Communications in Statistics-Simulation and Computation*, *54*(1), 23–43.

Batvandi, Z., Afshari, M., & Karamikabir, H. (2024). Bayesian estimation for mean vector of multivariate normal distribution on the linear and nonlinear exponential balanced loss based on wavelet decomposition. *International Journal of Wavelets, Multiresolution and Information Processing*, *22*(6), 2450031.

Batvandi, Z., Afshari, M., & Karamikabir, H. (2025). Bayesian shrinkage wavelet estimation of mean matrix of the matrix variate normal distribution with application in de-noising. *Computational and Applied Mathematics*, *44*(1), 1–26.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, *57*(1), 289–300.

Bock, M. E. (1982). Employing vague inequality information in the estimation of normal mean vectors. (Estimators that shrink to closed convex polyhedra.) Technical Report 81-42. Department of Statistics, Purdue University.

Cai, T. T., & Silverman, B. W. (2001). Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhya: The Indian Journal of Statistics, Series B*, *127*, 127–148.

Cao, M. X., & He, D. J. (2017). Admissibility of linear estimators of the common mean parameter in general linear models under a balanced loss function. *Journal of Multivariate Analysis*, *153*, 246–254.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224.

Fang, K. T., Kotz, S., & Ng, K. W. (1990). *Symmetric multivariate and related distributions.* Chapman and Hall.

Fourdrinier, D., Ouassou, I., & Strawderman, W. E. (2003). Estimation of a parameter vector when some components are restricted. *Journal of Multivariate Analysis*, *86*(1), 14–27.

Fourdrinier, D., Kortbi, O., & Strawderman, W. E. (2008). Bayes minimax estimators of the mean of a scale mixture of multivariate normal distributions. *Journal of Multivariate Analysis*, *99*(1), 74–93.

Fourdrinier, D., Kortbi, O., & Strawderman, W. E. (2014). Generalized Bayes minimax estimators of location vectors for spherically symmetric distributions with residual vector. *Metrika*, *77*(2), 285–296.

Karamikabir, H. (2024). Wavelet shrinkage estimation for mean matrix of matrix-variate elliptically contoured distributions. *Journal of Statistical Theory and Practice*, *18*(2), 23.

Karamikabir, H., & Afshari, M. (2020). Generalized Bayesian shrinkage and wavelet estimation of location parameter for spherical distribution under balance-type loss: Minimaxity and admissibility. *Journal of Multivariate Analysis*, *177*, 104583.

Karamikabir, H., & Afshari, M. (2021). New wavelet SURE thresholds of elliptical distributions under the balance loss. *Statistica Sinica*, *31*(4), 1829–1852.

Karamikabir, H., & Afshari, M. (2022). Wavelet shrinkage generalized bayes estimation for multivariate normal distribution mean vectors with unknown covariance matrix under balanced-LINEX loss. *Revista Colombiana de Estadística*, *45*(1), 107–123.

Karamikabir, H., Afshari, M., & Arashi, M. (2018). Shrinkage estimation of non-negative mean vector with unknown covariance under balance loss. *Journal of Inequalities and Applications*, *2018*(1), 1–11.

Karamikabir, H., Karamikabir, N., Khajeian, M. A., & Afshari, M. (2023). Bayesian wavelet Stein's unbiased risk estimation of multivariate normal distribution under reflected normal loss: Bayesian wavelet Stein's unbiased risk estimation. *Methodology and Computing in Applied Probability*, *25*(1), 23.

Karamikabir, H., Afshari, M., Karamikabir, N., & Kiaei, S. F. (2024). Generalized Bayes SURE for multivariate normal distribution under balanced-quadratic loss. *Journal of Mathematical Sciences*, *280*, 262–274.

Karamikabir, H., Sanati, A., & Hamedani, G. G. (2025). Low and high dimensional wavelet thresholds for matrix-variate normal distribution. *Communications in Statistics - Simulation and Computation*, *54*(7), 2742–2761.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693.

Marchand, E., & Strawderman, W. E. (2020). On shrinkage estimation for balanced loss functions. *Journal of Multivariate Analysis*, *175*, 104558.

Maruyama, Y. (2003). A robust generalized Bayes estimator improving on the James-Stein estimator for spherically symmetric distributions. *Statistics & Decisions*, *21*(1), 69–78.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(2), 463–479.

Ouassou, I., & Strawderman, W. E. (2002). Estimation of a parameter vector restricted to a cone. *Statistics & Probability Letters*, *56*(2), 121–129.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., & Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, *17*(4), 828–834.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, *9*(6), 1135–1151.

Zellner, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions. In *Bayesian and non-Bayesian estimation using balanced loss functions* (pp. 377–390). Springer New York.

Zinodiny, S., Rezaei, S., & Nadarajah, S. (2017). Bayes minimax estimation of the mean matrix of matrix-variate normal distribution under balanced loss function. *Statistics & Probability Letters*, *125*, 110–120.

# Chapter 11
# On Minimaxity of Shrinkage Estimators Under Concave Loss




**Stavros Zinonos and William E. Strawderman**

## 11.1 Introduction

We study minimax shrinkage estimators of a location vector for certain classes of spherically symmetric distributions under concave loss. In the most basic case, we observe

$$X \sim N_p(\theta, \sigma^2 I_p) \tag{11.1}$$

with $\sigma^2$ known and consider estimation of the location vector $\theta \in \Re^p$ under a loss function of the form

$$l(d, \theta) = c\|d - \theta\|^2 + \int_0^\infty \left(1 - \exp\{-b\|d - \theta\|^2\}\right) \mathrm{d}g(b) \tag{11.2}$$

where $c \geq 0$ and $g(\cdot)$ is a finite positive measure. The class of shrinkage estimators considered is the class of Baranchik-type estimators:

$$\delta_{a,r}(X) = \left(1 - \frac{a\sigma^2 r\left(\frac{\|X\|^2}{\sigma^2}\right)}{\|X\|^2}\right) X. \tag{11.3}$$

S. Zinonos (✉)
Global Biometrics and Data Managment, Pfizer Inc., New York, NY, USA
e-mail: stavros.zinonos@pfizer.com

W. E. Strawderman
Department of Statistics, Rutgers University, Newark, NJ, USA
e-mail: staw@stat.rutgers.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_11

Interest in estimation under concave loss has become more prominent in recent years, partly due to an increased interest in predictive density estimation. In particular, for plug-in density estimators, the loss under several common standard predictive density estimation losses becomes equivalent to a concave loss in a point estimation problem of the type we investigate. This is true for Kullback-Leibler, integrated $L_2$, integrated $L_1$, and also $\psi$-divergence losses. See George et al. (2006), Maruyama and Strawderman (2012), Kubokawa et al. (2015, 2017) for some examples of this equivalence and additional recent developments.

The basic result of Sect. 11.2 shows that an estimator given by Eq. (11.3) is minimax for $p \geq 3$, provided

$$\begin{aligned} &\text{(i) } r'(u) \geq 0, \\ &\text{(ii) } 0 \leq r(u) \leq 1, \\ &\text{(iii) } 0 \leq a \leq 2(p-2) \end{aligned} \tag{11.4}$$

with strict domination of the usual minimax estimator $\delta_0(X) = X$ if either (iii) is strict or either (i) or (ii) is strict on a set of positive measure. Interestingly, this is the same set of condition for minimaxity (domination) under quadratic loss:

$$l(d, \theta) = \|d - \theta\|^2. \tag{11.5}$$

In almost all results in the literature, (eg., Brandwein & Strawderman, 1991, 1980; Kubokawa et al., 2015), the maximum value of the shrinkage constant, $a$, is strictly smaller for classes of concave losses that are functions of quadratic loss. An exception is the work of Ghosh et al. (2008), which established for loss

$$l(d, \theta) = 1 - \exp\left\{ -\frac{b}{2\sigma^2}\|d - \theta\|^2 \right\}, \tag{11.6}$$

$b > 0$, $2(p-2)$ is the maximum value for the shrinkage constant, $a$, in Eq. (11.3) for minimaxity.

Extensions to the unknown scale case will be given. In particular,

$$\begin{pmatrix} X \\ U \end{pmatrix} \sim N_{p+k}\left( \begin{pmatrix} \theta \\ 0 \end{pmatrix}, \sigma^2 I_{p+k} \right) \tag{11.7}$$

where $\dim(X) = \dim(\theta) = p \geq 3$ and $\dim(U) = \dim(0) = k \geq 1$. Here, $(\|U\|^2, X)$ is a complete sufficient statistic for $\theta$ and $\sigma^2$. The loss studied is defined in Eq. (11.2), and the class of estimators studied is

$$\delta(X, U) = \left( 1 - \frac{a\,\|U\|^2\, r\left( \frac{\|X\|^2}{\|U\|^2} \right)}{(k+2)\|X\|^2} \right) X. \tag{11.8}$$

Section 11.3 establishes the estimator defined in (11.8) satisfying Eq. (11.4) will dominate the usual estimator $\delta_0(X) = X$ with strict domination if (i) is strict or either (ii) or (iii) is strict on a set of positive measure. Replacing $\|d - \theta\|^2$ with $\|d - \theta\|^2 / \sigma^2$ in Eq. (11.2) will result in the estimator $X$ being minimax and estimator defined in Eq. (11.8) dominating the estimator $X$ under conditions (i), (ii), and (iii). This section also studies the general class of scale mixtures of normal distributions when a residual vector is present. In particular,

$$\begin{pmatrix} X \\ U \end{pmatrix} |\sigma^2 \sim N_{p+k}\left(\begin{pmatrix} \theta \\ 0 \end{pmatrix}, \sigma^2 I_{p+k}\right) \tag{11.9}$$

with $\sigma^2 \sim F(\sigma^2)$ where $F(\cdot)$ is a distribution function on $(0, \infty)$ and $\mathrm{E}[\sigma^2]$ and $\mathrm{E}[1 / \sigma^2]$ are both finite. We show, as it does in the normal case, that estimator given by Eq. (11.8) improves upon the usual estimator $X$ under loss defined in Eq. (11.2), once Eq. (11.4) is satisfied. This result also gives improvement over $X$ under the same set of conditions as in the case of quadratic loss $\|d - \theta\|^2$.

A natural question to ask is if the results of Sect. 11.2 (the known scale case) extend to the general class of scale mixtures of normal distributions. In Sect. 11.4, we present some numerical simulations indicating this is not the case. Section 11.5 has some concluding remarks.

## 11.2 Normal Theory $\sigma^2$ Known

In this subsection, we study the minimaxity of estimators of the form Eq. (11.3) for the class of loss functions given by Eq. (11.2) when $X \sim N_p(\theta, \sigma^2 I_p)$ with $\sigma^2$ known. The results extend those of Ghosh et al. (2008), which give conditions for estimators given by Eq. (11.3) to strictly dominate the minimax estimator $\delta_0(X) = X$ (for proof of minimaxity of the estimator $X$, see Ghosh et al., 2008) for the class of losses given by

$$l_\beta(d, \theta) = 1 - \exp\left\{\frac{\beta}{\sigma^2}\|d - \theta\|^2\right\} \tag{11.10}$$

with $\beta > 0$. The following theorem is a direct consequence of combining Theorems 5.1 and 5.2 in Ghosh et al. (2008).

**Theorem 11.1 (Ghosh et al., 2008)** *Let $X \sim N_p(\theta, \sigma^2 I_p)$ with $\sigma^2$ known, and let $\delta_{a,r}(X)$ be an estimator for $\theta$ of the form given by Eq. (11.3), such that (11.4) holds. Then, the estimator $\delta_{a,r}$ will dominate the minimax estimator, $X$, under the loss defined in Eq. (11.10) for $\beta > 0$ and $p \geq 3$.*

**Remark 11.1** Setting

$$\frac{a\,\sigma^2\, r(u)}{u} = \frac{\tau\left(u / \sigma^2\right)}{u / \sigma^2}$$

results in the conditions

(i) $0 < \tau < 2(p-2)$

(ii) $\tau(u)$ is a differentiable nondecreasing function of $u$

for domination of the estimator $\delta_0(X) = X$.

Lemma 11.1 and Theorem 11.2 are the main results of this section. They extend the above result to losses of the form Eq. (11.2).

**Lemma 11.1** *If the estimator $\hat{\theta}(X)$ dominates $X$ for the loss $\|d-\theta\|^2$, as well as for losses $(1-\exp\{-b\|d-\theta\|^2\})$, $b>0$; then, $\hat{\theta}(X)$ dominates $X$ under loss given by Eq. (11.2) for all $c>0$ and measure $g(\cdot)$.*

***Proof*** Since $\mathrm{E}_\theta\left[\int_0^\infty \exp\{-b\|\hat{\theta}(X)-\theta\|^2\}\,dg(b)\right] \le \int_0^\infty dg(b) < \infty$, the interchange of the integral and expectation for

$$\mathrm{E}_\theta\left[\int_0^\infty \{\exp\{-b\|\hat{\theta}(X)-\theta\|^2\} - \exp\{-b\|X-\theta\|^2\}\}\,\mathrm{d}g(b)\right] = \int_0^\infty \mathrm{E}_\theta\left[\{\exp\{-b\|\hat{\theta}(X)-\theta\|^2\} - \exp\{-b\|X-\theta\|^2\}\}\right]\mathrm{d}g(b) \quad (11.11)$$

is valid, and the results follow from the assumption $\mathrm{E}_\theta\left[\|\hat{\theta}(X)-\theta\|^2\right] \le \mathrm{E}_\theta[\|X-\theta\|^2]$ and the positive expectation. □

**Theorem 11.2** *Let $X \sim N_p(\theta, \sigma^2 I_p)$ for $p \ge 3$. Let $g(\cdot)$ be any finite positive measure. Let $\delta_{a,r}(X)$ be an estimator of $\theta$ as given in Eq. (11.3). Then, for $c>0$, $\delta_{a,r}$ will be a minimax estimator of $\theta$ for each loss of the form (11.2) with strict domination over $\delta_0(X)=X$ if either (i) or (iii) is strict in Eq. (11.4).*

***Proof*** If the constant $a$ satisfies (i) and the function $r(\cdot)$ satisfies (ii) and (iii), then $\delta_{a,r}$ is a minimax estimator under quadratic loss, $\|d-\theta\|^2$, so that $R(\delta_{a,r},\theta) \le R(X,\theta)$ under quadratic loss with strict inequality if either (i) or (iii) is strict. Theorem 11.1 implies for any $b>0$,

$$\mathrm{E}\left[1-\exp\{-b\|\delta_{a,r}(X)-\theta\|^2\}\right] \le \mathrm{E}\left[1-\exp\{-b\|X-\theta\|^2\}\right] \quad (11.12)$$

where Eq. (11.12) is strict if either (i) or (iii) is strict. Therefore, the assumptions of Lemma 11.1 are satisfied so that $\delta_{a,r}$ is a minimax estimator of $\theta$ for the class of losses given by Eq. (11.2) as $X$ is also a minimax estimator. □

The following example illustrates Theorem 11.2 when $g(\cdot)$ is a $\Gamma(k_1,k_2)$ density.

**Example 11.1 (Folded t-loss)** Let $X \sim N_p(\theta,\sigma^2 I_p)$ with $\sigma^2$ known and $\delta$ be an estimator of $\theta$ of the form

$$\delta(X) = \left(1 - \frac{a\sigma^2}{\|X\|^2}\right) X. \tag{11.13}$$

Then, $\delta(X)$ will be a minimax estimator for the loss:

$$l(d,\theta) = c\|d-\theta\|^2 + 1 - (k_2\|d-\theta\|^2 + 1)^{-k_1}, \tag{11.14}$$

where $c \geq 0$ and $k_1, k_2 > 0$, once $0 \leq a \leq 2(p-2)$.

This follows from Theorem 11.2 by setting $g(b)$ to be density function of a $\Gamma(k_1, k_2)$ distribution with shape parameter $k_1$ and scale parameter $k_2$. Let $t = \|d-\theta\|^2$. The loss given by Eq. (11.2) is equivalent to

$$\begin{aligned} l(t) &= ct + 1 - \frac{1}{\Gamma(k_1)k_2^{k_1}} \int_0^\infty \exp\{-bt\} \exp\left\{-\frac{b}{k_2}\right\} b^{k_1-1}\, \mathrm{d}b \\ &= ct + 1 - \frac{1}{\Gamma(k_1)k_2^{k_1}} \int_0^\infty \exp\left\{-b(t+\frac{1}{k_2})\right\} b^{k_1-1}\, \mathrm{d}b \\ &= ct + 1 - \frac{1}{k_2^{k_1}(t+\frac{1}{k_2})^{k_1}} \\ &= ct + 1 - (k_2 t + 1)^{-k_1}. \end{aligned}$$

By Theorem 11.2 estimator defined by Eq. 11.13 is minimax once $0 \leq a \leq 2(p-2)$, as the function $r(\|X\|^2) = 1$ satisfies Eq. (11.4).

## 11.3 Extensions

Section 11.3.1 gives a preliminary result from Ghosh and Mergel (2009) that will be used in extending the results of Sect. 11.2 to the case that $(X^\top, U^\top)^\top$ is distributed as Eq. (11.7) with $\sigma^2$ unknown, $\dim(X) = \dim(\theta) = p \geq 3$, and $\dim(U) = \dim(0) = k \geq 1$. The estimators considered are of the form

$$\delta_{a,r}(X, S) = \left(1 - \frac{a\, S\, r\left(\frac{\|X\|^2}{S}\right)}{(k+2)\|X\|^2}\right) X \tag{11.15}$$

where $S = \|U\|^2$ in Eq. (11.15). The estimator $\delta_{a,r}(X, S)$ will dominate the estimator $\delta_0(X) = X$ for the class of losses given by Eq. (11.2) once conditions in Eq. (11.4) are satisfied. Note that conditions (i)–(iii) imply domination of Eq. (11.15) over the estimator $\delta_0(X) = X$ under quadratic loss.

Section 11.3.2 deals with the case

$$\begin{pmatrix} X \\ U \end{pmatrix} | \sigma^2 \sim N_{p+k} \left( \begin{pmatrix} \theta \\ 0 \end{pmatrix}, \sigma^2 I_{p+k} \right) \tag{11.16}$$

with $\sigma^2 \sim F(\sigma^2)$, where $F$ is a distribution function on $(0, \infty)$ such that $E[\sigma^2] < \infty$ and $\mathrm{E}\left[1/\sigma^2\right] < \infty$. For this case, the estimator $\delta_{a,r}$ given by Eq. (11.15) will dominate $X$ under loss given by Eq. (11.2). Under quadratic error loss, the results of Fourdrinier et al. (2006), which apply to the general class of spherically symmetric distributions with residuals, imply that estimator $\delta_{a,r}$ given by Eq. (11.15) will dominate the estimator $\delta_0(X) = X$ if (i)–(iii) are satisfied.

### *11.3.1 Normal with Scale Unknown*

The result of Ghosh and Mergel (2009) is presented below. As in the case in Sect. 11.2, Lemma 11.1 will be used to extend the results to the class of losses in Eq. (11.2).

**Lemma 11.2 (Ghosh & Mergel, 2009)** *Let*

$$\begin{pmatrix} X \\ U \end{pmatrix} \sim N_{p+k} \left( \begin{pmatrix} \theta \\ 0 \end{pmatrix}, \sigma^2 I_{p+k} \right)$$

*with $\sigma^2$ unknown, $dim(X) = dim(\theta) = p \geq 3$, and $dim(U) = dim(0) = k \geq 1$. Let $\delta_{a,r}(X, S)$ be the estimator of $\theta$ of the form Eq. (11.15). If conditions in Eq. (11.4) are satisfied, then*

$$R(\delta_\tau, \theta) \leq R(X, \theta)$$

*for the loss in Eq. (11.10).*

Theorem 11.3 extends the class of losses for which $\delta_{a,r}$ given in Lemma 11.2 will dominate the estimator $\delta_0(X) = X$, to the class of losses given in Eq. (11.2).

**Theorem 11.3** *Let $(X^\top, U^\top)^\top$ be distributed as in Lemma 11.2. Then, for losses as in Eq. (11.2), the estimator given by Eq. (11.15) will dominate the estimator, $X$, once conditions in Eq. (11.4) are satisfied with strict domination if inequality (i) or (ii) is strict.*

***Proof*** If conditions (i)–(iii) mentioned in Eq. (11.4) are satisfied, then Lemma 11.2 implies

$$\mathrm{E}\left[ \exp\left\{ -b\|X - \theta\|^2 \right\} \right] \geq \mathrm{E}\left[ \exp\left\{ -b\|\delta_{a,r} - \theta\|^2 \right\} \right] \tag{11.17}$$

for $b > 0$ with strict inequality if either (i) or (ii) is satisfied. The same set of conditions, under quadratic loss $\|d - \theta\|^2$, imply

$$R(\delta_{a,r}, \theta) \leq R(X, \theta)\,,$$

with strict inequality if either (i) or (ii) is strict. Therefore, Lemma 11.1 implies the result. □

### *11.3.2 Scale Mixture of Normals with Unknown Scale*

Theorem 11.4 gives conditions for which estimators of Eq. (11.15) dominate the estimator $\delta_0(X) = X$ under losses given by Eq. (11.2) when

$$\begin{pmatrix} X \\ U \end{pmatrix} |\sigma^2 \sim N_{p+k}\left(\begin{pmatrix} \theta \\ 0 \end{pmatrix}, \sigma^2 I_{p+k}\right)\,, \tag{11.18}$$

with $\sigma^2 \sim F(\sigma^2)$, where $F$ is a distribution function on $(0, \infty)$ having density $v(\sigma^2)$.

**Theorem 11.4** *Let* $(X^\top, U^\top)^\top|\sigma^2$ *be distributed as* (11.18) *where* $dim(X) = dim(\theta) = p \geq 3$ *and* $dim(U) = dim(0) = k \geq 1$ *and* $g$ *be a positive finite measure. For the loss defined in Eq.* (11.2)*, the risks for the class of estimators given in Eq.* (11.15) *satisfy* $R(X, \theta) \geq R(\delta_{a,r}, \theta)$ *uniformly in* $\theta$ *once conditions given by Eq.* (11.4) *are satisfied.*

***Proof*** Conditioning on $\sigma^2$ by Theorem 11.3, the difference in risk between the estimator $\delta_{a,r}$ and $X$ satisfies

$$\begin{aligned} &\mathrm{E}\left[\left\{c\|\delta_\tau - \theta\|^2 + \int_0^\infty (1 - \exp\{-b\|\delta_\tau - \theta\|^2\})\,\mathrm{d}g(b)\right\}\Big|\sigma^2\right] \\ &\quad - \mathrm{E}\left[\left\{c\|X - \theta\|^2 + \int_0^\infty (1 - \exp\{-b\|X - \theta\|^2\})\,\mathrm{d}g(b)\right\}\Big|\sigma^2\right] \leq 0 \end{aligned} \tag{11.19}$$

for all $\sigma^2 > 0$. Taking the expectation of the above establishes the result. □

## 11.4 Simulation Results

In this section, we investigate via simulations the behavior of the risk of the James-Stein estimator:

$$\delta_{\mathrm{JS},c}(X) = \left(1 - \frac{c}{\|X\|^2}\right) X \tag{11.20}$$

when the underlying distribution $X$ is a scale mixture of normal distributions without a residual, and the loss is given by

$$l_b(d, \theta) = 1 - \exp\left\{ - b\|d - \theta\|^2 \right\}. \tag{11.21}$$

From Sects. 11.2 and 11.3, one might suspect that for the class of estimators given by Eq. (11.20), which is a subclass of the estimators defined in Eq. (11.3), that domination of $X$ under quadratic loss implies dominance under the loss given by (11.10). Our simulation indicates this is not true in general.

Specifically, we assume

$$X|V \sim N_p(\theta, V I_p) \tag{11.22}$$

with $p = 6$, where $V$ has an inverse gamma distribution with the density

$$h(v) = \frac{1}{\Gamma(a)s^a}\, v^{-a-1} \exp\left\{ - \frac{1}{sv} \right\}; \quad a, s > 0 \tag{11.23}$$

with loss given by Eq. (11.21) for $b > 0$. We plot the risk of the estimator $\delta_{\mathrm{JS}}$ given by Eq. (11.20), when $c$ is chosen to be near the upper bound established by Strawderman (1974) for minimaxity under square error to see if it is minimax using the loss given by Eq. (11.21). In particular, for scale mixtures of normal distributions, Strawderman (1974) established that estimators given by Eq. (11.20) will be minimax for square error loss once

$$0 \le c \le \frac{2(p-2)}{\mathrm{E}\left[\frac{1}{V}\right]}, \tag{11.24}$$

and thus when the mixture distribution $V$ has an inverse gamma distribution given by Eq. (11.23), the estimator $\delta_{\mathrm{JS}}$ will be minimax under square error loss once

$$0 \le c \le \frac{2(p-2)}{\mathrm{E}\left[\frac{1}{V}\right]} = \frac{2(p-2)}{as} \tag{11.25}$$

as $(1 / V) \sim \Gamma(a, s)$.

Figure 11.1 plots the risk of various James-Stein estimators when the mixture has fixed shape parameter $a = 3$ with varying rate parameters $s$. Each James-Stein estimator plotted has a shrinkage constant, $c$, that is minimax under quadratic loss. More specifically, the constant $c$ was chosen to be at or near the value of $2(p-2) / ((a\, s))$. For $s = 10$ and $s = 4$, the James-Stein estimator appears to

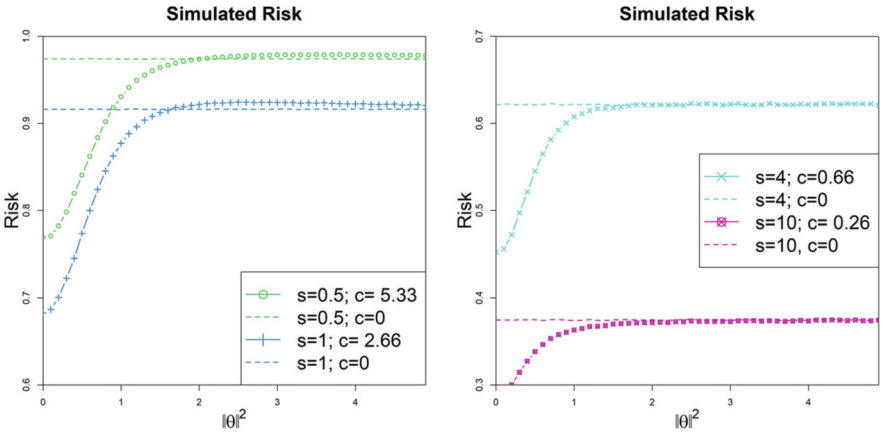



**Fig. 11.1** Risk of James-Stein estimator defined in Eq. (11.20) for various mixture parameter of the inverse gamma distribution in Eq. (11.23) as $s$, $a = 3$ and the loss function given by Eq. (11.21)

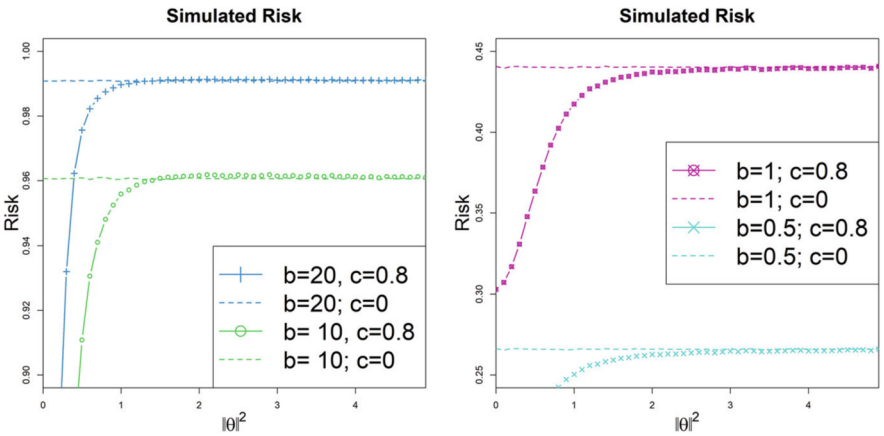



**Fig. 11.2** Simulated risk of the James-Stein estimator defined in Eq. (11.20) for varying $b$, mixture parameter of the inverse gamma distribution in Eq. (11.23) as $a = 10$ and $s = 1$ as well as the loss function given by Eq. (11.21)

dominate $\delta_0(X) = X$. For the distributions with heavier tails ($s = 0.5$ and $s = 1$), the risk of the estimator $X$ seems to be less than the risk of $\delta_{\text{JS}}$ for sufficiently large $\|\theta\|^2$.

Figure 11.2 investigates the behavior of the parameter $b$ on the risk of the James-Stein estimator with the shrinkage factor of $c = 0.8$ when the mixture parameters are $a = 10$ and $s = 1$. When $b = 0.5$ and $b = 1$, the risk of the James-Stein estimator approaches the risk of $X$ from below, while for $b = 10$ and $b = 20$, the James-Stein estimator approaches the risk of $X$ from above.

Based on the simulation in this section, it appears reasonable to conjecture that for a given mixture with shape parameter $a = a_0$ and rate parameter, $s = s_0$, there exists a $b^*$ such that for $b > b^*$ the James- Stein estimator with the shrinkage constant $c = 2(p-2)/(a_0\, s_0)$ has risk that starts below the risk of $X$ at the origin and asymptotically approaches the risk of $X$ from above. For $b < b^*$, the James-Stein estimator uniformly dominates the risk of $X$. Furthermore, given $b_0 < b^*$ and shape parameter $a = a_0$, the risk of the James-Stein estimator will be uniformly less than the risk of $X$ for all values of the scale parameter, $0 < s < s_0$ when the shrinkage constant $c = 2(p-2)/a_0 s$.

## 11.5 Conclusion

For the large class of bounded concave loss functions that are functions of quadratic loss, $\|d-\theta\|^2$, given in Sect. 11.2, we establish conditions for which classes of the Baranchik-type estimators are minimax and uniformly dominate the risk of $X$ for the Gaussian case. The conditions for minimaxity on the shrinkage factor, $a$, are identical to those for quadratic loss. Extensions to when the scale is unknown and a residual vector is present, and to the case the underlying density is a scale mixture with residual vector present, are given as well for the same class of losses. Further investigation is needed in the case that the underlying density is a scale mixture without a residual vector present in order to determine conditions for which the Baranchik-type estimators will be minimax for the classes of losses given in Sect. 11.2.

**Competing Interests** The authors have no conflicts of interest to declare that are relevant to the content of this chapter.

## References

Brandwein, A. C., & Strawderman, W. E. (1991). Generalizations of James-Stein estimators under spherically symmetry. *Annals of Statistics*, *19*(3), 1639–1650.

Brandwein, A. C., & Strawderman, W. E. (1980). Minimax Estimators of location parameters for spherically symmetric distribution with concave loss. *Annals of Statistics*, *8*, 279–284.

Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2006). Estimation of a location parameter with restrictions or “vague information" for spherically symmetric distributions. *Annals of the Institute of Statistical Mathematics*, *58*, 73–92.

George, E. I., Liang, F., & Xu, X. (2006). Improved minimax prediction under Kullback-Leibler loss. *The Annals of Statistics, 34*, 78–91.

Ghosh, M., Mergel, V., & Datta, G. S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *Journal of Multivariate Analysis*, *99*, 1941–1961.

Ghosh, M., & Mergel, V. (2009). On the Stein phonomenon under divergence loss and an unkown variance-covariance matrix. *Journal of Multivariate Analysis*, *100*, 2331–2336.

Kubokawa, T., Marchand, E., & Strawderman, W. E. (2015). On improved shrinkage estimators for concave loss. *Statistics and Probability Letters*, *96*, 241–246.

Kubokawa, T., Marchand, E., & Strawderman, W. E. (2015). On Predictive Density Estimation for Location Families under Integrated Squared Error Loss. *Journal of Multivariate Analysis*, *142*, 57–74.

Kubokawa, T., Marchand, E., & Strawderman, W. E. (2017). On predictive density estimation for location families under integrated absolute error loss. *Bernoulli*, *23*(4B), 3197–3212.

Maruyama, Y., & Strawderman, W. E. (2012). Bayesian predictive densities for linear models estimation under divergence loss: some results and open problems. In *Contemporary Developlents in Bayesian Analysis and Statistical Decision Theory; A festschrift for William E. Strawderman* (vol. 8, pp. 42–56). IMS Collections.

Strawderman, W. E. (1974) Minimax estimation of location parameters for certain spherically symmetric distributions. *Journal of Multivariate Analysis*, *4*, 255–264.

# Part II
# Penalized Estimation and Variable Selection

# Chapter 12
# Improved LASSO Estimator in Semiparametric Linear Measurement Error Models

**Hadi Emami and Kourosh Dadkhah**

## 12.1 Introduction

In standard linear regression models, miss-specifying the structures of models can cause severe modeling biases. As an alternative, nonparametric regression models are useful in exploring hidden structures and reducing modeling biases. However, when the large covariates are involved in the model, the nonparametric method is hindered by the curse of dimensionality. To overcome these shortcomings, semiparametric regression models (SPRMs) have been introduced, which are more flexible than standard and nonparametric regression models since it has a parametric and a nonparametric component. Generally, a SPRM can be written as

$$y_i = \mathbf{X}_i^\top \beta + g(t_i) + \epsilon_i \tag{12.1}$$

where $y_i$ is the response, $X_i$ is a $p$ vector-valued covariate, $t_i$ is a scalar covariate, the function $g(\cdot)$ is unknown, and the model errors $\epsilon_i$ are independent with conditional mean zero given the covariates. This model is also called a partially linear model or a partial spline model. Surveys regarding the estimation and application of model defined in Eq. (12.1) can be found in the monograph of Speckman (1988), Härdle et al. (2000), Ruppert et al. (2003), Horowitz (2009), and Harezlak et al. (2019), among others. When the measurement errors creep into the data, the usual statistical tools tend to lose their validity, and the main subject is how to obtain the consistent estimators of regression coefficients. In a model given by Eq. (12.1) when

H. Emami (✉) · K. Dadkhah
Department of Statistics, Faculty of Basic Sciences, University of Kurdistan, Sanandaj, Iran
e-mail: h.emami@uok.ac.ir; k.dadkhah@uok.ac.ir

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_12

the covariates $X_i$ are measured with error, instead of observing $X_i$, we observe

$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i \tag{12.2}$$

where the measurement errors $\mathbf{U}_i$ are independent and identically distributed, independent of $(y_i, X_i, t_i)$, with mean zero and covariance matrix $\Sigma_{uu}$. By some mild assumption and using the kernel weight with ordinary bandwidths of order $h \approx n^{\frac{-1}{5}}$, Liang et al. (1999) shows that

$$\sqrt{n}(\hat{\beta}_{0n} - \beta) \xrightarrow{\mathcal{D}} N_p(0, \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}), \tag{12.3}$$

where $\mathbf{C} = \mathrm{E}[\tilde{X}_i\tilde{X}_i] = \mathrm{var}(\tilde{X}_i)$, $\tilde{X}_i = X_i - \mathrm{E}[X_i|t]$, $\mathbf{\Lambda} = [\mathrm{E}[\epsilon - U^\top\beta]\tilde{\mathbf{X}}]^{\otimes 2} + \mathrm{E}[\mathbf{U}^\top\mathbf{U} - \mathbf{\Sigma}_{uu}\beta]^{\otimes 2} + \mathrm{E}[\mathbf{U}^\top\mathbf{U}\epsilon^2]$ and denoting $\mathbf{A}^{\otimes 2}$ by $\mathbf{A}^\top\mathbf{A}$ . One of the most powerful regression technique is the least absolute shrinkage and selection operator (LASSO) approach, which aims to induce sparseness in the linear models (Tibshirani, 1996). This property is obtained by minimizing an $l_1$-regularized least squares cost function, namely, a cost function consisting of a smooth quadratic plus a non-smooth $l_1$-norm term. The LASSO has been extensively used in regression models, and its statistical scope and limitations are well understood. The LASSO approach in measurement error has been studied by, e.g., Liang et al. (1999), Li et al. (2021), Sørensen et al. (2015), and Ali and Tibshirani (2019). It is well-known that the incorporation of prior information available in the form of restrictions provides better estimators than the ordinary Least Squares (OLS) estimator (see Toutenburg, 1982; Rao & Toutenburg, 2008). When the prior information is suspected, by combining the restricted subspace of parameters and OLS, we can have more efficient estimators, namely, the preliminary test's and Stein-type shrinkage estimators. The approach of preliminary test estimation has been studied by many authors, for example, Judge and Bock (1978), Saleh (2006), Ahmed et al. (2007), Kim and Saleh (2005), Kibria and Saleh (2012), Norouzirad and Arashi (2018), Yüzbaşı et al. (2019), and Roozbeh et al. (2020).

In this chapter, we introduce innovative preliminary test and Stein-type shrinkage estimators for the semiparametric linear measurement error (SLME) model within the context of LASSO. Unlike OLS, our approach utilizes LASSO as a foundational framework, allowing us to derive shrinkage estimators under specific model constraints. We derive the asymptotic properties of these new estimators, which enhance our understanding of their behavior compared to classical estimators. Notably, this is the first application of the shrinkage LASSO (or adaptive LASSO) preliminary test and Stein-type estimators to models affected by measurement errors in covariates, addressing a critical gap in the literature. While LASSO is known for its near mini-max optimality and consistency, it falls short of achieving the oracle property, which limits its performance against methods such as the smoothly clipped absolute deviation (SCAD) estimator. To address this limitation, we employ the Adaptive LASSO (ALASSO), originally developed by Zhou (2006) and Zhang and Lu (2007). ALASSO utilizes adaptive weights for penalization and successfully

attains the oracle property, maintaining the optimal convergence and near mini-max optimality characteristic of LASSO. In this study, we further extend the ALASSO framework by applying a modified version of the least angle regression (LARS) algorithm to define the ALASSO-based preliminary test and Stein-type shrinkage estimators specifically for SLME contexts.

The remainder of the paper is organized as follows. In Sect. 12.2, we discuss the asymptotic distributions of ordinary and restricted versions of LASSO and ALASSO and introduce a test statistic along with its asymptotic properties. We also propose the restricted LASSO, preliminary test, and Stein-type estimators and derive their asymptotic distributions. In Sect. 12.3, we analyze the asymptotic distributional bias, quadratic bias, mean squared error, and quadratic risks of the proposed LASSO estimators. In Sect. 12.4, the modified LARS algorithm is described in detail. Two simulation examples and a real data analysis are presented in Sects. 12.5 and 12.6, respectively. Finally, the conclusion is provided in Sect. 12.7.

## 12.2 Background and Definition

To get insights into the LASSO penalized least squares procedure, we consider the case in which $\mathbf{\Sigma}_{uu}$ is known, as in Li et al. (2021) among others. Define a penalized least squares function as

$$\frac{1}{2}\sum_{i=1}^{n}\{y_i - \mathbf{W}_i^{\top}\beta - g(t_i)\}^2 - \frac{n}{2}\beta^{\top}\mathbf{\Sigma}_{uu}\beta. \tag{12.4}$$

The second term in Eq. (12.4) is included to correct the bias in the squared loss function due to measurement error. Since $\mathrm{E}[\mathbf{U}|t] = 0$ and $g(t) = \mathrm{E}[y|t] - \mathrm{E}[\mathbf{W}|t]^{\top}\beta$, then the least squares function defined in Eq. (12.4) can be rewritten as

$$\frac{1}{2}\sum_{i=1}^{n}[\{y_i - \mathrm{E}[y_i|t_i]\} - \{\mathbf{W}_i - \mathrm{E}[\mathbf{W}_i|t_i]\}^{\top}\beta]^2 - \frac{n}{2}\beta^{\top}\mathbf{\Sigma}_{uu}\beta.$$

This motivates us to consider a penalized least squares method based on partial residuals. Let $\hat{g}_{w,h}(t)$ and $\hat{g}_{y,h}(t)$ be the kernel regressions with bandwidth $h$ to estimating $g_w(t) = \mathrm{E}[\mathbf{W}|t]$ and $g_y(t) = \mathrm{E}[Y|t]$, respectively. Then, the penalized least squares function based on partial residuals is defined as

$$\frac{1}{2}\sum_{i=1}^{n}[\{y_i - \hat{g}_{y,h}(t_i)\} - \{\mathbf{W}_i - \hat{g}_{w,h}(t_i)\}^{\top}\beta]^2 - \frac{n}{2}\beta^{\top}\mathbf{\Sigma}_{uu}\beta. \tag{12.5}$$

Now, by incorporating the $l_1$ penalty in the objective function in Eq. (12.5), the LASSO estimator of $\beta$ is defined as

$$\hat{\beta}_n^{\lambda} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} [\{y_i - \hat{g}_{y,h}(t_i)\} - \{W_i - \hat{g}_{w,h}(t_i)\}^{\top}\beta]^2 - n\beta^{\top}\boldsymbol{\Sigma}_{uu}\beta + \lambda_n \sum_{j=1}^{p} |\beta_j| \right\} \tag{12.6}$$

where $p_\lambda(\beta_j)$ is the penalty function on $\beta_j$ and $\lambda$ is the penalized tuning parameter, which may be chosen by a data-driven method. When $p_\lambda(\beta_j) = \lambda|\beta_j|$, then $\hat{\beta}_n^{\lambda}$ is the LASSO estimator. Since in LASSO the $l_1$ penalty forces the coefficients to be equally penalized, no estimator based on the LASSO can attain the oracle property. To overcome this problem, $\hat{\beta}_n^{\lambda}$ can be modified as an ALASSO estimator (Zhou, 2006) by setting $p_\lambda(\beta_j) = \lambda w_j|\beta_j|$ where $w_j$ can be chosen to be a proxy of the inverse of an initial estimator of $\beta_j$. It can be proved under a linear model setup that ALASSO possesses the oracle property. If $\boldsymbol{\Sigma}_{uu}$ is unknown, we can estimate it by computing partial replication, and we assume there exist replicated observations such that $\mathbf{W}_{is} = \mathbf{X}_i + \mathbf{U}_{is}$, $s = 1, 2, \ldots, N_i$. On the same approach outlined in Liang et al. (1999), we assume that a fraction $l$ of the data has such replicates. Denote $\bar{\mathbf{W}}_i$ as the sample mean of the replicates. Then, the consistent estimator of $\boldsymbol{\Sigma}_{uu}$ is (Carroll et al., 2006; Liang et al., 1999)

$$\hat{\boldsymbol{\Sigma}}_{uu} = \frac{\sum_{i=1}^{n}\sum_{s=1}^{N_i}(\mathbf{W}_{is} - \bar{\mathbf{W}}_i)(\mathbf{W}_{is} - \bar{\mathbf{W}}_i)^{\top}}{\sum_{i=1}^{n}(N_i - 1)}. \tag{12.7}$$

Thus, a feasible ALASSO estimator for $\boldsymbol{\beta}$ can be proposed as

$$\hat{\beta}_n^{\lambda} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} [\{y_i - \hat{g}_{y,h}(t_i)\} - \{W_i - \hat{g}_{w,h}(t_i)\}^{\top}\beta]^2 - n\left(1 - \frac{l}{2}\right)\beta^{\top}\hat{\boldsymbol{\Sigma}}_{uu}\beta \right.$$
$$\left. + \lambda_n \sum_{j=1}^{p} |\beta_j| \right\}. \tag{12.8}$$

To examine the asymptotic properties of $\hat{\beta}_n^{\lambda}$, we first define the random function $Z_n(\phi)$ as

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} [\{y_i - \hat{g}_{y,h}(t_i)\} - \{W_i - \hat{g}_{w,h}(t_i)\}^{\top}\Phi]^2 - \Phi^{\top}\boldsymbol{\Sigma}_{uu}\Phi + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\phi_j|, \tag{12.9}$$

which is minimized at $\Phi = \hat{\beta}_n^{\lambda}$. Consider the following lemma:

**Lemma 12.1** *According to Liang et al. (1999), we have*

*(i)* $\underset{n\to\infty}{\text{plim}} \quad n^{-1}\mathbf{C}_n = \mathbf{C}$ *where* $\mathbf{C}_n = (\mathbf{W}^*)^\top\mathbf{W}^* - n\boldsymbol{\Sigma}_{uu}$

*(ii)* $\underset{n\to\infty}{\text{plim}} \quad n^{-1}(\mathbf{W}^*)^\top\mathbf{y}^* = \mathbf{C}\beta$

*(iii)* $\underset{n\to\infty}{\text{plim}} \quad n^{-1}(\mathbf{y}^*)^\top\mathbf{y}^* = \beta^\top\mathbf{C}\beta + \sigma^2.$

*where* $y_i^* = y_i - \sum_{j=1}^n \omega_{nj}(t_i)y_j$, $\mathbf{W}_i^* = \mathbf{W}_i - \sum_{j=1}^n \omega_{nj}(t_i)\mathbf{W}_j$ *and* $\underset{n\to\infty}{\text{plim}}$ *denotes the probability in limit when n tends to* $\infty$.

The following theorem shows that $\hat{\beta}_n^\lambda$ is consistent provided $\lambda_n = o(n)$.

**Theorem 12.1** *If* $\mathbf{C}$ *is non-singular and* $(\lambda_n / n) \to \lambda_0 > 0$, *then* $\hat{\beta}_n^\lambda \xrightarrow{\mathcal{P}} \text{argmin}(Z)$, *where* $\mathbf{C}$ *and* $\mathbf{Z}$ *are as defined above.*

***Proof*** Using Lemma 12.1 as $n \to \infty$ and $(\lambda_n / n) \to \lambda_0$, we have

$$Z_n(\phi) = n^{-1}\left((\mathbf{y}^*)^\top\mathbf{y}^* - 2\phi(\mathbf{W}^*)^\top\mathbf{y}^* + \Phi^\top\mathbf{C}_n\Phi + \lambda_n\sum_{j=1}^p|\phi_j|\right) \xrightarrow{\mathcal{P}} Z(\Phi) + \sigma^2, \tag{12.10}$$

where $Z(\Phi) = (\Phi - \beta)^\top\mathbf{C}(\Phi - \beta) + \lambda_0\sum_{j=1}^p \mid \phi_j \mid$. Since $Z_n$ is convex, so from the point-wise convergence in probability of $Z_n(\Phi)$ to $Z(\Phi) + \sigma^2$ is obtained. By applying standard results (see Andersen & Gill, 1982; Knight & Fu, 2000), it follows that for any compact $K$,

$$\sup_{\phi\in K}\|Z_n(\Phi) - Z(\Phi) - \sigma^2\| \xrightarrow{\mathcal{P}} 0 \quad \text{and then} \quad \hat{\beta}_n = O_p(1). \tag{12.11}$$

Consequently, under Eq. (12.11), we have $\text{argmin}(Z_n) \xrightarrow{\mathcal{P}} \text{argmin}(Z)$. □

**Lemma 12.2** *Suppose* $\mathrm{E}(\epsilon^4 + \|\mathbf{U}\|^4) < \infty$ *and let* $\mathbf{S}_n = n^{\frac{-1}{2}}(\mathbf{W}^*)^\top\epsilon_w$ *where* $\epsilon_w = y^* - n\mathbf{W}^*((\mathbf{W}^*)^\top\mathbf{W}^*)^{-1}\mathbf{C}_n\beta$, *then* $\mathbf{S}_n \xrightarrow{\mathcal{D}} \mathbf{S}$, *where* $\mathbf{S} \sim N(0, \boldsymbol{\Lambda})$ *with* $\boldsymbol{\Lambda} = [\mathrm{E}[\epsilon - U^\top\beta]\tilde{\mathbf{X}}]^{\otimes 2} + \mathrm{E}[\mathbf{U}^\top\mathbf{U} - \boldsymbol{\Sigma}_{uu}\beta]^{\otimes 2} + \mathrm{E}[\mathbf{U}^\top\mathbf{U}\epsilon^2]$.

***Proof*** The proof is similar to the proof of Theorem 3.1 of Liang et al. (1999) and we omit it. □

Now, define the random function $V_n(e)$ as

$$V_n(e) = \left\|\epsilon_w - \frac{\mathbf{W}^*e}{\sqrt{n}}\right\|^2 - \|\epsilon_w^2\|^2 - e^\top\Sigma_{uu}\,e + \lambda_n\sum_{j=1}^p\left(\left|\beta_j + \frac{e_j}{\sqrt{n}}\right| + |\beta_j|\right) \tag{12.12}$$

where $e = (e_1, \ldots, e_p)^\top$, and $\mathbf{W}_i^* = \mathbf{W}_i - \sum_{j=1}^n \omega_{nj}(t_i)\mathbf{W}_j$ is the $i$-th column of $\mathbf{W}^*$. Note that $V_n$ is minimized at $\sqrt{n}(\hat{\beta}_n^\lambda - \beta)$. It can be shown that

$$\left\|\epsilon_w - \frac{\mathbf{W}^*e}{\sqrt{n}}\right\|^2 - \|\epsilon_w^2\|^2 - e^\top\Sigma_{uu}e \xrightarrow{\mathcal{D}} -2e^\top\mathbf{S} + e^\top\mathbf{C}e,$$

where $\mathbf{S} \sim N(0, \mathbf{\Lambda})$ and

$$\lambda_n \sum_{j=1}^{p} \left( \left| \beta_j + \frac{e_j}{\sqrt{n}} \right| + |\beta_j| \right) \to \lambda_0 \sum_{j=1}^{p} [e_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |e_j| I(\beta_j = 0)].$$

Therefore, we have the following theorem.

**Theorem 12.2** *If* $\mathbf{C}$ *is non-singular and* $(\lambda_n / \sqrt{n}) \to \lambda_0 \geq 0$, *then*

$$\sqrt{n}(\hat{\beta}_n^{\lambda} - \beta) \xrightarrow{\mathcal{D}} \text{argmin}(V),$$

*where*

$$V(e) = -2e^{\top}\mathbf{S} + e^{\top}\mathbf{C}e + \lambda_0 \sum_{j=1}^{p} [e_j \text{sgn}(\beta_j) I(\beta_j \neq 0) + |e_j| I(\beta_j = 0)]. \tag{12.13}$$

***Proof*** As defined above, we have $\text{argmin}(V_n) \xrightarrow{\mathcal{D}} \text{argmin}(V)$, also $V_n$, is convex and $V$ has a unique minimum; thus, $\text{argmin}(V_n) = \sqrt{n}(\hat{\beta}_n^{\lambda} - \beta) \xrightarrow{\mathcal{D}} \text{argmin}(V)$. □

Theorem 12.2 indicates that we need $\lambda_n = O(\sqrt{n})$ for $\sqrt{n}$-consistency. Note that when $\lambda_0 \to 0$, $\text{argmin}(V) = \mathbf{C}^{-1}\mathbf{S} \sim N(0, \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1})$.

In statistical research, there have been many attempts to provide better estimators by the incorporation of prior information available in the form of exact or stochastic restrictions. Suppose that there is auxiliary information on the vector of regression coefficients such that this prior information can be written in the form of exact linear restrictions

$$\mathbf{R}\beta = r \tag{12.14}$$

with $r$ a $q$-vector and $\mathbf{R}$ a given $q \times p$-matrix. Minimizing the sum of squared errors in Eq. (12.5) under the linear restriction in Eq. (12.14) with respect to $\beta$, the restricted LASSO (RL) estimator is given by

$$\hat{\beta}_n^{\text{R}\lambda} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} [\{y_i - \hat{g}_{y,h}(t_i)\} - \{W_i - \hat{g}_{w,h}(t_i)\}^{\top}\beta]^2 - n\beta^{\top}\mathbf{\Sigma}_{uu}\beta \right\}$$

$$\text{s.t.} \quad \mathbf{R}\beta = r$$

$$= \hat{\beta}_n^{\lambda} - \mathbf{C}_n^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}_n^{-1}\mathbf{R})^{-1}(\mathbf{R}\hat{\beta}_n^{\lambda} - r) \tag{12.15}$$

where $\mathbf{C}_n = \sum_{i=1}^{n} [W_i - \hat{g}_{w,h}(t_i)][W_i - \hat{g}_{w,h}(t_i)]^{\top} - n\mathbf{\Sigma}_{uu}$.

### 12.2.1 Test Statistics

Consider null hypothesis $\mathcal{H}_0 : \mathbf{R}\beta = r$ against local alternative hypothesis $\mathcal{H}_n : \mathbf{R}\beta = r + \xi / \sqrt{n}$ where $\xi \in R^q$. When $\lambda_n = o(\sqrt{n})$, to test $\mathcal{H}_0$, we use the following Wald LASSO test statistics:

$$\mathcal{L}_n = \frac{1}{n}(\mathbf{R}\hat{\beta}_n^{\lambda} - r)^{\top}[\mathbf{R}^{\top}\mathbf{C}_n^{-1}\mathbf{\Lambda}_n\mathbf{C}_n^{-1}\mathbf{R}]^{-1}(\mathbf{R}\hat{\beta}_n^{\lambda} - r) \tag{12.16}$$

where $\mathbf{\Lambda}_n = \sum_{i=1}^{n}[(W_i^*(y_i^* - (W^*)_i^{\top}\hat{\beta}_n^{\lambda}) + \mathbf{\Sigma}_{uu}\hat{\beta}_n^{\lambda}]^{\otimes 2}$ and $n^{-1}\mathbf{\Lambda}_n$ is a consistent estimator of $\mathbf{\Lambda}$.

**Theorem 12.3** *Under $\mathcal{H}_n$, as $n \to \infty$, $\mathcal{L}_n \xrightarrow{\mathcal{D}} \mathcal{L}$ which $\mathcal{L}$ follows the asymptotic noncentral $\chi^2(q, \Delta^2)$ distribution with $q$ degrees of freedom, and the noncentral parameter is $\Delta^2 = \xi^{\top}[\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{R}]^{-1}\xi$ and defined as*

$$\mathcal{L} = (\mathbf{RS} + \xi)^{\top}[\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{R}]^{-1}(\mathbf{RS} + \xi).$$

***Proof*** Under the condition of weak consistency, i.e., $\lambda_n = o(\sqrt{n})$, we have $\sqrt{n}(\hat{\beta}_n^{\lambda} - \beta) \xrightarrow{\mathcal{D}} \sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{D}} \mathbf{S}$, and the proof can be obtained using Slutsky's theorem. □

### 12.2.2 Proposed LASSO Estimators and Their Asymptotic Distributions

When the prior information is uncertain, the doubt on this prior information can be removed by using Fisher's recipe of testing the hypothesis $\mathcal{H}_0$ against $\mathcal{H}_n$. As a result of this test, $\hat{\beta}_n^{\lambda}$ or $\hat{\beta}_n^{\mathsf{R}\lambda}$ can be chosen. Accordingly, the estimator is written as

$$\hat{\beta}_n^{\mathsf{P}\lambda} = \hat{\beta}_n^{\lambda} - (\hat{\beta}_n^{\lambda} - \hat{\beta}_n^{\mathsf{R}\lambda})I(\mathcal{L}_n < \chi_{q,\alpha}^2) \tag{12.17}$$

and is called as the preliminary test LASSO estimator (PLE), where $I(A)$ is the indicator function of set. It can be shown that $\hat{\beta}_n^{\mathsf{P}\lambda}$ performs better than $\hat{\beta}_n^{\lambda}$ in some parts of the parameter space. The use of $\hat{\beta}_n^{\mathsf{P}\lambda}$ may, however, be limited due to its heavily dependence on $\alpha$, the level of significance, and also it yields the extreme results. To avoid the dependency of the estimator on $\alpha$, we devise the shrinkage Stein-type estimator, which is introduced by Stein (1956) and James and Stein (1961). This method of estimation is more sophisticated and improved than its precursor, the preliminary test estimation. Using the idea of Stein (1956), the Stein-type shrinkage LASSO estimator (SLASSO) is proposed as

$$\hat{\beta}_n^{\mathsf{S}\lambda} = \hat{\beta}_n^{\mathsf{R}\lambda} + (1 - k_n\mathcal{L}_n^{-1})(\hat{\beta}_n^{\lambda} - \hat{\beta}_n^{\mathsf{R}\lambda}), \qquad k_n = \frac{(q-2)(n-p)}{n-p+2}. \tag{12.18}$$

$\hat{\beta}_n^{\mathsf{S}\lambda}$ may not remain nonnegative. In an effort to fix this problem with $\hat{\beta}_n^{\mathsf{S}\lambda}$, we introduce a positive part shrinkage LASSO estimator:

$$\begin{aligned}\hat{\beta}_n^{+\mathsf{S}\lambda} &= \hat{\beta}_n^{\mathsf{R}\lambda} + (1 - k_n\mathcal{L}_n^{-1})I(\mathcal{L}_n > k_n)(\hat{\beta}_n^{\lambda} - \hat{\beta}_n^{\mathsf{R}\lambda}) \\ &= \hat{\beta}^{\mathsf{S}\lambda} - (1 - k_n\mathcal{L}_n^{-1})I(\mathcal{L}_n < k_n)(\hat{\beta}_n^{\lambda} - \hat{\beta}_n^{\mathsf{R}\lambda}). \qquad (12.19)\end{aligned}$$

Now, to overcome the difficulty of identical asymptotic distributions of LASSO estimators in large samples under fixed alternatives, we study the asymptotic risks under local alternative. To obtain meaningful asymptotic distributions of the proposed $\lambda$ estimators and the test statistics, $\mathcal{L}$, we consider Theorem 12.4.

**Corollary 12.1** *Under $\mathcal{H}_n$ and $\frac{\lambda_n}{n} \to \lambda_0 \geq 0$ as $n \longrightarrow \infty$, we have*

$$\begin{aligned}
&(i)\quad \hat{\beta}_n^{\mathsf{R}\lambda} \xrightarrow{\mathcal{P}} \operatorname{argmin}(Z) - \mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\operatorname{argmin}(Z) - r)\\
&(ii)\quad \hat{\beta}_n^{\mathsf{P}\lambda} \xrightarrow{\mathcal{P}} \operatorname{argmin}(Z) - \mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\operatorname{argmin}(Z) - r)I(\mathcal{L} < \chi^2_{q,\alpha})\\
&(iii)\quad \hat{\beta}_n^{\mathsf{S}\lambda} \xrightarrow{\mathcal{P}} \operatorname{argmin}(Z) - k\mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\operatorname{argmin}(Z) - r)\mathcal{L}^{-1}\\
&(iv)\quad \hat{\beta}_n^{+\mathsf{S}\lambda} \xrightarrow{\mathcal{P}} \operatorname{argmin}(Z) - [k\mathcal{L}^{-1} + (1-k)I(\mathcal{L} < \chi^2_{q,\alpha})]\\
&\qquad\qquad \times\mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\operatorname{argmin}(Z) - r)
\end{aligned}$$

**Corollary 12.2** *Under $\mathcal{H}_n$ and the assumptions of as $n \longrightarrow \infty$, we have*

$$\begin{aligned}
&(i)\quad \sqrt{n}(\hat{\beta}_n^{\mathsf{R}\lambda} - \beta) \xrightarrow{\mathcal{D}} \operatorname{argmin}(V) - \mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\operatorname{argmin}(V) + \xi)\\
&(ii)\quad \sqrt{n}(\hat{\beta}_n^{\mathsf{P}\lambda} - \beta) \xrightarrow{\mathcal{D}} \operatorname{argmin}(V)\\
&\qquad\qquad -\mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\ \operatorname{argmin}(V) + \xi)I(\mathcal{L} < \chi^2_{q,\alpha})\\
&(iii)\quad \sqrt{n}(\hat{\beta}_n^{\mathsf{S}\lambda} - \beta) \xrightarrow{\mathcal{D}} \operatorname{argmin}(V)\\
&\qquad\qquad -k\mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\ \operatorname{argmin}(V) + \xi)\mathcal{L}^{-1}\\
&(iv)\quad \sqrt{n}(\hat{\beta}_n^{+\mathsf{S}\lambda} - \beta) \xrightarrow{\mathcal{D}} \operatorname{argmin}(V)\\
&\qquad\qquad -[k\mathcal{L}^{-1} + (1-k)I(\mathcal{L} < \chi^2_{q,\alpha})]\\
&\qquad\qquad \times\mathbf{C}^{-1}\mathbf{R}^{\top}(\mathbf{R}^{\top}\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\ \operatorname{argmin}(V) + \xi)
\end{aligned}$$

**Theorem 12.4** *If $\lambda_n = o(\sqrt{n})$, under $\mathcal{H}_n$ and the assumptions of as $n \longrightarrow \infty$, we have:*

*(i)* $\sqrt{n}(\hat{\beta}_n^{\lambda} - \beta) \overset{\mathcal{D}}{\sim} \sqrt{n}(\hat{\beta}_n - \beta)$

*(ii)* $$\begin{pmatrix} \sqrt{n}(\hat{\beta}_n^{\lambda} - \beta) \\ \sqrt{n}(\hat{\beta}_n^{\mathsf{R}\lambda} - \beta) \\ \sqrt{n}(\hat{\beta}_n^{\lambda} - \hat{\beta}_n^{\mathsf{R}\lambda}) \end{pmatrix} \xrightarrow{\mathcal{D}} N_{3p}\left(\begin{pmatrix} 0 \\ -\delta \\ \delta \end{pmatrix}, \tilde{\mathbf{C}}\right) \text{ where}$$

$$\tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1} & (\mathbf{I}_p - \mathbf{A})\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1} & \mathbf{A}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1} \\ \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}(\mathbf{I}_p - \mathbf{A})^\top & (\mathbf{I}_p - \mathbf{A})\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}(\mathbf{I}_p - \mathbf{A})^\top & (\mathbf{I}_p - \mathbf{A})\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{A}^\top \\ \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{A} & \mathbf{A}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}(\mathbf{I}_p - \mathbf{A})^\top & \mathbf{A}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{A}^\top \end{pmatrix} \tag{12.20}$$

*with* $\mathbf{A} = \mathbf{C}^{-1}\mathbf{R}^\top(\mathbf{R}\mathbf{C}^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}$ *and* $\delta = \mathbf{C}^{-1}\mathbf{R}^\top(\mathbf{R}\mathbf{C}^{-1}\mathbf{R}^\top)^{-1}\xi$.

*(iii)*

$$\begin{aligned} &\lim_{n\to\infty} P(\sqrt{n}(\hat{\beta}_n^{\mathsf{P}\lambda} - \beta) \le x) = \\ &\int_{E_1(\delta)} \psi_p\Big(x + \delta - \mathbf{T}\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{R}^\top\mathbf{M}^{-1}\mathbf{J}, 0, \\ &\qquad \mathbf{T}\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{R}^\top\mathbf{M}^{-1}\mathbf{R}\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{T}\Big)\mathrm{d}\psi_p(x + \delta, 0, \mathbf{M}) \\ &+ \int_{E_2(\delta)} \psi_p\Big(x - \mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{J}, 0, \\ &\qquad \mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}[\mathbf{I}_p - \mathbf{R}^\top\mathbf{M}^{-1}\mathbf{R}\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}]\Big)\mathrm{d}\psi_p(x + \delta, 0, \mathbf{M}) \end{aligned} \tag{12.21}$$

*where* $\psi_p(\cdot, \mu, \Sigma)$ *is the c.d.f. of a p-variate normal distribution with mean* $\mu$ *and covariance matrix* $\Sigma$, $\mathbf{M} = \mathbf{H}\mathbf{C}^{-1}\mathbf{V}\mathbf{C}^{-1}\mathbf{H}^\top$, $\mathbf{J} \sim N(0, \mathbf{M})$, $E_1(\delta) = \{\mathbf{J}; \theta(\mathbf{J}) < \chi^2_{q,\alpha}\}$ *and* $E_2(\delta) = \{\mathbf{J}; \theta(\mathbf{J}) \ge \chi^2_{q,\alpha}\}$ *in which* $\theta(\mathbf{J}) = (\mathbf{J} + \delta)^\top\mathbf{M}^{-1}(\mathbf{J} + \delta)$.

*(iv)* $$\sqrt{n}(\hat{\beta}_n^{\mathsf{S}\lambda} - \beta) \overset{\mathcal{D}}{\sim} \mathbf{J} - k\frac{\mathbf{C}^{-1}\mathbf{R}^\top(\mathbf{R}^\top\mathbf{C}^{-1}\mathbf{H})^{-1}(\mathbf{R}\mathbf{J} + \xi)}{(\mathbf{R}\mathbf{J} + \xi)^\top\mathbf{M}^{-1}(\mathbf{R}\mathbf{J} + \xi)}$$

*(v)*

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n^{+\mathsf{S}\lambda} - \beta) \overset{\mathcal{D}}{\sim} \mathbf{J} &- k\frac{\mathbf{C}^{-1}\mathbf{R}^\top(\mathbf{R}^\top\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\mathbf{J} + \xi)}{(\mathbf{R}\mathbf{J} + \xi)^\top\mathbf{M}^{-1}(\mathbf{R}\mathbf{J} + \xi)} \\ &\times[\mathbf{C}^{-1}\mathbf{R}^\top(\mathbf{R}^\top\mathbf{C}^{-1}\mathbf{R})^{-1}(\mathbf{R}\mathbf{J} + \xi)] \\ \times\left\{1 - \frac{k}{(\mathbf{R}\mathbf{J} + \xi)^\top\mathbf{M}^{-1}(\mathbf{R}\mathbf{J} + \xi)}\right\} &I\left((\mathbf{R}\mathbf{J} + \xi)^\top\mathbf{M}^{-1}(\mathbf{R}\mathbf{J} + \xi) < k\right) \end{aligned}$$

## 12.3 Asymptotic Distributional Risk Properties

In order to investigate the asymptotic distributional bias (ADB), quadratic bias (ADQB), MSE (AMSE), and quadratic risks (ADQR) of the proposed LASSO estimators, we define the asymptotic distributions of these estimators and of the test statistic $\mathcal{L}_n$. Consider the generic estimator $\beta_n^*$ and a positive-semidefinite matrix $\mathbf{Q}$ defining the loss function:

$$L(\beta_n^*, \beta) = n(\beta_n^* - \beta)^\top (\beta_n^* - \beta) = \text{tr}(\mathbf{Q}n(\beta_n^* - \beta)^\top (\beta_n^* - \beta)). \tag{12.22}$$

The AMSE is $\mathbf{M}(\beta_n^*) = \lim_{n\to\infty} \text{E}[(\beta_n^* - \beta)(\beta_n^* - \beta)^\top]$ and then the ADQR of $\beta_n^*$ is given by $R(\beta_n^*, \beta) = \text{tr}(\mathbf{QM}(\beta_n^*))$. If the asymptotic distribution of $\beta_n^*$ exists, then the ADB and ADQB of $\beta_n^*$ are given by

$$\mathbf{b}(\beta_n^*) = \lim_{n\to\infty} \text{E}[\sqrt{n}(\beta_n^* - \beta)] = \int x \text{d}G(x),$$
$$\mathbf{Qb} = \mathbf{b}(\beta_n^*)^\top \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{b}(\beta_n^*),$$

respectively, where $G(x)$ is the asymptotic distribution of $\sqrt{n}(\beta_n^* - \beta)$ under $\mathcal{H}_n$. Lemma 12.3 is useful to obtain the bias, MSE, and risk expressions of the proposed LASSO-type estimator.

**Lemma 12.3 (Saleh, 2006)**

$$\text{E}[\chi_{q+2}^{-2}(\Delta)] = \exp\left\{\frac{-\Delta^2}{2}\right\} \sum_{r=0}^{\infty} \frac{1}{r!}\left(\frac{\Delta}{2}\right)^r \frac{1}{q+2r} = \text{E}_r(q+r)$$

$$\text{E}[\chi_{q+2}^{-2}(\Delta) I(\chi_{q+2}^2(\Delta) < c)] = \text{E}_r(q+2r)^{-1}\mathcal{H}_{q+2r}(c;0)$$

$$\text{E}[\chi_{q+2}^{-4}(\Delta)] = \exp\left\{\frac{-\Delta^2}{2}\right\} \sum_{r=0}^{\infty} \frac{1}{r!}\left(\frac{\Delta}{2}\right)^r \frac{1}{(q+2r)(q+2r-2)}$$
$$= \text{E}_r[(q+2r)(q+2r-2)]^{-1}$$

$$\text{E}[\chi_{q+2}^{-4}(\Delta) I(\chi_{q+2}^2(\Delta) < c)] = \text{E}_r[(q+2r)(q+2r-2)]^{-1}\mathcal{H}_{q+2r-2}(c;0)$$

*and* $\text{E}_r$ *stands for the expectation with respect to a Poisson variable* $r$ *with mean* $\Delta/2$.

Since matrices $\mathbf{C}^{-1/2}\mathbf{R}^\top(\mathbf{RC}^{-1}\mathbf{R}^\top)^{-1}\mathbf{RC}^{-1/2}$ and $\mathbf{C}^{-1/2}\mathbf{\Lambda}\mathbf{C}^{-1/2}$ are symmetric and idempotent with rank $q \le p$ and $p$, respectively, there exists an orthogonal matrix such that

$$\Gamma\mathbf{C}^{-1/2}\mathbf{R}^\top(\mathbf{RC}^{-1}\mathbf{R}^\top)^{-1}\mathbf{RC}^{-1/2}\Gamma^\top = \begin{pmatrix}\mathbf{I}_q & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$$

$$\Gamma \mathbf{C}^{-1/2} \mathbf{\Lambda} \mathbf{C}^{-1/2} \Gamma^{\top} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{12}^{\top} & \mathbf{G}_{22} \end{pmatrix}$$

$$\Gamma \mathbf{C}^{-1/2} \mathbf{Q} \mathbf{C}^{-1/2} \Gamma^{\top} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{12}^{\top} & \mathbf{F}_{22} \end{pmatrix} \tag{12.23}$$

The matrices $\mathbf{F}_{11}$ and $\mathbf{G}_{11}$ are of order $q$ and $\mathbf{F}_{12}$ and $\mathbf{G}_{12}$ are of order $p - q$, respectively. Now, setting $\delta_* = \Gamma \mathbf{C}^{-1/2} \mathbf{R}^{\top} (\mathbf{R} \mathbf{C}^{-1} \mathbf{R}^{\top})^{-1} \xi$ and $\delta_* = (\delta_{*1}, \delta_{*2})^{\top}$, we have the following theorem.

**Theorem 12.5** *Under $\sqrt{n}$-consistency of the proposed estimators, the bias and risk expressions of the five LASSO-type estimators are given as*

$(i)$ $\mathbf{b}_1(\hat{\beta}_n^{\lambda}) = 0,$

$\mathbf{Q}\mathbf{b}_1(\hat{\beta}_n^{\lambda}) = 0,$

$M_1(\hat{\beta}_n^{\lambda}) = \mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1},$

$R_1(\hat{\beta}_n^{\lambda}, \mathbf{Q}) = \operatorname{tr}(\mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1})$

$(ii)$ $\mathbf{b}_2(\hat{\beta}_n^{\mathsf{R}\lambda}) = -\delta$

$\mathbf{Q}\mathbf{b}_2(\hat{\beta}_n^{\mathsf{R}\lambda}) = \delta^{\top} \delta,$

$M_2(\hat{\beta}_n^{\mathsf{R}\lambda}) = (\mathbf{I}_p - \mathbf{A}) \mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1} (\mathbf{I}_p - \mathbf{A})^{\top} + \delta \delta^{\top}$

$R_2(\hat{\beta}_n^{\mathsf{R}\lambda}, \mathbf{Q}) = \operatorname{tr}(\mathbf{Q}(\mathbf{I}_p - \mathbf{A}) \mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1} (\mathbf{I}_p - \mathbf{A})^{\top}) + \delta^{\top} \mathbf{Q} \delta$

$(iii)$ $\mathbf{b}_3(\hat{\beta}_n^{\mathsf{P}\lambda}, \beta) = -\delta \mathcal{H}_{q+2}(\chi_{\alpha}^2; \Delta^2)$

$\mathbf{Q}\mathbf{b}_3(\hat{\beta}_n^{\mathsf{P}\lambda}) = \Delta^2 [\mathcal{H}_{q+2}(\chi_{\alpha}^2; \Delta^2)]^2$

$$M_3(\hat{\beta}_n^{\mathsf{P}\lambda}, \beta) = -\mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1} - \mathbf{A} \mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1} \mathbf{A}^{\top} \mathcal{H}_{q+2}(\chi_{q,\alpha}^2, \Delta^2) + \delta \delta^{\top} [2\mathcal{H}_{q+2}(\chi_{q,\alpha}^2, \Delta^2) - \mathcal{H}_{q+4}(\chi_{q,\alpha}^2, \Delta^2)]$$

$$R_3(\hat{\beta}_n^{\mathsf{P}\lambda}, \beta) = \operatorname{tr}(\mathbf{Q} \mathbf{C}^{-1} \mathbf{\Lambda} \mathbf{C}^{-1}) - \operatorname{tr}(\mathbf{F}_{11} \mathbf{G}_{11}) \mathcal{H}_{q+2}(\chi_{q,\alpha}^2; \Delta^2) + \delta_{*1}^{\top} \mathbf{F}_{11} \delta_{*1} [2\mathcal{H}_{q+2}(\chi_{q,\alpha}^2; \Delta^2) - \mathcal{H}_{q+4}(\chi_{q,\alpha}^2; \Delta^2)]$$

$$
\begin{aligned}
(iv)\quad & \mathbf{b}_4(\hat{\beta}_n^{\mathsf{S}\lambda}, \beta) = -k\delta\, \mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)] \\
& \mathbf{Qb}_4(\hat{\beta}_n^{\mathsf{P}\lambda}) = k^2\Delta^2\, \mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)]^2 \\
& M_4(\hat{\beta}_n^{\mathsf{S}\lambda}, \beta) = \mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1} \\
& \qquad -k\mathbf{A}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{A}^\top\left\{2\,\mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)] - k\,\mathrm{E}[\chi_{q+2}^{-4}(\Delta^2)]\right\} \\
& \qquad +k\delta\delta^\top\left\{2\,\mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)] - 2\,\mathrm{E}[\chi_{q+4}^{-2}(\Delta^2)] + k\,\mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)]\right\} \\
& R_4(\hat{\beta}_n^{\mathsf{S}\lambda}, \beta) = R_1(\hat{\beta}_n^{\lambda}, \beta) - k\mathrm{tr}(\mathbf{F}_{11}\mathbf{G}_{11})\left\{2\,\mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)] - k\,\mathrm{E}[\chi_{q+4}^{-4}(\Delta^2)]\right\} \\
& \qquad +k(k+4)\delta_{*1}^\top\mathbf{F}_{11}\delta_{*1}\,\mathrm{E}[\chi_{q+4}^{-4}(\Delta^2)] \\
& \mathbf{b}_5(\hat{\beta}_n^{+\mathsf{S}\lambda}, \beta) = \mathbf{b}_4(\hat{\beta}_n^{\mathsf{S}\lambda}, \beta) - \delta\,\mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2)I(\chi_{q+2}^2 < k))] \\
& \mathbf{Qb}_5(\hat{\beta}_n^{+\mathsf{S}\lambda}) = \Delta^2\left\{k\,\mathrm{E}[\chi_{q+2}^{-2}(\Delta^2)] - \mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2)I(\chi_{q+2}^2(\Delta^2) < k))]\right\}^2 \\
& M_5(\hat{\beta}_n^{+\mathsf{S}\lambda}, \beta) = M_4(\hat{\beta}_n^{\mathsf{S}\lambda}, \beta) \\
& \qquad -\mathbf{A}\mathbf{C}^{-1}\mathbf{\Lambda}\mathbf{C}^{-1}\mathbf{A}^\top\Big\{\mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) < k)] \\
& \qquad +\delta\delta^\top\,\mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) < k)]\Big\} \\
& \qquad -2\delta\delta^\top\,\mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2))I(\chi_{q+2}^2(\Delta^2) < k)] \\
& R_5(\hat{\beta}_n^{+\mathsf{S}\lambda}, \beta) = \mathcal{R}_4(\hat{\beta}_n^{+\mathsf{S}\lambda}, \beta) \\
& \qquad -\mathrm{tr}(\mathbf{F}_{11}\mathbf{G}_{11})\,\mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) < k)] \\
& \qquad +\delta_{*1}^\top\mathbf{F}_{11}\delta_{*1}\Big\{2\,\mathrm{E}[(1 - k\chi_{q+2}^{-2}(\Delta^2))I(\chi_{q+2}^2(\Delta^2) < k)] \\
& \qquad -\mathrm{E}[(1 - k\chi_{q+4}^{-2}(\Delta^2))^2 I(\chi_{q+4}^2(\Delta^2) < k)]\Big\}.
\end{aligned}
$$

The proof is an adaptation of proof of Theorem 7.2 in Saleh (2006) and thus is omitted.

## 12.4 Algorithm

There are some variable selection algorithm in linear model. The least angle regression (LARS) (Efron et al., 2004) and modified LARS (Wang et al., 2013) algorithms were applied to determine the solution paths for LASSO and ALASSO

in linear regression models, respectively, by Efron et al. (2004) and Zhou (2006). Recently, Wang et al. (2013) applied the modified LARS algorithm for variable selection in the context of varying coefficient linear measurement error models. This approach leverages the benefits of the ALASSO in terms of variable selection and estimation accuracy. The algorithm integrates a modified LARS technique to address the unique features of our model. For our model, we modify this algorithm as follows:

**Step 1: Initialization**

1. Compute the matrix

$$\mathbf{A} = \sum_{i=1}^{n} [\mathbf{W}_i - g_{\omega,h}(t_i)]^{\top} [\mathbf{W}_i - g_{\omega,h}(t_i)] - n\boldsymbol{\Sigma}_{uu}$$

2. Compute the vector

$$y = \mathbf{A}^{-1/2} \sum_{i=1}^{n} [\mathbf{W}_i - g_{\omega,h}(t_i)] Y_i$$

3. Standardize $y$ and $\mathbf{W}$.

**Step 2: Compute ALASSO Weights**

1. Estimate the coefficients using an initial consistent estimator, such as the ordinary least squares or the method of moments estimator, to obtain $\hat{\beta}_{\text{init}}$.
2. Compute the adaptive weights

$$\hat{w}_j = 1/|\hat{\beta}_{\text{init},j}|^{\gamma},$$

where $\gamma > 0$ is a tuning parameter.

**Step 3: Modified LARS Algorithm**

1. Initialize residuals $r = y$, active set $\mathbf{A} = \emptyset$, and inactive set $I = \{1, 2, \ldots, p\}$.
2. While the stopping criterion is not met,
   a. Compute the correlations $c = \mathbf{W}^{\top} r$.
   b. Identify the variable

$$j^* = \underset{j \in I}{\text{argmax}} |c_j \hat{w}_{\cdot j}|$$

   c. Update the active set $\mathbf{A} \leftarrow \mathbf{A} \cup \{j^*\}$ and the inactive set $I \leftarrow I \setminus \{j^*\}$.
   d. Solve the weighted least squares problem

$$\hat{\beta}_{\mathbf{A}} = \underset{\beta_{\mathbf{A}}}{\text{argmin}} \, \|y - \mathbf{W}_{\mathbf{A}} \beta_{\mathbf{A}}\|_2^2 + \lambda \sum_{j \in \mathbf{A}} \hat{w}_j |\beta_j|$$

e. Update the residuals

$$r \leftarrow y - \mathbf{W_A}\hat{\beta}_{\mathbf{A}}\,.$$

f. Adjust the penalty parameter $\lambda$ based on cross-validation or a data-driven method.

**Step 4: Solution Path and Selection**

1. Compute the entire solution path of the ALASSO estimates as a function of $\lambda$.
2. Use a model selection criterion, such as cross-validation, Bayesian information criterion (BIC), or Akaike information criterion (AIC), to select the optimal $\lambda$.

This algorithm extends the traditional LARS algorithm to accommodate the ALASSO framework, particularly suited for the partially linear measurement error models. By incorporating adaptive weights, the algorithm enhances variable selection by penalizing less relevant variables more heavily, which helps in reducing bias and improving the estimation of relevant coefficients. The modification for measurement error models ensures that the unique characteristics of the data are appropriately handled, leading to more accurate and reliable results.

The adaptive weights are crucial as they dynamically adjust the penalty applied to each coefficient based on initial estimates, allowing for better differentiation between significant and insignificant predictors. The use of cross-validation or information criteria for model selection ensures that the chosen model generalizes well to unseen data, balancing model complexity and fit.

In summary, this algorithm provides a robust framework for variable selection and coefficient estimation in complex models, leveraging the strengths of both ALASSO and LARS methodologies.

## 12.5 Simulation

Here, we investigate the small sample properties of the restricted LASSO and ALASSO estimators computed by the modified LARS algorithm via two simulation examples.

### *12.5.1 Example 1*

We simulated 5000 samples from the following data-generating process:

$$\begin{aligned} y &= \mathbf{X}\beta + g(t) + \epsilon \\ \mathbf{W} &= \mathbf{X} + \mathbf{U}\,, \end{aligned} \tag{12.24}$$

where following Fan and Li (2001) and Tibshirani (1996), the true parameters for $\beta$ are $(3, 1.5, 0, 0, 2, 0, 0, 0)^\top$ and $g(t)$ is considered as

$$g(t) = 3\big\{0.1\sin(2\pi t_i) + 0.2\cos(2\pi t_i) + 0.3(\sin(2\pi t_i))^2 + 0.4(\cos(2\pi t_i))^4 + 0.5(\sin(2\pi t_i))^3\big\} \tag{12.25}$$

where $t_i \in (0, 1)$. The chosen nonparametric function is difficult to be estimated and provides a good test case for the nonparametric regression. The vector $\epsilon$ generated from $N(0, 4)$ and matrices $\mathbf{X}$ and $\mathbf{U}$ generated from $N_p(0, \mathbf{I}_p)$ and $N_p(0, \sigma_u \mathbf{I}_p)$ where $p = 8$ and the standard deviation of measurement error $\sigma_u$ is considered for two cases: 0.5 and 1.

For restricted setting $\mathcal{H}_n$, we consider

$$\mathbf{H} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 2 & 0 & 0 & 0 \end{pmatrix}, \quad h = \mathbf{H}\beta, \quad \text{and} \quad \xi = (1, 1, 1)^\top.$$

The simulation studies were conducted using R programming software to generate data sets of $n = 60$ observations for comparing the performance of the proposed restricted estimators. We use the least squares estimator $\hat{\beta}$ defined in Eq. (12.8) as the consistent estimator $\beta$ included in the formula of the ALASSO estimator. The estimation of the model given by Eq. (12.24) also involves estimation of nonparametric term $g(t)$ even though our main interest centers on comparing the risk of estimators. We use local linear regression with the Gaussian kernel function as the basis for estimating the nonparametric term. In our implementation, we use direct plug-in methodology to select the bandwidth of a local linear Gaussian kernel regression estimate, as described by Ruppert et al. (1995). Furthermore, to select the tuning parameter $\lambda_n$, we use 5-fold cross-validation in both the LASSO and the ALASSO penalty functions. Tables 12.1 and 12.2 represent the results for $\alpha = 0.05$ and $\alpha = 0.15$, respectively, as the level of the test. We numerically calculate the estimates of parameters, squared bias of estimates, risk ($R(\cdot)$), and $\text{mse}(\hat{f}(t)) = (1/n)\sum_{i=1}^{n}(\hat{f}(t_i) - f(t_i))^2$. It can be seen that risks are stable within each class of estimators, with few exceptions, and generally, the ALASSO estimators always yield smaller risk and $\text{mse}(\hat{f})$ than the LASSO estimators at each level of $\sigma_u$ and $\alpha$. The usual LASSO and ALASSO estimators $\hat{\beta}_n^{\lambda}$ will generally have inferior performance compared with restricted estimators. In the class of restricted LASSO and ALASSO estimators, the smallest risk is achieved by RLE, followed by PLE, PSLE, and SLE, whereas the largest risk is obtained by SLE and PSLE. In other words, when $\mathcal{H}_0$ is true ($\xi$ is close to zero), the performance of RLE is remarkably good. As one may expect, by increasing the level of measurement error's variance, the risk of all estimators will be increased, and for low level of $\alpha$, the PLE performs better. The behavior of the risks for the two classes of estimators under $\mathcal{H}_n$

**Table 12.1** The finite sample performance of the LASSO and ALASSO restricted shrinkage-type estimators for $\alpha = 0.05$

| LASSO | $\sigma_u = 0.5$ | | | | | ALASSO | $\sigma_u = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ | | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ |
| 3.237 | 3.000 | 3.177 | 3.192 | 3.194 | | 3.262 | 3.000 | 3.205 | 3.217 | 3.218 |
| 1.582 | 1.537 | 1.569 | 1.573 | 1.573 | | 1.599 | 1.537 | 1.584 | 1.588 | 1.589 |
| 0.002 | 0.002 | 0.003 | 0.002 | 0.002 | | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| −0.001 | −0.000 | −0.001 | −0.001 | −0.001 | | −0.002 | −0.001 | −0.002 | −0.002 | −0.002 |
| 2.138 | 2.075 | 2.121 | 2.126 | 2.126 | | 2.163 | 2.075 | 2.144 | 2.148 | 2.149 |
| 0.001 | 0.002 | 0.001 | 0.002 | 0.002 | | −0.001 | 0.001 | −0.000 | −0.000 | −0.001 |
| 0.001 | 0.003 | 0.002 | 0.001 | 0.001 | | 0.000 | 0.002 | 0.001 | 0.000 | 0.000 |
| −0.002 | −0.002 | −0.003 | −0.002 | −0.002 | | −0.001 | −0.001 | −0.002 | −0.001 | −0.001 |
| 0.082 | 0.007 | 0.051 | 0.058 | 0.059 | ←-- bias$^2$ --→ | 0.105 | 0.007 | 0.070 | 0.077 | 0.078 |
| 0.274 | 0.062 | 0.220 | 0.232 | 0.228 | ←-- Risk --→ | 0.251 | 0.030 | 0.198 | 0.230 | 0.220 |
| 2.080 | 0.723 | 1.665 | 1.463 | 1.490 | ←-- mse$f(t)$ --→ | 2.530 | 0.392 | 1.237 | 1.938 | 1.621 |
| LASSO | $\sigma_u = 1$ | | | | | ALASSO | $\sigma_u = 1$ | | | |
| $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ | | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ |
| 3.195 | 3.000 | 3.027 | 3.129 | 3.138 | | 3.223 | 3.000 | 3.037 | 3.154 | 3.162 |
| 1.554 | 1.537 | 1.527 | 1.542 | 1.543 | | 1.570 | 1.537 | 1.533 | 1.555 | 1.555 |
| −0.003 | −0.003 | −0.001 | −0.003 | −0.003 | | 0.001 | 0.000 | 0.002 | 0.001 | 0.001 |
| 0.002 | 0.003 | 0.003 | 0.001 | 0.002 | | 0.003 | 0.004 | 0.004 | 0.002 | 0.003 |
| 2.098 | 2.075 | 2.063 | 2.085 | 2.085 | | 2.121 | 2.075 | 2.072 | 2.102 | 2.104 |
| 0.010 | 0.011 | 0.011 | 0.009 | 0.009 | | 0.012 | 0.013 | 0.014 | 0.012 | 0.012 |
| −0.003 | −0.003 | −0.002 | −0.003 | −0.003 | | −0.003 | −0.003 | −0.003 | −0.003 | −0.003 |
| 0.006 | 0.009 | 0.007 | 0.005 | 0.007 | | 0.004 | 0.008 | 0.006 | 0.005 | 0.005 |
| 0.051 | 0.007 | 0.006 | 0.026 | 0.028 | ←-- bias$^2$ --→ | 0.069 | 0.007 | 0.008 | 0.037 | 0.040 |
| 0.931 | 0.328 | 0.480 | 0.727 | 0.697 | ←-- Risk --→ | 0.869 | 0.237 | 0.411 | 0.659 | 0.631 |
| 11.258 | 3.47 | 5.139 | 8.623 | 8.308 | ←-- mse$f(t)$ --→ | 11.265 | 2.659 | 4.769 | 8.499 | 8.124 |

**Table 12.2** The finite sample performance of the LASSO and ALASSO restricted shrinkage-type estimators for $\alpha = 0.15$

| LASSO | $\sigma_u = 0.5$ | | | | | ALASSO | $\sigma_u = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ | | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ |
| 3.237 | 3.000 | 3.177 | 3.192 | 3.194 | | 3.262 | 3.000 | 3.205 | 3.217 | 3.218 |
| 1.582 | 1.537 | 1.569 | 1.573 | 1.573 | | 1.599 | 1.537 | 1.584 | 1.588 | 1.589 |
| 0.002 | 0.002 | 0.003 | 0.002 | 0.002 | | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| −0.001 | −0.000 | −0.001 | −0.001 | −0.001 | | −0.002 | −0.001 | −0.002 | −0.002 | −0.002 |
| 2.138 | 2.075 | 2.121 | 2.126 | 2.126 | | 2.163 | 2.075 | 2.144 | 2.148 | 2.149 |
| 0.001 | 0.002 | 0.001 | 0.002 | 0.002 | | −0.001 | 0.001 | −0.000 | −0.000 | −0.001 |
| 0.001 | 0.003 | 0.002 | 0.001 | 0.001 | | 0.000 | 0.002 | 0.001 | 0.000 | 0.000 |
| −0.002 | −0.002 | −0.003 | −0.002 | −0.002 | | −0.001 | −0.001 | −0.002 | −0.001 | −0.001 |
| 0.082 | 0.007 | 0.051 | 0.058 | 0.059 | ←-- bias$^2$ --→ | 0.105 | 0.007 | 0.070 | 0.077 | 0.078 |
| 0.274 | 0.062 | 0.220 | 0.235 | 0.227 | ←-- Risk --→ | 0.251 | 0.030 | 0.198 | 0.220 | 0.200 |
| 2.310 | 1.213 | 1.465 | 1.635 | 1.689 | ←-- mse$f(t)$ --→ | 2.230 | 0.922 | 0.383 | 1.444 | 1.121 |
| LASSO | $\sigma_u = 1$ | | | | | ALASSO | $\sigma_u = 1$ | | | |
| $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ | | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathrm{R}\lambda}$ | $\hat{\beta}_n^{\mathrm{P}\lambda}$ | $\hat{\beta}_n^{\mathrm{S}\lambda}$ | $\hat{\beta}_n^{+\mathrm{S}\lambda}$ |
| 3.186 | 3.000 | 3.103 | 3.131 | 3.135 | | 3.213 | 3.000 | 3.126 | 3.154 | 3.157 |
| 1.542 | 1.537 | 1.520 | 1.534 | 1.535 | | 1.557 | 1.537 | 1.533 | 1.547 | 1.547 |
| −0.008 | −0.006 | −0.005 | −0.008 | −0.007 | | −0.006 | −0.004 | −0.004 | −0.006 | −0.005 |
| 0.015 | 0.012 | 0.015 | 0.014 | 0.014 | | 0.014 | 0.010 | 0.013 | 0.012 | 0.013 |
| 2.094 | 2.075 | 2.077 | 2.081 | 2.083 | | 2.119 | 2.075 | 2.098 | 2.101 | 2.103 |
| −0.000 | −0.003 | −0.004 | −0.003 | −0.002 | | −0.002 | −0.005 | −0.005 | −0.005 | −0.004 |
| −0.015 | −0.012 | −0.015 | −0.015 | −0.015 | | −0.012 | −0.009 | −0.012 | −0.012 | −0.012 |
| 0.004 | 0.009 | 0.008 | 0.006 | 0.006 | | 0.004 | 0.009 | 0.007 | 0.006 | 0.006 |
| 0.046 | 0.007 | 0.018 | 0.025 | 0.027 | ←-- bias$^2$ --→ | 0.063 | 0.007 | 0.027 | 0.036 | 0.038 |
| 0.914 | 0.320 | 0.688 | 0.709 | 0.704 | ←-- Risk --→ | 0.847 | 0.232 | 0.621 | 0.654 | 0.636 |
| 10.678 | 3.344 | 7.317 | 8.706 | 7.373 | ←-- mse$f(t)$ --→ | 9.660 | 2.577 | 7.157 | 7.739 | 7.356 |

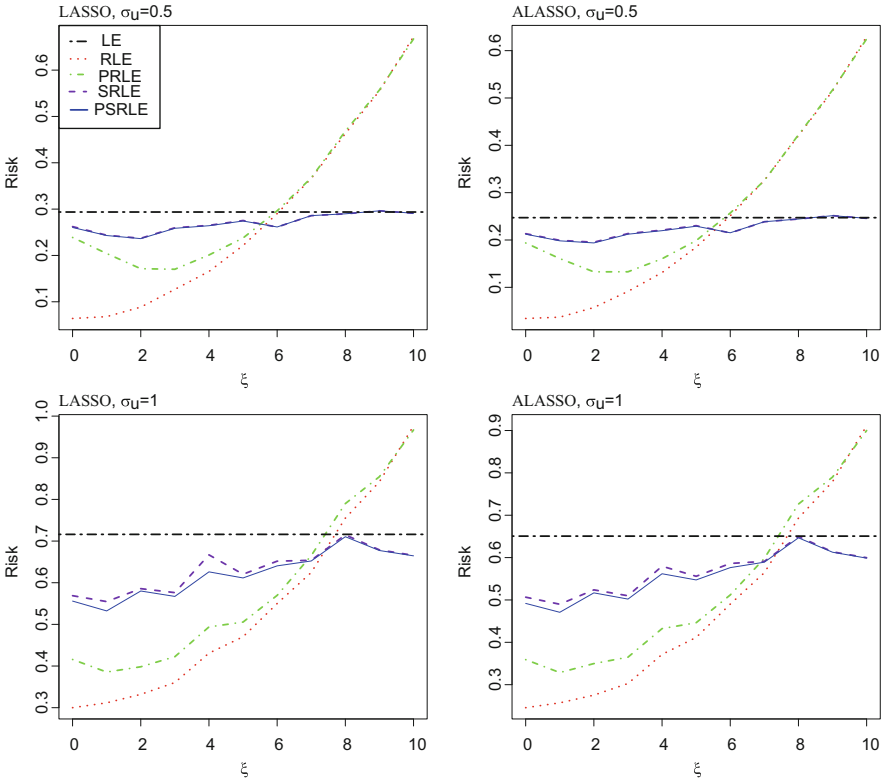



**Fig. 12.1** Risk of proposed estimators versus $\xi$ with $\alpha = 0.05$. Left panel: LASSO-type estimators. Right panel: ALASSO-type estimators

are shown in Figs. 12.1, 12.2, 12.3. The $\xi$ value ranges from 0 to 10. The risks of the LASSO and ALASSO estimators have a similar pattern. The risk of the unrestricted LASSO estimators is constant along $\xi$'s. When the values of $\xi$ are close to 0, the RLE and PLE achieve the smallest risk, but their risks increase quickly for large $\xi$. For small $\xi$ values, the risks of SLE and PSLE are higher than that of the RE and PLE. However, when $\xi$ moves from zero, say under $\mathcal{H}_n$, SLE and PSLE begin to dominate all other estimators. The nonparametric term of the model, corresponding to each estimate obtained from Tables 12.1 and 12.2, is illustrated in Fig. 12.4.

### *12.5.2 Example 2*

Again, we designed a simulation to study the relative efficiency with respect to various levels of sparsity. The data generation setup is the same as Example 1 except for $\beta$, which is modified to accommodate the number of nonzero $\beta$'s in the model

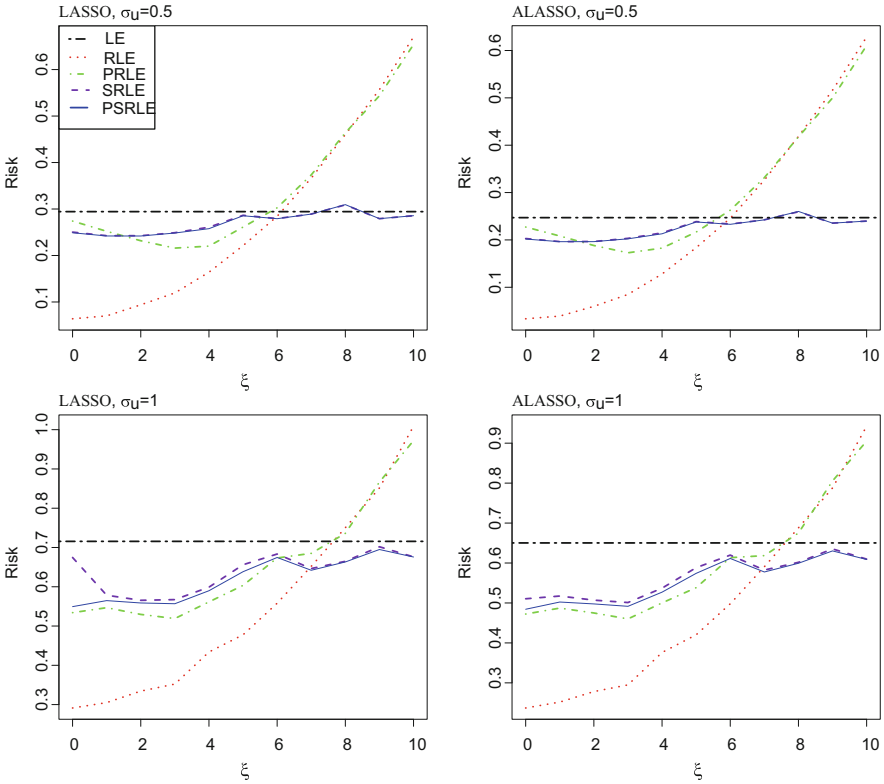



**Fig. 12.2** Risk of proposed estimators versus $\xi$ with $\alpha = 0.15$. Left panel: LASSO-type estimators. Right panel: ALASSO-type estimators

and considered as

$$\beta = (\underbrace{1, \ldots, 1}_{p-p_1}, \underbrace{0, \ldots, 0}_{p_1})^\top$$

where $p = 40$ and $p_1 = 10, 20$ and 30. For the linear restriction, we consider $\mathbf{R} = [\mathbf{0}_{p_1 \times (p-p_1)}, \mathbf{I}_{p_1}]$, $r = \mathbf{R}\beta$ and $\xi \in [0, 10]$. In order to have an easy comparison, at the $\alpha = 0.10$ level of significance, we calculate the relative risk efficiency of $\beta^*$ to the $\beta^\lambda$, which is given by

$$\text{RRE}(\beta^\lambda, \beta^*) = \frac{R(\beta^\lambda, \beta)}{R(\beta^*, \beta)}, \tag{12.26}$$

where $\beta^*$ is one of the restricted LASSO (or ALASSO) or restricted shrinkage LASSO (or ALASSO) estimators, which under $\mathcal{H}_0$ leads to a sparse model.

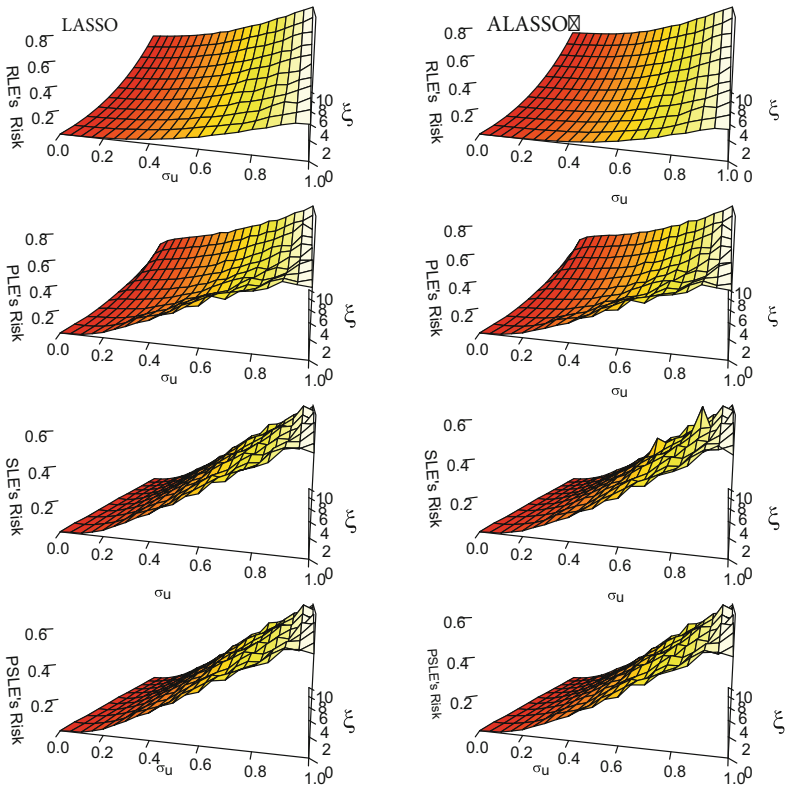



**Fig. 12.3** Response surface illustrating the risk of estimators against $\sigma_u$ and $\xi$. Left panel: LASSO-type estimators. Right panel: ALASSO-type estimators

By performing with 1000 replications, the results of Monte Carlo simulation are plotted in Fig. 12.5. It appears that at $\xi = 0$, say under $\mathcal{H}_0$, the ordinary LASSO (or ALASSO) estimator $\beta^{\lambda}$ performs much worse than the other estimators. On the other hand, when we move away from $\mathcal{H}_0$ by increasing $\xi$, the efficiency of RLE and PLE decreases sharply, specially for large $p_1$, and tends to the zero, while for all values of $p_1$, the efficiency of SLE and PSLE decreases slowly and converges to 1. Overall, for enough large $p_1$, the PSLE is better than other estimators for all values of $\xi$ in the sense of risk function.

## 12.6 Real Data

This section considers an application of the proposed shrinkage LASSO estimators to the Egyptian pottery data set. Smith et al. (1987) collected this data set from an archaeological survey of pottery production and distribution in the ancient Egyptian

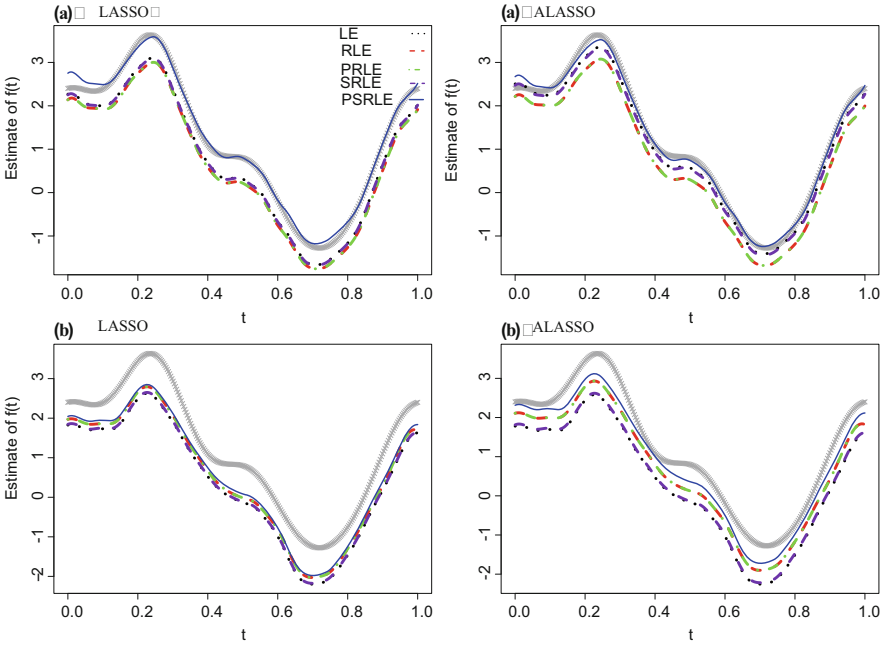



**Fig. 12.4** Estimation of nonparametric function (line with crosses is true values). (a): $\sigma_u = 0.5$ and (b): $\sigma_u = 1$

city of Al-Amarna. They applied two different methods, neutron activation analysis (NAA) and inductively coupled plasma spectrometry (ICP), for measuring the chemical contents of pottery samples. The set of pottery was collected from 28 different locations around the city. There are two main sources, namely, fabric code and location of pottery, for the group structure among the objects that based the group structure the selected vessels have been divided into 28 categories. For each category, there are various numbers of vessels from the same fabric code and provenance, which can essentially be regarded as replicated observations. In our model framework, this data consists of Na measured with NAA as response variable and Na ($x_1$), Al ($x_2$), K ($x_3$), Cr ($x_4$), and Mn ($x_5$) measured by ICP as continuous explanatory variables and location variable at 28 different locations (denoted by $x_i$, $i = 6, 7, \ldots, 33$) as a dummy explanatory variable. Using the mentioned continuous variables, a ridge linear measurement error model is fitted by Rasekh and Fieller (2003) to this data set to study the local influence analysis, and some restricted linear measurement error models based on the Liu estimation technique have been recently fitted by Ghapani and Babadi (2018) and Siray (2019). Here, we first standardized the variables and then fit a restricted shrinkage LASSO-type linear measurement error model to the data when the continuous variables are contaminated with measurement error. Among all continuous variables, “VICP" had minimum $s^2$ value, and so, it can be considered as a nonparametric part. From

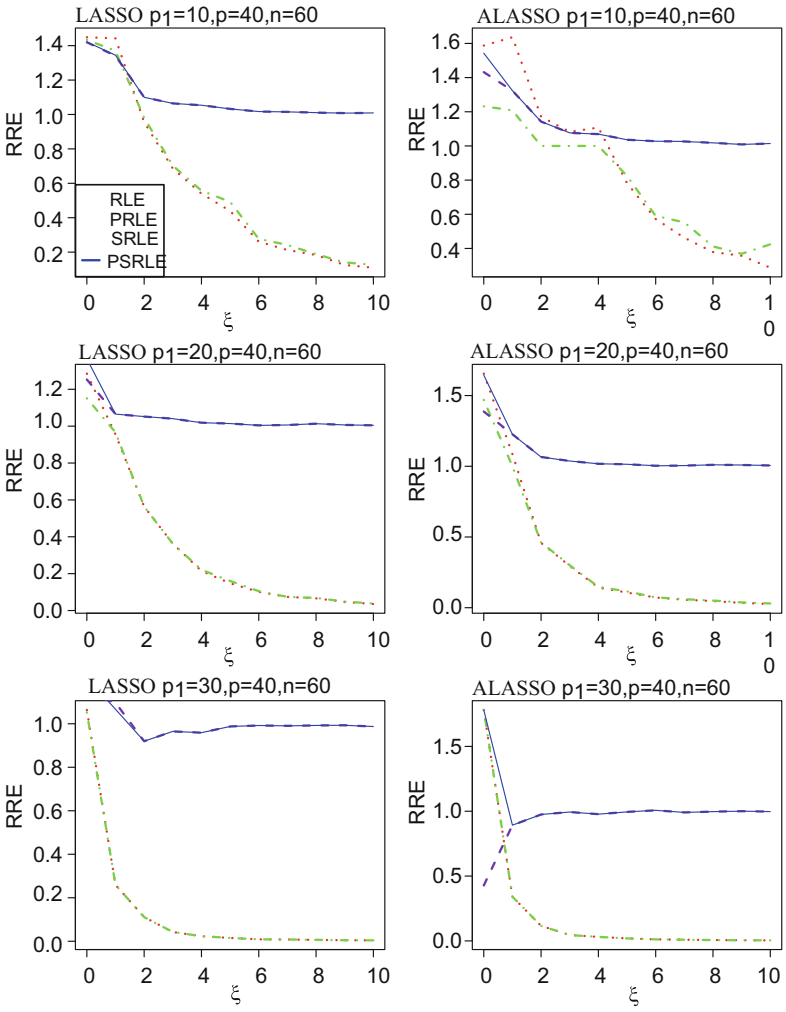



**Fig. 12.5** RRE against $\xi$. Left panel: LASSO-type estimators. Right panel: ALASSO-type estimators (Example 2)

previous researches, we consider the situation where restrictions in the form of that levels of location have the same linear effect on Na. Then, using advanced statistical methods for getting a reasonable estimate but weighting toward $\beta_6 = \beta_7 = \cdots = \beta_{33}$ makes some sense for the Egyptian pottery data set. Thus, we used the following specifications:

$$\mathbf{R}_0 = \begin{pmatrix} 0\,0\,0\,0\,0\,1\,-1\,\;0\;\ldots\,0 \\ 0\,0\,0\,0\,0\,0\;\;1\,-1\ldots\,0 \\ \vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\vdots\;\;\vdots\;\;\vdots\;\;\vdots \\ 0\,0\,0\,0\,0\,0\;\;0\;\;0\;\;\ldots\,1 \end{pmatrix}_{27\times 33}, \qquad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{12.27}$$

In Table 12.3, we reported LASSO and ALASSO estimators $\hat{\beta}_n^{\lambda}$, $\hat{\beta}_n^{\mathsf{R}\lambda}$, $\hat{\beta}_n^{\mathsf{P}\lambda}$, $\hat{\beta}_n^{\mathsf{S}\lambda}$, and $\hat{\beta}_n^{+\mathsf{S}\lambda}$ of our regression parameters based on 1000 bootstrap repetitions. We adopt the same methods of choosing the bandwidth and tuning parameter $\lambda$ as per the simulation exercises of Sect. 12.4. The solution paths of the LASSO and ALASSO based on the modified LARS algorithm are plotted in Fig. 12.6, where the various paths are labeled by the numbers corresponding to the $x_i$'s, $i = 1, \ldots, n$ and the vertical dotted line corresponds to the tuning parameters $\lambda_{\text{LASSO}} = 0.165$ and $\lambda_{\text{ALASSO}} = 0.122$ selected by 10-fold cross-validation for LASSO and ALASSO estimators, respectively. Based the LASSO and ALASSO estimations, the values of test statistics are obtained 27.738 and 14.215, respectively, and hence $\mathcal{H}_0$ is not rejected at 5% level of significance. The coefficient estimates based on tuning parameters are given in Table 12.3. Significant coefficients are highlighted. The maximum standard error value for bolded values is 0.015, while the minimum standard error for non-bolded values is 0.080. The results show that $x_1$ and $x_2$ and $x_{32}$ are the only covariates selected by LASSO and ALASSO. For estimation of the nonparametric component in Fig. 12.7, at first, we estimated the regression parameters by one of the proposed methods, and then, local polynomial approach was applied to fit $e_i = y_i - \mathbf{W}_i\hat{\beta}$ on $t_i, i = 1, \ldots, n$

In addition to variable selection, we also evaluate the performance of the estimators using the average tenfold cross-validation error. The $K$-fold cross-validation is a technique used to evaluate the performance of a machine learning model, particularly to guard against over-fitting. By partitioning the data into $K$ subsets and using each subset as a validation set while training on the remaining data, it helps estimate the model's effectiveness on independent data. This method mitigates over-fitting and provides a more accurate measure of model performance. Here, prediction errors are calculated as the squared differences between the observed values from the validation fold and the predicted values of the response variable from the training folds. By selecting a sample size of 1000 for the bootstrap simulation scheme, Table 12.4 presents the mean and standard deviation of the prediction errors. This table show that PSLE is the most efficient estimator in terms of reducing prediction errors. It is closely trailed by SLE. Additionally, raising the significance level for constructing PLE results in a reduction in prediction errors.

**Table 12.3** Results of real data example

| | LASSO | | | | | ALASSO | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathsf{R}\lambda}$ | $\hat{\beta}_n^{\mathsf{P}\lambda}$ | $\hat{\beta}_n^{\mathsf{S}\lambda}$ | $\hat{\beta}_n^{+\mathsf{S}\lambda}$ | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{\mathsf{R}\lambda}$ | $\hat{\beta}_n^{\mathsf{P}\lambda}$ | $\hat{\beta}_n^{\mathsf{S}\lambda}$ | $\hat{\beta}_n^{+\mathsf{S}\lambda}$ |
| $x_1$ | **0.184** | **0.277** | **0.277** | **0.267** | **0.267** | **0.216** | **0.272** | **0.272** | **0.312** | **0.322** |
| $x_2$ | **−0.182** | **−0.157** | **−0.157** | **−0.160** | **−0.160** | **−0.110** | **−0.124** | **−0.124** | **−0.134** | **−0.131** |
| $x_3$ | 0.001 | 0.022 | 0.022 | 0.020 | 0.020 | 0.000 | 0.028 | 0.028 | 0.049 | 0.029 |
| $x_4$ | 0.062 | 0.098 | 0.098 | 0.094 | 0.094 | 0.018 | 0.069 | 0.069 | 0.106 | 0.071 |
| $x_5$ | 0.129 | 0.037 | 0.037 | 0.047 | 0.047 | 0.070 | 0.019 | 0.019 | −0.019 | 0.019 |
| $x_6$ | 0.000 | −0.013 | −0.013 | −0.011 | −0.011 | −0.000 | −0.011 | −0.011 | −0.015 | −0.010 |
| $x_7$ | 0.019 | −0.013 | −0.013 | −0.009 | −0.009 | 0.000 | −0.011 | −0.011 | −0.019 | −0.010 |
| $x_8$ | 0.015 | −0.013 | −0.013 | −0.009 | −0.009 | 0.000 | −0.011 | −0.011 | −0.020 | −0.011 |
| $x_9$ | −0.032 | −0.013 | −0.013 | −0.015 | −0.015 | −0.012 | −0.011 | −0.011 | −0.010 | −0.010 |
| $x_{10}$ | 0.000 | −0.013 | −0.013 | −0.011 | −0.011 | 0.000 | −0.011 | −0.011 | −0.019 | −0.010 |
| $x_{11}$ | −0.016 | −0.013 | −0.013 | −0.013 | −0.013 | −0.011 | −0.011 | −0.011 | −0.010 | −0.011 |
| $x_{12}$ | −0.026 | −0.013 | −0.013 | −0.014 | −0.014 | −0.034 | −0.011 | −0.011 | 0.006 | −0.011 |
| $x_{13}$ | −0.036 | −0.013 | −0.013 | −0.015 | −0.015 | −0.023 | −0.011 | −0.011 | −0.002 | −0.010 |
| $x_{14}$ | 0.007 | −0.013 | −0.013 | −0.010 | −0.010 | 0.000 | −0.011 | −0.011 | −0.019 | −0.010 |
| $x_{15}$ | −0.061 | −0.013 | −0.013 | −0.018 | −0.018 | −0.050 | −0.011 | −0.011 | 0.018 | −0.009 |
| $x_{16}$ | −0.015 | −0.013 | −0.013 | −0.013 | −0.013 | −0.000 | −0.011 | −0.011 | −0.013 | −0.010 |
| $x_{17}$ | −0.049 | −0.013 | −0.013 | −0.017 | −0.017 | −0.034 | −0.011 | −0.011 | 0.006 | −0.011 |
| $x_{18}$ | 0.018 | −0.013 | −0.013 | −0.009 | −0.009 | 0.000 | −0.011 | −0.011 | −0.019 | −0.011 |
| $x_{19}$ | −0.089 | −0.013 | −0.013 | −0.021 | −0.021 | −0.070 | −0.011 | −0.011 | 0.032 | −0.011 |
| $x_{20}$ | 0.005 | −0.013 | −0.013 | −0.011 | −0.011 | 0.000 | −0.011 | −0.011 | −0.019 | −0.010 |
| $x_{21}$ | 0.065 | −0.013 | −0.013 | −0.004 | −0.004 | 0.000 | −0.011 | −0.011 | −0.025 | −0.010 |
| $x_{22}$ | −0.026 | −0.013 | −0.013 | −0.014 | −0.014 | −0.016 | −0.011 | −0.011 | −0.007 | −0.011 |
| $x_{23}$ | −0.059 | −0.013 | −0.013 | −0.018 | −0.018 | −0.039 | −0.011 | −0.011 | 0.010 | −0.011 |
| $x_{24}$ | 0.046 | −0.013 | −0.013 | −0.006 | −0.006 | 0.029 | −0.011 | −0.011 | −0.040 | −0.011 |
| $x_{25}$ | −0.000 | −0.013 | −0.013 | −0.011 | −0.011 | 0.000 | −0.011 | −0.011 | −0.019 | −0.009 |
| $x_{26}$ | 0.025 | −0.013 | −0.013 | −0.008 | −0.008 | 0.000 | −0.011 | −0.011 | −0.019 | −0.011 |
| $x_{27}$ | −0.029 | −0.013 | −0.013 | −0.014 | −0.014 | −0.016 | −0.011 | −0.011 | −0.007 | −0.011 |
| $x_{28}$ | 0.027 | −0.013 | −0.013 | −0.008 | −0.008 | 0.000 | −0.011 | −0.011 | −0.019 | −0.011 |
| $x_{29}$ | −0.060 | −0.013 | −0.013 | −0.018 | −0.018 | −0.056 | −0.011 | −0.011 | 0.022 | −0.012 |
| $x_{30}$ | 0.011 | −0.013 | −0.013 | −0.010 | −0.010 | 0.000 | −0.011 | −0.011 | −0.019 | −0.011 |
| $x_{31}$ | −0.038 | −0.013 | −0.013 | −0.015 | −0.015 | −0.043 | −0.011 | −0.011 | 0.012 | −0.011 |
| $x_{32}$ | **−0.101** | −0.013 | −0.013 | −0.023 | −0.023 | **−0.100** | −0.011 | −0.011 | 0.054 | −0.019 |
| $x_{33}$ | −0.064 | −0.013 | −0.013 | −0.018 | −0.018 | −0.053 | −0.011 | −0.011 | 0.020 | −0.011 |
| RSS | 2.089 | 0.973 | 0.973 | 0.886 | 0.738 | 0.973 | 0.925 | 0.925 | 0.647 | 0.631 |

## 12.7 Discussion

This chapter presents original improvements to the LASSO estimator by incorporating a model constraint in conjunction with preliminary testing and shrinkage techniques. Specifically, we introduce three innovative estimators designed for situ-

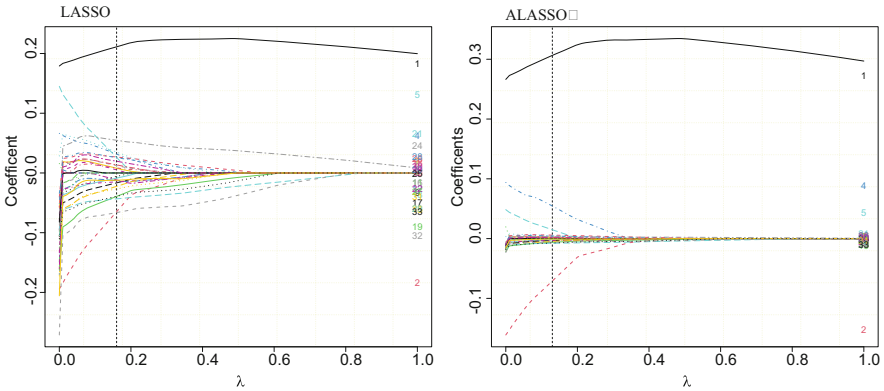



**Fig. 12.6** Solution paths of linear coefficients for Egyptian pottery data

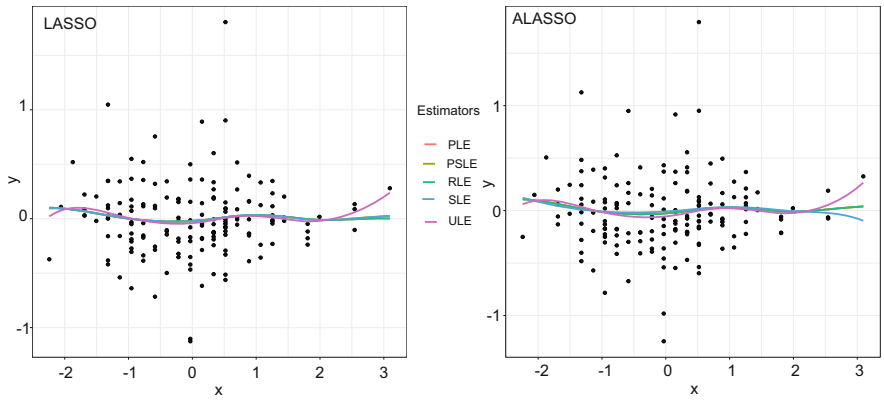



**Fig. 12.7** Estimation of nonparametric function

**Table 12.4** 10-fold cross validation average prediction errors and standard deviations for Egyptian pottery data

| Estimator type | | $\hat{\beta}_n^{\lambda}$ | $\hat{\beta}_n^{R\lambda}$ | $\hat{\beta}_n^{P\lambda}$ | $\hat{\beta}_n^{S\lambda}$ | $\hat{\beta}_n^{+S\lambda}$ |
|---|---|---|---|---|---|---|
| LASSO | mean | 9322 | 8331 | 10,157 | 8366 | 8352 |
| | sd | 310.49 | 275.13 | 473.99 | 301.60 | 299.73 |
| ALASSO | mean | 9136 | 8227 | 9725 | 8289 | 8239 |
| | sd | 307.81 | 255.64 | 437.02 | 270.56 | 261.47 |

ations involving a subspace restriction: the preliminary test LASSO and ALASSO (PLE), the Stein-type shrinkage LASSO and ALASSO (SLE), and the positive-rule shrinkage LASSO and ALASSO (PSLE). It is worth noting that this is the first application of shrinkage LASSO (or ALASSO) preliminary test and Stein-type estimators to models affected by measurement errors in the covariates. Additionally, we develop an algorithm inspired by LARS for determining the solution path

of the LASSO and ALASSO shrinkage estimators in practical applications. Our evaluation focuses on assessing the performance of these estimators under the null hypothesis in scenarios where the sample size ($n$) surpasses the number of features ($p$). The proposed methodology for enhancing the LASSO demonstrates its applicability in high-dimensional settings characterized by a large $p$ and a small $n$. In addition to the aforementioned theorems regarding the asymptotic behavior of the proposed estimators, we conducted a simulation study to numerically evaluate their performance across various configurations. These configurations consisted of different values of the parameter values ($p$) and variance of measurement errors. To facilitate a comparison between the improved estimators and the classical LASSO, we used the relative efficiency criterion. The performance of the estimators was evaluated for different non-centrality parameters $\xi$, levels of model misspecification, and numbers of nonzero values. Overall, the PSLE showed the best performance. When deviating from the null model, neither the PLE nor the SLE demonstrated clear superiority over the other; however, the PLE exhibited better performance as its value increased. The relative efficiency of the proposed estimators improved as more near-zero parameters were included in the model. As the null model was deviated from, the performance of the estimators worsened. As a final observation, it is evident from the figures and tables that the proposed shrinkage estimators of type ALASSO exhibit significantly superior performance in terms of risk reduction when compared to their type LASSO counterparts.

**Acknowledgments** We would like to express our sincere gratitude to the editors and reviewers for their constructive feedback and insightful comments, which significantly enhanced the quality of our manuscript.

## References

Ahmed, S. E., Doksum, K. A., Hossain, S., & You, J. (2007). Shrinkage, pretest and absolute penalty estimators in partially linear models. *Australian and New Zealand Journal of Statistics*, *49*(4), 435–454.

Ali, A., & Tibshirani, R. (2019). The generalized LASSO problem and uniqueness. *Electronic Journal of Statistics*, *13*(2), 2307–2347.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, *10*, 1100–1120.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman and Hall.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Ghapani, F., & Babadi, B. (2018). Mixed Liu estimator in linear measurement error models. *Communications in Statistics - Theory and Methods*, *47*, 1961–1970.

Harezlak, J., Ruppert, D., & Wand, M. P. (2019). *Semiparametric regression with R*. Springer.

Härdle, W. K., Liang, H., & Gao, J. (2000). *Partially linear models*. Physika Verlag.

Horowitz, J. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematics and Statistical Probability* (Vol. 1, pp. 361–379).

Judge, G. G., & Bock, M. E. (1978). *The statistical implications of pretest and Stein rule estimators in econometrics*. North-Holland Publications.

Kim, H. M., & Saleh, A. K. Md. E. (2005). Improved estimation of regression parameters in measurement error models. *Journal of Multivariate Analysis*, *95*, 273–300.

Knight, K., & Fu, W. (2000). Asymptotics for LASSO-type estimators. *Annals of Statistics*, *28*(5), 1356–1378.

Kibria, B. M., & Saleh, A. K. Md. E. (2012). Improving the estimators of the parameters of a probit regression model: A ridge regression approach. *Journal of Statistical Planning and Inference*, *142*(6), 1421–1435.

Li, M., Li, R., & Ma, Y. (2021). Inference in high-dimensional linear measurement error models. *Journal of Multivariate Analysis*, *184*, 104759.

Liang, H., Härdle, W., & Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics*, *27*, 1519–1535.

Norouzirad, M., & Arashi, M. (2018). Preliminary test and Stein-type shrinkage LASSO-based estimators. *SORT-Statistics and Operations Research Transactions*, *42*(1), 45–58.

Rao, C. R., & Toutenburg, H. (2008). *Linear models and generalizations*. Springer.

Rasekh, A. R., & Fieller, N. R. J. (2003). Influence functions in functional measurement error models with replicated data. *Statistics*, *37*, 169–178.

Roozbeh, M., Hamzah, N., & Mohamed, N. (2020). Using improved robust estimators to semiparametric models with high-dimensional data. In *Computational and Methodological Statistics and Biostatistics* (pp. 269–290).

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, *90*, 1257–1270.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression models*. Cambridge University Press.

Saleh, A. K. Md. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. Wiley.

Siray, Y. (2019). Liu estimation approach to the measurement error models. *Journal of Statistical Computation and Simulation*, *89*, 3453–3496.

Smith, D. M., Hart, F. A., & Symond, R. D. (1987). Analysis of Roman pottery from Colchester by inductively coupled plasma spectrometry. *Science and Archaeology Glasgow*, *196*, 41–55.

Sørensen, F., Frigessi, A., & Thoresen, M. (2015). Measurement error in LASSO: Impact and likelihood bias correction. *Statistica Sinica*, *25*, 809–829.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, *50*, 413–436.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematics and Statistical Probability* (Vol. 1, pp. 197–206).

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, *58*(1), 267–288.

Toutenburg, H. (1982). *Prior information in linear models*. Wiley.

Wang, H., Zou, G., & Wan, A. (2013). Adaptive LASSO for varying-coefficient partially linear measurement error models. *Journal of Statistical Planning and Inference*, *143*, 40–54.

Yüzbaşı, B., Arashi, M., & Ahmed, S. E. (2019). Shrinkage estimation strategies in generalized ridge regression models under low/high-dimension regime. *International Statistical Review*, *87*(3), 434–456.

Zhang, H., & Lu, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika*, *94*, 691–703.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

# Chapter 13
# Weighted-Average Least Squares Estimation of Panel Data Models




**Giuseppe De Luca and Jan R. Magnus**

## 13.1 Introduction

In a typical statistical project, the author considers a family of possible models, then plays a game of trial and error until one of these models comes out on top according to some criterion, and then estimates the preferred model. This, of course, is a highly dubious course of action and Saleh (2006) was one of the first to warn against the dour consequences of such procedures.

To understand what is wrong, we must recognize that two actions take place, namely, model selection and estimation, and that we should take the noise caused by the model selection procedure (pretesting) into account when reporting estimates and standard errors of the parameters in the preferred model. This, however, is seldom done, partly because it is difficult to do so; (see, e.g., Leeb & Pötscher, 2005; McCloskey, 2017).

The rapidly developing field of model averaging does make a serious attempt to combine model selection and estimation. At this time, there are three main branches of model averaging: Bayesian model averaging (BMA), frequentist model averaging (FMA), and weighted-average least squares (WALS). Useful overviews of these approaches can be found in Claeskens and Hjort (2008), Moral-Benito (2015), Magnus and De Luca (2016), and Steel (2020). Our contribution is based on the

G. De Luca
Department of Economics, Business and Statistics (SEAS), University of Palermo, Palermo, Italy
e-mail: giuseppe.deluca@unipa.it

J. R. Magnus (✉)
Department of Econometrics & Data Science, Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands
e-mail: jan@janmagnus.nl

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_13

WALS approach of Magnus et al. (2010), which was originally developed in the simplest possible framework, but has since been extended in many directions.

The extension studied in this chapter is to linear and static panel data models. A panel is a two-dimensional data set, where we have observations $y_{it}$ on units (individuals, firms, states, or other units) over a period of time. The units are indexed by $i = 1, \ldots, N$ and the time periods by $t = 1, \ldots, T$. In practice, we may not have observations on all units over all time periods (an unbalanced panel), and the periods may not be equally spaced over time. Our approach allows for both extensions, but we will not make this explicit in the notation. However, the proposed adjustments rest on the assumption of an ignorable missing-data mechanism (Verbeek & Nijman, 1996).

This chapter extends the WALS estimator to fixed-effects and random-effects panel data models with strictly exogenous regressors. In the fixed-effects model, we show that the WALS approach satisfies the Frisch–Waugh–Lovell theorem when partialling out an arbitrary subset of focus regressors from the unrestricted model. This result implies that, when the individual effects are treated as additional (nuisance) focus parameters, the fixed-effects WALS estimator can be easily constructed by means of the usual within transformations.

The random-effects WALS estimator builds on a special case of the feasible generalized least squares (FGLS) strategy proposed by Magnus et al. (2011) to perform WALS estimation of linear models with nonspherical errors. In this special case, we first estimate the variance components from the unrestricted model and then perform a WALS regression based on the FGLS transformations of the original data to account for the stable equicorrelation exhibited by the one-way errors of the same unit over time.

In addition to the basic setup with independent and identically distributed (i.i.d.) errors, we also use the FGLS transformations proposed by Bhargava et al. (1982) and Baltagi and Wu (1999) to analyze an extended setup of the fixed-effects and random-effects models where errors are allowed to follow a stationary AR(1) process and observations are allowed to be unequally spaced over time.

This chapter contributes to the rapidly growing literature on model-averaging methods for panel data models. BMA techniques for dynamic panel data models were developed by Moral-Benito (2012, 2013, 2015), Leon-Gonzalez and Montolio (2015), Mirestean and Tsangarides (2016), Chen et al. (2018), and Desbordes et al. (2018), among others. Similarly, FMA estimators for specialized panel data models have been proposed by Zhang et al. (2014), Gao et al. (2016), Wang et al. (2019), Yin et al. (2021), and Jiang et al. (2025). Our generalizations of the WALS approach do not extend to dynamic models, semi-parametric specifications, high-dimensional frameworks, or heteroskedastic and multifactor error structures for longitudinal data. However, the fixed-effects and random-effects models analyzed in this chapter remain the workhorse of applied econometrics. Furthermore, the WALS approach offers important theoretical and computational advantages over other model-averaging methods. Specifically, unlike FMA, the WALS weighting scheme is derived from a Bayesian perspective, ensuring desirable theoretical properties such as admissibility, bounded risk, robustness, and near optimality

in terms of minimax regret. In contrast to BMA, WALS employs a nonstandard Bayesian approach that places the prior on a scaled version of the model parameters, interpreted as "population $t$-ratios." This allows for a proper treatment of ignorance in the choice of the prior and guarantees other desirable sampling properties such as uniform $\sqrt{n}$-consistency. Additionally, WALS differs from both BMA and FMA in that it exploits a preliminary semi-orthogonal transformation of the regressors, which significantly reduces the computational burden associated with model-averaging estimation. This computational advantage enables us to explore a large model space without applying sophisticated MCMC methods or preliminary screening procedures.

This chapter is organized as follows: In Sect. 13.2, we summarize the WALS approach for cross-sectional data models with i.i.d. errors, and in Sect. 13.3, we extend this approach to nonspherical errors. In Sect. 13.4, we define the class of panel data models that we wish to study. In Sects. 13.5 and 13.6, we obtain the fixed-effects and random-effects WALS estimators for the case of i.i.d. errors, and in Sect. 13.7, we discuss the extended setup with AR(1) errors. In Sect. 13.8, we present an empirical application of the fixed-effects WALS estimator to analyze the impact of legalized abortion on crime using data from Donohue and Levitt (2020). The conclusion is in Sect. 13.9.

## 13.2 The WALS Approach

The basic framework in WALS is the homoskedastic linear model:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{13.1}$$

where $X_1$ and $X_2$ are nonstochastic $n \times k_1$ and $n \times k_2$ matrices such that $X = (X_1 : X_2)$ has full column-rank $k = k_1 + k_2 < n$, and $\epsilon$ is an $n \times 1$ vector of independent errors, identically distributed as $\mathcal{N}(0, \sigma_\epsilon^2 I_n)$, where $I_n$ denotes the identity matrix of order $n$. The two sets of regressors $X_1$ and $X_2$ play a different role: $X_1$ contains the $k_1$ "focus" regressors, which we want in the model on theoretical or other grounds, while $X_2$ contains the $k_2$ "auxiliary" regressors of which we are less certain. The parameters in $\beta_1$ and $\beta_2$ are called the focus and auxiliary parameters, respectively. The focus regressors are typically variables of which we wish to study the causal effects on $y$ or whose presence in the model is well-established from previous studies, while the auxiliary regressors are typically controls that are added to avoid omitted-variable bias, or transformations of and interactions between the regressors. We assume that $k_1 \geq 0$ and $k_2 \geq 1$. The case $k_2 = 0$ is excluded because it implies that there are no auxiliary regressors and hence no model uncertainty. The normality assumption is needed to obtain the finite-sample distribution of our estimator, but it can be relaxed in the asymptotic theory.

Our first step is to apply the following one-to-one transformations of $X_2$ and $\beta_2$:

$$Z_2 = X_2\Delta_2\Psi^{-1/2}, \qquad \gamma_2 = \Psi^{1/2}\Delta_2^{-1}\beta_2, \tag{13.2}$$

where $\Delta_2$ is a diagonal $k_2 \times k_2$ matrix such that the diagonal elements of the positive definite matrix $\Psi = \Delta_2 X_2^\top M_1 X_2 \Delta_2/n$ are all equal to one, $\Psi^{1/2}$ is the unique square root of $\Psi$, and $M_1 = I_n - X_1(X_1^\top X_1)^{-1}X_1^\top$. Next, we rescale $X_1$ and $\beta_1$ using

$$Z_1 = X_1\Delta_1, \qquad \gamma_1 = \Delta_1^{-1}\beta_1, \tag{13.3}$$

where $\Delta_1$ is a diagonal $k_1 \times k_1$ matrix such that the diagonal elements of $Z_1^\top Z_1/n$ are all equal to one. Since $Z_1\gamma_1 = X_1\beta_1$ and $Z_2\gamma_2 = X_2\beta_2$, we can rewrite the model defined in Eq. (13.1) as

$$y = Z_1\gamma_1 + Z_2\gamma_2 + \epsilon, \tag{13.4}$$

where $Z_2^\top M_1 Z_2/n = I_{k_2}$. The transformations in Eq. (13.2) ensure that the $k_2$ components of the least squares (LS) estimator of $\gamma_2$ in the model specified in Eq. (13.4) are independent, while the rescaling in Eq. (13.3) serves only to increase the numerical accuracy of the inversion and eigenvalue routines.

The WALS estimators of $\gamma_1$ and $\gamma_2$ are obtained by averaging the LS estimators $\hat{\gamma}_{1j}$ and $\hat{\gamma}_{2j}$ resulting from the $2^{k_2}$ models that contain all focus regressors and a subset of the auxiliary regressors in the model defined in Eq. (13.4):

$$\tilde{\gamma}_1 = \sum_{j=1}^{2^{k_2}} w_j\hat{\gamma}_{1j}, \qquad \tilde{\gamma}_2 = \sum_{j=1}^{2^{k_2}} w_j\hat{\gamma}_{2j}, \tag{13.5}$$

where $w_j$ are nonnegative model weights that depend only on $M_1 y$ and add up to one. Given $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$, the WALS estimators of $\beta_1$ and $\beta_2$ in model (13.1) are

$$\tilde{\beta}_1 = \Delta_1\tilde{\gamma}_1, \qquad \text{and} \qquad \tilde{\beta}_2 = \Delta_2\Psi^{-1/2}\tilde{\gamma}_2. \tag{13.6}$$

Additional insight is obtained by writing $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ as

$$\tilde{\gamma}_1 = \hat{\gamma}_{1r} - Q\tilde{\gamma}_2, \qquad \tilde{\gamma}_2 = \Lambda\hat{\gamma}_{2u}, \tag{13.7}$$

where $\hat{\gamma}_{1r} = (Z_1^\top Z_1)^{-1}Z_1^\top y$ is the LS estimator of $\gamma_1$ in the fully restricted model (with $\gamma_2 = 0$), $\hat{\gamma}_{2u} = Z_2^\top M_1 y/n$ is the LS estimator of $\gamma_2$ in the unrestricted model, $Q = (Z_1^\top Z_1)^{-1}Z_1^\top Z_2$, $\Lambda = \sum_j w_j W_j$, $W_j = I_{k_2} - S_j S_j^\top$ is a diagonal matrix whose diagonal elements are equal to zero or one, and $S_j$ is a $k_2 \times r_j$ selection matrix of rank $0 \le r_j \le k_2$ representing the $r_j$ exclusion restrictions implied by model $j$, that is, $S_j^\top = [I_{r_j} : 0]$ or a column-permutation thereof.

The "equivalence theorem" of Magnus and Durbin (1999) shows that the mean squared error (MSE) matrix of $\tilde{\gamma}_1$ depends only on the MSE matrix of $\tilde{\gamma}_2$, so attention can be restricted to the estimation of $\gamma_2$. The $k_2$ diagonal elements $\lambda_h$ of $\Lambda$, called the "WALS weights," are partial sums of the model weights $w_j$. Specifically, we have

$$\lambda_h = \sum_{j \in \mathcal{M}_h} w_j, \tag{13.8}$$

where $\mathcal{M}_h$ is the subspace of $2^{k_2-1}$ models that contain the $h$-th auxiliary regressor in Eq. (13.4). It follows that $0 \leq \lambda_h \leq 1$, and hence the $k_2$ components of $\tilde{\gamma}_2$ in Eq. (13.7) are shrinkage estimators of the components of $\gamma_2$. Like the posterior inclusion probabilities in the context of BMA, each $\lambda_h$ measures the importance of including the $h$th auxiliary regressor in the transformed model given by Eq. (13.4).

Since $Z_2^\top M_1 Z_2/n = I_{k_2}$, we know that the $k_2$ components $(\hat{\gamma}_{2u})_h$ of $\hat{\gamma}_{2u}$ are distributed independently as

$$(\hat{\gamma}_{2u})_h \sim \mathcal{N}\left(\gamma_{2h}, \sigma_\epsilon^2/n\right), \tag{13.9}$$

where $\gamma_{2h}$ denotes the $h$-th component of $\gamma_2$.[1] Thus, if we further restrict each $\lambda_h$ to depend only on $(\hat{\gamma}_{2u})_h$, then the components of $\tilde{\gamma}_2$ are also independent, and our $k_2$-dimensional estimation problem for $\gamma_2$ reduces to $k_2$ identical one-dimensional problems of the following type: given a single observation $x \sim \mathcal{N}\left(\theta, \sigma_\epsilon^2/n\right)$, find a "good" shrinkage estimator of the location parameter $\theta$. This stylized setting is known as the "normal location problem." Since risk properties of shrinkage estimators are little affected by the estimation of the variance parameter (see, e.g., Danilov, 2005), we assume that $\sigma_\epsilon^2$ is known. An equivalent representation of the normal location problem is $x^* \sim \mathcal{N}(\theta^*, 1)$, where, with $\sigma_\epsilon^2$ known, $x^* = \sqrt{n}x/\sigma_\epsilon$ is the $t$-ratio for testing the hypothesis $\theta = 0$ and $\theta^* = \sqrt{n}\theta/\sigma_\epsilon$ is the associated "population $t$-ratio."

An obvious estimator of $\theta$ in the normal location problem is $x$ itself, which is unbiased and consistent, and is sometimes called the "usual" estimator. Another estimator, sometimes called the "silly" estimator, is 0 (zero). Since $\text{MSE}(x) = \sigma_\epsilon^2/n$ and $\text{MSE}(0) = \theta^2$, we prefer the silly estimator if and only if $|\theta| < \sigma_\epsilon/\sqrt{n}$ or, equivalently, $|\theta^*| < 1$. No estimator dominates $x$ uniformly in the MSE sense. However, for $\theta$ sufficiently close to zero, shrinkage estimators of the form $m(x) = \lambda(x)x$ with $0 \leq \lambda(x) \leq 1$ may achieve substantial MSE improvements by trading some bias with a variance reduction. The closeness of $\theta$ to zero should be evaluated relative to the standard deviation $\sigma_\epsilon/\sqrt{n}$ of $x$. For the silly estimator, where $\lambda(x) = 0$, we obtain a lower MSE for $|\theta^*| < 1$. More generally, the region

[1] Notice that if we relax the normality of $\epsilon$ in model given by Eq. (13.1), then Eq. (13.9) holds asymptotically under the conditions of the central limit theorem.

of the parameter space where $\text{MSE}(m(x)) < \text{MSE}(x)$ depends on the population $t$-ratio $\theta^*$ and the shrinkage function $\lambda(x)$.

Based on theoretical considerations regarding admissibility, bounded risk, robustness, and optimality in terms of minimax regret, we adopt a Bayesian shrinkage approach that places a proper prior $\pi$ on the population $t$-ratio $\theta^*$. Under quadratic loss, the Bayes estimator of $\theta^*$ is the posterior mean $m^* = \text{E}(\theta^*|x^*)$, and the implied estimator of $\theta$ is $m = \sigma_\epsilon\, m^*/\sqrt{n}$. To ensure that $m$ enjoys attractive sampling properties as an estimator of $\theta$, we require that the prior $\pi$ on $\theta^*$ satisfies the following conditions:

(C1) $\pi$ is symmetric around zero.
(C2) $\pi$ is positive and nonincreasing on $(0, \infty)$.
(C3) $\pi$ is differentiable, except possibly at 0.
(C4) $\psi(\theta) = -\pi'(\theta)/\pi(\theta) \to \psi_0$ as $\theta \to \infty$, where $\psi_0 \geq 0$ is some finite constant.

As discussed in De Luca et al. (2025), (C1)–(C3) are mild regularity conditions, while (C4) ensures that the prior $\pi$ has bounded influence on $m^*$. If $\psi_0 = 0$, then (C4) implies the stronger property of "Bayesian robustness": $x^* - m^* \to 0$ as $x^* \to \infty$, which means that prior information gets essentially ignored when the data information is sufficiently strong.

Let $\Gamma(u)$ and $B(u, v)$ denote the gamma and beta functions, respectively. Examples of priors satisfying (C1)–(C4) are the class of reflected gamma-type priors:

$$\pi(\theta) = \frac{bc^{(1-a)/b}}{2\Gamma((1-a)/b)}|\theta|^{-a}\exp\left\{-c|\theta|^b\right\} \qquad 0 \leq a < 1,\ 0 < b \leq 1,\ c > 0, \tag{13.10}$$

the class of the reflected beta-type priors:

$$\pi(\theta) = \frac{c^{1/b}b}{2\,B(1/a - 1/b, 1/b)}(1 + c|\theta|^b)^{-1/a} \qquad 0 < a < b,\ c > 0, \tag{13.11}$$

the horseshoe prior (Carvalho et al., 2010):

$$\pi(\theta; c) = \frac{c\sqrt{2}}{\pi^{3/2}}\int_0^\infty \frac{\exp\{-t\}}{\theta^2 + 2c^2 t}\,\mathrm{d}t \qquad c > 0, \tag{13.12}$$

and the log prior, also known as horseshoe-like prior (Bhadra et al., 2017),

$$\pi(\theta) = \frac{1}{2\pi c}\ln(1 + c^2/\theta^2) \qquad c > 0. \tag{13.13}$$

As special cases of the class of reflected gamma-type priors defined in Eq. (13.10), we consider the Laplace ($a = 0, b = 1$), Weibull ($a + b = 1$), and Subbotin ($a = 0$) priors. The normal prior ($a = 0$, $b = 2$) is excluded because it violates (C4). As

special cases of the class of the reflected beta-type priors defined in Eq. (13.11), we consider the prior ($b = 1$) and the Student prior with $d$ degrees of freedom ($a = 2/(d+1)$, $b = 2$, $c = 1/d$). Our prior densities are also required to satisfy a notion of ignorance, which we call "neutrality," namely, not knowing whether the silly estimator of $\theta$ has a lower MSE than the usual estimator, that is, whether $|\theta^*| < 1$:

(C5) $\pi$ satisfies $\int_0^1 \pi(\theta)\,\mathrm{d}\theta = \int_1^\infty \pi(\theta)\,\mathrm{d}\theta = 1/4$.

Together with (C1), condition (C5) fixes one free prior parameter to ensure that the events $|\theta^*| < 1$ and $|\theta^*| > 1$ are equally likely a priori. The neutral Laplace and Weibull priors are obtained for $b = \ln(2)$, the neutral Student prior for $d = 1$ (the Cauchy distribution), and the neutral Pareto prior for $c = 2^{a/(1-a)} - 1$. Neutrality conditions for the Subbotin, horseshoe, and log priors are solved numerically. For the Subbotin, Weibull, and Pareto priors, we fix the other free prior parameter by the minimax regret criterion for $m^*$, where regret is defined as the difference between the MSE of $m^*$ and its lower bound (Magnus, 2002).

The Bayesian approach is used to construct our shrinkage estimators $m^*$ and $m$ of $\theta^*$ and $\theta$, but the properties of these estimators are then assessed in a classical frequentist setup. De Luca et al. (2022) studied the bias and variance of $m^*$ in repeated samples and proposed the plug-in maximum likelihood (ML) and double-shrinkage estimators of its sampling moments. More recently, De Luca et al. (2025) derived the finite-sample distribution of $m^* - \theta^*$ and showed that the finite-sample distribution of $m^* - \theta^*$ presents sizeable departures from normality, both in terms of skewness and of excess kurtosis. In the asymptotic theory, De Luca et al. (2025) also proved uniform $\sqrt{n}$-consistency of $m$ and obtained its asymptotic distribution. This estimator represents our estimator of the $h$-th component $\gamma_{2h}$ of $\gamma_2$ in the model defined in Eq. (13.4), and its sampling properties carry over, more or less straightforwardly, to the WALS estimators $\tilde{\gamma} = (\tilde{\gamma}_1^\top, \tilde{\gamma}_2^\top)^\top$ of $\gamma = (\gamma_1^\top, \gamma_2^\top)^\top$ given by Eq. (13.5) and the WALS estimators $\tilde{\beta} = (\tilde{\beta}_1^\top, \tilde{\beta}_2^\top)$ of $\beta = (\beta_1^\top, \beta_2^\top)^\top$ defined in Eq. (13.6). Specifically, under (C1)–(C4), the finite-sample distribution of the WALS estimator is generally nonnormal, and the choice of prior may have sizeable effects on the estimation bias.

In large samples, these estimators are uniformly $\sqrt{n}$-consistent, and their asymptotic distribution is normal only under certain conditions on $\beta_2$. The issue of inference has been addressed by De Luca et al. (2023), who proposed a simulation method for obtaining re-centered and asymmetric WALS confidence and prediction intervals based on the bias-corrected posterior mean. Although uniformly consistent estimators of the sampling distribution of shrinkage-type estimators may not exist (see Lemmas 3.1 and 3.5 in Leeb & Pötscher, 2006), the Monte Carlo simulations of De Luca et al. (2023) suggest that coverage errors of WALS intervals are small.

More detailed information on the WALS approach can be found in our 2016 survey (Magnus & De Luca, 2016) and our forthcoming monograph (De Luca & Magnus, 2025).

## 13.3 Nonspherical Errors

In the previous section, we provided a short summary of the WALS procedure in its simplest and most basic form. But WALS can be extended in many directions, and this chapter deals with the extension to panel data. Before we discuss panel data in more detail, we need to discuss how WALS deals with the situation where the errors are not identically and independently distributed as $\mathcal{N}(0, \sigma_\epsilon^2)$, but follow a more general distribution. Thus, we write

$$y = X_1\beta_1 + X_2\beta_2 + u, \tag{13.14}$$

where $u \sim \mathcal{N}(0, \Omega)$ and $\Omega$ is a positive definite (hence nonsingular) matrix. In the basic setup, we have $\Omega = \sigma_\epsilon^2 I_n$, and we estimate $\sigma_\epsilon^2$ from the unrestricted model, whereafter we consider it to be fixed (nonrandom).

We follow the same procedure in the more general case. Hence, we assume that $\Omega$ depends on a finite and fixed number of parameters $\alpha_1, \ldots, \alpha_p$, which we estimate from the unrestricted model, and then we transform

$$y^* = X_1^*\beta_1 + X_2^*\beta_2 + \epsilon, \tag{13.15}$$

where

$$y^* = \Omega^{-1/2}y, \quad X_1^* = \Omega^{-1/2}X_1, \quad X_2^* = \Omega^{-1/2}X_2, \quad \epsilon = \Omega^{-1/2}u. \tag{13.16}$$

Next, we apply the WALS approach described in Sect. 13.2 to the new model given by Eq. (13.15), and obtain estimates of $\beta_1$ and $\beta_2$.

## 13.4 Fixed and Random Effects

Our statistical panel data framework is

$$y_{it} = v_i + x_{it,1}^\top\beta_1 + x_{it,2}^\top\beta_2 + u_{it} \qquad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{13.17}$$

where $(y_{it}, x_{it,1}^\top, x_{it,2}^\top)$ are the observations on the outcome of interest $y_{it}$, the focus regressors $x_{it,1}$, and the auxiliary regressors $x_{it,2}$; and the $v_i$ are individual effects capturing unobserved and time-invariant heterogeneity. The total number of observations is denoted by $n = N\,T$.

We can write Eq. (13.17) in vector form by defining

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \quad X_{i,1} = \begin{pmatrix} x_{i1,1}^\top \\ \vdots \\ x_{iT,1}^\top \end{pmatrix}, \quad X_{i,2} = \begin{pmatrix} x_{i1,2}^\top \\ \vdots \\ x_{iT,2}^\top \end{pmatrix}, \quad u_i = \begin{pmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{pmatrix}, \tag{13.18}$$

and this leads to

$$y_i = \mathbf{1}_T \nu_i + X_{i,1}\beta_1 + X_{i,2}\beta_2 + u_i \qquad i = 1, \ldots, N, \tag{13.19}$$

where $\mathbf{1}_T$ denotes the $T \times 1$ vector of ones. Throughout we shall make the following assumption.

**Assumption 13.1** *The $T \times 1$ vectors $u_1, \ldots, u_N$ are i.i.d. error vectors with $\mathrm{E}[u_i] = 0$ and $\mathrm{var}(u_i) = \sigma_\epsilon^2 V_u$, where $\sigma_\epsilon^2$ is a positive scalar and $V_u$ is a $T \times T$ positive definite matrix.*

In Sects. 13.5 and 13.6, we shall first consider the basic setup with i.i.d. errors where $V_u = I_T$ and hence $u_i \equiv \epsilon_i$. In Sect. 13.7, we shall consider the extended setup with stationary AR(1) errors where $V_u$ is not equal to the identity matrix $I_T$. In either case, the assumption of nonrandom regressors amounts to assuming strict exogeneity of the regressors, that is, $u_{it}$ must be mean-independent of past, present, and future values of $x_{it} = (x_{it,1}^\top, x_{it,2}^\top)^\top$. This condition rules out dynamic panel data models where lags of the outcome variable are treated as predetermined regressors.

Stacking further, we let

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad X_1 = \begin{pmatrix} X_{1,1} \\ \vdots \\ X_{N,1} \end{pmatrix}, \quad X_2 = \begin{pmatrix} X_{1,2} \\ \vdots \\ X_{N,2} \end{pmatrix}, \tag{13.20}$$

and

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}, \quad \nu = \begin{pmatrix} \nu_1 \\ \vdots \\ \nu_N \end{pmatrix}, \quad D = I_N \otimes \mathbf{1}_T, \tag{13.21}$$

and arrive at

$$y = D\nu + X_1\beta_1 + X_2\beta_2 + u. \tag{13.22}$$

There are two classical approaches to the estimation of the parameters in the model given by Eq. (13.22), depending on which assumptions are made about the individual effects $\nu_i$ (see, e.g., Mundlak, 1978). In the *fixed-effects* approach, we treat $\nu_i$ as an additional subset of (nuisance) focus parameters to be estimated jointly with the other parameters of the model—an approach motivated by the fact that

unobserved time-invariant heterogeneity is often an important source of omitted-variable bias.

**Proposition 13.1** *Given Assumption 13.1, the error vector $u$ in the fixed-effects model defined in Eq. (13.22) satisfies* $\mathrm{E}[u] = 0$ *and* $\mathrm{var}(u) = \sigma_\epsilon^2 V$, *where* $V = I_N \otimes V_u$.

***Proof*** This follows immediately from Assumption 13.1. □

To implement the fixed-effects model, we need to regress $y$ on $\tilde{X}_1 = (X_1 : D)$ and $X_2$, where we note that the matrix $\tilde{X}_1$ is of order $NT \times (N+k_1)$. Here, a possible computational problem arises, caused by the fact that the matrix $\tilde{X}_1^\top \tilde{X}_1$ is of order $(N+k_1) \times (N+k_1)$, which will be large if $N$ is large, making inversion of $\tilde{X}_1^\top \tilde{X}_1$ computationally difficult or even infeasible. We shall show how to circumvent this problem in Sect. 13.5.

In contrast, the *random-effects* approach treats $v_i$ as random variables by writing Eq. (13.22) as

$$y = X_1\beta_1 + X_2\beta_2 + \zeta, \qquad \zeta = Dv + u. \tag{13.23}$$

**Proposition 13.2** *Assume, in addition to Assumption 13.1, that $v_i$ are i.i.d. with* $\mathrm{E}[v_i] = 0$ *and* $\mathrm{var}(v_i) = \sigma_v^2$ *and that $\{v_i\}$ and $\{u_{it}\}$ are independent. Then, the error vector $\zeta$ in the random-effects model defined in Eq. (13.23) satisfies* $\mathrm{E}[\zeta] = 0$ *and* $\mathrm{var}(\zeta) = \sigma_\epsilon^2 V$, *where*

$$V = I_N \otimes \left(V_u + \alpha_1 \mathbf{1}_T \mathbf{1}_T^\top\right)$$

*and* $\alpha_1 = \sigma_v^2/\sigma_\epsilon^2$.

***Proof*** It follows from our assumptions that $\mathrm{E}[v] = 0$, $\mathrm{E}[u] = 0$, and

$$\mathrm{var}(v) = \sigma_v^2 I_N, \qquad \mathrm{var}(u) = \sigma_\epsilon^2 (I_N \otimes V_u).$$

Hence, $\mathrm{E}[\zeta] = 0$ and

$$\begin{aligned}
\mathrm{var}(\zeta) &= \mathrm{var}(Dv + u) = D\,\mathrm{var}(v) D^\top + \mathrm{var}(u) \\
&= \sigma_v^2 (I_N \otimes \mathbf{1}_T)(I_N \otimes \mathbf{1}_T^\top) + \sigma_\epsilon^2 (I_N \otimes V_u) \\
&= \sigma_v^2 (I_N \otimes \mathbf{1}_T \mathbf{1}_T^\top) + \sigma_\epsilon^2 (I_N \otimes V_u) = I_N \otimes \left(\sigma_v^2 \mathbf{1}_T \mathbf{1}_T^\top + \sigma_\epsilon^2 V_u\right),
\end{aligned}$$

thus, it completes the proof. □

In the fixed-effects model, it is easy to obtain

$$V^{-1} = I_N \otimes V_u^{-1}, \qquad V^{-1/2} = I_N \otimes V_u^{-1/2}. \tag{13.24}$$

In the random-effects model, we obtain

$$V^{-1} = I_N \otimes \left( V_u^{-1} - \alpha_2 V_u^{-1} \mathbf{1}_T \mathbf{1}_T^\top V_u^{-1} \right), \tag{13.25}$$

where

$$\alpha_2 = \frac{\sigma_v^2}{\sigma_\epsilon^2 + \sigma_v^2 \, \mathbf{1}_T^\top V_u^{-1} \mathbf{1}_T} = \frac{\alpha_1}{1 + \alpha_1 \, \mathbf{1}_T^\top V_u^{-1} \mathbf{1}_T}, \tag{13.26}$$

but to obtain explicit expressions for the square root (and its inverse) of $V$ in the random-effects model is not straightforward, except in special cases. We shall encounter one such special case in Sect. 13.6.

## 13.5 WALS Estimation of the Fixed-Effects Model with i.i.d. Errors

In the case of i.i.d. errors, we have $V_u = I_T$ and $u_i \equiv \epsilon_i$. Thus, the fixed-effects model defined in Eq. (13.22) becomes

$$y = (D : X_1) \begin{pmatrix} \nu \\ \beta_1 \end{pmatrix} + X_2 \beta_2 + \epsilon, \tag{13.27}$$

with $\mathrm{E}[\epsilon] = 0$ and $\mathrm{var}(\epsilon) = \sigma_\epsilon^2 I_n$. We noted above that the matrix $\tilde{X}_1 = (D : X_1)$ has $N + k_1$ columns and that the inversion of $\tilde{X_1}^\top \tilde{X}_1$ may therefore not be feasible when $N$ is large. Fortunately, the next proposition shows that our model-averaging approach satisfies the Frisch–Waugh–Lovell theorem when partialling out an arbitrary subset of focus regressors from the unrestricted model.[2]

**Proposition 13.3** *Consider the model*

$$y = X_0 \beta_0 + X_1 \beta_1 + X_2 \beta_2 + \epsilon, \tag{13.28}$$

*where the parameter vectors $\beta_0$, $\beta_1$, and $\beta_2$ have dimensions $k_0$, $k_1$, and $k_2$, respectively, and the matrix $X = (X_0 : X_1 : X_2)$ has full column-rank. Premultiplying Eq. (13.28) by $M_0 = I_n - X_0 (X_0^\top X_0)^{-1} X_0^\top$, we obtain*

$$\breve{y} = \breve{X}_1 \beta_1 + \breve{X}_2 \beta_2 + \breve{\epsilon}, \tag{13.29}$$

[2] This trick only works if we think of the regressors in $D$ as focus regressors. The result does not extend to auxiliary regressors, because the auxiliary coefficients are subject to a nonlinear shrinkage step that invalidates the Frisch–Waugh–Lovell theorem.

*where*

$$\check{y} = M_0 y, \qquad \check{X}_1 = M_0 X_1, \qquad \check{X}_2 = M_0 X_2, \qquad \check{\epsilon} = M_0 \epsilon.$$

*Then, the WALS estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ of $\beta_1$ and $\beta_2$ in the two models are the same, and the WALS estimator of $\beta_0$ is given by*

$$\tilde{\beta}_0 = (X_0^\top X_0)^{-1} X_0^\top \left( y - X_1 \tilde{\beta}_1 - X_2 \tilde{\beta}_2 \right).$$

The proof is given in Appendix.

Specializing this result to the subset of focus regressors in the block-diagonal matrix $D$, this means that the individual effects can be partialled out from model defined in Eq. (13.22) using the familiar within-transformation

$$v_{it}^* = v_{it} - \bar{v}_i + \bar{v} \tag{13.30}$$

of a variable $v_{it}$, where $\bar{v}_i = T^{-1} \sum_t v_{it}$ and $\bar{v} = n^{-1} \sum_i \sum_t v_{it}$.

Applying this transformation to all variables in $(y, X_1, X_2)$, we first estimate $\beta_1$ and $\beta_2$ by fitting a WALS regression of $y^*$ on $X_1^*$ and $X_2^*$ with the error variance set equal to the residual sum of squares from the unrestricted model divided by the commonly used adjusted degrees of freedom $N(T-1) - (k-1)$ (Baltagi, 2021, Section 2.2). In WALS, this adjustment has a direct effect on the model-averaging estimates of $\beta_1$ and $\beta_2$, because the unbiased estimator of $\sigma_\epsilon^2$ affects the $t$-ratios in the unrestricted model. Moreover, since the WALS estimates depend nonlinearly on the $t$-ratios of the (transformed) auxiliary coefficients, such estimates must be specified at the outset of the estimation procedure. Given the WALS estimates $\tilde{\beta} = (\tilde{\beta}_1^\top, \tilde{\beta}_2^\top)^\top$ of $\beta = (\beta_1^\top, \beta_2^\top)^\top$, we can then estimate $v_i$ by

$$\tilde{v}_i = \bar{y}_i - \bar{x}_{i,1}^\top \tilde{\beta}_1 - \bar{x}_{i,2}^\top \tilde{\beta}_2, \tag{13.31}$$

where $\bar{x}_{i,1} = T^{-1} \sum_t x_{it,1}$ and $\bar{x}_{i,2} = T^{-1} \sum_t x_{it,2}$. The estimates resulting from this procedure are identical to those obtained from the computationally more demanding WALS regression of $y$ on $\tilde{X}_1 = (X_1, D)$ and $X_2$ in the model defined in Eq. (13.22).

## 13.6 WALS Estimation of the Random-Effects Model with i.i.d. Errors

In the case of random effects and i.i.d. errors, the variance in Proposition 13.2 specializes to

$$V = I_N \otimes \left( I_T + \alpha_1 \mathbf{1}_T \mathbf{1}_T^\top \right) = I_N \otimes \left( (T\alpha_1 + 1) J_T + (I_T - J_T) \right), \tag{13.32}$$

where $J_T = \mathbf{1}_T \mathbf{1}_T^\top / T$. Since $J_T$ and $I_T - J_T$ are idempotent and $J_T(I_T - J_T) = 0$, we obtain the following result.

**Proposition 13.4** *The matrix $V$, defined in Eq. (13.32), is positive definite, and*

$$V^{-1} = I_N \otimes \left( \frac{1}{T\alpha_1 + 1} J_T + (I_T - J_T) \right)$$

*and*

$$V^{-1/2} = I_N \otimes \left( \frac{1}{\sqrt{T\alpha_1 + 1}} J_T + (I_T - J_T) \right).$$

***Proof*** Let $\xi_1 > 0$ and $\xi_2 > 0$. Then,

$$\left(\xi_1^{-1} J_T + \xi_2^{-1}(I_T - J_T)\right)\left(\xi_1 J_T + \xi_2 (I_T - J_T)\right) = I_T$$

and

$$\left(\xi_1^{-1/2} J_T + \xi_2^{-1/2}(I_T - J_T)\right)\left(\xi_1^{-1/2} J_T + \xi_2^{-1/2}(I_T - J_T)\right) = \xi_1^{-1} J_T + \xi_2^{-1}(I_T - J_T).$$

Setting $\xi_1 = T\alpha_1 + 1$ and $\xi_2 = 1$ completes the proof. □

It follows from Proposition 13.4 that, for any two vectors $a$ and $b$ of appropriate orders,

$$V^{-1/2}(a \otimes b) = a \otimes \left( \frac{1}{\sqrt{T\alpha_1 + 1}} J_T b + b - J_T b \right) = a \otimes \left( b - \alpha_3 \bar{b}\, \mathbf{1}_T \right), \tag{13.33}$$

since $J_T b = \bar{b}\, \mathbf{1}_T$ and considering

$$\alpha_3 = 1 - \sqrt{\frac{\sigma_\epsilon^2}{T\sigma_v^2 + \sigma_\epsilon^2}}. \tag{13.34}$$

Hence, the appropriate transformation (Baltagi, 2021, Section 2.3) of a variable $v_{it}$ is now

$$v_{it}^* = v_{it} - \alpha_3 \bar{v}_i, \tag{13.35}$$

where $\bar{v}_i = T^{-1} \sum_t v_{it}$.

If the panel is unbalanced in the sense that $t = 1, \ldots, T_i$ for the $i$th unit and the $T_i$ are not all equal, then this transformation is still valid if we replace $\alpha_3$ with $\alpha_{3i}$, which we obtain from Eq. (13.34) by replacing $T$ with $T_i$. The most important threat posed by missing values in panel data arises from a nonignorable missing-data mechanism, such as selective attrition across waves of a longitudinal survey.

Even after adjustments for ignorable missing data, these endogeneity issues can still bias our estimates of the parameters of interest (Baltagi, 2021, Section 11.4).

## 13.7 The Baltagi–Wu Transformation

In Sects. 13.5 and 13.6, we assumed that the errors are independent and identically distributed. Let us now consider the case of serially correlated errors, where $V_u$ is not equal to the identity matrix. We shall only consider the model with stationary AR(1) errors, given by

$$u_{it} = \rho\, u_{i,t-1} + \epsilon_{it}, \tag{13.36}$$

where $|\rho| < 1$ and the $\epsilon_{it}$ are i.i.d. errors that follow a $\mathcal{N}(0, \sigma_\epsilon^2)$ distribution.

The variance matrix depends on how we specify the error in the first time period. The typical assumptions are

$$u_{i1} = \frac{\epsilon_{i1}}{\sqrt{1-\rho^2}} \qquad \text{or} \qquad u_{i1} = \epsilon_{i1}. \tag{13.37}$$

In the first (more common) case, the variance matrix is $\Omega = \sigma_\epsilon^2(I_N \otimes V_u)$, where

$$V_u = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-3} & \rho^{T-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho^{T-2} & \rho^{T-3} & \rho^{T-4} & \dots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & \rho & 1 \end{pmatrix} \tag{13.38}$$

and

$$V_u^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}. \tag{13.39}$$

There are various ways to estimate $\rho$. In principle, the theory of Sect. 13.3 applies whether or not the matrix $V_u$ is equal to the identity matrix, and it is sufficient to premultiply the system by $V^{-1/2} = I_N \otimes V_u^{-1/2}$. But in practice, it is more efficient to use a simple transformation without calculating the large matrix $V^{-1/2}$. Feasible GLS strategies for the case of AR(1) errors for both fixed effects and random effects were developed by Bhargava et al. (1982), Baltagi and Li (1991), and Baltagi and

Wu (1999). In particular, the random-effects estimator of Baltagi and Wu (1999) treats exogenously unbalanced panels with unequally spaced observations, where the missing-data mechanism is assumed to be ignorable.

Based on the Baltagi–Wu transformation, we proceed as follows:

*Fixed Effects*

1. Estimate $\rho$ and $\sigma_\epsilon^2$ in the unrestricted model.
2. Use the estimated correlation $\hat{\rho}$ from step 1 to compute the Baltagi–Wu AR(1) transformations $(y^*, X_1^*, X_2^*)$ of $(y, X_1, X_2)$.
3. Compute the within-transformations $(y^{**}, X_1^{**}, X_2^{**})$ of $(y^*, X_1^*, X_2^*)$ by excluding the $N$ observations corresponding to the first period of each unit.
4. Estimate $\beta_1$ and $\beta_2$ by performing a WALS regression of $y^{**}$ on $X_1^{**}$ and $X_2^{**}$ with error variance set equal to its unbiased estimate $\hat{\sigma}_\epsilon^2$ from step 1.

*Random Effects*

1. Estimate $\rho$, $\sigma_\epsilon^2$, and $\sigma_\nu^2$ in the unrestricted model.
2. Use the estimated correlation $\hat{\rho}$ from step 1 to compute the Baltagi–Wu AR(1) transformations $(y^*, X_1^*, X_2^*)$ of $(y, X_1, X_2)$.
3. Use the estimated variance components $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\nu^2$ from step 1 to compute the Baltagi–Wu feasible GLS transformations $(y^{**}, X_1^{**}, X_2^{**})$ of $(y^*, X_1^*, X_2^*)$.
4. Estimate $\beta_1$ and $\beta_2$ by performing a WALS regression of $y^{**}$ on $X_1^{**}$ and $X_2^{**}$.

## 13.8 The Impact of Legalized Abortion on Crime

As an empirical illustration, we revisit the abortion–crime hypothesis put forward by Donohue and Levitt (2001), which states that there exists a causal relationship between the legalization of abortion in the USA during the early 1970s and the subsequent decline in violent, property, and murder crimes observed in the 1990s. Donohue and Levitt, henceforth DL, provide two main arguments for their hypothesis: a cohort-size effect and a selection effect. First, holding the number of pregnancies constant, a higher abortion rate in a given year yields a smaller youth population in the next generation. Since younger individuals tend to commit more crimes than older individuals, this decrease in the share of younger people may result in lower crime rates as the cohorts exposed to legalized abortion reach the age when crimes are typically committed. Second, legalized abortion provides women with greater control over child planning, increasing the likelihood that children are born into more favorable environments for their development. This improved selection effect may further contribute to a reduction in crime rates, even if fertility rates remain constant.

The empirical findings of Donohue and Levitt (2001) attracted much attention, both within and outside the academic community, and led to many comments,

replies, and controversies, which were reviewed in Donohue and Levitt (2020). In this more recent study, DL also present additional empirical evidence for the causal impact of abortion on crime by extending their original state-level panel data analysis from the period 1985–1997 to the period 1985–2014. Although the new data differ from the original data in several important respects, the baseline models in the two studies are quite similar. In addition to state and year fixed effects, both studies rely on a limited set of control variables for eight time-varying and state-specific factors: log of lagged prisoners per capita, log of lagged police per capita, log of per capita income, per capita beer consumption, unemployment rate, poverty rate, generosity of the AFDC welfare program at time $t - 15$, and an indicator for the presence of concealed weapons laws.

As argued by Belloni et al. (2014a), a key concern with these parsimonious models is that state-level abortion rates in the early 1970s were not randomly assigned. Thus, failing to adequately control for factors associated with both state-level abortion and crime rates may introduce serious omitted-variable bias. Examples of such confounding factors include nonlinear aggregate trends and persistent state-to-state differences in policies, attitudes, and demographics. In the spirit of Belloni et al. (2014a); Belloni et al. (2014b), Hahn et al. (2018), and De Luca et al. (2022), our objective is to assess the robustness of the cross-state regression results reported by Donohue and Levitt (2020) with respect to additional controls for suitable transformations of the available explanatory variables and their interactions with a nonlinear trend. This leads us to consider an extended specification of the fixed-effects models estimated by Donohue and Levitt (2020) with a much richer set of control variables, thus strengthening the plausibility of the conditional independence assumption for the abortion rate. Furthermore, by using the fixed-effects WALS estimator, we can properly deal with the bias-precision trade-off in estimating the treatment effects of interest.

### *13.8.1 Preliminary Analysis of the Baseline Specifications*

Our analysis is based on a panel of 51 US states over the period 1985–2014, thus containing a total of $51 \times 30 = 1530$ state-year observations. We distinguish between four crime measures: violent crime, property crime, and murder crime based on both Uniform Crime Reports (UCR) and Vital Statistics (VS) data.[3] For each crime measure, Donohue and Levitt (2020) estimate a fixed-effects model with stationary AR(1) errors where the dependent variable is the log of the per capita crime rate in state $i$ at time $t$. The treatment variables of interest are obtained by interacting the "effective abortion rate" (EAR) for a given state, year, and crime

[3] Data are available at the website https://works.bepress.com/john_donohue/192/.

**Table 13.1** Fixed-effects estimates of the baseline specifications for log of per capita crime rates

| | | Type of crime[c] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method[a] | Variable[b] | Violent | | Property | | Murder (UCR) | | Murder (VS) | |
| (A) | EAR '85–'97 | −0.178 | ** | −0.152 | ** | −0.100 | * | −0.117 | ** |
| | | [0.022] | | [0.016] | | [0.040] | | [0.036] | |
| | EAR '98–'14 | −0.189 | ** | −0.168 | ** | −0.152 | ** | −0.164 | ** |
| | | [0.019] | | [0.015] | | [0.021] | | [0.019] | |
| | $n$ | 1517 | | 1517 | | 1517 | | 1499 | |
| | $k$ | 40 | | 40 | | 40 | | 40 | |
| | $\hat{\sigma}_\epsilon$ | 0.050 | | 0.037 | | 0.124 | | 0.109 | |
| | $\hat{\rho}$ | 0.904 | | 0.861 | | 0.430 | | 0.438 | |
| (B) | EAR '86–'97 | −0.214 | ** | −0.163 | ** | −0.131 | ** | −0.171 | ** |
| | | (0.035) | | (0.027) | | (0.042) | | (0.038) | |
| | EAR '98–'14 | −0.223 | ** | −0.176 | ** | −0.172 | ** | −0.194 | ** |
| | | (0.032) | | (0.026) | | (0.026) | | (0.025) | |
| | $n$ | 1466 | | 1466 | | 1466 | | 1455 | |
| | $k$ | 39 | | 39 | | 39 | | 39 | |
| | $\hat{\sigma}_\epsilon$ | 0.050 | | 0.035 | | 0.118 | | 0.102 | |
| | $\hat{\rho}$ | 0.891 | | 0.883 | | 0.633 | | 0.678 | |

[a] Method (A) provides our replication in Stata of the FGLS estimates from columns (2), (4), (6), and (8) of Table 4 in Donohue and Levitt (2020). Method (B) uses the same specifications as method (A), but the underlying estimation procedure differs in the choice of the weights and the estimation method for $\rho$, the use of Baltagi–Wu and Cochrane–Orcutt transformations, and the use of classical standard errors rather than robust standard errors. Details are given in the text

[b] Model statistics: $n$ represents the effective sample size in the second step of the FGLS procedure, $k$ is the number of regressors (excluding fixed effects), $\hat{\sigma}_\epsilon$ denotes the estimated standard deviation of the error term, and $\hat{\rho}$ is the estimated autocorrelation coefficient

[c] The symbols †, ∗, and ∗∗ denote significance levels of 10%, 5%, and 1%, respectively. Robust standard errors are presented in squared brackets, while classical standard errors are reported in curved brackets

category with binary indicators for the time periods 1985–1997 and 1998–2014.[4] Each model includes state and year fixed effects to control for time-invariant state-specific characteristics and national aggregate trends, respectively. In what follows, we focus on the specifications corresponding to columns (2), (4), (6), and (8) of Donohue and Levitt (2020, Table 4), which also include the eight time-varying and state-specific factors described above.

In the first panel (method A) of Table 13.1, we present our replication in Stata of their FGLS estimates. Except for a negligible difference in the estimated coefficient

[4] The EAR represents a measure of the abortion rate that is relevant for each type of crime, based on the typical ages at which individuals tend to commit these crimes. For a given crime category, the indicator is defined as a weighted average of the number of abortions per 1000 live births, with weights determined by the fractions of total arrests in 1985 across various ages (see Donohue & Levitt, 2020, Eq. (1)).

of EAR '85–'97 for murder crimes based on VS data, our estimates coincide with those reported in Donohue and Levitt (2020, Table 4). These fixed-effects regressions are based on a weighted two-step procedure that accounts for both state population weights and first-order serial correlation using the FGLS approach of Bhargava et al. (1982). In the first step, we compute the residuals from a weighted LS (WLS) regression with weights equal to the state population, and then estimate the correlation coefficients $\rho$ by $\hat{\rho} = 1 - d_p/2$, where $d_p$ is the panel data generalization of the Durbin–Watson statistics given in Bhargava et al. (1982, Eq. 4). In the second step, we perform a WLS regression with robust standard errors based on the Prais–Winsten transformations (Bhargava et al., 1982, Eqs. 16a and 16b). The estimated coefficients of the two EAR variables are negative and significant at the 1% level (except the first-period effect on UCR murder, which is significant only at the 5% level). Moreover, the estimated effects for the second period are always larger than those for the first period. The empirical findings thus seem to support the abortion–crime hypothesis.

The FGLS approach employed by DL is not entirely consistent with the approach used in the development of the WALS estimator for fixed-effects models with AR(1) errors. Therefore, to obtain meaningful comparisons, it is important to assess the influence of the underlying methodological differences. This is done in the second panel (method B) of Table 13.1, where we present the estimates resulting from our FGLS approach.

Method (B) differs from method (A) in five respects. First, the fixed-effects algorithm described in Sect. 13.7 can be easily extended to weighted data, provided that the weights are constant within units of the panel, so that the within-transformation effectively removes the fixed effects. In contrast, a dummy-variable specification of fixed effects allows for time-varying weights, but is feasible only when $N$ (the number of states) is relatively small. The weights employed in method (B) are the average state population over the period 1985–2014. Intermediate comparisons reveal that, other things being equal, this difference has only a small effect on the estimated coefficients of the two EAR variables and their standard errors.

Second, using weighted data restricts the available methods for the first-step estimation of the autocorrelation coefficient $\rho$. Since the method used by Donohue and Levitt (2020) is designed for unweighted data, we opt for another estimation method that properly accounts for the weighted data structure using a WLS regression of the first-step residuals on their lagged values.[5] The results are again similar, except in the estimates of $\rho$ in the two models for murder crimes.

Third, our fixed-effects approach relies on the Baltagi–Wu AR(1) transformation, which can accommodate unbalanced panels with unequally spaced observations over time. For equally spaced panels, this transformation is equivalent to the panel data generalization of the Prais–Winsten transformation proposed by Bhargava et al. (1982). However, when the panel is unequally spaced, the Prais–Winsten

[5] Our approach is one of the two estimation methods for $\rho$ available in the `xtregar` Stata command. The alternative method, involving a WLS regression of the residuals on their lead values, yields similar estimates of $\rho$.

transformation is invalid. In the current application, this issue may affect the estimates of the model for murder crime based on VS data, which are unbalanced and unequally spaced due to the presence of missing values in the outcome variable.[6]

Fourth, to ensure that the within-transformation effectively removes the fixed effects, we must exclude the $N$ observations corresponding to the first period of each unit. The differences in the sample size and the number of regressors between methods (A) and (B) are due to the use of the Cochrane–Orcutt transformation instead of the Prais–Winsten transformation. To highlight this adjustment, in method (B), we have relabeled the first-period treatment variable as EAR '86–'97.

Fifth, since method (B) will serve as the benchmark when we compare with the fixed-effects WALS estimates, we report classical standard errors. Our model-averaging approach does not currently support the calculation of robust standard errors. In fact, to our knowledge, the problem of computing robust standard errors for model-averaging estimators is still unexplored.

When comparing the estimated coefficients of the two EAR variables, we conclude that the methodological differences between methods (A) and (B) provide additional support for the abortion–crime hypothesis: the point estimates from method (B) are consistently larger (in absolute value) than those obtained with method (A). The underlying standard errors are also relatively larger, especially in the models with sizeable first-step estimates of $\rho$. However, all effects are now significant at the 1% level. The central issue remains to determine whether the strong negative association between EAR and crime rates reflects a causal relationship or whether it is driven by other confounding factors.

### *13.8.2 Expanding the Set of Control Variables*

To mitigate concerns about spurious correlations, we now consider an extended specification of the models estimated by Donohue and Levitt (2020), adding control variables for lagged values, squared terms, and interaction terms of the eight time-varying and state-specific factors already present in their baseline specifications.[7] Our set of control variables also contains interaction terms of a cubic trend with (i) the initial level of each crime-specific EAR variable; (ii) the current levels, lagged levels, initial levels, and within-state averages of the eight time-varying and state-specific factors considered by DL; and (iii) the squares of all the aforementioned variables. After excluding some regressors due to perfect collinearity, we have a

[6] The panels for the other three types of crime are unbalanced but equally spaced, because only few values are missing in the log of lagged prisoners per capita.

[7] We exclude the lagged generosity of the AFDC welfare program to avoid introducing additional missing values. Squared terms are limited to continuous regressors, while interaction terms are computed for both current and lagged levels of the original control variables.

**Table 13.2** Fixed-effects estimates of the extended models for log of per capita crime rates

| | | Type of crime[c] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method[a] | Variable[b] | Violent | | Property | | Murder (UCR) | | Murder (VS) | |
| (C) | EAR '86–'97 | 0.001 | | −0.068 | * | −0.146 | * | −0.184 | ** |
| | | (0.058) | | (0.034) | | (0.058) | | (0.049) | |
| | EAR '98–'14 | 0.000 | | −0.072 | * | −0.176 | ** | −0.208 | ** |
| | | (0.056) | | (0.034) | | (0.048) | | (0.040) | |
| | $n$ | 1466 | | 1466 | | 1466 | | 1455 | |
| | $k$ | 273 | | 273 | | 273 | | 273 | |
| | $\hat{\sigma}_\epsilon$ | 0.046 | | 0.032 | | 0.105 | | 0.089 | |
| | $\hat{\rho}$ | 0.762 | | 0.727 | | 0.178 | | 0.165 | |
| (D) | EAR '86–'97 | 0.027 | | −0.051 | † | −0.099 | * | −0.116 | ** |
| | | (0.040) | | (0.029) | | (0.043) | | (0.037) | |
| | EAR '98–'14 | 0.025 | | −0.056 | † | −0.125 | ** | −0.144 | ** |
| | | (0.038) | | (0.029) | | (0.034) | | (0.028) | |
| | $n$ | 1466 | | 1466 | | 1466 | | 1455 | |
| | $k_1$ | 31 | | 31 | | 31 | | 31 | |
| | $k_2^*$ | 62 | | 116 | | 98 | | 90 | |
| | $\hat{\sigma}_\epsilon$ | 0.045 | | 0.031 | | 0.103 | | 0.088 | |
| | $\hat{\rho}$ | 0.792 | | 0.769 | | 0.251 | | 0.258 | |
| (E) | EAR '86–'97 | −0.039 | | −0.085 | * | −0.130 | * | −0.171 | ** |
| | | (0.047) | | (0.028) | | (0.049) | | (0.043) | |
| | EAR '98–'14 | −0.042 | | −0.092 | * | −0.162 | ** | −0.201 | ** |
| | | (0.046) | | (0.027) | | (0.040) | | (0.036) | |
| | $n$ | 1466 | | 1466 | | 1466 | | 1455 | |
| | $k_1$ | 31 | | 31 | | 31 | | 31 | |
| | $k_2$ | 242 | | 242 | | 242 | | 242 | |
| | $\hat{\sigma}_\epsilon$ | 0.046 | | 0.032 | | 0.105 | | 0.089 | |
| | $\hat{\rho}$ | 0.762 | | 0.727 | | 0.178 | | 0.165 | |

[a] Method (C) employs the same fixed-effects FGLS approach as method (B). Method (D) is a pretest version of this approach based on a general-to-specific variable selection procedure. Method (E) applies the fixed-effects WALS approach with AR(1) errors and Pareto prior

[b] Model statistics: $n$ represents the effective sample size in the second step of the FGLS procedure, $k$ is the number of regressors (excluding fixed effects), $k_1$ is the number of focus regressors (excluding fixed effects), $k_2$ is the number of auxiliary regressors, $k_2^*$ is the number of selected auxiliary regressors, $\hat{\sigma}_\epsilon$ denotes the estimated standard deviation of the error term, and $\hat{\rho}$ is the estimated autocorrelation coefficient

[c] The symbols †, *, and ** denote a significance level of 10%, 5%, and 1%, respectively. Classical standard errors are presented in curved brackets

total of 323 explanatory variables: 1 constant term, 2 treatment variables, 50 state effects, 28 year effects, and 242 time-varying and state-specific controls.

Table 13.2 presents the fixed-effects estimates resulting from three alternative methods: method (C) uses the same fixed-effects FGLS approach as method (B) with the extended set of controls, method (D) is a pretest version of this approach

based on a general-to-specific variable selection procedure, while method (E) applies the fixed-effects WALS estimator. In methods (D) and (E), the focus regressors include the two EAR treatment variables, along with state and year fixed effects. Model selection and model averaging are performed on the remaining set of 242 auxiliary regressors. In method (C), which represents the unrestricted model, the distinction between focus and auxiliary regressors does not play a role.

Comparing the results from methods (B) and (C), we see that using a larger set of control variables leads to smaller estimates of both the error variance and the autocorrelation coefficient across all models. These differences are not entirely surprising, as method (C) uses a broader set of controls that also includes lagged levels of the original controls selected by DL and their interactions. The reduction in the first-step estimates of $\rho$ is more pronounced in the two models for murder crime, where $\hat{\rho}$ decreases from 0.633 to 0.178 for UCR data and from 0.678 to 0.165 for VS data. For violent crimes, we find that the estimated effects of the two EAR variables are positive and not significant at the conventional levels. For property crimes, the EAR coefficients decrease from $-0.163$ to $-0.068$ for the first time period and from $-0.176$ to $-0.072$ for the second time period. Moreover, each effect is significant only at the 5% level. On the other hand, for UCR and VS murder crimes, the negative effects of EAR become slightly stronger and remain significant at the 1% level.

An obvious concern with the unrestricted estimates in method (C) is the risk of overfitting, which is reflected in the large standard errors. Method (D) addresses this issue using an iterative general-to-specific variable selection procedure that retains all focus regressors and progressively eliminates the least significant auxiliary regressors if their $p$-value exceeds 0.05. The number of auxiliary regressors selected by this procedure ranges from a minimum of 62 controls for violent crimes to a maximum of 116 controls for property crimes. Looking at the fixed-effects FGLS estimates of the selected model for each crime measure, we see that the effects of EAR on violent crimes remain positive and not statistically significant at the conventional levels, while the effects on property and murder crimes are relatively smaller (in absolute value) than those obtained in method (C). Although method (D) produces smaller standard errors than method (C), the effects of the two EAR variables on property crimes are significant only at the 10% level and the effect of the first EAR variable on (UCR) murder crimes is significant only at the 5% level.

Model selection methods, like method (D), have several limitations. Most importantly, the properties of an estimator are often reported as if estimation had not been preceded by one or many model selection steps. However, disregarding the uncertainty due to the model selection process can lead to misleading inference (see, e.g., Leeb & Pötscher, 2005). The WALS approach developed in this chapter resolves some of the issues faced by other estimation methods. First, in addition to accounting for unobserved time-invariant heterogeneity and serial correlation in panel data models, the WALS approach relies on a preliminary semi-orthogonal transformation of the auxiliary regressors that greatly reduces the computational time required for model averaging. Second, it addresses the bias-precision trade-off arising in finite-sample estimation problems through a transparent Bayesian

shrinkage procedure that places a prior on the population $t$-ratios associated with the (transformed) auxiliary coefficients. Third, the WALS estimator enjoys desirable theoretical properties such as admissibility, bounded risk, robustness, minimax regret optimality, and $\sqrt{n}$-uniform consistency. Fourth, although its finite-sample distribution is not normal, it is relatively straightforward to derive plug-in estimators of its sampling moments and simulation-based confidence intervals.

In the last panel (method E) of Table 13.2, we present the fixed-effects WALS estimates based on a Pareto prior and the plug-in double shrinkage estimates of their standard errors. Significance levels of the estimated coefficients are derived from simulation-based confidence intervals constructed through 5000 Monte Carlo replications.[8] Our results suggest that the EAR coefficients are negative for all crime outcomes, but the magnitude and statistical significance of the estimates for violent and property crimes differ substantially from those obtained in method (B), which is based on the small set of controls selected by DL. In the model for violent crimes, the 90% confidence intervals for the EAR coefficients are $(-0.098, 0.089)$ for the first time period and $(-0.096, 0.085)$ for the second time period. In the model for property crimes, the EAR coefficients lie between those from methods (B) and (C), but are much closer to the unrestricted estimates from method (C). Although the WALS estimates are generally more precise than the unrestricted estimates, the estimated effects of EAR on property crimes are only significant at the 5% level. On the other hand, the estimated effects of EAR on UCR and VS murder crimes are quite close to the estimates obtained in method (B). Thus, even after conditioning on our large set of controls, the negative impact of legalized abortion on murder crimes remains strong and significant at the 1% level.

## 13.9 Concluding Remarks

In this chapter, we have extended the WALS approach of Magnus et al. (2010) to linear and static panel data models. We explored both fixed-effects and random-effects models, considering two alternative setups for the regression errors: one assuming i.i.d. errors and another accounting for AR(1) errors. Our fixed-effects estimation procedure exploits the fact that WALS satisfies the Frisch–Waugh–Lovell theorem when partialling out an arbitrary subset of focus regressors from the unrestricted model. Consequently, when treating the dummy variables for fixed effects as (nuisance) focus regressors that are needed to account for unobserved time-invariant heterogeneity, this subset can be partialled out from the unrestricted model using the familiar within-transformation.

The random-effects estimator and the extended setups for AR(1) errors build on special cases of the FGLS strategy for WALS estimation of linear models with

[8] Using a laptop with an Intel(R) Core(TM) i7-11800H CPU/2.30 GHz processor and 32 GB of RAM, the WALS estimates of each model are computed in about 1 minute.

nonspherical errors. Although these models are well-established in the panel data literature, the underlying WALS estimation procedure is challenging, particularly when dealing with unbalanced and unequally spaced panels.

To address these issues, we developed user-friendly Stata commands for fixed-effects and random-effects WALS estimation of the panel data models discussed in this chapter. Additional details about these commands will be provided in our forthcoming monograph (De Luca & Magnus, 2025), but the software is already freely available from the authors upon request.

In our empirical application, we employed the WALS estimator for fixed-effects panel data models with AR(1) errors to analyze the impact of legalized abortion on violent, property, and murder crimes, using data from Donohue and Levitt (2020). Our results provide further empirical support for Donohue and Levitt (2001, 2020)s hypothesis concerning the impact of legalized abortion on murder crimes. However, for violent and property crimes, the empirical evidence supporting this hypothesis is less definitive.

## Appendix

### *Proof of Proposition 13.3*

Consider first the model given by Eq. (13.29). Defining

$$\breve{M}_1 = I_n - \breve{P}_1, \qquad \breve{P}_1 = \breve{X}_1(\breve{X}_1^\top \breve{X}_1)^{-1}\breve{X}_1^\top,$$

we have

$$\breve{P}_1 = M_0 X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0$$

and

$$\breve{X}_2' \breve{M}_1 \breve{X}_2 = X_2^\top M_0 X_2 - X_2^\top M_0 X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 X_2. \tag{13.40}$$

Let $\breve{\Delta}_1$ and $\breve{\Delta}_2$ be diagonal matrices with diagonal elements

$$(\breve{\Delta}_1)_{jj} = \sqrt{\frac{n}{(\breve{X}_1^\top \breve{X}_1)_{jj}}} \qquad j = 1, \ldots, k_1$$

and

$$(\breve{\Delta}_2)_{hh} = \sqrt{\frac{n}{(\breve{X}_2^\top \breve{M}_1 \breve{X}_2)_{hh}}} \qquad h = 1, \ldots, k_2,$$

respectively, implying that the diagonal elements of the matrices $\breve{\Delta}_1 \breve{X}_1^\top \breve{X}_1 \breve{\Delta}_1 / n$ and $\breve{\Psi} = \breve{\Delta}_2 \breve{X}_2^\top \breve{M}_1 \breve{X}_2 \breve{\Delta}_2 / n$ are all equal to 1, and define

$$\breve{Z}_1 = \breve{X}_1 \breve{\Delta}_1, \qquad \breve{Z}_2 = \breve{X}_2 \breve{\Delta}_2 \breve{\Psi}^{-1/2}, \qquad \breve{\gamma}_1 = \breve{\Delta}_1^{-1} \beta_1, \qquad \breve{\gamma}_2 = \breve{\Psi}^{1/2} \breve{\Delta}_2^{-1} \beta_2.$$

Then, we obtain $\breve{Z}_1 \breve{\gamma}_1 = \breve{X}_1 \beta_1$ and $\breve{Z}_2 \breve{\gamma}_2 = \breve{X}_2 \beta_2$, which shows that the model defined in Eq. (13.29) can be written equivalently as

$$\breve{y} = \breve{Z}_1 \breve{\gamma}_1 + \breve{Z}_2 \breve{\gamma}_2 + \breve{\epsilon}, \tag{13.41}$$

where, by construction, $\breve{Z}_2^\top \breve{M}_1 \breve{Z}_2 / n = I_{k_2}$.

Next, consider the model defined in Eq. (13.28). Letting

$$X_* = (X_0 : X_1), \qquad \beta_* = (\beta_0^\top, \beta_1^\top)^\top,$$

we can write the model defined in Eq. (13.28) as

$$y = X_* \beta_* + X_2 \beta_2 + \epsilon.$$

Now,

$$(X_*^\top X_*)^{-1} = \begin{pmatrix} X_0^\top X_0 & X_0^\top X_1 \\ X_1^\top X_0 & X_1^\top X_1 \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma^{00} & \Sigma^{01} \\ \Sigma^{10} & \Sigma^{11} \end{pmatrix},$$

with

$$\Sigma^{00} = (X_0^\top X_0)^{-1} + (X_0^\top X_0)^{-1} X_0^\top X_1 \Sigma^{11} X_1^\top X_0 (X_0^\top X_0)^{-1},$$

$$\Sigma^{01} = -(X_0^\top X_0)^{-1} X_0^\top X_1 \Sigma^{11},$$

$$\Sigma^{11} = (X_1^\top M_0 X_1)^{-1}.$$

Letting $P_0 = I_n - M_0 = X_0 (X_0^\top X_0)^{-1} X_0^\top$ and $P_* = X_* (X_*^\top X_*)^{-1} X_*^\top$, we then obtain

$$\begin{aligned} P_* &= X_0 \Sigma^{00} X_0^\top + X_0 \Sigma^{01} X_1^\top + X_1 \Sigma^{10} X_0^\top + X_1 \Sigma^{11} X_1^\top \\ &= P_0 + P_0 X_1 \Sigma^{11} X_1^\top P_0 - P_0 X_1 \Sigma^{11} X_1^\top - X_1 \Sigma^{11} X_1^\top P_0 + X_1 \Sigma^{11} X_1^\top \\ &= P_0 + M_0 X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 \end{aligned}$$

and, using Eq. (13.40),

$$X_2^\top M_* X_2 = X_2^\top M_0 X_2 - X_2^\top M_0 X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 X_2 = \breve{X}_2^\top \breve{M}_1 \breve{X}_2, \tag{13.42}$$

where $M_* = I_n - P_*$. This shows that the semi-orthogonal transformation of $\breve{X}_2$ in the model defined in Eq. (13.29) has the same effect as the semi-orthogonal transformation of $X_2$ in the model given by Eq. (13.28).

In particular, letting $\Delta_*$ be the diagonal matrix

$$\Delta_* = \begin{pmatrix} \Delta_{*0} & 0 \\ 0 & \Delta_{*1} \end{pmatrix},$$

where

$$(\Delta_{*0})_{ii} = \sqrt{\frac{n}{(X_0^\top X_0)_{ii}}} \qquad i = 1, \ldots, k_0$$

and

$$(\Delta_{*1})_{jj} = \sqrt{\frac{n}{(X_1^\top X_1)_{jj}}} \qquad j = 1, \ldots, k_1,$$

we define

$$Z_* = X_* \Delta_*, \quad Z_2 = X_2 \Delta_2 \Psi^{-1/2}, \quad \gamma_* = \Delta_*^{-1} \beta_*, \quad \gamma_2 = \Psi^{1/2} \Delta_2^{-1} \beta_2,$$

and obtain

$$Z_* \gamma_* = X_* \beta_*, \qquad Z_2 \gamma_2 = X_2 \beta_2, \qquad \frac{Z_2^\top M_* Z_2}{n} = \frac{\breve{Z}_2^\top \breve{M}_1 \breve{Z}_2}{n} = I_{k_2}.$$

Model defined in Eq. (13.28) can thus be written equivalently as

$$y = Z_* \gamma_* + Z_2 \gamma_2 + \epsilon. \tag{13.43}$$

Consider next the WALS estimators of $\breve{\gamma}_2$ and $\gamma_2$ in models defined in Eqs. (13.41) and (13.43). The underlying LS estimators are the same, because

$$\begin{aligned} Z_2^\top M_* y &= Z_2^\top \left( M_0 - M_0 X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 \right) y \\ &= Z_2^\top M_0 \left( I_n - M_0 X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 \right) M_0 y = \breve{Z}_2^\top \breve{M}_1 \breve{y}, \end{aligned}$$

which ensures that, after taking into account the degrees of freedom in estimating $\sigma_\epsilon^2$, the $k_2$-vectors of $t$-ratios in the two models are the same and so are the underlying WALS estimators of $\gamma_2$. The same is true for the WALS estimators of

$\beta_2$ in models defined in Eqs. (13.28) and (13.29) because the transformations in Eqs. (13.41) and (13.43) are based on the same $\Delta_2$ and $\Psi$, in view of Eq. (13.42). Regarding the WALS estimators of the focus parameters, we note that

$$\breve{Q} = (\breve{Z}_1^\top \breve{Z}_1)^{-1} \breve{Z}_1^\top \breve{Z}_2 = \breve{\Delta}_1^{-1} (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 X_2 \breve{\Delta}_2 \breve{\Psi}^{-1/2},$$

$$\hat{\gamma}_{1r} = (\breve{Z}_1^\top \breve{Z}_1)^{-1} \breve{Z}_1^\top \breve{y} = \breve{\Delta}_1^{-1} (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 y,$$

leading to the WALS estimator of $\beta_1$ in model given by Eq. (13.29):

$$\tilde{\beta}_1 = \breve{\Delta}_1 (\hat{\gamma}_{1r} - \breve{Q} \tilde{\gamma}_2) = (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 (y - X_2 \tilde{\beta}_2).$$

In the model given by Eq. (13.28), we have

$$(Z_*^\top Z_*)^{-1} Z_*^\top = \Delta_*^{-1} \begin{pmatrix} Q_0 \\ (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 \end{pmatrix},$$

where

$$Q_0 = (X_0^\top X_0)^{-1} X_0^\top \left( I_n - X_1 (X_1^\top M_0 X_1)^{-1} X_1^\top M_0 \right),$$

and hence

$$Q_* = (Z_*^\top Z_*)^{-1} Z_*^\top Z_2 = \Delta_*^{-1} \begin{pmatrix} Q_0 X_2 \Delta_2 \Psi^{-1/2} \\ \breve{\Delta}_1 \breve{Q} \end{pmatrix},$$

$$\hat{\gamma}_{*r} = (Z_*^\top Z_*)^{-1} Z_*^\top y = \Delta_*^{-1} \begin{pmatrix} Q_0 y \\ \breve{\Delta}_1 \hat{\gamma}_{1r} \end{pmatrix},$$

leading to the WALS estimator of $\beta_*$ as

$$\tilde{\beta}_* = \Delta_* (\hat{\gamma}_{*r} - Q_* \tilde{\gamma}_2) = \begin{pmatrix} (X_0^\top X_0)^{-1} X_0^\top (y - X_1 \tilde{\beta}_1 - X_2 \tilde{\beta}_2) \\ \tilde{\beta}_1 \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix}.$$

This concludes the proof. □

## References

Baltagi, B. H. (2021). *Econometric analysis of panel data* (6th ed.). Springer.

Baltagi, B. H., & Li, Q. (1991). A transformation that will circumvent the problem of autocorrelation in an error-component model. *Journal of Econometrics*, *48*, 385–393.

Baltagi, B. H., & Wu, P. X. (1999). Unequally spaced panel data regressions with AR(1) disturbances. *Econometric Theory*, *15*, 814–823.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*, 608–650.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*, 29–50.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). Horseshoe regularization for feature subset selection. ArXiv preprint arXiv:1702.07400.

Bhargava, A., Franzini, L., & Narendranathan, W. (1982). Serial correlation and the fixed effects model. *Review of Economic Studies*, *49*, 533–549.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*, 465–480.

Chen, H., Mirestean, A., & Tsangarides, C. G. (2018). Bayesian model averaging for dynamic panels with an application to a trade gravity model. *Econometric Reviews*, *37*, 777–805.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging.* Cambridge University Press.

Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal*, *8*, 277–291.

De Luca, G., & Magnus, J. R. (2025). *Weighted-average least squares: From the normal location problem to the theory and practice of model averaging*. Springer.

De Luca, G., Magnus, J. R., & Peracchi, F. (2022). Sampling properties of the Bayesian posterior mean with an application to WALS estimation. *Journal of Econometrics*, *230*, 299–317.

De Luca, G., Magnus, J. R., & Peracchi, F. (2023). Weighted-average least squares (WALS): Confidence and prediction intervals. *Computational Economics*, *61*, 1637–1664.

De Luca, G., Magnus, J. R. & Peracchi, F. (2025). Bayesian estimation of the normal location model: A non-standard approach. *Oxford Bullettin of Economics and Statistics*, *87*(5), 913–923.

Desbordes, R., Koop, G., & Vicard, V. (2018). One size does not fit all…panel data: Bayesian model averaging and data poolability. *Economic Modelling*, *75*, 364–376.

Donohue, J. J., & Levitt, S. D. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics*, *116*, 379–420.

Donohue, J. J., & Levitt, S. D. (2020). The impact of legalized abortion on crime over the last two decades. *American Law and Economics Review*, *22*, 241–302.

Gao, Y., Zhang, X., Wang, S., & Zou, G. (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, *192*, 139–151.

Hahn, P. R., Carvalho, C. M., Puelz, D., & He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, *13*, 163–182.

Jiang, B., Lv, J., Li, J., & Cheng, M.-Y. (2025). Robust model averaging prediction of longitudinal response with ultrahigh-dimensional covariates. *Journal of the Royal Statistical Society (Series B)*. *87*(2), 337–361.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*, 21–59.

Leeb, H., & Pötscher, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, *22*, 69–97.

León-González, R., & Montolio, D. (2015). Endogeneity and panel data in growth regressions: A Bayesian model averaging approach. *Journal of Macroeconomics*, *46*, 23–39.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal*, *5*, 225–236.

Magnus, J. R., & De Luca, G. (2016). Weighted-average least squares: A review. *Journal of Economic Surveys*, *30*, 117–148.

Magnus, J. R., & Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica*, *67*, 639–643.

Magnus, J. R., Powell, O., & Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, *154*, 139–153.

Magnus, J. R., Wan, A. T. K., & Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics & Data Analysis*, *55*, 1331–1341.

McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics*, *200*, 17–25.

Mirestean, A., & Tsangarides, C. G. (2016). Growth determinants revisited using limited-information Bayesian model averaging. *Journal of Applied Econometrics*, *31*, 106–132.

Moral-Benito, E. (2012). Determinants of economic growth: A Bayesian panel data approach. *Review of Economics and Statistics*, *94*, 566–579.

Moral-Benito, E. (2013). Likelihood-based estimation of dynamic panels with predetermined regressors. *Journal of Business & Economic Statistics*, *31*, 451–472.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys*, *29*, 46–75.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*, 69–85.

Saleh, A. K. M. E. (2006). *Theory of preliminary test and stein-type estimation with applications*. Wiley.

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, *58*, 644–719.

Verbeek, M., & Nijman, T. E. (1996). Incomplete panels and selection bias. In Mátyás, L., & P. Sevestre (Eds.), *The econometrics of panel data: A handbook of the theory with applications* (pp. 449–490).

Wang, W., Zhang, X., & Paap, R. (2019). To pool or not to pool: What is a good strategy for parameter estimation and forecasting in panel regressions? *Journal of Applied Econometrics*, *34*, 724–745.

Yin, S.-Y., Liu, C.-A., & Lin, C.-C. (2021). Focused information criterion and model averaging for large panels with a multifactor error structure. *Journal of Business & Economic Statistics*, *39*, 54–68.

Zhang, X., Zou, G., & Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika*, *101*, 205–218.

# Chapter 14
# Performance of Some Test Statistics for Testing the Regression Coefficients for the One- and Two-Parameter Multicollinear Gaussian Multiple Linear Regression Models: An Empirical Comparison




**Md Ariful Hoque, Zoran Bursac, and B. M. Golam Kibria**

## 14.1 Introduction

In linear regression analysis, the assumption of independence among explanatory variables is pivotal, with the ordinary least squares (OLS) estimator traditionally considered the best linear unbiased estimator (BLUE). However, multicollinearity poses a significant challenge, rendering it difficult to estimate the unique effects of individual variables. The OLS may yield inefficient results with erroneous regression coefficients, accompanied by large standard errors and imprecise confidence intervals (Kibria, 2003). To address this issue, numerous scholars have proposed methods, employing alternative biasing estimators as substitutes for OLS and comparing their efficacy. Notable contributors to this body of research include Stein (1956), Hoerl and Kennard (1970), Swindel (1976), Saleh and Kibria (1993), Kibria (2003), Khalaf and Shukur (2005), Muniz and Kibria (2009), Kibria and Banik (2016), Kibria and Lukman (2020), Dawoud and Kibria (2020), Kibria (2022), and Hoque and Kibria (2023), including recent contribution such as Chand and Kibria (2024). Hypothesis testing is a crucial aspect of statistical inference. When it comes to a regression model, our interest is testing the statistical significance of its coefficients. This procedure helps identify which variables serve as significant predictors of the outcome and which do not.

Our primary aim is to ascertain the statistical significance of regression coefficients within the model through hypothesis testing. However, the existing literature

M. A. Hoque · Z. Bursac
Department of Biostatistics, Florida International University, Miami, FL, USA
e-mail: mhoqu010@fiu.edu; zbursac@fiu.edu

B. M. G. Kibria (✉)
Department of Mathematics and Statistics, Florida International University, Miami, FL, USA
e-mail: kibriag@fiu.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_14

on hypothesis testing for regression coefficients is limited. Halawa and El Bassiouni (2000) were among the first to propose non-exact t-tests for regression coefficients under ridge regression estimation, comparing empirical sizes and powers. Additionally, Cule et al. (2011) evaluated the performance of tests based on linear ridge and logistic ridge regression models. Gökpinar and Ebegil (2016) conducted an evaluation of $t$-test performance based on various estimators of the ridge parameter $k$. Furthermore, Kibria and Banik (2019) and Perez-Melo and Kibria (2020) analyzed the effectiveness of t-tests based on different ridge parameters. Despite these efforts, further research is needed to advance understanding and improve methodologies for testing the significance of regression coefficients.

In the context of the Liu estimator, Ullah et al. (2018) addressed the testing of specific coefficients. Moreover, Perez-Melo et al. (2022) extended this analysis by examining and comparing ridge, Liu, and Kibria-Lukman estimators together, contributing to a more comprehensive understanding of regression coefficient testing across different estimation methods. These studies highlight the importance of exploring various regression techniques and their implications for hypothesis testing in regression analysis.

This chapter aims to comprehensively compare various $t$-test statistics for testing regression coefficients. We will compare the performance of these tests across various estimation methods, including OLS, ridge regression, the Liu estimator, the one-parameter modified Liu estimator proposed by Lukman et al. (2020), the James-Stein estimator, the Kibria-Lukman estimator (Kibria & Lukman, 2020), the Yang-Chang estimator (Yang & Chang, 2010), the modified ridge-type (MRT) estimator (Lukman et al., 2019), and new two Stein estimators. This comparison will be achieved through Monte Carlo simulation, evaluating the empirical type I error and power properties of each test statistic. Based on the results, we aim to recommend suitable estimators for testing in regression models. This research contributes to advancing the understanding of regression coefficient testing methods and aids practitioners in informed decisions regarding model selection and interpretation in regression analysis.

This chapter is organized as follows: The statistical methodology is presented in Sect. 14.2, including different estimators for various parameters $k$ and $d$. A simulation study and its results are discussed in Sect. 14.3, and an example has been shown in Sect. 14.4. The chapter concludes with a summary and concluding remarks in Sect. 14.5.

## 14.2 Statistical Methodology

The linear regression model and different types of estimators are discussed in this section. The model and some existing estimators are provided in Sect. 14.2.1.

### 14.2.1 Model and Some Existing Estimators

We consider the multiple linear regression model,

$$Y = X\beta + \epsilon, \quad \epsilon \sim N\left(0, \sigma^2 I_n\right), \tag{14.1}$$

where $Y$ is an $n \times 1$ vector of response variable, $X$ is an $n \times p$ full rank matrix of explanatory variables, $\beta$ is a $p \times 1$ vector of unknown regression coefficients, and $\epsilon$ is an $n \times 1$ vector of residuals that are normally distributed with $\mathrm{E}[\epsilon] = 0$ and $\mathrm{var}(\epsilon) = \sigma^2 I_n$, where $I_n$ is an $n \times n$ identity matrix.

The OLS estimator of $\beta$ defined in Eq. (14.1) is defined as

$$\hat{\beta}^{\mathrm{OLS}} = S^{-1} X^\top Y, \tag{14.2}$$

where $S = X^\top X$.

To test whether the $i$-th component of $\beta$ equal to zero, i.e., $\mathcal{H}_0 : \beta_i = 0$, vs $\mathcal{H}_1 : \beta_i \neq 0$

$$t = \frac{\hat{\beta}_i^{\mathrm{OLS}}}{\mathrm{SE}\left(\hat{\beta}_i^{\mathrm{OLS}}\right)}, \tag{14.3}$$

where $\hat{\beta}_i$ is the $i$-th component of $\hat{\beta}^{\mathrm{OLS}}$ and $\mathrm{SE}\left(\hat{\beta}_i^{\mathrm{OLS}}\right)$ is the square root of the $i$-th diagonal element of $\mathrm{var}(\hat{\beta}^{\mathrm{OLS}}) = \sigma^2 \left(X^\top X\right)^{-1}$ where

$$\hat{\sigma}^2 = \frac{\left(Y - X\hat{\beta}^{\mathrm{OLS}}\right)^\top \left(Y - X\hat{\beta}^{\mathrm{OLS}}\right)}{n - p - 1}. \tag{14.4}$$

In Eq. (14.3), the test statistic represents the least squares test statistic, which, under the null hypothesis, follows a Student's $t$-distribution with $n - p - 1$ degrees of freedom. However, when the matrix $X^\top X$ becomes ill-conditioned due to multicollinearity, the OLS estimator defined in Eq. (14.2) produces estimates with excessively large sample variances. To address this issue of multicollinearity, Hoerl and Kennard (1970) introduced the ridge estimator, as described in Sect. 14.2.2.

### 14.2.2 One-Parameter Estimators

In this section, we are going to discuss different estimators with only one parameter either $k$ or $d$.

The ridge estimator (RE) of $\beta$ is defined as

$$\hat{\beta}^{\mathrm{RE}} = A_k \hat{\beta}^{\mathrm{OLS}}, \tag{14.5}$$

where

$$A_k = \left(X^\top X + kI\right)^{-1} X^\top X,$$

$k$ being a biasing parameter where $(k > 0)$ and $I$ denoting the identity matrix.

The expected vector and the covariance matrix of the ridge estimator are

$$\mathrm{E}\left[\hat{\beta}^{\mathrm{RE}}\right] = A_k \beta,$$

$$\mathrm{var}\left(\hat{\beta}^{\mathrm{RE}}\right) = \sigma^2_{\mathrm{RE}} A_k \left(X^\top X\right)^{-1} A_k^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}^2_{\mathrm{RE}} = \frac{\left(Y - X\hat{\beta}^{\mathrm{RE}}\right)^\top \left(Y - X\hat{\beta}^{\mathrm{RE}}\right)}{n - p - 1}. \tag{14.6}$$

Since the shrinkage parameter $k$ is unknown, it is necessary to estimate it from the observed data. Numerous methods exist in the literature for estimating the shrinkage parameter $k$. In this section, we will consider some ridge parameters $k$, which is provided in Table 14.1.

In Table 14.1, $\hat{\alpha}$ is defined as

$$\hat{\alpha} = R^\top \hat{\beta}^{\mathrm{OLS}},$$

where $R$ is an orthogonal matrix such that $R^\top X^\top X R = \Lambda = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)$, where $\lambda_j$ is the $j$-th eigenvalues of $X^\top X$.

Another estimator, as described in Kejian (1993), for the linear regression model, combined the Stein estimator with the ridge estimator and is defined as

$$\hat{\beta}^{\mathrm{Liu}} = A_d \hat{\beta}^{\mathrm{OLS}}, \tag{14.7}$$

where

$$A_d = \left(X^\top X + I\right)^{-1} \left(X^\top X + dI\right),$$

$d$ being a shrinkage parameter such that $0 \leq d \leq 1$.

**Table 14.1** Estimating of the ridge parameter $k$ with corresponding references

| Estimates of parameter $k$ | Reference |
|---|---|
| $k_1 = \hat{\sigma}^2 \sum_{j=1}^{p} \hat{\alpha}_j^2 \lambda_j^2 \,/\, (\sum_{j=1}^{p} \hat{\alpha}_j^2 \lambda_j)^2$ | Hocking et al. (1976) |
| $k_2 = \hat{\sigma}^2 \,/\, \max(\alpha_j^2)$ | Hoerl and Kennard (1970) |
| $k_3 = (p-2)\hat{\sigma}^2 \,/\, \sum_{j=1}^{p} \beta_j^2$ | Thisted (1977) |
| $k_4 = p\hat{\sigma}^2 \,/\, \sum_{j=1}^{p} \hat{\alpha}_j^2$ | Hoerl et al. (1975) |
| $k_5 = \text{median}(\hat{\sigma}^2 \,/\, \alpha_j^2)$ | Kibria (2003) |
| $k_6 = \text{geometric mean}(\hat{\sigma}^2 \,/\, \alpha_j^2)^{1/2}$ | Muniz and Kibria (2009) |
| $k_7 = \max\left(0, k_4 - 1 \,/\, (n\text{VIF}_{\max})\right)$ | Dorugade and Kashid (2010) |
| $k_8 = (\lambda_{\max}\hat{\sigma}^2) \,/\, \left((n-p)\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_{\max}^2\right)$ | Khalaf and Shukur (2005) |
| $k_9 = \max\left((\lambda_j\hat{\sigma}^2) \,/\, \left((n-p)\hat{\sigma}^2 + \lambda_j\alpha_j^2\right)\right)$ | Alkhamisi and Shukur (2008) |
| $k_{10} = \text{median}\left((\lambda_j\hat{\sigma}^2) \,/\, \left((n-p)\hat{\sigma}^2 + \lambda_j\alpha_j^2\right)\right)$ | Alkhamisi and Shukur (2008) |
| $k_{11} = 1 \,/\, \max\left(\alpha_j^2\right)$ | Schaefer et al. (1984) |
| $k_{12} = (p^2\hat{\sigma}^2) \,/\, \left(\lambda_{\max}^2 \sum_{j=1}^{p} \alpha_j^2\right)$ | Asar and Erisoglu (2016) |
| $k_{13} = (p^3\hat{\sigma}^2) \,/\, (\lambda_{\max}^3 \sum_{j=1}^{p} \alpha_j^2)$ | Asar and Erisoglu (2016) |
| $k_{14} = (p\hat{\sigma}^2) \,/\, \left(\lambda_{\max}^{1/3} \sum_{j=1}^{p} \alpha_j^2\right)$ | Asar and Erisoglu (2016) |
| $k_{15} = p^2\hat{\sigma}^2 \,/\, (\sum_{j=1}^{p} \sqrt{\lambda_j})^{1/3} \sum_{j=1}^{p} \alpha_j^2$ | Asar and Erisoglu (2016) |
| $k_{16} = 2p\hat{\sigma}^2 \,/\, \left(\lambda_{\max}^{\frac{1}{2}} \sum_{j=1}^{p} \alpha_j^2\right)$ | Asar and Erisoglu (2016) |
| $k_{17} = k_2 + 2 \,/\, (\lambda_{\max} + \lambda_{\min})$ | Khalaf (2012) |
| $k_{18} = \text{median}\left(2\hat{\sigma}^2 \,/\, \left(\lambda_{\max}\alpha_j^2\right)\right)$ | Dorugade (2014) |
| $k_{19} = \text{Harmonic mean}\left(2\hat{\sigma}^2 \,/\, \left(\lambda_{\max}\alpha_j^2\right)\right)$ | Dorugade (2014) |
| $k_{20} = \text{Geometric mean}\left(2\hat{\sigma}^2 \,/\, \left(\lambda_{\max}\alpha_j^2\right)\right)$ | Dorugade (2014) |
| $k_{21} = p\hat{\sigma}$ | Lukman and Olatunji (2018) |
| $k_{22} = p^{(1+1/p)}\hat{\sigma}$ | Shabbir et al. (2024) |
| $k_{23} = \Pi_{j=1}^{p}\left(1/q_j\right)^{1/p}$ where $q_i = \lambda_{\max}/\left((n-p)\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_j^2\right)$ | Kibria et al. (2012) |
| $k_{24} = \max\left(1 \,/\, q_j\right)$ | Kibria et al. (2012) |
| $k_{25} = \text{median}\left(1 \,/\, q_i\right)$ | Kibria et al. (2012) |

**Table 14.2** Estimates of the Liu parameter $d$ with corresponding references

| Estimates of parameter $d$ | Reference |
|---|---|
| $d_1 = \max\left(0, \frac{\alpha^2_{\max} - \hat{\sigma}^2}{\frac{1}{\lambda_{\max} + \hat{\alpha}^2 \max}}\right)$ | Månsson et al. (2014) |
| $d_2 = \max\left(0, \text{median}\left(\left(\alpha_j^2 - \hat{\sigma}^2\right) / \left(1/\lambda_j + \alpha_j^2\right)\right)\right)$ | Månsson et al. (2014) |
| $d_3 = \max\left(0, \text{mean}\left(\left(\alpha_j^2 - \hat{\sigma}^2\right) / \left(1/\lambda_j + \alpha_j^2\right)\right)\right)$ | Månsson et al. (2014) |
| $d_4 = \max\left(0, \max\left(\left(\alpha_j^2 - \hat{\sigma}^2\right) / \left(1/\lambda_j + \alpha_j^2\right)\right)\right)$ | Månsson et al. (2014) |
| $d_5 = \max\left(0, \text{median}\ \left(q_j\right)\right)$ where $q_j = \left(\alpha_j^2 - 1\right)/\left(\max\left(1/\lambda_j\right) + \alpha_j^2\right)$ | Månsson et al. (2014) |
| $d_6 = \max\left(0, \text{mean}(q_j)\right)$ | Månsson et al. (2014) |
| $d_7 = \max\left(0, \max(q_j)\right)$ | Månsson et al. (2014) |
| $d_8 = \max\left(0, \text{median}\left(\left(\alpha_j^2 - \hat{\sigma}^2\right) / \left(\max\left(\hat{\sigma}^2 / \lambda_j\right) + \hat{\alpha}^2_{\max}\right)\right)\right)$ | Qasim et al. (2020) |
| $d_9 = \max\left(0, \text{mean}\left(\left(\alpha_j^2 - \hat{\sigma}^2\right) / \left(\max\left(\hat{\sigma}^2 / \lambda_j\right) + \hat{\alpha}^2_{\max}\right)\right)\right)$ | Qasim et al. (2020) |
| $d_{10} = \max\left(0, \max\left(\left(\alpha_j^2 - \hat{\sigma}^2\right) / \left(\max\left(\hat{\sigma}^2 / \lambda_j\right) + \hat{\alpha}^2_{\max}\right)\right)\right)$ | Qasim et al. (2020) |

The expected vector and covariance matrix of the Liu estimator are, respectively,

$$\mathrm{E}\left[\hat{\beta}^{\mathrm{Liu}}\right] = A_d \beta \ and$$

$$\mathrm{var}\left(\hat{\beta}^{\mathrm{Liu}}\right) = \sigma^2_{\mathrm{Liu}} A_d \left(X^\top X\right)^{-1} A_d^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}^2_{\mathrm{Liu}} = \frac{\left(Y - X\hat{\beta}^{\mathrm{Liu}}\right)^\top \left(Y - X\hat{\beta}^{\mathrm{Liu}}\right)}{n - p - 1}. \tag{14.8}$$

Since the parameter $d$ is unknown, we need to estimate it from the observed data. We have provided some estimators of $d$ in Table 14.2.

Another estimator for the linear regression model is the one-parameter modified Liu estimator defined as

$$\hat{\beta}^{\mathrm{MLiu}} = A_M \hat{\beta}^{\mathrm{OLS}}, \tag{14.9}$$

where

$$A_M = \left(X^\top X + I\right)^{-1} \left(X^\top X - dI\right),$$

and $d$ is the shrinkage parameter where $0 \leq d \leq 1$.

The expected vector and covariance matrix of $\hat{\beta}^{\mathrm{MLiu}}$ are, respectively,

$$\mathrm{E}\left[\hat{\beta}^{\mathrm{MLiu}}\right] = A_M \beta,$$

$$\mathrm{var}\left(\hat{\beta}^{\mathrm{MLiu}}\right) = \sigma^2_{\mathrm{MLiu}} A_M \left(X^\top X\right)^{-1} A_M^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}^2_{\mathrm{MLiu}} = \frac{\left(Y - X\hat{\beta}^{\mathrm{MLiu}}\right)^\top \left(Y - X\hat{\beta}^{\mathrm{MLiu}}\right)}{n - p - 1}. \tag{14.10}$$

Kibria and Lukman (2020) also introduced an estimator defined as

$$\hat{\beta}^{\mathrm{KL}} = K\hat{\beta}^{\mathrm{OLS}}, \tag{14.11}$$

where

$$K = \left(X^\top X + kI\right)^{-1} \left(X^\top X - kI\right).$$

The expected vector and covariance matrix of $\hat{\beta}^{\mathrm{KL}}$ are

$$\mathrm{E}\left[\hat{\beta}^{\mathrm{KL}}\right] = K\beta,$$

$$\mathrm{var}\left(\hat{\beta}^{\mathrm{KL}}\right) = \sigma^2_{\mathrm{KL}} K \left(X^\top X\right)^{-1} K^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}^2_{\mathrm{KL}} = \frac{\left(Y - X\hat{\beta}^{\mathrm{KL}}\right)^\top \left(Y - X\hat{\beta}^{\mathrm{KL}}\right)}{n - p - 1}. \tag{14.12}$$

As the shrinkage parameter $k$ is unknown, it is essential to estimate it based on the observed data. The parameters are presented in Table 14.3.

### *14.2.3 Stein-Type Estimators*

In this section, we are going to discuss the Stein-type estimator with parameter $c$. And we will extend two new Stein estimators for the linear regression model.

The James-Stein (Stein, 1956) estimator (JSE) is also defined as

$$\hat{\beta}^{\mathrm{JSE}} = c\hat{\beta}^{\mathrm{OLS}}, \tag{14.13}$$

**Table 14.3** Estimates of the Kibria-Lukman parameter $k$ with corresponding references. Here, AM stands for the arithmetic mean

| Estimates of parameter $k$ | Reference |
|---|---|
| $K_1 = p\hat{\sigma}^2 / \sum_{j=1}^{p} \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2/\lambda_j\right)$ | Kibria and Lukman (2020) |
| $K_2 = \min\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2/\lambda_j\right)\right)$ | Kibria and Lukman (2020) |
| $K_3 = \text{median}\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2/\lambda_j\right)\right)$ | Kibria and Lukman (2020) |
| $K_4 = \text{Harmonic mean}\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2/\lambda_j\right)\right)$ | Kibria and Lukman (2020) |
| $K_5 = \text{Geometric mean}\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2/\lambda_j\right)\right)$ | Kibria and Lukman (2020) |
| $K_6 = \text{Arithmetic mean}\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2/\lambda_j\right)\right)$ | Kibria and Lukman (2020) |

where $0 < c < 1$. For the selection of $c$, we consider

$$c = \frac{(\hat{\beta}^{\text{OLS}})^\top \hat{\beta}^{\text{OLS}}}{(\hat{\beta}^{\text{OLS}})^\top \hat{\beta}^{\text{OLS}} + \sigma^2 \,\text{tr}\left(X^\top X\right)^{-1}}. \tag{14.14}$$

The expected vector and covariance matrix of $\hat{\beta}^{\text{JSE}}$ are, respectively,

$$\text{E}\left[\hat{\beta}^{\text{JSE}}\right] = c\beta \; and$$

$$\text{var}\left(\hat{\beta}^{\text{JSE}}\right) = c^2\sigma^2_{\text{JSE}}\left(X^\top X\right)^{-1},$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}^2_{\text{JSE}} = \frac{\left(Y - X\hat{\beta}^{\text{JSE}}\right)^\top \left(Y - X\hat{\beta}^{\text{JSE}}\right)}{n - p - 1}. \tag{14.15}$$

We considered two new Stein estimators for the logistics regression model (Hoque & Kibria, 2024), and we extended it for the linear regression model.

One of them is Stein's estimator for the linear ridge regression estimator defined as

$$\hat{\beta}^{\text{SRE}} = c_r\hat{\beta}^{\text{RE}} = c_r A_k \hat{\beta}^{\text{OLS}}, \tag{14.16}$$

where $0 < c_r < 1$. For the selection of $c_r$, we consider

$$c_r = \frac{(\hat{\beta}^{\text{OLS}})^\top A_k \hat{\beta}^{\text{OLS}}}{(\hat{\beta}^{\text{OLS}})^\top A_k^\top A_k \hat{\beta}^{\text{OLS}} + \sigma^2 \,\text{tr}\left(A_k \left(X^\top X\right)^{-1} A_k^\top\right)}. \tag{14.17}$$

**Table 14.4** Estimates of parameter $k$ for Stein-ridge estimator with corresponding references

| Estimates of parameter $k$ | Reference |
|---|---|
| $\text{sr}_1 = \hat{\sigma}^2\left(\sum_{j=1}^{p}\hat{\alpha}_j^2\lambda_j^2\right) / \left(\sum_{j=1}^{p}\hat{\alpha}_j^2\lambda_j\right)^2$ | Hocking et al. (1976) |
| $\text{sr}_2 = \hat{\sigma}^2 / \max(\alpha_j^2)$ | Hoerl and Kennard (1970) |
| $\text{sr}_3 = (p-2)\hat{\sigma}^2 / \sum_{j=1}^{p}\beta_j^2$ | Thisted (1977) |
| $\text{sr}_4 = \max\left(0, k_4 - 1/(n\text{VIF}_{\max})\right)$ | Dorugade and Kashid (2010) |
| $\text{sr}_5 = (\lambda_{\max}\hat{\sigma}^2) / \left((n-p)\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_{\max}^2\right)$ | Khalaf and Shukur (2005) |
| $\text{sr}_6 = \max\left(\lambda_j\hat{\sigma}^2 / \left((n-p)\hat{\sigma}^2 + \lambda_j\alpha_j^2\right)\right)$ | Alkhamisi and Shukur (2008) |
| $\text{sr}_7 = \text{median}\left(\lambda_j\hat{\sigma}^2 / \left((n-p)\hat{\sigma}^2 + \lambda_j\alpha_j^2\right)\right)$ | Alkhamisi and Shukur (2008) |
| $\text{sr}_8 = 1 / \max\left(\alpha_j^2\right)$ | Schaefer et al. (1984) |
| $\text{sr}_9 = p^2\hat{\sigma}^2 / \left(\lambda_{\max}^2\sum_{j=1}^{p}\alpha_j^2\right)$ | Asar and Erisoglu (2016) |
| $\text{sr}_{10} = p^3\hat{\sigma}^2 / \left(\lambda_{\max}^3\sum_{j=1}^{p}\alpha_j^2\right)$ | Asar and Erisoglu (2016) |
| $\text{sr}_{11} = p\hat{\sigma}^2 / \left(\lambda_{\max}^{1/3}\sum_{j=1}^{p}\alpha_j^2\right)$ | Asar and Erisoglu (2016) |
| $\text{sr}_{12} = \text{Harmonic mean}\left(2\hat{\sigma}^2 / \left(\lambda_{\max}\alpha_j^2\right)\right)$ | Dorugade (2014) |
| $\text{sr}_{13} = p\hat{\sigma}$ | Lukman and Olatunji (2018) |

The expected vector and covariance matrix of $\hat{\beta}^{\text{SRE}}$ are, respectively,

$$\text{E}\left[\hat{\beta}^{\text{SRE}}\right] = c_r A_k \beta\,,$$

$$\text{var}\left(\hat{\beta}^{\text{SRE}}\right) = c_r^2\sigma_{\text{SRE}}^2 A_k\left(X^\top X\right)^{-1}A_k^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}_{\text{SRE}}^2 = \frac{\left(Y - X\hat{\beta}^{\text{SRE}}\right)^\top\left(Y - X\hat{\beta}^{\text{SRE}}\right)}{n-p-1}. \tag{14.18}$$

Since this shrinkage parameter $k$ is unknown, we will consider different ridge parameters $k$ as provided in Table 14.4.

The Stein's estimator for the KL estimator is defined as

$$\hat{\beta}^{\text{SKE}} = c_k\hat{\beta}^{\text{RE}} = c_k K\hat{\beta}^{\text{OLS}}, \tag{14.19}$$

where $0 < c_k < 1$. Following the same procedure for the selection of $c_r$, we consider

$$c_k = \frac{(\hat{\beta}^{\text{OLS}})^\top K\hat{\beta}^{\text{OLS}}}{(\hat{\beta}^{\text{OLS}})^\top K^\top K\hat{\beta}^{\text{OLS}} + \sigma^2\,\text{tr}\left(K\left(X^\top X\right)^{-1}K^\top\right)}. \tag{14.20}$$

**Table 14.5** Estimates of the parameters for Stein-KL estimator with corresponding references

| Estimates of parameter | Reference |
|---|---|
| $\text{sk}_1 = \min\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2 / \lambda_j\right)\right)$ | Kibria and Lukman (2020) |
| $\text{sk}_2 = \text{AM}\left(\hat{\sigma}^2 / \left(2\hat{\alpha}_j^2 + \hat{\sigma}^2 / \lambda_j\right)\right)$ | Kibria and Lukman (2020) |

The expected vector and covariance matrix of $\hat{\beta}^{\text{SKE}}$ are, respectively,

$$\text{E}\left[\hat{\beta}^{\text{SKE}}\right] = c_k K\beta,$$

$$\text{var}\left(\hat{\beta}^{\text{SKE}}\right) = c_k^2 \sigma_{\text{SKE}}^2 K \left(X^\top X\right)^{-1} K^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}_{\text{SKE}}^2 = \frac{\left(Y - X\hat{\beta}^{\text{SKE}}\right)^\top \left(Y - X\hat{\beta}^{\text{SKE}}\right)}{n - p - 1}. \tag{14.21}$$

For this estimator, with the unknown parameter $k$, we will refer to Table 14.5.

### *14.2.4 Two-Parameter Estimators*

In this section, we will discuss two-parameter estimators containing both shrinkage parameters $k$ and $d$.

Yang and Chang (2010) introduced a two-parameter estimator to solve the multicollinearity problem, which is defined as

$$\hat{\beta}^{\text{Y}} = Y_{k,d}\hat{\beta}^{\text{OLS}}, \tag{14.22}$$

where

$$Y_{k,d} = \left(X^\top X + I\right)^{-1} \left(X^\top X + dI\right) \left(X^\top X + kI\right)^{-1} \left(X^\top X\right)$$

and $k$ and $d$ are biasing parameters.

The expected vector and covariance matrix of $\hat{\beta}^{\text{Y}}$ are, respectively,

$$\text{E}\left[\hat{\beta}^{\text{Y}}\right] = Y_{k,d}\beta,$$

$$\text{var}\left(\hat{\beta}^{\text{Y}}\right) = \sigma_{\text{Y}}^2 Y_{k,d} \left(X^\top X\right)^{-1} Y_{k,d}^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}_{\mathrm{Y}}^2 = \frac{\left(Y - X\hat{\beta}^{\mathrm{Y}}\right)^{\top}\left(Y - X\hat{\beta}^{\mathrm{Y}}\right)}{n - p - 1}. \tag{14.23}$$

In situations where both shrinkage parameters $k$ and $d$ are unknown, it becomes essential to estimate them from the available data. The following discussion will provide clarity and detail regarding the formulas for estimating these shrinkage parameters in various regression estimators considered in the simulation study for the test statistic given by Eq. (14.26).

For $k$, we fixed $d$, and based on Yang and Chang (2010), we get the optimal $k$ as

$$\widehat{k}_j = \frac{\hat{\sigma}^2\left(\lambda_j + d\right) - (1-d)\lambda_j\hat{\alpha}_j^2}{\left(\lambda_j + 1\right)\hat{\alpha}_j^2}.$$

We can apply different formulas for this parameter $k$, such as arithmetic mean, geometric mean, harmonic mean, and median of $k_j$ for $j = 1, \ldots, p$.

Then, we get

$$\hat{d}_{\mathrm{opt}} = \frac{\sum_{j=1}^{p}\left[\left((k+1)\lambda_j + k\right)\lambda_j\hat{\alpha}_j^2\right] / \left[\left(\lambda_j + 1\right)^2\left(\lambda_j + k\right)^2\right]}{\sum_{j=1}^{p}\left(\hat{\sigma}^2 + \lambda_j\hat{\alpha}_j^2\right)\lambda_j / \left[\left(\lambda_j + 1\right)^2\left(\left(\lambda_j + k\right)^2\right)\right]}.$$

The selection of the estimators of the parameters $k$ and $d$ in $\hat{\beta}^{\mathrm{Y}}$ is obtained iteratively as follows:

**Step 1:** Obtain an initial estimate of $d$ using $\hat{d} = \min\left(\hat{\sigma}^2 / \hat{\alpha}_j^2\right)$.

**Step 2:** Obtain $\hat{k}_j$ using $\hat{d}$ from step 1.

**Step 3:** Estimate $\hat{d}_{\mathrm{opt}}$ using $\hat{k}$ from step 2.

A modified ridge-type (MRT) estimator is proposed by Lukman et al. (2019), which also introduced a two-parameter estimator which is defined as

$$\hat{\beta}^{\mathrm{MRT}} = M_{k,d}\hat{\beta}^{\mathrm{OLS}}, \tag{14.24}$$

where

$$M_{k,d} = \left(X^{\top}X + k(1+d)I\right)^{-1}\left(X^{\top}X\right),$$

with $k$ and $d$ as biasing parameters.

The expected vector and covariance matrix of $\hat{\beta}^{\mathrm{MRT}}$ are, respectively, given by

$$\mathrm{E}[\hat{\beta}^{\mathrm{MRT}}] = M_{k,d}\beta,$$

$$\mathrm{var}\left(\hat{\beta}^{\mathrm{MRT}}\right) = \sigma^2_{\mathrm{MRT}} M_{k,d}\left(X^\top X\right)^{-1} M_{k,d}^\top,$$

and $\sigma^2$ is estimated as

$$\hat{\sigma}^2_{\mathrm{MRT}} = \frac{\left(Y - X\hat{\beta}^{\mathrm{MRT}}\right)^\top \left(Y - X\hat{\beta}^{\mathrm{MRT}}\right)}{n - p - 1}. \tag{14.25}$$

As the shrinkage parameters $k$ and $d$ are both unknown, it is necessary to estimate them from the observed data. The following discussion will provide clarity and detail regarding the formulas for various shrinkage parameter regression estimators.

To estimate, Lukman et al. (2019) proposed that for $j = 1, \ldots, p$,

$$\hat{k}_j = \frac{\hat{\sigma}^2}{(1+d)\hat{\alpha}_j^2}.$$

The harmonic mean of $k_j$s, denoted as $k_{\mathrm{HMP}}$, is defined as

$$\hat{k}_{\mathrm{HMP}} = \frac{p\hat{\sigma}^2}{\sum_{j=1}^{p}(1+d)\hat{\alpha}_j^2},$$

where

$$\hat{d}_j = \frac{\hat{\sigma}^2}{k\hat{\alpha}_j^2} - 1.$$

The harmonic mean version of $\widehat{d_j}$ proposed by Lukman et al. (2019) is defined as

$$\hat{d}_{\mathrm{MRT}} = \frac{p}{\sum_{j=1}^{p}\frac{1}{\hat{d}_j}}.$$

The selection of the estimators for the parameters $k$ and $d$ in $\hat{\beta}_{\mathrm{MRT}}$ is obtained iteratively as follows:

**Step 1:** Obtain an initial estimate of $d$ using $\hat{d} = \min\left(\hat{\sigma}^2 / \hat{\alpha}_j^2\right)$.

**Step 2:** Obtain $\hat{k}_{\mathrm{HMP}}$ using $\hat{d}$ from step 1.

**Step 3:** Estimate $\hat{d}_{\mathrm{MRT}}$ using $\hat{k}_{\mathrm{HMP}}$ from step 2.

We will also consider the median, arithmetic mean, and geometric mean of the $k$ and $d$ instead of the harmonic mean.

To test whether the $i$-th component of the parameter vector $\beta$ is equal to zero, the approach of Halawa and El Bassiouni (2000) is used with the $t$-test statistic as

$$t^* = \frac{\hat{\beta}_i^*}{\text{SE}\left(\hat{\beta}_i^*\right)}, \tag{14.26}$$

where $\hat{\beta}_i^*$ represents the $i$-th estimated coefficient obtained using one of the estimators ridge, Liu, MLiu, KL, JSE, SRE, SKE, Yang and Chang, and MRT, respectively. The corresponding standard error, $\text{SE}(\hat{\beta}_i^*)$, is the square root of the $i$-th element of $\text{Var}(\hat{\beta}_i^*)$. Under the null hypothesis, the test statistics given by Eq. (14.26) is approximately distributed as a Student's t-distribution with $n - p - 1$ degrees of freedom (Halawa and El Bassiouni, 2000).

## 14.3 Simulation Study

In this section, a Monte Carlo study has been conducted to compare the performance of the test statistics. First, we will discuss the empirical type I error of the tests in Sect. 14.3.1, followed by the empirical powers of the tests in Sect. 14.3.2.

### *14.3.1 Empirical Type I Error*

#### 14.3.1.1 Simulation Technique

The simulation procedure specifically involves generating the explanatory variables $X$ using the formula $H\Lambda^{1/2}G^\top$, where $H$ is any $n \times p$ matrix whose columns are orthogonal, $\Lambda$ is the diagonal matrix of eigenvalues of the correlation matrix, and $G$ is the matrix of normalized eigenvectors of the correlation matrix $X^\top X$. This procedure ensures that the explanatory variables are generated in a controlled manner, allowing for the systematic examination of the type I error in the context of regression analysis.

The study generates $n$ observations for the explanatory variables according to

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i. \tag{14.27}$$

Here, $\epsilon_i$ are independent normal random variables with mean 0 and variance $\sigma^2$. The performance of the test is compared in terms of their type I error rates under different biasing parameters. The comparison considers different sample sizes $n = 30, 50$, and 100, as well as different number of variables $p = 3, 5$, and 10.

Various correlations $\rho = 0.80, 0.90$, and 0.99 are chosen, along with a standard deviation of errors $\sigma = 1$. The experiment is replicated 5000 times. R software version 4.4 was used to conduct the simulation study.

Following Halawa and El Bassiouni (2000), the determination of the most and least favorable orientations ($\beta$) is based on the normalized eigenvectors corresponding to the largest and smallest eigenvalues of $X^\top X$ in correlation form. Thus, the most favorable (MF) orientation is $\beta = \left(1 / \sqrt{p}\right) 1_p$, where $1_p$ represents a vector of ones, and the least favorable (LF) orientation is any normalized vector orthogonal to $1_p$. In the MF orientation, all components of $\beta$ are equal, whereas in the LF orientation all components are equally likely. It has been noted in other studies investigating these tests that the LF orientation produces tests that maintain the nominal type I error rates regardless of the estimator used. Conversely, some tests exhibit higher type I error rates than expected under the MF orientation of $\beta$. Therefore, for practical purposes, the MF orientation allows us to discard tests that are excessively liberal in rejecting the null hypothesis. As a result, our simulations are conducted based on the MF orientation of $\beta$.

To commence, we compute the type I error rates of the test under $\mathcal{H}_0$. For this objective, the $i$-th component of the orientation vector of $\beta$ is substituted by zero as defined by $\beta = \left(1 / \sqrt{p}\right) 1_p$, $\beta_i = 0$, $i = 1, 2, \ldots, p$. Subsequently, the test statistics are computed from the models, and the type I error rates are estimated by the proportion of test statistic values exceeding the critical values calculated from the Student's $t$-distribution with $n - p - 1$ degrees of freedom. This procedure enables us to evaluate the probability of the test rejecting null hypotheses when they are true (type I error), thus providing its empirical size.

#### 14.3.1.2 Simulation Result Discussion

We provided the empirical sizes of tests for the MF orientation across correlation levels of 0.8, 0.9, and 0.99 in Tables 14.6, 14.7, and 14.8, respectively. Assuming a true type I error rate of 5% is expected to fall within the interval of comprising, 5000 runs, we estimate that the observed type I error rates $0.05 \pm 2\sqrt{(0.05 \times 0.95) / 5000}$, yielding approximately (4.4%, 5.6%). Tests whose average observed type I error is more than 0.06 will be omitted from the comparison to ensure consistency.

From Tables 14.6, 14.7, and 14.8, we observed that some of the estimators have type I errors above the 6% nominal size in the case of the MF orientation, and they cannot be recommended.

**Table 14.6** Type I error for $\rho = 0.80$ and $\alpha = 0.05$, MF orientation

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| OLS | 0.0442 | 0.0476 | 0.0495 | 0.0464 | 0.05 | 0.0482 | 0.0429 | 0.0481 | 0.0483 | 0.0472 |
| RE1 | 0.0517 | 0.0547 | 0.0577 | 0.0462 | 0.0532 | 0.0534 | 0.036 | 0.0465 | 0.0485 | 0.0498 |
| RE2 | 0.0473 | 0.0509 | 0.0517 | 0.0461 | 0.0508 | 0.0498 | 0.041 | 0.0478 | 0.0485 | 0.0482 |
| RE3 | 0.0455 | 0.0489 | 0.0503 | 0.0456 | 0.0508 | 0.0504 | 0.0362 | 0.0463 | 0.0478 | 0.0469 |
| RE4 | 0.0539 | 0.0575 | 0.0608 | 0.0457 | 0.0536 | 0.0542 | 0.0346 | 0.0461 | 0.048 | 0.0505 |
| RE5 | 0.0559 | 0.0637 | 0.0672 | 0.0546 | 0.0664 | 0.0713 | 0.0338 | 0.0534 | 0.061 | 0.0586 |
| RE6 | 0.0617 | 0.0704 | 0.0733 | 0.0595 | 0.074 | 0.0811 | 0.0308 | 0.0563 | 0.0676 | 0.0639 |
| RE7 | 0.0485 | 0.0549 | 0.0599 | 0.0476 | 0.0531 | 0.0529 | 0.0425 | 0.0483 | 0.0484 | 0.0507 |
| RE8 | 0.0449 | 0.0483 | 0.0492 | 0.046 | 0.0497 | 0.0487 | 0.0412 | 0.0478 | 0.0484 | 0.0471 |
| RE9 | 0.043 | 0.0475 | 0.0487 | 0.0408 | 0.0485 | 0.0484 | 0.0255 | 0.0433 | 0.0471 | 0.0436 |
| RE10 | 0.0443 | 0.0477 | 0.0494 | 0.0463 | 0.0499 | 0.0483 | 0.0428 | 0.0481 | 0.0482 | 0.0472 |
| RE11 | 0.0489 | 0.0515 | 0.0524 | 0.0462 | 0.051 | 0.0498 | 0.0405 | 0.0476 | 0.0484 | 0.0485 |
| RE12 | 0.0464 | 0.0497 | 0.0504 | 0.046 | 0.0501 | 0.0493 | 0.0423 | 0.048 | 0.0485 | 0.0479 |
| RE13 | 0.047 | 0.0502 | 0.0509 | 0.0462 | 0.0501 | 0.0496 | 0.0419 | 0.0481 | 0.0486 | 0.0481 |
| RE14 | 0.049 | 0.0525 | 0.0546 | 0.0454 | 0.0509 | 0.0505 | 0.0387 | 0.0471 | 0.0479 | 0.0485 |
| RE15 | 0.0728 | 0.0829 | 0.0857 | 0.074 | 0.0909 | 0.099 | 0.0488 | 0.0852 | 0.1014 | 0.0823 |
| RE16 | 0.0589 | 0.0625 | 0.0662 | 0.0456 | 0.0536 | 0.0536 | 0.037 | 0.0466 | 0.0476 | 0.0524 |
| RE17 | 0.0635 | 0.0718 | 0.0767 | 0.0472 | 0.0577 | 0.0645 | 0.0268 | 0.0427 | 0.0467 | 0.0553 |
| RE18 | 0.0519 | 0.0593 | 0.0621 | 0.0459 | 0.0543 | 0.055 | 0.0369 | 0.0465 | 0.0478 | 0.0511 |
| RE19 | 0.0495 | 0.0526 | 0.0555 | 0.046 | 0.0506 | 0.0495 | 0.0414 | 0.0479 | 0.0484 | 0.049 |
| RE20 | 0.0561 | 0.0622 | 0.066 | 0.0492 | 0.0588 | 0.0618 | 0.0359 | 0.0465 | 0.0483 | 0.0539 |
| RE21 | 0.0835 | 0.1013 | 0.1079 | 0.1395 | 0.1778 | 0.1948 | 0.2269 | 0.3484 | 0.4129 | 0.1992 |
| RE22 | 0.0881 | 0.1067 | 0.115 | 0.148 | 0.1904 | 0.2122 | 0.2416 | 0.3725 | 0.4412 | 0.2129 |
| RE23 | 0.0883 | 0.1141 | 0.1268 | 0.151 | 0.207 | 0.2468 | 0.157 | 0.3174 | 0.4551 | 0.2071 |
| RE24 | 0.0883 | 0.1141 | 0.1268 | 0.151 | 0.207 | 0.2468 | 0.157 | 0.3174 | 0.4551 | 0.2071 |
| RE25 | 0.0885 | 0.1139 | 0.1268 | 0.1498 | 0.206 | 0.2463 | 0.1427 | 0.2978 | 0.4441 | 0.2018 |
| LIU1 | 0.0445 | 0.0486 | 0.0496 | 0.0465 | 0.0501 | 0.0488 | 0.0426 | 0.0481 | 0.0483 | 0.0475 |
| LIU2 | 0.0451 | 0.0503 | 0.0532 | 0.0448 | 0.0551 | 0.0577 | 0.0285 | 0.0465 | 0.0522 | 0.0482 |
| LIU3 | 0.0337 | 0.0366 | 0.0423 | 0.0362 | 0.0465 | 0.0494 | 0.0253 | 0.0427 | 0.0475 | 0.04 |
| LIU4 | 0.0449 | 0.0488 | 0.0499 | 0.0456 | 0.05 | 0.049 | 0.0404 | 0.0475 | 0.0483 | 0.0472 |
| LIU5 | 0.0514 | 0.0581 | 0.0599 | 0.0485 | 0.0592 | 0.0618 | 0.0284 | 0.0476 | 0.0541 | 0.0521 |
| LIU6 | 0.051 | 0.0565 | 0.0589 | 0.045 | 0.0549 | 0.0577 | 0.0267 | 0.0443 | 0.0493 | 0.0494 |
| LIU7 | 0.0464 | 0.0501 | 0.0509 | 0.0459 | 0.0504 | 0.0493 | 0.0403 | 0.0475 | 0.0484 | 0.0477 |
| LIU8 | 0.0501 | 0.0568 | 0.0603 | 0.0479 | 0.0604 | 0.0664 | 0.0264 | 0.0493 | 0.0594 | 0.053 |
| LIU9 | 0.0483 | 0.0529 | 0.0577 | 0.042 | 0.053 | 0.0565 | 0.0246 | 0.0437 | 0.0501 | 0.0476 |
| LIU10 | 0.0461 | 0.0495 | 0.0505 | 0.0458 | 0.0502 | 0.0494 | 0.0408 | 0.0476 | 0.0483 | 0.0476 |
| MLiu1 | 0.0122 | 0.0305 | 0.048 | 0.0077 | 0.0269 | 0.0456 | 0.0002 | 0.0076 | 0.0264 | 0.0228 |
| MLiu2 | 0.0931 | 0.1137 | 0.1225 | 0.1276 | 0.1637 | 0.1786 | 0.1538 | 0.2176 | 0.2556 | 0.1585 |
| MLiu3 | 0.1165 | 0.1333 | 0.1391 | 0.1688 | 0.2049 | 0.2248 | 0.2128 | 0.2943 | 0.3471 | 0.2046 |
| MLiu4 | 0.0287 | 0.0528 | 0.0737 | 0.0256 | 0.0527 | 0.0726 | 0.0028 | 0.0163 | 0.0381 | 0.0404 |
| MLiu5 | 0.1172 | 0.1346 | 0.1457 | 0.14 | 0.1714 | 0.1873 | 0.1554 | 0.221 | 0.2523 | 0.1694 |

(continued)

**Table 14.6** (continued)

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| MLiu6 | 0.1172 | 0.1343 | 0.1433 | 0.1685 | 0.2067 | 0.2244 | 0.2317 | 0.3154 | 0.3635 | 0.2117 |
| MLiu7 | 0.0809 | 0.1119 | 0.1251 | 0.0512 | 0.0776 | 0.0969 | 0.0052 | 0.0187 | 0.0403 | 0.0675 |
| MLiu8 | 0.1083 | 0.122 | 0.1237 | 0.1219 | 0.1422 | 0.1546 | 0.0867 | 0.1347 | 0.1528 | 0.1274 |
| MLiu9 | 0.1269 | 0.143 | 0.1493 | 0.1898 | 0.2287 | 0.2478 | 0.2391 | 0.3268 | 0.3722 | 0.2248 |
| MLiu10 | 0.0766 | 0.1083 | 0.1241 | 0.0456 | 0.0734 | 0.0957 | 0.0047 | 0.0184 | 0.0396 | 0.0652 |
| KL1 | 0.083 | 0.0957 | 0.0963 | 0.0608 | 0.0708 | 0.0734 | 0.0347 | 0.0468 | 0.0489 | 0.0678 |
| KL2 | 0.0593 | 0.0658 | 0.0668 | 0.0473 | 0.0522 | 0.051 | 0.041 | 0.0478 | 0.0485 | 0.0533 |
| KL3 | 0.0973 | 0.1148 | 0.1157 | 0.109 | 0.1251 | 0.1354 | 0.049 | 0.0741 | 0.0838 | 0.1005 |
| KL4 | 0.083 | 0.0957 | 0.0963 | 0.0608 | 0.0708 | 0.0734 | 0.0347 | 0.0468 | 0.0489 | 0.0678 |
| KL5 | 0.0827 | 0.0959 | 0.0962 | 0.0727 | 0.0871 | 0.0941 | 0.0311 | 0.0486 | 0.0546 | 0.0737 |
| KL6 | 0.0819 | 0.1057 | 0.1143 | 0.0991 | 0.1316 | 0.1525 | 0.0549 | 0.0992 | 0.1272 | 0.1074 |
| JSE | 0.0414 | 0.0458 | 0.0485 | 0.0399 | 0.046 | 0.0464 | 0.0257 | 0.0384 | 0.0439 | 0.0418 |
| YNC1 | 0.0575 | 0.0742 | 0.0791 | 0.1162 | 0.1434 | 0.1597 | 0.2094 | 0.2914 | 0.3272 | 0.162 |
| YNC2 | 0.0652 | 0.0775 | 0.0792 | 0.0886 | 0.0945 | 0.1148 | 0.0333 | 0.0667 | 0.1286 | 0.0832 |
| YNC3 | 0.0426 | 0.0481 | 0.0549 | 0.0602 | 0.0742 | 0.0845 | 0.0651 | 0.0918 | 0.1057 | 0.0697 |
| YNC4 | 0.0403 | 0.0592 | 0.0685 | 0.0522 | 0.0691 | 0.078 | 0.031 | 0.0493 | 0.058 | 0.0562 |
| MRT1 | 0.0633 | 0.0761 | 0.0795 | 0.0972 | 0.1231 | 0.1374 | 0.1423 | 0.2303 | 0.273 | 0.1358 |
| MRT2 | 0.0699 | 0.0819 | 0.0899 | 0.0779 | 0.096 | 0.958 | 0.0286 | 0.0739 | 0.15 | 0.0767 |
| MRT3 | 0.0539 | 0.0577 | 0.0609 | 0.0461 | 0.0545 | 0.0548 | 0.0343 | 0.046 | 0.048 | 0.0507 |
| MRT4 | 0.0559 | 0.0637 | 0.672 | 0.546 | 0.664 | 0.713 | 0.338 | 0.534 | 0.061 | 0.0586 |
| SRE1 | 0.4528 | 0.4612 | 0.4749 | 0.371 | 0.3845 | 0.398 | 0.3287 | 0.3494 | 0.3484 | 0.3965 |
| SRE2 | 0.2694 | 0.2788 | 0.2932 | 0.1725 | 0.1873 | 0.1967 | 0.0979 | 0.1188 | 0.1297 | 0.1938 |
| SRE3 | 0.1745 | 0.1871 | 0.2058 | 0.2895 | 0.3026 | 0.3133 | 0.3027 | 0.3228 | 0.3237 | 0.2691 |
| SRE4 | 0.3633 | 0.4445 | 0.4979 | 0.24 | 0.3419 | 0.4083 | 0.1282 | 0.2471 | 0.324 | 0.3328 |
| SRE5 | 0.0885 | 0.0777 | 0.703 | 0.0942 | 0.0897 | 0.0771 | 0.0805 | 0.0911 | 0.0845 | 0.0837 |
| SRE6 | 0.1492 | 0.1073 | 0.0802 | 0.2263 | 0.1545 | 0.1007 | 0.3746 | 0.2763 | 0.1585 | 0.1808 |
| SRE7 | 0.049 | 0.051 | 0.0505 | 0.049 | 0.0512 | 0.0488 | 0.0379 | 0.0447 | 0.0465 | 0.0476 |
| SRE8 | 0.2655 | 0.2781 | 0.2905 | 0.1723 | 0.1873 | 0.197 | 0.1009 | 0.1207 | 0.13 | 0.1936 |
| SRE9 | 0.2676 | 0.2747 | 0.2922 | 0.1221 | 0.1375 | 0.1491 | 0.0561 | 0.0734 | 0.0814 | 0.1616 |
| SRE10 | 0.3129 | 0.3204 | 0.3367 | 0.1532 | 0.1681 | 0.1789 | 0.0644 | 0.0834 | 0.0938 | 0.1902 |
| SRE11 | 0.4241 | 0.4331 | 0.4415 | 0.2988 | 0.3123 | 0.3214 | 0.1868 | 0.2089 | 0.2145 | 0.3157 |
| SRE12 | 0.4425 | 0.4494 | 0.4567 | 0.2271 | 0.2432 | 0.2544 | 0.0843 | 0.1051 | 0.1155 | 0.2642 |
| SRE13 | 0.2607 | 0.2745 | 0.2867 | 0.1745 | 0.189 | 0.1986 | 0.0979 | 0.1188 | 0.1297 | 0.1923 |
| SKE1 | 0.2607 | 0.2745 | 0.2867 | 0.1745 | 0.189 | 0.1986 | 0.0979 | 0.1188 | 0.1297 | 0.1923 |
| SKE2 | 0.4309 | 0.4627 | 0.4904 | 0.526 | 0.5695 | 0.6027 | 0.5554 | 0.6412 | 0.6838 | 0.551 |

**Table 14.7** Type I error for $\rho = 0.90$ and $\alpha = 0.05$, MF orientation

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| OLS | 0.0412 | 0.0473 | 0.0499 | 0.0484 | 0.0489 | 0.0495 | 0.0445 | 0.0488 | 0.0467 | 0.0472 |
| RE1 | 0.0515 | 0.0639 | 0.0672 | 0.0487 | 0.0545 | 0.0607 | 0.0317 | 0.0427 | 0.048 | 0.0521 |
| RE2 | 0.0419 | 0.05 | 0.0513 | 0.0474 | 0.0492 | 0.0502 | 0.0427 | 0.0478 | 0.0462 | 0.0474 |
| RE3 | 0.0412 | 0.0485 | 0.0507 | 0.0468 | 0.0489 | 0.0505 | 0.0373 | 0.0445 | 0.0456 | 0.046 |
| RE4 | 0.0474 | 0.0593 | 0.0613 | 0.0461 | 0.0502 | 0.053 | 0.0357 | 0.0435 | 0.0454 | 0.0491 |
| RE5 | 0.0561 | 0.0709 | 0.0729 | 0.0565 | 0.0671 | 0.0752 | 0.0323 | 0.0486 | 0.0569 | 0.0596 |
| RE6 | 0.0717 | 0.0851 | 0.0912 | 0.0663 | 0.0855 | 0.0965 | 0.032 | 0.056 | 0.0708 | 0.0728 |
| RE7 | 0.0412 | 0.048 | 0.0549 | 0.0484 | 0.0493 | 0.0514 | 0.0445 | 0.0488 | 0.047 | 0.0482 |
| RE8 | 0.0412 | 0.0483 | 0.0501 | 0.0474 | 0.049 | 0.05 | 0.0428 | 0.0478 | 0.0463 | 0.047 |
| RE9 | 0.0381 | 0.0482 | 0.0499 | 0.0389 | 0.0447 | 0.0492 | 0.0214 | 0.0368 | 0.0438 | 0.0412 |
| RE10 | 0.0412 | 0.0472 | 0.0499 | 0.0482 | 0.0489 | 0.0494 | 0.0443 | 0.0486 | 0.0468 | 0.0472 |
| RE11 | 0.0429 | 0.0505 | 0.0518 | 0.0472 | 0.0492 | 0.0502 | 0.0419 | 0.0475 | 0.0462 | 0.0475 |
| RE12 | 0.0413 | 0.0489 | 0.0507 | 0.0476 | 0.0491 | 0.0503 | 0.0439 | 0.0483 | 0.0467 | 0.0474 |
| RE13 | 0.0415 | 0.0491 | 0.0509 | 0.0475 | 0.049 | 0.0504 | 0.0438 | 0.0483 | 0.0466 | 0.0475 |
| RE14 | 0.0418 | 0.0513 | 0.0523 | 0.0468 | 0.0489 | 0.0505 | 0.0405 | 0.0465 | 0.046 | 0.0472 |
| RE15 | 0.0749 | 0.0938 | 0.0957 | 0.0725 | 0.0878 | 0.101 | 0.0419 | 0.0701 | 0.0865 | 0.0805 |
| RE16 | 0.0527 | 0.0657 | 0.0681 | 0.0461 | 0.0497 | 0.0524 | 0.0387 | 0.0453 | 0.0457 | 0.0516 |
| RE17 | 0.0787 | 0.0961 | 0.0993 | 0.0526 | 0.0667 | 0.0759 | 0.0234 | 0.0375 | 0.0458 | 0.064 |
| RE18 | 0.0495 | 0.0629 | 0.0631 | 0.0472 | 0.0514 | 0.0558 | 0.0383 | 0.0453 | 0.046 | 0.0511 |
| RE19 | 0.0417 | 0.0515 | 0.0523 | 0.0472 | 0.049 | 0.0503 | 0.0431 | 0.0478 | 0.0463 | 0.0477 |
| RE20 | 0.0541 | 0.065 | 0.0671 | 0.0481 | 0.0541 | 0.0611 | 0.0369 | 0.0451 | 0.0454 | 0.053 |
| RE21 | 0.1117 | 0.1332 | 0.1375 | 0.2027 | 0.2474 | 0.2756 | 0.355 | 0.4928 | 0.5712 | 0.2808 |
| RE22 | 0.1141 | 0.1373 | 0.1447 | 0.2109 | 0.2607 | 0.2906 | 0.3676 | 0.5117 | 0.5934 | 0.2923 |
| RE23 | 0.1113 | 0.1421 | 0.1525 | 0.2128 | 0.2741 | 0.3148 | 0.3103 | 0.4869 | 0.6107 | 0.2906 |
| RE24 | 0.1113 | 0.1421 | 0.1525 | 0.2128 | 0.2741 | 0.3148 | 0.3103 | 0.4869 | 0.6107 | 0.2906 |
| RE25 | 0.1118 | 0.1423 | 0.1526 | 0.211 | 0.2726 | 0.3143 | 0.2875 | 0.466 | 0.6002 | 0.2843 |
| LIU1 | 0.0413 | 0.0475 | 0.0502 | 0.0484 | 0.049 | 0.0495 | 0.0445 | 0.0486 | 0.0467 | 0.0473 |
| LIU2 | 0.0399 | 0.0532 | 0.0554 | 0.046 | 0.0533 | 0.0608 | 0.0285 | 0.0427 | 0.0491 | 0.0477 |
| LIU3 | 0.0269 | 0.0344 | 0.038 | 0.0342 | 0.0407 | 0.0462 | 0.0249 | 0.038 | 0.0446 | 0.0364 |
| LIU4 | 0.0411 | 0.0487 | 0.0505 | 0.0469 | 0.0489 | 0.0501 | 0.0419 | 0.0475 | 0.0462 | 0.0469 |
| LIU5 | 0.0513 | 0.0657 | 0.0667 | 0.0526 | 0.0613 | 0.0686 | 0.0288 | 0.045 | 0.0521 | 0.0547 |
| LIU6 | 0.0457 | 0.0575 | 0.0603 | 0.0436 | 0.0484 | 0.0549 | 0.0263 | 0.0395 | 0.0457 | 0.0469 |
| LIU7 | 0.0417 | 0.0493 | 0.0512 | 0.0471 | 0.0489 | 0.0501 | 0.0418 | 0.0474 | 0.0462 | 0.0471 |
| LIU8 | 0.0525 | 0.0674 | 0.0711 | 0.0548 | 0.07 | 0.0787 | 0.0283 | 0.0502 | 0.064 | 0.0597 |
| LIU9 | 0.0427 | 0.0559 | 0.06 | 0.0413 | 0.0485 | 0.0556 | 0.0232 | 0.0383 | 0.0468 | 0.0458 |
| LIU10 | 0.0413 | 0.0489 | 0.0507 | 0.0472 | 0.0489 | 0.0501 | 0.0425 | 0.0477 | 0.0462 | 0.0471 |
| MLiu1 | 0.0039 | 0.0201 | 0.0371 | 0.0019 | 0.0132 | 0.0316 | 0.001 | 0.0032 | 0.0171 | 0.0142 |
| MLiu2 | 0.0507 | 0.0761 | 0.0857 | 0.0865 | 0.1156 | 0.1407 | 0.0927 | 0.1447 | 0.1812 | 0.1082 |
| MLiu3 | 0.0507 | 0.0761 | 0.0853 | 0.0851 | 0.1115 | 0.1375 | 0.0631 | 0.111 | 0.1488 | 0.0966 |
| MLiu4 | 0.0087 | 0.0283 | 0.0438 | 0.0045 | 0.0212 | 0.0394 | 0.0002 | 0.0064 | 0.0224 | 0.0194 |

(continued)

**Table 14.7** (continued)

| *p* | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *n* | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| MLiu5 | 0.1103 | 0.1392 | 0.1448 | 0.1403 | 0.17 | 0.1903 | 0.1326 | 0.1824 | 0.222 | 0.1591 |
| MLiu6 | 0.105 | 0.1327 | 0.14 | 0.1228 | 0.154 | 0.1758 | 0.073 | 0.1134 | 0.1456 | 0.1291 |
| MLiu7 | 0.0371 | 0.064 | 0.0753 | 0.013 | 0.029 | 0.0473 | 0.0004 | 0.0067 | 0.0224 | 0.0328 |
| MLiu8 | 0.1183 | 0.1426 | 0.1437 | 0.1595 | 0.1902 | 0.2121 | 0.1981 | 0.2769 | 0.3224 | 0.196 |
| MLiu9 | 0.1139 | 0.1434 | 0.1505 | 0.1513 | 0.1861 | 0.2129 | 0.1773 | 0.2477 | 0.2947 | 0.1864 |
| MLiu10 | 0.0325 | 0.0602 | 0.076 | 0.0112 | 0.0282 | 0.0468 | 0.0003 | 0.0063 | 0.0224 | 0.0315 |
| KL1 | 0.0824 | 0.1012 | 0.1003 | 0.061 | 0.0669 | 0.0734 | 0.0357 | 0.0437 | 0.0459 | 0.0678 |
| KL2 | 0.0527 | 0.0635 | 0.0655 | 0.0479 | 0.0502 | 0.0511 | 0.0427 | 0.0478 | 0.0463 | 0.052 |
| KL3 | 0.1088 | 0.126 | 0.1265 | 0.1041 | 0.1195 | 0.1298 | 0.0414 | 0.0586 | 0.0678 | 0.0981 |
| KL4 | 0.0824 | 0.1012 | 0.1003 | 0.061 | 0.0669 | 0.0734 | 0.0357 | 0.0437 | 0.0459 | 0.0678 |
| KL5 | 0.0958 | 0.1158 | 0.1153 | 0.0894 | 0.1076 | 0.1196 | 0.0317 | 0.0455 | 0.0529 | 0.086 |
| KL6 | 0.1114 | 0.1379 | 0.145 | 0.1435 | 0.177 | 0.1989 | 0.1328 | 0.1949 | 0.2314 | 0.1636 |
| JSE | 0.0385 | 0.0456 | 0.0489 | 0.0413 | 0.0447 | 0.048 | 0.0276 | 0.0386 | 0.0421 | 0.0417 |
| YNC1 | 0.0557 | 0.0682 | 0.0707 | 0.0999 | 0.1212 | 0.1365 | 0.1895 | 0.2503 | 0.2816 | 0.1415 |
| YNC2 | 0.0728 | 0.1014 | 0.0926 | 0.1062 | 0.1177 | 0.1337 | 0.05 | 0.1 | 0.1 | 0.0972 |
| YNC3 | 0.0397 | 0.0506 | 0.0502 | 0.0532 | 0.0644 | 0.0729 | 0.0556 | 0.0727 | 0.0849 | 0.0605 |
| YNC4 | 0.0294 | 0.0364 | 0.0382 | 0.0239 | 0.0282 | 0.036 | 0.008 | 0.0143 | 0.0201 | 0.0261 |
| MRT1 | 0.0665 | 0.0835 | 0.0872 | 0.1086 | 0.1372 | 0.1516 | 0.1731 | 0.252 | 0.2971 | 0.1508 |
| MRT2 | 0.0863 | 0.0865 | 0.0904 | 0.0721 | 0.1094 | 0.1333 | 0.05 | 0.15 | 0.15 | 0.1031 |
| MRT3 | 0.0481 | 0.0598 | 0.0617 | 0.0463 | 0.0506 | 0.0538 | 0.0353 | 0.0434 | 0.0454 | 0.0494 |
| MRT4 | 0.0561 | 0.0709 | 0.0729 | 0.0565 | 0.0671 | 0.0752 | 0.0323 | 0.0486 | 0.0569 | 0.0596 |
| SRE1 | 0.6243 | 0.6389 | 0.6458 | 0.5557 | 0.5738 | 0.5837 | 0.4973 | 0.5319 | 0.5518 | 0.5781 |
| SRE2 | 0.2941 | 0.3177 | 0.3209 | 0.1801 | 0.1898 | 0.2031 | 0.0992 | 0.1173 | 0.1274 | 0.2055 |
| SRE3 | 0.1923 | 0.2213 | 0.2308 | 0.3088 | 0.3237 | 0.3327 | 0.3102 | 0.3254 | 0.3327 | 0.2864 |
| SRE4 | 0.1814 | 0.2606 | 0.4109 | 0.0798 | 0.1292 | 0.2665 | 0.0334 | 0.0549 | 0.1436 | 0.1734 |
| SRE5 | 0.1067 | 0.1063 | 0.0848 | 0.1168 | 0.1104 | 0.0974 | 0.0898 | 0.1013 | 0.0992 | 0.1014 |
| SRE6 | 0.2777 | 0.1975 | 0.1228 | 0.4069 | 0.2916 | 0.1742 | 0.4978 | 0.4589 | 0.3011 | 0.3032 |
| SRE8 | 0.2903 | 0.3149 | 0.3204 | 0.1782 | 0.189 | 0.2019 | 0.1031 | 0.1187 | 0.128 | 0.2049 |
| SRE9 | 0.2434 | 0.2662 | 0.2755 | 0.1046 | 0.1147 | 0.1263 | 0.0503 | 0.0656 | 0.0724 | 0.1466 |
| SRE10 | 0.2699 | 0.2931 | 0.2981 | 0.1152 | 0.126 | 0.1387 | 0.0531 | 0.069 | 0.0755 | 0.1598 |
| SRE11 | 0.4671 | 0.4851 | 0.483 | 0.3094 | 0.3243 | 0.3334 | 0.183 | 0.2032 | 0.2119 | 0.3334 |
| SRE12 | 0.4695 | 0.4875 | 0.4849 | 0.2219 | 0.2361 | 0.245 | 0.0775 | 0.0961 | 0.1067 | 0.2695 |
| SRE13 | 0.2843 | 0.3101 | 0.3108 | 0.1804 | 0.19 | 0.203 | 0.0992 | 0.1173 | 0.1274 | 0.2025 |
| SKE1 | 0.2843 | 0.3101 | 0.3108 | 0.1804 | 0.19 | 0.203 | 0.0992 | 0.1173 | 0.1274 | 0.2025 |
| SKE2 | 0.2805 | 0.3319 | 0.3568 | 0.3836 | 0.4426 | 0.4886 | 0.46 | 0.5763 | 0.6577 | 0.442 |

**Table 14.8** Type I error for $\rho = 0.99$, and $\alpha = 0.05$ MF orientation

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| OLS | 0.0457 | 0.0475 | 0.0494 | 0.0473 | 0.048 | 0.0508 | 0.045 | 0.0489 | 0.0475 | 0.0478 |
| RE1 | 0.1207 | 0.1399 | 0.1475 | 0.1138 | 0.1457 | 0.1636 | 0.044 | 0.0858 | 0.1089 | 0.1189 |
| RE2 | 0.0481 | 0.0511 | 0.0518 | 0.0464 | 0.0473 | 0.0504 | 0.043 | 0.0481 | 0.047 | 0.0481 |
| RE3 | 0.0459 | 0.0484 | 0.0498 | 0.0452 | 0.0465 | 0.0497 | 0.0367 | 0.0443 | 0.0451 | 0.0457 |
| RE4 | 0.0497 | 0.0528 | 0.0553 | 0.0434 | 0.0457 | 0.0492 | 0.0346 | 0.0428 | 0.0444 | 0.0464 |
| RE5 | 0.0708 | 0.082 | 0.0867 | 0.0568 | 0.0645 | 0.0733 | 0.026 | 0.0392 | 0.0453 | 0.0605 |
| RE6 | 0.1069 | 0.1251 | 0.13 | 0.0963 | 0.1222 | 0.1364 | 0.0326 | 0.0645 | 0.0834 | 0.0997 |
| RE7 | 0.0457 | 0.0475 | 0.0501 | 0.0473 | 0.048 | 0.0509 | 0.045 | 0.0489 | 0.0475 | 0.0479 |
| RE8 | 0.0453 | 0.0471 | 0.0491 | 0.0463 | 0.0473 | 0.0504 | 0.0431 | 0.0482 | 0.047 | 0.0471 |
| RE9 | 0.0449 | 0.0457 | 0.0475 | 0.0487 | 0.0479 | 0.0471 | 0.043 | 0.0488 | 0.0443 | 0.0464 |
| RE10 | 0.0457 | 0.0476 | 0.0494 | 0.0472 | 0.048 | 0.0508 | 0.0448 | 0.0489 | 0.0475 | 0.0478 |
| RE11 | 0.0478 | 0.0517 | 0.0522 | 0.046 | 0.0472 | 0.0502 | 0.0424 | 0.0479 | 0.047 | 0.048 |
| RE12 | 0.046 | 0.0484 | 0.0499 | 0.047 | 0.0476 | 0.0508 | 0.0445 | 0.0487 | 0.0474 | 0.0478 |
| RE13 | 0.046 | 0.0485 | 0.0499 | 0.0469 | 0.0476 | 0.0508 | 0.0445 | 0.0487 | 0.0474 | 0.0478 |
| RE14 | 0.048 | 0.0511 | 0.0525 | 0.0452 | 0.0465 | 0.0497 | 0.0408 | 0.0467 | 0.0463 | 0.0474 |
| RE15 | 0.0571 | 0.0649 | 0.0683 | 0.0419 | 0.0499 | 0.0567 | 0.0186 | 0.0343 | 0.0441 | 0.0484 |
| RE16 | 0.0503 | 0.0537 | 0.0565 | 0.0439 | 0.0459 | 0.0493 | 0.0385 | 0.0455 | 0.0457 | 0.0477 |
| RE17 | 0.1401 | 0.1623 | 0.1657 | 0.1443 | 0.1758 | 0.1904 | 0.0262 | 0.0537 | 0.0699 | 0.1254 |
| RE18 | 0.0599 | 0.0671 | 0.0715 | 0.0453 | 0.048 | 0.0538 | 0.0382 | 0.0457 | 0.0454 | 0.0528 |
| RE19 | 0.0479 | 0.0512 | 0.0523 | 0.0461 | 0.0471 | 0.0502 | 0.0438 | 0.0485 | 0.0471 | 0.0482 |
| RE20 | 0.0459 | 0.055 | 0.0609 | 0.0399 | 0.0447 | 0.0502 | 0.0353 | 0.0435 | 0.0445 | 0.0467 |
| RE21 | 0.1523 | 0.1782 | 0.1835 | 0.2928 | 0.3511 | 0.3832 | 0.5473 | 0.6829 | 0.7489 | 0.3911 |
| RE22 | 0.1495 | 0.177 | 0.1845 | 0.289 | 0.3501 | 0.3852 | 0.5435 | 0.6834 | 0.7512 | 0.3904 |
| RE23 | 0.1328 | 0.1665 | 0.181 | 0.2662 | 0.3395 | 0.3824 | 0.5165 | 0.677 | 0.7528 | 0.3794 |
| RE24 | 0.1328 | 0.1665 | 0.181 | 0.2662 | 0.3395 | 0.3824 | 0.5165 | 0.677 | 0.7528 | 0.3794 |
| RE25 | 0.1347 | 0.1673 | 0.1813 | 0.2675 | 0.34 | 0.383 | 0.5099 | 0.6759 | 0.7529 | 0.3792 |
| LIU1 | 0.0457 | 0.0475 | 0.0494 | 0.0473 | 0.048 | 0.0508 | 0.045 | 0.0489 | 0.0475 | 0.0478 |
| LIU2 | 0.0394 | 0.0448 | 0.0487 | 0.0412 | 0.0462 | 0.0528 | 0.0242 | 0.0376 | 0.0425 | 0.0419 |
| LIU3 | 0.0241 | 0.0275 | 0.0297 | 0.0294 | 0.0332 | 0.0372 | 0.0216 | 0.0355 | 0.0405 | 0.031 |
| LIU4 | 0.0451 | 0.0471 | 0.0491 | 0.046 | 0.0471 | 0.0502 | 0.0424 | 0.0479 | 0.047 | 0.0469 |
| LIU5 | 0.0662 | 0.0743 | 0.0787 | 0.0544 | 0.0619 | 0.0691 | 0.0238 | 0.0378 | 0.0429 | 0.0566 |
| LIU6 | 0.0443 | 0.0497 | 0.0527 | 0.037 | 0.0423 | 0.0471 | 0.0234 | 0.0371 | 0.0413 | 0.0417 |
| LIU7 | 0.0469 | 0.0497 | 0.0512 | 0.046 | 0.0472 | 0.0503 | 0.0424 | 0.0479 | 0.047 | 0.0476 |
| LIU8 | 0.0691 | 0.0833 | 0.0909 | 0.0663 | 0.0804 | 0.0936 | 0.0234 | 0.0432 | 0.0563 | 0.0674 |
| LIU9 | 0.0424 | 0.0486 | 0.0517 | 0.0339 | 0.0406 | 0.0464 | 0.0174 | 0.0322 | 0.0397 | 0.0392 |
| LIU10 | 0.0467 | 0.0498 | 0.0513 | 0.0464 | 0.0474 | 0.0504 | 0.043 | 0.0481 | 0.047 | 0.0478 |
| MLiu1 | 0.0016 | 0.0111 | 0.0286 | 0.0001 | 0.0069 | 0.0234 | 0.0001 | 0.0013 | 0.0127 | 0.0095 |
| MLiu2 | 0.0175 | 0.0303 | 0.0435 | 0.0177 | 0.0335 | 0.0512 | 0.0041 | 0.0147 | 0.0322 | 0.0272 |
| MLiu3 | 0.0103 | 0.018 | 0.0251 | 0.0092 | 0.0205 | 0.0318 | 0.0017 | 0.013 | 0.0297 | 0.0177 |
| MLiu4 | 0.0031 | 0.015 | 0.0326 | 0.0006 | 0.011 | 0.027 | 0.0013 | 0.0028 | 0.0162 | 0.012 |
| MLiu5 | 0.0692 | 0.0887 | 0.0989 | 0.0522 | 0.0689 | 0.0886 | 0.0077 | 0.0193 | 0.0368 | 0.0589 |

(continued)

**Table 14.8** (continued)

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| MLiu6 | 0.0251 | 0.0387 | 0.0525 | 0.0108 | 0.0252 | 0.0386 | 0.0013 | 0.0114 | 0.028 | 0.0257 |
| MLiu7 | 0.0081 | 0.0212 | 0.038 | 0.0008 | 0.0113 | 0.0274 | 0.0013 | 0.0028 | 0.0162 | 0.014 |
| MLiu8 | 0.0869 | 0.1052 | 0.1143 | 0.0968 | 0.1219 | 0.1405 | 0.0526 | 0.0871 | 0.1106 | 0.1018 |
| MLiu9 | 0.0269 | 0.0396 | 0.0528 | 0.0145 | 0.0287 | 0.0414 | 0.0034 | 0.0174 | 0.0332 | 0.0287 |
| MLiu10 | 0.0075 | 0.0206 | 0.0371 | 0.0004 | 0.0109 | 0.0271 | 0.0011 | 0.0025 | 0.0159 | 0.0136 |
| KL1 | 0.0582 | 0.0627 | 0.0645 | 0.0496 | 0.0536 | 0.0564 | 0.0343 | 0.0427 | 0.0442 | 0.0518 |
| KL2 | 0.0486 | 0.0518 | 0.0525 | 0.0464 | 0.0473 | 0.0504 | 0.043 | 0.0481 | 0.047 | 0.0483 |
| KL3 | 0.0836 | 0.0952 | 0.1002 | 0.0641 | 0.0711 | 0.081 | 0.0263 | 0.039 | 0.0438 | 0.0671 |
| KL4 | 0.0582 | 0.0627 | 0.0645 | 0.0496 | 0.0536 | 0.0564 | 0.0343 | 0.0427 | 0.0442 | 0.0518 |
| KL5 | 0.0419 | 0.0501 | 0.059 | 0.0739 | 0.0898 | 0.1043 | 0.039 | 0.0581 | 0.0669 | 0.0648 |
| KL6 | 0.0035 | 0.0142 | 0.0322 | 0.0013 | 0.011 | 0.0266 | 0.0006 | 0.005 | 0.019 | 0.0126 |
| JSE | 0.0439 | 0.0459 | 0.0485 | 0.0407 | 0.0442 | 0.049 | 0.0264 | 0.0383 | 0.0425 | 0.0422 |
| YNC1 | 0.1722 | 0.189 | 0.1861 | 0.1896 | 0.2086 | 0.2148 | 0.1247 | 0.1736 | 0.2032 | 0.1846 |
| YNC2 | 0.053 | 0.0681 | 0.0737 | 0.0671 | 0.0823 | 0.0928 | 0.02 | 0.035 | 0.05 | 0.0563 |
| YNC3 | 0.0233 | 0.0249 | 0.0377 | 0.0197 | 0.0219 | 0.026 | 0.014 | 0.0196 | 0.0215 | 0.0232 |
| YNC4 | 0.0095 | 0.012 | 0.0113 | 0.003 | 0.0042 | 0.0046 | 0.0001 | 0.0001 | 0.0001 | 0.005 |
| MRT1 | 0.082 | 0.1011 | 0.1045 | 0.1119 | 0.1399 | 0.1573 | 0.1534 | 0.2164 | 0.2485 | 0.1461 |
| MRT2 | 0.0574 | 0.0686 | 0.0762 | 0.0733 | 0.0833 | 0.1474 | 0.075 | 0.0875 | 0.1 | 0.0854 |
| MRT3 | 0.0499 | 0.0529 | 0.0554 | 0.0431 | 0.0456 | 0.0492 | 0.034 | 0.0425 | 0.0442 | 0.0463 |
| MRT4 | 0.0708 | 0.082 | 0.0867 | 0.0568 | 0.0645 | 0.0733 | 0.026 | 0.0392 | 0.0453 | 0.0605 |
| SRE1 | 0.8659 | 0.879 | 0.8919 | 0.7792 | 0.8256 | 0.8613 | 0.5574 | 0.6849 | 0.7849 | 0.7922 |
| SRE2 | 0.354 | 0.3701 | 0.3839 | 0.1806 | 0.1953 | 0.2084 | 0.0982 | 0.1198 | 0.1291 | 0.2266 |
| SRE3 | 0.2304 | 0.2516 | 0.2727 | 0.3314 | 0.3439 | 0.355 | 0.3126 | 0.3279 | 0.3372 | 0.3070 |
| SRE4 | 0.0463 | 0.0497 | 0.0569 | 0.042 | 0.0445 | 0.0493 | 0.0296 | 0.0392 | 0.0427 | 0.0445 |
| SRE5 | 0.2504 | 0.2151 | 0.1789 | 0.1668 | 0.174 | 0.168 | 0.0973 | 0.1174 | 0.1248 | 0.1659 |
| SRE6 | 0.7893 | 0.7368 | 0.6139 | 0.7362 | 0.7553 | 0.7126 | 0.5702 | 0.6436 | 0.7242 | 0.698 |
| SRE7 | 0.0499 | 0.0505 | 0.0512 | 0.0496 | 0.0494 | 0.0516 | 0.0393 | 0.0448 | 0.0451 | 0.0479 |
| SRE8 | 0.3398 | 0.3607 | 0.3777 | 0.1807 | 0.1955 | 0.2063 | 0.1017 | 0.1205 | 0.1291 | 0.2236 |
| SRE9 | 0.2387 | 0.258 | 0.2804 | 0.088 | 0.1003 | 0.1091 | 0.0463 | 0.0604 | 0.0673 | 0.1387 |
| SRE10 | 0.2426 | 0.2613 | 0.2847 | 0.0888 | 0.1012 | 0.1105 | 0.0464 | 0.0606 | 0.0676 | 0.1404 |
| SRE11 | 0.5598 | 0.5674 | 0.5757 | 0.3243 | 0.337 | 0.3486 | 0.178 | 0.2012 | 0.2114 | 0.367 |
| SRE12 | 0.5497 | 0.5579 | 0.5657 | 0.2188 | 0.2347 | 0.2477 | 0.0727 | 0.0909 | 0.1017 | 0.2933 |
| SRE13 | 0.2934 | 0.3679 | 0.3816 | 0.1806 | 0.1953 | 0.2084 | 0.0982 | 0.1198 | 0.1291 | 0.2924 |
| SKE1 | 0.3506 | 0.3679 | 0.3816 | 0.1806 | 0.1953 | 0.2084 | 0.0982 | 0.1198 | 0.1291 | 0.2257 |
| SKE2 | 0.0051 | 0.0247 | 0.052 | 0.0031 | 0.0256 | 0.0588 | 0.0008 | 0.0211 | 0.0718 | 0.0292 |

### *14.3.2 Empirical Power of the Test*

#### 14.3.2.1 Simulation Technique for Statistical Power

In this section, we perform a comparative analysis of various test statistics with a focus on their power. Building upon the methodology outlined by Gökpinar and Ebegil (2016), our objective is to evaluate the efficacy of these tests by computing their empirical power. By assessing the ability of each test to correctly reject false null hypotheses, we gain insights into their relative performance and robustness in detecting true effects. This analysis provides valuable information for researchers and practitioners who are seeking to choose the most suitable test statistic for their regression models based on power considerations.

Based on the previous analysis of type I error, we have discarded the tests that significantly exceeded the nominal size of 5%. The remaining test statistics will now be compared in terms of power. To calculate power, we modify the $i$-th component of the $\beta$ vector by replacing it with $Jw(0)\sigma\beta_j$, where $J$ is a positive integer and $w^2(0) = (1 + (p - 2)\rho)/[(1 - \rho)(1 + (p - 1)\rho)]$. We choose $J$ such that for each combination of correlation level and number of predictors the maximum power achieved by the most powerful test is 100%. For this comparison, we select $J = 4$. This procedure allows us to evaluate the relative power of the remaining tests under various conditions, providing insights into their ability to detect true effects in the data.

Utilizing 5000 simulation runs, we calculate the power of these tests by determining the proportion of times the absolute value of the test statistic exceeds the critical value $t_{0.025,(n-p-1)}$. We explore all combinations of sample sizes, with $n = 30, 50,$ and 100, and numbers of regressors, with $p = 3, 5,$ and 10. These computations are conducted across correlation levels of $0.80, 0.90,$ and 0.99, respectively. By systematically varying these parameters, we obtain a comprehensive understanding of the power of each test under different conditions, enabling us to assess their relative effectiveness in detecting true effects in the data.

#### 14.3.2.2 Statistical Power Simulation Result Discussion

Based on the observations from Tables 14.9, 14.10 and 14.11 we note that as the sample size increases while keeping the number of variables constant, the power of the tests considered generally increases, as expected. Additionally, it is apparent that most of the tests exhibit greater power compared to the Wald $z$ test across correlation levels of $0.80, 0.90,$ and 0.99. Among all the estimators, the ridge tests consistently demonstrate superior performance in terms of power compared to the other tests across the simulated conditions. Furthermore, estimators such as KL, Liu, and other two-parameter estimators yield favorable results. Among them, RE15, YNC2, KL6, and SKE2 show greater power than the other estimators. These

**Table 14.9** Power of test for $\rho = 0.80$ and $\alpha = 0.05$

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| OLS | 0.5951 | 0.6149 | 0.6245 | 0.3897 | 0.4104 | 0.4188 | 0.2093 | 0.2325 | 0.2322 | 0.4142 |
| RE1 | 0.7210 | 0.7411 | 0.7481 | 0.4828 | 0.5167 | 0.5312 | 0.2322 | 0.2768 | 0.2895 | 0.5044 |
| RE2 | 0.7221 | 0.7426 | 0.7479 | 0.4562 | 0.4860 | 0.5014 | 0.2208 | 0.2502 | 0.2527 | 0.4867 |
| RE3 | 0.6749 | 0.6962 | 0.7069 | 0.4846 | 0.5194 | 0.5363 | 0.2352 | 0.2796 | 0.2918 | 0.4917 |
| RE4 | 0.8332 | 0.8511 | 0.8601 | 0.5686 | 0.6116 | 0.6291 | 0.2425 | 0.2926 | 0.3096 | 0.5776 |
| RE5 | 0.8807 | 0.8921 | 0.8985 | 0.6679 | 0.7041 | 0.723 | 0.3154 | 0.3912 | 0.4216 | 0.6549 |
| RE7 | 0.6867 | 0.7828 | 0.8291 | 0.4399 | 0.5299 | 0.5913 | 0.2172 | 0.2648 | 0.2912 | 0.5148 |
| RE8 | 0.6582 | 0.6627 | 0.6539 | 0.4328 | 0.4468 | 0.4427 | 0.2193 | 0.2461 | 0.2443 | 0.4452 |
| RE9 | 0.6641 | 0.6647 | 0.6549 | 0.4541 | 0.4616 | 0.4504 | 0.2303 | 0.2647 | 0.2564 | 0.4557 |
| RE10 | 0.6060 | 0.621 | 0.6269 | 0.3964 | 0.4144 | 0.4205 | 0.2116 | 0.2345 | 0.2331 | 0.4183 |
| RE11 | 0.7147 | 0.7407 | 0.747 | 0.4583 | 0.4861 | 0.5014 | 0.2211 | 0.251 | 0.2531 | 0.4859 |
| RE12 | 0.7001 | 0.7255 | 0.733 | 0.4329 | 0.4582 | 0.4702 | 0.2162 | 0.2422 | 0.2429 | 0.469 |
| RE13 | 0.7194 | 0.7432 | 0.7506 | 0.4402 | 0.4661 | 0.48 | 0.2171 | 0.2436 | 0.245 | 0.4784 |
| RE14 | 0.7748 | 0.7976 | 0.8041 | 0.4882 | 0.524 | 0.541 | 0.2265 | 0.2613 | 0.2674 | 0.5205 |
| RE15 | 0.9392 | 0.9475 | 0.9505 | 0.8344 | 0.862 | 0.8743 | 0.5311 | 0.6341 | 0.6816 | 0.8061 |
| RE16 | 0.8716 | 0.8847 | 0.8907 | 0.5624 | 0.6066 | 0.6246 | 0.232 | 0.2735 | 0.2836 | 0.5811 |
| RE18 | 0.8525 | 0.8664 | 0.8727 | 0.5406 | 0.577 | 0.5915 | 0.2297 | 0.2664 | 0.2732 | 0.5633 |
| RE19 | 0.7848 | 0.8076 | 0.8136 | 0.4621 | 0.4925 | 0.5089 | 0.2194 | 0.248 | 0.2493 | 0.5096 |
| RE20 | 0.8931 | 0.9051 | 0.9101 | 0.6131 | 0.6508 | 0.668 | 0.2407 | 0.2831 | 0.2942 | 0.6065 |
| LIU1 | 0.6181 | 0.6384 | 0.6475 | 0.4013 | 0.423 | 0.4318 | 0.2116 | 0.2365 | 0.2359 | 0.4271 |
| LIU2 | 0.8285 | 0.8507 | 0.8578 | 0.6059 | 0.6489 | 0.6708 | 0.2728 | 0.3416 | 0.3691 | 0.6051 |
| LIU3 | 0.7923 | 0.8199 | 0.8301 | 0.5579 | 0.6022 | 0.6229 | 0.2503 | 0.3147 | 0.3401 | 0.57 |
| LIU4 | 0.6247 | 0.6448 | 0.6527 | 0.4144 | 0.4378 | 0.4472 | 0.2171 | 0.2457 | 0.2477 | 0.4369 |
| LIU5 | 0.8329 | 0.8601 | 0.8687 | 0.6126 | 0.6612 | 0.6851 | 0.2785 | 0.3518 | 0.3839 | 0.615 |
| LIU6 | 0.841 | 0.8664 | 0.8809 | 0.6039 | 0.6552 | 0.6791 | 0.2606 | 0.3293 | 0.3594 | 0.6084 |
| LIU7 | 0.6819 | 0.7067 | 0.7168 | 0.4429 | 0.4691 | 0.4857 | 0.2186 | 0.2484 | 0.2506 | 0.469 |
| LIU9 | 0.8597 | 0.8761 | 0.8878 | 0.6323 | 0.6802 | 0.7013 | 0.2765 | 0.355 | 0.389 | 0.6287 |
| LIU10 | 0.6854 | 0.7089 | 0.7185 | 0.441 | 0.4693 | 0.4857 | 0.2183 | 0.2478 | 0.2505 | 0.4695 |
| MLiu1 | 0.003 | 0.0208 | 0.0535 | 0.0021 | 0.0245 | 0.0616 | 0 | 0.0108 | 0.045 | 0.0246 |
| MLiu2 | 0.2552 | 0.2767 | 0.3017 | 0.3627 | 0.3849 | 0.4156 | 0.3828 | 0.4328 | 0.4709 | 0.3648 |
| MLiu3 | 0.2518 | 0.2779 | 0.3002 | 0.3643 | 0.387 | 0.4131 | 0.4373 | 0.4903 | 0.5298 | 0.3835 |
| MLiu4 | 0.0033 | 0.0177 | 0.0429 | 0.0034 | 0.0236 | 0.0524 | 0.0004 | 0.0125 | 0.0427 | 0.0221 |
| MLiu5 | 0.2628 | 0.2793 | 0.2945 | 0.396 | 0.4221 | 0.4481 | 0.4883 | 0.5424 | 0.5782 | 0.4124 |
| MLiu6 | 0.369 | 0.3937 | 0.4135 | 0.4951 | 0.5213 | 0.5487 | 0.5381 | 0.5884 | 0.6249 | 0.4992 |
| MLiu7 | 0.0258 | 0.0476 | 0.0783 | 0.0292 | 0.0583 | 0.089 | 0.0053 | 0.0184 | 0.0502 | 0.0447 |
| MLiu8 | 0.5907 | 0.6084 | 0.624 | 0.8126 | 0.8338 | 0.8542 | 0.6403 | 0.7276 | 0.7693 | 0.7179 |
| MLiu9 | 0.4517 | 0.4859 | 0.5075 | 0.742 | 0.7745 | 0.7976 | 0.8007 | 0.8643 | 0.8899 | 0.7016 |
| MLiu10 | 0.0307 | 0.0497 | 0.0797 | 0.038 | 0.0638 | 0.0915 | 0.0079 | 0.0194 | 0.0512 | 0.048 |
| KL1 | 0.8875 | 0.8983 | 0.9011 | 0.635 | 0.6695 | 0.687 | 0.2495 | 0.3037 | 0.3221 | 0.6171 |
| KL2 | 0.7624 | 0.7798 | 0.7862 | 0.4671 | 0.4994 | 0.5154 | 0.2209 | 0.2503 | 0.2529 | 0.5038 |

(continued)

**Table 14.9** (continued)

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| KL4 | 0.8875 | 0.8983 | 0.9011 | 0.635 | 0.6695 | 0.687 | 0.2495 | 0.3037 | 0.3221 | 0.6171 |
| KL5 | 0.9155 | 0.922 | 0.9261 | 0.7021 | 0.7358 | 0.7483 | 0.2836 | 0.3528 | 0.3792 | 0.6628 |
| KL6 | 0.938 | 0.9407 | 0.9456 | 0.7754 | 0.802 | 0.8133 | 0.352 | 0.4341 | 0.4702 | 0.719 |
| JSE | 0.5451 | 0.5857 | 0.6111 | 0.3164 | 0.3688 | 0.3989 | 0.1233 | 0.1815 | 0.2077 | 0.3709 |
| YNC2 | 0.9763 | 0.9818 | 0.9832 | 0.8906 | 0.9007 | 0.9185 | 0.48 | 0.7000 | 0.6000 | 0.8257 |
| YNC3 | 0.4311 | 0.4571 | 0.4668 | 0.4115 | 0.4493 | 0.4711 | 0.2590 | 0.2976 | 0.3190 | 0.3958 |
| YNC4 | 0.6801 | 0.7276 | 0.7564 | 0.4389 | 0.5053 | 0.5331 | 0.1908 | 0.2565 | 0.2816 | 0.4856 |
| MRT3 | 0.8371 | 0.8536 | 0.8632 | 0.578 | 0.6197 | 0.6393 | 0.2456 | 0.2975 | 0.3156 | 0.5833 |
| MRT4 | 0.8807 | 0.8921 | 0.8985 | 0.6679 | 0.7041 | 0.7230 | 0.3154 | 0.3912 | 0.4216 | 0.6549 |
| SRE4 | 0.7712 | 0.8541 | 0.8882 | 0.5719 | 0.7230 | 0.7809 | 0.2998 | 0.505 | 0.5997 | 0.6660 |
| SRE7 | 0.5818 | 0.6067 | 0.6196 | 0.3583 | 0.3891 | 0.408 | 0.163 | 0.2013 | 0.2158 | 0.3937 |
| SKE2 | 0.9416 | 0.9415 | 0.9437 | 0.8804 | 0.8882 | 0.8868 | 0.753 | 0.783 | 0.7991 | 0.8686 |

**Table 14.10** Power of tests for $\rho = 0.90$ and $\alpha = 0.05$

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| OLS | 0.5924 | 0.6145 | 0.6255 | 0.3861 | 0.4097 | 0.419 | 0.2061 | 0.2263 | 0.2371 | 0.413 |
| RE1 | 0.7575 | 0.7739 | 0.7831 | 0.5218 | 0.5549 | 0.5736 | 0.2364 | 0.2929 | 0.3255 | 0.5355 |
| RE2 | 0.7455 | 0.7663 | 0.7747 | 0.4546 | 0.4854 | 0.5018 | 0.2151 | 0.2406 | 0.2552 | 0.4932 |
| RE3 | 0.6801 | 0.7088 | 0.7209 | 0.4858 | 0.5203 | 0.5404 | 0.2249 | 0.265 | 0.2884 | 0.4927 |
| RE4 | 0.8776 | 0.8883 | 0.8953 | 0.5864 | 0.6302 | 0.6482 | 0.2303 | 0.2762 | 0.3039 | 0.5929 |
| RE5 | 0.9016 | 0.9127 | 0.9183 | 0.6819 | 0.7179 | 0.7358 | 0.304 | 0.3783 | 0.4168 | 0.663 |
| RE7 | 0.5924 | 0.6145 | 0.7432 | 0.3861 | 0.4107 | 0.4867 | 0.2061 | 0.2266 | 0.2488 | 0.435 |
| RE8 | 0.6889 | 0.6953 | 0.6783 | 0.4382 | 0.4583 | 0.4562 | 0.2147 | 0.2391 | 0.2507 | 0.4577 |
| RE9 | 0.7004 | 0.7018 | 0.6813 | 0.4862 | 0.4955 | 0.4754 | 0.2346 | 0.2749 | 0.2776 | 0.4809 |
| RE10 | 0.6032 | 0.6203 | 0.6288 | 0.3927 | 0.4136 | 0.4214 | 0.2084 | 0.228 | 0.2381 | 0.4172 |
| RE11 | 0.7359 | 0.7593 | 0.7725 | 0.4564 | 0.4871 | 0.5004 | 0.2154 | 0.241 | 0.2554 | 0.4915 |
| RE12 | 0.6973 | 0.7247 | 0.7362 | 0.4209 | 0.4489 | 0.459 | 0.2109 | 0.2335 | 0.2449 | 0.464 |
| RE13 | 0.7058 | 0.7345 | 0.7452 | 0.4237 | 0.4524 | 0.4624 | 0.2113 | 0.2341 | 0.2457 | 0.4683 |
| RE14 | 0.8111 | 0.8286 | 0.8394 | 0.4861 | 0.5205 | 0.5407 | 0.2183 | 0.2496 | 0.2671 | 0.529 |
| RE15 | 0.963 | 0.9687 | 0.9722 | 0.8786 | 0.8999 | 0.9063 | 0.5576 | 0.6521 | 0.7014 | 0.8333 |
| RE16 | 0.9038 | 0.9119 | 0.918 | 0.5696 | 0.613 | 0.6321 | 0.2215 | 0.2575 | 0.2791 | 0.5896 |
| RE18 | 0.864 | 0.878 | 0.8857 | 0.5282 | 0.5616 | 0.5784 | 0.2203 | 0.2528 | 0.2715 | 0.5601 |
| RE19 | 0.8128 | 0.8303 | 0.841 | 0.4543 | 0.4848 | 0.4995 | 0.2137 | 0.2382 | 0.2513 | 0.514 |
| RE20 | 0.9077 | 0.9209 | 0.9211 | 0.6043 | 0.6433 | 0.6576 | 0.2276 | 0.2645 | 0.2869 | 0.6038 |
| LIU1 | 0.606 | 0.6287 | 0.6393 | 0.392 | 0.4157 | 0.4264 | 0.2073 | 0.228 | 0.2391 | 0.4203 |
| LIU2 | 0.8343 | 0.8608 | 0.8743 | 0.6086 | 0.6469 | 0.6714 | 0.262 | 0.3271 | 0.3628 | 0.6054 |
| LIU3 | 0.7829 | 0.8169 | 0.8353 | 0.5294 | 0.576 | 0.6004 | 0.2301 | 0.2883 | 0.3215 | 0.5534 |
| LIU4 | 0.6093 | 0.6319 | 0.6427 | 0.4028 | 0.4283 | 0.4377 | 0.2116 | 0.2362 | 0.2495 | 0.4278 |

(continued)

**Table 14.10** (continued)

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| LIU5 | 0.8469 | 0.8721 | 0.888 | 0.6248 | 0.6708 | 0.6956 | 0.2719 | 0.3468 | 0.3875 | 0.6227 |
| LIU6 | 0.8629 | 0.8873 | 0.9029 | 0.6067 | 0.6582 | 0.6865 | 0.2428 | 0.3088 | 0.3473 | 0.6115 |
| LIU7 | 0.6941 | 0.7229 | 0.7362 | 0.4393 | 0.4702 | 0.4834 | 0.2133 | 0.2392 | 0.2532 | 0.4724 |
| LIU9 | 0.901 | 0.9157 | 0.9244 | 0.6765 | 0.7243 | 0.7471 | 0.2796 | 0.3673 | 0.4118 | 0.6609 |
| LIU10 | 0.6985 | 0.7259 | 0.7387 | 0.4368 | 0.4692 | 0.4836 | 0.213 | 0.2391 | 0.253 | 0.4731 |
| MLiu1 | 0.0003 | 0.0184 | 0.0653 | 0.0004 | 0.0242 | 0.0771 | 0.0013 | 0.0087 | 0.0505 | 0.0272 |
| MLiu2 | 0.1451 | 0.1788 | 0.2215 | 0.1994 | 0.225 | 0.2635 | 0.161 | 0.1862 | 0.2188 | 0.1999 |
| MLiu3 | 0.0778 | 0.1105 | 0.1431 | 0.0744 | 0.0989 | 0.1293 | 0.043 | 0.0599 | 0.0879 | 0.0916 |
| MLiu4 | 0.0004 | 0.0181 | 0.0601 | 0.0009 | 0.0251 | 0.0724 | 0.0015 | 0.0119 | 0.0522 | 0.0268 |
| MLiu5 | 0.1472 | 0.1736 | 0.1964 | 0.1996 | 0.2194 | 0.2471 | 0.2033 | 0.2334 | 0.2649 | 0.2094 |
| MLiu6 | 0.1636 | 0.1933 | 0.2174 | 0.1724 | 0.1976 | 0.2255 | 0.098 | 0.1218 | 0.1512 | 0.1712 |
| MLiu7 | 0.0024 | 0.0224 | 0.0624 | 0.0026 | 0.0276 | 0.0736 | 0.0002 | 0.0123 | 0.0522 | 0.0284 |
| MLiu8 | 0.3535 | 0.3814 | 0.4049 | 0.6224 | 0.646 | 0.6567 | 0.769 | 0.8092 | 0.8274 | 0.6078 |
| MLiu9 | 0.2164 | 0.2392 | 0.2513 | 0.2823 | 0.3068 | 0.3331 | 0.3338 | 0.3737 | 0.4019 | 0.3043 |
| MLiu10 | 0.0015 | 0.0223 | 0.0619 | 0.0029 | 0.0276 | 0.0737 | 0.0003 | 0.0123 | 0.0525 | 0.0283 |
| KL1 | 0.9139 | 0.9221 | 0.9251 | 0.6491 | 0.6857 | 0.7015 | 0.2365 | 0.2853 | 0.3137 | 0.6259 |
| KL2 | 0.8004 | 0.8148 | 0.822 | 0.4668 | 0.5 | 0.5186 | 0.2151 | 0.2408 | 0.2552 | 0.5149 |
| KL4 | 0.9139 | 0.9221 | 0.9251 | 0.6491 | 0.6857 | 0.7015 | 0.2365 | 0.2853 | 0.3137 | 0.6259 |
| KL5 | 0.9325 | 0.9414 | 0.9433 | 0.7212 | 0.7537 | 0.7675 | 0.268 | 0.3348 | 0.3705 | 0.6703 |
| KL6 | 0.9507 | 0.9568 | 0.9581 | 0.8009 | 0.825 | 0.8316 | 0.3642 | 0.4417 | 0.4803 | 0.7344 |
| JSE | 0.5132 | 0.5637 | 0.6002 | 0.2841 | 0.3459 | 0.3866 | 0.1034 | 0.1605 | 0.2049 | 0.3514 |
| YNC2 | 0.9982 | 0.9983 | 0.9993 | 0.9426 | 0.9554 | 0.9596 | 0.45 | 0.75 | 0.7 | 0.8616 |
| YNC3 | 0.3422 | 0.3572 | 0.3618 | 0.2922 | 0.3086 | 0.3365 | 0.1923 | 0.2182 | 0.2442 | 0.2948 |
| YNC4 | 0.2342 | 0.2904 | 0.3482 | 0.0964 | 0.112 | 0.1608 | 0.0335 | 0.0486 | 0.0782 | 0.1558 |
| MRT3 | 0.8832 | 0.8943 | 0.8997 | 0.6 | 0.6412 | 0.6612 | 0.2337 | 0.2817 | 0.3101 | 0.6006 |
| MRT4 | 0.9016 | 0.9127 | 0.9183 | 0.6819 | 0.7179 | 0.7358 | 0.304 | 0.3783 | 0.4168 | 0.663 |
| SRE4 | 0.519 | 0.5768 | 0.7919 | 0.2952 | 0.3802 | 0.6086 | 0.1142 | 0.1784 | 0.3511 | 0.4239 |
| SRE7 | 0.5511 | 0.586 | 0.6103 | 0.3255 | 0.3676 | 0.3962 | 0.1411 | 0.1795 | 0.2126 | 0.3744 |
| SKE2 | 0.945 | 0.9479 | 0.9491 | 0.8854 | 0.8956 | 0.8954 | 0.7313 | 0.7805 | 0.8072 | 0.8708 |

findings underscore the effectiveness of ridge regression and related estimators in enhancing the power of hypothesis tests in regression analysis, particularly in the presence of multicollinearity.

**Table 14.11** Power of tests for $\rho = 0.99$ and $\alpha = 0.05$

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| OLS | 0.5962 | 0.6097 | 0.6205 | 0.3866 | 0.4134 | 0.4166 | 0.2076 | 0.2299 | 0.234 | 0.4127 |
| RE1 | 0.8002 | 0.8137 | 0.8225 | 0.6011 | 0.6368 | 0.6511 | 0.3024 | 0.3963 | 0.4431 | 0.6075 |
| RE2 | 0.774 | 0.7919 | 0.7994 | 0.4411 | 0.4748 | 0.485 | 0.2121 | 0.238 | 0.2457 | 0.4958 |
| RE3 | 0.6901 | 0.7147 | 0.7258 | 0.4732 | 0.5112 | 0.5292 | 0.2129 | 0.2509 | 0.2686 | 0.4863 |
| RE4 | 0.9053 | 0.9129 | 0.9181 | 0.5838 | 0.6235 | 0.6418 | 0.2137 | 0.2569 | 0.2787 | 0.5927 |
| RE5 | 0.9175 | 0.921 | 0.9265 | 0.6891 | 0.7207 | 0.7362 | 0.2757 | 0.351 | 0.3892 | 0.6585 |
| RE7 | 0.5962 | 0.6097 | 0.6205 | 0.3866 | 0.4134 | 0.4166 | 0.2076 | 0.2299 | 0.234 | 0.4127 |
| RE8 | 0.7635 | 0.7742 | 0.7683 | 0.4391 | 0.4709 | 0.4758 | 0.2121 | 0.238 | 0.2455 | 0.4875 |
| RE9 | 0.7891 | 0.7945 | 0.7888 | 0.5882 | 0.6114 | 0.6008 | 0.3001 | 0.3747 | 0.3855 | 0.5815 |
| RE10 | 0.6063 | 0.6153 | 0.6227 | 0.3913 | 0.4164 | 0.418 | 0.2094 | 0.2309 | 0.2342 | 0.4161 |
| RE11 | 0.7599 | 0.7837 | 0.7945 | 0.4446 | 0.4758 | 0.4861 | 0.2122 | 0.2383 | 0.2458 | 0.4934 |
| RE12 | 0.6915 | 0.7167 | 0.7275 | 0.4119 | 0.4418 | 0.4465 | 0.2104 | 0.2335 | 0.2385 | 0.4576 |
| RE13 | 0.6923 | 0.7178 | 0.7284 | 0.4121 | 0.442 | 0.4466 | 0.2105 | 0.2336 | 0.2385 | 0.458 |
| RE14 | 0.8441 | 0.8552 | 0.8643 | 0.4703 | 0.5076 | 0.5256 | 0.2121 | 0.2427 | 0.2542 | 0.5307 |
| RE15 | 0.9801 | 0.9803 | 0.9846 | 0.901 | 0.9188 | 0.9238 | 0.5316 | 0.6258 | 0.6737 | 0.8355 |
| RE16 | 0.9243 | 0.9299 | 0.9342 | 0.555 | 0.5969 | 0.6154 | 0.2125 | 0.2467 | 0.2616 | 0.5863 |
| RE18 | 0.3722 | 0.8787 | 0.8877 | 0.5077 | 0.5424 | 0.5562 | 0.2126 | 0.246 | 0.2584 | 0.5513 |
| RE19 | 0.8372 | 0.8488 | 0.8567 | 0.4388 | 0.4703 | 0.482 | 0.2119 | 0.2362 | 0.2429 | 0.5139 |
| RE20 | 0.9192 | 0.923 | 0.9293 | 0.582 | 0.6153 | 0.631 | 0.2147 | 0.2514 | 0.2657 | 0.5924 |
| LIU1 | 0.5975 | 0.6111 | 0.6217 | 0.387 | 0.4138 | 0.4169 | 0.2078 | 0.2299 | 0.234 | 0.4133 |
| LIU2 | 0.7811 | 0.8341 | 0.8642 | 0.5361 | 0.5966 | 0.6288 | 0.2199 | 0.2815 | 0.3142 | 0.5618 |
| LIU3 | 0.6865 | 0.7519 | 0.7932 | 0.4557 | 0.5096 | 0.5393 | 0.1989 | 0.2551 | 0.2829 | 0.497 |
| LIU4 | 0.5984 | 0.6118 | 0.6223 | 0.388 | 0.4155 | 0.4189 | 0.2091 | 0.2318 | 0.236 | 0.4146 |
| LIU5 | 0.7985 | 0.8519 | 0.8837 | 0.5675 | 0.6372 | 0.6745 | 0.2334 | 0.3128 | 0.3555 | 0.5906 |
| LIU6 | 0.8145 | 0.8731 | 0.9032 | 0.5346 | 0.6146 | 0.6579 | 0.2043 | 0.2721 | 0.3084 | 0.5759 |
| LIU7 | 0.6676 | 0.7135 | 0.7403 | 0.4164 | 0.4552 | 0.4676 | 0.2093 | 0.2365 | 0.2443 | 0.4612 |
| LIU9 | 0.8841 | 0.9167 | 0.9359 | 0.658 | 0.7305 | 0.7674 | 0.2502 | 0.3507 | 0.4037 | 0.6552 |
| LIU10 | 0.6728 | 0.7194 | 0.744 | 0.4152 | 0.4544 | 0.4665 | 0.2096 | 0.2364 | 0.2443 | 0.4625 |
| MLiu1 | 0.0001 | 0.0003 | 0.0176 | 0.0013 | 0.0051 | 0.0568 | 0.0015 | 0.0043 | 0.0514 | 0.0151 |
| MLiu2 | 0.0515 | 0.0555 | 0.0873 | 0.0441 | 0.0612 | 0.1149 | 0.0073 | 0.0266 | 0.0702 | 0.0576 |
| MLiu3 | 0.0001 | 0.0004 | 0.0183 | 0.0001 | 0.0138 | 0.0585 | 0.0004 | 0.0203 | 0.0635 | 0.0195 |
| MLiu4 | 0.0001 | 0.0003 | 0.0176 | 0.0011 | 0.0056 | 0.0573 | 0.0015 | 0.0055 | 0.0537 | 0.0156 |
| MLiu5 | 0.0515 | 0.0561 | 0.0855 | 0.0444 | 0.063 | 0.1061 | 0.0081 | 0.0278 | 0.0659 | 0.0565 |
| MLiu6 | 0.0043 | 0.0203 | 0.0587 | 0.008 | 0.0277 | 0.0711 | 0.0008 | 0.022 | 0.0608 | 0.0304 |
| MLiu7 | 0.0001 | 0.0003 | 0.0185 | 0.0011 | 0.0082 | 0.06 | 0.0015 | 0.0081 | 0.0579 | 0.017 |
| MLiu8 | 0.1063 | 0.1237 | 0.1561 | 0.1942 | 0.2285 | 0.2674 | 0.2062 | 0.2483 | 0.2801 | 0.2012 |
| MLiu9 | 0.0117 | 0.0357 | 0.0735 | 0.0161 | 0.0319 | 0.062 | 0.0023 | 0.0185 | 0.0426 | 0.0327 |
| MLiu10 | 0.0001 | 0.0003 | 0.0183 | 0.0011 | 0.0076 | 0.06 | 0.0012 | 0.0079 | 0.0578 | 0.0169 |
| KL1 | 0.9331 | 0.9369 | 0.9409 | 0.6421 | 0.6775 | 0.6928 | 0.2155 | 0.2609 | 0.2836 | 0.6204 |

(continued)

**Table 14.11** (continued)

| $p$ | 3 | | | 5 | | | 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 30 | 50 | 100 | 30 | 50 | 100 | 30 | 50 | 100 | Avg |
| KL2 | 0.8423 | 0.8542 | 0.8575 | 0.4533 | 0.4882 | 0.4997 | 0.2121 | 0.238 | 0.2457 | 0.5212 |
| KL4 | 0.9331 | 0.9369 | 0.9409 | 0.6421 | 0.6775 | 0.6928 | 0.2155 | 0.2609 | 0.2836 | 0.6204 |
| KL5 | 0.9489 | 0.9511 | 0.9544 | 0.7332 | 0.763 | 0.7739 | 0.2429 | 0.3098 | 0.3436 | 0.669 |
| KL6 | 0.9618 | 0.963 | 0.9666 | 0.8237 | 0.8444 | 0.8495 | 0.4735 | 0.5373 | 0.5671 | 0.7763 |
| JSE | 0.2994 | 0.3813 | 0.4816 | 0.1165 | 0.193 | 0.2758 | 0.0254 | 0.07 | 0.1301 | 0.2192 |
| YNC2 | 0.9961 | 0.9963 | 0.9966 | 0.9861 | 0.9863 | 0.9866 | 0.627 | 0.7481 | 0.831 | 0.9118 |
| YNC3 | 0.1265 | 0.1261 | 0.1362 | 0.0988 | 0.1157 | 0.1288 | 0.0636 | 0.0741 | 0.0817 | 0.1057 |
| YNC4 | 0.02 | 0.022 | 0.0259 | 0.0046 | 0.0046 | 0.0054 | 0.0003 | 0.0002 | 0.0003 | 0.0093 |
| MRT3 | 0.9134 | 0.9194 | 0.9241 | 0.6007 | 0.6412 | 0.6581 | 0.215 | 0.2603 | 0.2831 | 0.6017 |
| MRT4 | 0.9175 | 0.921 | 0.9265 | 0.6891 | 0.7207 | 0.7362 | 0.2757 | 0.351 | 0.3892 | 0.6585 |
| SRE4 | 0.3074 | 0.3864 | 0.4836 | 0.1278 | 0.1982 | 0.2772 | 0.0353 | 0.0757 | 0.1317 | 0.2248 |
| SRE7 | 0.3509 | 0.4133 | 0.4943 | 0.1603 | 0.2204 | 0.2874 | 0.0537 | 0.0878 | 0.139 | 0.2452 |
| SKE2 | 0.9469 | 0.952 | 0.9548 | 0.8112 | 0.868 | 0.8934 | 0.4257 | 0.6152 | 0.745 | 0.8014 |

## 14.4 Real-Life Example

In this section, we will consider the following pollution dataset which was originally published by McDonald and Schwing (1973). These data are modeling total age-adjusted mortality rate, which can be obtained for the years 1959–1961 for 201 Standard Metropolitan Statistical Areas (SMSA) from Duffy and Carrol (1967). It has 60 observations and 15 independent variables measuring demographic, socioeconomic, and environmental factors. Independent variables in this dataset exhibit high levels of correlation or multicollinearity. For the multicollinearity, we can use the variance inflation factor (VIF) of the explanatory variables which is given by

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, \quad i = 1, \ldots, p,$$

where $R_i^2$ is the square of the multiple correlation coefficient which we can obtain from regression of the explanatory variable. The VIF values are given in Table 14.12.

Based on the VIF values, we can suggest that there is a dependency among explanatory variables. To detect the multicollinearity, we obtained condition number value calculated as CN = (largest eigenvalue / smallest eigenvalue)$^{1/2}$ =

**Table 14.12** Variance inflation factor

| Variable | VIF | Variable | VIF |
|---|---|---|---|
| PREC | 4.113888 | NONW | 6.779599 |
| JANT | 6.143551 | WWDRK | 2.841582 |
| JULT | 3.967774 | POOR | 8.717068 |
| OVR65 | 7.470045 | HC | 98.63993 |
| POPN | 4.307618 | NOX | 104.9824 |
| EDUC | 4.860538 | SOx | 4.228929 |
| HOUS | 3.994781 | HUMID | 1.907092 |
| DENS | 1.658281 | | |

35406.49. As CN is greater than 10, it indicates severe multicollinearity among the regressors. As we want to test the null hypothesis for the significance of the regression coefficients, we calculate the $P$-values as $p = 2 \times P(t_{n-k-1} > |t_\alpha|)$. In the case of $p < \alpha$, we can reject the null hypothesis. The results are given in Table 14.13.

From Table 14.13, we observed that variables JANT and HUMDID are significant for the estimator RE15 and YNC2 but not for the estimator OLS, and also DENS variable is significant for KL6, while other estimators are not.

## 14.5 Some Concluding Remarks

This chapter explored various estimators derived from existing literature to address the challenge of multicollinearity in testing regression coefficients within the linear regression model. We considered several one-parameter, two-parameter, and Stein-type estimators. Through a simulation study conducted under diverse conditions, we empirically compared these estimators based on their performance in terms of empirical size and power. Our findings indicate that the ridge, Liu, and Kibria-Lukman estimators consistently outperformed other tests, demonstrating their effectiveness in handling multicollinearity and yielding more reliable results.

Additionally, we observed that two-parameter estimators also exhibited favorable performance in this context. RE15, YNC2, KL6, and SKE2 resulted in higher power for the linear regression model than other estimators. These results provide valuable insights into selecting appropriate test statistics for regression coefficient testing in the presence of multicollinearity, contributing to the advancement of regression analysis methodology and more accurate inference for practitioners.

**Table 14.13** Real-life example

| | Coeff | *SE* | t | P-value | Coeff | SE | t | P-value |
|---|---|---|---|---|---|---|---|---|
| Variable | OLS | | | | RE15 | | | |
| PREC | 1.1754 | 1.0602 | 1.1087 | 0.2736 | 1.3743 | 1.0612 | 1.2951 | 0.2021 |
| JANT | −1.5155 | 1.2914 | −1.1735 | 0.2469 | −2.5702 | 1.1062 | −2.3235 | 0.0248 |
| JULT | 1.3186 | 1.8192 | 0.7248 | 0.4724 | 2.4475 | 1.6902 | 1.4481 | 0.1547 |
| OVR65 | 11.1837 | 8.0077 | 1.3966 | 0.1695 | 7.705 | 7.6408 | 1.0084 | 0.3188 |
| POPN | 128.0358 | 45.0068 | 2.8448 | 0.0067 | 69.1871 | 24.7314 | 2.7975 | 0.0076 |
| EDUC | −1.4631 | 13.1124 | −0.1116 | 0.9117 | −2.3923 | 12.4308 | −0.1925 | 0.8483 |
| HOUS | 1.2206 | 1.9964 | 0.6114 | 0.5441 | 2.5345 | 1.801 | 1.4073 | 0.1664 |
| DENS | 0.0075 | 0.0046 | 1.6391 | 0.1083 | 0.007 | 0.0046 | 1.5107 | 0.138 |
| NONW | 4.1296 | 1.5504 | 2.6636 | 0.0108 | 3.9048 | 1.5464 | 2.525 | 0.0152 |
| WWDRK | 0.4473 | 1.936 | 0.2311 | 0.8183 | 0.3524 | 1.9043 | 0.185 | 0.8541 |
| POOR | 1.8858 | 3.7298 | 0.5056 | 0.6157 | 4.0319 | 3.4912 | 1.1549 | 0.2544 |
| HC | −0.3734 | 0.5681 | −0.6572 | 0.5145 | −0.3939 | 0.5758 | −0.6841 | 0.4975 |
| NOX | 0.8744 | 1.1691 | 0.7479 | 0.4585 | 0.9768 | 1.181 | 0.8271 | 0.4126 |
| SOx | 0.1595 | 0.1713 | 0.931 | 0.3569 | 0.1483 | 0.1724 | 0.8605 | 0.3942 |
| HUMID | 1.9152 | 1.2641 | 1.5151 | 0.1369 | 2.7465 | 1.1575 | 2.3728 | 0.0221 |
| | KL6 | | | | SKE2 | | | |
| PREC | 2.2911 | 3.2527 | 0.7044 | 0.4849 | 1.0557 | 2.7683 | 0.3813 | 0.7048 |
| JANT | 3.4481 | 4.1145 | 0.838 | 0.4065 | −1.3611 | 3.2382 | −0.4203 | 0.6763 |
| JULT | 4.6529 | 6.0108 | 0.7741 | 0.443 | 1.1842 | 4.6503 | 0.2546 | 0.8002 |
| OVR65 | −10.5305 | 27.2633 | −0.3863 | 0.7012 | 10.0441 | 20.6744 | 0.4858 | 0.6295 |
| POPN | −127.7511 | 153.5912 | −0.8318 | 0.41 | 114.9897 | 104.9036 | 1.0961 | 0.279 |
| EDUC | 2.2904 | 44.7106 | 0.0512 | 0.9594 | −1.314 | 34.001 | −0.0386 | 0.9693 |
| HOUS | 4.4065 | 6.6485 | 0.6628 | 0.5109 | 1.0962 | 5.0737 | 0.2161 | 0.8299 |
| DENS | 0.0595 | 0.0127 | 4.6899 | 0.001 | 0.0067 | 0.012 | 0.5629 | 0.5764 |
| NONW | −2.4519 | 5.0332 | −0.4871 | 0.6286 | 3.7088 | 4.048 | 0.9162 | 0.3646 |
| WWDRK | 2.7353 | 6.3779 | 0.4289 | 0.6701 | 0.4018 | 5.0469 | 0.0796 | 0.9369 |
| POOR | −0.3614 | 12.5951 | −0.0287 | 0.9772 | 1.6936 | 9.5597 | 0.1772 | 0.8602 |
| HC | −0.0343 | 1.7707 | −0.0194 | 0.9846 | −0.3353 | 1.4874 | −0.2254 | 0.8227 |
| NOX | −0.5661 | 3.6472 | −0.1552 | 0.8774 | 0.7853 | 3.0588 | 0.2567 | 0.7986 |
| SOx | 0.5285 | 0.5131 | 1.03 | 0.3086 | 0.1433 | 0.448 | 0.3198 | 0.7507 |
| HUMID | 2.6535 | 4.0053 | 0.6625 | 0.5111 | 1.7201 | 3.2201 | 0.5342 | 0.5959 |

(continued)

**Table 14.13** (continued)

| | YNC2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Coeff | SE | t | P-value | Coeff | SE | t | P-value |
| PREC | 1.5298 | 1.0676 | 1.4329 | 0.159 | | | | |
| JANT | −3.3118 | 1.0376 | −3.1917 | 0.0026 | | | | |
| JULT | 3.2515 | 1.6685 | 1.9487 | 0.0577 | | | | |
| OVR65 | 5.1695 | 7.3037 | 0.7078 | 0.4828 | | | | |
| POPN | 27.3418 | 10.0513 | 2.7202 | 0.0093 | | | | |
| EDUC | −2.6501 | 10.9125 | −0.2429 | 0.8092 | | | | |
| HOUS | 3.4331 | 1.7133 | 2.0038 | 0.0513 | | | | |
| DENS | 0.0067 | 0.0047 | 1.4141 | 0.1644 | | | | |
| NONW | 3.7309 | 1.5339 | 2.4323 | 0.0191 | | | | |
| WWDRK | 0.2501 | 1.8319 | 0.1365 | 0.892 | | | | |
| POOR | 5.5394 | 3.4183 | 1.6205 | 0.1123 | | | | |
| HC | −0.4057 | 0.5901 | −0.6875 | 0.4954 | | | | |
| NOX | 1.0437 | 1.2049 | 0.8662 | 0.3911 | | | | |
| SOx | 0.1424 | 0.1743 | 0.8172 | 0.4182 | | | | |
| HUMID | 3.3347 | 1.1323 | 2.9451 | 0.0051 | | | | |

**Acknowledgments** The authors are grateful to the reviewers for their constructive comments and suggestions, which have certainly improved the presentation and quality of the paper. Professor B. M. Golam Kibria is grateful to the late Professor A. K. Md. E. Saleh for his encouragement throughout his career, which has greatly contributed to his current position.

## References

Alkhamisi, M. A., & Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics-Theory and Methods, 37*(4), 544–564.

Asar, Y. & Erisoglu, M. (2016). Influence diagnostics in two-parameter ridge regression. *Journal of Data Science, 14*(1), 33–51.

Chand, S., & Kibria, B. M. G. (2024). New ridge type estimator and its performance for the linear regression model: Simulation and application. *Hacettepe Journal of Mathematics & Statistics*, *53*(3), 837–850.

Cule, E., Vineis, P., & De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC Bioinformatics, 12*, 1–15.

Dawoud, I., & Kibria, B. M. G. (2020). A new biased estimator to combat the multicollinearity of the Gaussian linear regression model. *Stats, 3*(4), 526-541.

Dorugade, A. V. (2014). A modified two-parameter estimator in linear regression. *Statistics in Transition. New Series, 15*(1), 23–36.

Dorugade, A. V., & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. *Applied Mathematical Sciences, 4*(9), 447–456.

Duffy, E. A., & Carroll, R. E. (1967). United States Metropolitan Mortality, 1959–1961, PHS Publication No. 999-AP-39, U. S. Public Health Service, National Center for Air Pollution Control.

Gökpinar, E., & Ebegil, M. (2016). A study on tests of hypothesis based on ridge estimator. *Gazi University Journal of Science, 29*(4), 769–781.

Halawa, A. M., & El Bassiouni, M. Y. (2000). Tests of regression coefficients under ridge regression models. *Journal of Statistical Computation and Simulation, 65*(1–4), 341–356.

Hocking, R. R., Speed, F. M., & Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics, 18*(4), 425–437.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics-Theory and Methods, 4*(2), 105–123.

Hoque, M. A., & Kibria, B. M. G. (2023). Some one and two-parameter estimators for the multicollinear Gaussian linear regression model: Simulations and applications. *Surveys in Mathematics and its Applications, 18*, 183–221.

Hoque, M. A., & Kibria, B. M. G. (2024). Performance of some estimators for the multicollinear logistic regression model: Theory, simulation, and applications. *Research in Statistics, 2*(1). https://doi.org/10.1080/27684520.2024.2364747.

Kejian, L. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods, b*(2), 393–402.

Khalaf, G. (2012). A proposed ridge parameter to improve the least square estimator. *Journal of Modern Applied Statistical Methods, 11*(2), 15.

Khalaf, G., & Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics-Theory and Methods, 34*(5), 1177–1182.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation, 32*(2), 419–435.

Kibria, B. M. G., & Banik, S. (2016). Some ridge regression estimators and their performances. *Journal of Modern Applied Statistical Methods, 15*(1), 206–238.

Kibria, G., & Banik, S. (2019). A simulation study on the size and power properties of some ridge regression Tests. *Applications and Applied Mathematics: An International Journal (AAM), 14*(2), 7.

Kibria, B. M. G., Månsson, K., & Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Computational Economics, 40*, 401–414.

Kibria, B. M. G., & Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica, 2020*, 9758378.

Kibria, B. M. G. (2022). More than hundred (100) estimators for estimating the shrinkage parameter in a linear and generalized linear ridge regression model. *Journal of Econometrics and Statistics, 2*(2), 233–252.

Lukman, A. F., & Olatunji, A. (2018). Newly proposed estimator for ridge parameter. an application to the Nigerian economy. *Pakistan Journal of Statistics, 34*(2), 91–98.

Lukman, A. F., Ayinde, K., Binuomote, S., & Clement, O. A. (2019). Modified ridge-type estimator to combat multicollinearity: Application to chemical data. *Journal of Chemometrics, 33*(5), e3125.

Lukman, A. F., Kibria, B. M. G., Ayinde, K., & Jegede, S. L. (2020). Modified one-parameter liu estimator for the linear regression model. *Modeling and Simulation in Engineering, 2020*, 9574304.

Månsson, K., Shukur, G., & Sjölander, P. (2014). A new ridge regression causality test in the presence of multicollinearity. *Communications in Statistics-Theory and Methods, 43*(2), 235–248.

McDonald, G. C., & Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics, 15*(3), 463–481.

Muniz, G., & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics-Simulation and Computation, 38*(3), 621–630.

Perez-Melo, S., & Kibria, B. M. G. (2020). On some test statistics for testing the regression coefficients in presence of multicollinearity: A simulation study. *Stats, 3*(1), 40–55.

Perez-Melo, S., Bursac, Z., & Kibria, B. M. G. (2022). Comparison of test statistics for testing the regression coefficients in the OLS, Ridge, Liu, and Kibria-Lukman linear regression model: A simulation study. In *JSM Proceedings, Biometrics Section*. American Statistical Association (pp. 59–80).

Qasim, M., Amin, M., & Omer, T. (2020). Performance of some new Liu parameters for the linear regression model. *Communications in Statistics-Theory and Methods, 49*(17), 4178–4196.

Saleh, A. K. Md. E., & Kibria, B. M. G. (1993). Performance of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics-Theory and Methods, 22*(10), 2747–2764.

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods, 13*(1), 99–113.

Shabbir, M., Chand, S., & Iqbal, F. (2024). A new ridge estimator for linear regression model with some challenging behavior of error term. *Communications in Statistics-Simulation and Computation, 53*(11), 5442–5452.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability, Volume 1: Contributions to the theory of statistics* (Vol. 3, pp. 197–207). University of California Press.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics-Theory and Methods, 5*(11), 1065–1075.

Thisted, R. A. (1977). *Ridge Regression, Minimax Estimation, and Empirical Bayes Methods*. Stanford University.

Ullah, M. I., Aslam, M., & Altaf, S. (2018). lmridge: A comprehensive R Package for ridge regression. *The R Journal, 10*(2), 326.

Yang, H., & Chang, X. (2010). A new two-parameter estimator in linear regression. *Communications in Statistics-Theory and Methods, 39*(6), 923–934.

# Chapter 15
# Ineffectiveness of Model Selection via t-Test in Regression with Collinearity

**Takeaki Kariya**

## 15.1 Introduction

Within the traditional ordinary least squares (OLS) framework, a linear regression model with

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \quad \text{with} \quad \mathrm{E}[\boldsymbol{u}] = \boldsymbol{0} \quad \text{and} \quad \mathrm{var}(\boldsymbol{u}) = \sigma^2 \boldsymbol{I} \tag{15.1}$$

is treated, where $(\boldsymbol{y}^\circ, \boldsymbol{X})$ is observed in advance. In this chapter, we consider the model and variable selection problem, where a selection procedure picks some $\boldsymbol{x}_j$s from the columns of $\boldsymbol{X} = (x_{nk}) = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K) : N \times K$ with $\boldsymbol{x}_1 = (1, \ldots, 1)^\top \equiv \boldsymbol{e}$ and $\mathrm{rank}(\boldsymbol{X}) = K$ to formulate a submodel $\boldsymbol{X}_j = (\boldsymbol{e}, \boldsymbol{x}_{j_2}, \ldots, \boldsymbol{x}_{j_m})$ and estimate it by the OLS, where the set of all the submodels that contain a constant term is denoted by $[\boldsymbol{X}] = \{\boldsymbol{X}_j\}$. The OLS estimator of $\boldsymbol{X}_j$ is denoted by

$$\hat{\boldsymbol{\beta}}(j) \equiv \hat{\boldsymbol{\beta}}(\boldsymbol{X}_j, \boldsymbol{y}) = \left(\hat{\beta}_k(j)\right) = (\boldsymbol{X}_j^\top \boldsymbol{X}_j)^{-1} \boldsymbol{X}_j^\top \boldsymbol{y}\,.$$

In this framework, we study the effectiveness of variable selection procedure (VSP) via $t$-testing (VSP-T) in view of the effectiveness of the traditional testing framework for VSP or model selection procedure (MSP). In particular, the effect of collinearity is considered in decisions in VSP-T. Furthermore, when $(\boldsymbol{y}^\circ, \boldsymbol{X})$ has been observed, the famous problem on the effectiveness of the VSP in the Neyman-Pearson testing theory (NPTT) is revisited, and it is questioned how submodels

T. Kariya (✉)
Institute of Economic Research, Hitotsubashi University, Kunitachi, Japan
e-mail: thekariya70@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_15

appearing in a repeated VSP can share a common error term $\boldsymbol{u}$ in each step of variable selection, which may be regarded as probability mass decomposition problem according to a significant level (see Sect. 15.5). One of the serious problems is that the order statistic concept is used for complicatedly correlated $t$-values in MSP via VSP-T. This chapter is a continuation of Kariya et al. (2024) where MSP in estimation problem with collinearity is comprehensively treated. Some concepts and results of the paper are adopted here (see Sect. 15.1.1).

Here, MSP and VSP are defined as follows:

(1) MSP is a process where a submodel is selected based on a total model performance measure such as Akaike information criterion (AIC) and adjusted coefficient of determination (CD), among others.
(2) VSP is a process in which each individual variable is selected based on a performance measure of individual variables such as $t$-test statistic and variance inflation factor (VIF) (Jolliffe, 1986). In this case, the total model performance is not carefully considered when each variable is included or excluded. VSP is sometimes included in MSP.

As terminology for VSP and MSP, $\boldsymbol{y}^{\circ}$MSP and $\boldsymbol{X}$MSP, respectively, denote the MSP that uses $(\boldsymbol{y}^{\circ}, \boldsymbol{X})$ and the MSP that uses $\boldsymbol{X}$ only. It is noted that traditional $y^{\circ}$MSP cannot identify and control the ill-effect of collinearity on *individual* OLS estimates, because $\boldsymbol{y}^{\circ}$ itself have no capacity to do it.

In the MSP, when $\boldsymbol{X}_j \in [\boldsymbol{X}]$ is said to be a submodel of the model given by Eq. (15.1), it should carry its own error term $\boldsymbol{u}_j$ as in

$$\boldsymbol{y} = \boldsymbol{X}_j \boldsymbol{\beta}(j) + \boldsymbol{u}_j,$$

which is independent of $\boldsymbol{u}$ in the parent model $\boldsymbol{X}$. In fact, if it would be the same, the two models would be expressed as $\boldsymbol{y} = \boldsymbol{X}_j \boldsymbol{\beta}(j) + \boldsymbol{u}$ and $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ for a common observation $\boldsymbol{y} = \boldsymbol{y}_0$. This implies a contradiction $\boldsymbol{X}_j \boldsymbol{\beta}(j) = \boldsymbol{X}\boldsymbol{\beta}$ unless $\boldsymbol{X}_j = \boldsymbol{X}$. Hence, each submodel carries its own error term in the context of MSP. Hence, two models are compared by AIC or adjusted CD without considering the interdependency of the two models. On the other hand, this problem is not well treated in the case of VSP-T, and it will be discussed in Sects. 15.4 and 15.5. In view of sampling theory, it is often the case that setting the observation $\boldsymbol{y}^{\circ}$ aside, most traditional VSP-T or MSPs ignore a nonlinear and conditional structure in a selected model and so it is out of the OLS framework. In addition, those MSPs mostly aim to select significant variables via testing and do not seek an optimal model because they just ignore the correlation structure in $\boldsymbol{X}$ as a whole.

In this chapter, we basically focus on the effectiveness of VSP-T in view of an effective MSP in the framework of empirically effective modeling in the sense that a selected model via VSP-T needs to have efficiency and stability of each individual OLS estimator, as is required in Kariya et al. (2024).

A brief summary of each Section is as follows:

Section 15.2 Discernability of null and nonnull hypotheses via VSP-T and the effect of collinearity on power function. Measures of discernability are proposed for one-sided and two-sided $t$-tests.

Section 15.3 Decomposition of the OLS estimator and the collinear structure of $X$ in view of MSP. Each $\hat{\beta}_k$ is shown to be not directly affected by the collinearity.

Section 15.4 UMPI tests on the equality of the means of two $t$-test statistics for one-sided and two-sided hypotheses. The correlation of any two statistics is shown to be always nonzero (a.e.) provided one of $\boldsymbol{x}_j$s is a continuous-type variable. This questions about the effectiveness of the deletion rule of variable in VSP-T.

Section 15.5 Some fundamental problems, rooted in the Neyman-Pearson testing theory (NPTT), are revisited in view of the common error term problem in selecting a submodel in VSP-T and nonlinear structure in estimation.

In the presentation of these results, some proofs are referred to the Appendix.

It is remarked that so long as the initial model in Eq. (15.1) is "true" with $\text{rank}(\boldsymbol{X}) = K$, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) in the $K \times K$ nonnegative definite ordering, in which no shrinkage-type estimator will beat the OLS estimator in risk matrix. No effect of collinearity is made on this basic optimality. In fact, similar to the case of BLUE optimality in the OLS estimation, no matter how strong collinearity may be, each individual $t_k$-test and $|t_k|$-test are shown to be respectively uniformly most powerful invariant (UMPI) for the problems (Lehmann & Romano, 2005).

However, strong collinearity makes variances of some *individual* OLS estimators large via the instability of the diagonal elements of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ so that some OLS estimators in a selected model will be instable in terms of standard errors (SEs) and the empirical effectiveness is lost. Draper and Smith (1998) and Belsley (1991) well describe serious ill-effects on individual estimates with empirical examples and provide some diagnostic methods.

### *15.1.1 Empirically Effective Modeling*

In Kariya et al. (2024), the following concept of "empirically effective modeling" (EEM) is defined, and the EEM methodology (EEM-M) is proposed for MSP and VSP. To discuss our problem as well as to introduce the notation, we briefly review their approach to MSP.

The EEM methodology is based on the EEM concept and forms a new integrated model selection process (MSP) consisting of *bundle model* concept (see Sect. 15.5), two different MSPs, new model comparison methodology, and algorithms for selecting inefficiency-controlled and collinearity-controlled models, where the two MSPs are the XMSP using $\boldsymbol{X}$ only and $y^\circ$MSP using $(\boldsymbol{y}^\circ, \boldsymbol{X})$. The whole picture of the EEM-M is symbolically described as an integrated process of the sub-processes

$$\text{(X1)}\ [\boldsymbol{X}] \implies \text{(X2)}\ \boldsymbol{X}\text{MSP}[[\boldsymbol{X}]] \implies \text{(X3)}\ \boldsymbol{y}^{\circ}\text{MSP}(\boldsymbol{y}^{\circ}, [[\boldsymbol{X}]]) \implies \text{(X4) Final Model}$$

where $\boldsymbol{X}$MSP is an MSP based on $\boldsymbol{X}$ only to select a class $[[\boldsymbol{X}]]$ of submodels and $y^{\circ}$MSP is an MSP based on $(\boldsymbol{y}^{\circ}, [[\boldsymbol{X}]])$, while the traditional MSP process is of the form

$$\text{(T1) OLSE via}(\boldsymbol{y}^{\circ}, \boldsymbol{X}) \longrightarrow \text{(T2) MSP\&VSP} \implies \text{(T3) Diagnosis} \implies \text{(T4) Final Model}$$

where there are a variety of criteria in (T2) such as AIC and CD as MSP and such as $t$-testing and VIF for individual variables as VSP. In the traditional approach, procedures (T2) and (T3) are often combined.

In Kariya et al. (2024), the EEM takes the XMSP approach and aims to formulate (X1)–(X2) procedure and derive a subclass $[[\boldsymbol{X}]]$ of empirically effective models (EE-models). It starts with the following concept:

**Definition 15.1** A model estimated by the OLS with data $(\boldsymbol{y}^{\circ}, \boldsymbol{X})$ is defined to be empirically effective if it is judged to hold the two properties, (a) and (b):

(a) Efficiency and stability of each "individual" OLS estimate in the estimated model
(b) Goodness or optimality in total model performance

In Kariya et al. (2024), the stability and efficiency of an estimated predictive model given by

$$\hat{y}_n = \hat{\beta}_1 + x_{n2}\hat{\beta}_2 + \cdots + x_{nK}\hat{\beta}_K \tag{15.2}$$

are considered directly related to that of $\{x_{nk}\hat{\beta}_k\}$s in a pre-sample view of $\hat{y}_n \equiv \hat{y}_n(\boldsymbol{y})$, and the measures $(I_k, C_k)$ of inefficiency and instability concerning the property (a) are selected from the individual predictive sampling variance (IndPSV) of each OLS *estimator* $\hat{\beta}_k$, which is defined by

$$\begin{aligned}\text{IndPSV}_k &\equiv \frac{1}{N}\sum_{n=1}^{N} \text{var}\left(x_{nk}\hat{\beta}_k\right) \\ &= \sigma^2 \frac{1}{N}\left[I_k \times C_k\right] \\ &\equiv \sigma^2 H(I_k, C_k). \end{aligned} \tag{15.3}$$

Here, $I_k$ and $C_k$ are respectively defined for $k \geq 2$ by

$$I_k \equiv \bar{x}_k^2 / s_k^2 \ = \ 1 + (\overline{v_k^2}) = 1 / (1 - q_k^2) \geq 1\,, \tag{15.4}$$

and

$$C_k \equiv \text{VIF}_k = 1 / (1 - \breve{R}_k^2) \geq 1, \tag{15.5}$$

which are, respectively, called *inefficiency index* $I_k$ and *collinearity index* $C_k$. The variables in Eqs. (15.4) and (15.5) are given by

$$v_{nk} = x_{nk} \,/\, s_{xk}, \quad \bar{x}_k = \sum_{n=1}^{N} x_{nk} \,/\, N, \quad \text{and} \quad s_{xk}^2 = \sum_{n=1}^{N} (x_{nk} - \bar{x}_k)^2 \,/\, N.$$

Here, $q_k^2$ is the squared raw correlation of $\boldsymbol{x}_1 = \boldsymbol{e}$ and $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{Nk})^\top$ given by

$$q_k^2 = (\boldsymbol{x}_k^\top \boldsymbol{e})^2 \,/\, \left[\boldsymbol{x}_k^\top \boldsymbol{x}_k \cdot \boldsymbol{e}^\top \boldsymbol{e}\right] = \cos^2(\theta_k), \tag{15.6}$$

where $\theta_k$ is the angle of the two vectors. Also, $\check{R}_k^2$ is the CD when $\boldsymbol{x}_k$ is regressed on $\check{\boldsymbol{X}}_k = \left\{\boldsymbol{x}_j \in \boldsymbol{X},\, j \neq k, k \geq 2\right\}$ in $\mathbb{R}^N$, and $q_k^2$ is a measure about how close $\boldsymbol{e}$ and $\boldsymbol{x}_k$ are. In our terminology, each $\boldsymbol{x}_k$ $(k \geq 2)$ is said to be collinear with $\check{\boldsymbol{X}}_k$ unless $\check{R}_k^2 = 0$ or $\check{R}_k^2 = 1$, and it is said to be perfectly collinear with $\check{\boldsymbol{X}}_k$ if $\check{R}_k^2 = 1$ and zero-collinear with $\check{\boldsymbol{X}}_k$ if $\check{R}_k^2 = 0$:

$$H(I_k, C_k) \equiv \tilde{H}(q_k^2, R_k^2) = 1 \,/\, (1 - q_k^2)(1 - \check{R}_k^2).$$

It is noted that IndPSV in Eq. (15.2) is related with the variance $\text{var}(\hat{\beta}_k)$ of the $k$-th OLS estimator as

$$\text{IndPSV}_k \equiv \sigma^2 \frac{\|\boldsymbol{x}_k\|^2 \,\text{var}(\hat{\beta}_k)}{N}, \tag{15.7}$$

where $\|\boldsymbol{x}_k\|^2 = x_{1k}^2 + \cdots + x_{nk}^2$ and

$$\text{var}(\hat{\beta}_k) = \sigma^2 \frac{1}{N s_{xk}^2 (1 - \check{R}_k^2)} = \frac{\sigma^2}{\text{EEF}_k^2}. \tag{15.8}$$

Here, $[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]_{kk} = 1/\text{EEF}_k^2$ with

$$\text{EEF}_k^2 = N s_{xk}^2 \left(1 - \check{R}_k^2\right) = \|\boldsymbol{x}_k\|^2 \left(1 - q_k^2\right) \left(1 - \check{R}_k^2\right), \tag{15.9}$$

where $\text{EEF}_k$ is called empirically effective factor. The larger EEF is, the smaller the variance $\text{var}(\hat{\beta}_k)$ is for given $\|\boldsymbol{x}_k\|^2$. Based on Eq. (15.8), $\text{VIF}_k = 1/(1 - \check{R}_k^2)$ is often used as a measure for selecting variables to be deleted (see O'Brien, 2007). The instability of the $k$-th OLS estimator via collinearity can be apparently detected by the VIF. However, it does not mean that a variable with the largest $\text{VIF}_k$ is the first variable to be deleted because the collinearity is a confounding relation of the variables in $\boldsymbol{X}$ that cannot be necessarily traced to a single variable. In the sequel, prefix EE- reads empirically effective.

Now, to find a subclass of EE-models in view of the EEM, it is necessary to control the pair $(I_k, C_k)$ $(k \geq 2)$ in such a way that for a pair of pre-specified $c$ and $d$

$$I_k \equiv I_k(p_k \boldsymbol{X}) \leq c \quad \text{and} \quad C_k \equiv (\boldsymbol{X}) \leq d \ (k \geq 2) \tag{15.10}$$

where $p_k \boldsymbol{X} = \boldsymbol{x}_k$, $c \geq 1$ and $d \geq 1$. This is equivalent to

$$q_k^2 \leq c_q = 1 - 1/c \quad \text{and} \quad \check{R}_k^2 \leq d_R = 1 - 1/d \quad (k \geq 2). \tag{15.11}$$

When this holds, $\text{IndPSV}_k \leq cd$ for any $k \geq 2$. Hence, define the set of EE-models at control level $(c, d)$ by

$$D^{\text{cd}} = \left\{ \boldsymbol{X}^* \in [\boldsymbol{X}] | I_k(p_k \boldsymbol{X}^*) \leq c, \ C_k(\boldsymbol{X}^*) \leq d, \ (k = 2, \ldots, J) \right\}. \tag{15.12}$$

Then, any model in this set satisfies the EE-condition or equivalently the condition in Eqs. (15.10) and (15.11) for any $k \geq 2$. Kariya et al. (2024) proposed two algorithms to select models that belong to class $D^{\text{cd}}$. Once such a class $D^{\text{cd}}$ of IC-controlled models is obtained, the second MSP in the EEM-M is a $y^\circ$MSP applied to data $(\boldsymbol{y}^\circ, \boldsymbol{X})$ for the property (b) in order to select an optimal model from class $D^{\text{cd}}$, where an algorithm with AIC, BIC, or adjusted CD with additional conditions such as sign condition may be developed.

It is noted that once $D^{\text{cd}}$ is obtained, an optimal model is obtained via a traditional MSP from the class of submodels

$$\text{RD}^{\text{cd}} = \left\{ \boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{u}_j | \boldsymbol{X}_j \in D^{\text{cd}}, \boldsymbol{u}_j \in [\boldsymbol{u}] \right\},$$

where $[\boldsymbol{u}]$ is the set of mutually independent error terms with $\text{E}[\boldsymbol{u}_j] = \boldsymbol{0}$, $\text{var}(\boldsymbol{u}_j) = \sigma^2 \boldsymbol{I}$ corresponding to each $\boldsymbol{y}_j$ before it is observed. Apply AIC, BIC, adjusted CD, etc. to each model with imposition of some conditions such as sign condition to get a final model.

A question may be raised about the relation of the result with $t$-values. We will consider it in Sect. 15.2.

### *15.1.2 A Brief Review on Literature*

There will be no literature on such an MSP that mitigates the ill-effect within the OLS framework of estimation and testing. However, a lot of hybrid remedies such as shrinkage-type estimation have been provided outside of the OLS framework. Besides, under strong collinearity, $\boldsymbol{y}^\circ$ loses its own capacity to evaluate model performances by an MSP, because OLSEs and maximum likelihood estimates suffer

from the ill-effect due to the instability of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, which invalidates such an MSP as the ones with AIC, BIC, CD, etc.

In view of the traditional OLS framework of testing and MSP, such hybrid methods are not treated: GLS method (Kariya & Kurata, 2003), elastic net-type (Zou & Hastie, 2005), Stein-type (Kubokawa & Srivastava, 2004), preliminary test estimator (Saleh, 2006), and rank-based (Saleh et al., 2022) ones. These methods basically aim to secure the stability of estimates under the presence of collinearity.

But there is not much literature on the effect of collinearity on the effectiveness of MSPs via the $t$-tests in the OLS framework. In fact, this problem has not been paid a careful attention to in the literature regardless of its importance, although in econometrics, environmental sciences, and engineering among others there are many textbooks that describe practical MSPs via testing (e.g., Sen & Srivastava, 1990; Draper & Smith, 1998). These textbooks mainly explain how to practically select models by the MSP via VSP-T but do not much discuss the effectiveness of the MSPs from a viewpoint of the NPTT, though they use the concept of type I and type II errors of test. While Mason and Perreault (1991) examined the collinearity effect on the type I and type II errors of test in their simulation models, using individual VIFs as a VSP, O'Brien (2007) made an overview on this area and discussed about effective thresholds for VIFs though the MSP is not based on $t$-testing but on individual CD $\check{R}_k^2$ of $\boldsymbol{x}_k$. The name of the VIF is said to have been coined by Marquardt (1970).

Fox and Monette (1992) generalized the concept of VIF to the case where $\boldsymbol{X}$ is divided into three categories of variables, among which a set of dummy variables is included. García et al. (2016) discussed the standardization of variables in association with the relationship between VIF and ridge regression. Lavery et al. (2019) studied the relation between the number of variables and collinearity. Aiken and West (1996) treated the OLS estimation via standardized variables. In general, the degree of collinearity increases as the column size in $\boldsymbol{X}$ increases one by one additively.

## 15.2 Effect of Collinearity on VSP-T

In this section, within the framework of VSP-T in the traditional Neyman-Pearson (NP) approach, the effect of collinearity on the effect of the $t$-testing is studied with a focus on the power function and the identifiability of null hypotheses in view of the structural relations of $t$-test statistics. In each stage of VSP-T, variables that are not rejected by $t$-testing are deleted from model, and the acceptance decision of a null hypothesis $\mathcal{H}_0$ is given a positive meaning for exclusion in formulating a model, though the original NPTT scheme does not have such implication. A subtle point lies in the decision scheme that gives this implication to the event that a single observation $\boldsymbol{y}^\circ$ falls into an acceptance region. In fact, when the collinearity is strong, the power of the $t$-test does not increase much in the neighborhood of $\mathcal{H}_0$,

and, hence, the probabilities of acceptance and rejection in the neighborhood are not much different in the power function, though it is uniformly most powerful invariant (UMPI) test. In other words, it is likely to accept each null hypothesis, when the PDF is large. To measure the strong effect on the power function in the neighborhood, the two measures $\mathrm{EEF}_k$ and $(\mathrm{EEF}_k)^2$ are respectively proposed for the effectiveness of one-sided test and two-sided test based on the slopes of the power functions at 0. In addition, various problems are indicated including the identifiability of null hypotheses.

In stepwise MSPs via VSP-T, a decision on the inclusion or exclusion of each variable $\boldsymbol{x}_k$ is frequently made based on observed $t_k$-value, and, hence, it is necessary to check if all the $\mathrm{EEF}_k$s in each stage are reasonably large in their components in Eq. (15.9) as will be discussed here.

### 15.2.1 Basic Setting of t-Testing

As is well known, the $t$-statistic for testing $\mathcal{H}_0 : \beta_k = 0$ is expressed as

$$t_k = \frac{\hat{\beta}_k \mathrm{EEF}_k}{\hat{\sigma}} = \hat{B}_k \, \|x_k\| \left[ (1 - q_k^2) \left( 1 - \check{R}_k^2 \right) \right]^{1/2}, \tag{15.13}$$

where $\mathrm{EEF}_k$ and $q_k^2$ are, respectively, given by Eqs. (15.9) and (15.6) and $\hat{B}_k = \hat{\beta}_k / \hat{\sigma}$ is the core part of $t_k$ carrying the information on $\beta_k$ through $\hat{\beta}_k$. Here, the variance $\sigma^2$ is unbiasedly estimated with the OLS residuals $\{\hat{u}_n\}$ by

$$\hat{\sigma}^2 = \frac{1}{N - K} \sum_{n=1}^{N} \hat{u}_n^2 \quad \text{with} \quad m\hat{\sigma}^2 \sim \sigma^2 \chi_m^2, \tag{15.14}$$

where $\hat{\beta}_k$ and $m\hat{\sigma}^2$ are independent and $\chi_m^2$ denotes chi-square distribution with degrees of freedom (d.f.) $m = N - K$. It follows from Eqs. (15.13) and (15.9) that the $t$-statistic directly depends on

$$\mathrm{EEF}_k = \left( \|\boldsymbol{x}_k\| \, I_k^{-1/2} \right) C_k^{-1/2} = \sqrt{N} s_{xk} \mathrm{PDF}_k \,, \tag{15.15}$$

where $\mathrm{PDF}_k = (1 - \check{R}_k^2)^{1/2} = C_k^{-1/2}$ is "power deflation factor" (PDF) of $\hat{\beta}_k$ due to collinearity. Also, call $I_k^{-1/2} = (1 - q_k^2)^{1/2} \equiv \mathrm{Eff}_k$ the efficiency (Eff) of $\hat{\beta}_k$. This implies that the smaller one of $\mathrm{EFF}_k$ and $\mathrm{PDF}_k$ is for a given $\|\boldsymbol{x}_k\|$, the smaller the $t$-value is, which in turn implies that a model selected by the acceptance of a hypothesis $\mathcal{H}_0 : \beta_k = 0$ may be simply due to the fact that the $\mathrm{EEF}_k$ is small. In addition, $\hat{\beta}_k$ itself is shown to be not structurally dependent on $\check{R}_k^2$ in Sect. 15.3.

In view of testing theory, as is well known, the power function $\pi_1$ (or $\pi_2$ resp.) of each $t$-test for testing

$$\mathcal{H}_0 : \beta_k = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \beta_k > 0 \ \ (\text{or} \quad \mathcal{H}_2 : \beta_2 \neq 0 \ \ \text{resp.}) \tag{15.16}$$

is increasing in $\mu_k$ (or $|\mu_k|$ resp.), where

$$\mu_k = \|\boldsymbol{x}_k\| \, \mathrm{Eff}_k \cdot \mathrm{PDF}_k \, \beta_k \equiv \mathrm{EEF}_k \beta_k. \tag{15.17}$$

Now, let us see how $\mathrm{Eff}_k$ can seriously affect the power performances of these tests. First, for each $k$, the power functions of the UMPI tests $\phi_1$ and $\phi_2$ are respectively given by

$$\pi_1(\mu_k \,/\, \sigma) = \mathrm{E}\,[\phi_1] = \mathrm{E}\,[I(t_k > a_1)] \tag{15.18}$$

$$\pi_2(|\mu_k| \,/\, \sigma) = \mathrm{E}\,[\phi_2] = \mathrm{E}\,[I(|t_k| > a_2)]\,, \tag{15.19}$$

where $I(A_i)$ is an indicator function of critical region $A_i$, $\phi_1 = \phi_1(t_k) = I(t_k > a_1)$ and $\phi_2 = \phi_2(t_k) = I(|t_k| > a_2)$. These tests are UMPI tests, and the power functions are respectively strictly increasing in $\mu_k \,/\, \sigma$ and $|\mu_k| \,/\, \sigma$ as well as in $\beta_k$ and $|\beta_k|$ (Lehmann & Romano, 2005).

Hence, particularly when the null hypothesis $\mathcal{H}_0 : \beta_k = 0$ is accepted by VSP-T, it is necessary to check if $\mathrm{EEF}_k$ for the OLSE $\hat{\beta}_k$ is large enough so that an accepted model via VSP-T is trustworthy. Otherwise, the VSP-T will lead us to the wrong model. As is discussed in Sect. 15.1.1, in estimation, small value of $I_k C_k$ index will give an indication of the effectiveness in terms of $\mathrm{IndPSV}_k$, while in $t$-testing the insensitivity of the tests typically appears in a neighborhood of $\beta_k = 0$. Hence, the local power of the test in a neighborhood of $\mathcal{H}_0 : \beta_k = 0$ is required to be checked in terms of sampling theory. The following proposition gives the two performance measures for VSP-T:

**Proposition 15.1 (Local Performance Measures (LPMs) for VSP-T)** *Let $m = N - K$ and let $a_i$ $(i = 1, 2)$ denote the cutoff point in Eqs.* (15.18) *and* (15.19) *that satisfies the size condition $\pi_i(0) = \alpha$.*

*(1) The slope of the power function of $\phi_1$ at $\beta_k = 0$ is given by*

$$\frac{\partial \pi_1(0)}{\partial \beta_k} = \mathrm{EEF}_k \cdot A_1(a_1, m) \tag{15.20}$$

*with*

$$A_1(a_1, m) = \frac{1}{\sqrt{2\pi}} \left[ \frac{a_1^2}{m} + 1 \right]^{-m/2} < 1.$$

*(2) The first and second derivatives of $\pi_2$ ($\mu_k$ / $\sigma$) at $\beta_k = 0$ are, respectively,* 0 *and*

$$\frac{\partial \pi_2(0)}{\partial \beta_k^2} = (\mathrm{EEF}_k)^2 \, A_2(a_2, m) \tag{15.21}$$

*with*

$$A_2(a_2, m) = \left(\frac{a_2}{\sqrt{m}}\right)\left(1 + \frac{a_2^2}{m}\right)^{-(m+1)/2} \frac{2}{\pi}\left[\Gamma\left(\frac{m+1}{2}\right) / \Gamma\left(\frac{m}{2}\right)\right] < 1.$$

The proof is given in appendix "Proof of Proposition 15.1".

Since the constants $A_1$ and $A_2$ do not depend on the structure of $\boldsymbol{X}$, the local performance measures of one-sided test $\phi_1$ and two-sided test $\phi_2$ in Eqs. (15.20) and (15.21) are, respectively, LPM1 = $\mathrm{EEF}_k$ and LPM2 = $(\mathrm{EEF}_k)^2$. Then, from the structure in Eq. (15.15), to make $\mathrm{EEF}_k$ larger, it is necessary that both $q_k^2$ and $\check{R}_k^2$ are closer to 0 or equivalently to the cases of $\boldsymbol{x}_k^\top \boldsymbol{e} = 0$ and $\boldsymbol{x}_k^\top \check{\boldsymbol{X}}_k = 0$ and that $\|\boldsymbol{x}_k\|$ be larger. However, since $x_{nk}\beta_k = \tilde{x}_{nk}\tilde{\beta}_k$ with $\tilde{x}_{nk} = 50x_{nk}$ and $\tilde{\beta}_k = \beta_k / 50$, $\mathcal{H}_0 : \beta_k = 0$ is not uniquely identified with $\mathcal{H}_0 : \tilde{\beta}_k = 0$ in view of $t$-testing, as will be discussed soon. Hence, the expression of $\|\boldsymbol{x}_k\|$ depends on the choice of scaling in $\beta_k$, and $\|\boldsymbol{x}_k\|$ cannot be compared to $\|\boldsymbol{x}_i\|$ for $k \neq i$ as physical units of these may be different. Nonetheless, it is important to have a model $\boldsymbol{X}$ with $q_k^2$ and $\check{R}_k^2$ smaller.

In particular, the LPM2 = $(\mathrm{EEF}_k)^2$ is of higher order in Eff and PDF than LPM1, which makes the power of $\phi_2$ insensitive in the neighborhood of $\beta_k = 0$. In other words, due to the larger the $\mathrm{PDF}_k$ (or $\mathrm{EFF}_k$), the two-sided $\mathcal{H}_2$ near $\beta_k = 0$ is harder to be distinguished from the null $\mathcal{H}_0$ than the one-sided alternative $\mathcal{H}_1$, and, hence, the two-sided $t$-value is harder to be recognized as significant in such a neighborhood. Hence, in VSP-T, when the realized $t$-value is made smaller by $\mathrm{PDF}_k$ or $\mathrm{EEF}_k$, it is more likely to lead us to no rejection of $\mathcal{H}_0 : \beta_k = 0$ in the decision regardless of values of $\beta_k$, and variable $\boldsymbol{x}_k$ is likely to be excluded from the model. In $t$-testing, whether or not $\mathrm{PDF}_k$ (or $\mathrm{Eff}_k$) may be close to 0, VSP-T will not take into account the effect of $\mathrm{Eff}_k$ though $|t|$-values tend to be small regardless of $\beta_k$, and so VSP-T without use of the information will make it likely to make a wrong decision of accepting $\mathcal{H}_0$ even if $\mathcal{H}_0$ does not hold.

From these facts, we draw some practically important implications:

(1) Since $\mathrm{EEF}_k^2 = \|\boldsymbol{x}_k\|^2 (1 - q_k^2)(1 - \check{R}_k^2)$, the computation of $\|\boldsymbol{x}_k\|$, $(1 - q_k^2)^{1/2}$ and $(1 - \check{R}_k^2)^{1/2}$ will be needed for each stage of VSP-T in order to make a legitimate decision on inclusion or exclusion of variables in the traditional MSP. In addition, it is necessary to compute correlations of $\{t_k\}$ when $\beta_k$s are zero (see Sect. 15.4).

(2) Through $\mu_k = \text{EEF}_k\,\beta_k$, the PDF deflates the power function uniformly. Then, e.g., for the two-sided problem, $\pi_2(|\mu_k|\,/\,\sigma)$ becomes flatter on the horizontal axis of $\beta_k$, and the observed $t$-value is likely to have been unreliably spread out in its realization with $\boldsymbol{y}^\circ$ by the $\text{EEF}_k$ effect. In such a case, each UMPI optimality of $t_k$-test does not guarantee the empirical effectiveness of the selected model, and so the collinearity in the parent model $\boldsymbol{X}$ needs to be controlled before it is used.

(3) The above observation questions about the identifiability of the null hypotheses between $\beta_k = 0$ and $\mu_k = 0$. Although $\text{rank}(\boldsymbol{X}) = K$ excludes the case $\text{EEF}_k = 0$, it can be very close to zero, $\text{EEF}_k \approx 0$, and so $\mu_k \approx 0$. Then, in neighborhood of $\beta_k = 0$, $\pi_2(|\mu_k|\,/\,\sigma)$ becomes close to the size of $\pi_2(0) = \alpha$ even for a relatively large fixed $|\beta_k|(> 0)$, because it loses the ability of detecting $\beta_k \neq 0$. It should be noted that this problem is not solved by replacing the null hypothesis $\beta_k = 0$ by $\mu_k = 0$ because $\mu_k = \text{EEF}_k\beta_k = 0$ cannot be separately distinguished between $\beta_k = 0$ and $\text{EEF}_k \approx 0$ by the $t_k$-test with $\boldsymbol{y}^\circ$. In fact, $\hat{\mu}_k = \text{EEF}_k\hat{\beta}_k \sim N(\mu_k, \sigma^2)$; then, the $t_k$-test statistic becomes $t_k = \hat{\mu}_k\,/\,\hat{\sigma}$, which is the same UMPI test statistic in Eq. (15.15).

## 15.3 Comprehensive Framework for Analyzing Collinearity Effect on Testing

In this section, a comprehensive analytical framework is formulated for studying more on the structure of $t$-statistic and the effect of the EEF on each individual $t$-test. The framework is useful in studying individual sampling properties of each OLSE and its associated $t$-test, where complicated inversions of matrices are avoided once the format is set up.

In particular, using an orthogonal decomposition of $\hat{\boldsymbol{\beta}}$ in Eq. (15.2), each $\hat{\beta}_k$ is shown to be independent of the coordinate system of $\check{\boldsymbol{X}}_k$, implying that $\hat{\boldsymbol{\beta}}_k$ itself does not include the effect of $\check{R}_k^2$ within the estimate. In addition, the set of the OLSEs is shown to be obtained by a simple regression, and various statistics are characterized in our framework including $\check{R}_k^2$, $\text{PDF}_k$, and $t_k$. These will base the analytical results for Sect. 15.4.

### *15.3.1 Simple Regression Models for Individual OLSEs*

For a given matrix $\boldsymbol{W} : N \times L$, let $L(\boldsymbol{W})$ and $L^{\perp}(\boldsymbol{W})$, respectively, denote the linear subspace spanned by the columns of $\boldsymbol{W}$ and its orthogonally complementary subspace in $R^N$. As before, $\check{\boldsymbol{X}}_k = \{\boldsymbol{x}_j \in \boldsymbol{X},\, j \neq k\}$ is interchangeably used as a set of $(K-1)$ vectors or as the $N \times (K-1)$ matrix $(k \neq 1)$. Recall that $\boldsymbol{X}$ and its submodels always include $\boldsymbol{x}_1 \equiv \boldsymbol{e}$.

We begin by summarizing seven facts on the OLSEs, some of which are presented as Proposition 15.2:

(a) For $(\boldsymbol{y}, \boldsymbol{X})$ given, let $\mathbf{M}_X = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\mathbf{M}_{\boldsymbol{e}} = \boldsymbol{e}(\boldsymbol{e}^\top \boldsymbol{e})^{-1}\boldsymbol{e}$,

$$\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}} \tag{15.22}$$

with $\hat{\boldsymbol{y}} = \mathbf{M}_{\boldsymbol{X}}\boldsymbol{y}$ and $\hat{\boldsymbol{u}} = (\boldsymbol{I} - \mathbf{M}_{\boldsymbol{X}})\boldsymbol{y}$ implying $\hat{\boldsymbol{y}} \in L(\boldsymbol{X})$ and $\hat{\boldsymbol{u}} \in L^\perp(\boldsymbol{X})$.

(b) Let $\check{\mathbf{M}}_k = \check{\boldsymbol{X}}_k \left(\check{\boldsymbol{X}}_k^\top \check{\boldsymbol{X}}_k\right)^{-1} \check{\boldsymbol{X}}_k^\top$ $(k = 2, \ldots, K)$. Then,

$$\boldsymbol{y} = \hat{\boldsymbol{y}}_k + \hat{\boldsymbol{u}}_k \quad \text{with} \quad \hat{\boldsymbol{y}}_k = \check{\mathbf{M}}_k \boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{u}}_k = (\boldsymbol{I} - \check{\mathbf{M}}_k)\boldsymbol{y}\,, \tag{15.23}$$

$$\boldsymbol{x}_k = \hat{\boldsymbol{x}}_k + \hat{\boldsymbol{v}}_k \quad \text{with} \quad \hat{\boldsymbol{x}}_k = \check{\mathbf{M}}_k \boldsymbol{x}_k \quad \text{and} \quad \hat{\boldsymbol{v}} = (\boldsymbol{I} - \check{\boldsymbol{M}}_{\boldsymbol{X}})\boldsymbol{x}_k\,, \tag{15.24}$$

implying $\hat{\boldsymbol{y}}_k, \hat{\boldsymbol{x}}_k \in L_k = L(\check{\boldsymbol{X}}_k)$ and $\hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_k \in L_k^\perp = L^\perp(\check{\boldsymbol{X}}_k)$.
Here, for each $k$, $R^N = L_k + L_k^\perp$ is a direct sum in $R^N$ and $L^\perp(\boldsymbol{X}) = \cap_k L^\perp(\check{\boldsymbol{X}}_k)$.

(c) $\check{\boldsymbol{M}}_k \boldsymbol{M}_{\boldsymbol{e}} = \boldsymbol{M}_{\boldsymbol{e}}$, $L(\boldsymbol{X}) \supset L(\check{\boldsymbol{X}}_k) \supset L(\boldsymbol{e})$, and $L^\perp(\boldsymbol{X}) \subset L^\perp(\check{\boldsymbol{X}}_k) \subset L^\perp(\boldsymbol{e})$.
In addition, $\hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{u}} = \boldsymbol{x}_k^\top (\boldsymbol{I} - \check{\boldsymbol{M}}_k)(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{u} = \boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{u} = 0$ for any $k$.

**Proposition 15.2 (Decomposition of Regression Model)**

*(d) For each $k \geq 2$, the OLSE $\hat{\beta}_k$ of $\beta_k$ is expressed as*

$$\hat{\beta}_k = \hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{u}}_k / \hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{v}}_k \sim N\left(\beta_k, \sigma^2 / \hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{v}}_k\right) \tag{15.25}$$

*with*

$$f_{kk} \equiv \hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{v}}_k = N s_{xk}^2 \left(1 - \check{R}_k^2\right) = \mathrm{EEF}_k^2. \tag{15.26}$$

*(e) For each $k \geq 2$, $\hat{\beta}_k$ in Eq. (15.25) is obtained via the simple regression model in $L_k^\perp$;*

$$\hat{\boldsymbol{u}}_k = \hat{\boldsymbol{v}}_k \beta_k + \boldsymbol{w}_k \quad \textit{with} \quad \boldsymbol{w}_k = \left(\boldsymbol{I} - \check{\boldsymbol{M}}_k\right)\boldsymbol{u}. \tag{15.27}$$

*The set of $\hat{\beta}_k$s for $k \geq 2$ in Eqs. (15.25) and (15.26) is equivalent to the whole OLSEs in Eq. (15.2) with*

$$\hat{\beta}_1 = \bar{y} - \bar{x}_2 \hat{\beta}_2 - \bar{x}_3 \hat{\beta}_3 - \cdots - \bar{x}_K \hat{\beta}_K.$$

*(f) The covariance matrix of $\hat{\boldsymbol{u}}_k$ and $\hat{\boldsymbol{u}}_j$ is*

$$\operatorname{cov}\left(\widehat{u}_k, \hat{u}_j\right) = \operatorname{cov}\left(\boldsymbol{w}_k, \boldsymbol{w}_j\right) = \sigma^2 \left(I - \check{\boldsymbol{M}}_k\right)\left(\boldsymbol{I} - \check{\boldsymbol{M}}_j\right), \tag{15.28}$$

*which is singular. In particular, when $\boldsymbol{u}$ follows $N_N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$, $\boldsymbol{w}_k$ and $\boldsymbol{w}_j$ are jointly distributed as $N_{2N}\left(\boldsymbol{0}, \left[\operatorname{cov}\left(\boldsymbol{w}_k, \boldsymbol{w}_j\right)\right]\right)$. In addition,*

$$\operatorname{cov}\left(\hat{\boldsymbol{v}}_k' \hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_j' \hat{\boldsymbol{u}}_j\right) = \hat{\boldsymbol{v}}_k' \operatorname{cov}\left(\boldsymbol{w}_k, \boldsymbol{w}_j\right) \hat{\boldsymbol{v}}_j = \sigma^2 \hat{\boldsymbol{v}}_k' \hat{\boldsymbol{v}}_j \equiv \sigma^2 f_{kj}\,. \tag{15.29}$$

*(g) If $\boldsymbol{u} \sim N_N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$, then*

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K\right)^\top \sim N_K\left(\boldsymbol{\beta}, \sigma^2 \boldsymbol{G}\right) \tag{15.30}$$

*where $\boldsymbol{G} = \left(g_{kj}\right)$ with $g_{kj} = f_{kj}/f_{kk} f_{jj}$ and $m\hat{\sigma}^2 \sim \sigma^2 \chi_m^2$ independently of Eq. (15.29).*

The proof is given in appendix "Proof of Proposition 15.2".

**Proposition 15.3** *Each estimator $\hat{\beta}_k$ and the residual variance $\hat{\sigma}^2$ are independent of the coordinates $\check{\boldsymbol{X}}_k$ in $L_k = L\left(\check{\boldsymbol{X}}_k\right)$.*

***Proof*** In fact, since both $\hat{\boldsymbol{u}}_k$ and $\hat{\boldsymbol{v}}_k$ belong to $L_k^\perp\left(\check{\boldsymbol{X}}_k\right)$, $\hat{\beta}_k$ in Eq. (15.27) is independent of $L_k = L\left(\check{\boldsymbol{X}}_k\right)$, and so is $\hat{\sigma}^2$ since $\hat{\boldsymbol{u}} \in L^\perp(\boldsymbol{X}) \subset L_k^\perp\left(\check{\boldsymbol{X}}_k\right)$ for any $k$. Consequently, each $\hat{\beta}_k$ is a function of $\hat{\boldsymbol{u}}_k$ and $\hat{\boldsymbol{v}}_k$ only that belong to $L^\perp\left(\check{\boldsymbol{X}}_k\right)$. Here, both $\hat{\boldsymbol{u}}_k$ and $\hat{\boldsymbol{v}}_k$ are linearly independent of $\check{\boldsymbol{X}}_k$, and $\hat{\boldsymbol{u}}_k$ is dependent on $\boldsymbol{x}_k$ through Eq. (15.27) if $\beta_k \neq 0$.

Therefore, $\hat{\beta}_k$ via Eq. (15.27) is the OLSE in the space $L_k^\perp$, implying the information carrying core-part $\hat{B}_k$ in Eq. (15.13) gets the information on $\beta_k$ only though $\hat{\boldsymbol{u}}_k$, implying that the information on $\beta_k$ is captured in $\hat{\boldsymbol{u}}_k$. □

In particular, for testing $\beta_k = 0$, it follows directly from Proposition 15.2 (d) that the $t$-statistic in Eq. (15.13) is expressed as

$$t_k = \frac{\hat{\beta}_k \sqrt{f_{kk}}}{\hat{\sigma}} = \frac{\hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{u}}_k}{\hat{\sigma}\sqrt{f_{kk}}} = \frac{\hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{u}}_k}{\hat{\sigma}\sqrt{\hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{v}}_k}}\,. \tag{15.31}$$

When an inference on $\left(\beta_k, \beta_j\right)$ is concerned, it suffices to focus on $\left(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{v}}_i\right)$ $(i = k, j)$ with its covariance in Eq. (15.29) by the principle of invariance. Similarly, for testing $\beta_k = \beta_i$, $\left(\hat{\boldsymbol{u}}_k, \hat{\boldsymbol{u}}_j, \hat{\sigma}\right)$ is an invariantly sufficient statistic, the fact of which will be used in Sect. 15.4. In particular, this result is used in testing the equality of the means of two $t$-test statistics.

**Proposition 15.4** *The following results are important in our context:*

*(1) The means of the elements of* $\hat{\boldsymbol{u}}_k$ *and* $\hat{\boldsymbol{v}}_k$ *are both 0, for* $\boldsymbol{e}^\top \hat{\boldsymbol{u}}_k = 0$ *and* $\boldsymbol{e}^\top \hat{\boldsymbol{v}}_k = 0$.
*(2) If* $\hat{\boldsymbol{v}}_k = 0$, *then* $\boldsymbol{x}_k = \hat{\boldsymbol{x}}_k \in L_k$, *implying the perfect collinearity with* $\check{R}_k^2 = 1$.
*(3) If* $\hat{\boldsymbol{v}}_k = \boldsymbol{x}_k (\neq \boldsymbol{0})$, $\boldsymbol{x}_k$ *is linearly independent of* $\check{\boldsymbol{X}}_k$, *and* $\check{R}_k^2 = 0$.

### 15.3.2 Collinearity Effect on t-Test

Now, to express $t_k$ in terms of the variables introduced in Sect. 15.3.1, note that $t_k$ in Eq. (15.25) consists of $\left(\hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_k, \hat{\sigma} : f_{kk}\right)$ $(k = 2, \ldots, K)$ with $f_{kk}$ in Eq. (15.26). First, let us express $\check{R}_k^2$ and $\mathrm{PDF}_k$ in these terms. By its definition and by Eq. (15.26),

$$\check{R}_k^2 = \frac{s_{\hat{x}k}^2}{s_{xk}^2} = \frac{\hat{\boldsymbol{x}}_k^\top (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{e}}) \hat{\boldsymbol{x}}_k}{\boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{e}}) \boldsymbol{x}_k} = \frac{\boldsymbol{x}_k^\top \left(\check{\boldsymbol{M}}_k - \boldsymbol{M}_{\boldsymbol{e}}\right) \boldsymbol{x}_k}{\boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{e}}) \boldsymbol{x}_k} \tag{15.32}$$

$$\mathrm{PDF}_k^2 = \hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{v}}_k / N s_{xk}^2 = \boldsymbol{x}_k^\top \left(\boldsymbol{I} - \check{\boldsymbol{M}}_k\right) \boldsymbol{x}_k / \boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{e}}) \boldsymbol{x}_k \,. \tag{15.33}$$

It is noted that $\boldsymbol{x}_k^\top \left(\boldsymbol{I} - \check{\boldsymbol{M}}_k\right) \boldsymbol{x}_k + \boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{e}}) \boldsymbol{x}_k = \boldsymbol{x}_k^\top \left(\check{\boldsymbol{M}}_k - \boldsymbol{M}_{\boldsymbol{e}}\right) \boldsymbol{x}_k$.

**Theorem 15.1** *In the decomposition of* $\boldsymbol{x}_k = \hat{\boldsymbol{v}}_k + \hat{\boldsymbol{x}}_k$ *in Eq.* (15.24), $\hat{\boldsymbol{x}}_k = \check{\boldsymbol{M}}_k \boldsymbol{x}_k = \bar{x}_k \boldsymbol{e} \in L(\boldsymbol{e})$ *is a necessary and sufficient condition for zero collinearity* $\mathrm{PDF}_k = 1$. *In particular, when* $\boldsymbol{x}_k = \hat{\boldsymbol{v}}_k$, *then* $\mathrm{PDF}_k = 1$, *and so* $\boldsymbol{x}_k = \hat{\boldsymbol{v}}_k$ *is zero-collinear with* $\check{\boldsymbol{X}}_k$.

***Proof*** By Eq. (15.33), $\mathrm{PDF}_K = 1$ if and only if $\boldsymbol{x}_k^\top \left(\boldsymbol{I} - \check{\boldsymbol{M}}_k\right) \boldsymbol{x}_k = \boldsymbol{x}_k^\top \left(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{e}}\right) \boldsymbol{x}_k$, which in turn holds if and only if $\boldsymbol{x}_k^\top \left(\check{\boldsymbol{M}}_k - \boldsymbol{M}_{\boldsymbol{e}}\right) \boldsymbol{x}_k = 0$. Noting $\left(\check{\boldsymbol{M}}_k - \boldsymbol{M}_{\boldsymbol{e}}\right)^2 = \left(\check{\boldsymbol{M}}_k - \boldsymbol{M}_{\boldsymbol{e}}\right)$, this condition is equivalent to $\left(\check{\boldsymbol{M}}_k - \boldsymbol{M}_{\boldsymbol{e}}\right) \boldsymbol{x}_k = 0$. This holds if and only if $\check{\boldsymbol{M}}_k \boldsymbol{x}_k = \boldsymbol{M}_{\boldsymbol{e}} \boldsymbol{x}_k = \boldsymbol{e} \bar{x}_k$. Hence, the result follows: When $\bar{x}_k = 0$ in $\hat{\boldsymbol{x}}_k = \bar{x}_k \boldsymbol{e}$, then $\boldsymbol{x}_k = \hat{\boldsymbol{v}}_k$ since $\boldsymbol{x}_k = \hat{\boldsymbol{v}}_k + \check{\boldsymbol{M}}_k \boldsymbol{x}_k$, or equivalently $\boldsymbol{x}_k$ is orthogonal to $\check{\boldsymbol{X}}_k$ and so linearly independent of $\check{\boldsymbol{X}}_k$, implying zero collinearity. □

It is interesting to know that no collinearity effect on the $k$-th variable exists if and if $\hat{\boldsymbol{x}}_k \in L(\boldsymbol{e})$ or equivalently $\hat{\boldsymbol{x}}_k = \check{\boldsymbol{M}}_k \boldsymbol{x}_k = \boldsymbol{e} \bar{x}_k = \boldsymbol{M}_{\boldsymbol{e}} \boldsymbol{x}_k$, which holds if and only if $\boldsymbol{x}_k^\top \boldsymbol{x}_j = 0$ for $j \neq k$ and $j \geq 2$.

## 15.4 Test for Comparability of $t$-Values and Correlation Structure of $t$-Statistics

In this section, a fundamental question about the comparability of $t$-values in a regression model is raised from a theoretical standpoint, which seems a critical problem for stepwise MSPs via VSP-T by ordering $t$- (for equivalently partial $F$-) values. Here, in addition, we raise the correlation problem among $t$-values. In a stepwise MSP, it is often the case that in each step after addition or deletion of a variable, observed $\boldsymbol{y}^0$ is regressed on the remaining variables in the next step, and, again, the explanatory variables are ordered by the size of the $|t|$-values, and then the variable that has the least insignificant $|t|$-value is usually deleted. In such a case, if there are correlations in the $t$-values, each step of this process must concern the correlations among $t$-values, while the collinearity effect on $t$-values via PDFs needs to be treated as well, described in Sect. 15.2. In this section, the correlation problem is considered, and the problem of the collinearity effect on the correlations is also discussed.

To find an explicit correlation coefficient of two $t$-statistics $\left(t_k, t_j\right)$, first, let us consider the situation of empirical decision-making in variable selection.

Both two-sided $t$-values of the $k$-th and $j$-th variables are close to each other around 2 with $t_j = 1.8 < 2 < t_k = 2.2$ and $m > 25$ where $\alpha = 0.05$. (15.34)

In such a case, in VSP-T, it is often suggested that $\boldsymbol{x}_j$ be deleted from the model because $\boldsymbol{x}_j$ has the smaller $t$-value without paying any attention to the collinearity factors and to the fact that $t_k$ and $t_j$ were simultaneously and stochastically realized in a correlated manner through a realization of $\boldsymbol{y}$ where $\boldsymbol{X}$ is given. Since the correlations are determined by $\boldsymbol{X}$ as will be shown soon, it is not easy to treat the situation logically with $t$-values without knowing the correlation.

### *15.4.1 Distribution of $\left(t_k, t_j\right)$*

Now, let us consider the structure of the distribution of $T \equiv (t_2, t_3, \ldots, t_K)$. For simplicity, it suffices to treat the pairwise case of $\left(t_k, t_j\right)$.

Let $f_{kj} = \hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{v}}_j$ as in Eq. (15.29) and

$$h_{kj} = \frac{f_{kj}}{\sqrt{f_{kk} f_{jj}}} \quad \text{for any} \quad k, j = 2, \ldots, K. \tag{15.35}$$

Let $\mathbf{H} = \left(h_{kj}\right)$ with $h_{ii} = 1$ $(i = k, j)$, which is a correlation matrix, since

$$\left\| \hat{\boldsymbol{v}}_k / \sqrt{f_{kk}} \right\| = 1 \quad \text{and} \quad \left| f_{kj} \right|^2 \leq f_{kk} f_{jj}. \tag{15.36}$$

**Lemma 15.1 (Decomposition of the Joint Distribution of $\left(t_k, t_j\right)$ via Conditioning)**

*(A) The conditional distribution of $\left(t_k, t_j\right)$ given $\hat{\sigma}^2$ is given by the conditional normal distribution,*

$$\begin{pmatrix} t_k \\ t_j \end{pmatrix} = \begin{pmatrix} \dfrac{\hat{\beta}_k \mathrm{EEF}_k}{\sigma} \\ \dfrac{\hat{\beta}_j \mathrm{EEF}_j}{\sigma} \end{pmatrix} = \begin{pmatrix} \dfrac{\hat{\beta}_k \sqrt{f_{kk}}}{\sigma} \\ \dfrac{\hat{\beta}_j \sqrt{f_{jj}}}{\hat{\sigma}} \end{pmatrix} \sim N \left( \begin{pmatrix} \dfrac{\beta_k \sqrt{f_{kk}}}{\hat{\sigma}} \\ \dfrac{\beta_j \sqrt{f_{jj}}}{\hat{\sigma}} \end{pmatrix}, \frac{\sigma^2}{\hat{\sigma}^2} \mathbf{H} \right), \tag{15.37}$$

*where the conditional covariance matrix is $\left(\sigma^2/\hat{\sigma}^2\right)\mathbf{H}$. And $z \equiv m\hat{\sigma}^2/\sigma^2$ is distributed as $\chi_m^2$ independent of $(\hat{\beta}_k, \hat{\beta}_j)$. It is noted that the fact in Eq. (15.37) is straightforwardly extended to the general case $(t_2, t_3, \ldots, t_K)$ in an obvious manner.*

*(B) The conditional correlation of $t_k$ and $t_j$ given $\hat{\sigma}^2$ is $h_{kj} = f_{kj}/\sqrt{f_{kk} f_{jj}}$ independent of any unknown parameter and $\hat{\sigma}^2$, and it satisfies $h_{kj} < 1$. When the joint hypothesis $\beta_k = \beta_j = 0$ holds, $h_{kj}$ is the unconditional correlation coefficient.*

***Proof*** The result (A) follows from Proposition 15.2 (g) by a conditional argument. The first part of (B) follows from (A). To show $h_{kj} < 1$, it is easy to see that $h_{kj} = 1$ if and only if $\hat{\boldsymbol{v}}_k = a\hat{\boldsymbol{v}}_j$ for some $a \neq 0$. We show that this cannot happen. By the definitions in (a) and (b) in Sect. 15.3, the decomposition $\boldsymbol{x}_k = \hat{\boldsymbol{x}}_k + \hat{\boldsymbol{v}}_k$ corresponds to the direct sum $L(\boldsymbol{X}) = L\left(\check{\boldsymbol{X}}_k\right) + L\left(\hat{\boldsymbol{v}}_k\right)$ in $R^N$. Here, $\check{\boldsymbol{X}}_k = \{\boldsymbol{x}_i \in \boldsymbol{X}, i \neq k\}$ implies $\hat{\boldsymbol{v}}_k = \left(\boldsymbol{I} - \check{\boldsymbol{M}}_k\right)\boldsymbol{x}_k \notin L\left(\check{\boldsymbol{X}}_k\right)$. Hence, since $\hat{\boldsymbol{v}}_j = \left(\boldsymbol{I} - \check{\boldsymbol{M}}_j\right)\boldsymbol{x}_j \in L\left(\check{\boldsymbol{X}}_k\right)$, if $\hat{\boldsymbol{v}}_k = a\hat{\boldsymbol{v}}_j$, then $\boldsymbol{x}_k = \hat{\boldsymbol{x}}_k + a\hat{\boldsymbol{v}}_j \in L\left(\check{\boldsymbol{X}}_k\right)$. This contradicts $\mathrm{rank}(\boldsymbol{X}) = K$. This completes the proof. □

The unconditional distribution of the conditional multivariate normal variates in Eq. (15.37) is known as multivariate noncentral $t$-distribution (see Johnson & Kotz 1972).

**Means, Covariances, and Correlations**

To investigate the collinearity effect on the unconditional means, variance, and correlations of $t_k$ and $t_j$, let $z = m\hat{\sigma}^2/\sigma^2 \sim \chi_m^2$ with d.f. $m$, which is independent

of all $\hat{\beta}_k$s. Let $\mu_k = \sqrt{f_{kk}}\beta_k$ with $\sqrt{f_{kk}} = \sqrt{N}s_{xk}\mathrm{PDF}_k$, and let

$$c_m = \mathrm{E}\left[z^{-1/2}\right] = 2^{-\frac{1}{2}}\Gamma\left(\frac{m-1}{2}\right)/\Gamma\left(\frac{m}{2}\right) \quad \text{and} \quad d_m = \sqrt{m-2}\,c_m. \tag{15.38}$$

**Lemma 15.2 (Unconditional Mean, Variance, Covariance, and Correlation)** *Mean:*

$$\mathrm{E}[t_k] = \left(\frac{\sqrt{m}}{\sigma}\right) c_m \mu_k \tag{15.39}$$

*Variance and Covariance:*

$$\mathrm{cov}\left(t_k, t_j\right) = \frac{m}{m-2}\left\{h_{kj} + \frac{\mu_k \mu_j}{\sigma^2}\left[1 - d_m^2\right]\right\} \tag{15.40}$$

*Correlation: for each pair* $\left(t_k, t_j\right)$

$$\mathrm{cor}\left(t_k, t_j\right) = \frac{h_{kj} + \frac{\mu_k \mu_j}{\sigma^2}\left[1 - d_m^2\right]}{\left(\left[1 + \left(\frac{\mu_k}{\sigma}\right)^2\left[1 - d_m^2\right]\right]\left[1 + \left(\frac{\mu_j}{\sigma}\right)^2\left[1 - d_m^2\right]\right]\right)^{1/2}} \tag{15.41}$$

*Clearly,* $\mathrm{cor}\left(t_k, t_j\right) = h_{kj}$ *under* $\beta_k = \beta_j = 0$ *in Eq.* (15.41).

The proof is outlined in appendix "Proof of Lemma 15.2".

**Lemma 15.3**

*(I) Under* $\mathrm{rank}(\boldsymbol{X}) = K$, *the conditional correlation* $h_{kj} = f_{kj}/\sqrt{f_{kk}f_{jj}}$ *of* $\left(t_k, t_j\right)$ $(j \neq k)$ *is* 0 *if and only if* $\hat{\boldsymbol{x}}_k^\top \hat{\boldsymbol{x}}_j = \boldsymbol{x}_k^\top \boldsymbol{x}_j$, *which is equivalent to* $\cos\theta_{kj} = \cos\omega_k \cos\omega_j \cos\hat{\theta}_{kj}$, *where* $\theta_{kj}$ *measures the angle of* $\boldsymbol{x}_k$ *and* $\boldsymbol{x}_j$ *in* $R^N$, $\hat{\theta}_{kj}$ *the angle of* $\hat{\boldsymbol{x}}_k \in L\left(\check{\boldsymbol{X}}_k\right)$ *and* $\hat{\boldsymbol{x}}_j \in L\left(\check{\boldsymbol{X}}_j\right)$ *in* $R^N$, *and* $\cos\omega_i = \|\widehat{\mathbf{x}}_i\| / \|\mathbf{x}_i\|$ $(i = k, j)$. *If one of* $\mathbf{x}_j$*s has a continuous distribution, the probability that the condition holds is zero.*

*(II) The unconditional correlation* $\mathrm{cor}\left(t_k, t_j\right)$ *is* 0 *if* $h_{kj} = 0$ *and* $\beta_k = \beta_j = 0$. *In fact, it is* 0 *if and only if*

$$\frac{f_{kj}}{f_{kk}f_{ij}} = (1 - d_m^2)\frac{\beta_k \beta_j}{\sigma^2}, \tag{15.42}$$

*which does not hold a.e. if one of* $\mathbf{x}_j$*s is generated by a continuous distribution.*

***Proof***

(I) Since $\boldsymbol{x}_j \in L\left(\check{\boldsymbol{X}}_k\right)$ implies $\check{\boldsymbol{M}}_k \boldsymbol{x}_j = \boldsymbol{x}_j$ $(j \neq k)$ and vice versa,

$$f_{kj} = \boldsymbol{x}_k^\top \left(I - \check{\boldsymbol{M}}_k\right)\left(I - \check{\boldsymbol{M}}_j\right)\boldsymbol{x}_j = x_k^\top \check{\boldsymbol{M}}_k \check{\boldsymbol{M}}_j \boldsymbol{x}_j - \boldsymbol{x}_k^\top \boldsymbol{x}_j = \hat{\boldsymbol{x}}_k^\top \hat{\boldsymbol{x}}_j - \boldsymbol{x}_k^\top \boldsymbol{x}_j = 0.$$

This holds if and only if $\left\|\hat{\boldsymbol{x}}_k\right\| \left\|\hat{\boldsymbol{x}}_i\right\| \cos\hat{\theta}_{ki} = \|\boldsymbol{x}_k\| \|\boldsymbol{x}_i\| \cos\theta_{ki}$, implying the result.

(II) Combining $\mu_i = \sqrt{f_{ii}}\beta_i$ and $h_{kj}$ in Eq. (15.35) with the numerator in Eq. (15.42) yields $g_{kj} + \gamma_{ki} = 0$ with $g_{kj}$ in Proposition 15.2(g) and $\gamma_{kj} \equiv \left[1 - d_m^2\right]\beta_k \beta_l / \sigma^2$, which is the result.

□

Some remarks follow. These results are important in understanding the moment structure of $T \equiv (t_2, t_3, \ldots, t_K)$ so that we will be able to get a whole picture about how $t_j$s are distributed under given $\beta_j$s and how $\text{PDF}_k$s affect the moment structure through $\sqrt{f_{jj}}\beta_j$. In particular, in the moment structure, a realization of the pair $\left(t_k, t_j\right)$ depends on $\sqrt{f_{kk}}\beta_k$ and $\sqrt{f_{jj}}\beta_j$. Consequently, without making a reasonable insight into these values and without considering the correlation structure, it will not be easy to say that the $j$-th variable in the problem given by Eq. (15.34) be deleted.

In fact, all these moments and correlations depend on $\mu_i = \sqrt{f_{ii}}\beta_i$ $(i = 2, \ldots, K)$, and if $\text{PDF}_i$ is close to 0, $\mu_i$ is close to 0 even if $\beta_i$ may not be close to 0, and the mean, covariance, and correlation, are, respectively close to 0, $mh_{kj}/(m-2)$, and $h_{kj}$. These hold exactly and unconditionally under $\beta_k \beta_j = 0$ as in Lemma 15.3(II). In addition, the condition in Eq. (15.41) hardly ever holds, Hence, except for the case (15.42) or without loss of generality (a.e.), $\text{cor}\left(t_k, t_j\right) \neq 0$ unconditionally, while Lemma 15.3(I) gives a geometric structure for which the result holds in terms of *four* angles in cosine functions.

In the traditional stepwise MSPs, it is often the case that the relative contributions of variables to the model are evaluated by the sizes of $t$-values in each step and a variable with the least $t$-value is likely to be first deleted. But as Lemmas 15.2 and 15.3 show, those MPSs are not theoretically effective because the $t$-values that are simultaneously realized with $y$ are not directly compared in terms of the explanatory power. In addition, the conditional correlations and $\text{PDF}_i$s in each step will be desired to be taken into the stepwise MSP analysis. In any case, it will be necessary to evaluate the values of $\sqrt{f_{ii}} = \text{PDF}_i \sqrt{N} s_{xi} = \text{EEF}_i$ and the conditional correlations of $h_{kj}$s in each step.

### 15.4.2 Testing the Equality of the Means of $t_k$ and $t_j$

Now, let us consider the problem of testing the equality of the means of $t_k$ and $t_j$ in (15.39) or equivalently the problems of the forms

$$\mathcal{H}_0 : \mu_k = \mu_j \quad \text{vs.} \quad \mathcal{H}_1 : \mu_k > \mu_j \quad \text{and} \quad \mathcal{H}_0 : \mu_k = \mu_j \quad \text{vs.} \quad \mathcal{H}_2 : \mu_k \neq \mu_j \tag{15.43}$$

when the joint distribution of $\left(t_k, t_j\right)$ is derived as the marginal distribution of the joint decomposed conditional normal p.d.f. $\zeta\left(t_k, t_j \mid \hat{\sigma}^2, \left(\mu_k, \mu_j, \sigma^2\right)\right)$ of $\left(t_k, t_j\right)$ given $\hat{\sigma}^2$ in Eq. (15.37) and the unconditional p.d.f. $\xi\left(\hat{\sigma}^2 \mid \sigma^2\right)$ of $\hat{\sigma}^2$ with $z = m\hat{\sigma}^2/\sigma^2 \sim \chi_m^2$. But the sufficient statistic $(t_k, t_j, \hat{\sigma}^2)$ is equivalent to the sufficient statistic $\left(b_k, b_j, \hat{\sigma}^2\right)$, where

$$\begin{pmatrix} b_k \\ b_j \end{pmatrix} \equiv \begin{pmatrix} \hat{\beta}_k \sqrt{f_{kk}} \\ \hat{\beta}_j \sqrt{f_{\mu j}} \end{pmatrix} \equiv \begin{pmatrix} \hat{\mu}_k \\ \mu_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_k \\ \mu_j \end{pmatrix}, \sigma^2 H\right) \quad \text{and} \quad m\hat{\sigma}^2 \sim \sigma^2 \chi_m^2 . \tag{15.44}$$

Now, to test the problems in Eq. (15.43), we apply invariance principle to each testing problem. Let the group $G_1 = \left\{(a, c) \in R^+ \times R\right\}$ (or $G_2 = \{(a, c) \in R^* \times R\}$ resp.) act on $\left(b_k, b_j, \hat{\sigma}^2\right)$ by $b_i \rightarrow ab_i + c$ and $\hat{\sigma} \rightarrow a\hat{\sigma}$ for $\mathcal{H}_0$ vs. $\mathcal{H}_1$ with $(a, c) \in G_1$ (or for $\mathcal{H}_0$ vs. $\mathcal{H}_2$ with $(a, c) \in G_2$ resp.), where $R^+ = \{a \in R : a > 0\}$ and $R^+ = \{a \in R : a \neq 0\}$. Then it is easy to see that a maximal invariant is $t_k - t_j$ for $\mathcal{H}_0$ vs. $\mathcal{H}_1$ (or $\left|t_k - t_j\right|$ for $\mathcal{H}_0$ vs. $\mathcal{H}_2$ resp.) and a maximal invariant parameter is $\gamma_{kj}$ for $\mathcal{H}_0$ vs. $\mathcal{H}_1$ (or $\left|\gamma_{kj}\right|$ for $\mathcal{H}_0$ vs. $\mathcal{H}_2$ resp.), where

$$\gamma_{kj} = \sqrt{m}\delta_{kj}/\sigma \quad \text{with} \quad \delta_{kj} = \frac{\mu_k - \mu_j}{\sqrt{2\left(1 - h_{kj}\right)}} . \tag{15.45}$$

Therefore, if the distribution of $t_k - t_j$ is obtained, then the testing problems are invariantly reduced to testing $\mathcal{H}_0$ vs. $\mathcal{H}_1 : \gamma_{kj} > 0$ (or $\mathcal{H}_2 : \left|\gamma_{kj}\right| \neq 0$ resp). Since for given $\hat{\sigma}$ or equivalently $z = m\hat{\sigma}^2/\sigma^2$,

$$t_{kj} = \frac{t_k - t_j}{\sqrt{2\left(1 - h_{kj}\right)}} \sim N\left(\frac{\gamma_{kj}}{\sqrt{z}}, \frac{m}{z}\right) , \tag{15.46}$$

where $z = m\hat{\sigma}^2/\sigma^2 \sim \chi_m^2$. Then, as is well-known, the unconditional distribution of $t_{kj}$ is clearly noncentral $t$ distribution with d.f. $m$ and noncentrality $\gamma_{kj} = \sqrt{m}\delta_{kj}/\sigma$, denoted by $t\left(m, \gamma_{kj}\right)$, and the null distribution under $\gamma_{ki} = 0$ is central $t$-distribution $t(m, 0)$.

Consequently, in the reduced problems via invariance, $t_{kj}$ is a test statistic for testing $\mathcal{H}_0 : \gamma_{kj} = 0$ vs. $\mathcal{H}_1 : \gamma_{kj} > 0$, and $\left|t_{kj}\right|$ is a test statistic for the two-sided alternative $\mathcal{H}_2 : \gamma_{kj} \neq 0$. Hence, obviously, the following results follow:

**Theorem 15.2** *The tests $\phi_{kj1}$ and $\phi_{kj2}$ with c.r. $t_{kj} > c$ and $F = t_{kj}^2 > c$ are, respectively, UMPI for $\mathcal{H}_1 : \mu_k > \mu_j$ and $\mathcal{H}_2 : \mu_k \neq \mu_j$, and the power functions an increasing function of $\gamma_{kj}$ and $\gamma_{kj}^2$, respectively, where $\gamma_{kj} = \sqrt{m}\delta_{kj}/\sigma$. And $t_{kj}$ is distributed as noncentral t distribution $t\left(m, \gamma_{kj}\right)$ with d.f. and noncentrality $\gamma_{kj}$, and $F = t_{kj}^2$ is distributed as noncentral F-distribution with d.f. $(1, m)$ and noncentrality $\gamma_{kj}^2 = \left(\sqrt{m}\delta_{kj}/\sigma\right)^2$.*

In fact, in the one-sided case, since the p.d.f. of $t_{kj}$ is of the form $k\left(t_{kj}|\gamma_{ki}\right) = k\left(t_{kj} \mid 0\right)\ell\left(t_{kj} \mid \gamma_{kj}\right)$ and, for any fixed $\gamma_{kj} > 0$, then $\ell\left(t_{kj} \mid \gamma_{kj}\right) > c$ implies $t_{kj} > c$, the result follows. The two-sided case is similar with $F = t_{ki}^2$. Therefore, the critical points for one-sided and two-sided tests are obtained by the central $t$-distribution, implying the results. Also, the power functions for $\mathcal{H}_1 : \mu_k > \mu_j$ and $\mathcal{H}_2 : \mu_k \neq \mu_j$ are known to be increasing in $\gamma_{kj}$ and $\left|\gamma_{kj}\right|$, respectively.

Then, the two $t$-values $\left(t_k, t_j\right)$ are compared, and the basic correlation $h_{kj}$ affects the denominators of the test statistic $t_{kj}$ and the parameter $\gamma_{kj}$ in Eq. (15.45) through which it affects a decision on VSP-T.

In fact, for the two-sided problem $\mathcal{H}_0$ vs. $\mathcal{H}_2 : \mu_k \neq \mu_j$ in Theorem 15.2, the test $\phi_{kj2}$ with c.r. $\left|t_{kj}\right| > c$ is UMPI, and its power function is increasing in $\gamma_{kj}^2$. But $\gamma_{kj}^2$ in the power function contains not only variate's own factor $\sqrt{N}s_{xi}$ and collinearity factor $\text{PDF}_i$ together with $\beta_i$ $(i = k, j)$ through $\mu_i$ but also the correlation factor $h_{kj}$. These factors create various values of $\gamma_{kj}^2$, which in turn create various realizations of $t_{kj} \equiv t_{kj}\left(t_k, t_j\right)$ in a stochastic way.

The set of pairs of $\left(t_k, t_j\right)$ in which the UMPI test statistics do not reject the null hypothesis

$$D = \left\{\left(t_k, t_j\right) : \left|t_{kj}\right| \leq c, k, j = 2, \ldots, K, j < k\right\},$$

where $\text{P}\left(\left|t_{kj}\right| \leq c\right) = 1 - \alpha$ under the null central $t$-distribution. Each pair of this set will not be comparable and orderable at significance level $\alpha$. In other words, for $\left(t_k, t_j\right) \in D$, the two explanatory variables $\boldsymbol{x}_k$ and $\boldsymbol{x}_j$ are not sufficiently separated in view of their $t$-values with probability $1 - \alpha$, and they are only such randomly realized values with the effects of PDFs and correlations in each step of a stepwise MSP.

## 15.5 Common Error Term Problem on MSP via VSP-T in View of NPTT

Many textbooks such as Draper and Smith (1998) often describe VSP-T as an MSP in a systematic way. But in the MSP via VSP-T, the well-known problem of type I error and type II error is shared repeatedly with submodels in each repetition of MSP via VSP-T so that a selected model is valid with not big probability in the sample space $R^N$. This problem is rooted in the NPTT that requires each submodel

to share the common error term $u$ of the parent model, where the probability mass on $R^N$ is decomposed in each step so that the probability for which the selected model holds becomes smaller.

In this section, we will scrutinize the data-scientific legitimacy of the traditional MSP via VSP-T that is repeatedly applied in the framework of NPTT. From the results and arguments in Sects. 15.2, 15.3, and 15.4, it follows that the discernability of $t$-test on the significant explanatory power of a variable $\boldsymbol{x}_k$ for $\boldsymbol{y}$ in the model depends greatly on the two factors: EEFs of each $t$-test and correlations of $t$-statistics. Here, they are summarized, and additional problems to be discussed are raised:

(a) The EEF value of each $t$-statistic makes it highly possible to make a "wrong decision" in each step of the MSP to end up with an ineffective model where $t$-values are significant and adjusted CD is satisfactory. To avoid this risk, the XMSP method is proposed in Kariya et al. (2024) so that produces all the submodels with EEFs controlled without using $\boldsymbol{y}$, though in this case the MSP via VSP-T may not be needed. It is noted that $\boldsymbol{y}$ does not have a capacity to identify collinearity.

(b) In each step, the correlations of $t$-test statistics in Eq. (15.41) make the $t$-values used in selecting variables directly incomparable. In fact, as in Eq. (15.41), the correlation of two $t$-statistics depends on not only the raw correlation $h_{kj}$ of $\hat{\boldsymbol{v}}_k$ and $\hat{\boldsymbol{v}}_j$ as a function of $\boldsymbol{x}_k$ and $\boldsymbol{x}_j$ but also $\beta_k$, $\beta_j$, and $\sigma$. But it is often the case that a typical MSP via VSP-T compares the $t$-values in each step and a variable with the minimal insignificant $|t|$-value in the step is excluded in a variable-increasing method or a variable with maximal significant $|t|$-value is included in a variable-increasing method. Here, $|t|$ denotes the absolute value of $t$-value.

(c) The third problem to be discussed here is about the fact that those $|t|$-values are ordered without considering the EEF effect in (a) and the correlation problem in (b). In each step, the MSP via VSP-T first computes the $|t|$-values of all the variables and uses the minimum (or maximum) to decide which variable to be excluded (or included). In other words, the order statistic is used in terms of $|t|$-values. In this section, we discuss this problem in Sect. 15.5.1.

(d) A remark on indeterminacy of null distribution in NPTT is made in Sect. 15.5.2.

(e) These are summarized as follows: As was pointed out, VSP-T does not consider the whole correlation structure of explanatory variables $\{\boldsymbol{x}_k\}$, but it only looks into the *marginal t-values* through NPTT to make a decision of inclusion or exclusion of each variable without respect to the problems (a) and (b). This approach will make us blind what a good model should be in view of the aim of regression analysis, unless an overall optimal concept is set up in advance.

Before we discuss the order statistic problem, it is remarked that when error terms in submodels are independent, there will be no such problems in (b) and (c). Let us consider the model in Eq. (15.1) and its submodel with $\beta_K = 0$, where $\boldsymbol{x}_1 = \boldsymbol{e}$:

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \cdots + \boldsymbol{x}_k\beta_k + \cdots + \boldsymbol{x}_K\beta_K + \boldsymbol{u}\,, \tag{15.47}$$

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \cdots + \boldsymbol{x}_k\beta_k + \cdots + \boldsymbol{x}_{K-1}\beta_{K-1} + \boldsymbol{v}\,. \tag{15.48}$$

Suppose that there is no serious collinearity problem in $\boldsymbol{X}$. Then, if the error terms $\boldsymbol{u}$ and $\boldsymbol{v}$ are independent, the two models are directly compared by AIC, adjusted CD, etc. to give observation $(\boldsymbol{y}^\circ, \boldsymbol{X})$, from which a model can be selected. Even the problem can be treated in the framework of NPTT with $\boldsymbol{u}$ and $\boldsymbol{v}$ being independent, where the test will be based on the ratio of the maximum likelihoods.

On the other hand, if $\boldsymbol{u}$ and $\boldsymbol{v}$ are assumed to be equal, the two models cannot be independent, and for testing $\mathcal{H}_0$: Eq. (15.48) vs. $\mathcal{H}_1$: Eq. (15.47), the NPTT suggests to reduce the problem to a simpler problem via invariance principle, which gives $t$-test statistic $t_K(\boldsymbol{y})$ and leads us to the binary decision problem that if $\boldsymbol{y}^\circ \in A(1-\alpha) = \{\boldsymbol{y} : |t_K(\boldsymbol{y})| \le c(\alpha)\}$ with $P\,(A(1-\alpha)) = 1-\alpha$ for a given, say $\alpha = 0.05$, $\mathcal{H}_0$ is not rejected or the model defined in Eq. (15.48) is accepted along the view of MSP via VSP-T. This implies that we decide to select (15.48) with probability 0.95, meaning that the model given by Eq. (15.47) is effective with probability 0.05 (type I error).

### *15.5.1 MSP via VSP-T Within the Framework of NPTT*

To treat the problem in (c), for simplicity, we here consider a typical MSP via VSP-T in the case of variable reduction method. Let $|t_k(\boldsymbol{y})|$ be the absolute value of $t_k \equiv t_k(\boldsymbol{y})$ statistic before $\boldsymbol{y}^\circ$ is observed. It is noted that $|t_k(\boldsymbol{y})| \neq \left|t_j(\boldsymbol{y})\right|$ (a.e.) for $k \neq j$. Then, it is noted that using the minimum of $|t_k\,(\boldsymbol{y}^\circ)|$s is nothing but using the order statistics. Assume that $N$ is large and $\alpha = 0.05$, so that, as usual, acceptance region in the MSP via VSP-T is given $A(0.95) = \{\boldsymbol{y} : |t_K(\boldsymbol{y})| \le 2\}$ and $\mathrm{P}(A(0.95)) = 0.95$ approximately, where no consideration is made into the ordering of $|t_k\,(\boldsymbol{y}^\circ)|$s.

**Step 1** Let $|t_k(\boldsymbol{y})|$ be denoted as $|t_k(1, \boldsymbol{y})|$ in the first step variable selection. The first step of a typical MSP via VSP-T is that the $|t_k\,(1, \boldsymbol{y}^\circ)|$ values of all the $K$ variables except for the case of $\boldsymbol{x}_1 = \boldsymbol{e}$ in the parametrization $B(1) = \{\beta_k(1) : k = 2, \ldots, K\}$ are computed and ordered for individual null hypotheses $\mathcal{H}_0 : \beta_k(1) = 0$ in $B(1)$ for each $k$, based on the test statistics in

$$T(1, \boldsymbol{y}) = \{|t_k(1, \boldsymbol{y})| : k = 2, \ldots, K\}\,.$$

Then, in the MSP VSP-T, based on the observed $|t_k(1, \boldsymbol{y}^\circ)|$s, a variable that has a minimum value of

$$\mathrm{Min}T(1, \boldsymbol{y}^\circ) = \min\left\{\left|t_k(1, \boldsymbol{y}^\circ)\right| : k = 2, \ldots, K\right\} \equiv \left|t_{(1)}(1, \boldsymbol{y}^\circ)\right| \leq 2$$

is excluded if it is not greater than 2. In other words, if $\boldsymbol{y}^\circ$ falls in the region $G_{\min}(1) \equiv \{\boldsymbol{y} : \mathrm{Min}\ T(1, \boldsymbol{y}) \leq 2\}$, the variable that gives the minimum is deleted from the model, where $\left|t_{(k)}(1, \boldsymbol{y}^\circ)\right|$s are observations of order statistics of $\{|t_k(1, \boldsymbol{y})|\}$ :

$$\left|t_{(1)}(1, \boldsymbol{y})\right| \leq \cdots \leq \left|t_{(K-1)}(1, \boldsymbol{y})\right| .$$

But to be strict, here, the cutoff point $c(1)$ is no longer 2, but it should satisfy $\mathrm{P}(\mathrm{Min}T(1, \boldsymbol{y}) \leq c(1)) = 0.95$ under the null hypothesis that the corresponding parameter is zero in the marginal distribution of $\left|t_{(1)}(1, \boldsymbol{y})\right|$, though the derivation problem of the marginal distribution is left open here because it is intractably difficult by the reasons of (a) and (b). The joint distribution of $\{t_k(\boldsymbol{y})\}$ is described in Lemma 15.1. Hence, $c(1)$ is unknown and so is $\mathrm{P}(\mathrm{Min}T(1, \boldsymbol{y}) \leq c(1))$.

Here, it should be noted that $t_k(\boldsymbol{y})$s are always correlated when $\boldsymbol{X}$ includes a continuous-type variable, as has been shown in Sect. 15.4. Besides, the correlations are affected by the EEFs of the $t$-test statistics and $(\beta_k, \beta_j)$ and $\sigma$.

**Step 2** The second step is similarly discussed. Let $\left|t_{(1)}(1, \boldsymbol{y}^\circ)\right|$ correspond to $\boldsymbol{x}_K$ without loss of generality after arranging the order of the variables. Then, $K - 2$ individual null hypotheses $\mathcal{H}_0 : \beta_k(2) = 0$ for $B(2) = \{\beta_k(2) : k = 2, \ldots, K - 1\}$ are treated, where the test statistic is given by

$$T(2, \boldsymbol{y}) = \{|t_k(2, \boldsymbol{y})| : k = 2, \ldots, K - 1\}.$$

Then, in the MSP via VSP-T, if

$$\mathrm{Min}T(2, \boldsymbol{y}^\circ) = \min\left\{\left|t_k(2, \boldsymbol{y}^\circ)\right| : k = 2, \ldots, K - 1\right\} \equiv \left|t_{(1)}(2, \boldsymbol{y}^\circ)\right| \leq 2,$$

the corresponding variable satisfying this is deleted if any. Without loss of generality, let it be $\boldsymbol{x}_{K-1}$. Here, again, the acceptance region should be $\mathrm{Min}T(2, \boldsymbol{y}) \leq c(2)$ that satisfies $\mathrm{P}(\mathrm{Min}T(2, \boldsymbol{y}) \leq c(2)) = 0.95$.

**Step 3** This procedure will be stopped at $m$ if no variable in the selected model is insignificant. But the selected model is effective with probability $(0.95)^m$ at most, even if $c(j)$s are known. In fact, letting the acceptance region at each step be denoted by

$$A(j : \boldsymbol{y}) = \{\boldsymbol{y} : \mathrm{Min}T(j, \boldsymbol{y}) \leq c(j)\}.$$

then, $\boldsymbol{y}^\circ \in \prod_{j=1}^m A(j:\boldsymbol{y})$ and

$$\left[\mathrm{P}\left(\prod_{j=1}^m A(j:\boldsymbol{y})\right)\right]^m \le \prod_{j=1}^m \mathrm{P}(A(j:\boldsymbol{y}))\,.$$

But since the distribution is not known, $c(j)$s are not known and depend on the correlation matrix of $t_k(\boldsymbol{y})$s of the remaining variables at each step, which involves the elements of $\boldsymbol{X}$. Therefore, the probability that a selected model is effective is not known.

This uncomfortable situation is caused by the NPTT framework where different submodels share a common error term $\boldsymbol{u}$. In other words, a submodel in Eq. (15.47) with null hypothesis $\mathcal{H}_0 : \beta_K = 0$ is compared with the parent model under the same error term, which in fact the NPTT requires as its essential part of the power maximization problem. Consequently, from a data-scientific viewpoint, it is hard to support this traditional MSP via VSP-T approach without much sure evidence, though it has been used for a long time. And the selected model will not be trustworthy.

### *15.5.2 A Remark on Indeterminacy of Null Distribution*

Another problem is the indeterminacy problem of the null distribution in NPTT. For simplicity, consider $t$-testing with the null hypothesis $\mathcal{H}_0 : \beta_K = 0$ against $\beta_K \neq 0$ in the model defined in Eq. (15.1) or Eq. (15.47), where the normality is assumed. Then, if $\mathcal{H}_0$ is accepted by its $t$-value with $c = 2$ and probability $1 - \alpha = 0.95$ under the assumed null distribution, variable $\boldsymbol{x}_K$ is excluded from model $\boldsymbol{X}$, and the model $\boldsymbol{X}^* = \check{\boldsymbol{X}}_K$ with $\boldsymbol{x}_K$ excluded (see Sect. 15.3.1 for notation) is expected to be correct with probability 0.95 by the stipulation of the NPTT, whether $\beta_K = 0$ may be true or not in reality. But since the theory allows the model to be incorrect, i.e., $\beta_K \neq 0$, with probability 0.05 under the null distribution, strictly speaking, $\boldsymbol{y}$ still follows (a) $N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$, but not (b) $N\left(\boldsymbol{X}^*\boldsymbol{\beta}^*, \sigma^2\boldsymbol{I}\right)$. Hence, this implies that the OLSE $\hat{\boldsymbol{\beta}}^*$ is biased because (a) cannot be rejected completely by this framework. To be more specific, suppose that an observed value $\boldsymbol{y}^\circ$ was realized from (a) by the model assumption and the observed $t$-value $t\,(\boldsymbol{y}^\circ)$ in $t \equiv t(\boldsymbol{y})$ simply fell into the acceptance region $A(\boldsymbol{y}) = \{\boldsymbol{y} : |t(\boldsymbol{y})| \le 2\}$. Then, the probability that $t(\boldsymbol{y})$ falls into $A(\boldsymbol{y})$ is 0.95 under the (assumed) null distribution. However, it does not mean that $\boldsymbol{y}^\circ$ was realized from the model (b), because it can also happen under the nonnull $\beta_K \neq 0$ in model (a) with considerably high probability, particularly when $\beta_K$ is in a neighborhood of $\beta_K = 0$ (or when $\check{R}_K^2$ (CD of $\boldsymbol{x}_K$ on $\check{\boldsymbol{X}}_K$) is large with $\beta_K \neq 0$, as was shown in Sect. 15.2). In fact, for such a $\beta_K \neq 0$ but close to $\beta_K = 0$, $\mathrm{P}\,(A(\boldsymbol{y}) \mid \beta_K \neq 0)$ may be large.

The missing point is that the $t$-value $t(\boldsymbol{y}^\circ)$ alone cannot determine that $\boldsymbol{y}^\circ$ is realized from (b), but the NPTT requires us to assume that it is true, probably because by assumption it is a binary choice problem between (b) and (a). But, in reality, the fact that $t(\boldsymbol{y}^\circ)$ fell into $A(\boldsymbol{y})$ does not imply that $\boldsymbol{y}^\circ$ came from (b). In other words, strictly speaking from a viewpoint of data science, based on the fact that $t(\boldsymbol{y}^\circ)$ fell into the $A(\boldsymbol{y})$, it is incorrect to judge that (b) is a correct distribution that generated $\boldsymbol{y}^\circ$. In view of sampling scheme, it is not correct to claim that the OLSEs under (b) are unbiased, as has been well known in the area of pretest estimation problem (see, for example, Saleh, 2006).

## Appendix

### *Proof of Proposition 15.1*

We have the following:

(1) For testing $\mathcal{H}_0$ vs. $\mathcal{H}_1$ in Eq. (15.18), the power function of the one-sided UMPI test $\phi_1$ is

$$\pi_1(\mu_k/\sigma) = \mathrm{E}[\phi_1] = \mathrm{P}(t_k > a_1) = \mathrm{E}\left[\mathrm{P}\left(t_k > a_1 : \hat{\sigma}\right)\right]$$
$$= 1 - \mathrm{E}\left[\Phi\left(\frac{a_1}{\sqrt{m}}\sqrt{z} - \frac{\mu_k}{\sigma}\right)\right]$$

where $\mathrm{E}\left[A : \hat{\sigma}\right]$ (resp. $\mathrm{P}(A : \hat{\sigma})$) is the conditional expectation (resp. probability) given $\hat{\sigma}$, where $z = (\sqrt{m}\hat{\sigma}/\sigma)^2 \sim \chi^2_m$ and $\Phi(\cdot)$ is the c.d.f. of $N(0, 1)$. Here, $a_1$ is determined by $\pi_1(0) = P(t_k > a_1) = \alpha$ for level $\alpha$. Then, the result follows just by differentiating the power function at 0.

(2) For testing $\mathcal{H}_0$ vs. $\mathcal{H}_1$ in Eq. (15.19), the power function of the two-sided UMPI test $\phi_2$ is

$$\pi_2(\mu_k/\sigma) = \mathrm{E}[\phi_2] = 1 - \mathrm{P}(|t_k| \le a_2)$$
$$= 1 - \left(\mathrm{E}\left[\Phi\left(\frac{a_2}{\sqrt{m}}\sqrt{z} - \frac{\mu_k}{\sigma}\right)\right] - \mathrm{E}\left[\Phi\left(-\frac{a_2}{\sqrt{m}}\sqrt{z} - \frac{\mu_k}{\sigma}\right)\right]\right), \tag{15.49}$$

where $a_2$ is determined by the size condition $\pi_2(0) = \alpha$. The result in Proposition 15.1 follows by setting the first derivative at 0 to be zero and by differentiating the first derivative again at 0.

□

## *Proof of Proposition 15.2*

The result (d) follows from a direct evaluation of the inverse matrix $\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. In fact, letting $k = K$ and using $\boldsymbol{C} = (C_{ij})$ $(i, j = 1, 2)$ with $C_{22} : 1 \times 1$ yield

$$\hat{\boldsymbol{\beta}}_K = \boldsymbol{C}_{21} \check{\boldsymbol{X}}_K' \boldsymbol{y} + \boldsymbol{C}_{22} \boldsymbol{x}_K^\top \boldsymbol{y} = \boldsymbol{x}_K^\top (\boldsymbol{I} - \check{\boldsymbol{M}}_K) \boldsymbol{y} / \boldsymbol{x}_K^\top (\boldsymbol{I} - \check{\boldsymbol{M}}_K) \boldsymbol{x}_K \,. \tag{15.50}$$

For (e) and (f), this is obtained from the definition of $\hat{\boldsymbol{u}}_K$ as

$$\hat{\boldsymbol{u}}_k = (\boldsymbol{I} - \check{\boldsymbol{M}}_k)(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) + (\boldsymbol{I} - \check{\boldsymbol{M}}_k)\boldsymbol{x}_k \beta_k + (\boldsymbol{I} - \check{\boldsymbol{M}}_k)\boldsymbol{u},$$

since $(\boldsymbol{I} - \check{\boldsymbol{M}}_k)\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{I} - \check{\boldsymbol{M}}_k)\boldsymbol{x}_k \beta_k$. The rest is straightforward.

For (g), by (d) and (f), $\operatorname{cov}(\hat{\beta}_k, \hat{\beta}_j) = \operatorname{cov}(\hat{\boldsymbol{v}}_k^\top \hat{\boldsymbol{u}}_k, \hat{\boldsymbol{v}}_j^\top \hat{\boldsymbol{u}}_j)/(f_{kk} f_{jj})$. This completes the proof. □

## *Proof of Lemma 15.2*

First, note that $\mathrm{E}[z^{-1}] = c_m$ and $\mathrm{E}[z^{-2}] = 1 \,/\, (m-2)$. The mean of $t_k$ follows as

$$\mathrm{E}[t_k] = \mathrm{E}\left[\frac{\hat{\beta}_k \sqrt{N} s_{xk} \mathrm{CDF}_k}{\hat{\sigma}}\right] = \left(\frac{\sqrt{m}}{\sigma}\right) \mathrm{E}\left[z^{-1}\right] \beta_k a_k = \left(\frac{\sqrt{m}}{\sigma}\right) \beta_k a_k c_m \,.$$

The variance of $t_k$ is $\operatorname{var}(t_k) = I + II$ with

$$\begin{aligned} I &= \mathrm{E}[\operatorname{var}(t_k : z)] = \mathrm{E}\left[m z^{-2}\right] = \frac{m}{m-2} \,. \\ II &= \operatorname{var}(\mathrm{E}[t_k : z]) = \left(\frac{\beta_k a_k}{\sigma}\right)^2 m \,\mathrm{E}\left[z^{-1} - \mathrm{E}[z^{-1}]\right]^2 \\ &= \left(\frac{\beta_k a_k}{\sigma}\right)^2 \frac{m}{m-2} \left[1 - (m-2) c_m^2\right]. \end{aligned}$$

The covariance of $(t_k, t_j)$ is expressed as

$$\begin{aligned} \operatorname{cov}(t_k, t_j) &= \mathrm{E}\left[\operatorname{cov}(t_k, t_j : z)\right] + \mathrm{E}\left[\{\mathrm{E}[t_k : z] - \mathrm{E}[t_k]\}\left\{\mathrm{E}[t_j : z] - \mathrm{E}[t_j]\right\}\right] \\ &= III + IV, \end{aligned}$$

where

$$III = \mathrm{E}[mz^{-1}h_{kj}] = \frac{m}{m-2}h_{kj}$$

$$IV = \frac{\beta_k a_k \beta_j a_j}{\sigma^2} m\,\mathrm{E}\left[z^{-1} - \mathrm{E}[z^{-1}]\right]^2 = \frac{\beta_k a_k \beta_j a_j}{\sigma^2}\frac{m}{m-2}\left[1 - (m-2)c_m^2\right].$$

Hence, the unconditional correlation of $(t_k, t_j)$ directly follows from (b) and (c). □

**Acknowledgments** The author acknowledges the Japan Society for the Promotion of Science, Grant Number 21K01431, for its research support.

## References

Aiken, L. S., & West, S. G. (1996). *Multiple regression: Testing and interpreting interaction*. SAGE Publication.

Belsley, D. A. (1991). *Conditioning diagnosis, collinearity and weak data in regression*. Wiley.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). Wiley.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178–183.

García, J., Salmerón, R., García, C., & Martín, M. M. L. (2016). Standardization of variables and collinearity diagnostic in ridge regression. *International Statistical Review, 84*, 245–266.

Johnson, N. L., & Kotz, S. (1972). *Distributions in statistics—continuous multivariate distributions*. Wiley.

Jolliffe, I. T. (1986). *Principal component analysis*. Springer.

Kariya, T., & Kurata, H. (2003). *Generalized least squares*. Wiley.

Kariya, T., Kurata, H., & Hayashi, T. (2024). A modelling framework for regression with collinearity. *Journal of Statistical Planning and Inference, 228*, 95–115.

Kubokawa, T., & Srivastava, M. S. (2004). Improved empirical Bayes ridge regression estimators under multicollinearity. *Communications in Statistics-Theory and Methods, 33*(8), 1943–1973.

Lavery, R. M., Acharya, P., Sivo, A. S., & Xu, L. (2019). Number of predictors and multicollinearity: What are their effects on error and bias in regression?. *Communications in Statistics—Simulation and Computation, 48*(1), 27–38.

Lehmann, E. L., & Romano, J. R. (2005). *Testing statistical hypotheses*. Springer.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics, 12*(3), 591–612.

Mason, H. C., & Perreault, W. D. Jr. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research, 28*(3), 268–280.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factor. *Quality and Quantity, 41*, 673–690.

Saleh, A. K. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. Wiley.

Saleh, A. K. M. E., Arashi, M., Saleh, R. A., & Norouzirad, M. (2022). *Rank-based methods for Shrinkage and selection: With application to machine learning*. Wiley.

Sen, A., & Srivastava, M. S. (1990). *Regression analysis-theory, methods and applications*. Springer.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 67*(2), 301–320.

# Chapter 16
# A New Ridge-Based Biased Prediction Technique in Linear Mixed Models




**Özge Kuran and M. Revan Özkale**

## 16.1 Introduction

The linear regression model is the most frequently preferred statistical model. However, there are instances when we encounter data structures that require models incorporating both fixed and random effects. Fixed effects are parameters linked to an entire population or specific repeatable levels of experimental factors, while random effects pertain to individual experimental units that are randomly selected from a population. This type of model, which includes both fixed and random effects, is known as a linear mixed model (LMM) (Pinheiro & Bates, 2000).

LMMs have gained significant recognition as essential tools for analyzing clustered or longitudinal data, making them applicable across diverse fields such as agriculture, sociology, economics, biology, and medicine (Demidenko, 2013; Jiang, 2007; Laird & Ware, 1982; Verbeke & Molenberghs, 2000).

The LMM is considered as

$$y_i = X_i\beta + Z_i u_i + \varepsilon_i, \quad i = 1, \ldots, m,$$

where $y_i$ is an $n_i \times 1$ vector of response variables measured on subject $i$; $\beta$ is a $p \times 1$ parameter vector of fixed effects; $X_i$ and $Z_i$ are $n_i \times p$ and $n_i \times q$ known design matrices of the fixed and random effects, respectively; and $u_i$ is a $q \times 1$ random vector, the components of which are called random effects , while $\varepsilon_i$ is an $n_i \times 1$

Ö. Kuran (✉)
Faculty of Science, Department of Statistics, Dicle University, Diyarbakır, Türkiye
e-mail: ozge.kuran@dicle.edu.tr

M. R. Özkale
Faculty of Science and Letters, Department of Statistics, Çukurova University, Adana, Türkiye
e-mail: mrevan@cu.edu.tr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_16

random vector of errors. Usually, one assumes that

$$u_i \overset{iid}{\sim} N_q\left(0, \sigma^2 \Gamma\right) \quad \text{and} \quad \varepsilon_i \overset{iid}{\sim} N_{n_i}\left(0, \sigma^2 W_i\right), \qquad i = 1, \ldots, m,$$

where $u_i$ and $\varepsilon_i$ are independent and $\Gamma$ and $W_i$ are known positive definite (p.d.) matrices of dimensions $q \times q$ and $n_i \times n_i$, respectively.

Let $y = (y_1^\top, \ldots, y_m^\top)^\top$, $X = (X_1^\top, \ldots, X_m^\top)^\top$, and $Z = \oplus_{i=1}^m Z_i$, where $\oplus$ represents the direct sum, allowing us to combine them into a single matrix. Also, $u = (u_1^\top, \ldots, u_m^\top)^\top$ and $\varepsilon = (\varepsilon_1^\top, \ldots, \varepsilon_m^\top)^\top$. Then, we can write the model more compactly as

$$y = X\beta + Zu + \varepsilon, \tag{16.1}$$

which implies that

$$\begin{pmatrix} u \\ \varepsilon \end{pmatrix} \sim N_{qm+n}\left(\begin{pmatrix} 0_{qm} \\ 0_n \end{pmatrix}, \sigma^2 \begin{pmatrix} G & 0 \\ 0 & W \end{pmatrix}\right),$$

where $n = \sum_{i=1}^{m} n_i$, $G = I_m \otimes \Gamma$, and $W = \oplus_{i=1}^m W_i$. The $\otimes$ denotes the Kronecker product, and $I_m$ represents the identity matrix of order $m$. Then, under model defined in Eq. (16.1), we derive $y \sim N\left(X\beta, \sigma^2 H\right)$, where $H = ZGZ^\top + W$. It is assumed that the matrices $G$ and $W$ are known for ease of theoretical computations. However, in real-world data applications, the matrices $G$ and $W$ are often unknown; therefore, we substitute them with their maximum likelihood (ML) or restricted maximum likelihood (REML) estimates.

Henderson (1950) and Henderson et al. (1959) respectively found the best linear unbiased estimator (BLUE) ($\hat{\beta}$) for $\beta$ and the best linear unbiased predictor (BLUP) ($\hat{u}$) for $u$ expressed as

$$\hat{\beta} = (X^\top H^{-1} X)^{-1} X^\top H^{-1} y, \tag{16.2}$$

$$\hat{u} = GZ^\top H^{-1}(y - X\hat{\beta}). \tag{16.3}$$

Although the variables in the fixed effects design matrix are typically assumed to be nearly independent, practical scenarios often reveal a tendency toward dependence. In such cases, the matrix $X^\top H^{-1} X$ becomes ill-conditioned , meaning $\left|X^\top H^{-1} X\right| \approx 0$, where $|\cdot|$ represents the determinant of a matrix. When $X^\top H^{-1} X$ is ill-conditioned, $\hat{\beta}$ becomes sensitive to errors, and at least one element on the main diagonal of $(X^\top H^{-1} X)^{-1}$ may be considerably large. Given that $\text{var}(\hat{\beta}) = \sigma^2 (X^\top H^{-1} X)^{-1}$, this implies that at least one element of $\hat{\beta}$ may have a large variance, potentially causing it to deviate significantly from its true value. Consequently, making meaningful statistical inferences becomes very difficult for practitioners dealing with such models. To address these multicollinearity issues,

several methods are available in the literature, among which ridge regression is one of the most popular. In this context, Liu and Hu (2013) and Özkale and Can (2017) extended the ridge estimator originally developed for the linear regression models by Hoerl and Kennard (1970) to be applicable within the framework of LMMs. They subsequently proposed the ridge estimator and ridge predictor in LMMs, respectively, as

$$\hat{\beta}_k = (X^\top H^{-1} X + kI_p)^{-1} X^\top H^{-1} y, \tag{16.4}$$

$$\hat{u}_k = GZ^\top H^{-1}(y - X\hat{\beta}_k), \tag{16.5}$$

where $k > 0$ is called the ridge biasing parameter. Özkale and Can (2017) derived Eqs. (16.4) and (16.5) by augmenting the model given by Eq. (16.1) with the equation $0 = \sqrt{k}\beta + \psi$, where $\psi \sim N(0, I_*)$ with $I_* = \sigma^2 I_p$. They then applied the penalized log-likelihood approach. As $k$ increases, the distance between $\sqrt{k}\beta$ and 0 expands. However, augmenting the equation $0 = \sqrt{k}\beta + \psi$ to the model defined in Eq. (16.1) introduces additional bias to ridge regression. Consequently, practitioners are generally advised to select a smaller value for $k$ to minimize bias. Nonetheless, even with a small $k$, $X^\top H^{-1} X + kI_p$ may still be ill-conditioned, which can lead to potentially unstable ridge estimators and predictors. The condition number of $X^\top H^{-1} X + kI_p$ decreases as $k$ increases. Therefore, a larger $k$ is generally favored to maintain a smaller condition number, which enhance the stability of the estimation process.

Liu (1993) introduced an alternative approach specifically designed to improve estimation in linear regression models, subsequently naming it the Liu estimator. Building upon Liu's study, Kaçıranlar et al. (1999) made further refinements to the method by proposing the restricted Liu estimator and examining its stochastic properties. In extending Liu's methodology from linear regression models to LMMs, Özkale and Kuran (2020) adopted a penalized log-likelihood approach. Their work was inspired by Swindel (1976) and further built upon the findings of Özkale and Kaçıranlar (2007). Consequently, Özkale and Kuran (2020) respectively put forward the Liu estimator and the Liu predictor in the context of LMMs as

$$\hat{\beta}_d = (X^\top H^{-1} X + I_p)^{-1} (X^\top H^{-1} y + d\hat{\beta}), \tag{16.6}$$

$$\hat{u}_d = GZ^\top H^{-1}(y - X\hat{\beta}_d), \tag{16.7}$$

where $0 < d < 1$ is the Liu biasing parameter and $\hat{\beta}$ is given by Eq. (16.2). Although not explicitly mentioned in Özkale and Kuran (2020), the Liu estimator and predictor can be derived by augmenting model defined in Eq. (16.1) with the equation $d\hat{\beta} = \beta + \psi$ where $0 < d < 1$. This framework involves applying the penalized log-likelihood approach. The rationale is that in ridge prediction, as defined by Özkale and Can (2017), the difference between $\sqrt{k}\beta$ and 0 in the equation $0 = \sqrt{k}\beta + \psi$ increases as $k$ approaches infinity. In contrast, the equation $d\hat{\beta} = \beta + \psi$ solves this problem by taking $d$ between 0 and 1.

Some of the other biased estimators and predictors proposed in LMM are as follows: Kuran (2020) suggested the modified ridge estimator and predictor in LMMs. Kuran and Özbay (2021) defined a two-parameter estimator and predictor with the help of the penalized log-likelihood approach in LMMs. Kuran (2023) proposed the modified Liu estimator and predictor by unifying the advantages of both ridge and Liu prediction methods.

In this chapter, we aim to propose a novel ridge-based biased prediction technique that encompasses BLUE/BLUP, ridge estimator/predictor, and Liu estimator/predictor. In Sect. 16.2, this new prediction technique offers an alternative approach to addressing multicollinearity. In Sect. 16.3, we show that the new ridge-based biased prediction technique exhibits superior properties in terms of mean squared error compared to the ridge and Liu prediction methods. In Sect. 16.4, the selection of biasing parameters is handled. Section 16.5 provides a real dataset analysis to illustrate these findings. In Sect. 16.6, we give some conclusions.

## 16.2 A New Ridge-Based Biased Prediction of Fixed and Random Effects

In this section, we propose new ridge-based biased parameter vectors for fixed and random effects using the penalized log-likelihood approach.

Assuming model defined in Eq. (16.1), $u$ and $y$ follow a joint Gaussian distribution as

$$\begin{pmatrix} u \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \sigma^2 \begin{pmatrix} G & GZ^T \\ ZG & H \end{pmatrix}\right).$$

The conditional distribution of $y$ given $u$ is $y|u \sim N\left(X\beta + Zu, \sigma^2 W\right)$. Henderson et al. (1959) formulated equations that provide both the BLUE and the BLUP. They achieved this by maximizing the joint density of $y$ and $u$ given by

$$\begin{aligned} f(y,u) &= f(y|u)\, f(u) \\ &= (2\pi\sigma^2)^{-(n+qm)/2} |W|^{-1/2} |G|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}[(y - X\beta - Zu)^\top W^{-1} (y - X\beta - Zu) + u^\top G^{-1} u]\right\}. \end{aligned}$$

These resulting equations are referred to as Henderson's mixed model equations (MMEs) (Henderson et al., 1959).

We aim to identify a new estimator that is closer in magnitude to $\beta$ than $\hat{\beta}_k$. To achieve this, it may be advantageous to minimize the distance between an estimator and a specified point, distinct from the origin. According to Swindel (1976), this point can be determined based on prior information. Following the approaches of Özkale and Kaçıranlar (2007) and Sakallıoğlu and Kaçıranlar (2008) in linear regression models, we select the prior information to be $d\hat{\beta}_k$. Then, the stochastic linear restriction $d\hat{\beta}_k = \beta + \psi$, where $\psi \sim N(0, \sigma^2 I_p)$ with $k > 0$ and $-\infty < d < \infty$, can be unified with model defined in Eq. (16.1). This results in $y_* = X_*\beta + Z_* u + \epsilon_*$, where

$$y_* = \begin{pmatrix} y \\ d\hat{\beta}_k \end{pmatrix}, \quad X_* = \begin{pmatrix} X \\ I \end{pmatrix}, \quad Z_* = \begin{pmatrix} Z \\ 0_{p\times mq} \end{pmatrix}$$

and $\epsilon_* \sim N(0, W_*)$ with $W_* = \begin{pmatrix} W & 0 \\ 0 & I_p \end{pmatrix}$. Then, $u$ and $y_*$ are jointly distributed as

$$\begin{pmatrix} u \\ y_* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ X_*\beta \end{pmatrix}, \sigma^2 \begin{pmatrix} G & G\,Z_*^\top \\ Z_* G & H_* \end{pmatrix}\right),$$

where $H_* = Z_*\, G\, Z_*^\top + W_* = \begin{pmatrix} H & 0 \\ 0 & I_p \end{pmatrix}$. Then, the conditional distribution of $y_*$ given $u$ is $y_*|u \sim N\left(X_*\beta + Z_* u, W_*\right)$ (see Searle, 1982, p. 354). To derive the estimator of $\beta$ and the predictor of $u$ simultaneously, we need to maximize the joint density of $y$ and $u$, given by

$$\begin{aligned} f(y,u) &= f(y|u)\, f(u) \\ &= \frac{\exp\left\{-\frac{1}{2}[(y_* - X_*\beta - Z_* u)^\top W_*^{-1}(y_* - X_*\beta - Z_* u) + u^\top G^{-1} u]\right\}}{(2\pi)^{(n+p+mq)/2}\, |W_*|^{1/2}\, |G|^{1/2}}. \end{aligned}$$

The log-joint distribution of $f\,(y, u)$ is derived as

$$\begin{aligned} \log f\,(y,u) &= \log f\,(y|u) + \log f\,(u) \\ &= -\frac{1}{2}\Big\{(n+p+qm)\log(2\pi) + (p+n+qm)\log\sigma^2 \\ &\quad + \log|W_*| + \log|G| \\ &\quad + [(y_* - X_*\beta - Z_* u)^\top\, W_*^{-1}\,(y_* - X_*\beta - Z_* u) + u^\top G^{-1} u]\,/\,\sigma^2\Big\}. \end{aligned}$$

By omitting the constant term and substituting $y_*$, $X_*$, $Z_*$, and $W_*$, we get

$$\begin{aligned}\log f(y,u) &= -\frac{1}{2\sigma^2}\Big\{(y-X\beta-Zu)^\top W^{-1}(y-X\beta-Zu)+u^\top G^{-1}u \\ &\quad +(d\hat{\beta}_k-\beta)^\top(d\hat{\beta}_k-\beta)\Big\} \\ &= -\frac{1}{2\sigma^2}\Big\{(y-X\beta)^\top W^{-1}(y-X\beta)+(d\hat{\beta}_k-\beta)^\top(d\hat{\beta}_k-\beta)\Big\} \\ &\quad -\frac{1}{2\sigma^2}\{u^\top(Z^\top W^{-1}Z+G^{-1})u-2(y-X\beta)^\top W^{-1}Zu\}, \quad (16.8)\end{aligned}$$

where the first term implements a new biased estimator based on ridge estimation given by Sakallıoğlu and Kaçıranlar (2008) in linear regression models and the second term applies the prediction process for the random effects.

Setting the partial derivatives of Eq. (16.8) with respect to the elements of $\beta$ and $u$ to zero and using $\hat{\beta}_{k,d}$ and $\hat{u}_{k,d}$ to represent the solutions, respectively, result in

$$X^\top W^{-1}(y-X\hat{\beta}_{k,d})+d\hat{\beta}_k-\hat{\beta}_{k,d}-X^\top W^{-1}Z\hat{u}_{k,d}=0 \quad (16.9)$$

$$Z^\top W^{-1}(y-X\hat{\beta}_{k,d})-(Z^\top W^{-1}Z+G^{-1})\hat{u}_{k,d}=0. \quad (16.10)$$

For convenience and clarity, Eqs. (16.9) and (16.10) can be concisely expressed in matrix form as

$$\begin{pmatrix} X^\top W^{-1}X+I_p & X^\top W^{-1}Z \\ Z^\top W^{-1}X & Z^\top W^{-1}Z+G^{-1}\end{pmatrix}\begin{pmatrix}\hat{\beta}_{k,d}\\ \hat{u}_{k,d}\end{pmatrix}=\begin{pmatrix}X^\top W^{-1}y+d\hat{\beta}_k\\ Z^\top W^{-1}y\end{pmatrix}. \quad (16.11)$$

Applying Gilmour et al. (1995)'s method, Eq. (16.11) can be formulated as

$$C\hat{\Psi}=\omega^\top W^{-1}y+\kappa, \quad (16.12)$$

where $\hat{\Psi}=(\hat{\beta}_{k,d}^\top,\hat{u}_{k,d}^\top)^\top$, $\omega=(X,Z)$, $\kappa=(d\hat{\beta}_k^\top,0^\top)^\top$, and $C=\omega^\top W^{-1}\omega+\mathcal{G}^+$ are full rank with

$$\mathcal{G}^+=\begin{pmatrix}I_p & 0\\ 0 & G^{-1}\end{pmatrix} \text{ and } \mathcal{G}=\begin{pmatrix}I_p & 0\\ 0 & G\end{pmatrix},$$

where $\mathcal{G}^{+}$ denotes the Moore-Penrose inverse of $\mathcal{G}$. By solving Eq. (16.12), we find

$$\hat{\Psi} = C^{-1}\omega^{\top}W^{-1}y + C^{-1}\kappa, \tag{16.13}$$

where $C^{-1}$ is determined using the inverse formula for partitioned matrices (Searle, 1982) as

$$C^{-1} = \begin{pmatrix} \acute{N} & -\acute{N}X^{\top}H^{-1}ZG \\ -G\,Z^{\top}H^{-1}X\acute{N} & \Upsilon + GZ^{\top}H^{-1}X\acute{N}X^{\top}H^{-1}ZG \end{pmatrix},$$

with $\acute{N} = (X^{\top}H^{-1}X + I_p)^{-1}$ and $\Upsilon = (Z^{\top}W^{-1}Z + G^{-1})^{-1}$.

Substituting $C^{-1}$ into Eq. (16.13) and simplifying algebraically, we obtain the new estimator and predictor, referred to as the $k-d$ estimator and the $k-d$ predictor, respectively, as

$$\hat{\beta}_{k,d} = (X^{\top}H^{-1}X + I_p)^{-1}(X^{\top}H^{-1}y + d\hat{\beta}_k) \tag{16.14}$$

$$\hat{u}_{k,d} = GZ^{\top}H^{-1}(y - X\hat{\beta}_{k,d}), \tag{16.15}$$

where $-\infty < d < \infty$ is the biasing parameter and $\hat{\beta}_k$ is given by Eq. (16.4). For further analysis, Eq. (16.14) can also be rewritten as

$$\hat{\beta}_{k,d} = (X^{\top}H^{-1}X + I_p)^{-1}(X^{\top}H^{-1}X + (k+d)I_p)(X^{\top}H^{-1}X + kI_p)^{-1}X^{\top}H^{-1}y.$$

The newly proposed $k - d$ estimator and predictor, given by Eqs. (16.14) and (16.15), are derived from replacing the BLUE with the ridge estimator in the Liu equation $d\hat{\beta} = \beta + \psi$. By substituting $\hat{\beta}_k$ for $\hat{\beta}$, which is sensitive to multicollinearity , we anticipate that $\hat{\beta}_{k,d}$ and $\hat{u}_{k,d}$ will provide advantages similar to those of the Liu estimator and predictor. Moreover, a large $k$ can be chosen to adjust another parameter $d$, thereby enhancing the expected advantages of $\hat{\beta}_{k,d}$ over the ridge estimator.

Estimator $\hat{\beta}_{k,d}$ and predictor $\hat{u}_{k,d}$ include other estimators for the special case of biasing parameters:

1. $\hat{\beta}_{0,1} = \hat{\beta}$, $\hat{u}_{0,1} = \hat{u}$: the BLUE and BLUP given by Eqs. (16.2) and (16.3)
2. $\hat{\beta}_{k,1-k} = \hat{\beta}_k$, $\hat{u}_{k,1-k} = \hat{u}_k$: the ridge estimator and the ridge predictor given by Eqs. (16.4) and (16.5)
3. $\hat{\beta}_{0,d} = \hat{\beta}_d$, $\hat{u}_{0,d} = \hat{u}_d$, $0 < d < 1$: the Liu estimator and the Liu predictor given by Eqs. (16.6) and (16.7).

## 16.3 Examination of Mean Square Error Performance

The prediction of linear combinations of $\beta$ and $u$ can be expressed as $\mu = L^\top \beta + M^\top u$ for specific matrices $L \in, \mathbb{R}^{p\times 1}$ and $M \in \mathbb{R}^{q\times 1}$ (Pereira & Coelho, 2012; Robinson, 1991; Yang et al., 2014). Yang et al. (2014) and Özkale and Kuran (2020) obtained the matrix mean square error (MMSE) for $\hat{\mu}$, $\hat{\mu}_k$, and $\hat{\mu}_d$ by following Štulajter (1997) and given by

$$\text{MMSE}(\hat{\mu}) = \mathbb{Q}\text{MMSE}(\hat{\beta})\mathbb{Q}^\top + \sigma^2 M^\top (G - G\,Z^\top H^{-1} ZG)M, \quad (16.16)$$

$$\text{MMSE}(\hat{\mu}_k) = \mathbb{Q}\text{MMSE}(\hat{\beta}_k)\mathbb{Q}^\top + \sigma^2 M^\top (G - GZ^\top H^{-1} ZG)M, \quad (16.17)$$

$$\text{MMSE}(\hat{\mu}_d) = \mathbb{Q}\text{MMSE}(\hat{\beta}_d)\mathbb{Q}^\top + \sigma^2 M^\top (G - GZ^\top H^{-1} ZG)M, \quad (16.18)$$

where $\mathbb{Q} = L^\top - M^\top GZ^\top H^{-1} X$ and

$$\hat{\mu} = L^\top \hat{\beta} + M^\top \hat{u} = \mathbb{Q}\hat{\beta} + M^\top G\, Z^\top H^{-1}\, y$$

$$\hat{\mu}_k = L^\top \hat{\beta}_k + M^\top \hat{u}_k = \mathbb{Q}\hat{\beta}_k + M^\top GZ^\top H^{-1} y$$

$$\hat{\mu}_d = L^\top\, \hat{\beta}_d + M^\top\, \hat{u}_d = \mathbb{Q}\hat{\beta}_d + M^\top\, G\, Z^\top\, H^{-1}\, y\,.$$

Also,

$$\text{MMSE}(\hat{\beta}) = \sigma^2 A^{-1}, \quad (16.19)$$

$$\text{MMSE}(\hat{\beta}_k) = \sigma^2 A_k^{-1} A A_k^{-1} + k^2 A_k^{-1} \beta\beta^\top A_k^{-1}, \quad (16.20)$$

$$\text{MMSE}(\hat{\beta}_d) = \sigma^2 A_p^{-1} A_d A^{-1} A_d A_p^{-1} + (1-d)^2 A_p^{-1} \beta\beta^\top A_p^{-1}, \quad (16.21)$$

where $A = X^\top H^{-1} X$, $A_k = A + k\, I_p$, $A_d = A + d\, I_p$, and $A_p = A + I_p$.

Using Eqs. (16.14) and (16.15), our new prediction technique allows us to express the predictor of $\mu$ as

$$\hat{\mu}_{k,d} = L^\top \hat{\beta}_{k,d} + M^\top \hat{u}_{k,d} = \mathbb{Q}\,\hat{\beta}_{k,d} + M^\top\, G\, Z^\top\, H^{-1}\, y.$$

We find the MMSE for $\hat{\mu}_{k,d}$ via Štulajter (1997) as

$$\text{MMSE}(\hat{\mu}_{k,d}) = \mathbb{Q}\text{MMSE}(\hat{\beta}_{k,d})\mathbb{Q}^\top + \sigma^2\, M^\top\, (G - G\, Z^\top\, H^{-1}\, Z\, G)M, \quad (16.22)$$

where

$$\begin{aligned}\text{MMSE}(\hat{\beta}_{k,d}) &= \sigma^2 A_p^{-1} A_{k,d} A_k^{-1} A A_k^{-1} A_{k,d} A_p^{-1} \\ &+ \left[A_p^{-1} A_{k,d} A_k^{-1} A - I_p\right] \beta\beta^\top \left[A_p^{-1} A_{k,d} A_k^{-1} A - I_p\right]^\top,\end{aligned} \tag{16.23}$$

where $A_{k,d} = A + (k+d)\, I_p$.

The superiority of $\text{MMSE}(\hat{\mu}_{k,d})$ over $\text{MMSE}(\hat{\mu})$, $\text{MMSE}(\hat{\mu}_k)$, and $\text{MMSE}(\hat{\mu}_d)$, as given by Eqs. (16.16)–(16.18) and Eq. (16.22), is equivalent to the superiority of $\text{MMSE}(\hat{\beta}_{k,d})$ over $\text{MMSE}(\hat{\beta})$, $\text{MMSE}(\hat{\beta}_k)$, and $\text{MMSE}(\hat{\beta}_d)$, as provided in Eqs. (16.19)–(16.21) and Eq. (16.23).

The MMSE values for the estimators $\hat{\mu}$, $\hat{\mu}_k$, $\hat{\mu}_d$, and $\hat{\mu}_{k,d}$ can be simplified by transforming model defined in Eq. (16.1) into a canonical form through an orthogonal transformation. For this purpose, we employ the canonical form. Since $H$ is p.d., there exists a nonsingular symmetric matrix $N$ such that $H = N^\top N$. By premultiplying both sides of Eq. (16.1) by $N^{-1}$, we get

$$y^* = X^* \beta + Z^* u + \varepsilon^*, \tag{16.24}$$

where $y^* = N^{-1} y$, $X^* = N^{-1} X$, $Z^* = N^{-1} Z$, $\varepsilon^* = N^{-1} \varepsilon$, and $\text{var}(y^*) = \sigma^2 I$.

Using the spectral decomposition of the matrix, we can write $X^\top H^{-1} X = P^\top \Lambda P$, where $\Lambda = \text{diag}(\lambda_i)$ is a $p \times p$ diagonal matrix containing the eigenvalues of $X^\top H^{-1} X$ (with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$) and $P = \left[P_1, \ldots, P_p\right]$ is a $p \times p$ orthogonal matrix comprising the normalized eigenvectors corresponding to these eigenvalues. Consequently, the model in Eq. (16.24) can be rewritten as $y^* = K^* \alpha + Z^* u + \varepsilon^*$, where $K^* = X^* P^\top$ and $\alpha = P\beta$. In this transformed model, MMSE of $\tilde{\alpha}$ is given by $\text{MMSE}(\tilde{\alpha}) = P[\text{MMSE}(\tilde{\beta})]P^\top = P[\text{var}(\tilde{\beta}) + \text{Bias}(\tilde{\beta})\text{Bias}(\tilde{\beta})^\top]\, P^\top$ for any estimator of $\tilde{\alpha}$. Therefore, using Eqs. (16.19)–(16.21) and Eq. (16.23), we derive that

$$\text{MMSE}(\hat{\alpha}) = \sigma^2 \Lambda^{-1}, \tag{16.25}$$

$$\text{MMSE}(\hat{\alpha}_k) = \sigma^2\, \Lambda_k^{-1}\, \Lambda\, \Lambda_k^{-1} + k^2 \Lambda_k^{-1} \alpha\alpha^\top \Lambda_k^{-1}, \tag{16.26}$$

$$\text{MMSE}(\hat{\alpha}_d) = \sigma^2\, \Lambda_p^{-1}\, \Lambda_d\, \Lambda^{-1}\, \Lambda_d\, \Lambda_p^{-1} + (1-d)^2 \Lambda_p^{-1} \alpha\alpha^\top \Lambda_p^{-1}, \tag{16.27}$$

$$\begin{aligned}\text{MMSE}(\hat{\alpha}_{k,d}) &= \sigma^2 \Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda \Lambda_k^{-1} \Lambda_{k,d} \Lambda_p^{-1} \\ &+ [\Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda - I_p] \alpha\, \alpha^\top [\Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda - I_p]^\top,\end{aligned} \tag{16.28}$$

where $\Lambda_k = (\Lambda + k\, I_p)$, $\Lambda_d = (\Lambda + d\, I_p)$, $\Lambda_p = (\Lambda + I_p)$, and $\Lambda_{k,d} = (\Lambda + (k + d) I_p)$.

To facilitate the upcoming proofs, we present two key lemmas:

**Lemma 16.1 (Farebrother, 1976)** *Let $A$ be a p.d. matrix, $b$ be a nonzero vector, and $d$ be a positive scalar. Then, $d\,A - b\,b^\top$ is p.d. if and only if $b^\top A^{-1} b < d$.*

**Lemma 16.2 (Trenkler, 1980)** *Let $\tilde{\beta}_j = A_j\, y$ for $j = 1, 2$ denote two homogeneous linear estimators of $\beta$. Assume that $Var(\tilde{\beta}_1) - Var(\tilde{\beta}_2) = \sigma^2 C = \sigma^2 (A_1 A_1^\top - A_2 A_2^\top)$ is a p.d. matrix. If $\beta^\top (A_2 X - I_p)^\top C^{-1} (A_2 X - I_p)\beta < \sigma^2$, then* $\mathrm{MMSE}(\tilde{\beta}_1) - \mathrm{MMSE}(\tilde{\beta}_2)$ *is a p.d. matrix, where* $\mathrm{Bias}(\tilde{\beta}_2) = (A_2 X - I_p)\beta$.

### *16.3.1 The k − d Estimator vs. the BLUE*

**Theorem 16.1** *For $k > 0$, if*

$$\max\left\{-\frac{(\lambda_i + k)(2\lambda_i + 1)}{\lambda_i}\right\}_i^p < d < \min\left\{\frac{\lambda_i + k}{\lambda_i}\right\}_{i=1}^p,$$

*then* $\mathrm{MMSE}(\hat{\alpha}) - \mathrm{MMSE}(\hat{\alpha}_{k,d})$ *is p.d. if and only if*

$$\alpha^\top [\Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda - I_p]^\top [\Lambda^{-1} - \Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda \Lambda_k^{-1} \Lambda_{k,d} \Lambda_p^{-1}]^{-1}$$
$$[\Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda - I_p]\alpha < \sigma^2.$$

***Proof*** With the help of Eqs. (16.25) and (16.28), we get

$$\mathrm{var}(\hat{\alpha}) - \mathrm{var}(\hat{\alpha}_{k,d}) = \sigma^2(\Lambda^{-1} - \Lambda_p^{-1} \Lambda_{k,d} \Lambda_k^{-1} \Lambda \Lambda_k^{-1} \Lambda_{k,d} \Lambda_p^{-1})$$
$$= \sigma^2\, \mathrm{diag}\left\{\frac{1}{\lambda_i} - \frac{(\lambda_i + k + d)^2\, \lambda_i}{(\lambda_i + 1)^2 (\lambda_i + k)^2}\right\}_{i=1}^p.$$

The variance difference $\mathrm{var}(\hat{\alpha}) - \mathrm{var}(\hat{\alpha}_{k,d})$ is a p.d. matrix if and only if $f(d) = (\lambda_i + 1)^2(\lambda_i + k)^2 - \lambda_i^2(\lambda_i + k + d)^2 > 0$ for a fixed $k > 0$. Since

$$f(d) = -d^2\, \lambda_i^2 - 2d\lambda_i^2\, (\lambda_i + k) + (\lambda_i + k)^2(2\lambda_i + 1)$$

is a quadratic function defined over $-\infty < d < \infty$ for a fixed $k > 0$, we find that it is positive within the bounds

$$-\frac{(\lambda_i + k)(2\lambda_i + 1)}{\lambda_i} < d < \frac{\lambda_i + k}{\lambda_i} \quad \forall i.$$

Therefore, by applying the conclusions of Lemma 16.1, the proof is completed. □

### 16.3.2 The $k - d$ Estimator vs. the Ridge Estimator

**Theorem 16.2** *For $k > 0$,*

$$\max\left\{-(\lambda_i + k) - \sqrt{(\lambda_i(\lambda_i + 2))}\right\}_i^p < d < \min\left\{-(\lambda_i + k) + \sqrt{(\lambda_i(\lambda_i + 2))}\right\}_i^p,$$

$\mathrm{MMSE}(\hat{\alpha}_k) - \mathrm{MMSE}(\hat{\alpha}_{k,d}) > 0$ *is p.d. if*

$$\alpha^\top[\Lambda_p^{-1}\Lambda_{k,d}\Lambda_k^{-1}\Lambda - I_p]^\top[B\,\Lambda\,\mathrm{diag}\,((1-d-k)(2\lambda_i + 1 + d + k))\;B]^{-1} \times[\Lambda_p^{-1}\Lambda_{k,d}\Lambda_k^{-1}\Lambda - I_p]\alpha < \sigma^2 \tag{16.29}$$

*where* $B = \Lambda_p^{-1}\Lambda_k^{-1}$.

***Proof*** With the help of Eqs. (16.26) and (16.28), the variances of $\hat{\alpha}_k$ and $\hat{\alpha}_{k,d}$ are given as

$$\mathrm{var}(\hat{\alpha}_k) - \mathrm{var}(\hat{\alpha}_{k,d}) = \sigma^2\, B\,\mathrm{diag}\,\{(1-d-k)\lambda_i\,(2\lambda_i + 1 + d + k)\}_{i=1}^p\; B^\top,$$

where $B = \Lambda_p^{-1}\Lambda_k^{-1}$. It is p.d. if and only if $f(d) = (1-d-k)(2\lambda_i+1+d+k) > 0$. Since $f(d)$ is a quadratic function in $-\infty < d < \infty$, it is positive for

$$-(\lambda_i + k) - \sqrt{(\lambda_i(\lambda_i + 2))} < d < -(\lambda_i + k) + \sqrt{(\lambda_i(\lambda_i + 2))} \quad \forall i.$$

By following Lemma 16.2, we complete the proof. □

### 16.3.3 The $k - d$ Estimator vs. the Liu Estimator

**Theorem 16.3** *For $k > 0$, either $d > 0$ or $d < -\max\left\{(\lambda_i^2 + \lambda_i k)\,/\,(\lambda_i + \frac{k}{2})\right\}_{i=1}^p$,* $\mathrm{MMSE}(\hat{\alpha}_d) - \mathrm{MMSE}(\hat{\alpha}_{k,d}) > 0$ *is p.d. if*

$$\alpha^\top[\Lambda_p^{-1}\Lambda_{k,d}\Lambda_k^{-1}\Lambda - I_p]^\top\left[\Lambda_p^{-1}\,\mathrm{diag}\,\left\{\frac{k}{2}[d^2(\lambda_i + \frac{k}{2}) + d\,(\lambda_i^2 + \lambda_i\,k)]\right\}_{i=1}^p \times\Lambda_p^{-1}\right]^{-1}[\Lambda_p^{-1}\,\Lambda_{k,d}\,\Lambda_k^{-1}\Lambda - I_p]\alpha < \sigma^2.$$

***Proof*** With the help of Eqs. (16.27) and (16.28), the variances of $\hat{\alpha}_d$ and $\hat{\alpha}_{k,d}$ are given by

$$\begin{aligned}\text{var}(\hat{\alpha}_d) - \text{var}(\hat{\alpha}_{k,d}) &= \sigma^2(\Lambda_p^{-1}\Lambda_d\Lambda^{-1}\Lambda_d\Lambda_p^{-1} - \Lambda_p^{-1}\Lambda_{k,d}\Lambda_k^{-1}\Lambda\Lambda_k^{-1}\Lambda_{k,d}\Lambda_p^{-1}) \\ &= \sigma^2\Lambda_p^{-1}\,\text{diag}\left\{\frac{k}{2}[d^2(\lambda_i+\frac{k}{2})+d(\lambda_i^2+\lambda_i k)]\right\}_{i=1}^{p}\Lambda_p^{-1}.\end{aligned}$$

By applying principles of calculus, the conditions that make the $f(d) = d^2(\lambda_i + \frac{k}{2}) + d(\lambda_i^2 + \lambda_i k) > 0$ function positive are also the conditions that make the difference $\text{var}(\hat{\alpha}_d) - \text{var}(\hat{\alpha}_{k,d})$ a p.d. matrix. Since $f(d)$ is a quadratic function in $d$ for fixed $k$, it is positive for $d > 0$ or $d < ((\lambda_i^2 + \lambda_i k)/(\lambda_i + k/2))$ for all $i$. Then by following Lemma 16.2, the proof is completed. □

## 16.4 Selection of the Biasing Parameters

Our newly proposed $k - d$ estimator and $k - d$ predictor, described in Eqs. (16.14) and (16.15), are determined by two biasing parameters. The effectiveness of the estimator and predictor is closely associated with the selection of these biasing parameters. Hence, we are compelled to search for optimal methods for selecting the biasing parameters.

One aspect of the information contained in the MMSE is that its diagonal elements represent the mean square error of any given estimator $\tilde{\beta}_i$ for the parameter $\beta_i$ where $i$ ranges from 1 to $p$. Therefore, the mean square errors of all possible linear combinations of the $\tilde{\beta}_i$s, used to estimate the corresponding linear combinations of $\beta_i$s, can be derived from the MMSE. Consequently, the trace of the MMSE defines the expected squared distance of the vector estimator $\tilde{\beta}$ from $\beta$ itself, expressed mathematically as $\text{SMSE}(\tilde{\beta}) = \text{tr}(\text{MMSE}(\tilde{\beta}))$ where SMSE denotes the scalar mean square error of the estimator $\tilde{\beta}$.

By holding $k > 0$ constant and applying the trace operator to Eq. (16.28), we obtain

$$\text{SMSE}(\hat{\alpha}_{k,d}) = \sum_{i=1}^{p}\frac{((1-d)\lambda_i + k)^2 + \alpha_i^2}{(\lambda_i+1)^2(\lambda_i+k)^2} + \sigma^2\sum_{i=1}^{p}\frac{\lambda_i(\lambda_i+(d+k))^2}{(\lambda_i+1)^2(\lambda_i+k)^2}. \tag{16.30}$$

By minimization of Eq. (16.30) with respect to $d$ while keeping $k$ fixed, we find that the optimal $d$ is

$$d_{\text{opt}} = \frac{\sum_{i=1}^{p} \frac{\lambda_i(\alpha_i^2 - \sigma^2)}{(\lambda_i+1)^2(\lambda_i+k)}}{\sum_{i=1}^{p} \frac{\lambda_i(\lambda_i\alpha_i^2 + \sigma^2)}{(\lambda_i+1)^2(\lambda_i+k)^2}}. \tag{16.31}$$

Unfortunately, $d_{\text{opt}}$ given by Eq. (16.31) depends on unknown model parameters. So, for practical purposes, we must replace these unknown parameters with appropriate estimates. Substituting $\alpha_i^2$ and $\sigma^2$ by unbiased estimates which are, respectively, as $\hat{\alpha}_i^2 - (\hat{\sigma}^2 / \lambda_i)$ (Liu, 1993; Schaefer et al., 1984; Özkale & Arıcan, 2015) and $\hat{\sigma}^2 = (y - X\hat{\beta})^\top H^{-1}(y - X\hat{\beta}) / (n - p)$ where $\hat{\alpha}$ and $\hat{\sigma}^2$ are the BLUE estimates, we derive the estimated optimal $d$ as

$$\hat{d}_{\text{opt}} = \frac{\sum_{i=1}^{p} \frac{\lambda_i(\hat{\alpha}_i^2 - \frac{\hat{\sigma}^2}{\lambda_i} - \hat{\sigma}^2)}{(\lambda_i+1)^2(\lambda_i+k)}}{\sum_{i=1}^{p} \frac{\lambda_i(\lambda_i(\hat{\alpha}_i^2 - \frac{\hat{\sigma}^2}{\lambda_i}) + \hat{\sigma}^2)}{(\lambda_i+1)^2(\lambda_i+k)^2}}. \tag{16.32}$$

## 16.5 Real Data Analysis of Greenhouse Gases

Greenhouse gases, including carbon dioxide, methane, and nitrous oxide, act as heat-trapping agents, contributing to global warming. Over the last 150 years, human activities, such as industrialization and deforestation, have been the main cause of the significant rise in greenhouse gases in Earth's atmosphere. According to the United States Environmental Protection Agency (EPA), the combustion of fossil fuels for electricity, heating, and transportation is the largest source of greenhouse gas emissions in the United States. In 2016, the transportation sector alone accounted for about 28.5% of the total emissions in the country. Most of the greenhouse gas emissions in this sector come from burning fossil fuels in cars, trucks (both light and heavy), buses, motorcycles, and railways. Other forms of transportation, such as commercial aircraft, ships, and boats, contribute to a smaller share of emissions within the transportation sector (Eurostat, 2018).

In this analysis, we employ a dataset consisting of 297 randomly selected measurements of fuel combustion in transportation (in million tons) from 27 regions. These regions include Austria, Belgium, Bulgaria, Croatia, Denmark, Estonia, the European Union, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, the Netherlands, Norway, Poland, Portugal, Romania,

Slovakia, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom. The dataset was systematically collected over a decade, from 2006 to 2016, covering the entire period.

To assess fuel combustion in transportation (measured in million tons), we gather repeated measurements across different categories such as cars, light-duty trucks, heavy-duty trucks-buses, motorcycles, and railways. This dataset is obtained from the official Eurostat website (Eurostat, 2018).

The response variable ($y$) is defined as fuel combustion in transportation (in million tons), with explanatory variables including fuel combustion in cars ($x_1$), light-duty trucks ($x_2$), heavy-duty trucks-buses ($x_3$), motorcycles ($x_4$), and railways ($x_5$). In this analysis, fuel combustion in cars, light-duty trucks, heavy-duty trucks-buses, motorcycles, and railways are considered fixed effects. Since the 27 regions are randomly selected from a broader geographical area, the regional influences on the response variable are treated as random effects. Therefore, we model this as a random intercept and slope model (RISM), which represents a specific case of an LMM.

Therefore, for $i = 1, \ldots, 27$ and $j = 1, \ldots, 11$,

$$y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + \beta_5 x_{ij5} + u_1 + u_2 t_{ij} + \varepsilon_{ij},$$

where $y_{ij}$ represents the $i$-th observation of the $j$-th area of the response, $x_{ij}$s denote the $i$-th observation of the $j$-th area of the explanatory variable $x_s$, $s = 1, \ldots, 5$, and $t_{ij}$ signifies time corresponding to $y_{ij}$.

We examine both unstructured (UN) and diagonal (UN(1)) variance-covariance models within the heterogeneous category, as well as variance components (VC) and compound symmetry (CS) models within the homogeneous variance-covariance category. Model selection is based on comparing criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). Subsequently, the unknown parameters in the covariance matrices are estimated using both maximum likelihood (ML) and restricted maximum likelihood (REML) methods. The variance-covariance structures obtained using RISM are detailed in Table 16.1.

**Table 16.1** Findings on the suitability of the model (statistical metrics for model fit)

| Covariance structures | Estimated covariance parameter methods | AIC | BIC |
|---|---|---|---|
| UN | ML | **337.30** | 374.24 |
| | REML | **362.03** | 398.76 |
| UN(1) | ML | 339.42 | **372.67** |
| | REML | 362.87 | **395.93** |
| VC | ML | 391.56 | 421.11 |
| | REML | 416.72 | 446.11 |
| CS | ML | 393.42 | 426.67 |
| | REML | 418.60 | 451.66 |

Based on the results shown in Table 16.1, the UN variance-covariance model performs best according to the AIC criterion, whereas the BIC criterion favors the UN(1) model for representing the structure of the variance-covariance matrix related to the response variable. In accordance with the principle established by Kass and Raftery (1995) that BIC provides consistent model selection within a fixed set of models (including the true model), while AIC tends to prefer more complex models as the amount of data increases, we choose the UN(1) model based on the BIC criterion. The estimated UN(1) variance-covariance matrix is then obtained as follows,

$$\hat{G}_{\text{ML}} = \begin{bmatrix} 2.1913 & 0 \\ 0 & 0.0755 \end{bmatrix}, \quad \hat{G}_{\text{REML}} = \begin{bmatrix} 2.4122 & 0 \\ 0 & 0.0756 \end{bmatrix},$$

and $\hat{W}_{\text{ML}} = 0.25451\, I_{297}$, $\hat{W}_{\text{REML}} = 0.25637\, I_{297}$. Subsequently, the estimates $\hat{H}_{\text{ML}}$ and $\hat{H}_{\text{REML}}$ are derived from the formula $H = ZGZ^{\top} + W$. However, since UN(1) under ML produces smaller AIC/BIC values compared to REML and, typically, there is no significant difference between the covariance estimates obtained from ML and REML (Özkale & Can, 2017), we choose the ML for our dataset.

The eigenvalues of the matrix $X^{\top}\hat{H}_{\text{ML}}^{-1}X$ are determined as

$$\begin{aligned} &\lambda_1 = 1.4326 \times 10^{+7}, \quad \lambda_2 = 1.5085 \times 10^{+4}, \quad \lambda_3 = 4.7251 \times 10^{+3}, \\ &\lambda_4 = 247.7243, \quad \lambda_5 = 41.5100. \end{aligned} \tag{16.33}$$

As a result, the condition number is determined to be $\lambda_{\max} / \lambda_{\min} = 345{,}120$. Since the condition number is a measure of multicollinearity and it exceeds 1000, this dataset has a severe degree of multicollinearity.

To estimate the value of $k$, we follow the approach of Özkale and Can (2017), by selecting the ridge parameter that minimize the generalized cross-validation (GCV) criterion for LMM, resulting in $\hat{k} = 9.9910$. For the estimated value of $d$, using Eq. (16.32), we find $\hat{d} = \hat{d}_{\text{opt}} = 0.9588$ with $\hat{\sigma}^2$ calculated as 5.1729. Table 16.2 summarizes the parameter estimates and SMSE values for $\hat{k} = 9.991$ and $\hat{d} = 0.9588$ where SMSE stands for the scalar MSE which is obtained by taking the trace of MMSE.

Table 16.2 clearly shows that it is evident that $\hat{\beta}_{\hat{k},\hat{d}}$ outperforms $\hat{\beta}$, $\hat{\beta}_{\hat{k}}$, and $\hat{\beta}_{\hat{d}}$ in terms of SMSE values for $\hat{k} = 9.991$ and $\hat{d} = 0.9588$. Additionally, we assess whether the theorems outlined in Sect. 16.3 hold true. Considering Theorem 16.1, we find that $\hat{k} = 9.991 > 0$ and $-1.0548 \times 10^{+2} < \hat{d} = \hat{d}_{\text{opt}} = 0.9588 < 1$ with $2.7619 < \hat{\sigma}^2 = 5.1729$. Since the conditions hold, we confirm that Theorem 16.1 is satisfied, yielding $\text{SMSE}(\hat{\beta}_{\hat{k},\hat{d}}) = 0.1462 < \text{SMSE}(\hat{\beta}) = 0.1469$.

For Theorem 16.2, we have $\hat{k} = 9.991 > 0$, $-74.0204 < \hat{d} = \hat{d}_{\text{opt}} = 0.9588 < 10.9792$, and $-16.5211 < \hat{\sigma}^2 = 5.1729$. The conditions are satisfied. Hence,

**Table 16.2** Parameter estimates of fixed and random effects and SMSE values for $\hat{k} = 9.991$ and $\hat{d} = \hat{d}_{\text{opt}} = 0.9588$

| | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | SMSE | | $u_1$ | $u_2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}$ | 1.0247 | 1.0500 | 0.9330 | 3.3436 | 3.6789 | 0.1469 | $\hat{u}$ | 0.5488 | − 0.0780 |
| $\hat{\beta}_{\hat{k}}$ | 1.0532 | 1.0640 | 0.9110 | 2.7696 | 2.9192 | 1.0097 | $\hat{u}_{\hat{k}}$ | 0.5919 | − 0.0843 |
| $\hat{\beta}_{\hat{d}}$ | 1.0248 | 1.0501 | 0.9329 | 3.3407 | 3.6751 | 0.1467 | $\hat{u}_{\hat{d}}$ | 0.5490 | − 0.0780 |
| $\hat{\beta}_{\hat{k},\hat{d}}$ | 1.0255 | 1.0504 | 0.9324 | 3.3281 | 3.6578 | 0.1462 | $\hat{u}_{\hat{k},\hat{d}}$ | 0.5500 | − 0.0782 |

Theorem 16.2 is confirmed, with the SMSE for $\hat{\beta}_{\hat{k},\hat{d}}$ being 0.1462, which is lower than the SMSE for $\hat{\beta}_{\hat{k}}$ at 1.0097.

In relation to Theorem 16.3, we observe that $\hat{k} = 9.991 > 0$, $\hat{d} = \hat{d}_{\text{opt}} = 0.9588 > 0$, or $\hat{d} = \hat{d}_{\text{opt}} = 0.9588 < 1.4326 \times 10^{+7}$, $3.5424 \times 10^{-8} < \hat{\sigma}^2 = 5.1729$. The conditions are met. Thus, Theorem 16.3 is validated, and $\text{SMSE}(\hat{\beta}_{\hat{k},\hat{d}}) = 0.1462 < \text{SMSE}(\hat{\beta}_{\hat{d}}) = 0.1467$.

The analysis of greenhouse gas data reveals that the selected values of $k$ and $d$, using the estimator $\hat{\beta}_{k,d}$, outperform $\hat{\beta}$, $\hat{\beta}_k$, and $\hat{\beta}_d$ by yielding lower SMSE values.

## 16.6 Conclusion

In this chapter, we introduce a novel ridge-based biased prediction technique that integrates best linear unbiased estimator (BLUE)/best linear unbiased predictor (BLUP), ridge estimators/predictors, and Liu estimators/predictors. We then compare these methods in terms of matrix mean squared error (MSE) and identify the optimal biasing parameters. The study is further supported by an analysis of greenhouse gas data.

Our findings illustrate that the $k - d$ estimator/predictor can serve as a viable alternative to ridge and Liu estimators/predictors in linear mixed models (LMMs), especially in the context of multicollinearity. Additionally, we show that the $k - d$ approach achieves a lower MSE—indicating better performance—compared to BLUE, ridge, and Liu estimators when the appropriate biasing parameters are chosen.

**Disclosure statement** The authors declare that there are no reported potential conflicts of interest.

## References

Demidenko, E. (2013). *Mixed models: Theory and applications with R*. Wiley.
Eurostat. (2018). Greenhouse gas emissions by source sector (source: EEA, env_air_gge). http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=env_air_gge&lang=en

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological), 38*, 248–250.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics, 51*, 1440–1450.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics, 21*, 309–310.

Henderson, C. R., Kempthorne, O., Searle, S. R., & Von Krosig, C. N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics, 15*, 192–218.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer.

Kaçıranlar, S., Sakallıoğlu, S., Akdeniz, F., Styan, G. P. H., & Werner, H. J. (1999). A new biased estimator in linear regression and a detailed analysis of the widely analyzed dataset on Portland cement. *Sankhya A—The Indian Journal of Statistics, 61B*, 443–459.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kuran, Ö. (2020). The modified ridge prediction method in the linear mixed models. In *Proceedings of the 3rd international congress on statistics, mathematics and analytical, Istanbul, Turkey* (pp. 33–43).

Kuran, Ö., & Özbay, N. (2021). Improving prediction by means of a two-parameter approach in linear mixed models. *Journal of Statistical Computation and Simulation, 91*, 3721–3743.

Kuran, Ö. (2023). A new two-parameter biased prediction class in linear mixed models. In *Proceedings of the 9th international conference on advances in statistics*, Azerbaijan-Baku (pp. 39–47).

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963–974.

Liu, K. J. (1993). A new class of biased estimate in linear regression. *Communications in Statistics—Theory and Methods, 22*, 393–402.

Liu, X.-Q., & Hu, P. (2013). General ridge predictors in a mixed linear model. *Statistics, 47*, 363–378.

Özkale, M. R., & Kaçıranlar, S. (2007). The restricted and unrestricted two-parameter estimators. *Communications in Statistics—Theory and Methods, 36*, 2707–2725.

Özkale, M. R., & Arıcan, E. (2015). First-order r-d class estimator in binary logistic regression model. *Statistics & Probability Letters, 106*, 19–29.

Özkale, M. R., & Can, F. (2017). An evaluation of ridge estimator in linear mixed models: An example from kidney failure data. *Journal of Applied Statistics, 44*, 2251–2269.

Özkale, M. R., & Kuran, Ö. (2020). A further prediction method in linear mixed models: Liu prediction. *Communications in Statistics—Simulation and Computation, 49*, 3171–3195.

Pereira, L. N., & Coelho, P. S. (2012). A small area predictor under area-level linear mixed models with restrictions. *Communications in Statistics—Theory and Methods, 41*, 2524–2544.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects in S and S-Plus*. Springer.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science, 6*, 15–51.

Sakallıoğlu, S., & Kaçıranlar, S. (2008). A new biased estimator based on ridge estimation. *Statistical Papers, 49*, 669–689.

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics—Theory and Methods, 13*, 99–113.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. Wiley.

Štulajter, F. (1997). Predictions in nonlinear regression models. *Acta Mathematica Universitatis Comenianae, 66*, 71–81.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics—Theory and Methods, A5*, 1065–1075.

Trenkler, G. (1980). Generalized mean square error comparisons of biased regression estimators. *Communications in Statistics—Theory and Methods, 9*, 1247–1259.
Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.
Yang, H., Ye, H., & Xue, K. (2014). A further study of predictions in linear mixed models. *Communications in Statistics—Theory and Methods, 43*, 4241–4252.

# Chapter 17
# L-Estimation of Location: Shrinkage and Selection




**A. K. Md. Ehsanes Saleh and Bashir U. Khan**

## 17.1 Introduction

The L-statistics is a linear combination of order statistics. The L-statistics form an important class of estimators in nonparametric statistics. Its members include trimmed means and sample quantiles. The trimmed means provide a robust alternative to sample mean for estimating a location parameter (Wilcox, 2012). The sample quantiles are used for nonparametric estimation of population quantiles and their functions such as the interquartile range (David & Ragaraja, 2003). The main benefits of L-estimators are that they are often extremely simple and often robust statistics: assuming sorted data, they are very easy to calculate and interpret and often resistant to outliers.

There is an extensive literature on L-statistics and their asymptotic properties in the case of independent and identically distributed (i.i.d.) observations. See, for example, Serfling (1980), Huber (1981), and Saleh (1992). The non-i.i.d. cases have also received considerable attention. See, for example, Bickel (1973), Koenker and Bassett (1978), Welsh (1987), and Koenker and Portony (1987), to name a few, who have established many asymptotic properties of L-estimators in linear regression problems. While L-statistics are not as efficient as other statistics, they often have high relative efficiency and show that a large fraction of the information used in estimation can be obtained using only a few points – as few as one, two, or three.

A. K. Md. E. Saleh
School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada
e-mail: esaleh@math.carleton.ca

B. U. Khan (✉)
Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS, Canada
e-mail: bkhan@smu.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_17

Consider a random ordered sample $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ from a known continuous distribution $F_o((y-\theta)/\sigma)$ with location parameter $\theta$ and scale parameter $\sigma$ having p.d.f. $(1/\sigma) f_o((y-\theta)/\sigma)$.

Let us consider a predecided size $k$ ($\leq n$) of the spacings $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k \leq 1$. Define $y_{[n\lambda_j]+1}$, $j = 1, \ldots, k$ as the sample quantile of order $\lambda_j$. It is well-known that the joint distribution of $(y_{[n\lambda_j]+1} | j = 1, \ldots, k)^\top$ is $k$-dimensional normal distribution with asymptotic mean $(\theta + \sigma Q_o(\lambda_j) | j = 1, \ldots, k)$ and $k \times k$ covariance matrix

$$\sigma^2 \Omega = \sigma^2 \frac{\lambda_i^\top \Lambda \lambda_j - \lambda_i \lambda_j}{q_o(\lambda_i) q_o(\lambda_j)}, \tag{17.1}$$

where $Q_o(\lambda_j) = F_o^{-1}(\lambda_j)$ and $q_o(\lambda_j) = f_o(Q_o(\lambda_j))$ are the quantile and quantile density functions of order $\lambda_j$, respectively, as $n \to \infty$. We can write

$$Y_{[n\lambda_j]+1} = \theta + \sigma e_{[n\lambda_j]+1}$$

and

$$\mathrm{E}[Y_{[n\lambda_j]+1}] = \theta + \sigma\, \mathrm{E}[e_{[n\lambda_j]+1}],$$

where $e_{[n\lambda_j]+1}$ is the estimation error associated with the sample quantile of order $\lambda_j$, $[n\lambda_j]$ stands for the largest integer of $n\lambda_j$, and $\mathrm{E}[Y_{[n\lambda_j]+1}] = Q_0(\lambda_j)$ as $n \to \infty$ (see David, 1981). Thus, we can write

$$\mathrm{E}\left[Y_{[n\lambda_j]+1}\right] = \theta + \sigma\ Q_0(\lambda_j), \quad j = 1, \ldots, k\,. \tag{17.2}$$

The unrestricted L-estimators (ULE) of $\theta$ and $\sigma$ are obtained from minimizing the following expression,

$$(\tilde{\theta}_n, \tilde{\sigma}_n)^\top == \left(\boldsymbol{y}_{(\cdot)} - \theta \mathbf{1}_k - \sigma \boldsymbol{u}\right)^\top \Omega^{-1} \left(\boldsymbol{y}_{(\cdot)} - \theta \mathbf{1}_k - \sigma \boldsymbol{u}\right), \tag{17.3}$$

where $\boldsymbol{y}_{(\cdot)} = (y_{[n\lambda_1]+1}, \ldots, y_{[n\lambda_k]+1})^\top$ and $\boldsymbol{u} = (u_1, \ldots, u_k)^\top$ with $u_j = Q_o(\lambda_j)$.

Differentiating the right-hand side of Eq. (17.3) with respect to $\theta$ and $\sigma$, we get the following equations:

$$\begin{aligned} (\mathbf{1}_k^\top \Omega^{-1} \mathbf{1}_k)\theta + (\mathbf{1}_k^\top \Omega^{-1} \boldsymbol{u})\sigma &= \mathbf{1}^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} \\ (\boldsymbol{u}^\top \Omega^{-1} \mathbf{1}_k)\theta + (\boldsymbol{u}^\top \Omega^{-1} \boldsymbol{u})\sigma &= \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} \end{aligned} \tag{17.4}$$

Solving the equations in (17.4), we obtain

$$\begin{pmatrix} \tilde{\theta}_n \\ \tilde{\sigma}_n \end{pmatrix} = \begin{pmatrix} K_1 & K_3 \\ K_3 & K_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} \\ \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} \end{pmatrix}, \tag{17.5}$$

where $K_1 = \mathbf{1}_k^\top \Omega^{-1} \mathbf{1}_k$, $K_2 = \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{u}$, and $K_3 = \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{u}$. Explicitly, we obtain

$$\begin{pmatrix} \tilde{\theta}_n \\ \tilde{\sigma}_n \end{pmatrix} = \frac{1}{K_1 K_2 - K_3^2} \begin{pmatrix} K_2 \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} - K_3 \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} \\ K_1 \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} - K_3 \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{y}_{(\cdot)} \end{pmatrix}, \tag{17.6}$$

where

$$K_1 = \sum_{j=1}^{k} \frac{\big(q_o(\lambda_j) - q_o(\lambda_{j-1})\big)^2}{(\lambda_j - \lambda_{j-1})},$$

$$K_2 = \sum_{j=1}^{k} \frac{\big(q_o(\lambda_j) Q_o(\lambda_j) - q_o(\lambda_{j-1}) Q_o(\lambda_{j-1})\big)^2}{(\lambda_j - \lambda_{j-1})},$$

and

$$K_3 = \sum_{j=1}^{k} \frac{\big(q_o(\lambda_j) - q_o(\lambda_{j-1})\big)\big(q_o(\lambda_j) Q_o(\lambda_j) - q_o(\lambda_{j-1}) Q_o(\lambda_{j-1})\big)}{(\lambda_j - \lambda_{j-1})}.$$

The variance-covariance matrix of $\tilde{\theta}_n$ and $\tilde{\sigma}_n$ is given by

$$\operatorname{cov}(\tilde{\theta}_n, \tilde{\sigma}_n) = \frac{\sigma^2}{n} \frac{1}{K_1 K_2 - K_3^2} \begin{pmatrix} K_2 & K_3 \\ K_3 & K_1 \end{pmatrix}. \tag{17.7}$$

The two estimators, $\tilde{\theta}_n$ and $\tilde{\sigma}_n$, can also be written as

$$\tilde{\theta}_n = a_1 y_{[n\lambda_1]+1} + \cdots + a_k y_{[n\lambda_k]+1}$$

and

$$\tilde{\sigma}_n = b_1 y_{[n\lambda_1]+1} + \cdots + b_k y_{[n\lambda_k]+1},$$

where

$$a_j = \frac{q_o(\lambda_j) K_2}{K_1 K_2 - K_3^2} \left\{ \frac{q_o(\lambda_j) - q_o(\lambda_{j-1})}{\lambda_j - \lambda_{j-1}} - \frac{q_o(\lambda_{j+1}) - q_o(\lambda_j)}{\lambda_{j+1} - \lambda_j} \right\}$$

$$- \frac{q_o(\lambda_j) K_3}{K_1 K_2 - K_3^2} \left\{ \frac{q_o(\lambda_j) Q_o(\lambda_j) - q_o(\lambda_{j-1}) Q_o(\lambda_{j-1})}{\lambda_j - \lambda_{j-1}} \right.$$

$$\left. - \frac{q_o(\lambda_{j+1}) Q_o(\lambda_{j+1}) - q_o(\lambda_j) Q_o(\lambda_j)}{\lambda_{j+1} - \lambda_j} \right\}$$

and for $j = 1, \ldots, k$

$$b_j = \frac{q_o(\lambda_j)K_1}{K_1K_2 - K_3^2}\left\{\frac{q_o(\lambda_j)Q_o(\lambda_j) - q_o(\lambda_{j-1})Q_o(\lambda_{j-1})}{\lambda_j - \lambda_{j-1}}\right.$$
$$\left. - \frac{q_o(\lambda_{j+1})Q_o(\lambda_{j+1}) - q_o(\lambda_j)Q_o(\lambda_j)}{\lambda_j - \lambda_{j-1}}\right\}$$
$$- \frac{q_o(\lambda_j)K_3}{K_1K_2 - K_3^2}\left\{\frac{q_o(\lambda_j) - q_o(\lambda_{j-1})}{\lambda_j - \lambda_{j-1}} - \frac{q_o(\lambda_{j+1}) - q_o(\lambda_j)}{\lambda_{j+1} - \lambda_j}\right\}.$$

If the distribution $F$ is symmetric about 0, then $K_3 = 0$, and all the above expressions can be written in a simplified form as

$$\tilde{\theta}_n = \frac{1}{K_1}\mathbf{1}_k\Omega^{-1}\boldsymbol{y}_{(\cdot)} \quad \text{and} \quad \tilde{\sigma}_n = \frac{1}{K_2}\boldsymbol{u}\Omega^{-1}\boldsymbol{y}_{(\cdot)}. \tag{17.8}$$

$a_j$ and $b_j$ reduce to

$$a_j = \frac{q_o(\lambda_j)}{K_1}\left\{\frac{q_o(\lambda_j) - q_o(\lambda_{j-1})}{\lambda_j - \lambda_{j-1}}\right\}$$

and

$$b_j = \frac{q_o(\lambda_j)}{K_2}\left\{\frac{q_o(\lambda_j)Q_o(\lambda_j) - q_o(\lambda_{j-1})Q_o(\lambda_{j-1})}{\lambda_j - \lambda_{j-1}}\right\},$$

respectively, for $j = 1, \cdots k$. Further, the variance-covariance matrix given in Eq. (17.7) becomes

$$\text{cov}(\tilde{\theta}_n, \tilde{\sigma}_n) = \frac{\sigma^2}{n}\begin{pmatrix} \frac{1}{K_1} & 0 \\ 0 & \frac{1}{K_2} \end{pmatrix}. \tag{17.9}$$

If the distribution and density functions are unknown, we then use their consistent estimators $F_n(Q_0(\lambda_j))$ and $f_n(Q_0(\lambda_j))$ of the quantile density function.

In this chapter, we propose L-estimators for the estimation of location parameter $\theta$ of a continuous distribution. We use $L_2$-risk criterion to find the asymptotic distribution risk expressions of the proposed estimators. In order to gauge the performances of these estimators, we compare their asymptotic distributional risks and the asymptotic relative efficiencies. In Sect. 17.2, we define the L-estimators of the location parameter $\theta$ and derive the expressions for their asymptotically distributed biases, the $L_2$-risks, and the relative efficiencies. The performance analyses of the proposed L-estimators based on their asymptotic distributional risks

and the relative efficiencies are discussed in Sect. 17.3. Section 17.4 provides some tabulated results for the optimum spacing considering the logistic distribution as an example. Some concluding remarks are presented in Sect. 17.5.

## 17.2 Asymptotic Distributional Biases and $L_2$-Risk of the Proposed L-Estimators

In this section, we derive and discuss the asymptotic distributional bias (ADB) and the asymptotic distributional $L_2$-risk (ADR) expressions of the estimators of $\theta$.

In this study, $\tilde{\theta}_n$ is the unrestricted L-estimator (ULE) of $\theta$. The ADB and ADR of $\tilde{\theta}_n$ are given by

$$\text{ADB}(\sqrt{n}\tilde{\theta}_n) = 0, \quad \text{and} \quad \text{ADR}(\sqrt{n}\tilde{\theta}_n) = \frac{\sigma^2}{K_1}, \tag{17.10}$$

where $K_1$ depends on the spacing $0 < \lambda_1 < \cdots < \lambda_k < 1$. To obtain the optimum spacings, we minimize $K_1^{-1}$; i.e., maximize $K_1$ with respect to $\lambda_1, \ldots, \lambda_k$, and present some results for the optimum spacings when the underlying $F_o$ is the logistic or normal distribution.

The restricted L-estimator (RLE) of the parameter $\theta$ is obtained when we utilize the nonsample information in the estimation procedure. Without loss of generality, assuming that $\theta$ may be 0, then the restricted L-estimator of $\theta$ is also 0. The ADB and ADR of the restricted L-estimator are given by

$$\text{ADB}(0) = \sqrt{n}(0 - \theta_{(n)}) = -\delta, \quad \text{where} \quad \theta_{(n)} = n^{-\frac{1}{2}}\,\delta.$$

$$\text{ADR}(0) = \delta^2 = \frac{\sigma^2}{K_1}\Delta^2.$$

The asymptotic distributional risk efficiency (ADRE) of the restricted L-estimator relative to the unrestricted L-estimator (ULE) is

$$\text{ADRE}(0\,;\sqrt{n}\tilde{\theta}_n) = \Delta^{-2},$$

where $\Delta^2 = K_1\delta^2\,/\,\sigma^2$.

### 17.2.1 Shrinkage L-Estimator of $\theta$

The shrinkage L-estimator, $\hat{\theta}_n(c)$, is defined as

$$\hat{\theta}_n(c) = c\,\tilde{\theta}_n, \quad 0 < c < 1. \tag{17.11}$$

The ADB and ADR functions of $\hat{\theta}_n(c)$ are respectively given by

$$\begin{aligned}
\text{ADB}(\sqrt{n}\hat{\theta}_n(c)) &= \sqrt{n}\; c\; (\tilde{\theta}_n - \theta_{(n)}) \\
&= \sqrt{n}(1-c)\theta_{(n)} \\
&= -(1-c)\delta \\
\text{ADR}(\sqrt{n}\hat{\theta}_n(c)) &= c^2\; \text{var}(\sqrt{n}\tilde{\theta}_n) + (1-c)^2\delta^2 \\
&= c^2\frac{\sigma^2}{K_1} + (1-c)^2\delta^2 \\
&= \frac{\sigma^2}{K_1}\Big(c^2 + (1-c)^2\Delta^2\Big).
\end{aligned} \tag{17.12}$$

Replacing $c$ by its optimum value $c^* = \Delta^2 / (1 + \Delta^2)$, the ADR in Eq. (17.12) can be written as

$$\text{ADR}(\hat{\theta}_n(c^*)) = \frac{\sigma^2}{K_1}\,\frac{\Delta^2}{1+\Delta^2}\,. \tag{17.13}$$

Thus, the ADRE of $\hat{\theta}_n(c^*)$ relative to $\tilde{\theta}_n$ is given by

$$\text{ADRE}(\hat{\theta}_n(c^*)\,;\tilde{\theta}_n) = 1 + \frac{1}{\Delta^2}\,. \tag{17.14}$$

### 17.2.2 Ridge-Type L-Estimator of $\theta$

If we replace $c$ by $1 / (1 + \kappa)$ in Eq. (17.11), what we obtained is the ridge-type L-estimator of $\theta$ denoted by $\hat{\theta}_n^{\text{R}}(\kappa)$ and is defined as

$$\hat{\theta}_n^{\text{R}}(\kappa) = \frac{\tilde{\theta}_n}{1+\kappa}, \quad 0 < \kappa < \infty. \tag{17.15}$$

In order to derive Eq. (17.15), we consider

$$\underset{\theta,\sigma}{\text{argmin}}\Big\{\,\big(\boldsymbol{y}_{(\cdot)} - \theta\mathbf{1}_k - \sigma\boldsymbol{u}\big)^\top\,\Omega^{-1}\,\big(\boldsymbol{y}_{(\cdot)} - \theta\mathbf{1}_k - \sigma\boldsymbol{u}\big) + \frac{1}{2}K_1\,\kappa\,\theta^2\Big\}.$$

Differentiating the expression inside the braces with respect to $\theta$ and $\sigma$, we get the following equations,

$$-\mathbf{1}_k^\top \Omega^{-1} \boldsymbol{y} + \theta \mathbf{1}_k^\top \Omega^{-1} \mathbf{1}_k + \sigma \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{u} + K_1 \kappa \theta = 0$$

$$-\boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y} + \theta \boldsymbol{u}^\top \Omega^{-1} \mathbf{1}_k + \sigma \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{u} = 0$$

or, equivalently,

$$(1+\kappa) K_1 \theta + K_3 \sigma = \mathbf{1}_k^\top \Omega^{-1} \mathbf{y}$$

$$K_3 \theta + \sigma K_2 = \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y}.$$

The solutions of the above equations for $\theta$ and $\sigma$ are as follows,

$$\begin{pmatrix} \hat{\theta}_n^{\mathrm{R}}(\kappa) \\ \tilde{\sigma}_n \end{pmatrix} = \begin{pmatrix} (1+\kappa) K_1 & K_3 \\ K_3 & K_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{y} \\ \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \tilde{\theta}_n \,/\, (1+\kappa) \\ \tilde{\sigma}_n \end{pmatrix},$$

for $K_3 = 0$ assuming $F_o$ is symmetric about 0.

The ADB and ADR of $\hat{\theta}_n^{\mathrm{R}}(\kappa)$ are given by

$$\mathrm{ADB}\left(\sqrt{n}\, \hat{\theta}_n^{\mathrm{R}}(\kappa)\right) = -\frac{\kappa\, \delta}{1+\kappa} \tag{17.16}$$

and

$$\mathrm{ADR}\left(\sqrt{n}\, \hat{\theta}_n^{\mathrm{R}}(\kappa)\right) = \frac{\sigma^2}{K_1} \frac{1+\kappa^2 \Delta^2}{(1+\kappa)^2}. \tag{17.17}$$

The optimum value of $\kappa$ is $\kappa_{\mathrm{opt}} = \Delta^{-2}$, and the ADR becomes

$$\mathrm{ADR}\left(\hat{\theta}_n^{\mathrm{R}}(\kappa_{\mathrm{opt}})\right) = \frac{\sigma^2}{K_1} \frac{\Delta^2}{1+\Delta^2}. \tag{17.18}$$

The ADRE of $\hat{\theta}_n^{\mathrm{R}}(\kappa_{\mathrm{opt}})$ relative to $\tilde{\theta}_n$ can be written as

$$\mathrm{ADRE}\left(\hat{\theta}_n^{\mathrm{R}}(\kappa_{\mathrm{opt}})\,;\, \tilde{\theta}_n\right) = 1 + \frac{1}{\Delta^2}. \tag{17.19}$$

### 17.2.3 LASSO L-Estimator of $\theta$

Following Donoho and Johnstone (1994), we define least absolute selection and shrinkage (LASSO) or soft-threshold L-estimator as

$$\hat{\theta}_n^{\text{LASSO}}(\kappa) = \text{sgn}(\sqrt{n}\tilde{\theta}_n)\left(|\sqrt{n}\tilde{\theta}_n| - \kappa\frac{\sigma}{\sqrt{K_1}}\right) I\left(|\sqrt{n}\tilde{\theta}_n| > \kappa\ \frac{\sigma}{\sqrt{K_1}}\right)$$
$$= \frac{\sigma}{\sqrt{K_1}}\text{sgn}(Z_n)\,(|Z_n| - \kappa)\ I\,(|Z_n| > \kappa) \tag{17.20}$$

where $\kappa$ is the regularization parameter that controls the amount of shrinkage, $I(A)$ is the indicator function of $A$, and

$$Z_n = \frac{\sqrt{nK_1}\,\tilde{\theta}_n}{\sigma} \sim N(\Delta, 1) \quad \text{as} \quad n \to \infty.$$

The ADB and ADR of the LASSO L-estimator of $\theta$ are respectively given by

$$\text{ADB}\left(\hat{\theta}_n^{\text{LASSO}}(\kappa)\right) = -\frac{\sigma}{\sqrt{K_1}}\Big(\Delta H_3\left(\kappa^2; \Delta^2\right) - \kappa\ \left(\Phi(\kappa - \Delta) - \Phi(\kappa + \Delta)\right)\Big) \tag{17.21}$$

and

$$\text{ADR}\left(\hat{\theta}_n^{\text{LASSO}}(\kappa)\right) = \frac{\sigma^2}{K_1}\,\gamma_{\text{st}}(\kappa; \Delta) \tag{17.22}$$

where

$$\gamma_{\text{st}}(\kappa; \Delta) = \Big(1 - H_3(\kappa^2; \Delta^2)\Big) + \Delta^2\ \Big(2H_3(\kappa^2; \Delta^2) - H_5(\kappa^2; \Delta^2)\Big)$$
$$+\kappa^2\Big(1 - H_3(\kappa^2; \Delta^2)\Big) - 2\kappa\Delta\big\{\Phi(\kappa - \Delta) - \Phi(\kappa + \Delta)\big\} + 4\kappa\phi(\Delta),$$

where $H_\nu(\kappa^2, \Delta^2)$ is the c.d.f. of a chi-square distribution with $\nu$ degrees of freedom and the noncentrality parameter $\Delta^2$ and $\phi$ and $\Phi$ are the p.d.f and c.d.f. of $N(\Delta; 1)$, respectively.

The ADRE of LASSO L-estimator relative to $\tilde{\theta}_n$ is

$$\text{ADRE}\left(\hat{\theta}_n^{\text{LASSO}}(\kappa); \tilde{\theta}_n\right) = (\gamma_{\text{st}}(\kappa; \Delta))^{-1}. \tag{17.23}$$

### 17.2.4 Preliminary Test L-Estimator (PTLE) of θ

The preliminary test L-estimator of $\theta$ can be written as

$$\sqrt{n}\,\hat{\theta}_n^{\mathrm{PT}}(\kappa) = \sqrt{n}\,\tilde{\theta}_n - \sqrt{n}\,\tilde{\theta}_n\, I(|Z_n| \le \kappa)\,. \tag{17.24}$$

Hence, under the local alternatives $K_{(n)} : \theta_{(n)} = -\frac{1}{2}n\delta$, the ADB and ADR of $\hat{\theta}_n^{\mathrm{PT}}(\kappa)$ are given by

$$\begin{aligned}
&\mathrm{ADB}\left(\sqrt{n}\,\hat{\theta}_n^{\mathrm{PT}}(\kappa)\right) = -\frac{\sigma}{\sqrt{K_1}}\Delta H_3(\kappa^2;\Delta^2) \qquad (17.25)\\
&\mathrm{ADR}\left(\sqrt{n}\,\hat{\theta}_n^{\mathrm{PT}}(\kappa)\right) = \frac{\sigma^2}{\sqrt{K_1}}\Big[\Big(1 - H_3(\kappa^2;\Delta^2)\Big)\\
&+\Delta^2\big\{2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\big\}\Big].
\end{aligned}$$

And the ADRE of $\hat{\theta}_n^{\mathrm{PT}}(\kappa)$ relative to $\tilde{\theta}_n$ is

$$\begin{aligned}
\mathrm{ADRE}\left(\hat{\theta}_n^{\mathrm{PT}}(\kappa)\,;\tilde{\theta}_n\right) = &\Big[\Big(1 - H_3(\kappa^2;\Delta^2)\Big)\\
&+\Delta^2\big\{2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\big\}\Big]^{-1}. \qquad (17.26)
\end{aligned}$$

### 17.2.5 Saleh-Type L-Estimator of θ

The Saleh-type L-estimator (Saleh, 2006) of $\theta$ is defined as

$$\sqrt{n}\,\hat{\theta}_n^{\mathrm{S}}(\kappa) = \frac{\sigma}{\sqrt{K_1}}\left(Z_n - \kappa\,\frac{Z_n}{|Z_n|}\right) = \frac{\sigma}{\sqrt{K_1}}\left(Z_n - \kappa\,\mathrm{sgn}(Z_n)\right), \tag{17.27}$$

where $Z_n = \sqrt{n}\tilde{\theta}_n$ and $\kappa$ is an appropriate nonnegative bounded constant. The asymptotic distributional bias (ADB) and the asymptotic distributional risk (ADR) of $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ are given by

$$\mathrm{ADB}\left(\hat{\theta}_n^{\mathrm{S}}(\kappa)\right) = -\kappa\frac{\sigma}{\sqrt{K_1}}\left(2\Phi(\Delta) - 1\right) \tag{17.28}$$

and

$$\mathrm{ADR}\left(\hat{\theta}_n^{\mathrm{S}}(\kappa)\right) = \frac{\sigma^2}{K_1}\left[1 + \kappa^2 - 2\kappa\sqrt{\frac{2}{\pi}}\,\exp\{-\Delta^2/2\}\right]. \tag{17.29}$$

For the optimum value of $\kappa$, that is, $\kappa_{\text{opt}} = \sqrt{\frac{2}{\pi}}$ which occurs at $\Delta = 0$, the asymptotic distributional risk of $\hat{\theta}_n^{\text{S}}(\kappa)$ can be written as

$$\text{ADR}\left(\hat{\theta}_n^{\text{S}}(\kappa)\right) = \frac{\sigma^2}{K_1}\Big[1 - \frac{2}{\pi}\,\Big(2\exp\{-\Delta^2/2\} - 1\Big)\Big]. \tag{17.30}$$

The asymptotic distributional risk efficiency of $\hat{\theta}_n^{\text{S}}(\kappa)$ relative to $\tilde{\theta}_n$ is

$$\text{ADRE}\left(\hat{\theta}_n^{\text{S}}(\kappa)\,;\,\tilde{\theta}_n\right) = \Big[1 - \frac{2}{\pi}\,\Big(2\exp\{-\Delta^2/2\} - 1\Big)\Big]^{-1}. \tag{17.31}$$

### *17.2.6 Elastic Net L-Estimator of θ*

The Elastic Net L-estimator is a convex combination of ridge L-estimator and LASSO L-estimator. We define the elastic net L-estimator of $\theta$ subject to $K_3 = 0$ as

$$\begin{aligned}(\hat{\theta}_n^{\text{Enet}}(\kappa;\alpha), \hat{\sigma}_n^{\text{Enet}}(\kappa;\alpha))^\top = \underset{\theta,\sigma}{\text{argmin}}\Big\{ &\left(\boldsymbol{y}_{(\cdot)} - \theta\mathbf{1}_k - \sigma\boldsymbol{u}\right)^\top \Omega^{-1}\left(\boldsymbol{y}_{(\cdot)} - \theta\mathbf{1}_k - \sigma\boldsymbol{u}\right) \\ &+ \kappa\left(\sqrt{K_1}\,\sigma\,\alpha\,\theta + \frac{K_1}{2}(1-\alpha)\,\theta^2\right)\Big\}.\end{aligned} \tag{17.32}$$

The result in Eq. (17.32) is obtained by solving the following two equations:

$$\begin{aligned}\big[1 + \kappa(1-\alpha)\big]K_1\theta &= \mathbf{1}_k^\top\Omega^{-1}\boldsymbol{y} - K_1\kappa\,\alpha\,\sigma\,\text{sgn}(\theta) \\ K_2\sigma &= \boldsymbol{u}^\top\Omega^{-1}\boldsymbol{y}\end{aligned}$$

Equation (17.32) can also be written as

$$\begin{aligned}\begin{pmatrix}\hat{\theta}_n^{\text{Enet}}(\kappa;\alpha)\\ \\ \hat{\sigma}_n^{\text{Enet}}(\kappa;\alpha)\end{pmatrix} &= \begin{pmatrix}\big[1+\kappa(1-\alpha)\big]K_1 & \mathbf{0}\\ \\ \mathbf{0} & K_2\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}_k^\top\Omega^{-1}\boldsymbol{y} - \sqrt{K_1}\,\kappa\alpha\sigma\,\text{sgn}(\theta)\\ \\ \boldsymbol{u}^\top\Omega^{-1}\boldsymbol{y}\end{pmatrix} \\ &= \begin{pmatrix}\big[1+\kappa(1-\alpha)\big]^{-1}\left[\tilde{\theta}_n - \frac{\kappa\alpha\sigma}{\sqrt{K_1}}\,\text{sgn}(\tilde{\theta}_n)\right]\\ \\ \hat{\sigma}_n^{\text{Enet}}.\end{pmatrix}\end{aligned}$$

Using the above, we can write the elastic net L-estimator of $\theta$ as

$$\hat{\theta}_n^{\text{Enet}}(\kappa;\alpha) = \frac{\sigma}{\sqrt{K_1}}\big[1+\kappa(1-\alpha)\big]^{-1}\ \text{sgn}(Z_n)\,(|Z_n|-\kappa\alpha)^+, \tag{17.33}$$

where $\kappa > 0$ and $0 < \alpha < 1$. The $\alpha$ is a mixing parameter between ridge and LASSO L-estimators. Note that for $\alpha = 0$

$$\hat{\theta}_n^{\text{Enet}}(\kappa;0) = \frac{\sigma}{\sqrt{K_1}}\big(1+\kappa\big)^{-1} Z_n$$

is a ridge L-estimator and for $\alpha = 1$

$$\hat{\theta}_n^{\text{Enet}}(\kappa;1) = \frac{\sigma}{\sqrt{K_1}}\ \text{sgn}(Z_n)\,(|Z_n|-\kappa)^+$$

is a LASSO L-estimator. The ADB and ADR expressions of the elastic net L-estimator of $\theta$ can respectively be written as

$$\begin{aligned}\text{ADB}\left(\hat{\theta}_n^{\text{Enet}}(\kappa;\alpha)\right) = -\frac{\sigma}{\sqrt{K_1}}\big(1+\kappa(1-\alpha)\big)^{-1}\Big[&\Delta H_3\left(\kappa^2\alpha^2;\Delta^2\right)\\ &+\kappa(1-\alpha)\Delta-\kappa\alpha\big\{\Phi(\kappa\alpha-\Delta)-\Phi(\kappa\alpha+\Delta)\big\}\Big]\end{aligned}$$

and

$$\begin{aligned}\text{ADR}\left(\hat{\theta}_n^{\text{Enet}}(\kappa;\alpha)\right) =\frac{\sigma^2}{K_1}\,[1+\kappa\alpha)]^{-2}\Big[&\gamma_{\text{st}}(\kappa\alpha,\,\Delta)+\kappa^2(1-\alpha)^2\Delta^2\\ &+2\kappa(1-\alpha)\Delta^2 H_3(\kappa^2\alpha^2;\,\Delta^2)\\ &-2\kappa^2\alpha(1-\alpha)\Delta\big\{\Phi(\kappa\alpha-\Delta)-\Phi(\kappa\alpha+\Delta)\big\}\Big],\end{aligned} \tag{17.34}$$

where $\gamma_{\text{st}}(\kappa,\,\Delta)$ is already defined in Eq. (17.22), and replacing "$\kappa$" by "$\kappa\alpha$" in that equation results in $\gamma_{\text{st}}(\kappa\alpha;\,\Delta)$. Note that if $\alpha = 0$, then

$$\text{ADR}\left(\hat{\theta}_n^{\text{Enet}}(\kappa;0)\right) = \frac{\sigma^2}{K_1}(1+\kappa)^{-2}(1+\kappa^2\Delta^2)$$

and if $\alpha = 1$, then

$$\text{ADR}\left(\hat{\theta}_n^{\text{Enet}}(\kappa;1)\right) = \frac{\sigma^2}{K_1}\,\gamma_{\text{st}}(\kappa;\,\Delta).$$

It is clear from Table 17.1 that the ridge-type L-estimator, $\hat{\theta}_n^{\text{R}}(\kappa)$, uniformly dominates all the other estimators of $\theta$. The performance of preliminary test L-

**Table 17.1** Asymptotic distributional risk efficiencies of all the estimators of $\theta$

| $\Delta^2$ | $\hat{\theta}_n$ | $\hat{\theta}_n^{\mathrm{R}}$ | $\hat{\theta}_n^{\mathrm{PT}}$ | $\hat{\theta}_n^{\mathrm{S}}$ | $\hat{\theta}_n^{\mathrm{LASSO}}$ | $\hat{\theta}_n^{\mathrm{Enet}}$ |
|---|---|---|---|---|---|---|
| 0 | $\infty$ | $\infty$ | 1.411 | 2.752 | 4.303 | 1.977 |
| 0.01 | 100.000 | 101.000 | 1.399 | 2.705 | 4.199 | 1.965 |
| 0.02 | 50.000 | 51.000 | 1.387 | 2.659 | 4.101 | 1.953 |
| 0.03 | 33.333 | 34.333 | 1.375 | 2.615 | 4.007 | 1.942 |
| 0.04 | 25.000 | 26.000 | 1.364 | 2.573 | 3.918 | 1.930 |
| 0.05 | 20.000 | 21.000 | 1.353 | 2.533 | 3.833 | 1.919 |
| 0.1 | 10.000 | 11.000 | 1.301 | 2.350 | 3.463 | 1.865 |
| 0.2 | 5.000 | 6.000 | 1.213 | 2.064 | 2.916 | 1.770 |
| 0.3 | 3.333 | 4.333 | 1.141 | 1.849 | 2.532 | 1.688 |
| 0.4 | 2.500 | 3.500 | 1.082 | 1.683 | 2.248 | 1.616 |
| 0.5 | 2.000 | 3.000 | 1.031 | 1.550 | 2.029 | 1.553 |
| 1 | 1.000 | 2.000 | 0.869 | 1.157 | 1.417 | 1.329 |
| $2\ln(2)$ | 0.721 | 1.721 | 0.801 | 1.000 | 1.187 | 1.218 |
| 2 | 0.500 | 1.500 | 0.743 | 0.856 | 0.980 | 1.103 |
| 3 | 0.333 | 1.333 | 0.713 | 0.739 | 0.813 | 0.998 |
| 4 | 0.250 | 1.250 | 0.720 | 0.683 | 0.730 | 0.942 |
| 5 | 0.200 | 1.200 | 0.746 | 0.653 | 0.685 | 0.910 |
| 10 | 0.100 | 1.100 | 0.908 | 0.614 | 0.619 | 0.867 |
| 20 | 0.050 | 1.050 | 0.996 | 0.611 | 0.611 | 0.863 |
| 30 | 0.033 | 1.033 | 1.000 | 0.611 | 0.611 | 0.863 |
| 40 | 0.025 | 1.025 | 1.000 | 0.611 | 0.611 | 0.863 |
| 50 | 0.020 | 1.020 | 1.000 | 0.611 | 0.611 | 0.863 |

estimator, $\hat{\theta}_n^{\mathrm{PT}}(\kappa)$, is better than Saleh-type L-estimator, $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ in the interval $[0, 0.316] \cup [2.236, \infty]$, while outside this interval, the performance of $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ is better than $\hat{\theta}_n^{\mathrm{PT}}(\kappa)$. We also noticed from Table 17.1 that $\hat{\theta}_n^{\mathrm{LASSO}}(\kappa)$ performs better than $\tilde{\theta}_n$, $\hat{\theta}_n^{\mathrm{PT}}(\kappa)$, and $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ whenever $\Delta$ is in the interval $[0, 1.177]$.

## 17.3 Risk Comparisons of the Different L-Estimators of $\boldsymbol{\theta}$

In this section, we provided the risk comparisons of the proposed L-estimators with the unrestricted L-estimator (ULE), $\tilde{\theta}_n$.

### 17.3.1 Comparison of Restricted L-Estimator (RLE) and ULE

The difference between the asymptotic distributional risk (ADR) functions of the two estimators is given by

$$\text{ADR}(\tilde{\theta}_n) - \text{ADR}(\hat{\theta}_n) = \frac{\sigma^2}{K_1} - \delta^2 = \frac{\sigma^2}{K_1}(1 - \Delta^2). \tag{17.35}$$

The difference given in Eq. (17.35) is nonnegative whenever $0 < \Delta^2 < 1$, and hence, the RLE, $\hat{\theta}_n$, outperforms the ULE, $\tilde{\theta}_n$, in this interval. Whenever $\Delta^2 > 1$, then the performance of ULE is better than RLE.

### 17.3.2 Comparison of Preliminary Test L-Estimator (PTLE) and ULE

In this case, the ADR-difference of the two estimators is given by

$$\begin{aligned}\text{ADR}(\tilde{\theta}_n) - \text{ADR}(\hat{\theta}_n^{\text{PT}}) = \frac{\sigma^2}{K_1}\Big[ \Big\{1 - \Big(1 - H_3(\kappa^2; \Delta^2)\Big)\Big\} \\ + \Delta^2 \Big\{2H_3(\kappa^2; \Delta^2) - H_5(\kappa^2; \Delta^2)\Big\}\Big].\end{aligned} \tag{17.36}$$

The ADR difference expression in Eq. (17.36) is nonnegative whenever

$$0 < \Delta^2 < \frac{H_3(\kappa^2; \Delta^2)}{\big\{2H_3(\kappa^2; \Delta^2) - H_5(\kappa^2; \Delta^2)\big\}}.$$

This implies that PTLE outperforms ULE whenever $\Delta^2$ is in this interval. Since $H_3(\kappa^2, \Delta^2) > H_5(\kappa^2, \Delta^2)$, therefore, the PTLE outperforms ULE whenever $\Delta^2 \in (0, 1)$. On the contrary, ULE outperforms PTLE whenever $\Delta^2$ is outside the interval. But as $\Delta^2 \to \infty$, both L-estimators are equivalent in terms of their ADRs.

### 17.3.3 Comparison of the Saleh-Type L-Estimator and ULE

The ADR difference in this case is given by

$$\begin{aligned}\text{ADR}\left(\tilde{\theta}_n\right) - \text{ADR}\left(\hat{\theta}_n^{\text{S}}(\kappa_{\text{opt}})\right) &= \frac{\sigma^2}{K_1}\Big[1 - \Big\{1 - \frac{2}{\pi}\ \Big(2\exp\{-\Delta^2/2\} - 1\Big)\Big\}\Big] \\ &= \frac{\sigma^2}{K_1}\frac{2}{\pi}\Big(2\exp\{-\Delta^2/2\} - 1\Big).\end{aligned} \tag{17.37}$$

The difference in Eq. (17.37) is a decreasing function of $\Delta^2$ which attain a maximum value of $(\sigma^2 / K_1)(2/\pi)$ at $\Delta^2 = 0$ and then start decreasing to $-(\sigma^2 / K_1)(2/\pi)$ after crossing the one line at $\Delta^2 = 2\ln(2)$. Thus, for $0 < \Delta^2 \leq 2\ln(2)$, $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ uniformly dominates $\tilde{\theta}_n$, while for $\Delta^2 > 2\ln(2)$, the performance of $\tilde{\theta}_n$ is better than $\hat{\theta}_n^{\mathrm{S}}(\kappa)$. Neither of the two estimators outperform the other uniformly. The ADR of $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ decreases from $(1 - 2/\pi)$ at $\Delta = 0$ and crosses the line at $\Delta^2 = 2\ln(2)$. Thus, for $0 < \Delta^2 < 2\ln(2)$, the estimator $\hat{\theta}_n^{\mathrm{S}}$ performs better than $\tilde{\theta}_n$.

### 17.3.4 Risk Comparison of Ridge-Type L-Estimator and ULE

In this case, the ADR-difference of the two estimator is given by

$$\begin{aligned}\mathrm{ADR}(\tilde{\theta}_n) - \mathrm{ADR}\left(\hat{\theta}_n^{\mathrm{R}}(\kappa_{\mathrm{opt}})\right) &= \frac{\sigma^2}{K_1}\left[1 - \frac{\Delta^2}{1+\Delta^2}\right] \\ &= \frac{\sigma^2}{K_1(1+\Delta^2)} \geq 0 \quad \forall\, \Delta^2 \in R^+. \end{aligned} \tag{17.38}$$

Hence, the ridge-type L-estimator performs better than $\tilde{\theta}_n$ uniformly.

### 17.3.5 Comparison of LASSO L-Estimators and ULE

The ADR-difference of the two estimators is given by

$$\begin{aligned}\mathrm{ADR}\left(\hat{\theta}_n^{\mathrm{LASSO}}(\kappa)\right) - \mathrm{ADR}\left(\tilde{\theta}_n\right) = \frac{\sigma^2}{K_1}\Big[ &- H_3(\kappa^2;\Delta^2) \\ &+ \Delta^2\left\{2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\right\} \\ &+ \kappa^2\left\{1 - H_1(\kappa^2;\Delta^2)\right\} \\ &- 2\kappa\Delta\Big\{\left(\Phi(\kappa-\Delta) - \Phi(\kappa+\Delta)\right)\Big\} \\ &+ 4\kappa\phi(\Delta)\Big]. \end{aligned} \tag{17.39}$$

This shows that LASSO performs better than ULE whenever

$$0 \le \Delta^2 \le \frac{H_3(\kappa^2;\Delta^2) - \kappa^2\left\{1 - H_1(\kappa^2;\Delta^2)\right\} + 2\kappa\Delta\left(\Phi(\kappa-\Delta) - \Phi(\kappa+\Delta)\right) + 4\kappa\phi(\Delta)}{2[H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)]}.$$

Otherwise, ULE performs better than LASSO for other variation of $\Delta$. Hence, neither LASSO L-estimator nor ULE outperforms the other uniformly.

### 17.3.6 Comparison of LASSO and Ridge-Type L-Estimators

From Lemma 1 of Donoho and Johnstone (1994), we have $1 < \gamma_{\mathrm{st}}(\kappa;\Delta) \le 1+\Delta^2$. The difference between the ADRs of the two estimators is

$$\begin{aligned}
&\mathrm{ADR}\left(\hat{\theta}_n^{\mathrm{LASSO}}(\kappa)\right) - \mathrm{ADR}\left(\hat{\theta}_n^{\mathrm{R}}(k)\right) = \frac{\sigma^2}{K_1}\left[\gamma_{st}(\kappa;\Delta) - \frac{\Delta^2}{1+\Delta^2}\right] \\
&= \frac{\sigma^2}{K_1}\Big[\frac{1}{1+\Delta^2} - \Big(H_3(\kappa^2;\Delta^2) - \Delta^2\left\{2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\right\} \\
&\quad - \kappa^2\left\{1 - H_1(\kappa^2;\Delta^2)\right\} + 2\kappa\Delta\left\{\Phi(\kappa-\Delta) + \Phi(\kappa+\Delta)\right\} + 4\kappa\phi(\Delta)\Big)\Big],
\end{aligned} \tag{17.40}$$

which is nonnegative for all $\Delta^2 \in R^+$. Since $1/(1+\Delta^2)$ and the function in the parenthesis are both decreasing in $\Delta^2$ on top of the other, hence, the ridge-type L-estimator outperform LASSO L-estimator uniformly.

### 17.3.7 Comparison of LASSO and Saleh-Type L-Estimators

The ADR-difference of these two estimators can be written as

$$\begin{aligned}
&\mathrm{ADR}(\hat{\theta}_n^{\mathrm{S}}(\kappa)) - \mathrm{ADR}(\hat{\theta}_n^{\mathrm{LASSO}}(\kappa)) \\
&= \frac{\sigma^2}{K_1}\Big[\Big\{1 - \frac{2}{\pi}\ (2\exp\{-\Delta/2\} - 1)\Big\} - \gamma_{\mathrm{st}}(\kappa;\Delta)\Big]. \\
&= \frac{\sigma^2}{K_1}\Big[\left\{1 - \gamma_{\mathrm{st}}(\kappa;\Delta)\right\} - \frac{2}{\pi}\ (2\exp\{-\Delta/2\} - 1)\Big] \\
&\le 0.
\end{aligned} \tag{17.41}$$

Since both the functions, in the braces and in the parenthesis, are decreasing in $\Delta^2$ on top of the other, hence, LASSO L-estimator outperforms Saleh-type L-estimator uniformly.

### 17.3.8 Comparison of LASSO L-Estimator and PTLE

The ADR-difference of $\hat{\theta}_n^{\text{LASSO}}(\kappa)$ and $\hat{\theta}_n^{\text{PT}}(\kappa)$ is

$$\begin{aligned}
&\text{ADR}(\hat{\theta}_n^{\text{LASSO}}(\kappa)) - \text{ADR}(\hat{\theta}_n^{\text{PT}}(\kappa)) \\
&= \frac{\sigma^2}{K_1}\Big[\gamma_{st}(\kappa;\Delta) - \Big(1 - H_3(\kappa^2;\Delta^2)\Big) - \Delta^2\Big(2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\Big)\Big] \\
&= \frac{\sigma^2}{K_1}\Big[\kappa^2\big\{1 - H_1(\kappa^2;\Delta^2)\big\} - 2\kappa\Delta\big\{\Phi(\kappa-\Delta) + \Phi(\kappa+\Delta)\big\} + 4\kappa\phi(\Delta)\Big] \\
&\geq 0, \qquad (17.42)
\end{aligned}$$

which is a decreasing function of $\Delta^2$. Hence, PTLE outperforms LASSO L-estimator uniformly.

### 17.3.9 Comparison of PTLE and Ridge-Type L-Estimator

The ADR difference between the two estimators is given by the following expression:

$$\begin{aligned}
&\text{ADR}(\hat{\theta}_n^{\text{PT}}(\kappa)) - \text{ADR}(\hat{\theta}_n^{\text{R}}(k_{\text{opt}})) \\
&= \frac{\sigma^2}{K_1}\left[\Big(1 - H_3(\kappa^2;\Delta^2)\Big) + \Delta^2\big\{2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\big\} - \frac{\Delta^2}{1+\Delta^2}\right] \\
&= \frac{\sigma^2}{K_1}\left[\frac{1}{1+\Delta^2} - \Big\{H_3(\kappa^2;\Delta^2) + \Delta^2\Big(2H_3(\kappa^2;\Delta^2) - H_5(\kappa^2;\Delta^2)\Big)\Big\}\right] \\
&\geq 0 \qquad \forall \Delta^2 \in R^+ \qquad (17.43)
\end{aligned}$$

Since $1/(1+\Delta^2)$ is decreasing from 1 to 0 as $\Delta^2 \to \infty$ and so the expression in the braces is also decreasing from $H_3(\kappa^2;0)$ to $\infty$ below the graph of $1/(1+\Delta^2)$, hence, the expression is nonnegative for all $\Delta^2 \in R^+$. Thus, the ridge L-estimator outperforms PTLE uniformly.

### 17.3.10 Comparisons of Ridge-Type and Saleh-Type L-Estimators

The ADR-difference of $\hat{\theta}_n^{\mathrm{S}}(\kappa)$ and $\hat{\theta}_n^{\mathrm{R}}(k_{\mathrm{opt}})$ is

$$\begin{aligned}
&\mathrm{ADR}(\hat{\theta}_n^{\mathrm{S}}(\kappa)) - \mathrm{ADR}(\hat{\theta}_n^{\mathrm{R}}(k_{\mathrm{opt}})) \\
&= \frac{\sigma^2}{K_1}\Big[\Big\{1 - \frac{2}{\pi}\,\Big(2\exp\{-\Delta^2/2\} - 1\Big)\Big\} - \frac{\Delta^2}{1+\Delta^2}\Big] \\
&= \frac{\sigma^2}{K_1}\Big[\frac{1}{1+\Delta^2} - \frac{2}{\pi}\,\Big(2\exp\{-\Delta^2/2\} - 1\Big)\Big] \\
&= \frac{\sigma^2}{K_1}\Big[\frac{2}{\pi} + \Big\{\frac{1}{1+\Delta^2} - \frac{4\exp\{-\Delta^2/2\}}{\pi}\Big\}\Big]. \qquad (17.44)
\end{aligned}$$

Since both terms in the braces are decreasing functions of $\Delta^2$ for all $\Delta^2 \in R^+$, hence, the difference is nonnegative. Thus, the ridge-type L-estimator outperforms the Saleh-type L-estimator uniformly.

## 17.4 Application Using Logistic Regression

The L-estimators discussed in Sects. 17.2 and 17.3 depend on the quantile density function, $q_o = F_o^{-1}(\kappa)$, which may be estimated by the nonparametric methods of p.d.f. estimation (see, for example, Koul, 1992). However, in this section, we demonstrate the application using logistic distribution defined by

$$f(x) = \frac{\exp\left\{\dfrac{-\pi x}{\sqrt{3}}\right\}}{\sqrt{3}\left(1 + \dfrac{-\pi x}{\sqrt{3}}\right)^2}, \quad -\infty < x < \infty. \qquad (17.45)$$

For a sample of size $n$, the information matrix of the parameters $(\theta, \sigma)^\top$ is given by

$$\frac{n}{\sigma^2}\begin{pmatrix} \dfrac{\pi^2}{3} & 0 \\ 0 & \dfrac{3+\pi^2}{9} \end{pmatrix}. \qquad (17.46)$$

The $\kappa$-quantile of the distribution is

$$Q(\kappa) = \theta + \frac{\sqrt{3}}{\pi}\,\sigma\,\ln\left(\frac{\kappa}{1-\kappa}\right), \quad 0 < \kappa < 1. \qquad (17.47)$$

The quantile density function of $\kappa$ is then given by

$$q_o(\kappa) = \frac{\pi}{\sqrt{3}}\kappa(1-\kappa). \tag{17.48}$$

Using Eq. (17.6), the estimators of $(\theta, \sigma)^\top$ are then given by

$$\begin{pmatrix} \tilde{\theta}_n \\ \tilde{\sigma}_n \end{pmatrix} = \begin{pmatrix} K_1 & K_3 \\ K_3 & K_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_k^\top \Omega^{-1} \boldsymbol{y} \\ \boldsymbol{u}^\top \Omega^{-1} \boldsymbol{y} \end{pmatrix}. \tag{17.49}$$

The explicit forms of $K_1$, $K_2$, and $K_3$ are given by

$$K_1 = \sum_{i=1}^{k+1} \frac{\pi^2}{3}(\lambda_i - \lambda_{i-1})(\lambda_i + \lambda_{i-1} - 1)^2,$$

$$K_2 = \sum_{i=1}^{k+1} \frac{\lambda_i(\lambda_i - 1)\ \ln\left(\frac{\lambda_i}{1-\lambda_i}\right) - \lambda_{i-1}(\lambda_{i-1} - 1)\ \ln\left(\frac{\lambda_{i-1}}{1-\lambda_{i-1}}\right)}{(\lambda_i - \lambda_{i-1})},$$

and

$$K_3 = \sum_{i=1}^{k+1} \frac{\pi}{\sqrt{3}}(\lambda_i + \lambda_{i-1} - 1)\,(\lambda_i(\lambda_i - 1))\ \ln\left(\frac{\lambda_i}{1-\lambda_i}\right)$$
$$- \lambda_{i-1}(\lambda_{i-1} - 1)\ \ln\left(\frac{\lambda_{i-1}}{1-\lambda_{i-1}}\right).$$

Table 17.2 gives the optimum spacing $\lambda_i$ $(i = 1, \ldots, k)$ for $k = 2, \ldots, 9$, the coefficients $a_i$ for $\tilde{\theta}_n$, $b_i$ for $\tilde{\sigma}$, the variances of the estimates, and their asymptotic relative efficiencies. The $\text{var}(\tilde{\theta}_n)$ is in terms of $\sigma^2 / (a^2 n)$, where $a = \pi/\sqrt{3}$, and $\text{var}(\tilde{\sigma})$ is in terms of $\sigma^2 / n$. The asymptotic relative efficiency is defined as the ratio between the Crameér-Rao lower bound for the variance of the estimator and of the particular estimate. The terms $3\sigma^2 / (a^2 n)$ and $\{9 / (3 + \pi^2)\}(\sigma^2/n)$ are the bounds for the estimators of $\theta$ and $\sigma^2$, respectively. It should be noted that Table 17.2 is from the paper of Hassanien (1969).

## 17.5 Conclusion

In this chapter, we have defined six different L-estimators for estimating the location parameter $\theta$ of a continuous distribution. The performances of the estimators have been analyzed via $L_2$-risk criterion. The performance of the restricted L-estimator $\hat{\theta}_n$ is better than the unrestricted L-estimator $\tilde{\theta}_n$ whenever $\Delta^2 \in (0, 1)$, but $\tilde{\theta}_n$ is

**Table 17.2** Asymptotically optimum spacing $\lambda_i$, the best linear unbiased estimators $\tilde{\theta}_n = a_1 y_{[n\lambda_1]+1} + \cdots + a_k y_{[n\lambda_k]+1}$ and $\tilde{\sigma}_n = b_1 y_{[n\lambda_1]+1} + \cdots + b_k y_{[n\lambda_k]+1}$, the variances, and the asymptotic relative efficiencies of the estimators for the logistic distribution. (This table is from Hassanien (1969)'s paper on logistic regression)

| $k$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| $\lambda_1$ | 0.238 | 0.153 | 0.104 | 0.073 | 0.053 | 0.039 | 0.030 | 0.023 |
| $a_1$ | 0.500 | 0.2138 | 0.1043 | 0.0549 | 0.0308 | 0.0175 | 0.0107 | 0.0066 |
| $b_1$ | -0.7794 | -0.5300 | -0.3444 | -0.2329 | -0.1638 | -0.1175 | -0.0882 | -0.0666 |
| $\lambda_2$ | 0.762 | 0.500 | 0.351 | 0.259 | 0.197 | 0.152 | 0.120 | 0.096 |
| $a_2$ | 0.5000 | 0.5724 | 0.3957 | 0.2565 | 0.1670 | 0.1098 | 0.0733 | 0.0497 |
| $b_2$ | 0.7794 | 0.0000 | -0.2686 | -0.2998 | -0.2697 | -0.2282 | -0.1883 | -0.1549 |
| $\lambda_3$ | - | 0.847 | 0.649 | 0.500 | 0.394 | 0.316 | 0.257 | 0.211 |
| $a_3$ | - | 0.2138 | 0.3957 | 0.3772 | 0.3024 | 0.2313 | 0.1735 | 0.1294 |
| $b_3$ | - | 0.5300 | 0.2686 | 0.000 | -0.1291 | -0.1783 | -0.1881 | -0.1803 |
| $\lambda_4$ | - | - | 0.896 | 0.741 | 0.606 | 0.500 | 0.417 | 0.351 |
| $a_4$ | - | - | 0.1043 | 0.2565 | 0.3024 | 0.2828 | 0.2425 | 0.2007 |
| $b_4$ | - | - | 0.3444 | 0.2998 | 0.1291 | 0.0000 | -0.7636 | -0.1166 |
| $\lambda_5$ | - | - | - | 0.927 | 0.803 | 0.684 | 0.583 | 0.500 |
| $a_5$ | - | - | - | 0.0549 | 0.1670 | 0.2313 | 0.2425 | 0.2272 |
| $b_5$ | - | - | - | 0.2329 | 0.2697 | 0.1783 | 0.7636 | 0.0000 |
| $\lambda_6$ | - | - | - | - | 0.947 | 0.848 | 0.743 | 0.649 |
| $a_6$ | - | - | - | - | 0.0306 | 0.1098 | 0.1735 | 0.2007 |
| $b_6$ | - | - | - | - | 0.1638 | 0.2282 | 0.1881 | 0.1166 |
| $\lambda_7$ | - | - | - | - | - | 0.961 | 0.880 | 0.789 |
| $a_7$ | - | - | - | - | - | 0.0175 | 0.0733 | 0.1294 |
| $b_7$ | - | - | - | - | - | 0.1175 | 0.1883 | 0.1803 |
| $\lambda_8$ | - | - | - | - | - | - | 0.970 | 0.904 |
| $a_8$ | - | - | - | - | - | - | 0.0107 | 0.0497 |
| $b_8$ | - | - | - | - | - | - | 0.0882 | 0.1549 |
| $\lambda_9$ | - | - | - | - | - | - | - | 0.977 |
| $a_9$ | - | - | - | - | - | - | - | 0.0068 |
| $b_9$ | - | - | - | - | - | - | - | 0.0666 |
| var($\tilde{\theta}_n$) | 3.6181 | 3.2993 | 3.1876 | 3.1303 | 3.0962 | 3.0747 | 3.0594 | 3.0486 |
| asymp. rel. eff.($\tilde{\theta}_n$) | 0.8292 | 0.9093 | 0.9411 | 0.9584 | 0.9689 | 0.9757 | 0.9806 | 0.9841 |
| var($\tilde{\sigma}_n$) | 1.4001 | 1.0975 | 0.9369 | 0.8638 | 0.8210 | 0.7931 | 0.7740 | 0.7603 |
| asymp.rel. eff.($\tilde{\sigma}_n$) | 0.4995 | 0.6478 | 0.7464 | 0.8096 | 0.8518 | 0.8818 | 0.9035 | 0.9198 |

a better choice for $\Delta^2 > 1$. The preliminary test L-estimator $\hat{\theta}_n^{\mathrm{PT}}$ performs better than the $\tilde{\theta}_n$ over a range of $\Delta^2$ values, but out of that range, the performance of $\tilde{\theta}_n$ is better. As $\Delta^2 \to \infty$, the performance of both the estimators is the same. The Saleh-type L-estimator $\hat{\theta}_n^{\mathrm{S}}$ outperforms $\tilde{\theta}_n$ whenever $0 < \Delta^2 < 2\ln(2)$, but for $\Delta^2 > 2\ln(2)$, the performance of $\tilde{\theta}_n$ is better than $\hat{\theta}_n^{\mathrm{S}}$. The ridge-type L-estimator

$\hat{\theta}_n^{\mathrm{R}}$ uniformly dominates $\tilde{\theta}_n$, $\hat{\theta}_n^{\mathrm{PT}}$, $\hat{\theta}_n^{\mathrm{S}}$, and the LASSO L-estimator $\hat{\theta}_n^{\mathrm{LASSO}}$. Neither $\hat{\theta}_n^{\mathrm{LASSO}}$ nor $\tilde{\theta}_n$ performs better than the other uniformly. The $\hat{\theta}_n^{\mathrm{PT}}$ outperforms $\hat{\theta}_n^{\mathrm{LASSO}}$ uniformly for all $\Delta^2 \in R^+$.

**Acknowledgments** The authors are thankful to the unknown referees for their valuable comments on the earlier draft. The authors are also thankful to Dr. Mina Noruzirad for her help in the current format of the manuscript.

## References

Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *The Annals of Statistics, 1*, 597–616.

David, H. A. (1981). *Order statistics* (2nd ed.). Wiley.

David, H. A., & Ragaraja, H. N. (2003). *Order statistics*(3rd ed.) Wiley.

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika, 81*, 425–455.

Hassanien, K. M. (1969). Estimation of the parameters of the logistic distribution by sample quantiles. *Biometrika, 56*(3), 684–687.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*, 33–50.

Koenker, R., & Portony, S. (1987). L-estimation for linear models. *Journal of the American Statistical Association, 82*(399), 851–857.

Koul, H. A. (1992). *Weighted empiricals and linear models*. IMS. Lecture notes—monograph series 21.

Saleh, A. K. Md. E. (1992). *Contributions to order statistics*. School of Mathematics and Statistics, Carleton University.

Saleh, A. K. Md. E. (2006). *Theory of preliminary test and stein-type estimation with applications*. Wiley.

Serfling, R. J. (1980). *Approximation theorem of mathematical statistics*. Wiley.

Welsh, A. H. (1987). The trimmed mean in the linear model. *The Annals of Statistics, 15*, 20–45.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press.

# Chapter 18
# Variable Selection in Regression Models with Dependent and Asymmetrically Distributed Error Term




**Yetkin Tuaç and Olcay Arslan**

## 18.1 Introduction

An outstanding number of studies have been conducted on regression analysis under the assumption that observations are independent of each other. However, if the data are correlated in time and ignoring the correlation of the error terms when it exists will result in the estimate of the variance of the model parameters are incorrect. Regarding this issue, many studies based on symmetric distributions have been conducted for variable selection of regression models with auto-regressive error terms. However, in case of skewness in the data encountered in real life, it is seen that variable selection and parameter estimation studies regarding regression models with auto-regressive error terms are quite limited.

Some studies in the literature regarding parameter estimation of the dependent error term regression model that use the normal distribution as the error distribution assumption can be listed below. Alpuim and El-Shaarawi (2008) estimates the parameters of the regression model using the $p$-th order auto-regressive (AR($p$)) error term using the ordinary least squares (OLS) estimation method. They also use the maximum likelihood (ML) and conditional maximum likelihood (CML) estimation methods under the assumption of normality and examine the asymptotic properties of the resulting estimators. Beach and MacKinnon (1978) used the ML method to estimate the parameters of the AR(1) error term regression models. However, the assumption of normality is quite restrictive and sensitive to the presence of outliers in the data. This situation attracted the attention of researchers, and in order to relax the strict conditions of the normal distribution assumption, they developed parameter estimation methods by assuming that the error terms are

Y. Tuaç (✉) · O. Arslan
Faculty of Science, Department of Statistics, Ankara University, Ankara, Türkiye
e-mail: ytuac@ankara.edu.tr; oarslan@ankara.edu.tr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_18

distributed with alternative distributions to the normal distribution. Tiku et al. (1999) made parameter estimates using the modified maximum likelihood (MML) method in regression models with AR(1) error terms, under the assumption of gamma and generalized logistic distributions without concern for variable selection. However, the skew distribution used in this study does not have sufficient flexibility to model moderate skewness. Recently, Tuaç et al. (2018) suggested using the symmetric $t$ distribution as the distribution of error terms and used the CML method to maintain the estimation of model parameters in the auto-regressive error term regression model as a robust alternative to the normal distribution. Furthermore, Tuaç et al. (2020) used skew alternatives to normal and t distributions as error distributions in the AR($p$) error term regression model. However, variable selection was not handled in any of the studies discussed.

Regarding regression models with auto-regressive error terms, few papers have studied penalty-based variable selection methods to identify significant variables. For example, Wang et al. (2007) evaluated linear regression with auto-regressive errors using a modified LASSO-type penalty on both regression and auto-regressive coefficients, and (Yoon et al., 2013) discussed adaptive LASSO, SCAD, and bridge methods in AR($p$) error term regression models under normally distributed errors. These studies are conducted under the least squares method or the assumption of normal distribution, which is known to be sensitive to skewness and/or outliers in the data. Concerning the variable selection part of this study, we employed penalized regression methods, such as LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), ridge (Hoerl & Kennard, 1970), bridge (Frank & Friedman, 1993), and elastic net (Zou & Hastie, 2005). These methods select variables and estimate model parameters simultaneously by providing a smaller prediction error and stability compared to the subset selection methods.

For datasets with potential skewness, selecting variables and estimating parameters using methods developed for the normal distribution can lead to ineffective estimates. Using skew distributions as an alternative to symmetric ones will make the predictions obtained more effective for those datasets. Therefore, in this study, Azzalini-type skew distributions (Azzalini, 1985, 1986; Azzalini & Capitanio, 2003) were used as error distribution to give flexibility to the assumption of symmetric distribution to suit real-life data. This flexibility is provided with a skewness parameter in the distributions. The distributions become their symmetric alternatives when the skewness parameter is equal to zero.

Since the expectation maximization (EM) algorithm (Dempster et al., 1977) can be easily applied to Azzalini-type skew distributions compared to other skew distributions, it becomes easier to obtain the estimates for the parameters of interest. In this study, variable selections and parameter estimations are obtained by assuming that errors in the regression model with auto-regressive error terms are distributed according to Azzalini-type skew normal and skew t distributions. We propose an expectation conditional maximization (ECM) (Meng & Rubin, 1993) algorithm to calculate the ML estimates of the parameters and variable selection of the penalized regression model with AR($p$) error terms for skew normal and skew $t$

distribution assumptions. A simulation study and a real data example are provided to compare the performance of the parameter estimates in different scenarios.

The remainder of this chapter is organized as follows: In Sect. 18.2, we introduce the $p$-th order auto-regressive error term regression model and discuss penalizing the model with brief information about the penalized methods used. In Sect. 18.3, we propose an ECM-type algorithm to calculate the ML estimate of the parameters and the selection of variables of the penalized regression model with AR($p$) error terms for each of the five methods under the assumption of Azzalini-type skew normal and skew $t$. Section 18.4 presents the numerical proof of the proposed methods with a simulation study and an example of real data. Section 18.5 concludes the chapter.

## 18.2 Regression Model

Consider the linear regression model as

$$y_t = \sum_{i=1}^{M} x_{ti}\beta_i + e_t \quad t = 1, \ldots, N, \tag{18.1}$$

where $y_t$ is the response variable, $x_{ti}$ are independent variables, $\beta_i$ are regression parameters, and $e_t$ is the error term. We will assume that $e_t$ has a stationary auto-regressive error process of order $p$, i.e., AR($p$), which $a_t = e_t - \phi_1 e_{t-1} - \cdots - \phi_p e_{t-p}$. Here, $\phi_j$, for $j = 1, \ldots, p$, are the auto-regressive parameters, and $a_t$ is the new i.i.d. error term with $\mathrm{E}[a_t] = 0$ and $\mathrm{var}(a_t) = \sigma^2$.

For simplicity, the following backshift operator $B$ can be used:

$$a_t = \Phi(B)e_t. \tag{18.2}$$

With the help of backshift operator, the regression model given by Eq. (18.1) with auto-regressive error terms can be rewritten as

$$\Phi(B)\, y_t = \sum_{i=1}^{M} \beta_i \Phi(B)\, x_{ti} + a_t \quad t = 1, \ldots, T, \tag{18.3}$$

where

$$\Phi(B)y_t = y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} \tag{18.4}$$

and

$$\Phi(B)x_{ti} = x_{ti} - \phi_1 x_{t-1,i} - \cdots - \phi_p x_{t-p,i}\ . \tag{18.5}$$

Here, first $p$ observations are known, so $T = N - p$.

Let us turn our attention to the simultaneous parameter estimation and variable selection in this model. One way to perform parameter estimation and variable selection simultaneously is to minimize the following penalized objective function. This approach has already been used by Yoon et al. (2013) under the assumption of normality of errors $a_t$ ($a_t \sim N(0, \sigma^2)$). In that paper, the following penalized objective function was minimized,

$$Q(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2) = -\ln L(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2 | a_t) + \sum_{i=1}^{M} p_\lambda(\beta_i) + \sum_{j=1}^{p} p_\delta(\phi_j)\,, \tag{18.6}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^\top$ are regression and $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_p\right)^\top$ are auto-regressive parameters with $p$ auto-regressive degree and $M$ regression coefficients. $N$ is the sample size, $p_\lambda(\beta_i)$ is the penalty function for $\boldsymbol{\beta}$, $p_\delta(\phi_j)$ is the penalty function for $\boldsymbol{\phi}$, and $\ln L(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2 | a_t)$ is the following log-likelihood function:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2 | a_t) = c - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=p+1}^{N} \left[ \Phi(B) y_t - \sum_{i=1}^{M} \beta_i \Phi(B) x_{ti} \right]^2 . \tag{18.7}$$

Note that, for ease of computation, penalizing the auto-regressive parameters $p_\delta(\phi_j)$ is ignored. In this study, we calculate the estimations of the regression and auto-regression parameters but penalize only the regression parameters. We will use the following commonly used penalty functions:

1. $p_\lambda^{\text{LASSO}}(\beta) = \lambda \sum_{i=1}^{M} |\beta_i|, \quad \lambda > 0$
2. $p_\lambda^{\text{SCAD}}(\beta_i) = \begin{cases} \lambda |\beta_i| & \text{if} \quad |\beta_i| \le \lambda \\ \frac{-|\beta_i|^2 - 2\alpha\lambda|\beta_i| + \lambda^2}{2(\alpha - 1)} & \text{if} \quad \lambda < |\beta_i| \le \alpha\lambda, \quad i = 1, \ldots, M \\ \frac{\alpha+1}{2}\lambda^2 & \text{if} \quad |\beta_i| > \alpha\lambda \end{cases}$
3. $p_\lambda^{\text{Bridge}}(\beta) = \lambda \sum_{i=1}^{M} |\beta_i|^\gamma, \quad 0 < \gamma \le 1$
4. $p_\lambda^{\text{Ridge}}(\beta) = \lambda \sum_{i=1}^{M} \beta_i^2, \quad \lambda > 0$
5. $p_\lambda^{\text{E-Net}}(\beta) = \lambda_1 \sum_{i=1}^{M} |\beta_i| + \lambda_2 \sum_{i=1}^{M} \beta_i^2$

Since penalty functions are not differentiable at the origin, it is impossible to obtain an analytical solution to the maximization problem. Therefore, numerical methods must be used to tackle this minimization problem. An approach to address

this problem is to use the local quadratic approximation (LQA) method, which is a special case of the Taylor expansion, typically involving the first two terms, the constant term and the quadratic term, as proposed by Fan and Li (2001), to approximate the penalty function. In LQA, penalty functions are locally approximated at a given $\beta^{(0)} = \left(\beta_1^0, \ldots, \beta_M^0\right)$ using the following quadratic function,

$$\begin{aligned} p_{\lambda_i}(|\beta_i|) &\approx p_\lambda\left(\left|\beta_i^{(0)}\right|\right) + \frac{1}{2}\frac{p_{\lambda_i}^{'}\left(\left|\beta_i^{(0)}\right|\right)}{\left(\left|\beta_i^{(0)}\right|\right)}(\beta_i^2 - \beta_i^{(0)2}) \\ &= \frac{1}{2}\frac{p_{\lambda_i}^{'}\left(\left|\beta_i^{(0)}\right|\right)}{\left(\left|\beta_i^{(0)}\right|\right)}\beta_i^2 + p_{\lambda_i}\left(\left|\beta_i^{(0)}\right|\right) - \frac{1}{2}\frac{p_{\lambda_i}^{'}\left(\left|\beta_i^{(0)}\right|\right)}{\left(\left|\beta_i^{(0)}\right|\right)}\beta_i^{(0)2}, \end{aligned} \tag{18.8}$$

where $p_\lambda^{'}$ is the derivative of the penalty function. Next, utilizing Eq. (18.6) and disregarding the penalization of the auto-regressive parameters, the objective function for skew distributions can be constructed.

## 18.3 Penalized Maximum Likelihood Estimation with Skew Distributions

In this section, after giving brief information about Azzalini-type skew normal and skew $t$ distributions, model parameter estimates and variable selections under these distribution assumptions will be discussed.

### *18.3.1 Skew Normal Distribution*

The Azzalini-type skew normal distribution with the parameters $\mu$, $\sigma$, and $\lambda$ was defined by Azzalini (1985) with the probability density function (p.d.f.) as

$$h(y; \mu, \sigma, \lambda) = \frac{2}{\sigma} f\left(\frac{y-\mu}{\sigma}\right) F\left(\zeta\frac{y-\mu}{\sigma}\right), \tag{18.9}$$

where $f$ and $F$ are the p.d.f. and the cumulative distribution function (c.d.f.) of the univariate standard normal distribution, respectively. Here, $\mu \in \mathbb{R}$ is the location, $\sigma \in \mathbb{R}^+$ is the scale, and $\zeta \in \mathbb{R}$ is the skewness parameter.

If the random variable $Y$ has a skew normal distribution, $Y \sim SN(\mu, \sigma, \zeta)$, it has the stochastic representation as

$$Y = \mu + \sigma\delta|U_1| + \sigma\sqrt{(1-\delta^2)}U_2, \tag{18.10}$$

where $U_1$ and $U_2$ are independent random variables with standard normal distribution and $\delta = \zeta / \sqrt{1+\zeta^2}$ (Azzalini, 1986; Henze, 1986). This stochastic representation will be very useful for the implementation of the EM algorithm.

We also have the following hierarchical representation with the help of conditional distributions:

$$\begin{aligned} Y|\gamma &\sim N(\mu + \sigma\delta\gamma, (1-\delta^2)\sigma^2)\,, \\ \gamma &\sim TN(0, 1; (0, \infty)). \end{aligned} \tag{18.11}$$

Here, $\gamma = |U_1|$ and TN$(\cdot)$ corresponds to the truncated normal distribution. Further, using conditional distributions, the joint p.d.f. of random variables $Y$ and $\gamma$ is obtained as

$$f(y, \gamma) = \frac{1}{\pi\sqrt{\kappa^2}} \exp\left(-\frac{(y-\mu-\alpha\gamma)^2}{2\kappa^2} - \frac{\gamma^2}{2}\right). \tag{18.12}$$

Here, $\kappa^2 = \sigma^2(1-\delta^2)$ and $\alpha = \sigma\delta$. By proportioning the joint p.d.f. to the marginal p.d.f. of $Y$, the conditional distribution of $\gamma|y$ is obtained as

$$f(\gamma|y) = \frac{f(y,\gamma)}{f(y)} = \frac{1}{\sqrt{2\pi(1-\delta^2)}F(\zeta\frac{y-\mu}{\sigma})} \exp\left(-\frac{\left(\gamma - \frac{\delta(y-\mu)}{\sigma}\right)^2}{2(1-\delta^2)}\right). \tag{18.13}$$

This conditional p.d.f. can be expressed as $\gamma|y \sim TN((\delta\eta, 1-\delta^2); (0, \infty))$. Here, $\eta = (y-\mu)/\sigma$.

The following proposition will allow us to compute the conditional expectations of $\gamma$ and $\gamma^2$ given $Y$, which will be encountered in computing the estimates using the ECM algorithm.

**Proposition 18.1** *If $X \sim TN((\mu, \sigma^2); (\alpha_1, \alpha_2))$ is a random variable with distribution, the p.d.f. is*

$$f(x; \mu, \sigma^2) = \{F(\alpha_2) - F(\alpha_1)\}^{-1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad \alpha_1 < x < \alpha_2\,. \tag{18.14}$$

*Here, $\alpha_i = (\alpha_i - \mu)/\sigma$, $i = 1, 2$. Using Eq. (18.14), the expected values for random variables $X$ and $X^2$ are obtained as*

$$\mathrm{E}[X] = \mu - \sigma\frac{f(\alpha_2) - f(\alpha_1)}{F(\alpha_2) - F(\alpha_1)} \tag{18.15}$$

*and*

$$\mathrm{E}[X^2] = \mu^2 + \sigma^2 - \sigma^2 \frac{\alpha_2 f(\alpha_2) - \alpha_1 f(\alpha_1)}{F(\alpha_2) - F(\alpha_1)} - 2\mu\sigma \frac{f(\alpha_2) - f(\alpha_1)}{F(\alpha_2) - F(\alpha_1)} . \quad (18.16)$$

Then, using Proposition 18.1, we can obtain the following conditional expectations,

$$\mathrm{E}[\gamma|y] = \delta\eta + \sqrt{1-\delta^2}\frac{f(\zeta\eta)}{F(\zeta\eta)} , \quad (18.17)$$

$$\mathrm{E}[\gamma^2|y] = 1 - \delta^2 + \delta^2\eta\,\mathrm{E}[\gamma|y] , \quad (18.18)$$

where $\eta = (y-\mu)/\sigma$.

Note that these results will be used to obtain the estimators for the parameters of model defined in Eq. (18.3). Unlike the symmetric distribution, we will have $\mathrm{E}[a_t] \neq 0$ since $\mathrm{E}[a_t] = \sqrt{2/\pi}\delta\sigma$. Therefore, under the assumption of skew normal distribution, the estimator for the intercept parameter should be modified with respect to this expectation.

#### 18.3.1.1 CML Estimation for Skew Normal Distribution

Let us consider the regression model given by Eq. (18.1) with AR($p$) error terms and assume that the errors are $a_t \sim SN(0, \sigma^2, \zeta)$ (Azzalini-type skew normally distributed). Then, given the observations $(x_{p+1}, y_{p+1})$, $(x_{p+2}, y_{p+2})$, ..., $(x_N, y_N)$, the CML estimators of the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \zeta)$ can be obtained by maximizing the following log-likelihood function:

$$\begin{aligned}\ln L\left(\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \zeta|a_t\right) &= T\,(\ln(2) - \ln(\sigma)) - \frac{T}{2}\ln(2\pi) \\ &\quad - \sum_{t=p+1}^{N} \frac{1}{2}\frac{\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)^2}{\sigma^2} \\ &\quad + \sum_{t=p+1}^{N} \ln\left(F\left(\frac{\zeta\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)}{\sigma}\right)\right) . \end{aligned} \quad (18.19)$$

However, since CML estimators cannot be obtained analytically, numerical methods must be used to obtain estimates. Due to the stochastic representation of the skew normal distribution, the EM-type algorithm can be easily applied to obtain estimates. To implement the EM-type algorithm, define $\boldsymbol{\gamma} = (\gamma_{p+1}, \ldots, \gamma_N)$ as missing data, $\mathbf{y} = (y_{p+1}, \ldots, y_N)$ as observed data, and $(y_t, \gamma_t)$ for $t = p+1, \ldots, N$ as complete data. Then, using the stochastic and hierarchical representations of the skew normal distribution given in Sect. 18.3.1, we can write the following complete data log-likelihood function:

$$l_c(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\gamma}) = -T\left(\frac{\ln(\kappa^2)}{2} + \ln(\pi)\right) - \frac{1}{\pi\sqrt{\kappa^2}} \sum_{t=p+1}^{N} \frac{\left(\Phi(B)y_t - \Phi(B)x_t^\top \boldsymbol{\beta} - \alpha\gamma_t\right)^2}{2\kappa^2} - \sum_{t=p+1}^{N} \frac{\gamma_t^2}{2}. \tag{18.20}$$

Since this function is very similar to the log-likelihood function obtained from normal distribution, it is relatively easy to maximize. However, as the random variable $\boldsymbol{\gamma}$ is a latent variable and the resulting estimators will depend on $\boldsymbol{\gamma}$, these estimators will be unfeasible to use. To solve the latency problem, we should find the conditional expectation of the complete data log-likelihood function given $\boldsymbol{y}$ and the current estimates for the parameters of interest. The conditional expectation of the complete data log-likelihood function given observed data will be as

$$\begin{aligned} \mathrm{E}[l_c(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\gamma}) \,|\Phi(B)y_t] = &-\frac{T}{2}\ln\kappa^2 - \sum_{t=p+1}^{N} \frac{\left(\Phi(B)y_t - \Phi(B)x_t^\top \boldsymbol{\beta}\right)^2}{2\kappa^2} \\ &+ \sum_{t=p+1}^{N} \alpha \frac{\left(\Phi(B)y_t - \Phi(B)x_t^\top \boldsymbol{\beta}\right)^2}{\kappa^2} \mathrm{E}[\gamma_t|\Phi(B)y_t] \\ &- \sum_{t=p+1}^{N} \frac{\alpha^2}{2\kappa^2} \mathrm{E}\left[\gamma_t^2|\Phi(B)y_t\right]. \end{aligned} \tag{18.21}$$

Using the conditional expectations

$$S_{1t} = \mathrm{E}\left[\gamma_t|\Phi(B)y_t\right] = \delta\eta_t + \sqrt{1-\delta^2}\frac{f(\zeta\eta_t)}{F(\zeta\eta_t)},$$

$$S_{2t} = \mathrm{E}\left[\gamma_t^2|\Phi(B)y_t\right] = 1 - \delta^2 + \delta^2\eta_t S_{1t},$$

$$\eta_t = \frac{\left(\Phi(B)y_t - \Phi(B)x_t^T\boldsymbol{\beta}\right)^2}{\sigma^2},$$

we can write the following objective function that should be maximized to obtain the estimates of the unknown parameters:

$$Q(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}) = -\frac{T}{2}\ln\kappa^2 - \sum_{t=p+1}^{N} \frac{\left(\Phi(B)y_t - \Phi(B)x_t^\top \boldsymbol{\beta}\right)^2}{2\kappa^2}$$

$$+\sum_{t=p+1}^{N}\alpha\frac{\left(\Phi(B)y_t-\Phi(B)x_t^{\top}\boldsymbol{\beta}\right)^2}{\kappa^2}S_{1t}$$
$$-\sum_{t=p+1}^{N}\frac{\alpha^2}{2\kappa^2}S_{2t} \tag{18.22}$$

#### 18.3.1.2 Penalized Maximum Likelihood Estimation for Skew Normal Distribution

Since our main goal is to carry out parameter estimation and variable selection simultaneously, we can now combine Eq. (18.22) with penalty function to form our penalized log-likelihood function as

$$Q_{\mathrm{PL}}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})=-\frac{T}{2}\ln\kappa^2-\sum_{t=p+1}^{N}\frac{\left(\Phi(B)y_t-\Phi(B)x_t^{\top}\boldsymbol{\beta}\right)^2}{2\kappa^2}$$
$$+\sum_{t=p+1}^{N}\alpha\frac{\left(\Phi(B)y_t-\Phi(B)x_t^{\top}\boldsymbol{\beta}\right)^2}{\kappa^2}S_{1t}$$
$$-\frac{\alpha^2}{2\kappa^2}\sum_{t=p+1}^{N}S_{2t}+T\sum_{i=1}^{M}p_{\lambda}\left(|\beta_i|\right)+T\sum_{j=1}^{p}p_{\delta}\left(\left|\phi_j\right|\right). \tag{18.23}$$

By maximizing this function with respect to the unknown parameters, we can carry on the parameter estimation and variable selection. In this study, the $\boldsymbol{\phi}$ parameter, which determines the auto-regressive degree of the error terms, will not be included in the model selection process to avoid complexity. Taking the derivatives of $Q_{\mathrm{PL}}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})$ with respect to the parameters $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\phi},\sigma^2,\zeta)$ and setting the resulting equations to zero, we get the following updating equations to get estimates,

$$\hat{\boldsymbol{\beta}}_{\mathrm{Sn}}=\left(\sum_{t=p+1}^{N}\hat{\Phi}(B)x_t\hat{\Phi}(B)x_t^{\top}+\hat{\kappa}^2\Omega\left(p_{i_{\lambda_i}}(\beta)\right)\right)^{-1}$$
$$\times\left(\sum_{t=p+1}^{N}\hat{\Phi}(B)y_t-\hat{\alpha}\hat{S}_{1t}\hat{\Phi}(B)x_t\right), \tag{18.24}$$

where

$$\Omega\left(p_{i_{\lambda_i}}(\beta)\right)=\mathrm{diag}\left(\frac{p_{\lambda}^{'}\beta_1^{(0)}}{\beta_1^{(0)}},\ldots,\frac{p_{\lambda}^{'}\beta_M^{(0)}}{\beta_M^{(0)}}\right),$$

with $\beta_i^{(0)} \neq 0$ as an initial point for $\beta_i$ and $p_\lambda^{'}$ as the derivative of the penalty function. The estimators for the auto-regressive parameters, the error scale, and skewness parameters can be obtained as follows. Note that since the following equations will not be used in variable selection case, they will remain the same in each penalty method,

$$\hat{\boldsymbol{\phi}} = \mathbf{R}^{-1}(\hat{\boldsymbol{\beta}})\Big(R_0(\hat{\boldsymbol{\beta}}) + \hat{\alpha} \sum_{t=p+1}^{N} \hat{S}_{1t}\Big) \tag{18.25}$$

where

$$R_0(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \sum_{t=p+1}^{N} e_t\, e_{t-1} \\ \sum_{t=p+1}^{N} e_t\, e_{t-2} \\ \vdots \\ \sum_{t=p+1}^{N} e_t\, e_{t-p} \end{bmatrix} \tag{18.26}$$

and

$$\mathbf{R}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \sum_{t=p+1}^{N} e_{t-1}^2 & \sum_{t=p+1}^{N} e_{t-1}\, e_{t-2} & \cdots & \sum_{t=p+1}^{N} e_{t-1}\, e_{t-p} \\ \sum_{t=p+1}^{N} e_{t-2}\, e_{t-1} & \sum_{t=p+1}^{N} e_{t-2}^2 & \cdots & \sum_{t=p+1}^{N} e_{t-2}\, e_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=p+1}^{N} e_{t-p}\, e_{t-1} & \sum_{t=p+1}^{N} e_{t-p}\, e_{t-2} & \cdots & \sum_{t=p+1}^{N} e_{t-p}^2 \end{bmatrix} \tag{18.27}$$

as well as

$$\hat{\alpha}^2 = \frac{\sum_{t=p+1}^{N} \hat{S}_{1t} \left(\hat{\Phi}(B) y_t - \hat{\Phi}(B) x_t^\top \hat{\boldsymbol{\beta}}\right)^2}{\sum_{t=p+1}^{N} \hat{S}_{2t}}, \tag{18.28}$$

$$\hat{\sigma}^2 = \hat{\kappa}^2 + \hat{\alpha}^2, \tag{18.29}$$

$$\hat{\zeta} = \hat{\delta}\left(1-\hat{\delta}^2\right)^{-1/2}, \tag{18.30}$$

where

$$\hat{\kappa}^2 = \frac{1}{T}\left(\sum_{t=p+1}^{N}\left(\hat{\Phi}(B)y_t - \hat{\Phi}(B)x_t^\top\hat{\boldsymbol{\beta}}\right)^2 - 2\hat{\alpha}\hat{S}_{1t}\left(\hat{\Phi}(B)y_t - \hat{\Phi}(B)x_t^\top\hat{\boldsymbol{\beta}}\right) + \hat{\alpha}^2\hat{S}_{2t}\right) \tag{18.31}$$

and $\hat{\delta} = \hat{\alpha}^2 / \hat{\sigma}^2$.

Now, we can give the steps of the ECM algorithm to compute the estimates of the parameters under the assumption that the error terms are skew-normally distributed.

### 18.3.2 Steps of the ECM Algorithm

1. Determine the initial value $\boldsymbol{\theta}^{(1)} = \left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\phi}^{(1)}, \alpha^{(1)}, \kappa^{(1)}\right)$ of the parameter vector $\boldsymbol{\theta}$, and stop criterion $\epsilon$.
2. ***E step.***
   Compute the following conditional expectations using the current estimates $\boldsymbol{\theta}^{(k)} = \left(\boldsymbol{\beta}^{(k)}, \boldsymbol{\phi}^{(k)}, \alpha^{(k)}, \kappa^{(k)}\right)$ for $k = 1, 2, \ldots$:

$$S_{1t}^{(k)} = \mathrm{E}\left[\gamma_t|\Phi(B)y_t, \boldsymbol{\theta}^{(k)}\right] = \delta^{(k)}\eta_t^{(k)} + \sqrt{1-\delta^{2(k)}}\frac{f(\zeta^{(k)}\eta_t^{(k)})}{F(\zeta^{(k)}\eta_t^{(k)})}, \tag{18.32}$$

$$S_{2t}^{(k)} = \mathrm{E}\left[\gamma_t^2|\Phi(B)y_t, \boldsymbol{\theta}^{(k)}\right] = 1 - \delta^{2(k)}\delta^{2(k)}\eta_t^{(k)}S_{1t}^{(k)}. \tag{18.33}$$

   Here, $\eta_t^{(k)} = \left(\Phi(B)y_t - \Phi(B)\,x_t^\top\boldsymbol{\beta}^{(k)}\right)/\sigma^{(k)}$.
   Given the current estimate of $\boldsymbol{\theta}^{(k)} = \left(\boldsymbol{\beta}^{(k)}, \boldsymbol{\phi}^{(k)}, \alpha^{(k)}, \kappa^{(k)}\right)$ at the $k$-th iteration, the penalized expected complete data log-likelihood function or the $Q$ function is

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\right) = -\frac{T}{2}\ln\kappa^{2(k)} - \sum_{t=p+1}^{N}\frac{\left(\Phi^{(k)}(B)y_t - \Phi^{(k)}(B)x_t^\top\boldsymbol{\beta}^{(k)}\right)^2}{2\kappa^{2(k)}} + \sum_{t=p+1}^{N}\alpha^{(k)}S_{1t}^{(k)}\frac{\left(\Phi^{(k)}(B)y_t - \Phi^{(k)}(B)x_t^\top\boldsymbol{\beta}^{(k)}\right)^2}{\kappa^{2(k)}}$$

$$-\frac{\alpha^{2(k)}}{2\kappa^{2(k)}}\sum_{t=p+1}^{N} S_{2t}^{(k)} + T\sum_{i=1}^{M} p_{\lambda}\left(\left|\beta_i^{(k)}\right|\right) . \tag{18.34}$$

3. ***CM step.***
Maximize $Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right)$ with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\phi},\alpha,\kappa)$ to obtain $\boldsymbol{\theta}^{(k+1)} = \left(\boldsymbol{\beta}^{(k+1)},\boldsymbol{\phi}^{(k+1)},\alpha^{(k+1)},\kappa^{(k+1)}\right)$. This step yields the following updating equations,

$$\boldsymbol{\beta}_{\mathrm{Sn}}^{(k+1)} = \left(\sum_{t=p+1}^{N}\Phi^{(k)}(B)x_t\Phi^{(k)}(B)x_t^{\top} + \kappa^{2(k)}\Omega\left(p_{i_{\lambda_i}}(\beta^{(k)})\right)\right)^{-1} \times\left(\sum_{t=p+1}^{N}\Phi^{(k)}(B)y_t - \alpha^{(k)}S_{1t}^{(k)}\Phi^{(k)}(B)x_t\right) , \tag{18.35}$$

where $\Omega\left(p_{i_{\lambda_i}}(\beta^{(k)})\right) = \mathrm{diag}\left(\frac{p_{\lambda}^{'}\beta_1^{(0)}}{\beta_1^{(0)}},\ldots,\frac{p_{\lambda}^{'}\beta_M^{(0)}}{\beta_M^{(0)}}\right)$.
Also,

$$\boldsymbol{\Phi}^{(k+1)} = \mathbf{R}^{-1}(\boldsymbol{\beta}^{(k+1)})\left(R_0(\boldsymbol{\beta}^{(k+1)}) + \alpha^{(k)}\sum_{t=p+1}^{N} S_{1t}^{(k)}\right) \tag{18.36}$$

where $R_0(\boldsymbol{\beta}^{(k+1)})$ and $\mathbf{R}(\boldsymbol{\beta}^{(k+1)})$ are defined the same as Eqs. (18.26) and (18.27), respectively, as well as

$$\alpha^{2(k+1)} = \frac{\sum_{t=p+1}^{N} S_{1t}^{(k)}\left(\Phi^{(k+1)}(B)y_t - \Phi^{(k+1)}(B)x_t^{\top}\boldsymbol{\beta}^{(k+1)}\right)^2}{\sum_{t=p+1}^{N} S_{2t}^{(k)}} . \tag{18.37}$$

In addition,

$$\kappa^{2(k+1)} = \frac{1}{T}\left(\sum_{t=p+1}^{N}\left(\Phi^{(k+1)}(B)y_t - \Phi^{(k+1)}(B)x_t^{\top}\boldsymbol{\beta}^{(k+1)}\right)^2\right) -\frac{1}{T}\left(2\alpha^{(k+1)}\sum_{t=p+1}^{N} S_{1t}^{(k)}\left(\Phi^{(k+1)}(B)y_t - \Phi^{(k+1)}(B)x_t^{\top}\boldsymbol{\beta}^{(k+1)}\right)\right)$$

$$+\frac{1}{T}\left(\alpha^{2(k+1)}\sum_{t=p+1}^{N}S_{2t}^{(k)}\right). \tag{18.38}$$

4. ***CM step.*** By using the results of $\alpha^{2(k+1)}$ and $\kappa^{2(k+1)}$, the updating equations for $\sigma^2$ and $\zeta$ can be obtained as

$$\sigma^{2(k+1)} = \kappa^{2(k+1)} + \alpha^{2(k+1)}, \tag{18.39}$$

$$\zeta^{(k+1)} = \delta^{(k+1)}\left(1-\delta^{2(k+1)}\right)^{-1/2}, \tag{18.40}$$

   where $\delta^{(k+1)} = \alpha^{2(k+1)}/\sigma^{2(k+1)}$.
5. The E and CM steps are repeated until the convergence criterion $|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}| < \epsilon$ is met.

### *18.3.3 Skew t Distribution*

In regression analysis, real data may not always be symmetric and may even have outliers. In order to make parameter estimates and variable selection statistically reliable in regression analysis, it is necessary to use different approaches that can deal with outliers and skewness, simultaneously. One of these approaches is to use heavy-tailed skew distributions as the error distribution alternative to the normal or skew normal distributions. In this section, we will consider the regression models with auto-regressive error terms under the assumption that the errors have Azzalini-type skew t distribution. With this assumption, we again carry on parameter estimation and variable selection using penalized maximum likelihood estimation method.

The skew $t$ distribution with $\nu > 0$ degrees of freedom can be defined as a scale mixture of the skew normal and the chi-square distribution (Azzalini, 1985, 1986, 2005; Azzalini & Capitanio, 2003; Branco & Dey, 2001). The p.d.f. of the skew $t$ distribution has the following form,

$$f_{st}(u;\lambda) = \frac{2}{\sigma}t_{\nu}\left(\frac{u-\mu}{\sigma}\right)T_{\nu+1}\left(\zeta\frac{u-\mu}{\sigma}\sqrt{\frac{\nu+1}{\nu+\left(\frac{u-\mu}{\sigma}\right)^2}}\right) \tag{18.41}$$

where $t_\nu$ is the p.d.f. of the t distribution with $\nu$ degrees of freedom and $T_{\nu+1}$ is the c.d.f. of the $t$ distribution with $\nu+1$ degrees of freedom. Here, $\mu \in \mathbb{R}$ is the location, $\sigma \in \mathbb{R}^+$ is the scale, and $\zeta \in \mathbb{R}$ is the skewness parameter. Additionally, $\nu$ (the degree of freedom parameter) is the shape parameter that controls the tail thickness.

We will use the notation $U \sim St_\nu(\mu,\sigma,\zeta)$ if the random variable $U$ has a skew $t$ distribution. The expectation and the variance of $U$ are (Azzalini & Capitanio, 2003)

$$\mathrm{E}[U] = \mu + \sigma\left(\sqrt{\frac{\nu}{\pi}}\zeta\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\sqrt{\zeta^2+1}\Gamma\left(\frac{\nu}{2}\right)}\right) \quad \text{for} \quad \nu > 1 \; and \tag{18.42}$$

$$\mathrm{var}(U) = \sigma^2\left(\frac{\nu}{\nu-2} - \frac{\nu}{\pi}\frac{\zeta^2}{\zeta^2+1}\left(\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^2\right) \quad \text{for} \quad \nu > 2. \tag{18.43}$$

If $Z \sim SN(\lambda)$ and $\tau \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ are two independent random variables, then $Y = \mu + \sigma\, Z / \sqrt{\tau}$ will have $St_\nu(\mu, \sigma^2, \zeta, \nu)$. Using the stochastic representation of the skew normal distribution given by Eq. (18.12), we can have the following hierarchical representation of the skew t distribution,

$$Y|\gamma, \tau \sim N\left(\mu + \sigma\delta\gamma, \frac{(1-\delta^2)}{\tau}\sigma^2\right), \tag{18.44}$$

$$\gamma|\tau \sim TN\left(0, \frac{1}{\tau}; (0, \infty)\right) \tag{18.45}$$

where

$$\gamma = \frac{|U_1|}{\sqrt{\tau}}$$
$$\tau \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

Further, using this hierarchical representation, the joint p.d.f.s of $(y, \gamma, \tau)$ and $(y, \tau)$ can be obtained as

$$f(y, \gamma, \tau) = \frac{1}{\pi\sigma\sqrt{1-\delta^2}}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\tau^{\frac{\nu}{2}}\exp\left(-\frac{\tau\eta^2}{2(1-\delta^2)} - \frac{\nu\tau}{2}\right) \times \exp\left(-\frac{\gamma^2}{2\tau(1-\delta^2)} + \frac{\gamma\tau\delta(y-\mu)}{\sigma\sqrt{1-\delta^2}}\right) \tag{18.46}$$

and

$$f(y, \tau) = \sqrt{\frac{2}{\pi}}\frac{\tau^{\frac{\nu-1}{2}}}{\sigma}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\frac{1}{\Gamma\left(\frac{\nu}{2}\right)}\exp\left(-\frac{\tau}{2}\left(\eta^2+\nu\right)\right)F\left(\zeta\sqrt{\tau}\right), \tag{18.47}$$

respectively. Also, the conditional distribution of $\gamma|\tau, y$ is obtained as

$$f(\gamma|\tau, y) = \frac{\sqrt{\tau}}{\sqrt{2\pi(1-\delta^2)}}\exp\left(-\frac{\tau\left(\gamma - \delta\frac{(y-\mu)}{\sigma}\right)^2}{2(1-\delta^2)}\right)F^{-1}\left(\zeta\eta\sqrt{\tau}\right). \tag{18.48}$$

This conditional distribution is $TN\left(\delta\frac{y-\mu}{\sigma\sqrt{1-\delta^2\tau}};0,\infty\right)$. Similarly, the conditional distribution of $\tau|y$ can be obtained as

$$f(\tau|y)=b\tau^{\frac{\nu+1}{2}}\exp\left(-\tau\frac{(\eta^2+\nu)}{2}\right)F\left(\zeta\eta\sqrt{\tau}\right), \tag{18.49}$$

where $b$ is the normalizing constant and defined as

$$b=\left(\frac{\eta^2+\nu}{2}\right)^{\frac{\nu+1}{2}}\left(\Gamma\left(\frac{\nu+1}{2}\right)T_{\nu+1}\left(\zeta\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)\right)^{-1}.$$

Using the conditional distribution of $\tau$ given $Y$, the conditional expectations of $\tau,\gamma\tau,\gamma^2\tau$, and $\ln\tau$ given $Y=y$, which will be necessary in the EM implementation to compute the estimates, can be obtained as (Lin et al., 2007)

$$\mathrm{E}[\tau|y]=\left(\frac{\nu+1}{\eta^2+\nu}\right)\frac{T_{\nu+3}\left(M\sqrt{\frac{\nu+3}{\nu+1}}\right)}{T_{\nu+1}(M)}, \tag{18.50}$$

$$\mathrm{E}[\gamma\tau|y]=\delta\frac{(y-\mu)}{\sigma}\mathrm{E}[\tau|y]+\frac{\sqrt{1-\delta^2}}{\pi\sigma f(y)}\left(\frac{\eta^2}{\nu(1-\delta^2)}+1\right)^{(-\frac{\nu}{2}+1)}, \tag{18.51}$$

$$\begin{aligned}\mathrm{E}[\gamma^2\tau|y]=&\delta^2\frac{(y-\mu)^2}{\sigma^2}\mathrm{E}[\tau|y]+\sqrt{1-\delta^2}\\&+\delta\sqrt{1-\delta^2}\frac{(y-\mu)}{\pi\sigma^2 f(y)}\left(\frac{\eta^2}{\nu(1-\delta^2)}+1\right)^{(-\frac{\nu}{2}+1)},\end{aligned} \tag{18.52}$$

$$\begin{aligned}\mathrm{E}[\log\tau|y]=&\mathrm{DG}\left(\frac{\nu+1}{2}\right)-\log\left(\frac{\eta^2+\nu}{2}\right)\\&+\left(\frac{\nu+1}{\eta^2+\nu}\right)\left(\frac{T_{\nu+3}\left(\zeta\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}{T_{\nu+1}\left(\zeta\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}-1\right)\\&+\frac{\zeta\eta\left(\eta^2-1\right)}{\sqrt{(\nu+1)(\eta^2+\nu)^3}}\frac{t_{\nu+1}\left(\zeta\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}{T_{\nu+1}\left(\zeta\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}\\&+\frac{1}{T_{\nu+1}\left(\zeta\eta\sqrt{\frac{\nu+1}{\eta^2+\nu}}\right)}\int_{-\infty}^{M}g_\nu(x)t_{\nu+1}(x)\mathrm{d}x\end{aligned} \tag{18.53}$$

where

$$M=\zeta\eta\sqrt{\frac{\nu}{\eta+\nu}},$$

$$f(y) = \frac{2}{\sigma} t_\nu(\eta) T_{\nu+1}(M)\,,$$

$$g_\nu(x) = \mathrm{DG}\left(\frac{\nu+2}{2}\right) - \mathrm{DG}\left(\frac{\nu+1}{2}\right) - \ln\left(1+\frac{x^2}{\nu+1}\right) + \left(\frac{x^2(\nu+1)-\nu-1}{(\nu+1)(\nu+1+x^2)}\right)$$

with $\mathrm{DG}(x)$ as the *digamma function* with $\mathrm{DG}(x) = \Gamma'(x)\,/\,\Gamma(x)$.

#### 18.3.3.1 CML Estimation for Skew t Distribution

Now, consider that the error term of the regression model with AR($p$) structure has an Azzalini-type skew t distribution, that is, $a_t \sim \mathrm{ST}(0, \sigma^2, \zeta, \nu)$. Given the observations $(x_{p+1}, y_{p+1})$, $(x_{p+2}, y_{p+2})$, …, $(x_N, y_N)$, the CML estimators of the unknown parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2, \zeta, \nu)$ can be obtained by maximizing the following log-likelihood function:

$$\ln L = T\ln(2c_\nu) - T\ln(\sigma) - \frac{\nu+1}{2}\sum_{t=p+1}^{N}\ln\left(\nu + a_t^2\right) + \sum_{t=p+1}^{N}\ln\left(T_{\nu+1}\left(\zeta a_t\sqrt{\frac{\nu+1}{\nu+a_t^2}}\right)\right). \tag{18.54}$$

However, since CML estimators cannot be obtained analytically as in skew normal case, numerical methods must be used to obtain estimates. Due to the scale mixture representation of the skew $t$ distribution, the EM-type algorithm can be easily applied to obtain the estimates. To implement the EM-type algorithm, we define $\boldsymbol{\gamma} = (\gamma_{p+1}, \ldots, \gamma_N)$ and $\boldsymbol{\tau} = (\tau_{p+1}, \ldots, \tau_N)$ as missing data, $\mathbf{y} = (y_{p+1}, \ldots, y_N)$ as observed data, and $(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau})$ for $t = p+1, \ldots, N$ as complete data. Then, using the stochastic representation of the skew $t$ distribution given in Sect. 18.3.3, we can write the following complete data log-likelihood function:

$$\ell_c(\boldsymbol{\Theta}; \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = -\frac{T\nu}{2}\ln\left(\frac{\nu}{2}\right) + T\ln\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\sum_{t=p+1}^{N}\ln(\tau_t) - \frac{\nu}{2}\sum_{t=p+1}^{N}\tau_t - T\ln\pi - \frac{T}{2}\ln(\kappa^2) - \sum_{t=p+1}^{N}\left[-\frac{\left(\Phi(B)y_t - \Phi(B)x_t^\top\beta - \alpha\gamma_t\right)^2}{2\kappa^2/\tau_t} + \frac{\gamma_t^2}{2\tau_t}\right]. \tag{18.55}$$

This function is very similar to the log-likelihood function obtained from normal distribution; it is relatively easy to maximize. However, since the random variables $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ are latent variables and the resulting estimators will depend on them, these estimators will be unfeasible to use. To solve the latency problem, we should find the conditional expectation of the complete data log-likelihood function given $\boldsymbol{y}$ and the current estimates for the parameters of interest. The conditional expectation of the complete data log-likelihood function given observed data will be as

$$\begin{aligned}
\mathrm{E}[\ell_c(\Theta; \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\tau})|\Phi(B)y_t] = &-\frac{T\nu}{2}\ln\left(\frac{\nu}{2}\right) + T\ln\Gamma\left(\frac{\nu}{2}\right) \\
&+\frac{\nu}{2}\sum_{t=p+1}^{N}\mathrm{E}\left[\ln\tau_t|y_t\right] - \frac{\nu}{2}\sum_{t=p+1}^{N}\mathrm{E}[\tau_t|y_t] \\
&-\frac{T}{2}\ln\kappa^2 - \sum_{t=p+1}^{N}\frac{\alpha^2\,\mathrm{E}[\gamma_t^2\tau_t|y_t]}{2\kappa^2} \\
&-\sum_{t=p+1}^{N}\frac{\mathrm{E}[\tau_t|y_t]\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)^2}{2\kappa^2} \\
&+\sum_{t=p+1}^{N}\frac{\alpha\,\mathrm{E}[\gamma_t\tau_t|y_t]\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)}{\kappa^2}.
\end{aligned} \tag{18.56}$$

Using the conditional expectations given by Eq. (18.50)–(18.53), we can write the following objective function that should be maximized to obtain the estimates:

$$\begin{aligned}
Q(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}) = &-\frac{T\nu}{2}\ln\left(\frac{\nu}{2}\right) + T\ln\Gamma\left(\frac{\nu}{2}\right) \\
&+\frac{\nu}{2}\sum_{t=p+1}^{N}t_{4t} - \frac{\nu}{2}\sum_{t=p+1}^{N}t_{1t} \\
&-\frac{T}{2}\ln\kappa^2 - \sum_{t=p+1}^{N}\frac{\alpha^2 t_{3t}}{2\kappa^2} \\
&-\sum_{t=p+1}^{N}\frac{t_{1t}\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)^2}{2\kappa^2} \\
&+\sum_{t=p+1}^{N}\frac{\alpha t_{2t}\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)}{\kappa^2}.
\end{aligned} \tag{18.57}$$

Here, $t_{1t} = \mathrm{E}[\tau|y]$, $t_{2t} = \mathrm{E}[\gamma\tau|y]$, $t_{3t} = \mathrm{E}[\gamma^2\tau|y]$, and $t_{4t} = \mathrm{E}[\ln\tau|y]$.

#### 18.3.3.2 Penalized Maximum Likelihood Estimation for Skew *t* Distribution

As in the skew normal distribution, our main goal is to carry out parameter estimation and variable selection simultaneously; we can now combine this function with the penalty function to form our penalized log-likelihood function as

$$\begin{aligned} Q_{\mathrm{PL}}(\boldsymbol{\Theta};\hat{\boldsymbol{\Theta}}) = & -\frac{T\nu}{2}\ln\left(\frac{\nu}{2}\right)+T\ln\Gamma\left(\frac{\nu}{2}\right)+\frac{\nu}{2}\sum_{t=p+1}^{N}t_{4t}-\frac{\nu}{2}\sum_{t=p+1}^{N}t_{1t} \\ & -\frac{T}{2}\ln\kappa^2-\sum_{t=p+1}^{N}\frac{\alpha^2 t_{3t}}{2\kappa^2} \\ & -\sum_{t=p+1}^{N}\frac{t_{1t}\left(\Phi(B)y_t-\Phi(B)x_t^{\top}\boldsymbol{\beta}\right)^2}{2\kappa^2} \\ & +\sum_{t=p+1}^{N}\frac{\alpha t_{3t}\left(\Phi(B)y_t-\Phi(B)x_t^{\top}\boldsymbol{\beta}\right)}{\kappa^2} \\ & +T\sum_{i=1}^{M}p_{\lambda}\left(|\beta_i|\right)+T\sum_{j=1}^{p}p_{\delta}\left(\left|\phi_j\right|\right) \end{aligned} \tag{18.58}$$

where $p_\lambda$ and $p_\delta$ are the penalty functions. By maximizing this function with respect to the unknown parameters, we can carry on with parameter estimation and variable selection. In this study, the parameter $\boldsymbol{\phi}$, which determines the auto-regressive degree of the error terms, will not be included in the model selection process to avoid complexity. By taking the derivatives of $Q_{\mathrm{PL}}(\boldsymbol{\Theta};\hat{\boldsymbol{\Theta}})$ with respect to the parameters $\boldsymbol{\Theta}=(\boldsymbol{\beta},\boldsymbol{\phi},\sigma^2,\zeta,\nu)$ and setting the resulting equations to zero, we obtain the following updating equations for parameter estimation,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{St}} = & \left(\sum_{t=p+1}^{N}\hat{t}_{1t}\hat{\Phi}(B)x_t\hat{\Phi}(B)x_t^{\top}+\hat{\kappa}^2\Omega\left(p_{i_\lambda}(\beta)\right)\right)^{-1} \\ & \times\left(\sum_{t=p+1}^{N}\hat{t}_{1t}\hat{\Phi}(B)y_t-\hat{\alpha}\hat{t}_{2t}\hat{\Phi}(B)x_t\right) \end{aligned} \tag{18.59}$$

where $\Omega\left(p_{i_\lambda}(\beta)\right)=\mathrm{diag}\left(p_\lambda^{'}\beta_1^{(0)}/\beta_1^{(0)},\ldots,p_\lambda^{'}\beta_M^{(0)}/\beta_M^{(0)}\right)$, $\beta_i^{(0)}\neq 0$ is an initial point for $\beta_i$, and $p_\lambda^{'}$ is the derivative of the penalty function. Estimating equations are obtained by using the penalty terms of the method of interest in Eq. (18.59). Furthermore, estimators for the auto-regressive parameters, the error scale, and the

skewness parameters can be obtained as follows. Since the following equations will not be used in variable selection case, they will remain the same in each penalty method,

$$\hat{\boldsymbol{\Phi}} = \hat{t}_1 \mathbf{R}^{-1}\left(\hat{\boldsymbol{\beta}}\right)\left(\hat{t}_1 R_0\left(\hat{\boldsymbol{\beta}}\right) + \hat{\alpha}\hat{t}_2\right) \tag{18.60}$$

where $R_0(\hat{\beta})$ and $\mathbf{R}^{-1}\left(\hat{\boldsymbol{\beta}}\right)$ are defined in Eqs. (18.26) and (18.27), respectively.

The updating equations for the error scale and the skewness parameters are

$$\hat{\sigma}^2 = \hat{\kappa}^2 + \hat{\alpha}^2\,, \tag{18.61}$$

$$\hat{\zeta} = \frac{\hat{\delta}}{\left(1-\hat{\delta}^2\right)^{-1/2}}\,, \tag{18.62}$$

where

$$\begin{aligned}\hat{\kappa}^2 = \frac{1}{T}\Bigg(&\sum_{t=p+1}^{N}\left(\hat{t}_{1t}\hat{\Phi}(B)y_t - \hat{\Phi}(B)x_t^{\top}\hat{\boldsymbol{\beta}}\right)^2 \\ &-2\hat{\alpha}\hat{t}_{2t}\left(\hat{\Phi}(B)y_t - \hat{\Phi}(B)x_t^{\top}\hat{\boldsymbol{\beta}}\right) + \hat{\alpha}^2\hat{t}_{3t}\Bigg)\end{aligned} \tag{18.63}$$

and $\hat{\delta} = \hat{\alpha}^2 / \hat{\sigma}^2$.

In this study, the parameter $\nu$ is considered as known to ensure robustness. This parameter can be estimated from the solution of the following equation if desired:

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^k)}{\partial \nu} &= \frac{T}{2}\ln\left(\frac{\nu}{2}+1\right) - \frac{\Gamma'(\nu/2)}{\Gamma(\nu/2)} + \frac{1}{2}\sum_{t=p+1}^{N}\hat{t}_{4t} - \frac{1}{2}\sum_{t=p+1}^{N}\hat{t}_{1t} \\ &= \ln\left(\frac{\nu}{2}\right) + 1 - \mathrm{DG}\left(\frac{\nu}{2}\right)\frac{1}{T}\sum_{t=p+1}^{N}\hat{t}_{4t} - \hat{t}_{1t} = 0.\end{aligned} \tag{18.64}$$

The estimator defined in Eq. (18.59) will be used to estimate simultaneous parameters with variable selection in the auto-regressive error term regression model under the assumption of a skew $t$ distribution.

Now, we can give the steps of the ECM algorithm to compute the estimates of the parameters under the assumption that the error terms are skew t distributed.

Steps of the ECM Algorithm

1. Determine the initial value $\boldsymbol{\Theta}^{(1)} = \left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\phi}^{(1)}, \sigma^{2(1)}, \zeta^{(1)}, \nu^{(1)}\right)$ of the parameter vector $\boldsymbol{\Theta}$, and stop criterion $\epsilon$.

2. *E step.* To calculate the $(k+1)$-th iteration for $k = 1, 2, \ldots$, the following conditional expected values are calculated for $t = p+1, \ldots, N$:

$$
\begin{aligned}
t_{1t}^{(k)} &= \mathrm{E}[\tau_t | \Phi(B) y_t, \boldsymbol{\Theta}^{(k)}] \\
&= \frac{\nu^{(k)} + 1}{\eta_t^{2(k)} + \nu^{(k)}} \frac{T_{\nu^{(k)}+3}\left(M^{(k)} \sqrt{\frac{\nu^{(k)}+3}{\nu^{(k)}+1}}\right)}{T_{\nu^{(k)}+1}(M^{(k)})}, \qquad (18.65) \\
t_{2t}^{(k)} &= \mathrm{E}[\gamma_t \tau_t | \Phi(B) y_t, \boldsymbol{\Theta}^{(k)}] \\
&= \delta^{(k)} \frac{\Phi(B)(y_t) - \Phi(B) x_t^{\top} \beta^{(k)}}{\sigma^{(k)}} t_{1t}^{(k)} \\
&\quad + \frac{\sqrt{1-\delta^{2(k)}}}{\pi \sigma^{(k)} f(y_t^{(k)})} \left( \frac{\eta_t^{2(k)}}{\nu^{(k)}(1-\delta^{2(k)})} + 1 \right)^{\left(-\frac{\nu^{(k)}}{2}+1\right)}, \qquad (18.66) \\
t_{3t}^{(k)} &= \mathrm{E}[\gamma_t^2 \tau_t | \Phi(B) y_t, \boldsymbol{\Theta}^{(k)}] \\
&= \delta^{(k)} \frac{(\Phi(B) y_t - \Phi(B) x_t^{\top} \beta^{(k)})}{\sigma^{(k)}} t_{1t}^{(k)} + (1 - \delta^{2(k)}) \\
&\quad + \frac{\delta^{(k)} \sqrt{1-\delta^{2(k)}} (\Phi(B) y_t - \Phi(B) x_t^{\top} \beta^{(k)})}{\pi \sigma^{(k)} f(y_t^{(k)})} \\
&\quad \times \left( \frac{\eta_t^{2(k)}}{\nu^{(k)}(1-\delta^{2(k)})} + 1 \right)^{-\left(\frac{\nu^{(k)}}{2}+1\right)}, \qquad (18.67) \\
t_{4t}^{(k)} &= \mathrm{E}[\log \tau_t | \Phi(B) y_t, \boldsymbol{\Theta}^{(k)}] \\
&= \mathrm{DG}\left( \frac{\nu^{(k)} + 1}{2} \right) - \log \left( \frac{\eta^{2(k)} + \nu^{(k)}}{2} \right) \\
&\quad + \left( \frac{\nu^{(k)} + 1}{\eta^{2(k)} + \nu^{(k)}} \right) \left( \frac{T_{\nu^{(k)}+3}\left(\zeta^{(k)} \eta^{(k)} \sqrt{\frac{\nu^{(k)}+1}{\eta^{2(k)}+\nu^{(k)}}}\right)}{T_{\nu^{(k)}+1}\left(\zeta^{(k)} \eta^{(k)} \sqrt{\frac{\nu^{(k)}+1}{\eta^{2(k)}+\nu^{(k)}}}\right)} - 1 \right) \\
&\quad + \frac{\zeta^{(k)} \eta^{(k)} \left(\eta^{2(k)} - 1\right)}{\sqrt{(\nu^{(k)}+1)(\eta^{2(k)}+\nu^{(k)})^3}} \frac{t_{\nu^{(k)}+1}\left(\zeta^{(k)} \eta^{(k)} \sqrt{\frac{\nu^{(k)}+1}{\eta^{2(k)}+\nu^{(k)}}}\right)}{T_{\nu^{(k)}+1}\left(\zeta^{(k)} \eta^{(k)} \sqrt{\frac{\nu^{(k)}+1}{\eta^{2(k)}+\nu^{(k)}}}\right)} \\
&\quad + \frac{1}{T_{\nu^{(k)}+1}\left(\zeta^{(k)} \eta^{(k)} \sqrt{\frac{\nu^{(k)}+1}{\eta^{2(k)}+\nu^{(k)}}}\right)} \int_{-\infty}^{M^{(k)}} g_{\nu}^{(k)}(x) t_{\nu^{(k)}+1}(x) \mathrm{d}x \,. \qquad (18.68)
\end{aligned}
$$

Here,

$$\eta_t^{(k)} = \frac{\Phi(B)y_t - \Phi(B)x_t^\top \beta^{(k)}}{\sigma^{(k)}} ,$$

$$\delta^{(k)} = \frac{\zeta^{(k)}}{\sqrt{1+\zeta^{2(k)}}}$$

$$M_t^{(k)} = \zeta^{(k)}\eta_t^{(k)}\frac{\sqrt{\nu^{(k)}}}{\eta_t^{(k)}+\nu^{(k)}}$$

$$f(y_t^{(k)}) = \frac{2}{\sigma^{(k)}} t_{\nu^{(k)}}(\eta_t^{(k)}) T_{\nu^{(k)}+1}(M_t^{(k)})$$

$$g_{\nu^{(k)}}(x) = \mathrm{DG}\left(\frac{\nu^{(k)}+2}{2}\right) - \mathrm{DG}\left(\frac{\nu^{(k)}+1}{2}\right) - \ln\left(1+\frac{x^2}{\nu^{(k)}+1}\right) + \left(\frac{x^2(\nu^{(k)}+1)-\nu^{(k)}-1}{(\nu^{(k)}+1)(\nu^{(k)}+1+x^2)}\right),$$

with $\mathrm{DG}(x)$ as the *digamma function* with $\Gamma'(x)/\Gamma(x)$. We can now form our penalized objective function, $Q_{\mathrm{PL}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k)})$, as

$$\begin{aligned} Q_{\mathrm{PL}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k)}) = {} & \frac{T\nu}{2}\ln\left(\frac{\nu}{2}\right) + T\ln\Gamma\left(\frac{\nu}{2}\right) \\ & + \frac{\nu}{2}\sum_{t=p+1}^{N} t_{4t}^{(k)} - \frac{\nu}{2}\sum_{t=p+1}^{N} t_{1t}^{(k)} - \frac{T}{2}\ln\kappa^2 \\ & - \sum_{t=p+1}^{N}\frac{\alpha^2 t_{3t}^{(k)}}{2\kappa^2} - \sum_{t=p+1}^{N}\frac{t_{1t}^{(k)}\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)^2}{2\kappa^2} \\ & + \sum_{t=p+1}^{N}\frac{\alpha t_{2t}^{(k)}\left(\Phi(B)y_t - \Phi(B)x_t^\top\boldsymbol{\beta}\right)}{\kappa^2} \\ & + T\sum_{i=1}^{M} p_\lambda(|\beta_i|) + T\sum_{j=1}^{p} p_\delta\left(\left|\phi_j\right|\right) . \end{aligned} \tag{18.69}$$

We can update $\boldsymbol{\Theta}^{(k)}$ by maximizing $Q_{\mathrm{PL}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k)})$ over $\boldsymbol{\Theta}$ which leads to the following updating equations.

3. *CM step.*
   Update $\boldsymbol{\beta}^{(k)}$ by maximizing $Q_{\mathrm{PL}}(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k)})$ over $\boldsymbol{\beta}$, leading to

$$\boldsymbol{\beta}^{(k+1)} = \left( \sum_{t=p+1}^{N} t_{1t}^{(k)} \Phi^{(k)}(B) x_t \Phi^{(k)}(B) x_t^{\top} \right)^{-1} \times \left( \sum_{t=p+1}^{N} t_{1t}^{(k)} \Phi^{(k)}(B) y_t - \alpha^{(k)} t_{2t}^{(k)} \Phi^{(k)}(B) x_t \right) . \tag{18.70}$$

4. *CM step.*
Fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$ and update $\boldsymbol{\Phi}^{(k)}$ using the following updating equation:

$$\boldsymbol{\Phi}^{(k+1)} = \sum_{t=p+1}^{N} t_{1t}^{(k)} \mathbf{R}(\boldsymbol{\beta})^{-1} \left( t_{1t}^{(k)} \mathbf{R}_0(\boldsymbol{\beta}) + \alpha^{(k)} t_{2t}^{(k)} \right) . \tag{18.71}$$

5. *CM step.*
Fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{(k+1)}$ update $\alpha^{2(k)}$ using

$$\alpha^{2(k+1)} = \frac{\sum_{t=p+1}^{N} t_2^{(k)} \left( \Phi^{(k)}(B) y_t - \Phi^{(k)}(B) x_t^{\top} \boldsymbol{\beta}^{(k)} \right)^2}{\sum_{t=p+1}^{N} t_{3t}^{(k)}} . \tag{18.72}$$

6. *CM step.*
Fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{(k+1)}$, and $\alpha^2 = \alpha^{2(k+1)}$, update $\kappa^{2(k)}$ using the following equation:

$$\kappa^{2(k+1)} = \frac{1}{T} \sum_{t=p+1}^{N} t_{1t}^{(k)} \left( \Phi^{(k)}(B) y_t - \Phi^{(k)}(B) x_t^{\top} \boldsymbol{\beta}^{(k)} \right)^2 - \frac{2}{T} \sum_{t=p+1}^{N} \alpha^{(k)} t_{2t}^{(k)} \left( \Phi^{(k)}(B) y_t - \Phi^{(k)}(B) x_t^{\top} \boldsymbol{\beta}^{(k)} \right) + \alpha^{2(k)} t_{3t}^{(k)} . \tag{18.73}$$

7. *CM step.*
Fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{(k+1)}$, $\alpha^2 = \alpha^{2(k+1)}$, and $\kappa^2 = \kappa^{2(k+1)}$, and update $\sigma^{2(k)}$ by using

$$\sigma^{2(k+1)} = \kappa^{2(k+1)} + \alpha^{2(k+1)} . \tag{18.74}$$

8. *CM step.*
Fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\Phi} = \boldsymbol{\Phi}^{(k+1)}$, $\alpha^2 = \alpha^{2(k+1)}$, $\kappa^2 = \kappa^{2(k+1)}$, and $\sigma^2 = \sigma^{2(k+1)}$ to obtain $\zeta^{(k+1)}$ as the solution of

$$T\delta\left(1-\delta^2\right)\delta\sum_{t=p+1}^{N}t_{1t}^{(k)}\left(\Phi^{(k+1)}(B)y_t-\Phi^{(k+1)}(B)x_t^{\top}\boldsymbol{\beta}^{(k+1)}\right)^2/\sigma^{2(k+1)}$$
$$+\sum_{t=p+1}^{N}\frac{t_{3t}^{(k)}}{\sigma^{2(k+1)}} \tag{18.75}$$
$$+\left(1+\delta^2\right)\sum_{t=p+1}^{N}t_{2t}^{(k)}\left(\Phi^{(k+1)}(B)y_t-\Phi^{(k+1)}(B)x_t^{\top}\boldsymbol{\beta}^{(k+1)}\right)/\sigma^{(k+1)}=0$$

or the following equation can be used,

$$\zeta^{(k+1)}=\frac{\delta^{(k+1)}}{\left(1-\delta^{2(k+1)}\right)^{-1/2}} \tag{18.76}$$

where $\delta^{(k+1)}=\alpha^{2(k+1)}/\sigma^{2(k+1)}$.

9. *CM step.*
Fix $\boldsymbol{\beta}=\boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\phi}=\boldsymbol{\phi}^{(k+1)}$, $\alpha^2=\alpha^{2(k+1)}$, $\kappa^2=\kappa^{2(k+1)}$, $\sigma^2=\sigma^{2(k+1)}$, and $\zeta=\zeta^{(k+1)}$ to obtain $\nu^{(k+1)}$ as a solution of

$$\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^k)}{\partial\nu}=\frac{T}{2}\ln\left(\frac{\nu}{2}+1\right)-\frac{\Gamma'(\nu/2)}{\Gamma(\nu/2)}+\frac{1}{2}\sum_{t=p+1}^{N}t_{4t}^{(k)}-\frac{1}{2}\sum_{t=p+1}^{N}t_{1t}^{(k)}$$
$$=\ln\left(\frac{\nu}{2}\right)+1-\mathrm{DG}\left(\frac{\nu}{2}\right)\frac{1}{T}\sum_{t=p+1}^{N}t_{4t}^{(k)}-t_{1t}^{(k)}=0\,. \tag{18.77}$$

10. These steps are repeated until the convergence criterion $\left\|\boldsymbol{\Theta}^{(k+1)}-\boldsymbol{\Theta}^{(k)}\right\|<\epsilon$ is met.

## 18.4 Numerical Examples

We provide a simulation study and a real data example to compare the performance of parameter estimation and variable selection of the OLS, LASSO, SCAD, ridge, bridge, and elastic net methods for auto-regressive error term regression models under the assumption that the errors are skew normal and skew $t$ distributed. The entirety of these simulations and the real data are performed using the R software (version 4.2.3) (R Core Team, 2020).

### *18.4.1 Simulation*

In this subsection, which is dedicated to simulation studies, simulation scenarios and results will be presented, respectively.

#### 18.4.1.1 Simulation Design

Explanatory variables for the auto-regressive error term regression model to be used in the simulation are produced using the multivariate normal distribution. $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0M})$ and $\boldsymbol{\phi}_0 = (\phi_{01}, \phi_{02}, \ldots, \phi_{0p})$ denote the regression and auto-regressive model parameter vectors, respectively. During the construction of the model, the auto-regressive model parameters were determined to ensure the stationary of the auto-regressive part of the model. The values of the response variable $(y_t)$ were obtained using model defined in Eq. (18.3).

The error term $a_t$ is assumed to be distributed with normal, Azzalini-type skew normal and skew $t$. Here, the degrees of freedom for the skew $t$ distribution are considered as constant, with $\nu = 3$. For cases without outliers, the variance of the error term is considered to be $\sigma^2 = 1$. The optimal value of the parameter $\lambda$ for the LASSO method is obtained by performing a grid search on $[0, 2]$ and calculating the BIC values for each $\lambda$ value. The model with the smallest BIC value is selected as the best model, and the corresponding $\lambda$ value is selected as the optimal value. Similarly, for the bridge method, the optimal values of parameters $\gamma$ and $\lambda$ are obtained using two-dimensional grid search. The model corresponding to the smallest calculated BIC value was considered the best model. For SCAD, the value of $\alpha = 3.7$ is used as suggested by Fan and Li (2001). For the elastic net method, $\lambda_1$ and $\lambda_2$ were scanned simultaneously, and the smallest of the BIC values for each result was obtained as the best result. For this method, for $\lambda_1$, the range is taken as in the LASSO method, and the values of $[0, 0.5, 1, 1.5]$ are used for case $\lambda_2$. The sample sizes for all cases were taken as $N = 50, 100, 300$. The simulation results reported were obtained from 100 replications.

**Estimation Performance** In order to measure the estimation performances of the proposed methods, the following root mean squared error (RMSE) values of the parameters are calculated and reported in the tables:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{18.78}$$

**Variable Selection Performance** In order to measure the performance of the model selection criteria, the correctly predicted zero coefficients were counted under each case, and their ratios were calculated (Correct). Similarly, the coefficients that were estimated to be zero, although they should not have been zero in the model, were also counted, and their ratios were calculated (Incorrect). At the same time,

the correct model ratio obtained in 100 repetitions was calculated by counting the correct zero amounts given in the real model (Corr. fit). For the auto-regressive model, in each case, four extra zeros were added for the coefficient of the auto-regressive degree, and the number of correct auto-regressive degrees was counted in 100 repetitions (AR degree).

Three different simulation scenarios with different settings are designed to evaluate the performance of the methods used in terms of parameter estimation and variable selection:

*Case I*

In this case, we consider eight covariates for the model defined in Eq. (18.3) where $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. The variables $x_t = (x_{t1}, \ldots, x_{t8})^\top$ are independently produced from a multivariate normal distribution such that $\mathbf{x}_{ti} \sim \mathrm{N}_M(\mathbf{O}, \boldsymbol{\Sigma})$ and the pairwise correlations between the variables $x_{t,i_1}$ and $x_{t,i_2}$ are $(0.5)^{|i_1 - i_2|}$. Error terms were generated from a skew normal distribution as $a_t \sim \mathrm{SN}(\lambda = 0.2, 0, 1)$ and from a skew $t$ distribution as $a_t \sim \mathrm{St}_3(\lambda = 0.2, 0, 1)$. Here, the order of the AR process is taken as $p = 1$ and $p = 2$. The actual values of AR($p$) model parameters are taken as $\phi = 0.8$ for first-order AR and $\boldsymbol{\phi} = (1.2, -0.7)^\top$ for second-order AR.

*Case II*

In order to measure the performance of the methods under high-dimensional case, a regression model with a total of 40 variables, 20 statistically significant and 20 insignificant, was produced with the following real parameter values:

$$\boldsymbol{\beta} = \left( \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10} \right) \tag{18.79}$$

The parameters of the AR($p$) model were used as $\phi = 0.8$ in the case of $q = 1$ and $\boldsymbol{\Phi} = (1.2, -0.7)^\top$ in the case of $p = 2$. Explanatory variables, response variables, and error distribution assumptions were considered as explained in *Case I*.

*Case III*

In order to demonstrate the robustness of estimators based on the skew $t$ distribution against outliers, the model designed in Case I was reconsidered with outliers. Let $\varepsilon$ represent the contamination rate with a value of 0.1, and let $m = [\varepsilon N]$ represent the exact value of the contaminated data, where $N$ is the sample size. $m$ of data contaminated with $t = p+1, p+2, \ldots, p+m$ were generated from the distribution

$N(50, 1)$; the remaining data $t = p+m+1, p+m+2, \ldots, N$ was $N(0, 1)$. The real value of the model parameters $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$ and AR($p$) is $\phi = 0.8$ for $p = 1$.

#### 18.4.1.2 Simulation Results

All the simulation results on the selection of variables and parameter estimations are given in Tables 18.1, 18.2, 18.3, 18.4, 18.5, 18.6, 18.7, 18.8, 18.9, and 18.10. In Tables 18.1, 18.2, 18.3, and 18.4, the results for Case I are summarized. Tables 18.1 and 18.2 show the results of the calculations made with the skew normal and skew $t$ assumptions, respectively, with first-order auto-regressive error terms. From Table 18.1, it can be observed that all the methods used have close performances regarding the parameter estimation according to the RMSE values. It can be easily seen that LASSO, SCAD, bridge, and elastic net methods have equally superior performances compared to OLS and ridge methods regarding variable selection and determining the correct degree of AR. In Table 18.2, the results are presented under the assumption of a skew $t$ distribution to represent heavy-tailed and skew cases. From this table, it can be said that both the parameter estimation and variable selection performances of the all methods are affected by heavy-tailedness. On the other hand, SCAD, bridge, and elastic net methods are less affected by heavy-tailed errors than OLS and ridge methods in terms of variable selection. Tables 18.3 and

**Table 18.1** Skew normal results for Case I and AR(1)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 0.75672 | 0.19 | 0.00 | 0.00 | 0 |
| | LASSO | 0.84227 | 4.98 | 0.00 | 1.00 | 100 |
| | SCAD | 0.83358 | 4.97 | 0.00 | 1.00 | 100 |
| | Ridge | 0.94107 | 0.17 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma = 0.5$) | 0.77641 | 4.99 | 0.00 | 1.00 | 100 |
| | Elastic net | 0.78873 | 4.99 | 0.00 | 1.00 | 100 |
| 100 | OLS | 0.62813 | 0.32 | 0.00 | 0.00 | 0 |
| | LASSO | 0.70687 | 4.99 | 0.00 | 1.00 | 100 |
| | SCAD | 0.71119 | 4.99 | 0.00 | 1.00 | 100 |
| | Ridge | 0.88413 | 0.39 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma = 0.5$) | 0.72519 | 4.99 | 0.00 | 1.00 | 100 |
| | Elastic net | 0.72274 | 4.99 | 0.00 | 1.00 | 100 |
| 300 | OLS | 0.21292 | 0.63 | 0.00 | 0.00 | 0 |
| | LASSO | 0.14151 | 5.00 | 0.00 | 1.00 | 100 |
| | SCAD | 0.14096 | 5.00 | 0.00 | 1.00 | 100 |
| | Ridge | 0.14926 | 1.72 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma = 0.6$) | 0.14586 | 5.00 | 0.00 | 1.00 | 100 |
| | Elastic net | 0.14644 | 5.00 | 0.00 | 1.00 | 100 |

**Table 18.2** Skew *t* results for Case I and AR(1)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 6.27004 | 0.07 | 0.00 | 0.00 | 1 |
| | LASSO | 5.38272 | 4.92 | 0.00 | 1.00 | 75 |
| | SCAD | 4.36782 | 4.91 | 0.00 | 1.00 | 79 |
| | Ridge | 9.15211 | 0.03 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma = 0.3$) | 1.32212 | 4.95 | 0.00 | 1.00 | 95 |
| | Elastic net | 1.42161 | 4.94 | 0.00 | 1.00 | 92 |
| 100 | OLS | 4.71634 | 0.12 | 0.00 | 0.00 | 3 |
| | LASSO | 3.36237 | 4.94 | 0.00 | 1.00 | 79 |
| | SCAD | 2.64797 | 4.95 | 0.00 | 1.00 | 82 |
| | Ridge | 5.79192 | 0.14 | 0.00 | 0.00 | 3 |
| | Bridge ($\gamma = 0.3$) | 1.23198 | 4.99 | 0.00 | 1.00 | 99 |
| | Elastic net | 1.28862 | 4.98 | 0.00 | 1.00 | 97 |
| 300 | OLS | 1.54206 | 0.46 | 0.00 | 0.00 | 5 |
| | LASSO | 2.38521 | 5.00 | 0.00 | 1.00 | 87 |
| | SCAD | 2.08271 | 5.00 | 0.00 | 1.00 | 90 |
| | Ridge | 3.38121 | 0.55 | 0.00 | 0.00 | 6 |
| | Bridge ($\gamma = 0.3$) | 0.88275 | 5.00 | 0.00 | 1.00 | 100 |
| | Elastic net | 1.01174 | 5.00 | 0.00 | 1.00 | 100 |

**Table 18.3** Skew normal results for Case I and AR(2)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 1.32874 | 0.09 | 0.00 | 0.00 | 0 |
| | LASSO | 2.68375 | 4.92 | 0.00 | 1.00 | 99 |
| | SCAD | 1.63358 | 4.92 | 0.00 | 1.00 | 99 |
| | Ridge | 2.84107 | 0.10 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma$=0.7) | 1.30554 | 4.94 | 0.00 | 1.00 | 100 |
| | Elastic net | 1.16273 | 4.93 | 0.00 | 1.00 | 100 |
| 100 | OLS | 0.94135 | 0.12 | 0.00 | 0.00 | 0 |
| | LASSO | 1.52066 | 4.95 | 0.00 | 1.00 | 100 |
| | SCAD | 1.38245 | 4.96 | 0.00 | 1.00 | 100 |
| | Ridge | 2.12141 | 0.19 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma$=0.8) | 1.13119 | 4.98 | 0.00 | 1.00 | 100 |
| | Elastic net | 0.82691 | 4.99 | 0.00 | 1.00 | 100 |
| 300 | OLS | 0.65248 | 0.31 | 0.00 | 0.00 | 0 |
| | LASSO | 1.39091 | 5.00 | 0.00 | 1.00 | 100 |
| | SCAD | 0.94215 | 5.00 | 0.00 | 1.00 | 100 |
| | Ridge | 1.33245 | 0.42 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma$=0.7) | 0.70577 | 5.00 | 0.00 | 1.00 | 100 |
| | Elastic net | 0.73907 | 5.00 | 0.00 | 1.00 | 100 |

**Table 18.4** Skew $t$ results for Case I and AR(2)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 9.11264 | 0.12 | 0.00 | 0.00 | 1 |
| | LASSO | 7.53555 | 4.94 | 0.00 | 1.00 | 70 |
| | SCAD | 7.23121 | 4.98 | 0.00 | 1.00 | 71 |
| | Ridge | 10.8359 | 0.04 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma = 0.4$) | 3.37744 | 4.99 | 0.00 | 1.00 | 94 |
| | Elastic net | 2.74349 | 4.98 | 0.00 | 1.00 | 91 |
| 100 | OLS | 6.74062 | 0.19 | 0.00 | 0.00 | 2 |
| | LASSO | 6.31065 | 5.00 | 0.00 | 1.00 | 78 |
| | SCAD | 4.25052 | 5.00 | 0.00 | 1.00 | 82 |
| | Ridge | 8.28776 | 0.21 | 0.00 | 0.00 | 1 |
| | Bridge ($\gamma = 0.5$) | 2.67432 | 5.00 | 0.00 | 1.00 | 98 |
| | Elastic net | 2.41269 | 5.00 | 0.00 | 1.00 | 94 |
| 300 | OLS | 4.77396 | 0.25 | 0.00 | 0.00 | 3 |
| | LASSO | 4.84763 | 5.00 | 0.00 | 1.00 | 82 |
| | SCAD | 3.67702 | 5.00 | 0.00 | 1.00 | 89 |
| | Ridge | 6.67764 | 0.36 | 0.00 | 0.00 | 5 |
| | Bridge ($\gamma = 0.3$) | 1.70667 | 5.00 | 0.00 | 1.00 | 100 |
| | Elastic net | 1.84872 | 5.00 | 0.00 | 1.00 | 100 |

**Table 18.5** Skew normal results for Case II and AR(1)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 153.371 | 0.05 | 0.02 | 0.00 | 0 |
| | LASSO | 18.9886 | 19.90 | 0.00 | 1.00 | 100 |
| | SCAD | 23.8413 | 19.91 | 0.00 | 1.00 | 100 |
| | Ridge | 25.4831 | 0.24 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.8$) | 5.25402 | 19.96 | 0.00 | 1.00 | 100 |
| | Elastic net | 4.96496 | 19.95 | 0.00 | 1.00 | 100 |
| 100 | OLS | 9.84961 | 0.12 | 0.01 | 0.00 | 0 |
| | LASSO | 14.0061 | 19.92 | 0.00 | 1.00 | 100 |
| | SCAD | 15.1936 | 19.93 | 0.00 | 1.00 | 100 |
| | Ridge | 3.50364 | 0.30 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.8$) | 3.20183 | 19.98 | 0.00 | 1.00 | 100 |
| | Elastic net | 2.54721 | 19.97 | 0.00 | 1.00 | 100 |
| 300 | OLS | 77.0228 | 0.23 | 0.01 | 0.00 | 0 |
| | LASSO | 7.92019 | 19.95 | 0.00 | 1.00 | 100 |
| | SCAD | 11.0226 | 19.95 | 0.00 | 1.00 | 100 |
| | Ridge | 11.2441 | 0.41 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.7$) | 1.93828 | 20.00 | 0.00 | 1.00 | 100 |
| | Elastic net | 1.82779 | 20.00 | 0.00 | 1.00 | 100 |

**Table 18.6** Skew *t* results for Case II and AR(1)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 596.370 | 0.01 | 0.02 | 0.00 | 0 |
| | LASSO | 62.9105 | 19.91 | 0.00 | 1.00 | 80 |
| | SCAD | 43.7098 | 19.90 | 0.00 | 1.00 | 81 |
| | Ridge | 72.9647 | 0.03 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.5$) | 6.40076 | 19.92 | 0.00 | 1.00 | 87 |
| | Elastic net | 6.20525 | 19.93 | 0.00 | 1.00 | 88 |
| 100 | OLS | 257.696 | 0.03 | 0.01 | 0.00 | 1 |
| | LASSO | 42.7374 | 19.93 | 0.00 | 1.00 | 85 |
| | SCAD | 24.3421 | 19.95 | 0.00 | 1.00 | 89 |
| | Ridge | 52.9388 | 0.06 | 0.01 | 0.00 | 1 |
| | Bridge ($\gamma = 0.6$) | 5.26127 | 19.96 | 0.00 | 1.00 | 90 |
| | Elastic net | 5.36708 | 19.96 | 0.00 | 1.00 | 90 |
| 300 | OLS | 179.401 | 0.08 | 0.01 | 0.00 | 1 |
| | LASSO | 29.9325 | 19.95 | 0.00 | 1.00 | 93 |
| | SCAD | 19.9322 | 19.95 | 0.00 | 1.00 | 94 |
| | Ridge | 43.2943 | 0.12 | 0.01 | 0.00 | 1 |
| | Bridge ($\gamma = 0.6$) | 4.97501 | 19.98 | 0.00 | 1.00 | 96 |
| | Elastic net | 5.25141 | 19.98 | 0.00 | 1.00 | 96 |

**Table 18.7** Skew normal results for Case II and AR(2)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 201.527 | 0.02 | 0.02 | 0.00 | 0 |
| | LASSO | 24.6611 | 19.92 | 0.00 | 1.00 | 100 |
| | SCAD | 30.5723 | 19.91 | 0.00 | 1.00 | 100 |
| | Ridge | 28.2912 | 0.20 | 0.01 | 0.00 | 0 |
| | Bridge $\gamma = 0.7$ | 5.72688 | 19.93 | 0.00 | 1.00 | 100 |
| | Elastic net | 6.21428 | 19.92 | 0.00 | 1.00 | 100 |
| 100 | OLS | 132.550 | 0.06 | 0.01 | 0.00 | 0 |
| | LASSO | 21.7181 | 19.94 | 0.00 | 1.00 | 100 |
| | SCAD | 21.6325 | 19.94 | 0.00 | 1.00 | 100 |
| | Ridge | 23.0624 | 0.24 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.7$) | 5.06114 | 19.96 | 0.00 | 1.00 | 100 |
| | Elastic net | 5.39975 | 19.95 | 0.00 | 1.00 | 100 |
| 300 | OLS | 95.0699 | 0.34 | 0.01 | 0.00 | 0 |
| | LASSO | 13.5405 | 19.96 | 0.00 | 1.00 | 100 |
| | SCAD | 15.1031 | 19.96 | 0.00 | 1.00 | 100 |
| | Ridge | 13.6138 | 0.46 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.8$) | 4.24545 | 19.98 | 0.00 | 1.00 | 100 |
| | Elastic net | 4.37516 | 19.97 | 0.00 | 1.00 | 100 |

**Table 18.8** Skew *t* results for Case II and AR(2)

| *N* | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 871.672 | 0.01 | 0.02 | 0.00 | 0 |
| | LASSO | 78.2475 | 19.84 | 0.00 | 1.00 | 45 |
| | SCAD | 80.2884 | 19.85 | 0.00 | 1.00 | 72 |
| | Ridge | 101.886 | 0.18 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.5$) | 23.2888 | 19.9 | 0.00 | 1.00 | 98 |
| | Elastic net | 22.9305 | 19.89 | 0.00 | 1.00 | 97 |
| 100 | OLS | 668.253 | 0.06 | 0.01 | 0.00 | 0 |
| | LASSO | 69.9621 | 19.86 | 0.00 | 1.00 | 51 |
| | SCAD | 71.3488 | 19.87 | 0.00 | 1.00 | 80 |
| | Ridge | 88.3965 | 0.24 | 0.01 | 0.00 | 0 |
| | Bridge ($\gamma = 0.6$) | 21.7774 | 19.92 | 0.00 | 1.00 | 99 |
| | Elastic net | 21.1160 | 19.91 | 0.00 | 1.00 | 99 |
| 300 | OLS | 518.381 | 0.17 | 0.00 | 0.00 | 0 |
| | LASSO | 60.0242 | 19.89 | 0.00 | 1.00 | 62 |
| | SCAD | 63.2577 | 19.89 | 0.00 | 1.00 | 91 |
| | Ridge | 79.5837 | 0.35 | 0.00 | 0.00 | 0 |
| | Bridge ($\gamma = 0.6$) | 19.2554 | 19.94 | 0.00 | 1.00 | 100 |
| | Elastic net | 19.2577 | 19.94 | 0.00 | 1.00 | 100 |

**Table 18.9** Skew normal results for Case III and AR(1)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 8595.69 | 0.00 | 0.21 | 0.00 | 0 |
| | LASSO | 2594.26 | 3.88 | 0.02 | 0.88 | 32 |
| | SCAD | 172.308 | 3.75 | 0.09 | 0.75 | 46 |
| | Ridge | 1635.02 | 0.00 | 0.20 | 0.00 | 0 |
| | Bridge ($\gamma = 0.8$) | 99.3096 | 4.20 | 0.02 | 0.99 | 89 |
| | Elastic net | 95.9513 | 4.03 | 0.03 | 0.97 | 82 |
| 100 | OLS | 4261.36 | 0.00 | 0.18 | 0.00 | 0 |
| | LASSO | 1604.50 | 3.96 | 0.01 | 0.88 | 45 |
| | SCAD | 132.109 | 3.87 | 0.06 | 0.75 | 59 |
| | Ridge | 881.294 | 0.00 | 0.12 | 0.00 | 0 |
| | Bridge ($\gamma = 0.7$) | 68.4108 | 4.65 | 0.01 | 0.99 | 92 |
| | Elastic net | 66.2998 | 4.56 | 0.01 | 0.97 | 91 |
| 300 | OLS | 1105.40 | 0.00 | 0.00 | 0.00 | 0 |
| | LASSO | 1059.21 | 4.55 | 0.08 | 1.00 | 72 |
| | SCAD | 88.4068 | 4.31 | 0.00 | 0.98 | 79 |
| | Ridge | 468.694 | 0.00 | 0.02 | 0.00 | 0 |
| | Bridge ($\gamma = 0.8$) | 57.2758 | 4.82 | 0.00 | 1.00 | 100 |
| | Elastic net | 58.3881 | 4.81 | 0.00 | 1.00 | 100 |

**Table 18.10** Skew $t$ results for Case III and AR(1)

| N | Method | RMSE | Correct | Incorrect | Corr. fit | AR degree |
|---|---|---|---|---|---|---|
| 50 | OLS | 345.803 | 0.02 | 0.26 | 0.00 | 0 |
| | LASSO | 306.994 | 4.00 | 0.00 | 1.00 | 45 |
| | SCAD | 101.811 | 4.25 | 0.01 | 0.97 | 76 |
| | Ridge | 103.138 | 0.05 | 0.19 | 0.00 | 0 |
| | Bridge ($\gamma = 0.4$) | 51.5142 | 4.92 | 0.00 | 1.00 | 95 |
| | Elastic net | 55.3098 | 4.90 | 0.00 | 1.00 | 92 |
| 100 | OLS | 189.274 | 0.08 | 0.21 | 0.00 | 0 |
| | LASSO | 169.982 | 4.12 | 0.00 | 1.00 | 62 |
| | SCAD | 85.3696 | 4.34 | 0.00 | 1.00 | 81 |
| | Ridge | 94.4173 | 0.11 | 0.10 | 0.00 | 0 |
| | Bridge ($\gamma = 0.5$) | 21.1368 | 4.95 | 0.00 | 1.00 | 98 |
| | Elastic net | 25.3511 | 4.96 | 0.00 | 1.00 | 99 |
| 300 | OLS | 47.1527 | 0.20 | 0.17 | 0.00 | 0 |
| | LASSO | 65.6419 | 4.71 | 0.00 | 1.00 | 75 |
| | SCAD | 79.2785 | 4.85 | 0.00 | 1.00 | 90 |
| | Ridge | 71.6095 | 0.22 | 0.04 | 0.00 | 0 |
| | Bridge ($\gamma = 0.5$) | 10.1527 | 5.00 | 0.00 | 1.00 | 100 |
| | Elastic net | 15.1679 | 5.00 | 0.00 | 1.00 | 100 |

18.4 show the results where the error terms have second-degree autocorrelation. When Tables 18.3 and 18.4 are examined, it can be said that as the degree of auto-regressive structure increases, small deteriorations are observed in both parameter estimates and model selection. This can be observed from the worsening of the RMSE values in the tables.

The residual analysis was performed for this model after applying the OLS method, and the results of the residual analysis are shown in Fig. 18.1.

The simulation results obtained as a result of the high-dimensional design are shown in Tables 18.5, 18.6, 18.7, and 18.8. From Table 18.5, it can be said that the bridge and elastic net methods have better performance in terms of parameter estimation performances according to their RMSE values. The model selection performances of the LASSO, SCAD, bridge, and elastic net methods have shown that they can identify significant variables without difficulty even under a high-dimensional situation. In Table 18.6, high-dimensional situations can be seen together with the skewed and heavy-tailed error assumption. From Table 18.6, the negative effect of heavy-tailedness on parameter estimation performance can be seen on OLS, LASSO, and ridge methods. On the other hand, SCAD, bridge, and elastic net methods have better results on parameter estimation performance under the effects of heavy-tailedness, skewness, and high dimensions. Tables 18.7 and 18.8 show the assumptions of skew normal and skew t distribution with second-order auto-regressive error terms, respectively. Similar to Case I, we can say that the parameter estimation performance of the mentioned methods worsens as the auto-

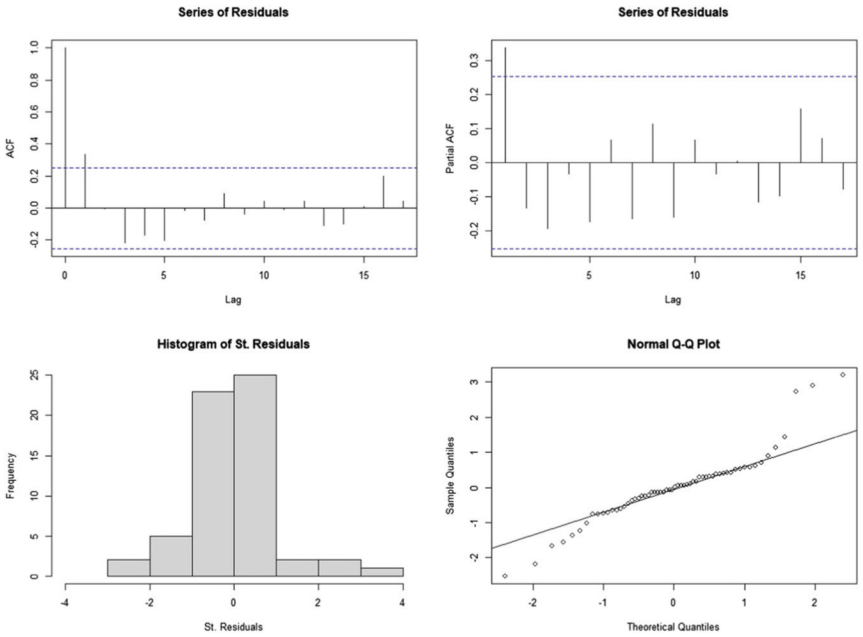



**Fig. 18.1** Residual analysis. ACF (top left), PACF (top right), histogram (bottom left), and Q-Q plot (bottom right)

regressive effect increases. However, bridge and elastic net methods generally show better performance in parameter estimation and variable selection. The most striking part about this simulation is seen in the correct determination of the auto-regressive degree among the methods. Generally, bridge and elastic net methods have better performance among the others in terms of determining the correct auto-regressive degree. Although LASSO and SCAD methods have equal performance to bridge and elastic net in terms of determining significant variables in the regression model, bridge and elastic net have slightly better performance than other methods.

Tables 18.9 and 18.10 contain the simulation results for the data obtained with 10% contamination and design with AR(1) error terms. From Table 18.9, it is clearly seen that the parameter estimation and variable selection performances of all methods deteriorate if there are some vertical outliers in the data under the skew normal distribution assumption. Table 18.10 contains the results obtained from the skew t distribution. From Table 18.10, all the methods are less affected from the outliers in the data. Among these methods, bridge and elastic net perform better than others for both parameter estimation and variable selection performances. Especially when the sample size increased, the variable selection performances of the bridge and elastic net methods improved for both the selection of significant variables for the regression model and the correct AR degree.

**Table 18.11** Data variables and explanations

| Variable | Explanation |
|---|---|
| y | Follow-up return rate |
| $x_1$ | Unemployment rate |
| $x_2$ | Inflation |
| $x_3$ | Interest rates |
| $x_4$ | Total industrial production index |
| $x_5$ | Exchange rate |
| $x_6$ | GDP |
| $x_7$ | Export |
| $x_8$ | Import |
| $x_9$ | Credits |
| $x_{10}$ | Deposit |
| $x_{11}$ | Active |
| $x_{12}$ | Deposit/active |
| $x_{13}$ | Credit/deposit |
| $x_{14}$ | Credits/active |

### *18.4.2 Real Data*

For the application of real-world data, it has been investigated on which data types the loan return rates in Türkiye may depend economically, using the loan return rate data obtained from the Central Bank of the Republic of Türkiye (TCMB) (Retrieved from https://evds2.tcmb.gov.tr/ on 2021-09-30). The data used were collected in quarters between 2005 and 2019 with a sample size $N = 42$. The resulting dataset contains 1 response and 14 explanatory variables. The names of the variables are given in Table 18.11.

According to the ACF and PACF graphs in Fig. 18.1, the errors show a first-order auto-regressive structure. This result is also supported by Durbin-Watson test (autocorrelation $= 0.9044997$, D-W $= 0.0362028$, $P$-value $< 10^{-16}$). The histogram of the residuals suggests the presence of skewness or heavy tails in the data. This observation is supported by the Shapiro-Wilk test, which confirms that the residuals are nonnormally distributed ($W = 0.93921$, $P$-value $= 0.02011$). Finally, it was determined that there might be some outliers observed in the data from the Q-Q graph. According to these results, it has been seen that modeling the data under the assumption of a first-order autoregressive error term (AR(1)) and a heavy-tailed skew distribution will be the most suitable solution. The results of the model selection and analysis under these assumptions are given in Table 18.12.

As a result of the real data analysis, according to BIC values, it can be said that the analyses performed under the skew $t$ distribution assumption modeled the data more successfully than the analyses performed under the skew normal distribution due to outliers in the data. In terms of model selection, it has been observed that the bridge method provides better modeling than other methods according to BIC values. Accordingly, the variables selected by bridge regression for this data with

**Table 18.12** Results for real data application

| | Skew Normal | | Skew $t$ ($\nu = 4$) | |
|---|---|---|---|---|
| | Sel. variables | BIC | Sel. variables | BIC |
| LASSO | (3,4,5,9,11) | 95.12 | (3,4,5,9,11) | 84.22 |
| SCAD | (3,4,5,9) | 87.25 | (3,4,5,9) | 79.57 |
| Bridge | (3,4,5,11) | 71.19 | (3,4,5,11) | 62.59 |
| Elastic net | (3,4,5,9) | 90.25 | (3,4,5,9) | 85.14 |

skewed and outlier values with auto-regressive error term are $x_3$, interest rates; $x_4$, total industrial production index; $x_5$, exchange rate; and $x_{11}$, active with parameter estimates—$x_3 = -0.02514$, $x_4 = 1.25149$, $x_5 = -0.98741$, and $x_{11} = 1.74136$.

## 18.5 Conclusion

In this chapter, the penalized maximum likelihood method based on skewed distributions has been investigated to perform parameter estimation and variable selection simultaneously in regression models with auto-regressive error terms when datasets have skewed or skewed heavy-tailed distributions. We have assumed that the error terms of our regression model have both Azzalini-type skew normal and Azzalini-type skew $t$ distributions. To form the penalized log-likelihood function, the penalty functions of LASSO, SCAD, bridge, and elastic net methods have been used. To obtain estimates of model parameters, an EM-type algorithm is proposed that uses stochastic representations of the distributions for both the Azzalini-type skew normal and Azzalini-type skew $t$ distribution assumptions. A simulation study under different scenarios and a real-world application have been given to demonstrate the variable selection and parameter estimation performances of the resulting estimators. We have also examined the ability of the proposed methods to accurately detect the auto-regressive degree. The results of the numerical studies have shown that all the variable selection methods considered in this chapter have similar performance for datasets with moderate skewness and no outliers. In the presence of outliers and skewness, the bridge and elastic net estimators obtained under the skew $t$ assumption showed superior results than the other methods.

**Acknowledgments** The authors would like to express sincere gratitude to the editor and the two anonymous reviewers for their insightful comments and constructive feedback, which greatly contributed to improving the quality of this study. This study is part of the Ph.D. dissertation of Tuaç (2020) at Ankara University, Graduate School of Natural and Applied Sciences, Thesis number: 651682.

## References


Alpuim, T., & El-Shaarawi, A. (2008). On the efficiency of regression analysis with AR(p) errors. *Journal of Applied Statistics, 35*(7), 717–737.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, 12*(2), 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica, 46*(2), 199–208.

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*, 367–389.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion). *Scandinavian Journal of Statistics, 32*(2), 159–188.

Beach, C. M., & MacKinnon, J. G. (1978). A maximum likelihood procedure for regression with autocorrelated errors. *Econometrica: Journal of the Econometric Society, 46*(1), 51–58.

Branco, M. D., & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis, 79*(1), 79–113.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1–38.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*(2), 301–320.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution. *Scandinavian Journal of Statistics, 13*(4), 271–275.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Lin, T. I., Lee, J. C., & Hsieh, W. J. (2007). Robust mixture modeling using the skew t-distribution. *Statistics and Computing, 17*, 81–92.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*(2), 267–278.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Tiku, M. L., Wong, W. K., & Bian, G. (1999). Time series models with asymmetric innovations. *Communications in Statistics—Theory and Methods, 28*(6), 1331–1360.

Tuaç, Y., Güney, Y., Şenoğlu, B., & Arslan, O. (2018). Robust parameter estimation of regression model with AR(p) error terms. *Communications in Statistics—Simulation and Computation, 47*(8), 2343–2359.

Tuaç, Y., Güney, Y., & Arslan, O. (2020). Parameter estimation of regression model with AR(p) error terms based on skew distributions with EM algorithm. *Soft Computing, 24*(5), 3309–3330.

Tuaç, Y. (2020). *Robust parameter estimation and model selection in autoregressive error term regression models (Unpublished doctoral dissertation)*. Ankara University, Graduate School of Natural and Applied Sciences, Department of Statistics.

Wang, H., Li, G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-LASSO. *Journal of Business & Economic Statistics, 25*, 347–355.

Yoon, Y. J., Park, C., & Lee, T. (2013). Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation, 83*(9), 1756–1772.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

# Part III
# Robust Estimation and Nonparametrics Methods

# Chapter 19
# Shrinkage Estimator for Spatial Auto-regressive Model with Endogenous Covariates




**Anoop Chaturvedi, Christian Heumann, and Shalabh**

## 19.1 Introduction

The shrinkage estimation in various models has played a prominent role in econometrics. The family of Stein-rule estimators (SREs) proposed by James and Stein (1960) has smaller variability than ordinary least squares (OLS) estimator under quadratic risk with a simple condition that the number of explanatory variables is more than two. The use of SRE in real data problems has also received considerable attention in the literature, e.g., Knight et al. (1992), Knight et al. (1993a, 1993b). The book by Saleh (2006) is a valuable resource for understanding the broader landscape of shrinkage estimation techniques and their applications in econometrics and statistics. The book provides a comprehensive review of shrinkage methods, discussing various approaches and their theoretical underpinnings.

The weighted ordinary least squares and two-stage least squares (2SLS) estimators have been derived under a local misspecification framework, where the assumed model might not perfectly match the true data-generating process in the vicinity of the true parameter values. Sawa (1973) initiated the idea of combining OLS and 2SLS estimators with the objective of reducing the bias. Employing the small sigma asymptotic theory, Maasoumi (1978) developed a Stein-like estimator for

A. Chaturvedi
Department of Statistics, University of Allahabad, Allahabad, India
e-mail: anoopchaturv@gmail.com

C. Heumann
Department of Statistics, Ludwig Maximilians University, Munich, Germany
e-mail: Christian.Heumann@lmu.de

Shalabh (✉)
Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, India
e-mail: shalab@iitk.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi. M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_19

reduced form coefficients in simultaneous equations that stemmed from applying Stein shrinkage with the weight depending inversely on the Hausman (1978) statistic for exogeneity, leading to a shrinkage estimator with finite moments. Ullah and Srivastava (1988) analyzed the properties of a Stein-type estimator of the coefficients of the structural simultaneous equations model.

Model averaging has indeed gained traction in both applied and theoretical research as a valuable tool for estimating regression models. It is particularly useful when dealing with uncertainty about which model specification is the best fit for the data or when there's concern about model selection bias. Weighted-average least squares (WALS), considered by Magnus and Durbin (1999) within the framework of estimating some of the parameters of a linear regression model when other coefficients are of no interest, is one such model averaging method with their proposed approach being a Bayesian combination of frequentist ordinary least squares and restricted least squares estimators (see also Danilov & Magnus, 2004). They also established an equivalence theorem, which simplifies the problem of estimating the coefficients of interest to that of obtaining an optimal estimator of the vector of coefficients of no interest given a single observation from a normal distribution. Zou et al. (2007) considered the model with non-normal disturbances and extended the equivalence theorem for it. Chaturvedi et al. (2012) further extended the framework of Magnus and Durbin (1999) for the model with nonspherical disturbances.

Hansen (2017) revisits Maasoumi's idea and considered a shrinkage estimator that combines OLS and 2SLS estimators, weighted by the Hausman (1978) statistic for exogeneity. He investigated the efficacy of a Stein-like shrinkage estimator using asymptotic tools and demonstrated the potential for risk reduction compared to traditional 2SLS estimation. Hensen's estimator can be viewed as a smoothed version of the hard-threshold Hausman pretest estimator. His finding aligns with classic results by Stein (1956) and James and Stein (1960), showing that shrinkage can reduce risk, particularly when the dimensionality of the problem exceeds two. Clarke (2017) specifically addresses situations where there is ambiguity regarding the endogeneity of one or more regressors and utilizes the framework established by Danilov and Magnus (2004) along with a Bayesian approach to form weighted-average least squares, which combines the OLS and 2SLS estimators.

When some of the regressors have errors in variables, the OLS estimator is inconsistent, and typically, the instrumental variable (IV) estimator is recommended. Motivated by Hansen (2017), Huan et al. (2017) utilized the Hausman (1978) test for errors in variables to combine OLS and IV and explore the asymptotic properties and risk of this combined estimator within a local asymptotic framework. They, further, demonstrate that the resulting combined estimator outperforms the IV estimator when the dimension of regressors exceeds two.

Spatial econometric models have been specifically designed to account for spatial interactions, externalities, and interdependence in causal relationships among spatially referenced data. A crucial component of these models is the spatial weight matrix, which quantifies the extent of spatial interdependence between different geographic regions or economic agents. Spatial auto-regressive (SAR)

models stand out in this domain by assuming that the response variable's level depends on the levels of the same variable in neighboring regions, thereby capturing spatial spillover effects. Anselin (1988) laid the groundwork for understanding spatial econometrics by delving into its theoretical aspects. Anselin and Bera (1998) contributed significantly by providing comprehensive overviews of spatial econometric models. LeSage and Pace (2009) comprehensively discussed various spatial econometric models, in which SAR, the prominent ones, are spatial Durbin model (SDM) and spatial error model (SEM). These models not only capture spatial dependence but also offer insights into how interactions among spatial units influence the variable under study. They discussed the theoretical aspects of spatial econometric models and provided practical guidance on classical and Bayesian estimation procedures for these models.

This chapter considers the spatial auto-regressive (SAR) model with endogenous and exogenous regressors. The plan of this chapter is as follows: The SAR model with endogenous and exogenous regressors is described in Sect. 19.2. The combined shrinkage estimator and its asymptotic distribution are presented in Sect. 19.3. The comparison of asymptotic risks is demonstrated in Sect. 19.4. The results from a Monte Carlo simulation are presented in Sect. 19.5 followed by some conclusions in Sect. 19.6.

## 19.2 SAR Model with Endogenous and Exogeneous Regressors

Let $X_i$, $i \in \{1, \ldots, \}$ be $(m \times 1)$ observation vectors on endogenous regressors and $Z_i = (Z_{1i} \quad Z_{2i})$ vectors of exogenous regressors of order $(l + k) \times 1$ with $\mathrm{E}[Z_i^\top \epsilon_i] = 0$. Consider the SAR model involving endogenous and exogenous regressors,

$$y_i = \lambda \sum_{j \neq i} w_{ij} y_j + X_i^\top \beta + Z_{1i}^\top \gamma + \epsilon_i; \quad i \in \{1, \ldots, n\} \tag{19.1}$$

where $\beta$ and $\gamma$ are the regression coefficients associated with $X_i$s and $Z_i$s, respectively, $\lambda$ is the parameter associated with the spatial weights, and $\epsilon_i$ is the disturbance term with

$$\mathrm{E}[\epsilon_i] = 0,$$

$$\mathrm{E}[\epsilon_i \epsilon_j] = \begin{cases} \sigma^2, & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The reduced form equation for the endogenous regressors is

$$X_i = \pi_1^\top Z_{1i} + \pi_2^\top Z_{2i} + v_i; \quad \mathrm{E}[Z_i \, v_i^\top] = 0. \tag{19.2}$$

Here, $\pi_1$ and $\pi_2$ are the $(l \times m)$ and $(k \times m)$ matrices of reduced form coefficients. Following Hansen (2017), we write the structural equation error $\epsilon_i$ as the following linear function of the reduced form error $v_i^\top$ and an orthogonal error $\eta_i$,

$$\epsilon_i = v_i^\top \rho + \eta_i \tag{19.3}$$

where $\eta_i$ are identically and independently distributed with the following properties:

$$\begin{aligned}
\mathrm{E}[\eta_i] &= 0, \\
\mathrm{E}[\eta_i^2] &= \sigma_\eta^2, \\
\mathrm{E}[v_i \eta_i] &= 0, \\
\mathrm{E}[v_i v_i^\top] &= \Sigma, \\
\mathrm{E}[v_i v_j^\top] &= 0 \quad \forall\, i \neq j
\end{aligned}$$

Here, $\Sigma$ is a positive definite and symmetric matrix. Then,

$$\begin{aligned}
\mathrm{E}[\epsilon_i^2] &= \sigma^2 \\
&= \mathrm{E}[(v_i^\top \rho + \eta_i)^\top (v_i^\top \rho + \eta_i)] \\
&= \rho^\top \Sigma \rho + \sigma_\eta^2
\end{aligned}$$

and for $i \neq j$

$$\mathrm{E}[\epsilon_i^\top \epsilon_j] = \mathrm{E}[(v_i^\top \rho + \eta_i)^\top (v_j^\top \rho + \eta_j)] = 0.$$

Further, we assume $\rho = n^{-\frac{1}{2}}\delta$, where $\delta$ is a $(l \times 1)$ vector. For the coefficient $\beta$ to be identified, we require the condition $\mathrm{rank}(\pi_2) = m$. Let $\hat{\lambda}$ be a consistent estimator of $\lambda$, so that

$$\sqrt{n}(\hat{\lambda} - \lambda) \approx o_p(n^{-\frac{1}{2}}).$$

Then,

$$\begin{aligned}
\mathrm{E}[X_i \epsilon_i] &= \mathrm{E}[(\pi_1^\top Z_{1i} + \pi_{2i}^\top Z_{2i} + v_i)(v_i^\top \rho + \eta_i)] \\
&= \Sigma\, \rho \\
&= n^{-\frac{1}{2}}\, \Sigma \delta.
\end{aligned}$$

If $\delta = 0$, then $X_i$ is exogeneous, whereas if $\delta \neq 0$, then $X_i$ is endogenous. We also assume that

$$\mathrm{E}[Z_i\, Z_i^\top] = Q$$

where $Q$ is a finite positive definite matrix.

Stacking the $n$ in Eqs. (19.1) and (19.2), we can write

$$\begin{aligned} y &= \lambda W y + X\beta + Z_1\gamma + \epsilon \\ X &= Z_1\pi_1 + Z_2\pi_2 + V; \quad \mathrm{E}[ZV] = 0. \end{aligned} \tag{19.4}$$

Here, $y$ and $\epsilon$ are $(n \times 1)$ vectors; $W = \left(w_{ij}\right)$ is an $n \times n$ matrix of known spatial weights; $X$ is an $n \times m$ matrix of observations on exogenous variables; $Z_1$ and $Z_2$ are $(n \times l)$ and $(n \times k)$ matrices, respectively, of observations on endogenous variables; $Z = (Z_1 \quad Z_2)$ is a matrix of order $(n \times (l + k))$; and $V = \left(v_1^\top, v_2^\top, \ldots, v_n^\top\right)^\top$ is a matrix of order $(n \times m)$.

For a matrix $A$, let us define

$$P_A = A\left(A^\top A\right)^{-1} A^\top,$$

$$\bar{P}_A = I - A\left(A^\top A\right)^{-1} A^\top.$$

Then, the OLS estimator of $\beta$ is given by

$$\hat{\beta}^{\mathrm{OLS}} = b^{1,\mathrm{LS}} - \hat{\lambda} b^{2,\mathrm{LS}}$$

with

$$b^{1,\mathrm{LS}} = \left(X^\top \bar{P}_{Z_1} X\right)^{-1} X^\top \bar{P}_{Z_1} y,$$

$$b^{2,\mathrm{LS}} = \left(X^\top \bar{P}_{Z_1} X\right)^{-1} X^\top \bar{P}_{Z_1}\, W\, y.$$

In practice, we may take a consistent estimator of $\lambda$ as

$$\hat{\lambda} = \frac{y^\top \bar{P}_Z\, W\, y}{y^\top \bar{P}_Z\, y}.$$

Further, the 2SLS estimator of $\beta$ is

$$\hat{\beta}^{\mathrm{TLS}} = b^{1,\mathrm{TLS}} - \hat{\lambda}\, b^{2,\mathrm{TLS}}$$

with

$$b^{1,\mathrm{TLS}} = \left(\hat{X}^\top \bar{P}_{Z_1} \hat{X}\right)^{-1} \hat{X}^\top \bar{P}_{Z_1}\, y,$$

$$b^{2,\mathrm{TLS}} = \left(\hat{X}^\top \bar{P}_{Z_1} \hat{X}\right)^{-1} \hat{X}^\top \bar{P}_{Z_1}\, W\, y,$$

where

$$\begin{aligned}\hat{X} &= Z\left(Z^\top Z\right)^{-1} Z^\top X \\ &= P_Z\, X.\end{aligned}$$

We can write

$$\begin{aligned}\hat{\beta}^{\mathrm{OLS}} &= b^{1,\mathrm{LS}} - \hat{\lambda} b^{2,\mathrm{LS}} \\ &= b^{1,\mathrm{LS}} - \lambda b^{2,\mathrm{LS}} - (\hat{\lambda} - \lambda) b^{2,\mathrm{LS}},\end{aligned}$$

$$\begin{aligned}\hat{\beta}^{\mathrm{TLS}} &= b^{1,\mathrm{TLS}} - \hat{\lambda} b^{2,\mathrm{TLS}} \\ &= b^{1,\mathrm{TLS}} - \lambda b^{2,\mathrm{TLS}} - \left(\hat{\lambda} - \lambda\right) b^{2,\mathrm{TLS}}.\end{aligned}$$

Let us write

$$M = \bar{P}_{Z_1} - \bar{P}_{Z_1} X \left(X^\top \bar{P}_{Z_1}\, X\right)^{-1} X^\top \bar{P}_{Z_1}.$$

Then, an estimator of $\sigma^2$ is given as

$$\hat{\sigma}^2 = \frac{(y - \hat{\lambda}\, W\, y)^\top M\, (y - \hat{\lambda}\, W\, y)}{\mathrm{tr}(M)}.$$

## 19.3 Combined Shrinkage Estimator and Asymptotic Distribution

The Hausman-Wu statistic for testing the null hypothesis of exogeneity is given as

$$H_n = n(\hat{\beta}^{\mathrm{TLS}} - \hat{\beta}^{\mathrm{LS}})^\top (V_2 - V_1)^{-1} (\hat{\beta}^{\mathrm{TLS}} - \hat{\beta}^{\mathrm{LS}}).$$

Then, the combined estimator is obtained as

$$\hat{\beta}^{\mathrm{S}} = w\hat{\beta}^{\mathrm{LS}} + (1 - w)\hat{\beta}^{\mathrm{TLS}}$$

with

$$w = \min\left\{\frac{\tau}{H_n}, 1\right\}$$

where $\tau$ is the nonnegative shrinkage parameter and $w$ is the weight associated with $\hat{\beta}^{\mathrm{LS}}$ in the combined estimator. Following Hansen (2017), we can write

$$\sqrt{n}(\hat{\beta}^{\mathrm{OLS}} - \beta,\ \hat{\beta}^{\mathrm{TLS}} - \beta)^{\top} \xrightarrow{d} h + \xi$$

where $h = \sigma^{-2} V_1 \Sigma \delta$ and $\xi \sim N(0, V)$.

The matrix $V$ is defined as

$$V = \begin{pmatrix} V_1 & V_1 \\ V_1 & V_2 \end{pmatrix}$$

with the components $V_1$ and $V_2$ given by

$$\begin{aligned} V_1 &= \sigma^2 \left\{ \mathrm{E}[X_i X_i^{\top}] - \mathrm{E}[X_i Z_{1i}^{\top}] \left(\mathrm{E}[Z_{1i} Z_{1i}^{\top}]\right)^{-1} E\left(Z_{1i} X_i^{\top}\right) \right\}^{-1}, \\ V_2 &= \sigma^2 \left\{ \mathrm{E}[X_i Z_i^{\top}] \left(\mathrm{E}[Z_{1i} Z_{1i}^{\top}]\right)^{-1} \mathrm{E}[Z_i X_i^{\top}] \right. \\ &\qquad \left. - \mathrm{E}[X_i Z_{1i}^{\top}](\mathrm{E}[Z_{1i} Z_{1i}^{\top}])^{-1} \mathrm{E}[Z_{1i} X_i^{\top}] \right\}^{-1}. \end{aligned}$$

We can consistently estimate $V_1$ and $V_2$ by

$$\hat{V}_1 = \hat{\sigma}^2 \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i X_i^{\top} - \frac{1}{n} \sum_{i=1}^{n} X_i Z_{1i}^{\top} \left( \frac{1}{n} \sum_{i} Z_{1i} Z_{1i}^{\top} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_{1i} X_i^{\top} \right\}^{-1}$$

and

$$\begin{aligned} \hat{V}_2 = \hat{\sigma}^2 \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i Z_i^{\top} \left( \frac{1}{n} \sum_{i=1}^{n} Z_{1i} Z_i^{\top} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_i X_i^{\top} \right. \\ \left. - \frac{1}{n} \sum_{i=1}^{n} X_i Z_{1i}^{\top} \left( \frac{1}{n} \sum_{i=1}^{n} Z_{1i} Z_{1i}^{\top} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_{1i} X_i^{\top} \right\}^{-1}, \end{aligned}$$

respectively. Then

$$H_n \underset{d}{\to} (h + \xi)^{\top} P (h + \xi)$$

where

$$P = G(V_2 - V_1)^{-1}G^\top = GUG^\top,$$

$$U = (V_2 - V_1)^{-1}$$

$$G = (-I \;\; I)^\top$$

$$G_2 = (0 \;\; I)^\top$$

$$(a)_1 = \min\{1, a\}.$$

Hence, we can write

$$\begin{aligned}\sqrt{n}(\hat{\beta}^{\mathrm{S}} - \beta) &= \sqrt{n}\left[(w\hat{\beta}^{\mathrm{LS}} - \beta) + (1-w)(\hat{\beta}^{\mathrm{TLS}} - \beta)\right] \\ &= \sqrt{n}\left[(\hat{\beta}^{\mathrm{TLS}} - \beta) - w(\hat{\beta}^{\mathrm{TLS}} - \beta^{\mathrm{LS}})\right] \\ &\underset{d}{\to} G_2^\top \xi - \left(\frac{\tau}{(h+\xi)^\top P(h+\xi)}\right)_1 G^\top(h+\xi).\end{aligned}$$

## 19.4 Comparison of Asymptotic Risks

Now, we observe combining the OLS estimator $\hat{\beta}^{\mathrm{LS}}$ and the 2SLS estimator $\hat{\beta}^{\mathrm{TLS}}$ can result in any improvement in estimation precision.

Let

$$R\left(\hat{\beta}^{\mathrm{S}}\right) = n\ \mathrm{E}\left[\left(\hat{\beta}^{\mathrm{S}} - \beta\right)^\top \left(\hat{\beta}^{\mathrm{S}} - \beta\right)\right]. \tag{19.5}$$

Then, without positive part trimming,

$$\begin{aligned}R\left(\hat{\beta}^{\mathrm{S}}\right) = R\left(\hat{\beta}^{\mathrm{TLS}}\right) &- \mathrm{E}\left[\frac{\tau}{(h+\xi)^\top\ P\ (h+\xi)} 2\xi^\top\ G_2\ U\ G^\top\ (h+\xi)\right] \\ &+ \mathrm{E}\left[\frac{\tau^2}{\left\{(h+\xi)^\top\ P\ (h+\xi)\right\}^2}\ (h+\xi)^\top\ G\ U\ G^\top\ (h+\xi)\right]. \end{aligned} \tag{19.6}$$

Let

$$\eta(x) = \frac{x}{x^\top\ P\ x}.$$

Then

$$\frac{\partial}{\partial x}(\eta(x))^{\top} = \frac{1}{x^{\top}Px} I - \frac{2}{\left(x^{\top}Px\right)^2} Px\,x^{\top}.$$

Further, we can easily verify that

$$G^{\top}VG = G_2^{\top}VG = V_2 - V_1 = U^{-1} \tag{19.7}$$

$$GUG_2^{\top}VP = GUG_2^{\top}VGUG^{\top} = GUG^{\top} = P. \tag{19.8}$$

Hence,

$$\mathrm{E}[\xi\eta(h+\xi)^{\top}] = \mathrm{E}\left[\frac{\partial}{\partial\xi}\frac{(h+\xi)^{\top}}{(h+\xi)^{\top}P(h+\xi)}\right]$$

$$= \frac{1}{(h+\xi)^{\top}P(h+\xi)} I - \frac{2}{\left((h+\xi)^{\top}P(h+\xi)\right)^2} P(h+\xi)(h+\xi)^{\top}.$$

Therefore,

$$\mathrm{E}[\eta(h+\xi)^{\top}GUG_2^{\top}\xi] = \frac{1}{(h+\xi)^{\top}P(h+\xi)}\,\mathrm{tr}\left(GUG_2^{\top}V\right)$$

$$-\frac{2}{\left((h+\xi)^{\top}P(h+\xi)\right)^2}(h+\xi)^{\top}GUG_2^{\top}VP(h+\xi).$$

Using the results (19.7) and (19.8), we obtain

$$\mathrm{tr}(GUG_2^{\top}V) = \mathrm{tr}(UG_2^{\top}VG) = \mathrm{tr}(UU^{-1}) = \mathrm{tr}(I_m) = m$$

and

$$(h+\xi)^{\top}GUG^{\top}(h+\xi) = (h+\xi)^{\top}P(h+\xi),$$

so that

$$\mathrm{E}[\eta(h+\xi)^{\top}GUG_2^{\top}\xi] = \frac{m-2}{(h+\xi)^{\top}P(h+\xi)}.$$

Hence,

$$R(\hat{\beta}^{\mathrm{TLS}}) - R(\hat{\beta}^{\mathrm{S}}) = \mathrm{E}\left[\frac{2\tau(m-2)}{(h+\xi)^{\top}P(h+\xi)}\right] - \mathrm{E}\left[\frac{\tau^2}{(h+\xi)^{\top}P(h+\xi)}\right]$$

$$= \tau\,\mathrm{E}\left[\frac{1}{(h+\xi)^{\top}P(h+\xi)}\right]\{2(m-2)-\tau\} \tag{19.9}$$

which is nonnegative as long as

$$0 < \tau < 2(m-2), \quad m > 2.$$

## 19.5 Monte Carlo Simulation

The large sample properties of the estimator give an idea about the performance of the estimator when the sample size grows large. To understand the finite sample behavior of the estimator, we conducted a simulation study. For the simulation study, the data is created according to the reduced form in Eq. (19.2) and the error structure in Eq. (19.3). To incorporate the spatial dependence of the response variable, we choose a block-diagonal structure. To understand the way it is used in the simulation, we consider the following example to explain it. Consider a sample size of $n = 50$ and a block size $q = 5$; then, the weight matrix $W$ is calculated as a block-diagonal matrix with $n/q$ blocks where each block is generated as $(1/(q-1))(1,\ 1^\top - I_q)$.

The following parameters are chosen for the simulation study:

- The simulation is repeated 1000 times for each parameter setting.
- The number of columns of $X_i$ is chosen as $m = 4$.
- The elements of the coefficient vector $\beta$ are set to $\beta = (1.0, 1.2, 1.2, 1.2)^\top$.
- The covariance matrix $\Sigma$ is the positive definite (correlation) matrix with 1s on the diagonal elements and 0.4 on the off-diagonal elements.
- The multivariate random variables $v_i$ are simulated as multivariate normal variables with mean zero and covariance matrix $\Sigma$.
- The error terms $\eta_i$ are generated as independent standard normal variables.
- The vector $\rho$ is computed as a constant vector with elements in the open interval $(0,\ 1)$.
- The matrix $\pi_1$ is set to zero while $\pi_2$ is the constant $c = \sqrt{R_2/(1-R_2)}$ where $R_2$ is set to 0.4, that is, $X_i = c\, Z_{2i} + v_i$, where the elements of $Z_{2i}$ are generated as standard normal variables and $Z_{2i}$ has $m$ columns.
- $Z_{1i}$ is generated as a column vector (matrix with one column) where the elements are standard normal and $\gamma = 0.1$.
- The global dependence parameter $\lambda$ is varied on a grid from 0 to 0.9 with a step size of 0.1.
- The response is generated using the inversion $Y = (I - \lambda W)^{-1}(X\beta + Z_1\gamma + \epsilon)$, where $\epsilon$ is generated according to Eq. (19.3).

The results for the simulation in Tables 19.1 and 19.2 are the median risk of the estimators based on 1000 replications. The median risk is computed as the median of the scalar mean squared error, e.g., $\left(\hat{\beta}^{\mathrm{S}} - \beta\right)^\top \left(\hat{\beta}^{\mathrm{S}} - \beta\right)$ for the shrinkage estimator, which is computed for each of the 1000 simulation runs. For this configuration, the shrinkage estimator has the lowest median risk for all values of $\lambda$. Figure 19.1 shows the results graphically.

**Table 19.1** Empirical median risk (EMR) of $\hat{\beta}^{OLS}$, $\hat{\beta}^{TLS}$, and $\hat{\beta}^{S}$ with $n = 50$ and $\rho = 0.9$

| $\lambda$ | EMR($\hat{\beta}^{OLS}$) | EMR($\hat{\beta}^{TLS}$) | EMR($\hat{\beta}^{S}$) |
|---|---|---|---|
| 0.0 | 1.932 | 0.909 | 0.826 |
| 0.1 | 1.947 | 0.891 | 0.818 |
| 0.2 | 1.958 | 0.953 | 0.877 |
| 0.3 | 1.975 | 0.919 | 0.850 |
| 0.4 | 2.042 | 1.000 | 0.920 |
| 0.5 | 2.065 | 0.973 | 0.886 |
| 0.6 | 1.960 | 0.999 | 0.884 |
| 0.7 | 1.842 | 0.907 | 0.823 |
| 0.8 | 1.587 | 0.901 | 0.794 |
| 0.9 | 1.215 | 1.195 | 0.894 |

**Table 19.2** Empirical median risk (EMR) of $\hat{\beta}^{OLS}$, $\hat{\beta}^{TLS}$, and $\hat{\beta}^{S}$ with $n = 50$ and $\rho = 0.5$

| $\lambda$ | EMR($\hat{\beta}^{OLS}$) | EMR($\hat{\beta}^{TLS}$) | EMR($\hat{\beta}^{S}$) |
|---|---|---|---|
| 0.0 | 0.624 | 0.368 | 0.324 |
| 0.1 | 0.622 | 0.379 | 0.329 |
| 0.2 | 0.626 | 0.393 | 0.341 |
| 0.3 | 0.646 | 0.393 | 0.351 |
| 0.4 | 0.669 | 0.419 | 0.370 |
| 0.5 | 0.677 | 0.410 | 0.367 |
| 0.6 | 0.626 | 0.423 | 0.357 |
| 0.7 | 0.583 | 0.408 | 0.334 |
| 0.8 | 0.481 | 0.455 | 0.365 |
| 0.9 | 0.380 | 0.681 | 0.431 |

It is clear from the results in Table 19.1 that as the values of $\lambda$ increase in $[0, 0.5]$, the median risk of $\hat{\beta}^{OLS}$ increases and then decreases as the values of $\lambda$ increase in $(0.5, 0, 9]$. It is lowest when $\lambda = 0.9$. The shrinkage estimator always has a lower median risk than $\hat{\beta}^{OLS}$ and $\hat{\beta}^{TLS}$ for all values of $\lambda$. The estimator $\hat{\beta}^{TLS}$ always has a smaller risk than $\hat{\beta}^{OLS}$ for all values of $\lambda$.

The similar results based on Table 19.1 are graphically presented in Fig.19.1. The black (solid line), red (longer breaks), and green (shorter breaks) lines indicate the median risks of $\hat{\beta}^{OLS}$, $\hat{\beta}^{TLS}$, and $\hat{\beta}^{S}$, respectively. The line corresponding to $\hat{\beta}^{S}$ is lying in the lowest part indicating that $\hat{\beta}^{S}$ has the lowest median risk. The black (solid line) sharply declines after $\lambda \in (0.5, 1]$ indicating the behavior of the median risk of $\hat{\beta}^{OLS}$. The red (longer breaks) indicates that the median risks of $\hat{\beta}^{TLS}$ lies between the median risks of $\hat{\beta}^{OLS}$ and $\hat{\beta}^{S}$, which, in general, increases with the increase in the values of $\lambda$.

The results in Table 19.2 provide similar conclusions as in Table 19.1 about the empirical median risks of $\hat{\beta}^{OLS}$, $\hat{\beta}^{TLS}$, and $\hat{\beta}^{S}$. The median risk of $\hat{\beta}^{OLS}$ increases and then decreases for $\lambda$ in $[0, 0.5]$ and in $(0.5, 0, 9]$, respectively, and is lowest when $\lambda = 0.9$. The shrinkage estimator always has a lower median risk than $\hat{\beta}^{OLS}$ (except $\lambda = 0.9$) and $\hat{\beta}^{TLS}$ for all values of $\lambda$, whereas $\hat{\beta}^{TLS}$ always has a smaller risk than $\hat{\beta}^{OLS}$ for all values of $\lambda$ (except $\lambda = 0.9$).

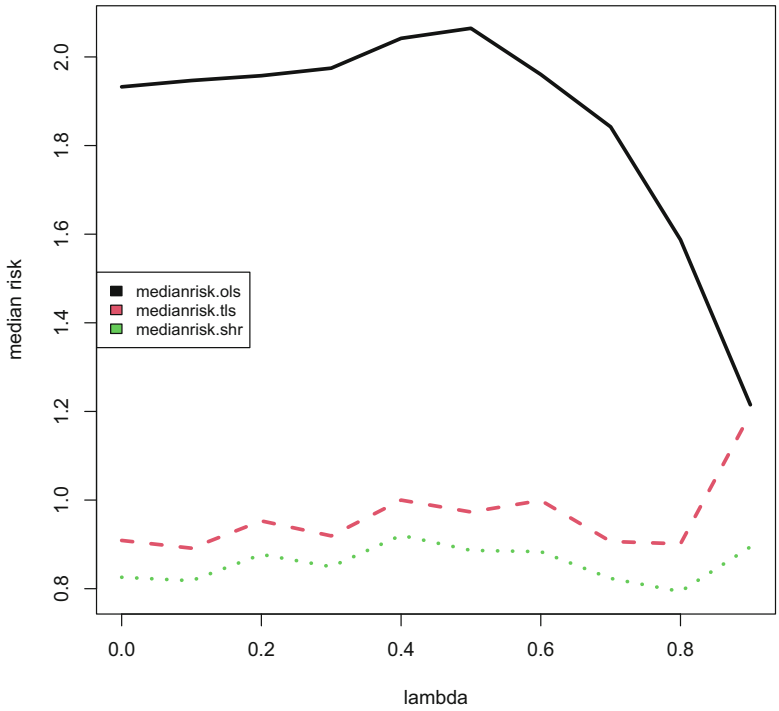



**Fig. 19.1** Empirical median risk of $\hat{\beta}^{\mathrm{OLS}}$, $\hat{\beta}^{\mathrm{TLS}}$, and $\hat{\beta}^{\mathrm{S}}$ with $n = 50$ and $\rho = 0.9$ with $\lambda$ on $x$-axis and median risks on $y$-axis

The results similar to those in Table 19.2 are graphically presented in Fig. 19.2 with the same annotations as in Fig. 19.1. The line corresponding to $\hat{\beta}^{\mathrm{S}}$ is lying in the lowest part, indicating that $\hat{\beta}^{\mathrm{S}}$ has the lowest median risk, but it is usually higher and more stable than in Fig. 19.1. The black (solid line) sharply declines after $\lambda \in (0.5, 1]$ indicating that behavior of the median risk of $\hat{\beta}^{\mathrm{OLS}}$, but the declination is steeper than in Fig. 19.1. It further crosses the median risks of $\hat{\beta}^{\mathrm{TLS}}$ and $\hat{\beta}^{\mathrm{S}}$ when approximately $\lambda \geq 0.8$. Such a behavior is not visible from Fig. 19.1. The red (longer breaks) indicates the median risks of $\hat{\beta}^{\mathrm{TLS}}$ have a sharper increase when approximately $\lambda \geq 0.7$ is more than the median risks of $\hat{\beta}^{\mathrm{OLS}}$ and $\hat{\beta}^{\mathrm{S}}$ when approximately $\lambda \geq 0.9$.

## 19.6 Conclusions

This chapter presents a sophisticated approach to addressing the challenges posed by endogenous regressors in spatial auto-regressive (SAR) models, and stemming from the concept of model averaging and Hansen's (Hansen, 2017) work, the

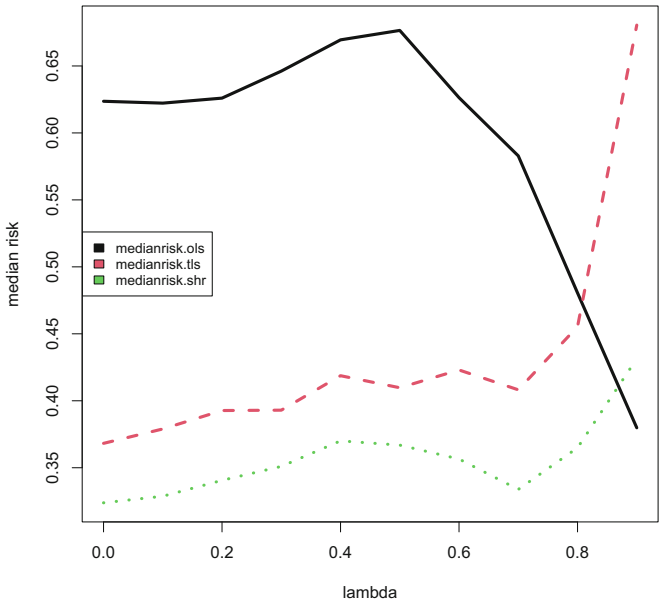



**Fig. 19.2** Empirical median risk of $\hat{\beta}^{\text{OLS}}$, $\hat{\beta}^{\text{TLS}}$, and $\hat{\beta}^{\text{S}}$ with $n = 50$ and $\rho = 0.5$ with $\lambda$ on $x$-axis and median risks on $y$-axis

chapter introduces a shrinkage estimation technique that blends OLS and 2SLS estimators. The method uses the Hausman-Wu statistic for exogeneity to determine the weighting scheme, with weights inversely proportional to the Hausman statistic. This Stein-like shrinkage estimator combines the strengths of both OLS and 2SLS, aiming to improve robustness and accuracy. The Monte Carlo simulations further validate the practical effectiveness of the shrinkage estimator and highlight that the proposed estimator reduces the finite sample median squared error compared to the standard 2SLS estimator, especially in small sample sizes.

The measurement error in covariates usually leads to endogeneity. The SAR model with measurement error in a subset of covariates may be considered as a special case of SAR model with endogenous covariates in Eq. (19.1).

Suppose we have two sets of covariates: $\tilde{X}_i$ $(i = 1, \ldots, m)$ which are subject to measurement error, and $Z_{1i}$ $(i = 1, \ldots, l)$ which are observable and exogenous. Further,

$$\tilde{X}_i = X_i + v_i \quad \text{and} \quad \text{cov}(v_i | X_i, \epsilon_i) = \Sigma.$$

We can obtain the shrinkage estimator and develop the asymptotic results along similar lines.

## References


Anselin, L. (1988). A test for spatial autocorrelation in seemingly unrelated regressions. *Economics Letters, 28*(2), 335–341.

Anselin, L., & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah (Ed.), *Handbook of applied economic statistics* (pp. 237–290). CRC Press.

Chaturvedi, A,, Gupta, S., & Bhatti, M. I. (2012). Confidence ellipsoids based on a general family of shrinkage estimators for a linear model with non-spherical disturbances. *Journal of Multivariate Analysis, 104*(1), 140–158.

Clarke, J. A. (2017). Model averaging OLS and 2SLS: An application of the WALS procedure. *Econometrics Working Papers* (Vol. 1701). Department of Economics, University of Victoria.

Danilov, D., & Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics, 122*, 27–46.

Luo, G., Wu, M., & Pang, Z. (2022). Estimation of spatial autoregressive models with covariate measurement errors. *Journal of Multivariate Analysis, 192*, 105093.

Hansen, B. E. (2017). Stein-like 2SLS estimator. *Econometric Reviews, 36*(6–9), 840–852.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251–1271.

Huan, L., Lee, T. H., & Ullah, A. (2017). A combined estimator of regression models with measurement errors. *Indian Economic Review, 52*(1), 73–91.

James, W., & C. Stein (1960). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Sciences and Probability*. University of California Press (pp. 361–379).

Knight, J. R., Hill, R. C., & Sirmans, C. F. (1992). Biased prediction of housing values. *Real Estate Economics, 20*, 427–456.

Knight, J. R., Hill, R. C., & Sirmans, C. F. (1993a). Estimation of hedonic housing price models using nonsample information: A Monte Carlo study. *Journal of Urban Economics, 34*(3), 319–346.

Knight, J. R., Hill, R. C., & Sirmans, C. F. (1993b). Stein-rule estimation in real estate appraisal. *Appraisal Journal, 15*(4), 539–544.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to spatial econometrics*. CRC Press.

Maasoumi, E. (1978). A modified stein-like estimator for the reduced form coefficients of simultaneous equations. *Econometrica, 46*(3), 695–703.

Magnus R., & Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica, 67*, 639–643.

Saleh, A. K. Md. E. (2006). *Theory of preliminary test and stein-type estimation with applications*. Wiley.

Sawa, T. (1973). Almost unbiased estimator in simultaneous equations systems. *International Economic Review, 14*(1), 97–106.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J. (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (pp. 197–206).

Ullah, A., & Srivastava, V. K., (1988). On the improved estimation of structural coefficients. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002), 50*(1), 111–118.

Zou, G., Wan, A. T. K., Wu, X., & Chen, T., (2007). Estimation of regression coefficients of interest when other regression coefficients are of no interest: The case of non-normal errors. *Statistics & Probability Letters, 77*(8), 803–810.

# Chapter 20
# Regularization of Robust Neural Networks: Bayesian Connections and Outlier Detection




**Jan Kalina**

## 20.1 Introduction

This work investigates recently proposed robust versions of neural networks and particularly their regularized versions, which are recommendable for regression models with a large number of parameters. Regularization brings benefits for models with a large number of parameters, and it is often used for neural networks to avoid overfitting (Poggio & Smale, 2003). $L_2$-regularization remains to be the most common regularization type used for shallow neural networks (Hastie et al., 2015). $L_2$-regularized neural networks, which are based on a loss function penalized by the sum of squared estimates of the parameters, were repeatedly found to be effective for small data (Piotrowski et al., 2020). While regularization is known to improve local robustness to small changes of the data, it cannot in general ensure a high resistance against severely outlying data points.

Bayesian reasoning represents an intuitive approach to considering conditional probabilities (Hooten & Hefley, 2019), and Bayesian estimation is especially useful for solving regression problems, e.g., for high-dimensional data (Greene, 2018). From the theoretical point of view, Bayesian estimates have been proven to dominate standard (e.g., maximum likelihood) estimates as overviewed in the monograph by Saleh et al. (2019) for a variety of statistical methods. Bayesian neural networks have also found many recent applications revealing their advantages compared to standard (non-Bayesian) neural networks (Goan & Fookes, 2020).

J. Kalina (✉)
The Czech Academy of Sciences, Institute of Computer Science, Prague, Czech Republic

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
e-mail: kalina@cs.cas.cz

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_20

Bayesian procedures are often exploited for evaluating and controlling uncertainty of measurements within metrology as the science about measurement (Cheng et al., 2018). Penalized (shrinkage) methods that can be naturally derived in a Bayesian setup are also intended to be incorporated in later editions of the Guide to the Expression of Uncertainty in Measurement (GUM) methodology in the next 5 years (Wübbeler et al., 2020). Still, the methods of Bayesian linear regression or Bayesian neural networks are often considered in practice as specific self-standing tools, without acknowledging their narrow connections to intuitive thinking and regularized (but non-Bayesian) estimation (Jospin et al., 2022).

Regularization may be recommended also for robust neural networks, which are based on a robust loss function (Kalina et al., 2024) aimed to yield resistant results also for data contaminated by outliers. Regularized versions of robust neural networks, which are based on ranks of residuals, were recently proposed in the papers of Gupta et al. (2020); Kalina et al. (2022, 2024), which did not however study the connections between Bayesian thinking and regularization.

Multilayer perceptrons are considered in this chapter as the most important class of shallow neural networks. First, regularized multilayer perceptrons are discussed in Sect. 20.2, and connections related to the Bayesian estimation of parameters are discussed. Then, the linear model is considered in Sect. 20.3 as a special case. The rest of the paper is devoted to robust regularized neural networks, which are resistant to the presence of outliers in the data. In Sect. 20.4, the Bayesian connections are extended to the robust methods, which are based on penalized versions of robust loss functions. For these robust networks, a procedure for outlier detection, which is based on a sequential instance selection, is proposed and illustrated on experiments in Sect. 20.5. Section 20.6 concludes the chapter.

## 20.2 Bayesian Interpretation of Regularized Multilayer Perceptrons

This section is devoted to Bayesian estimation of parameters of multilayer perceptrons, which leads to penalized estimators. The ideas developed here will be used in later sections.

### *20.2.1 The Bayes Theorem*

Let us recall the Bayes formula (Bayes theorem) for the continuous data in a schematic form for a very general setup. The likelihood denoted $L(\text{data}|\theta)$ depending on some $d$-dimensional parameter $\theta \in \mathbb{R}^d$ and the prior density $f(\theta)$ are assumed to be known. The formula gives the posterior density $f(\theta|\text{data})$ in the form

$$f(\theta|\text{data}) = \frac{L(\text{data}|\theta)\, f(\theta)}{\int L(\text{data}|\theta)\, f(\theta)}, \quad \theta \in \mathbb{R}^d \tag{20.1}$$

if this expression is well defined. Because the denominator represents a normalizing constant ensuring $\int f(\theta|\text{data}) = 1$, the evaluation of the posterior density and the subsequent estimation of $\theta$ require only to evaluate the numerator. Therefore, we use the notation $\propto$ to denote that the posterior density is proportional to the product of the likelihood and the prior, i.e., that the constant is neglected here. In this notation, we rewrite Eq. (20.1) as

$$f(\theta|\text{data}) \propto L(\text{data}|f(\theta))\, f(\theta), \quad \theta \in \mathbb{R}^d. \tag{20.2}$$

Let us remark that it is common in this context to denote all the densities by $f$ so that $f$ plays the role of different functions with different parameters.

### *20.2.2 Multilayer Perceptrons*

To introduce the nonlinear regression model, a continuous random variable $Y_i \in \mathbb{R}$ is considered as a response of the vector of regressors (independent variables) $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$, which are $p$-dimensional random vectors available for the total number of $n$ measurements. The nonlinear regression model is formally expressed as

$$Y_i = \varphi(X_i) + e_i, \quad i = 1, \ldots, n, \tag{20.3}$$

with an unknown function $\varphi$ and random errors $e_1, \ldots, e_n$, which are independent on the regressors. The matrix $X$ will denote the matrix of all the regressors with the $i$-th measurement $(X_{i1}, \ldots, X_{ip})^\top$ in the $i$-th row.

Let the function $\varphi$ be fully specified up to a $\theta = (\theta_1, \ldots, \theta_d)^\top \in \mathbb{R}^d$ and the aim of the training be reliable estimation of $\theta$. In contrary to the linear model of Sect. 20.3, the dimension $d$ of $\theta$ does not have to coincide with $p$. Let us consider the model of multilayer perceptrons (MLPs) as an example of such $\varphi$; however, we do not use any specific model, i.e., we may (but do not have to) consider the model of an MLP with a fixed architecture. To stress the dependence on $\theta$, we will denote the values $\varphi(X_i)$ as $\varphi(X_i, \theta)$ for $i = 1, \ldots, n$. Training the model is typically based on the optimization task

$$\min u_i^2(\hat{\theta}) \quad \text{over} \quad \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)^\top \in \mathbb{R}^d, \tag{20.4}$$

where $u_1(\hat{\theta}), \ldots, u_n(\hat{\theta})$ are residuals computed for a given estimate $\hat{\theta}$ of $\theta$, i.e.,

$$u_i(\hat{\theta}) = Y_i - \varphi(X_i, \hat{\theta}), \quad i = 1, \ldots, n. \tag{20.5}$$

Most commonly, no probabilistic assumptions on the random noise are needed for training MLPs. Our likelihood-based considerations of Sects. 20.2.4 and 20.2.3

however need the assumption that $e_i \sim \mathsf{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. Then, we have

$$Y_i \sim \mathsf{N}(\varphi(X_i, \theta), \sigma^2) \tag{20.6}$$

for $i = 1, \ldots, n$. Using the notation $\propto$ to denote proportionality, the likelihood is expressed (up to a normalizing constant) as

$$L(\text{data}|\theta, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \varphi(X_i, \theta))^2\right\}. \tag{20.7}$$

### *20.2.3 $L_1$-Regularization*

Assuming again the likelihood given by Eq. (20.7), let us formulate the prior for $\theta$ with the density

$$f(\theta) = \prod_{j=1}^{d} f(\theta_j) = \left(\frac{1}{2\tau}\right)^d \exp\left\{-\frac{1}{\tau}\sum_{j=1}^{d}|\theta_j|\right\}, \tag{20.8}$$

which means that $\theta_1, \ldots, \theta_d$ are independent random variables following Laplace distribution with variance $\text{var}(\theta_j) = 2\tau^2$ for $j = 1, \ldots, d$. The posterior density has the form

$$\begin{aligned} f(\theta|\text{data}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}u_i^2\right\}\exp\left\{-\frac{1}{\tau}\sum_{j=1}^{d}|\theta_j|\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}u_i^2 + \frac{2\sigma^2}{\tau}\sum_{j=1}^{d}|\theta_j|\right]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}u_i^2 + \lambda\sum_{j=1}^{d}|\theta_j|\right]\right\} \end{aligned} \tag{20.9}$$

with $\lambda = 2\sigma^2/\tau$. The Bayesian estimator of the parameters is obtained by maximizing $f(\theta|\text{data})$ in Eq. (20.9), which corresponds to finding parameters of $L_1$-regularized MLP denoted here as $L_1$-MLP.

$L_1$-regularized robust MLPs are based on the search for the most relevant information in terms of parameters and observations. The robust loss function ensures the selection of relevant observations, and the regularization ensures the selection of the relevant parameters. In the context of neural networks, however,

unlike linear regression, the practice of relying solely on relevant parameters and discarding redundant ones is generally not well-regarded (Hssayni et al., 2023). To explain this, pruning redundant parameters can harm the performance of neural network models by reducing its capacity to capture complex patterns and interactions in the data. Unlike linear models, neural networks rely on a large number of parameters to effectively learn intricate, nonlinear relationships (Kalina, 2013). Aggressive parameter reduction can limit generalization, especially in deep networks where even seemingly redundant parameters may contribute to stability and performance.

### 20.2.4 $L_2$-Regularization

We assume the prior information in the form $\theta \sim \mathsf{N}_d(0, \delta^2 \mathcal{I}_d)$, i.e., a $d$-variate Gaussian distribution, where $\delta^2 > 0$ and $\mathcal{I}_d$ denotes the unit matrix of size $d \times d$. Just like in Sect. 20.3.1, the assumption of a zero expectation is without loss of generality and without limiting the applicability of the results. This prior can be alternatively expressed as a product density of individual Gaussian densities corresponding to $\theta_j \sim \mathsf{N}(0, \delta^2)$ for every $j = 1, \ldots, d$, where $\theta_1, \ldots, \theta_d$ are mutually uncorrelated (and thus independent thanks to the normality) random variables. We stress that the prior information assumes not only normality of the errors but also homoscedasticity in the model. Using this prior information, the posterior density given by the Bayes formula has the form

$$
\begin{aligned}
f(\theta|\text{data}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i^2\right\} \exp\left\{-\frac{1}{2\delta^2}\sum_{j=1}^{d}\theta_j^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} u_i^2 + \frac{\sigma^2}{\delta^2}\sum_{j=1}^{d}\theta_j^2\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} u_i^2 + \lambda\sum_{j=1}^{d}\theta_j^2\right]\right\}, \quad \theta \in \mathbb{R}^d, \qquad (20.10)
\end{aligned}
$$

with the notation for the regularization parameter $\lambda = \sigma^2 / \delta^2$. The Bayesian estimator of $\theta$ is based on maximizing the posterior density, i.e., on solving the optimization task

$$
\min\left\{\sum_{i=1}^{n}\left(Y_i - \varphi(X_i, \hat{\theta})\right)^2 + \lambda\sum_{j=1}^{d}\hat{\theta}_j^2\right\} \qquad (20.11)
$$

over $\hat{\theta} \in \mathbb{R}^d$, and the network with the obtained estimate of $\theta$ is exactly a regularized MLP using the $L_2$-regularization.

**Remark 20.1** The value of $\lambda$ is inversely proportional to $\delta^2$, which means that increasing $\lambda$ corresponds to the prior belief about smaller variability of $\theta$, i.e., larger certainty (stronger prior belief) about $\theta$ being close to its expected value.

**Remark 20.2** If the prior $\theta \sim \mathsf{N}_d(\xi, \delta^2 \mathcal{I}_d)$ is assumed with an expectation $\xi \in \mathbb{R}^d$, which is possibly nonzero, the posterior density defined in Eq. (20.10) is replaced by

$$f(\theta|\text{data}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} u_i^2 + \lambda(\theta-\xi)^\top \sigma^{-2}\mathcal{I}(\theta-\xi)\right]\right\}, \quad \theta \in \mathbb{R}^d \tag{20.12}$$

and the optimization task given by Eq. (20.11) is replaced by

$$\min_{\hat{\theta}\in\mathbb{R}^d}\left\{\sum_{i=1}^{n} u_i^2 + \lambda \sum_{j=1}^{d}(\hat{\theta}_j - \xi_j)^2\right\}, \tag{20.13}$$

yielding a regularized version shrunken toward $\xi$.

### *20.2.5 General Remarks*

Considering the parameters $\theta \in \mathbb{R}^d$ to be a random vector is meaningful also with regard to the possibly different results of the training if performed repeatedly. The estimates of $\theta$ may (and often do) depend on the choice of initial weights, which are often chosen at random or on the optimization algorithm, where, for example the gradient descent method for optimization of the parameters does not ensure a global convergence (Shwetha et al., 2024).

The considerations of this section do not exploit the particular model $\varphi$, e.g., it is not necessary to use an MLP model. In an analogous manner, radial basis function (RBF) networks can be considered, for which however the regularization is typically considered for the weights and not for the data centers (Vang-Mata, 2020). The residuals of RBF networks again have the form (20.5). In fact, regularization can be obtained within a Bayesian approach for any model in nonlinear regression based on minimizing the sum of squares residuals, including, for example, the regularization networks described in Poggio and Smale (2003).

In this chapter, Bayesian reasoning is applied to standard (non-Bayesian, frequentist) neural networks to provide a probabilistic interpretation of the parameters and predictions. Such an approach treats the parameters as probability distributions rather than fixed values and allows it to quantify uncertainty in predictions. It needs to be stressed that all the tools of this chapter are strictly different from

Bayesian neural networks (BNNs). BNNs explicitly integrate Bayesian principles into their structure and training process, treat weights as distributions, and update these distributions based on observed data. This inherent integration allows BNNs to quantify uncertainty in a better (and much more complex) way and make more informed predictions compared to standard neural networks that apply Bayesian reasoning only as a supplemental explanation of the regularization principles.

## 20.3 Bayesian Estimation in Linear Regression

This section is devoted to Bayesian estimation of parameters of linear regression, which leads to penalized estimators, and of a generalized linear model. We will see that the ideas and results for linear regression can be understood as special cases of those formulated for the nonlinear model given by Eq. (20.3) of Sect. 20.2.

### *20.3.1 Bayesian Linear Regression: Basic Model*

We consider the standard linear regression model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n, \tag{20.14}$$

which may be expressed in the matrix notation as $Y = X\beta + e$. Here, we consider $p$ fixed regressors (predictors) available for $n$ measurements (observations). In our notation, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the vector of parameters, where the $i$-th row of $X$ will be denoted as $X_i = (X_{i1}, \ldots, X_{ip})^\top$ for $i = 1, \ldots, n$. Let us stress that the parameters are $p$-dimensional in the entire Sect. 20.3.1, because the number of parameters equals the number of predictors in the linear regression model.

Under the assumption $e_i \sim \mathsf{N}(0, \sigma^2)$, we have $Y_i \sim \mathsf{N}(X_i^\top \beta, \sigma^2)$ for $i = 1, \ldots, n$, and the likelihood can be expressed as the product of Gaussian (normal) densities

$$L(\text{data}|\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (Y_i - X_i^\top \beta)^2}{2\sigma^2}\right\}. \tag{20.15}$$

Let us consider the prior distribution about the parameter $\beta \in \mathbb{R}^p$ to be $p$-variate Gaussian $\beta \sim \mathsf{N}_p(\xi, \Psi)$ with a known positive definite matrix $\Psi$ and with $\xi \in \mathbb{R}^p$. Equation (20.1) requires the prior to be formulated by means of the joint density $f(\beta, \sigma^2)$, while $\sigma^2$ represents a nuisance parameter here. The prior in the form of the joint density is simply obtained as

$$f(\beta, \sigma^2) = f(\sigma^2) f(\beta|\sigma^2), \quad \beta \in \mathbb{R}^p, \quad \sigma^2 > 0, \tag{20.16}$$

under the usual assumption that $f(\sigma^2)$ is the density of the inverse gamma distribution (Hans, 2009). Here, the prior for $\beta$ does not actually depend on $\sigma^2$, and its density is obtained according to

$$f(\beta) = (2\pi)^{-p/2} (\det(\Psi))^{-1/2} \exp\left\{-\frac{1}{2}(\beta - \xi)^\top \Psi^{-1}(\beta - \xi)\right\}, \quad \beta \in \mathbb{R}^p. \tag{20.17}$$

The posterior distribution of $\beta$ conditioned on the given data and on $\sigma$ given by the Bayes theorem is the Gaussian distribution, and the Bayesian estimator of $\beta$ is given by the mean of the posterior distribution. This estimator

$$\hat{\beta} = \left(X^\top X + \sigma^2 \Psi^{-1}\right)^{-1} (X^\top Y + \sigma^2 \Psi^{-1} \xi) \tag{20.18}$$

represents a special case of the general form in Eq. (20.13), and we note that it depends on $\sigma$.

**Remark 20.3** In the special case with $\xi = 0$, the Bayesian estimator given by (20.18) is reduced to the commonly considered formula

$$\hat{\beta} = \left(X^\top X + \sigma^2 \Psi^{-1}\right)^{-1} (X^\top Y), \tag{20.19}$$

commonly denoted as the ridge regression estimator (Hastie et al., 2015).

**Remark 20.4** If $\xi = 0$ and if the prior information was obtained as $R$ previously performed measurements measured under the same conditions as the currently available data, it is now natural to choose $\Psi^{-1} = (X^\top X)^{-1} / R$, and the resulting $\hat{\beta}$ defined in Eq. (20.18) has the form

$$\hat{\beta} = \left(X^\top X + R X^\top X\right)^{-1} X^\top Y = \frac{1}{R+1} \hat{\beta}_{\mathrm{LS}}, \tag{20.20}$$

where $\hat{\beta}_{\mathrm{LS}}$ denotes the least squares estimator $\hat{\beta}_{\mathrm{LS}} = (X^\top X)^{-1} X^\top Y$. In this case, the variance of the Bayesian estimator defined in Eq. (20.20) is

$$\mathrm{var}(\hat{\beta}) = \frac{1}{(R+1)^2} \mathrm{var}(\hat{\beta}_{\mathrm{LS}}) = \frac{1}{(R+1)^2} \sigma^2 (X^\top X)^{-1}, \tag{20.21}$$

which is strictly smaller already for $R \geq 1$ compared to the variance of the least squares estimator

$$\mathrm{var}(\hat{\beta}_{\mathrm{LS}}) = \sigma^2 (X^\top X)^{-1}. \tag{20.22}$$

**Remark 20.5** If Jeffreys uninformative prior is used instead of $f(\beta)$ given by Eq. (20.17), the Bayesian estimator coincides with the least squares.

### 20.3.2 Bayesian Linear Regression: Generalized Model

Sometimes, the Bayesian linear regression considers a slightly more general model in the form

$$Y_i = (\phi(X_i))^\top \beta + e_i, \quad i = 1, \ldots, n, \tag{20.23}$$

with $\beta \in \mathbb{R}^p$, where however the response depends on $X_i$ through a known nonlinear function $\phi$. The notation

$$\phi(X_i) = \big(\phi(X_{i1}), \ldots, \phi(X_{ip})\big)^\top \in \mathbb{R}^p, \quad i = 1, \ldots, n, \tag{20.24}$$

is used here, and $\Phi$ is the matrix with elements

$$\Phi_{ij} = \phi(X_{ij}) \quad \text{for } i = 1, \ldots, n \quad \text{and} \quad j = 1, \ldots, p. \tag{20.25}$$

The model given by Eq. (20.23) is an extension of the model (20.14) replacing $X_i$ by $\phi(X_i)$ for $i = 1, \ldots, n$ so that the likelihood has the form

$$L(\text{data}|\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n [Y_i - (\phi(X_i))^\top \beta]^2}{2\sigma^2}\right\}. \tag{20.26}$$

Let us assume the prior information in the form $\beta \sim \mathsf{N}_p(0, \sigma^2\Psi^{-1})$ just like in Sect. 20.3.1. The posterior density of $\beta$ given the data is normal with expectation $\sigma^{-2}\Omega\Phi^\top Y$, where $\Omega = \left(\sigma^{-2}\Phi^\top\Phi + \Psi^{-1}\right)^{-1}$. The parameter $\beta$ can be expressed in the form

$$\hat{\beta} = \left(\Phi^\top\Phi + \lambda\Psi^{-1}\right)^{-1} \Phi^\top Y, \tag{20.27}$$

which is a direct extension of the estimator given by Eq. (20.18). The Bayesian linear regression turns out to be suitable also for models contaminated by measurement errors (Klauenberg et al., 2022).

## 20.4 Robust Regularized Multilayer Perceptrons

Because of high sensitivity of standard MLPs with respect to outliers, robust training is recommendable whenever it is desirable to perform resistant training under the presence of possible outliers in the data. The aim of this subsection is to recall available regularized versions of robust MLPs. The versions regularized in the $L_2$-norm may be derived in the Bayesian context as explained in Sect. 20.2. The regularized versions of robust MLPs overviewed in this section were proposed only

recently and therefore were not discussed in the monograph (Saleh et al., 2022) on rank-based methods in machine learning.

We return to the notation of Sect. 20.2, considering again the model (20.3) with parameters $\theta \in \mathbb{R}^d$. For all the robust networks of this section, the architecture is the same MLP architecture as in Sect. 20.2, and the residuals have the form (20.5). The training of the networks may be based on standard back-propagation algorithm but replacing the standard loss function by that given in (20.30). Some of the versions of the robust MLPs defined below are novel; the methods previously studied in numerical experiments yield promising results (Kalina et al., 2022, 2024).

The Bayesian explanation of the regularization, which was formulated in Sect. 20.2, is general and holds also for robust versions of MLPs. The likelihood given by Eq. (20.7) may be replaced by a robust version. In such cases, it represents a quasi-likelihood instead of the plain likelihood, and all the subsequent steps of the Bayesian reasoning are valid to obtain Eq. (20.11) as a regularized version.

The quantile regression MLPs represent a method for estimating quantiles of nonlinear trend inspired by Petneházi (2021). Here, QR-MLP($\tau$) denotes the quantile with $\tau \in (0, 1)$ defined by

$$\underset{\hat{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \zeta_\tau(u_i) + \lambda \sum_{j=1}^{d} \theta_j^2 \right\}, \tag{20.28}$$

i.e., an MLP with the loss function $\zeta_\tau$ with a given $\tau$. The so-called check function $\zeta_\tau$ is used, which is defined as

$$\zeta_\tau(x) = \begin{cases} (\tau - 1)x, & \text{if } x < 0, \\ \tau x, & \text{if } x \geq 0. \end{cases} \tag{20.29}$$

The following robust versions of MLPs are inspired by the least trimmed squares (LTS) and the least weighted squares (LWS) estimators of parameters in the linear regression model (Čížek, 2011; Víšek, 2011). The idea of trimming away or down-weighting the outliers may be applied also in connection with estimators based on the sum of the absolute values (rather than squares) of residuals. In this spirit, we may consider a version of MLP based on the least trimmed absolute (LTA) value estimator proposed already in Gentle (1977) but later almost forgotten. All these belong to methods based on ranks of residuals, which have been known to perform reliably also under the presence of measurement errors as investigated by Saleh et al. (2012). Particularly the LWS estimator in linear regression is able to combine efficiency, global robustness (with respect to outliers), and local robustness with respect to small changes of all data points. Let us refer to (Jurečková et al., 2019, p. 30) for a discussion of local and global concepts of statistical robustness.

$L_2$-LWS-MLP is defined as an MLP with parameters obtained by minimizing the loss

$$\min_{\hat{\theta}\in\mathbb{R}^d}\left\{\sum_{i=1}^{n}\psi\left(\frac{i-1/2}{n}\right)u^2_{(i)}(\hat{\theta})+\lambda\sum_{j=1}^{d}\hat{\theta}_j^2\right\} \tag{20.30}$$

with the notation $u^2_{(1)}(\hat{\theta})\le\cdots\le u^2_{(n)}(\hat{\theta})$, where $\psi$ is a given weight function. This is a nonincreasing function $\psi:[0,1]\to[0,1]$. If the loss of the LTS estimator (Rousseeuw & Van Driessen, 2006; Rusiecki, 2013) is used instead of the LWS, the optimization task solved within the training of the robust network denoted as $L_2$-LTS-MLP has the form

$$\min_{\hat{\theta}\in\mathbb{R}^d}\left\{\sum_{i=1}^{h}u^2_{(i)}(\hat{\theta})+\lambda\sum_{j=1}^{d}\hat{\theta}_j^2\right\} \tag{20.31}$$

for a chosen trimming constant $h\in[n/2,n)$. Such an MLP version is a special case of $L_2$-LWS-MLP in Eq. (20.30).

$L_2$-LTA-MLP is an MLP with parameters obtained by solving

$$\min_{\hat{\theta}\in\mathbb{R}^d}\left\{\sum_{i=1}^{h}|u(\hat{\theta})|_{(i)}+\lambda\sum_{j=1}^{d}\hat{\theta}_j^2\right\}, \tag{20.32}$$

where we use the notation

$$|u(\hat{\theta})|_{(1)}\le\cdots|u(\hat{\theta})|_{(n)}. \tag{20.33}$$

A novel extension allowing to down-weight the weights based on the LWA estimator denoted as $L_2$-LWA-MLP may be defined by solving

$$\min_{\hat{\theta}\in\mathbb{R}^d}\left\{\sum_{i=1}^{n}\psi\left(\frac{i-1/2}{n}\right)|u(\hat{\theta})|_{(i)}+\lambda\sum_{j=1}^{d}\hat{\theta}_j^2\right\}. \tag{20.34}$$

In the computations, we use LWS-based MLPs with the weight function

$$\psi(t)=\exp\left\{-\frac{t^2}{2\tau^2}\right\}\mathbb{1}[t<\alpha],\quad t\in[0,1],\quad \tau=0.8,\quad \alpha=0.75. \tag{20.35}$$

LTS-based MLPs are computed with $h=\lfloor n/2\rfloor+1$, where $\lfloor x\rfloor$ denotes the integer part of $x\in\mathbb{R}$. As an alternative to the robust MLPs given in this section, robust MLPs regularized in the $L_1$-norm may be considered. Also, it is possible to define $L_2$- or $L_1$-regularized versions of radial basis function (RBF) networks (Dey et al., 2019) or of other types of shallow neural networks. An overview of the different versions of MLPs of this section is presented in Table 20.1.

**Table 20.1** Overview of different versions of multilayer perceptrons of Sect. 20.4

| Method | Explanation | Loss function minimized over $\hat{\theta} \in \mathbb{R}^d$ |
|---|---|---|
| MLP | Standard training | $\min \sum_{i=1}^{n} u_i^2(\hat{\theta})$ |
| $L_2$-LTS-MLP | Least trimmed squares | $\min \left\{ \sum_{i=1}^{h} u_{(i)}^2(\hat{\theta}) + \lambda \sum_{j=1}^{d} \hat{\theta}_j^2 \right\}$ |
| $L_2$-LTA-MLP | Least trimmed absolute values | $\min \left\{ \sum_{i=1}^{h} \lvert u(\hat{\theta}) \rvert_{(i)} + \lambda \sum_{j=1}^{d} \hat{\theta}_j^2 \right\}$ |
| $L_2$-LWS-MLP | Least weighted squares | $\min \left\{ \sum_{i=1}^{n} \psi\left(\frac{i-1/2}{n}\right) u_{(i)}^2(\hat{\theta}) + \lambda \sum_{j=1}^{d} \hat{\theta}_j^2 \right\}$ |
| $L_2$-LWA-MLP | Least weighted absolute values | $\min \left\{ \sum_{i=1}^{n} \psi\left(\frac{i-1/2}{n}\right) \lvert u(\hat{\theta}) \rvert_{(i)} + \lambda \sum_{j=1}^{d} \hat{\theta}_j^2 \right\}$ |

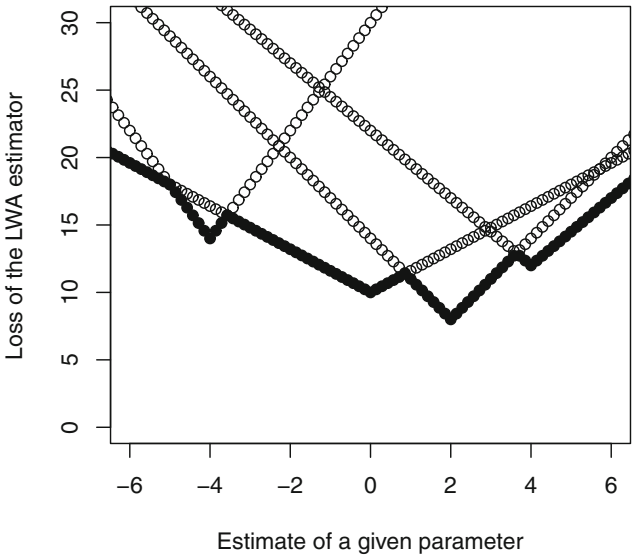



**Fig. 20.1** The loss function of the LWA estimator in a linear regression example is obtained as the minimum across all the individual weighted least absolute deviation estimators

Let us accompany the formal description of the robust training by a numerical illustration. For the sake of simplicity, we consider here the linear regression model given by Eq. (20.14) and a dataset from Kalina (2015), where domestic investments in the United States are modeled as a response of a single predictor, namely, the real gross domestic product. The model contains an intercept $\beta_0$ and slope $\beta_1$ related to the predictor. We consider the LWA estimator and the weight function defined in Eq. (20.35). For a given permutation of the weights, the loss of the LWA estimator is one of the V-shaped trends in Fig. 20.1. The estimator considers all

possible permutations of the weights. Different V-shaped loss curves shown in the figure correspond to different permutations of the weights. The estimator takes the permutation leading to the minimal loss. This is the minimum across the different V-shaped curves, and this is shown in the figure as the bold curve shown across different estimates $\hat{\beta}_1$ of $\beta_1$.

## 20.5 Outlier Detection for Robust Regularized Multilayer Perceptrons

The LTS-based robust training of MLPs described in Sect. 20.4 was proposed in Rusiecki (2013). LTS-MLP and particularly $L_2$-LTS-MLP turned out to yield successful results in numerical experiments in Kalina et al. (2022, 2024) in terms of cross-validation mean squared error in various regression tasks. Still, there remains the practical question how to find the suitable value of the trimming constant $h$ for $L_2$-LTS-MLP (and also for LTS-MLP, which is only its special case). In this section, a forward-search approach is proposed for the task, which is based on a sequence instance selection, i.e., on a sequential search of the most reliable subset of observations, which is tailor-made for rank-based models based on the LTS or LWS.

The Forward Search (FS) is a versatile methodology for outlier detection previously elaborated in a number of papers (Atkinson et al., 2010; Riani et al., 2015; Cerioli et al., 2018) for regression or classification tasks. While methods of sequential statistics are most often connected with hypothesis testing (Mukhopadhyay, 2021), they are only rarely exploited for estimation in the regression context (Lüdtke et al., 2020), and the Forward Search is one of rare sequential tools for robust regression estimation. It is a technique for practical regression modeling for contaminated data motivated to accompany robust estimation (Riani et al., 2014). Although it is a computational technique, the Forward Search has been justified by interesting theoretical results (Atkinson et al., 2010; Cerioli et al., 2014). The main idea is to repeatedly compute S- or MM-estimation while sequentially increasing their breakdown point, i.e., the set of relevant data points is sequentially increasing. Formally, the breakdown point as a fundamental measure of robustness is defined as the minimal fraction of data that can drive an estimator beyond all bounds when set to arbitrary values (Jurečková et al., 2019). LTS-MLP with the trimming constant $h$ achieves the breakdown point equal to $1 - h/n$. So far, the idea of the Forward Search has not been exploited within the context of neural networks.

Let us propose the novel approach for the optimal selection of $h$ for $L_2$-LTS-MLP, which is denoted as the sequential instance selection based on the Forward Search (FS-SIS). The procedure is based on estimates of the variance $\sigma^2$ of the random errors in the regression model (20.3) obtained within Algorithm 20.1 for $L_2$-LTS-MLP models with different values of $h$ starting from $n$ and going down to $\lfloor n/2 \rfloor$. The variance $\sigma^2$ is estimated only from the "good" (non-discarded) data in

**Algorithm 20.1** FS-SIS: Sequential instance selection based on the Forward Search for $L_2$-LTS-MLP

**Input:** Response $Y_1, \ldots, Y_n$, features $X_{i1}, \ldots, X_{ip}$ for $i = 1, \ldots, n$

**Input:** Hyperparameters selected by the user: number of hidden layers and number of their neurons

**Output:** Optimal value $h^*$ of the trimming constant $h$

Use cross-validation to find $\lambda^*$ as the optimal $\lambda$ for $L_2$-LTS-MLP for $h = \lfloor n/2 \rfloor$

**for** $h = n$ downto $\lfloor n/2 \rfloor$ **do**

Train $L_2$-LTS-MLP with $\lambda^*$ for the given trimming constant $h$

$u = (u_1, \ldots, u_n)^\top :=$ residuals of $L_2$-LTS-MLP

$$\hat{\sigma}_h^2 := \frac{1}{h} \sum_{i=1}^{n} u_i^2 \mathbb{1}[u_i^2 \leq u_{(h)}^2] \tag{20.36}$$

**end for**

Subjectively find $h^*$ based on visualizations of the values $\hat{\sigma}_n^2, \hat{\sigma}_{n-1}^2, \ldots, \hat{\sigma}_{\lfloor n/2 \rfloor}^2$

each step of the algorithm, and the notation $\hat{\sigma}_h^2$ is used for an estimate based on exactly $h$ observations. The regularization parameter $\lambda$ is determined only at the beginning for the sake of saving computational burden; this should be sufficient, because the search for a suitable $\lambda$ should not be influenced by the presence of outliers. Also, the architecture should be determined without being influenced by the outliers.

Let us describe finding the suitable value of $h$ based on the outputs of Algorithm 20.1; the obtained value of $h$ will be denoted as $h^*$. Finding $h^*$ based on the graph of $\hat{\sigma}_{\lfloor n/2 \rfloor}^2$ in dependence on the breakdown point is subjective and can be only explained on illustrative numerical examples. The decision criterion is based on finding an abrupt change in the graph in an analogy to the reasoning of Riani et al. (2014) based however on different graphs.

Figure 20.3 shows typical monitoring plots allowing to find the optimal value $h^*$ very clearly even in spite of the subjectivity of the approach. Let us stress that for a low breakdown point, the estimator of $\sigma^2$ is highly influenced by outliers. On the other hand, for a high breakdown point, the estimator is based (mainly) on the "good" (reliable) data points. Finally, the point corresponding to the true contamination level is exactly the *smallest* breakdown point, where the estimate of $\sigma^2$ is not influenced by the presence of outliers. Such an approach turns out to be much more reliable compared to attempts to train $L_2$-LTS-MLP just once and to estimate values of $\sigma_h^2$ from the single training.

The proposed approach can also be used for other types of networks or for other types of regularization. If LWS-based versions of MLPs are used, it is necessary to specify the weight function; in such a case, however, we need to modify the given weight function $\psi$ within Algorithm 20.1 to trim away the given percentage of observations according to

$$\psi^*(t) = \begin{cases} \psi(t), & t \in \left[0, \ \frac{h-1/2}{n}\right), \\ 0, & t \in \left[\frac{h-1/2}{n}, \ 1\right]. \end{cases} \tag{20.37}$$

In the experiments, we use Eq. (20.37) with a linearly decreasing weight function $\psi$; such an approach has the breakdown point equal to $(n - h)/n$ exactly. As an alternative approach for LWS-based versions of MLPs, it would be possible to consider a robust estimate of $\sigma^2$, e.g., the LWS-based estimate proposed in Víšek (2010) for the linear model, which is meaningful for the nonlinear setup as well.

### 20.5.1 Experiment: Artificial Data with $p = 1$

The aim of the experiments is to illustrate the performance of the proposed FS-SIS approach for finding the suitable $h$ for $L_2$-LTS-MLP for real data. We start with an experiment with only $p = 1$, because it allows a visualization of the estimated nonlinear trend. We used the TensorFlow library (Abadi et al., 2015) of Python for the computations and R software for the visualization of the results.

The dataset with $n = 101$ observations and a single regressor ($p = 1$) is shown in Fig. 20.2. It was constructed as a mixture of 81 observations generated by a given nonlinear (but homoscedastic) model and 20 outliers with the value of the regressor exceeding 3. We use various versions of MLPs always with *one* hidden layer and *eight* neurons. We use a sigmoid activation function in the hidden layer and a linear output layer. Estimates of $\sigma^2$ depending on the breakdown point of $L_2$-LTS-MLP, as obtained in Algorithm 20.1, are shown in Fig. 20.3 (left). There is a clear change of the trend around the value 0.2 corresponding to 20% of outliers. Thus, we can estimate that the suitable $h$ corresponds to 20% of outliers in the data, i.e., that

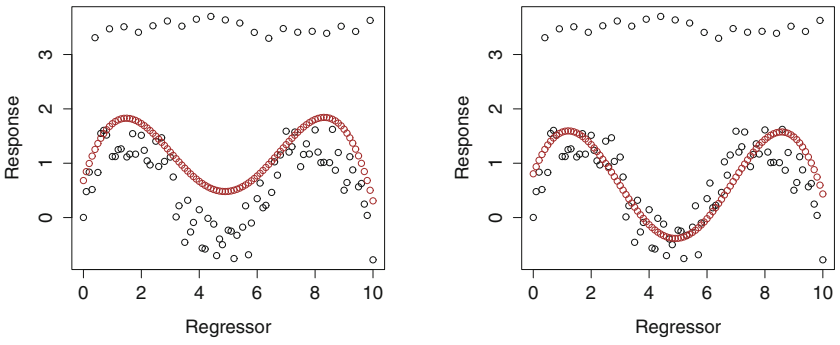



**Fig. 20.2** Results for the artificial dataset of Sect. 20.5.1. Left: results of $L_2$-MLP. Right: results of $L_2$-LTS-MLP with $h^* = 0.8n$, which was determined as the optimal value by FS-SIS

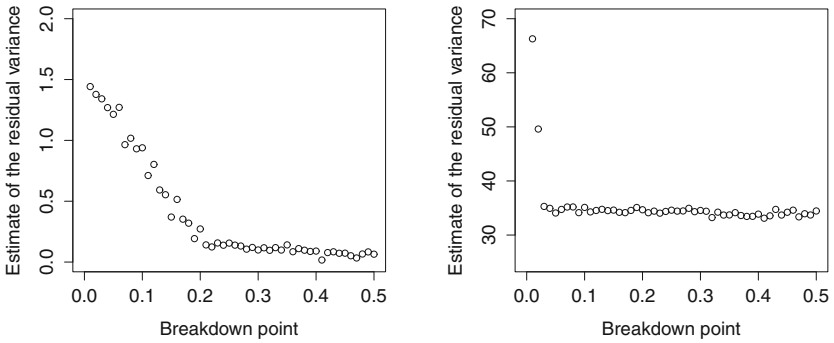



**Fig. 20.3** Estimates of $\sigma^2$ for $L_2$-LTS-MLP in dependence on the breakdown point. Left: artificial dataset of Sect. 20.5.1. Right: auto-MPG dataset of Sect. 20.5.2. The figures suggest to take the trimming constants $h^* = 0.2$ (left) and $h^* = 0.02$ (right)

$h = 0.8n$, which actually corresponds to the true contamination level in the data. The trend in Fig. 20.3 is not monotone; the variability of the values violating the monotonicity can be explained by random results of training LTS-MLP using the back-propagation algorithm.

Let us now consider $L_2$-LTS-MLP with the trimming constant $h = 0.8n$ corresponding to the estimated level of 20% of outliers. Figure 20.2 compares $L_2$-MLP (left) with $L_2$-LTS-MLP with $h = 0.8n$ (right) for the given dataset. Results of the mean squared error (MSE) and trimmed mean squared error (TMSE) evaluated in 5-fold cross-validation are presented in Table 20.2 for various versions of MLPs.

To discuss the results of Table 20.2 for the artificial dataset, we can see that all the robust MLPs outperform the standard MLP in terms of TMSE. The regularization brings an improvement of TMSE in most cases. This improvement is however only minor for $L_2$-regularization and almost negligible for $L_1$-regularization. Using $h^*$ as the trimming constant generally yields much better results than taking $h = \lfloor n/2 \rfloor$. The best result with TMSE equal to 0.16 is achieved for $L_2$-LWS-MLP with $h^*$ compared to the standard MLP with TMSE equal to 0.36. Values of MSE are heavily influenced by outliers, and we do not consider them relevant for comparison of the methods.

### *20.5.2 Experiments: A Real Dataset*

Further, we analyze the Auto-MPG dataset as a real dataset coming from the UCI public repository (Dua & Graff, 2019). We consider the consumption of cars in miles per gallon (MPG) as response, explained by four continuous features, which are technical specifications of cars: displacement, horsepower, weight, and acceleration. As we omitted the whole observations with some missing values (i.e., observations with index 33, 127, 331, 337, 355, and 375), we work with $n = 392$ and $p = 4$.

**Table 20.2** Results for the artificial dataset (Sect. 20.5.1) and for the Auto-MPG dataset (Sect. 20.5.2) evaluated in a fivefold cross-validation. The notation $h^*$ is used for the optimal $h$ found by the FS-SIS method

| Method | $h$ | Artificial MSE | Artificial TMSE | Auto-MPG MSE | Auto-MPG TMSE |
|---|---|---|---|---|---|
| MLP | – | 1.44 | 0.36 | 61.4 | 28.4 |
| LTS-MLP | $\lfloor n/2 \rfloor$ | 2.02 | 0.24 | 80.8 | 17.9 |
| LWS-MLP | $\lfloor n/2 \rfloor$ | 1.96 | 0.25 | 78.3 | 15.8 |
| LTA-MLP | $\lfloor n/2 \rfloor$ | 1.99 | 0.26 | 77.1 | 18.4 |
| LTS-MLP | $h^*$ | 1.93 | 0.21 | 77.6 | 16.2 |
| LWS-MLP | $h^*$ | 1.85 | 0.20 | 75.4 | 15.7 |
| LTA-MLP | $h^*$ | 1.87 | 0.23 | 78.0 | 16.6 |
| LWA-MLP | $h^*$ | 1.90 | 0.22 | 77.3 | 16.0 |
| $L_2$-LTS-MLP | $\lfloor n/2 \rfloor$ | 1.79 | 0.20 | 69.7 | 15.0 |
| $L_2$-LWS-MLP | $\lfloor n/2 \rfloor$ | 1.76 | 0.19 | 73.5 | 16.1 |
| $L_2$-LTA-MLP | $\lfloor n/2 \rfloor$ | 1.82 | 0.21 | 70.2 | 17.2 |
| $L_2$-LWA-MLP | $\lfloor n/2 \rfloor$ | 1.79 | 0.20 | 72.4 | 16.9 |
| $L_2$-LTS-MLP | $h^*$ | 1.82 | 0.17 | 67.2 | 14.1 |
| $L_2$-LWS-MLP | $h^*$ | 1.81 | 0.16 | 71.2 | 13.8 |
| $L_2$-LTA-MLP | $h^*$ | 1.85 | 0.19 | 72.7 | 14.7 |
| $L_2$-LWA-MLP | $h^*$ | 1.85 | 0.19 | 71.7 | 14.3 |
| $L_1$-LTS-MLP | $\lfloor n/2 \rfloor$ | 1.98 | 0.22 | 67.9 | 16.8 |
| $L_1$-LWS-MLP | $\lfloor n/2 \rfloor$ | 2.04 | 0.22 | 71.4 | 16.5 |
| $L_1$-LTA-MLP | $\lfloor n/2 \rfloor$ | 1.96 | 0.24 | 72.5 | 17.1 |
| $L_1$-LWA-MLP | $\lfloor n/2 \rfloor$ | 1.93 | 0.23 | 70.8 | 16.9 |
| $L_1$-LTS-MLP | $h^*$ | 1.83 | 0.19 | 65.0 | 16.3 |
| $L_1$-LWS-MLP | $h^*$ | 1.88 | 0.19 | 69.1 | 15.0 |
| $L_1$-LTA-MLP | $h^*$ | 1.91 | 0.22 | 71.3 | 16.4 |
| $L_1$-LWA-MLP | $h^*$ | 1.93 | 0.21 | 70.9 | 16.0 |

There are no missing values in the dataset, and an exploratory data analysis reveals that there is no problem with heteroscedasticity in the regression model.

Our aim is to find a suitable $h$ for the $L_2$-LTS-MLP model, which is used with 2 hidden layers with 16 and 6 neurons. We use a sigmoid activation function in every hidden layer and a linear output layer. Estimates of $\sigma^2$ depending on the breakdown point of the $L_2$-LTS-MLP are shown in Fig. 20.3 (right) as obtained using Algorithm 20.1. Based on the figure, we can estimate that the suitable $h$ corresponds to 2% of severe outliers in the data, i.e., we obtain $h = 0.98n$. This percentage corresponds to visualizations of the data (not shown here) as well as to the results presented in the literature for this popular benchmarking dataset (Tsai et al., 2023). Results of the MSE and TMSE evaluated in fivefold cross-validation are presented in Table 20.2 for various versions of MLPs.

The results of Table 20.2 reveal for the Auto-MPG dataset that all the robust MLPs outperform the standard MLP in terms of TMSE. The benefit of the regu-

larization is more remarkable here compared to Sect. 20.5.1, because we analyze a larger dataset now. Here, $L_2$-regularization leads to smaller values of TMSE than $L_1$-regularization in most situations. We can see that $L_1$-regularization would actually be better than $L_2$-regularization if considering MSE; we do not however reflect the results with MSE, which is non-robust. Using $h^*$ as the trimming constant generally yields much better results than taking $h = \lfloor n/2 \rfloor$. The best result with TMSE equal to 13.8 is achieved for $L_2$-LWS-MLP with $h^*$ compared to the standard MLP with TMSE equal to 28.4, where the standard MLP is apparently highly influenced by outliers in the dataset. LWS-based robust training are more effective compared to the LTS-based ones, and LWA-based methods are superior to LTA-based.

## 20.6 Conclusion

This chapter deals with multilayer perceptrons, which represent the most important class of shallow neural networks, but analogous ideas can be formulated also for radial basis function (RBF) networks (Kalina et al., 2024) or other types of nonlinear regression models. Bayesian estimation is a common methodology for linear regression, for which we have to keep in mind the assumptions on normality and homoscedasticity, and Bayesian thinking allows to incorporate previously performed measurements in a very intuitive way. The Bayesian reasoning is extended here to training of robust MLPs, where the training is expressed as a maximization of certain quasi-likelihood functions, i.e., as an optimization task very analogous to likelihood-based statistical estimation problems. The presented considerations are valid even in the situation that no explicit formulas for the robust estimates of $\theta$ are available.

The Bayesian reasoning is applicable in metrological scenarios either with available prior knowledge or under errors in the estimation computations (e.g., approximate computing in a low-precision arithmetic). On the other hand, the Bayesian methodology is known for its large skepticism against hypothesis testing about the parameters, as admitted in Section 16.4.3 of Greene (2018). Future theoretical research should extend the considerations to deep learning or to develop (non-Bayesian) likelihood-based hypothesis testing for neural networks.

The novel sequential instance selection (FS-SIS) approach of Sect. 20.5, which is based on a graphical visualization with a clear interpretation, turns out to yield meaningful results for the considered datasets containing outliers. The method based on evaluating estimates of $\sigma^2$ for subsets of data is feasible also for regularized models for high-dimensional data. Robust MLPs based on the LWS are more flexible compared to those based on the LTS, but the former require the user to choose a specific weight function. The search for the optimal trimming constant $h^*$ turns out to improve the prediction ability in the experiments; it allows a reasonable trade-off between high robustness (for contaminated data) and high efficiency (for non-contaminated data). Alternative possibilities could consider estimates of

selected parameters of the MLPs or correlations of adjacent sets of residuals according to Riani et al. (2014) or MLPs based on robust multivariate estimates of the covariance matrix (Rusiecki, 2008; Kalina, 2012). Finding the suitable trimming constant is a problem encountered but not sufficiently investigated also, for example, in robust regularized regression (Alfons et al., 2013) or robust regularized estimation of the covariance matrix (Boudt et al., 2020).

FS-SIS deserves to be investigated in experiments also in other contexts, such as in linear regression or covariance matrix estimation (Boudt et al., 2020). While heteroscedasticity in the model given by Eq. (20.3) may represent an obstacle for the proposed approach, we have to note that homoscedasticity is always a key requirement for robust training of neural networks as such. Also, the Forward Search for linear regression has never been examined on heteroscedastic data, so there is a gap of knowledge regarding the performance of the Forward Search for heteroscedastic data even for the simplest linear model. For heteroscedastic models, however, it seems better to consider regularized regression quantiles based on neural networks (rather than versions of LTS-MLP or LWS-MLP) as proposed in Kalina et al. (2024).

**Acknowledgments** The research was supported by the project 24-11146S ("Maximum entropy portfolio") of the Czech Science Foundation. The author would like to thank Petra Vidnerová for discussion.

## References

Abadi, M., et al. (2015) TensorFlow: A system for large-scale machine learning. http://tensorflow.org

Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics, 7*, 222–248.

Atkinson, A. C., Riani, M., & Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society, 39*, 117–134.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing, 30*, 113–128.

Cerioli, A., Farcomeni, A., & Riani, M. (2014). Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter. *Journal of Multivariate Analysis, 126*, 167–183.

Cerioli, A., Riani, M., Atkinson, A. C., & Corbellini, A. (2018). The power of monitoring: How to make the most of a contaminated multivariate sample. *Statistical Methods & Applications, 27*, 559–587.

Cheng, Y. B., Chen, X. H., Li, H. L., Cheng, Z. Y., Jiang, R., Lü, J., & Fu, H. D. (2018). Analysis and comparison of Bayesian methods for measurement uncertainty evaluation. *Mathematical Problems in Engineering, 2018*, 7509046.

Čížek, P. (2011). Semiparametrically weighted robust estimation of regression models. *Computational Statistics & Data Analysis, 55*, 774–788.

Dey, P., Gopal, M., Pradhan, P., & Pal, T. (2019). On robustness of radial basis function network with input perturbation. *Neural Computing and Applications, 31*, 523–537.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Gentle, J. E. (1977). Least absolute values estimation: An introduction. *Communications in Statistics—Simulation and Computation, 6*, 313–328.

Goan, E., & Fookes, C. (2020). Bayesian neural networks: An introduction and survey. *Lecture Notes in Mathematics, 2259*, 45–87.

Greene, W. H. (2018). Econometric analysis (8th ed.) Pearson Education Limited.

Gupta, D., Hazarika, B. B., & Berlin, M. (2020). Robust regularized extreme learning machine with asymmetric Huber loss function. *Neural Computing and Applications, 32*, 12971–12998.

Hans, C. (2009). Bayesian lasso regression. *Biometrika, 96*, 835–845.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity. The lasso and generalizations*. CRC Press.

Hooten, M. B., & Hefley, T. (2019). Bringing Bayesian models to life. CRC Press.

Hssayni, E. H., Joudar, N. E., & Ettaouil, M. (2023). Localization and reduction of redundancy in CNN using $L1$-sparsity induction. *Journal of Ambient Intelligence and Humanized Computing, 14*, 13715–13727.

Jospin, L.V., Laga, H., Boussaid, F., Buntine, W., & Bennamoun, M. (2022). Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Computational Intelligence Magazine, 17*, 29–48.

Jurečková, J., Picek, J., & Schindler, M. (2019). *Robust statistical methods with R* (2nd ed.). CRC Press.

Kalina, J. (2012). Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering, 32*(2), 3–16.

Kalina, J. (2013). Highly robust methods in data mining. *Serbian Journal of Management, 8*, 9–24.

Kalina, J. (2015). Three contributions to robust regression diagnostics. *Journal of Applied Mathematics, Statistics and Informatics, 11*(2), 69–78.

Kalina, J., Tumpach, J., & Holeňa, M. (2022). On combining robustness and regularization in training multilayer perceptrons over small data. *Proceedings IJCNN 2022* (pp. 1–8).

Kalina, J., Vidnerová, P., & Janáček, P. (2024). Highly robust training of regularized radial basis function networks. *Kybernetika, 60*, 38–59.

Klauenberg, K., Martens, S., Bošnjaković, A., Cox, M.G., van der Veen, A. M. H., & Elster, C. (2022). The GUM perspective on straight-line errors-in-variables regression. *Measurement, 187*, 110340.

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods, 25*, 157–181.

Mukhopadhyay, N. (2021). Purely sequential point estimation of a function of the mean in an exponential distribution. *Journal of Statistical Theory and Practice, 15*, 56.

Petneházi, G. (2021). Quantile convolutional neural networks for value at risk forecasting. *Machine Learning with Applications, 6*, 100096.

Piotrowski, A. P., Napiorkowski, J. J., & Piotrowska, A. E. (2020). Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling. *Earth-Science Reviews, 201*, 103076.

Poggio, T., & Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS), 50*, 537–544.

Riani, M., Cerioli, A., Atkinson, A. C., & Perrotta, D. (2014). Monitoring robust regression. *Electronic Journal of Statistics, 8*, 646–677.

Riani, M., Perrotta, D., & Cerioli, A. (2015). The forward search for very large datasets. *Journal of Statistical Software, 67*, Code Snippet 1.

Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery, 12*, 29–45.

Rusiecki, A. (2008). Robust MCD-based backpropagation learning algorithm. *Lecture Notes in Artificial Intelligence, 5097*, 154–163.

Rusiecki, A. (2013). Robust learning algorithm based on LTA estimator. *Neurocomputing, 120*, 624–632.

Saleh, A. K. M. E., Arashi, M., & Kibria, B. M. G. (2019). Theory of ridge regression estimation with applications. Wiley.

Saleh, A. K. M. E., Picek, J., & Kalina, J. (2012). R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika, 75*, 311–328.

Saleh, A. K. M. E., Arashi, M., Saleh, R. A., & Norouzirad, M. (2022). Rank-based methods for shrinkage and selection with application to machine learning. Wiley.

Shwetha, N., Priyatham, M., & Gangadhar, N. (2024). Artificial neural network based channel equalization using battle royale optimization algorithm with different initialization strategies. *Multimedia Tools and Applications, 83*, 15565–15590.

Tsai, S. H., Wu, C. Y., Chen, Y. H. (2023). A novel type-2 fuzzy identification method based on the improved membership function. *International Journal of Fuzzy Systems, 25*, 1818–1833.

Vang-Mata, R. (2020). Multilayer perceptrons: Theory and applications. Nova Science Publishers.

Víšek, J. Á. (2010). Robust error-term-scale estimate. *IMS Collections, 7*, 254–267.

Víšek, J. Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika, 47*, 179–206.

Wübbeler, G., Marschall, M., & Elster, C. (2020). A simple method for Bayesian uncertainty evaluation in linear models. *Metrologia, 57*, 065010.

# Chapter 21
# Estimating Finite Mixture Models Using Component Self-Paced Learning




**Andre R. Kleynhans, Frans H. J. Kanfer, and Salomon M. Millard**

## 21.1 Introduction

The finite mixture model (FMM) is tractable (Zhang et al., 2018; Peel & McLachlan, 2000a) and is widely used in unsupervised learning tasks, such as modeling density functions, while also being applicable in supervised learning scenarios, including mixtures of regressions and combinations of expert classifiers. In this chapter, we will focus on modeling density functions. Generally, FMMs are fitted by finding the maximum likelihood estimate of the parameters using the expectation maximization (EM) algorithm. The EM algorithm guarantees convergence to some optima and can result in sub-optimal parameter estimates when the algorithm does not converge to the global optima (Bishop, 2006; Dempster et al., 1977; McLachlan & Peel, 1998; Hipp & Bauer, 2006; Shireman et al., 2017).

The accuracy of the parameter estimates can also be distorted by both mild and gross outliers. Mild outliers are non-typical observations generated from a different distribution or far from the presumed distribution, whereas gross outliers are unpredictable observations not modeled by any distribution (Punzo & McNicholas, 2016; McLachlan & Peel, 1998). We refer to such observations as non-typical. There are many approaches to reduce the impact of non-typical observations. One approach is using a mixture of heavier-tailed distributions, for example, the Student's noncentral multivariate $t$ mixture model (TMM) (Peel & McLachlan, 2000b). Another method is constructing a distributional framework for non-typical observations by adding additional components for non-typical data (Banfield &

A. R. Kleynhans (✉) · F. H. J. Kanfer · S. M. Millard
Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa
e-mail: kleynhans.andre@up.ac.za; frans.kanfer@up.ac.za; sollie.millard@up.ac.za

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_21

Raftery, 1993; Henning, 2004). The mixture of contaminated normals (CNM) is an example of this approach (Punzo & McNicholas, 2016).

Self-paced learning (SPL) is a supervised learning method that introduces observations into the training process in a systematic sequence (Li et al., 2022; Kumar et al., 2010). This results in a method that can mitigate the presence of outliers (Meng et al., 2017). Zhang et al. (2018) have shown that a TMM fitted using the SPL algorithm can be more robust than fitted using the EM algorithm. This study examined one contamination scenario, a data set with 5% random non-typical observations. No indications of bias and mean squared error (MSE) are given. Furthermore, no verification is provided on the accuracy of the correct number of clusters and the correct classification of the observations.

Among other, we aim to bridge this gap by using a simulation study to test the efficacy of the SPL algorithm for different contamination scenarios. The simulation study showed the SPL algorithm provides some resilience to the impact of non-typical data but is prone to more degenerate solutions than compared to the EM algorithm. Degenerate solution refers to a parameter estimate that is not meaningful or theoretically possible, such as singular covariance matrices (Biernacki & Chrétien, 2003). We found that the rule of observation selection for the SPL algorithm in the context of FMMs is at the root of this problem.

We propose an improved SPL algorithm for estimating FMMs that determines the order in which observations are introduced at component level (SPCL). The simulation study shows that when fitting a Gaussian mixture model (GMM) on highly contaminated data with partially overlapped components, the SPCL algorithm had lower bias and MSE compared to a GMM fitted using the EM algorithm and the original SPL algorithm. SPCL also outperformed more robust models such as the CNM and TMM in these cases. The SPCL method also reduced the number of degenerate solutions than compared to the original SPL algorithm. We also fitted a GMM using the original SPL and SPCL methods on the well-known Iris and Wine data set to demonstrate their ability to cluster observations based on their true label.

The structure of this chapter is as follows: In Sect. 21.2, we formally define FMMs and explain how the EM algorithm is used to fit them. Section 21.3 covers traditional SPL, outlining its key concepts and definitions. In Sect. 21.4, we introduce our new method, SPCL, and provide the updating equation for fitting a GMM with SPCL. Section 21.5 presents the simulation study and discusses the results. In Sect. 21.6, we demonstrate each method's clustering performance on real data. Finally, Sect. 21.7 summarizes the findings and conclusions of this study.

## 21.2 Finite Mixture Models

In this section, a general definition of a FMMs is provided and specific cases are used in this chapter is provided. A description of the EM algorithm and how to fit a FMM using the EM algorithm is also provided.

### 21.2.1 Definition of the Finite Mixture Model

The probability density function or probability mass function of a random vector $\boldsymbol{x} \in \mathbb{R}^P$ from a $K$-component FMM with parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and $\boldsymbol{\Pi} = (\pi_1, \ldots, \pi_K)$ is

$$\Pr(\boldsymbol{x}|\boldsymbol{\Pi}, \boldsymbol{\Theta}) = \sum_{j=1}^{K} \pi_j \cdot f_j(\boldsymbol{x}|\boldsymbol{\theta}_j), \tag{21.1}$$

where $\pi_j$ the mixing proportion of the $j$-th component, $\pi_j > 0$ $\sum_{j=1}^{K} \pi_j = 1$ and $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j)$ the probability density of the $j$-th component with parameter $\boldsymbol{\theta}_j$.

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a random sample of size $n$ from model (21.1) and let $\boldsymbol{Z} = (z_1, \ldots, z_n)$, $z_i = (z_{i1}, \ldots, z_{iK})$ be the collection of latent indicator variables with

$$z_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ belongs to cluster } j, \\ 0 & \text{otherwise,} \end{cases}$$

with $\Pr(z_{ij} = 1|\boldsymbol{\Pi}, \boldsymbol{\Theta}) = \pi_j$.

The complete data are $(\boldsymbol{x}, z) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, z_1, \ldots, z_n)$ with corresponding complete log-likelihood function:

$$l(\boldsymbol{\Pi}, \boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \Big\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \Big\}. \tag{21.2}$$

In this chapter, special FMM cases will be considered:

- A GMM, when $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j) = \phi_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma_j})$, a $p$-dimensional Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma_j}$ for component $j$.
- A TMM, when $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j) = t_j(\boldsymbol{x}|\nu_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma_j})$, a $p$-dimensional noncentral $t$ distribution with degrees of freedom $\nu_j$, location $\boldsymbol{\mu}_j$, and scale matrix $\boldsymbol{\Sigma_j}$.
- If $f_j(\boldsymbol{x}|\boldsymbol{\theta}_j) = \alpha_j \phi_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma_j}) + (1 - \alpha_j)\phi_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \eta_j \cdot \boldsymbol{\Sigma_j})$ , a $p$-dimensional contaminated normal distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma_j}$, degree of contamination $\eta_j > 1$ and proportion of contamination $0 < \alpha < 1$.

### 21.2.2 Expectation Maximization Algorithm

The EM algorithm is a numerical optimization algorithm that iterates between the expectation (E) step and the maximization (M) step. The EM algorithm is typically

used to obtain MLEs in the presence of latent data. In the E-step we estimate each $z_{ij}$ value by using the conditional expected values:

$$\begin{aligned}\gamma_{ij} &= \mathrm{E}_{z_{ij}}[z_{ij}|\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{\Pi}] \\ &= \frac{\pi_j \cdot f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)}{\sum_{o=1}^{K} \pi_o \cdot f_o(\boldsymbol{x}_i|\boldsymbol{\theta}_j)}.\end{aligned} \tag{21.3}$$

The M-step consists of finding the parameter estimates $\boldsymbol{\theta}$ that maximizes $Q(\cdot|\cdot)$, the conditional expected log likelihood:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Pi}|\boldsymbol{x}) = \sum_{i=1}^{n}\sum_{j=1}^{K} \gamma_{ij}\Big\{\ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j))\Big\},$$

with respect to parameters $\pi_j$ and $\boldsymbol{\theta}_j$ for $j = 1, \ldots, K$ resulting in the following updating equations

$$\pi_j = \max_{\pi_j}\Big\{Q(\boldsymbol{\Theta}, \boldsymbol{\Pi}|\,\boldsymbol{x})\Big\} = \frac{\sum_{i=1}^{n}\gamma_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{K}\gamma_{ij}}, \tag{21.4}$$

$$\boldsymbol{\theta}_j = \max_{\boldsymbol{\theta}_j}\Big\{Q(\boldsymbol{\Theta}, \boldsymbol{\Pi}|\,\boldsymbol{x})\Big\}. \tag{21.5}$$

Algorithm 21.1 shows the EM algorithm for fitting a FMM.

**Algorithm 21.1** Fitting a FMM using the EM algorithm

**Input:** Determine the initial set of parameters $\boldsymbol{\Theta}^{(\text{old})}$ and $\boldsymbol{\Pi}^{(\text{old})}$
  **while** *(parameter estimates have not converged)* **do**
    **E-step**: Update $\gamma_{ij}^{(\text{new})}$ using Eq. (21.3)
    **M-step**:
    Update $\pi_j^{(\text{new})}$ using Eq. (21.4)
    Update $\boldsymbol{\theta}_j^{(\text{new})}$ using Eq. (21.5)
    Set $\pi_j^{(\text{old})} \leftarrow \pi_j^{(\text{new})}$
    Set $\boldsymbol{\theta}_j^{(\text{old})} \leftarrow \boldsymbol{\theta}_j^{(\text{new})}$
    Set $\gamma_{ij}^{(\text{old})} \leftarrow \gamma_{ij}^{(\text{new})}$
  **end while**

The expectation-conditional maximization (ECM) algorithm is a variant of the EM algorithm frequently used in cases where there are two or more latent variables.

The ECM algorithm uses a series of conditional maximization (CM) steps as part of the maximization step (McLachlan & Krishnan, 2008).

## 21.3 Self-Paced Learning

In this section, we discuss the traditional SPL and define key SPL notation and concepts. By introducing observations in a meaningful order, SPL mimics a learning strategy that starts with easy examples and gradually includes more difficult ones. Following the SPL approach, we use Eq. (21.6) to determine the order in which observations are introduced (Zhang et al., 2018). Observations with a higher contribution to the complete log-likelihood are considered less typical and are gradually incorporated during the later stages of MLE estimation.

Let $\boldsymbol{v} = (v_1, \ldots, v_n)$ be the set of SPL weights, where

$$v_i = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ is typical,} \\ 0 & \text{if } \boldsymbol{x}_i \text{ is non-typical.} \end{cases}$$

The SPL minimizes the penalized, weighted negative complete log-likelihood function to guide the learning process (Kumar et al., 2010; Li et al., 2022):

$$E(\boldsymbol{v}, \boldsymbol{\Theta}, \boldsymbol{\Pi}, \lambda) = \sum_{i=1}^{n} v_i \sum_{j=1}^{K} -z_{ij} \Big\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \Big\} + g(\boldsymbol{v}, \lambda), \tag{21.6}$$

where $g(\boldsymbol{v}, \lambda)$ is the self-paced regularizer (SP regularizer) and $\lambda$ is the age parameter. The age parameter controls the number of observations included in the learning process. As training progresses, $\lambda$ increases, gradually incorporating more non-typical observations. In this chapter, we focus solely on the hard regularizer:

$$g(\boldsymbol{v}, \lambda) = \lambda \cdot \sum_{i=1}^{n} |v_i|.$$

The SPL estimates the parameters by iteratively minimizing (21.6) with respect to $v_i$ and $\boldsymbol{\Theta}$:

$$\begin{aligned} v_i &= \min_{v_i} \{E(\boldsymbol{v}, \boldsymbol{\Theta}, \boldsymbol{\Pi}, \lambda)\} \\ &= \begin{cases} 1 & \text{if } \sum_{j=1}^{K} z_{ij} \Big\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \Big\} \geq \lambda, \\ 0 & \text{if } \sum_{j=1}^{K} z_{ij} \Big\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \Big\} < \lambda, \end{cases} \end{aligned} \tag{21.7}$$

$$\pi_j = \min_{\pi_j} \{E(\boldsymbol{v}, \boldsymbol{\Theta}, \boldsymbol{\Pi}, \lambda)\}$$
$$= \frac{\sum_{i=1}^{n} v_i \cdot \gamma_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{K} v_i \cdot \gamma_{ij}}, \tag{21.8}$$

and

$$\boldsymbol{\Theta} = \min_{\boldsymbol{\theta}} \{E(\boldsymbol{v}, \boldsymbol{\Theta}, \boldsymbol{\Pi}, \lambda)\}. \tag{21.9}$$

The age parameter $\lambda$ is increased by

$$\lambda = \lambda \cdot a, \tag{21.10}$$

where $a > 1$ (Zhou & Zhang, 2012).

This process continues until all observations are included or the model parameters converge. Algorithm 21.2 shows the process of fitting a FMM using SPL.

**Algorithm 21.2** Fitting a FMM using SPL

**Input:** Initial set of parameters $\boldsymbol{\Theta}^{(\text{old})}$, $\boldsymbol{\Pi}^{(\text{old})}$, $\lambda^{(\text{old})} > 0, a > 1$
  **while** (*while at least one* $v_i^{(\text{old})} = 0$) **or** (*parameter estimates have not converged*) **do**
    Update $v_i^{(\text{new})}$ using Eq. (21.7)
    **while** *parameter estimates have not converged* **do**
      **E-Step:** Update $\gamma_{ij}^{(\text{new})}$ using Eq. (21.3)
      **M-Step:**
        Update $\pi_j^{(\text{new})}$ using Eq. (21.8)
        Update $\boldsymbol{\theta}_j^{(\text{new})}$ using Eq. (21.9)
    **end while**
    Update $\lambda^{(\text{old})}$ using Eq. (21.10)
    Set $\mathbf{v}^{(\text{old})} \leftarrow \mathbf{v}^{(\text{new})}$
    Set $\pi_j^{(\text{old})} \leftarrow \pi_j^{(\text{new})}$
    Set $\boldsymbol{\theta}_j^{(\text{old})} \leftarrow \boldsymbol{\theta}_j^{(\text{new})}$
    Set $\lambda^{(\text{old})} \leftarrow \lambda^{(\text{new})}$
  **end while**

## 21.4 Self-Paced Component Leaning

In this section, we present the proposed SPCL approach and detail the updating equations required for fitting a GMM using SPCL.

We propose fitting a FMM using SPL at the component level. For each observation, SPCL uses the negative log-likelihood contribution of each component to the complete log-likelihood as a measure of how typical the observation is for that component. Observation $i$ is assigned a self-paced learning weight $v_{ij}$ for component $j$.

Let $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, where $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iK})$, represent the SPCL weights, with

$$v_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ is a typical observation for component } j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ represent the set of age parameters for each component, where $\lambda_j$ is the age parameter of component $j$. SPCL minimizes the penalized negative log-likelihood function:

$$E(\boldsymbol{\Theta}, \boldsymbol{\Pi}, \boldsymbol{v}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{j=1}^{K} -v_{ij} z_{ij} \Big\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \Big\} + \sum_{j=1}^{K} \lambda_j \sum_{i=1}^{n} |v_{ij}|. \tag{21.11}$$

In each iteration of SPCL, the following equations are updated:

$$\begin{aligned} v_{ij} &= \min_{v_{ij}} \{E(\boldsymbol{\Theta}, \boldsymbol{\Pi}, \boldsymbol{v}, \boldsymbol{\lambda})\} \\ &= \begin{cases} 1 & \text{if } z_{ij} \left\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \right\} \geq \lambda_j, \\ 0 & \text{if } z_{ij} \left\{ \ln(\pi_j) + \ln(f_j(\boldsymbol{x}_i|\boldsymbol{\theta}_j)) \right\} < \lambda_j, \end{cases} \end{aligned} \tag{21.12}$$

$$\begin{aligned} \pi_j &= \min_{\pi_j} \{E(\boldsymbol{\Theta}, \boldsymbol{\Pi}, \boldsymbol{v}, \boldsymbol{\lambda})\} \\ &= \frac{\sum_{i=1}^{n} v_{ij} \cdot \gamma_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{K} v_{ij} \cdot \gamma_{ij}}, \end{aligned} \tag{21.13}$$

and

$$\boldsymbol{\Theta} = \min_{\boldsymbol{\Theta}} \{E(\boldsymbol{\Theta}, \boldsymbol{\Pi}, \boldsymbol{v}, \boldsymbol{\lambda})\}. \tag{21.14}$$

The age parameter $\lambda_j$ is updated as

$$\lambda_j = \lambda_j \cdot a, \tag{21.15}$$

with $a > 1$.

Fitting a FMM using SPCL is summarized in Algorithm 21.3.

**Algorithm 21.3** Fitting a FMM using SPCL

**Input:** Initial set of parameters $\boldsymbol{\Theta}^{(\text{old})}$, $\boldsymbol{\Pi}^{(\text{old})}$, $\lambda^{(\text{old})} > 0, a > 1$
  **while** (*while at least one* $v_{ij}^{(\text{old})} = 0$) **or** (*parameter estimates have not converged*) **do**
    **while** (at least one $v_{ij}^{(\text{old})} = 0$) **or** (parameters have not converged) **do**
      **E-Step:** Update $\gamma_{ij}^{(\text{new})}$ using Eq. (21.3)
      **M-Step:**
        Update $v_{ij}^{(\text{new})}$ using Eq. (21.12)
        Update $\pi_j^{(\text{new})}$ using Eq. (21.13)
        Update $\boldsymbol{\theta}_j^{(\text{new})}$ using Eq. (21.14)
      Set $v_{ij}^{(\text{old})} \leftarrow v_{ij}^{(\text{new})}$
      Set $\pi_j^{(\text{old})} \leftarrow \pi_j^{(\text{new})}$
      Set $\boldsymbol{\theta}_j^{(\text{old})} \leftarrow \boldsymbol{\theta}_j^{(\text{new})}$
    **end while**
    Update $\lambda^{(\text{old})}$ using Eq. (21.10)
    Set $\lambda^{(\text{old})} \leftarrow \lambda^{(\text{new})}$
  **end while**

The SPCL approach ensures that each component receives enough observations for accurate representation. It achieves this by using $n \times K$ self-paced component learning weights, instead of the $n$ weights used in standard self-paced learning. These weights are based solely on each component's likelihood, providing a more precise SPL weight for every component. The SPCL weights are updated after each EM iteration to reflect the current parameter estimates.

The following updating equations are used to fit a GMM using SPCL:

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{n} v_{ij} \cdot \gamma_{ij} \cdot \boldsymbol{x}_i}{\sum_{i=1}^{n} v_{ij} \cdot \gamma_{ij}}, \tag{21.16}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{n} v_{ij} \cdot \gamma_{ij} \cdot (\boldsymbol{x}_i - \boldsymbol{\mu}_j) \cdot (\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top}{\sum_{i=1}^{n} v_{ij} \cdot \gamma_{ij}}. \tag{21.17}$$

The SPCL approach for fitting a GMM is summarized in Algorithm 21.4.

**Algorithm 21.4** Fitting a GMM using SPCL

**Input:** Initial set of parameters $\boldsymbol{\mu}_j^{(\text{old})}$, $\boldsymbol{\Sigma}_j^{(\text{old})}$, $v_{ij}^{(\text{old})}$, $\lambda_j^{(\text{old})}$, $a > 1$

**while** (while at least one $v_{ij}^{(\text{old})} = 0$) **or** (parameters have not converged) **do**

  **while** (while at least one $v_{ij}^{(\text{old})} = 0$) **or** (parameters have not converged) **do**

    **Perform E-step:**

      Update $\gamma_{ij}^{(\text{new})}$ using Eq. (21.3)

    **Perform M-step:**

      Update $v_{ij}^{(\text{new})}$ using Eq. (21.12)

      Update $\boldsymbol{\pi}_j^{(\text{new})}$ using Eq. (21.13)

      Update $\boldsymbol{\mu}_j^{(\text{new})}$ using Eq. (21.16)

      Update $\boldsymbol{\Sigma}_j^{(\text{new})}$ using Eq. (21.17)

    Set $v_{ij}^{(\text{old})} \leftarrow v_{ij}^{(\text{new})}$

    Set $\pi_j^{(\text{old})} \leftarrow \pi_j^{(\text{new})}$

    Set $\boldsymbol{\mu}_j^{(\text{old})} \leftarrow \boldsymbol{\mu}_j^{(\text{new})}$

    Set $\boldsymbol{\Sigma}_j^{(\text{old})} \leftarrow \boldsymbol{\Sigma}_j^{(\text{new})}$

  **end while**

  Update $\boldsymbol{\lambda}^{(\text{new})}$ using Eq. (21.15)

  Set $\lambda^{(\text{old})} \leftarrow \lambda^{(\text{new})}$

**end while**

## 21.5 Simulation Study

The SPL and SPCL are evaluated through a simulation study using R 4.4.1. The strengths and limitations of each method are discussed. The goal is to assess their robustness compared to the EM algorithm by fitting a GMM with SPCL, SPL, and the EM.

### *21.5.1 Data Generation*

Contamination models are used to generate various contamination scenarios by mixing observations from a main distribution with effects from a contamination distribution. The probability density or mass function of a random vector $\boldsymbol{x} \in \mathbb{R}^p$ from a contamination model with parameters $\boldsymbol{\theta} = (\boldsymbol{\theta_f}, \boldsymbol{\theta_h})$ is given by

$$g(\boldsymbol{x}|\boldsymbol{\theta}, \beta) = \beta \cdot f(\boldsymbol{x}|\boldsymbol{\theta}_f) + (1 - \beta) \cdot h(\boldsymbol{x}|\boldsymbol{\theta}_h), \tag{21.18}$$

where $0 < \beta \leq 1$ is the proportion of contamination. The function $f(\boldsymbol{x}|\boldsymbol{\theta}_f)$ represents the probability density of the main distribution with parameters $\boldsymbol{\theta_f}$, and $h(\boldsymbol{x}|\boldsymbol{\theta}_h)$ represents the probability density of the contamination distribution with parameters $\boldsymbol{\theta_h}$.

This chapter also uses a finite mixture of contamination models for intricate contamination scenarios. The probability density function of a random vector $\boldsymbol{x} \in \mathbb{R}^p$ from a $K$-component finite mixture of contamination models, with parameters $(\boldsymbol{\Pi}, \boldsymbol{\Theta}, \boldsymbol{\beta})$ where $\boldsymbol{\Pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ is

$$\begin{aligned} g(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\Pi}, \boldsymbol{\beta}) &= \sum_{j=1}^{K} \pi_j g_j(\boldsymbol{x}|\boldsymbol{\theta}_{fj}, \boldsymbol{\theta}_{hj}) \\ &= \sum_{j=1}^{K} \pi_j \left[\beta_j \cdot f_j(\boldsymbol{x}|\boldsymbol{\theta}_{fj}) + (1-\beta_j) \cdot h_j(\boldsymbol{x}|\boldsymbol{\theta}_{hj})\right], \quad (21.19) \end{aligned}$$

where $\pi_j$ is the mixing proportion of the $j$th component, $\beta_j$ is the proportion of contamination for component $j$, $f_j(\boldsymbol{x}|\boldsymbol{\theta}_{fj})$ is the density of the main distribution of the $j$th component with parameters $\boldsymbol{\theta}_{fj}$, and $h_j(\boldsymbol{x}|\boldsymbol{\theta}_{hj})$ is the density of the contamination distribution of the $j$th component with parameters $\boldsymbol{\theta}_{hj}$.

We generate data from six contamination scenarios, each with $p = 2$ dimensions and $K = 2$ clusters. All six data generation processes have the following shared parameters:

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad \pi_1 = 0.3.$$

For each scenario, we consider partially overlapped components with means $\boldsymbol{\mu}_1 = (0, -1)^\top$, and we use well-separated components with $\boldsymbol{\mu}_1 = (0, -3)^\top$.

Data from the following scenarios were generated:

(a) **Data with no contamination**: Using Eq. (21.19) with parameters $\boldsymbol{\beta} = (1, 1)^\top$ and $f_j(\boldsymbol{x}|\boldsymbol{\theta}_{fj}) = \phi_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with $\boldsymbol{\theta}_j = \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the mean vector and covariance matrix of the $j$ th component, respectively.

(b) **Component level contamination** Using Eq. (21.19) with parameters $\boldsymbol{\beta} = (0.9, 0.8)^\top$ and

$$f_j(\boldsymbol{x}|\boldsymbol{\theta}_{fj}) = \phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

$$h_j(\boldsymbol{x}; \boldsymbol{\theta}_{hj}) = \phi(\boldsymbol{x}|\boldsymbol{\mu}_j, \eta_j \boldsymbol{\Sigma}_j),$$

with $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ the mean and covariance matrix parameters and the degree of contamination for each parameter is $\eta_1 = 20$ and $\eta_2 = 30$, respectively.

(c) **Components with heavy-tailed distributions**: Using Eq. (21.19) with parameters $\boldsymbol{\beta} = (1, 1)^\top$ and $f_j(\boldsymbol{x}|\boldsymbol{\theta}_{fj}) = t(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu_j)$, where $\boldsymbol{\theta}_{fj} = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu_j)$,

$\boldsymbol{\mu}_j$, and $\boldsymbol{\Sigma}_j$ are the location and scale parameters, and $\nu_j$ is the degrees of freedom. The degrees of freedom for each respective component are $\nu_1 = 4$ and $\nu_2 = 10$.

(d) **Data with sparse outlier contamination**: Using Eq. (21.18) with parameter $\beta = 0.99$, $f(\boldsymbol{x}|\boldsymbol{\theta}_f)$ is a GMM, and $h(\boldsymbol{x}; \boldsymbol{\theta}_h) = \mathcal{U}(x_{i2} \mid 10, 15)$, where $\mathcal{U}$ denotes the continuous uniform distribution with parameters 10 and 15. Spares non-typical data are generated with the coordinates $(0, x_{i2})$.

(e) **Data with** 5% **uniform noise contamination**: Using Eq. (21.18) with parameter $\beta = 0.95$. Similar to the scenario (d), this model introduces a controlled level of external contamination by replacing typical observations with uniform noise. This means $h(\boldsymbol{x})$ is a bivariate uniform distribution with density on $-10 < x_1 < 10$ and $-10 < x_2 < 10$.

(f) **Data with** 10% **uniform noise contamination**: Similar to scenario (e), with $\beta = 0.9$.

### *21.5.2 Model Fitting*

A GMM with unconstrained covariance matrices was fitted using the standard EM, SPL, and SPCL approaches, with initial values obtained through $k$-means clustering. The `kmeans()` function from the `Mclust` package was used in this study. We used the `Mclust()` function to fit a GMM using the standard EM (Scrucca et al., 2016). In the results tables, we refer to a GMM fitted using the standard EM algorithm as GMM.

For SPL, we calculate the initial age parameter as

$$\lambda^{(\text{old})} = \text{median}\left(\left\{\sum_{j=1}^{K} \gamma_{ij}^{(\text{old})}\left(\ln \pi_j^{(\text{old})} + \phi_j(\boldsymbol{x}_i|\boldsymbol{\mu}_j^{(\text{old})}, \boldsymbol{\Sigma}_j^{(\text{old})})\right) \middle| i = 1, 2, \ldots, n\right\}\right).$$

This approach is consistent with the method found in Zhang et al. (2018).

We propose initializing each SPCL age parameter as the median of the top $\pi_j^{(\text{old})} \times 100\%$ of observations that contribute the most to the component's complete likelihood. This ensures that each component is fitted using the most representative data, reducing the influence of outliers. It also prevents data from larger components from influencing the estimation of smaller components. We chose $a = 1.05$ for both the SPL and SPCL approaches. Alternatively, the age parameters and $a$ could be selected using cross-validation.

Due to the robustness of the CNM and TMM models in handling non-typical observations, we fit these models for comparison purposes. The CNM was fitted using the `CNmixt()` function from the `ContaminatedMixt` package (Punzo et al., 2018), where the degree of contamination varies across groups and is estimated. Similarly, TMM was fitted using the `teigen()` function from the `teigen` package (Andrews et al., 2018), with the degrees of freedom being estimated and varying

across groups. Both CNM and TMM were fitted using unconstrained covariance matrix models.

### *21.5.3 Simulation Study Results*

We will discuss the simulation study results. This is done in two parts: First, we discuss the number of degenerate solutions calculated by each method, and then we consider the bias and MSE values of each method's parameter estimates under different contamination scenarios.

#### 21.5.3.1 Degenerate Solutions

All approaches estimated some form of degenerate solution; however, an approach with more frequent degenerate solutions is less desirable. For the simulation study, only data sets where all approaches produced nondegenerate solutions were used, ensuring a fair comparison. A total of 2,684 (7.5%) of the simulated data sets had at least 1 approach estimate a degenerate solution.

Table 21.1 shows that SPL estimated the most degenerate solutions, with 2,653 (7.37%) of the data sets. SPCL estimated the second most, with 353, and TMM the third, with 310. Since SPL selects data subsets based on the complete log-likelihood values, it biases the selection toward observations from larger components. Smaller components, having lower likelihood values, are less likely to be chosen during the initial subset selection.

This bias can result in an overrepresentation of larger components, while observations from smaller components may be underrepresented or, in the worst case, completely omitted. As a result, SPL may capture the characteristics of the dominant component, reducing the smaller components to a single observation and leading to a degenerate solution. SPCL, in contrast, reduces the number of degenerate solutions by applying SPL weights to each component, ensuring that each component receives enough observations to be accurately fitted.

#### 21.5.3.2 Performance Assessment

The bias and MSE for the parameter $\theta$ are calculated respectively as

$$\text{Bias}_{\hat{\boldsymbol{\theta}}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})$$

**Table 21.1** Total number of degenerate solutions

| CNM | GMM | SPCL | SPL | TMM |
|---|---|---|---|---|
| 213 | 49 | 353 | 2653 | 310 |

and

$$\text{MSE}_{\hat{\boldsymbol{\theta}}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^2,$$

where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ is the set of parameter estimates for each data set $i$. Both bias and MSE values closer to zero signify better estimation performance.

We compare the bias and MSE values for each method's estimated mixing proportion $\pi_1$ and the univariate means of each component; $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Label switching is a well-known problem in finite mixture modeling (Punzo & McNicholas, 2016). We label the estimated parameters based on the lowest least squared distance from the true parameter values.

Table 21.2 shows the bias and MSE values for the parameter estimates of each method for scenario (a) with well-separated components and a sample size of $n = 500$. All methods had low bias and MSE values for their parameter estimates. SPL had the highest bias and MSE values, but this is to be expected, as it is well known that self-paced learning algorithms does not work as well on clean data sets (Meng et al., 2017). SPCL however had lower bias values for the mean parameter estimates than the EM algorithm and SPL. The GMM, CNM, and TMM had the lowest MSE values for their mean and proportion estimates. SPL and SPCL have higher MSE values.

Table 21.3 reports on scenario (b) with well-separated components and $n = 500$. GMM fitted using the EM algorithm had the highest overall bias values for its parameter estimates than compared to a GMM fitted using SPL and SPCL. SPCL and TMM had lower bias and MSE for the mean parameter estimates for both components. SPCL had the overall lowest bias estimates. SPL had the lowest MSE values for mean components, and SPCL had the lowest bias and MSE values for the proportion estimates.

Table 21.4 reports on scenario (d) with well-separated components. SPCL had the lowest overall bias values for the mean and proportion parameters. GMM has the highest overall bias values for the mean parameter estimate. TMM has the lowest MSE values for the mean parameter estimates.

Scenario (f), with partially overlapped components, is reported in Table 21.5. GMM had the highest bias in both the mean and proportion estimates. SPCL had

**Table 21.2** Scenario (a) with well-separated components, $n = 500$

| | Bias | | | | | MSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GMM | SPCL | SPL | CNM | TMM | GMM | SPCL | SPL | CNM | TMM |
| $\mu_{11}$ | 0.0049 | **0.0025** | 0.0035 | 0.0047 | 0.0051 | **0.0070** | 0.0093 | 0.0101 | **0.0070** | **0.0070** |
| $\mu_{12}$ | 0.0005 | −0.0005 | 0.0035 | 0.0008 | **0.0001** | **0.0064** | 0.0089 | 0.0102 | **0.0064** | **0.0064** |
| $\mu_{21}$ | −0.0021 | **−0.0010** | −0.0029 | −0.0020 | −0.0022 | **0.0026** | 0.0035 | 0.0036 | **0.0026** | **0.0026** |
| $\mu_{22}$ | −0.0004 | −0.0004 | −0.0016 | **−0.0002** | −0.0008 | **0.0026** | 0.0035 | 0.0035 | **0.0026** | **0.0026** |
| $\pi_1$ | 0.0001 | 0.0007 | −0.1037 | 0.0001 | **0.0000** | **0.0000** | **0.0000** | 0.0206 | **0.0000** | **0.0000** |

**Table 21.3** Scenario (b) with well-separated components, $n = 500$

| | Bias | | | | | MSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GMM | SPCL | SPL | CNM | TMM | GMM | SPCL | SPL | CNM | TMM |
| $\mu_{11}$ | 0.0339 | **0.0061** | −0.0529 | 0.0221 | −0.0171 | 0.1899 | 0.7216 | **0.0437** | 0.2447 | 0.0595 |
| $\mu_{12}$ | 2.2728 | **0.2887** | 0.4056 | 2.5602 | 0.6641 | 5.6076 | 2.0658 | **0.8211** | 6.9073 | 1.6734 |
| $\mu_{21}$ | 0.0514 | 0.0108 | 0.0082 | 0.0742 | **0.0055** | 0.0966 | **0.0076** | 0.0143 | 0.1408 | 0.0310 |
| $\mu_{22}$ | −0.4721 | −0.0066 | **−0.0008** | −0.7644 | −0.1547 | 0.5219 | 0.0345 | **0.0313** | 0.8413 | 0.1913 |
| $\pi_1$ | 0.2426 | **0.0390** | 0.0717 | 0.2477 | 0.1087 | 0.1249 | **0.0090** | 0.0207 | 0.1634 | 0.0472 |

**Table 21.4** Scenario (d) with well-separated components, $n = 500$

| | Bias | | | | | MSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GMM | SPCL | SPL | CNM | TMM | GMM | SPCL | SPL | CNM | TMM |
| $\mu_{11}$ | −0.0471 | **−0.0021** | −0.0046 | 0.0212 | 0.0155 | 0.0121 | 0.0080 | 0.0099 | 0.0071 | **0.0067** |
| $\mu_{12}$ | 1.0368 | **0.0005** | 0.0029 | −0.0436 | −0.0233 | 1.8164 | 0.0087 | 0.0108 | 0.0089 | **0.0076** |
| $\mu_{21}$ | 0.0330 | **−0.0017** | −0.0030 | −0.0109 | −0.0055 | 0.0063 | 0.0036 | 0.0036 | **0.0031** | 0.0032 |
| $\mu_{22}$ | 0.0954 | 0.0062 | 0.0124 | 0.1006 | **0.0017** | 0.0126 | 0.0040 | 0.0044 | 0.0134 | **0.0032** |
| $\pi_1$ | 0.0616 | **−0.0044** | −0.0651 | −0.0093 | −0.0070 | 0.0070 | **0.0001** | 0.0109 | **0.0001** | **0.0001** |

**Table 21.5** Scenario (f) with partially overlapped components, $n = 500$

| | Bias | | | | | MSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GMM | SPCL | SPL | CNM | TMM | GMM | SPCL | SPL | CNM | TMM |
| $\mu_{11}$ | 0.0582 | −0.1059 | −0.1073 | **0.0377** | −0.0429 | 0.1881 | 0.9051 | 0.1099 | 0.1152 | **0.0125** |
| $\mu_{12}$ | 0.8136 | **0.1178** | 0.6195 | 0.7915 | 0.1402 | 0.8494 | 0.8629 | 1.3811 | 0.7752 | **0.0356** |
| $\mu_{21}$ | **−0.0041** | 0.0097 | 0.0057 | −0.0123 | 0.0372 | **0.0042** | 0.0050 | 0.0258 | 0.0046 | 0.0053 |
| $\mu_{22}$ | −0.2899 | **−0.0002** | −0.0447 | −0.1779 | 0.0560 | 0.3322 | 0.0254 | 0.0887 | 0.1734 | **0.0075** |
| $\pi_1$ | 0.0537 | **0.0113** | 0.0528 | 0.0822 | 0.0928 | 0.0215 | **0.0031** | 0.0342 | 0.0171 | 0.0089 |

the lowest bias and MSE for the proportion estimates. While both SPCL and TMM had the lowest bias for the mean parameters, TMM had the lowest MSE values. SPL and CNM had higher overall bias and MSE values compared to SPCL.

#### 21.5.3.3 Simulation Study Conclusion

From the simulation study, we conclude that SPCL effectively mitigates the impact of non-typical observations. A GMM fitted using SPL or SPCL is more robust than one fitted with the EM algorithm. In particular, a GMM fitted with SPCL provides the best parameter estimates on highly contaminated data sets, without requiring additional parametric assumptions.

## 21.6 Real Data Analysis

In this section we fit a GMM using the SPL and SPCL to demonstrate their ability to cluster observations based on their true label. We also consider the CNM and TMM for our demonstration. We fit the methods on the Iris and Wine data set. For the sake of simplicity, we consider the unconstrained covariance structure of each model.

### *21.6.1 Iris Data Set*

The Iris data set consists of $n = 150$ observations and captured 4 morphological measurements $p = 4$, namely, the sepal length, sepal width, petal length, and petal width of 3 iris species, *setosa*, *versicolor*, and *virginica* (Fisher, 1922). There are 50 observations for each species. We perform a clustering analysis on the measurements by ignoring the species label. We are studying how well each method can find the correct species label for each observation. We fit the unconstrained CNM and TMM models.

Table 21.6 shows the Bayesian information criteria (BIC) value to determine the number of components for each method, where a lower BIC value corresponds to a better fit. BIC is calculated using the formula $\text{BIC} = -2l + m \ln(n)$ where $l$ is the complete log-likelihood and $m$ is the number of free parameters estimated.

From Table 21.6, we see the GMM fitted using SPCL was able to correctly identify three labels present in the data set. All other approaches did not find the correct number of labels. Both CNM and TMM fitted two components, while SPL fitted four components. Missing BIC values indicate computational difficulties, and therefore could not calculate the BIC values.

Table 21.7 shows the clustering results for the Iris data set. SPCL correctly identified all the *setosa* and *versicolor* observations. Five observations of the species *virginica* were incorrectly clustered as *versicolor*. SPL clustered the *setosa* observations in two groups. Finally CNM and TMM clustered all the *versicolor* and *virginica* observations as one group.

**Table 21.6** BIC values for different numbers of clusters ($K$) across various algorithms for the Iris data set

| $K$ | SPCL | SPL | CNM | TMM |
|---|---|---|---|---|
| 1 | 1046.67 | 1046.67 | 939.09 | – |
| 2 | 715.54 | 715.54 | 790.70 | 583.81 |
| 3 | 700.28 | 690.53 | 797.52 | 593.64 |
| 4 | 730.38 | 689.72 | 841.69 | 646.84 |

**Table 21.7** Clustering results for each competing method on the Iris data set

| | SPCL | | | SPL | | | | CNM | | TMM | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 1 | 2 | 1 | 2 |
| *setosa* | 50 | 0 | 0 | 26 | 0 | 24 | 0 | 50 | 0 | 0 | 50 |
| *versicolor* | 0 | 50 | 0 | 0 | 46 | 0 | 4 | 0 | 50 | 50 | 0 |
| *virginica* | 0 | 5 | 45 | 0 | 2 | 0 | 48 | 0 | 50 | 50 | 0 |

**Table 21.8** BIC values for different numbers of clusters ($K$) across various algorithms for the Wine data set

| $K$ | SPCL | SPL | CNM | TMM |
|---|---|---|---|---|
| 1 | 5728.30 | 5728.30 | 6008.88 | – |
| 2 | 4944.27 | 4865.58 | 5713.33 | 5531.60 |
| 3 | 4465.89 | 4597.49 | 5794.51 | 5740.93 |
| 4 | 4500.22 | – | 6008.88 | 6133.88 |

**Table 21.9** Clustering results for each competing method on Wine data set

| | SPCL | | | SPL | | | CNM | | TMM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 1 | 2 |
| *Barbera* | 0 | 48 | 0 | 0 | 0 | 48 | 0 | 48 | 0 | 48 |
| *Barolo* | 14 | 0 | 45 | 59 | 0 | 0 | 59 | 0 | 59 | 0 |
| *Grignolino* | 71 | 0 | 0 | 48 | 23 | 0 | 28 | 43 | 68 | 3 |

### *21.6.2 Wine Data*

The Wine data set contains $n = 178$ Italian wines, each with $p = 13$ measurements of chemical and physical properties, from 3 cultivars: Barbera, Barolo, and Grignolino (Forina et al., 1998). Table 21.8 shows the Bayesian information criteria (BIC) for each approach fitted on the Wine data set.

From Table 21.8, we see the GMM fitted using SPCL and SPL was able to correctly identify three labels present in the data set. CNM and TMM found two components.

A GMM fitted using SPCL correctly clustered all the Gringnolino and Barbera observations together. SPCL incorrectly clustered 14 observations from the Barolo cultivar with the Grignolino observations. SPL incorrectly clustered most of the Barolo and Grignolino observations together. TMM clustered all the Barolo and Grignolino observations together, whereas CNM clustered most of the Barrbera and Grignolino observations together.

From Table 21.9, we see the GMM fitted using SPCL and SPL was able to correctly identify three labels present in the data set. CNM and TMM found two components. A GMM fitted using SPCL correctly clustered all the Gringnolino and Barbera observations together. SPCL incorrectly clustered 14 observations from the Barolo cultivar with the Grignolino observations. SPL incorrectly clustered most of the Barolo and Grignolino observations together. TMM clustered all the Barolo and Grignolino observations together, whereas CNM clustered most of the Barrbera and Grignolino observations together.

## 21.7 Discussion and Conclusion

We studied the properties of SPL when fitting FMMs on different contamination scenarios. We found that the MLE estimates for a GMM using SPL reduced the impact of outliers. However, SPL estimates more degenerate solutions than the EM algorithm, and we identified that the observation selection criteria are the main reason for this.

We also proposed SPCL to address the key issues present in SPL. We studied the properties of SPCL and found it also reduced the impact of outliers on highly contaminated data sets. A GMM fitted using SPCL provided better MLE estimates than SPL and the EM algorithm, and SPCL estimated fewer degenerate solutions than SPL. Additionally, SPCL shows robustness similar to models like the TMM, without additional parametric assumptions. We fitted GMMs on real data using SPCL and SPL to show their ability to cluster data according to their true labels. SPCL identified the correct number of clusters for each data set and grouped the most observations of the same true label together.

This research shows that the proposed approach SPCL is a robust method for parameter estimation by reducing the impact of non-typical observations. Future work may explore a self-paced learning algorithm for FMM problems, aiming to enhance performance on clean data and further minimize degenerate solutions.

**Acknowledgments** This study was funded by a grant from the Statistics Industry Hub and Statomet, Department of Statistics, University of Pretoria.

## References

Andrews, J. L., Wickins, J. R., Boers, N. M., McNicholas, P. D. (2018). teigen: An R package for model-based clustering and classification via the multivariate t distribution. *Journal of Statistical Software, 83*(7), 1–32.

Banfield, J. D., Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*(3), 803–821.

Biernacki, C., & Chrétien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM. *Statistics & Probability Letters, 61*, 373–382.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222*, 309–368.

Forina, M., Leardi, R., Armanino, C., & Lanteri, S. (1998). PARVUS: *An extendible package for data exploration, classification and correlation*. Technical Report, Institute of Pharmaceutical and Food Analysis and Technologies, Genoa, Italy.

Henning, C. (2004). Breakdown points for maximum likelihood estimators of location–scale mixtures. *The Annals of Statistics, 32*, 1313–1340.

Hipp, J. R., Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychol Methods, 11*(1), 36–53.

Kumar, M., Packer, B., Koller, D. (2010). Self-paced learning for latent variable models. In: J. Lafferty, C. Williams, J. Shawe-Taylor, Zemel, R., & Culotta, A. (eds.). *Advances in Neural Information Processing Systems* (Vol. 23). Curran Associates, Inc.

Li, W., Chen, H., Li, T., Wan, J., & Sang, B. (2022). Unsupervised feature selection via self-paced learning and low-redundant regularization. *Knowledge-Based Systems, 240*, 108150.

McLachlan, G. J., & Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley.

McLachlan, G. J., & Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In: A. Amin, D. Dori, P. Pudil, Freeman, H. (Eds.) *Advances in pattern recognition. SSPR/SPR 1998. Lecture Notes in Computer Science* (Vol. 1451, pp. 658–66). Springer.

Meng, D., Zhao, Q., & Jiang, L. (2017). A theoretical understanding of self-paced learning. *Information Sciences, 414*, 319–328.

Peel, D., & McLachlan, G. (2000a). *Finite mixture models*. Wiley.

Peel, D., & McLachlan, G. J. (2000b). Robust mixture modelling using the t distribution. *Statistics and Computing, 10*, 339–348.

Punzo, A., Mazza, A., & McNicholas, P. D. (2018). ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *Journal of Statistical Software, 85*(10), 1–25.

Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal, 58*, 1506–1537.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 289–317.

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods, 49*, 282–293.

Zhang, Y., Tang, Q., Niu, L., Dai, T., Xiao, X., & Xia, S.-T. (2018). Self-paced mixture of t distribution model. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada (pp. 2796–2800).

Zhou, H., Zhang, Y. (2012). EM vs MM: A case study. *Computational Statistics & Data Analysis, 56*(12), 3909–3920.

# Chapter 22
# Shrinkage Estimation in Generalized CIR Processes with Change-Point




**Yunhong Lyu and Sévérien Nkurunziza**

## 22.1 Introduction

The Cox-Ingersoll-Ross (CIR) model is a cornerstone of financial research, widely recognized for its ability to model interest rate dynamics by incorporating key features such as mean reversion and volatility clustering (Cox et al., 1985a,b). This model has found extensive applications in bond pricing, option valuation, and risk management (Duffie & Kan, 1996). However, its classical formulation assumes a constant mean-reverting level, which may not adequately reflect the complexities of real-world financial markets.

Further, the classical CIR process might not align with data where the mean-reverting level can vary over time. This is especially notable in phenomena influenced by factors that dynamically change over time. For example, financial markets are subject to structural changes and shifts in dynamics, which can significantly impact interest rate behavior. These change-points can arise due to various factors such as monetary policy shifts, financial crises, changes in market regimes, or macroeconomic events (Hamilton, 1989; Bai & Perron, 1998). Statistical inference has struggled with change-point identification in stochastic processes over the years. Ignoring these change-points in the CIR model can lead to biased parameter estimates, insufficient volatility modeling, and suboptimal

Y. Lyu
Mathematics and Statistics Department, University of Windsor, Windsor, ON, Cananda

Department of Mathematics and Statistics, Trent University, Peterborough, ON, Canada
e-mail: lyu111@uwindsor.ca;yunhonglyu@trentu.ca

S. Nkurunziza (✉)
Mathematics and Statistics Department, University of Windsor, Windsor, ON, Cananda
e-mail: severien@uwindsor.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_22

pricing of financial instruments. Capturing these structural changes is not merely a theoretical concern; it is critical for accurate risk assessment and decision-making in finance. For example, during the 2008 global financial crisis, unprecedented market conditions significantly altered interest rate dynamics, rendering traditional models inadequate for capturing these shifts. Similarly, shifts in monetary policies, such as the transition from expansionary to contractionary measures, introduce discontinuities that challenge the assumptions of constant parameters in traditional CIR models. The need for models that can account for such structural changes is increasingly pressing in today's rapidly evolving financial landscape.

Hull and White (1990) first extended the constant mean reversion level to a time-dependent mean reversion function in the context of the O-U process. Dehling et al. (2014) considered the generalized O-U process with periodic mean-reverting term and one known change-point. Further, Nkurunziza and Zhang (2018) studied an inference problem in generalized O–U processes with an unknown change-point when the drift parameter is suspected to satisfy a linear restriction. However, the CIR model—a pivotal tool in financial modeling—has not yet been generalized to account for structural changes, leaving a significant gap in the literature.

Motivated by these gaps and the limitations of the classical CIR model, this chapter introduces a Generalized CIR model designed to incorporate periodic long-term means and structural breaks. By addressing the challenges posed by change-point, this framework enables more realistic modeling of interest rate dynamics and other financial phenomena. Our focus is on developing shrinkage estimation method within the GCIR model with change-point. To evaluate the effectiveness of the proposed method, we conduct extensive experiments using simulated data as well as interest rate real datasets from various financial markets.

The rest of this chapter is structured as follows. In Sect. 22.2, we present the statistical model and some preliminary findings. Section 22.3 focuses on inference within the context of a known change-point. Here, we derive both the Unrestricted Maximum Likelihood Estimator (UMLE) and the Restricted Maximum Likelihood Estimator (RMLE) for the drift parameters. Furthermore, we establish their joint asymptotic normality under a set of local restrictions. Moving on to Sect. 22.4, we focus on the shrinkage estimator in the case of an unknown change-point. We show that the shrinkage estimators (SEs) optimally blend the UMLE and the RMLE. This section also gives a comparison of the relative efficiency among the estimators. Section 22.5 presents the numerical results from the application of the proposed method to real financial market historical data. Finally, in Sect. 22.6, we give some concluding remarks. Further, for the convenience of the reader, we also include in the supplementary file technical results and simulation results.

## 22.2 Statistical Model and Preliminary Results

In this section, we present the statistical model and some preliminary results that are useful for this chapter. First, let us introduce the concepts and notations used throughout the remainder of the paper. Let $(\Omega^*, \mathscr{F}, \mathrm{P})$ be a probability space where

$\mathscr{F}$ is $\sigma-$field on the sample space $\Omega^*$ and P is a probability measure. Further, let $L^d$ denote the space of measurable $d-$integrable function, for some $d \geqslant 1$. We denote $\xrightarrow[T\longrightarrow\infty]{\mathrm{D}}$, $\xrightarrow[T\to\infty]{L^d}$, $\xrightarrow[T\to\infty]{a.s.}$, $\xrightarrow[T\to\infty]{\mathrm{P}}$ the convergence in distribution, in $L^d-$space, almost surely, and in probability, respectively, as $T$ tends to infinity. Let $\mathbb{R}_+ = [0, \infty)$. Further, we say that a stochastic process $\{Y_t,\ t \geqslant 0\}$ is $L^d$ bounded if there exists $K > 0$ such that $\mathrm{E}(|Y_t|^d) < K$, for all $t \geqslant 0$. Thereafter, let $I_n$ stand for the $n-$dimensional identity matrix and let $\mathbb{I}(A)$ denote the indicator function of the event $A$, $\boldsymbol{A}^\top$ denote the transpose of a matrix $\boldsymbol{A}$, let $\theta = (\theta_1^\top, \theta_2^\top)^\top \in \Theta \subset \mathbb{R}^{2(p+1)}$ with $\theta_j = (\mu_{j1}, \ldots, \mu_{jp}, \alpha_j)^\top$, for $j = 1, 2$. $\mu_{jk}$ is a real number with $k = 1, 2, \ldots, p$. Let $\varphi(t) = \big(\varphi_1(t), \varphi_2(t), \varphi_3(t), \ldots, \varphi_p(t)\big)$, $t \geqslant 0$ where, for each $i = 1, \ldots, p$, the function $\varphi_i(t)$ is a real-valued function of $t$. Let $\{B_t, t \geqslant 0\}$ be a one-dimensional standard Brownian motion defined on $(\Omega^*, \mathscr{F}, \mathrm{P})$ and let $\sigma > 0$. Let $\mathbb{N}_+ = \{1, 2, \ldots\}$ and let $||\cdot||$ represent Frobenius norm of a matrix.

The classical CIR model with one change-point is given by the stochastic differential equation (SDE):

$$\mathrm{d}x(t) = \big(\alpha_1(\beta_1 - x(t))\mathbb{I}(t \leqslant t^*) + \alpha_2(\beta_2 - x(t))\mathbb{I}(t > t^*)\big)\,\mathrm{d}t + \sigma\sqrt{x(t)}\mathrm{d}B_t,$$
$$x(0) > 0, \quad (22.1)$$

where $\beta_j$, $\alpha_j$, and $\sigma$ are parameters, as well as $t^* = sT$ with $s \in (0, 1)$ is the change-point. The parameter $\alpha_j$ represents the speed of adjustment to the mean $\beta_j$, and $\sigma$ is the volatility. An interest rate of zero is also precluded if the condition $2\alpha_j\beta_j \geqslant \sigma^2$ holds. The quantity $4\alpha_j\beta_j/\sigma^2$ plays a critical role in the behavior of the process. It is known as the *dimension* of the process $\{x(t), t \geqslant 0\}$.

Lyu (2023) has studied the inference problems in generalized CIR model without considering the structural changes. In this chapter, we consider observing the process $\{X(t), t \geqslant 0\}$ which is a solution of SDE as

$$\mathrm{d}X(t) = \big(S(\theta_1, t, X(t))\mathbb{I}(t \leqslant t^*) + S(\theta_2, t, X(t))\mathbb{I}(t > t^*)\big)\,\mathrm{d}t + \sigma\sqrt{X(t)}\mathrm{d}B_t,$$
$$0 \leqslant t \leqslant T, \quad (22.2)$$

where for $j = 1, 2$,

$$S(\theta_j, t, X(t)) = L_j(t) - \alpha_j X(t), \quad (22.3)$$

with $L_j(t) = \varphi(t)\theta_j$, $j = 1, 2$. Let $\{\mathscr{F}_t, t \geqslant 0\}$ denote the natural filtration of the Brownian motion given in the SDE defined in Eq. (22.2) up to time $t$. In this chapter, $\theta = (\theta_1^\top, \theta_2^\top)^\top$ is the target parameter and $t^*$ is the unknown nuisance parameter to be estimated. As mentioned above, we suppose that $t^* = sT$ where s is unknown with $0 < s < 1$.

Additionally, we assume that we have prior information concerning the parameter $\theta$. In particular, as in Nkurunziza and Zhang (2018); Lyu and Nkurunziza (2023), we consider the case where $\theta$ may or may not satisfy the following restriction:

$$\mathcal{H}_0 : M\,\theta = r \tag{22.4}$$

where, for a fixed $p$, $M$ is a known $q \times 2(p+1)$ full rank matrix with $q < 2(p+1)$, $r$ is a known $q$ column vector. This restriction leads to the hypothesis testing problem:

$$\mathcal{H}_0 : M\theta = r \quad \text{vs.} \quad \mathcal{H}_1 : M\theta \neq r. \tag{22.5}$$

**Example 22.1** Let $A \otimes B$ denote the Kronecker product between the matrices $A$ and $B$, suppose that $p = 4$, let $M = (1, -1) \otimes \boldsymbol{I}_5$ and let $r = 0$. The restriction in Eq. (22.4) corresponds the absence of a change-point, and thus the testing problem in Eq. (22.5) corresponds to testing the absence of a change-point. In this investigation, we only focus on shrinkage estimation method, and thus, the testing problem is beyond the scope of this paper. Nevertheless, such an interesting problem is part of our ongoing research.

As mentioned in Lyu (2023), we consider the following assumptions.

**Assumption 22.1** *The distribution of the initial value, $X_0$, of the SDE in Eq. (22.2) does not depend on the drift parameter $\theta$. Further, $X_0$ is positive a.s. and independent of $\{B_t : t \geqslant 0\}$ and $\mathrm{E}(|X_0|^m) < \infty$ for some $m \geqslant 2$.*

This assumption ensures that the starting point of the process is well-behaved (positive and at least $L^2$-bounded). We require that the distribution of the initial value does not depend on the parameter $\theta$ in order to simplify the likelihood function of the SDE. Note that in SDE, it is common to suppose that the initial value $X_0$ and the processes $\{B_t : t \geqslant 0\}$ are independent.

**Assumption 22.2** *For any $T > 0$, the base function $\varphi(t)$ is Riemann-integrable on $[0,\ T]$ and possesses:*

*I Periodicity: for some $\upsilon > 0$, $\varphi(t + \upsilon) = \varphi(t)$, for all $t \in [0, T]$.*
*II Orthogonality in $L^2([0, \upsilon], (1\ /\ \upsilon)d\lambda)$ : $\int_0^{\upsilon} \varphi^{\top}(t)\varphi(t)\mathrm{d}t = \upsilon I_p$.*

The periodicity and orthogonality conditions are typical in modeling phenomena with recurring cycles, such as seasonal trends in time series or oscillatory systems in physics. This assumption also ensures numerical stability and analytical tractability. To avoid negative values in the process $\{X(t), t \geqslant 0\}$, certain constraints on the parameters and the mean-reverting term $L_j(t)$ for $j = 1, 2$ of this GCIR model are necessary.

**Assumption 22.3** *The parameter $\alpha_j$, $j = 1, 2$ is positive, and the function $L_j(t)$ satisfies the condition $L_j(t) \geqslant \max\{\alpha_j\beta_j, \sigma^2\}$ for all $t \geqslant 0$. Further, we suppose that $m^* = 2 \min\limits_{\{i=1,2\}} \left\{\alpha_i\beta_i\ /\ \sigma^2\right\} \geqslant 2$.*

These constraints are designed to maintain positivity in $X(t)$, crucial in applications like finance (e.g., interest rates or stock prices) and population dynamics, where negative values would be nonsensical. The conditions on $\alpha_j$, $\beta_j$, and $\sigma^2$ provide a balance between the mean-reverting force and the variability introduced by randomness.

In this chapter, just like in Dehling et al. (2014), Nkurunziza and Zhang (2018), Lyu (2023), and other related references, we assume, without any loss of generality, that the period $\upsilon$ is known and equals to 1. As mentioned in the referenced paper, the base function $\varphi(t)$ is also supposed to be known. According to Proposition 3.2.1 in Lyu (2023), the GCIR model in Eq. (22.2) admits a strong and unique nonnegative solution on $[0, T]$, for $0 \leqslant T < \infty$.

**Proposition 22.1** *Suppose that Assumptions 22.1–22.3 hold, then the GCIR process in Eq. (22.2) admits a strong and unique nonnegative solution on* $[0, T]$ *for* $0 \leqslant T < \infty$.

This proposition guarantees the existence and uniqueness of the solution of the SDE in Eq. (22.2). It also guarantees that the solution remains nonnegative over the entire time interval. This is important because it ensures that the process does not take a negative value, which is often a necessary property for models involving quantities like interest rates, stock prices, or physical quantities (e.g., concentrations or temperatures) that cannot realistically take negative values. The proof of this result follows from *Yamada-Watanabe theorem* (Röckner et al., 2008). As stated above, the solution is nonnegative. Further, by Assumption 22.3, an interest rate of zero is precluded. The expression of the process $\{X(t), t \geqslant 0\}$ as the solution to the SDE in Eq. (22.2) is given in the following proposition.

**Proposition 22.2** *Suppose that Assumptions 22.1–22.3 hold. Then, the solution of SDE in Eq. (22.2) is given by*

$$X(t) = X_1(t)\mathbb{I}(t \leqslant t^*) + X_2(t)\mathbb{I}(t > t^*)$$

*with*

$$X_1(t) = \exp\{-\alpha_1 t\}X_0 + r_1(t) + \tau_1(t),$$

$$X_2(t) = \exp\{-\alpha_2(t - t^*)\}X^{t^*}(0) + r_2^{t^*}(t - t^*) + \tau_2^{t^*}(t - t^*)$$

*where for* $0 \leqslant t \leqslant t^*$,

$$r_1(t) = \exp\{-\alpha_1 t\} \int_0^t \exp\{\alpha_1 s\}L_1(s)\mathrm{d}s$$

$$\tau_1(t) = \sigma \exp\{-\alpha_1 t\} \int_0^t \exp\{\alpha_1 s\}\sqrt{X_1(s)}\,\mathrm{d}B_s, \tag{22.6}$$

*and for* $t^* \leqslant t \leqslant T$,

$$r_2^{t^*}(t) = \exp\{-\alpha_2 t\} \int_0^t \exp\{\alpha_2 s\}L_2(s + t^*)\mathrm{d}s$$

$$\tau_2^{t^*}(t) = \sigma \exp\{-\alpha_2 t\} \int_0^t \exp\{\alpha_2 s\}\sqrt{X(s)}\mathrm{d}B_s^{t^*}$$

$$X^{t^*}(t) = X(t + t^*)$$

$$\varphi^{t^*}(t) = \varphi(t + t^*)$$

$$B_t^{t^*} = B(t + t^*) - B(t^*) \qquad t \geqslant 0,\ t^* \geqslant 0.$$

The proof of this proposition can be found in Appendix "Proof of Proposition 22.2".

The solution describes how the process $X(t)$ behaves over time. For the first part of the time interval (before $t^*$), it follows a process $X_1(t)$ influenced by both deterministic and random factors. After time $t^*$, the process switches to a new form $X_2(t)$, with the previous value $X_1(t^*)$ serving as its starting point. The process continues to evolve with both deterministic and stochastic terms, accounting for its past behavior and randomness.

As in Nkurunziza and Zhang (2018), Lyu (2023), to get the likelihood function of a diffusion process, it is sufficient to guarantee that

$$\mathrm{P}\left(\int_0^T S^2(\theta, t, X(t)) \Big/ (\sigma^2 X(t))\mathrm{d}t < \infty\right) = 1, \quad \forall T \in [0, \infty], \forall \theta \in \Theta.$$

In the following, we derive this sufficient condition for the likelihood function to exist.

**Corollary 22.1** *If Assumptions 22.1–22.3 hold, then*

$$\mathrm{P}\left(\int_0^T \frac{S^2(\theta, t, X(t))}{\sigma^2 X(t)}\mathrm{d}t < \infty\right) = 1, \qquad \forall T \in [0, \infty], \forall \theta \in \Theta$$

***Proof*** Using Eq. (22.3) leads to $S(\theta_j, t, X(t)) = L_j(t) - \alpha_j X(t)$ and $L_j(t) = \sum_{i=1}^p \mu_{ij}\varphi_i(t)$, for $j = 1, 2$. Then, together with Assumptions 22.1–22.3

$$\left(\frac{S(\theta_j, t, X(t))}{\sigma\sqrt{X(t)}}\right)^2 < \frac{L_j^2(t)}{\sigma^2 X(t)} + \frac{X(t)}{\sigma^2}$$

where, from Assumption 22.3 and

$$\sup_{t\geqslant 0} \mathrm{E}\left[X^{-\Lambda}(t)\right] \leqslant \sup_{t\geqslant 0} \mathrm{E}\left[x^{-\Lambda}(t)\right] < \infty, \quad \text{for } -\Lambda \in \left[-m^*, \infty\right), \tag{22.7}$$

we have $\sup_{t\geqslant 0} \mathrm{E}\left[1/X(t)\right] < \infty$. Further, we have $\sup_{t\geqslant 0} \mathrm{E}[X(t)] < \infty$, which implies that

$$\sup_{t\geqslant 0} \mathrm{E}\left(\left(|S(\theta_1, t, X(t))|^2 + |S(\theta_2, t, X(t))|^2\right) \Big/ \left|\sigma^2 X(t)\right|\right) < \infty.$$

Then,

$$\mathrm{P}\left(\int_0^T \left(|S(\theta_1, t, X(t))|^2 + |S(\theta_2, t, X(t))|^2\right) \Big/ \left|\sigma^2 X(t)\right| \, \mathrm{d}t < \infty\right) = 1,$$
$$\forall T \in [0, \infty).$$

This completes the proof. □

Based on Proposition 22.2, we obtain the following result which is crucial in deriving the likelihood function of the GCIR model given by Eq. (22.2) when dealing with a known change-point. In this context, we assume, as an intermediate step, that the change-point $t^* = sT$ is known, i.e., that the location parameter $s$ is known. For $j = 1, 2$, let $\sum_{i=1}^{p} |\mu_{ij}| \leqslant K_\mu$ and $||\varphi(t)|| \leqslant K_\varphi$, for $j = 1, 2$, and some positive constants $K_\mu$, $K_\varphi$; then,

$$0 < L_j(t) \leqslant K_\varphi K_\mu. \tag{22.8}$$

Further, let

$$K_1 = \mathrm{E}\left[X_0^2\right] + \frac{K_\varphi K_\mu}{\alpha_1^2} + \sigma^2 \mathrm{E}[X_0] \frac{1}{4\alpha_1} + \sigma^2 K_\varphi K_\mu \left(\frac{1}{4\alpha_1^2}\right) + \mathrm{E}[X_0] K_\varphi K_\mu \frac{1}{2\alpha_1}, \tag{22.9}$$

and

$$K_2 = K_1 + \frac{K_\varphi^2 K_\mu^2}{\alpha_2^2} + \sigma^2 \sqrt{K_1} \frac{1}{4\alpha_2} + \sigma^2 K_\varphi K_\mu \left(\frac{1}{2\alpha_2^2}\right) + \sqrt{K_1} K_\varphi K_\mu \frac{1}{2\alpha_2}. \tag{22.10}$$

We have the following proposition which shows that the solution of the SDE in Eq. (22.2) is bounded.

**Proposition 22.3** *If Assumptions 22.1–22.3 hold, then* $\sup_{t \geqslant 0} \mathrm{E}[X^2(t)] < \infty$.

The proof of this proposition follows from the fact that Itô integrals are local martingales (see Panik, 2017, p. 119, Section 4.2)). We also refer to Lyu (2023). Moreover, let

$$C_m = m\left(K_\mu K_\varphi + \frac{1}{2}\sigma^2(m-1)\right),$$

$$K_{m1} = \mathrm{E}[X_0^m] + \max\left\{1, \sqrt{K_1 + K_2}\right\} \times \sum_{j=1}^{m} \left(\prod_{k=1}^{j} C_{m-(k-1)} \, \mathrm{E}[X_0^{m-(k-1)}] \frac{1}{\alpha_1(m-(k-1))}\right), \tag{22.11}$$

$$K_{m2} = K_{m1} + \max\{1, \sqrt{K_1 + K_2}\} \sum_{j=1}^{m} \left( \prod_{k=1}^{j} C_{m-(k-1)} K_{m1}^{\frac{m-(k-1)}{m}} \frac{1}{\alpha_2(m-(k-1))} \right) \tag{22.12}$$

**Proposition 22.4** *Let* $K_m = \max\{K_{m1}, K_{m2}\}$, *if Assumptions 22.1–22.3 hold, then*

$$\sup_{t \geqslant 0} \mathrm{E}[X^m(t)] \leqslant K_m. \tag{22.13}$$

The proof is thoroughly explained in Appendix "Proof of Proposition 22.4".

**Remark 22.1** Let $t^* = sT$. By Jensen's inequality, it is obvious that

$$\mathrm{E}\left[\frac{1}{X(t)}\right] \geqslant \frac{1}{\mathrm{E}[X(t)]} = \begin{cases} 1 \Big/ \left(\exp\{-\alpha_1 t\}\, \mathrm{E}[X_0] + h_1(t)\right) & \text{if } t \in [0, sT], \\ 1 \Big/ \left(\exp\{-\alpha_2(t-t^*)\}\, \mathrm{E}[X_0^{t^*}] + h_2(t)\right) & \text{if } t \in [sT, T]. \end{cases}$$

One of the main challenges we face is that the process $\{X(t), t \geqslant 0\}$ is not stationary, except in the special case where the dimension of the GCIR process is a positive integer. Moreover, there is no explicit expression for the solution $\{X(t), t \geqslant 0\}$. To overcome this difficulty, we introduce an auxiliary process below, which is strictly stationary and ergodic. Additionally, we prove that the distance between the auxiliary process and the solution of the SDE in Eq. (22.2) converges to 0 both in $L_1$ and almost surely. This convergence allows us to derive the asymptotic distributions of the UMLE and RMLE by utilizing the ergodicity of the auxiliary process. The limiting distributions of both UMLE and RMLE are crucial in testing the imposed restrictions.

Let the process

$$\left\{ X^*(t) = (X^{(1*)}(t), X^{(2*)}(t)) : t \geqslant 0 \right\},$$

and

$$\begin{aligned} X^{(1*)}(t) &= X_1^{(1*)}(t)\mathbb{I}(0 \leqslant t \leqslant t^*) + X_2^{(1*)}(t)\mathbb{I}(t^* < t) \\ X^{(2*)}(t) &= X_1^{(2*)}(t)\mathbb{I}(0 \leqslant t \leqslant t^*) + X_2^{(2*)}(t)\mathbb{I}(t^* < t) \end{aligned} \tag{22.14}$$

satisfy the following SDEs

$$\begin{aligned} \mathrm{d}X^{(1*)}(t) &= \left( \left( \frac{\sigma^2}{2} - \alpha_1 X^{(1*)}(t) \right) \mathbb{I}(0 \leqslant t \leqslant t^*) \right. \\ &\quad \left. + \left( \frac{\sigma^2}{2} - \alpha_2 X^{(1*)}(t) \right) \mathbb{I}(t^* < t) \right) \mathrm{d}t + \sigma \sqrt{X^{(1*)}(t)} d B_t^{(1*)}; \quad (22.15) \\ \mathrm{d}X^{(2*)}(t) &= \left( \left( L_1(t) - \frac{\sigma^2}{2} - \alpha_1 X^{(2*)}(t) \right) \mathbb{I}(0 \leqslant t \leqslant t^*) \right. \\ &\quad \left. + \left( L_2(t) - \frac{\sigma^2}{2} - \alpha_2 X^{(2*)}(t) \right) \mathbb{I}(t^* < t) \right) \mathrm{d}t \\ &\quad + \sigma \sqrt{X^{(2*)}(t)} d B_t^{(2*)}; \quad (22.16) \end{aligned}$$

where the initial value $X^{(1*)}(0) = X_0/2$, $X^{(2*)}(0) = X_0/2$, and $B_t^{(*)} = (B_t^{(1*)}, B_t^{(2*)})$ is a two-dimensional standard Brownian motion defined on the probability space $(\Omega^*, \mathscr{F}, \mathrm{P})$. Proposition 22.1 indicates that SDEs in Eqs. (22.15) and (22.16) admit strong solutions. We establish the following proposition that showcases the relationship between the processes $X^{(1*)}(t)$, $X^{(2*)}(t)$ and $X(t)$, where $X(t)$ is the solution to SDE in Eq. (22.2).

**Proposition 22.5** *Let $X^{(1*)}(t)$ and $X^{(2*)}(t)$ be the strong solution to SDE in Eqs. (22.15) and (22.16), respectively. For $t \geqslant 0$, $X^{(1*)}(t) + X^{(2*)}(t)$ is the strong solution to SDE in Eq. (22.2), i.e., $X^{(1*)}(t) + X^{(2*)}(t) = X(t)$ almost surely.*

The proof of this result is similar to the ones given in Lyu (2023).

Essentially, by combining the two processes $\{X^{(1*)}(t),\ t \geqslant 0\}$ and $\{X^{(2*)}(t),\ t \geqslant 0\}$, we get a new process $\{X(t),\ t \geqslant 0\}$ that behaves according to the combined dynamics described by SDE in Eq. (22.2). Further, for the convenience of the reader, we also outline the proof in Appendix "Proof of Proposition 22.5". In deriving the main result of this paper, we leverage the significance of the following proposition. This proposition plays a crucial role in establishing the subsequent result, demonstrating that its solution is positive with probability 1.

**Proposition 22.6** *Suppose that Assumptions 22.1–22.3 hold, then, for all $w \in \Omega^*$, we have $\int_0^T \mathbb{I}(A_t)(w)\mathrm{d}t = 0$ a.s. for $0 \leqslant T < \infty$, where $A_t(w) = \{\omega : X^{(2*)}(t, w) = 0, t \geqslant 0\}$.*

The proof is similar to the proof of Proposition 3.2.5 in Lyu (2023).

We also derive the following proposition which is useful in constructing a stationary and ergodic auxiliary process.

**Proposition 22.7** *Suppose that Assumptions 22.1–22.3 hold, then the solution of SDE defined in Eq. (22.15) is given as* $X^{(1*)}(t) = \sum_{j=1}^{2} Y_j^2(t)$, *where* $Y_j(t) = Y_{1j}(t)\mathbb{I}(0 < t \leqslant t^*) + Y_{2j}(t)\mathbb{I}(t^* < t \leqslant T)$, *and*

$$
\begin{aligned}
Y_{1j}(t) &= \exp\left\{-\frac{\alpha_1}{2}t\right\}Y_{1j}(0) + \frac{1}{2}\sigma\exp\left\{-\frac{\alpha_1}{2}t\right\}\int_0^t \exp\left\{\frac{\alpha_1}{2}s\right\}\mathrm{d}B_j(s),\\
Y_{2j}(t) &= \exp\left\{-\frac{\alpha_2}{2}(t-t^*)\right\}Y_{2j}^{t^*}(0)\\
&\quad + \frac{1}{2}\sigma\exp\left\{-\frac{\alpha_2}{2}(t-t^*)\right\}\int_0^{(t-t^*)} \exp\left\{\frac{\alpha_2}{2}s\right\}\mathrm{d}B_j^{t^*}(s),
\end{aligned}
$$

*where* $Y_{2j}^{t^*}(0) = Y_{1j}(t^*)$. *Further, let*

$$
\begin{aligned}
\bar{\bar{Y}}_1(t) &= \exp\left\{-\frac{\alpha_1}{2}t\right\}\sqrt{X^{(2*)}(0)}\\
&\quad + \exp\left\{-\frac{\alpha_1}{2}t\right\}\int_0^t \exp\left\{\frac{\alpha_1}{2}s\right\}\frac{4L_1(s) - 3\sigma^2}{8\sqrt{X^{(2*)}(s)}}\mathrm{d}s\\
&\quad + \frac{\sigma}{2}\exp\left\{-\frac{\alpha_1}{2}t\right\}\int_0^t \exp\left\{\frac{\alpha_1}{2}s\right\}\mathrm{d}B_s{}^{(2*)},
\end{aligned}
$$

*and*

$$
\begin{aligned}
\bar{\bar{Y}}_2(t) &= \exp\left\{-\frac{\alpha_2}{2}(t-t^*)\right\}\sqrt{(X^{(2*)}(0))^{t^*}}\\
&\quad + \exp\left\{-\frac{\alpha_2}{2}(t-t^*)\right\}\int_0^{t-t^*} \exp\left\{\frac{\alpha_2}{2}s\right\}\frac{4L_2(s+t^*) - 3\sigma^2}{8\sqrt{X^{(2*)}(s)}}\mathrm{d}s\\
&\quad + \frac{\sigma}{2}\exp\left\{-\frac{\alpha_2}{2}(t-t^*)\right\}\int_0^{t-t^*} \exp\left\{\frac{\alpha_2}{2}s\right\}\mathrm{d}(B_s^{(2*)})^{t^*},
\end{aligned}
$$

*where* $(X^{(2*)}(0))^{t^*} = X^{(2*)}(t^*)$.

*Further, the solution of SDE in Eq. (22.16) can be rewritten as* $X^{(2*)}(t) = (\bar{\bar{Y}}_1(t)\mathbb{I}(t^* < t \leqslant T) + \bar{\bar{Y}}_2(t)\mathbb{I}(t^* < t \leqslant T))^2$.

The proof of this result is similar to the one given for Proposition 3.2.6 in Lyu (2023).

This proposition essentially shows how to break down the solution of the two SDEs into more manageable parts, each governed by simpler dynamics. This decomposition is especially useful for studying the long-term behavior of the system, enabling the construction of a stationary and ergodic auxiliary process that can better reflect the underlying structure of the system's behavior over time.

## 22.3 Estimation in the Case of One Known Change-Point

In this section, we derive both the UMLE and the RMLE. Going beyond derivation, we focus on establishing the joint asymptotic normality of these estimators. This involves a comprehensive exploration of the statistical properties, ensuring a robust understanding of their behavior in various scenarios. Through this chapter, we not only unravel the intricacies of individual estimators but also shed light on their collective behavior, providing a deeper insight into their reliability and performance in statistical modeling.

### *22.3.1 The UMLE $\hat{\theta}_T$ and the RMLE $\tilde{\theta}_T$*

Let $\mathrm{P}_{X^T}^{(\theta)}$ denote the measure induced by the observable realizations $X^T = \{X(t),\ t \geqslant 0\}$ on the measurable space $(C_{[0,T]}, \mathcal{B}_{[0,T]})$, where $C_{[0,T]}$ is the space of continuous, real-valued functions on $[0, T]$ and $\mathcal{B}_{[0,T]}$ is the associated Borel $\sigma$-algebra. Further, let $\mathrm{P}_B$ be the measure generated by the Brownian motion on $(C_{[0,T]}, \mathcal{B}_{[0,T]})$. Then, the Radon-Nikodym derivative of observations $X^T$ is given by

$$\mathcal{L}(\theta, X^T) = \frac{\mathrm{dP}_{X^T}^{(\theta)}}{\mathrm{dP}_B(X^T)}.$$

Thus, the UMLE is derived by minimizing the functional $\theta \longmapsto \mathcal{L}(\theta, X^T)$, i.e., $\hat{\theta}_T = \mathrm{argmax}_\theta\, \mathcal{L}(\theta, X^T)$.

To simplify some notations, let

$$Q_{[a,b]} = \begin{bmatrix} \int_a^b \frac{\varphi^\top(t)\varphi(t)}{X(t)}\mathrm{d}t & -\int_a^b \varphi^\top(t)\mathrm{d}t \\ -\int_a^b \varphi(t)\mathrm{d}t & \int_a^b X(t)\mathrm{d}t \end{bmatrix}_{(p+1)\times(p+1)}, \quad 0 < a < b \tag{22.17}$$

$$R_{[a,b]} = \left(\int_a^b \frac{\varphi(t)}{X(t)}\mathrm{d}X(t), -\int_a^b \frac{X(t)}{X(t)}\mathrm{d}X(t)\right)^\top,$$

$$W_{[a,b]} = \left(\int_a^b \frac{\varphi(t)}{\sqrt{X(t)}}\mathrm{d}B_t, -\int_a^b \frac{X(t)}{\sqrt{X(t)}}\mathrm{d}B_t\right)^\top, \tag{22.18}$$

and

$$Q(s, T) = \begin{bmatrix} Q_{[0,sT]} & 0 \\ 0 & Q_{[sT,T]} \end{bmatrix},$$

$$R(s,T) = \begin{bmatrix} R_{[0,sT]} \\ R_{[sT,T]} \end{bmatrix}, \tag{22.19}$$

$$W(s,T) = \begin{bmatrix} W_{[0,sT]} \\ W_{[sT,T]} \end{bmatrix}.$$

For the purpose of deriving the UMLE and RMLE, we need to ensure that the matrix $Q_{[0,T]}$ is invertible. In the following proposition, we prove that the matrix $Q_{[0,T]}$ is positive definite provided that $T \geqslant 1$.

**Proposition 22.8** *If Assumptions 22.1–22.3 hold, then $Q(s,T)$ is a positive definite matrix provided that $T \geqslant 1/\min\{s, 1-s\}$.*

The proof of this proposition is given in Appendix "Proof of Proposition 22.8".

Given that the proposed method's optimality is asymptotic, going forward, we assume, without loss of generality, that the condition $T \geqslant 1/\min\{s, 1-s\}$ holds. Let $\tilde{\theta}_T(s)$ be the RMLE and $G(s,T) = Q^{-1}(s,T)M^\top(MQ^{-1}(s,T)M^\top)^{-1}$.

**Proposition 22.9** *If Assumptions 22.1–22.3 hold along with Eq. (22.5), then $\hat{\theta}_T(s) = Q^{-1}(s,T)R(s,T)$ and $\tilde{\theta}_T(s) = \hat{\theta}_T(s) - G(s,T)(M\hat{\theta}_T(s) - r)$.*

The proof of this proposition is detailed in Appendix "Proof of Proposition 22.9".

In essence, this mathematical setup ensures that the statistical estimation methods can be applied correctly and optimally, provided that certain conditions on the time horizon $T$ are met. The positive definiteness of $Q(s,T)$ guarantees that these methods work well in practice.

### 22.3.2 Auxiliary Stationary and Ergodic Process

The auxiliary process is constructed in the case of integer dimension in Appendix "Some Important Results with Auxiliary Processes", and general dimensions are given below, respectively. The stationarity and ergodicity of the special auxiliary process are proved in Appendix "Some Important Results with Auxiliary Processes". In the following, we give proof of the stationarity and ergodicity in the general case.

Let $\tilde{X}^{(1*)}(t) = \tilde{X}_1^{(1*)}(t) + \tilde{X}_2^{(1*)}(t)$, with

$$\tilde{X}_1^{(1*)}(t) = \sum_{j=1}^{2} \left(\tilde{Y}_{1j}(t)\mathbb{I}(0 < t \leqslant t^*)\right)^2; \tilde{X}_2^{(1*)}(t) = \sum_{j=1}^{2} \left(\tilde{Y}_{2j}(t)\mathbb{I}(t^* < t \leqslant T)\right)^2$$

where $\tilde{Y}_{1j}(t)$ and $\tilde{Y}_{2j}(t)$ are defined in Appendix "Some Important Results with Auxiliary Processes".

Further, let

$$\mathcal{V}_1(s) = \frac{4L_1(s) - 3\sigma^2}{8}\mathbb{I}(s \leqslant 0) + \frac{4L_1(s) - 3\sigma^2}{8\sqrt{X^{(2*)(s)}\mathbb{I}(0 < t \leqslant t^*)}}\mathbb{I}(s > 0)\,,$$

$$\mathcal{V}_2(s) = \frac{4L_2(s) - 3\sigma^2}{8}\mathbb{I}(s \leqslant 0) + \frac{4L_2(s) - 3\sigma^2}{8\sqrt{X^{(2*)(s)}\mathbb{I}(0 < t \leqslant t^*)}}\mathbb{I}(s > 0)\,.$$

In addition, let $\{B_s^{(2*)}, s \geqslant 0\}$ and $\{\bar{B}_s^{(2*)}, s \geqslant 0\}$ be two independent Brownian motions and $\tilde{B}_s^{(2*)}$ be a bilateral Brownian motion with

$$\tilde{B}_s^{(2*)} = B_s^{(2*)}\mathbb{I}(s > 0) + \bar{B}_{-s}^{(2*)}\mathbb{I}(s \leqslant 0).$$

Let

$$\tilde{\tilde{Y}}_1(t) = \exp\left\{-\frac{\alpha_1}{2}t\right\}\int_{-\infty}^{t}\exp\left\{\frac{\alpha_1}{2}s\right\}\mathcal{V}_1(s)ds + \frac{\sigma}{2}\exp\left\{-\frac{\alpha_1}{2}t\right\}\int_{-\infty}^{t}\exp\left\{\frac{\alpha_1}{2}s\right\}d\tilde{B}_s^{(2*)},$$

$$\tilde{\tilde{Y}}_2(t) = \exp\left\{-\frac{\alpha_2}{2}t\right\}\int_{-\infty}^{t}\exp\left\{\frac{\alpha_2}{2}s\right\}\mathcal{V}_2(s)ds + \frac{\sigma}{2}\exp\left\{-\frac{\alpha_2}{2}t\right\}\int_{-\infty}^{t}\exp\left\{\frac{\alpha_2}{2}s\right\}d\tilde{B}_s^{(2*)},$$

and $\tilde{X}_1^{(2*)}(t) = \tilde{\tilde{Y}}_1^2(t)\mathbb{I}(0 < t \leqslant t^*)$; $\tilde{X}_2^{(2*)}(t) = \tilde{\tilde{Y}}_2^2(t)\mathbb{I}(t^* < t \leqslant T)$, *and* $\tilde{X}^{(2*)}(t) = \tilde{X}_1^{(2*)}(t) + \tilde{X}_2^{(2*)}(t)$. We have

**Proposition 22.10** *If Assumptions 22.1–22.3 hold, then for each* $i = 1, 2$*, the stochastic process* $\tilde{X}_i^{(2*)}(t+k-1), t \in (0, 1]\}_{k\in\mathbb{N}_+}$ *is strictly stationary and ergodic.*

The proof of this result is similar to that given for Proposition 3.3.3 in Lyu (2023). In short, crucial step consists in the use of the extension of Dubins-Schwarz theorem along with the fact that Itô integrals are local martingales (Panik, 2017, p. 119, Section 4.2). In the sequel, let $\tilde{X}_i(t) = (\tilde{X}_i^{(1*)}(t) + \tilde{X}_i^{(2*)}(t))$, for $i = 1$. Below, we prove that this is a stationary process.

**Proposition 22.11** *If Assumptions 22.1–22.3 hold, then for* $i = 1, 2$ $\left\{\tilde{X}_i(t + k - 1), 0 < t \leqslant 1\right\}_{k=1}^{\infty}$ *is stationary and ergodic.*

***Proof*** From the definition, it is clear that $\tilde{Y}_{1j}(t)$ and $\tilde{Y}_{2j}(t)$ are independent with $\tilde{Y}_{ij}(t) \sim \mathcal{N}\left(0, \sigma^2 / 4\alpha_i\right)$, $i = 1, 2$. This implies that

$$\tilde{X}_1^{(1*)}(t) \sim \frac{\sigma^2}{4\alpha_1}\chi_2^2 \quad \tilde{X}_2^{(1*)}(t) \sim \frac{\sigma^2}{4\alpha_2}\chi_2^2$$

and then $\{\tilde{X}^{(1*)_i}(t), t \geqslant 0\}$ is strictly stationary and ergodic. Proposition 22.10 implies that $\{\tilde{X}_i^{(2*)}(t+k-1), t \in (0,1]\}_{k\in\mathbb{N}_+}$ is strictly stationary and ergodic; we conclude that $\tilde{X}_i(t+k-1)$ is strictly stationary and ergodic. □

In simple terms, these propositions tell us that the processes in question behave predictably over time (stationary) and that their long-term behavior can be inferred from a single realization of the process (ergodic).

By using this proposition, we derive the asymptotic normality of the estimators. As an intermediate result, we derive the following proposition to state the relation between the auxiliary process and the solution of the SDE in Eq. (22.2).

**Proposition 22.12** *If Assumptions 22.1–22.3 hold, then:*

*(1)* $\tilde{X}(t) - X(t) \xrightarrow[t\to\infty]{a.s.} 0$,

*(2)* $\tilde{X}(t) - X(t) \xrightarrow[t\to\infty]{L^1} 0$.

The proof of this proposition is similar to the one given for Proposition 3.3.4 in Lyu (2023).

This proposition indicates that the two processes eventually become almost identical in the long run, both in terms of individual paths (almost surely) and in terms of their average behavior (in $L^1$).

**Proposition 22.13** *If Assumptions 22.1–22.3 hold, then* $\sup_{t\geqslant 0} \mathrm{E}\left[\left|\tilde{X}(t)\right|\right] < \infty$.

***Proof*** Since we have

$$\left|\tilde{X}(t)\right| = \left|\tilde{X}(t) - X(t) + X(t)\right| \leq \left(\left|\tilde{X}(t) - X(t)\right| + |X(t)|\right).$$

Then,

$$\sup_{t\geq 0} \mathrm{E}\left[\left|\tilde{X}(t) - X(t)\right|\right] \leqslant \sup_{t\geq 0}\Bigg(\mathrm{E}\left[\left|\tilde{X}^{(1*)}(t) - X^{(1*)}(t)\right|\right] + \mathrm{E}\left[\left|\tilde{X}^{(2*)}(t) - X^{(2*)}(t)\right|\right]\Bigg).$$

Equation (22.57) implies that $\sup_{t\geq 0}\mathrm{E}\left[\left|\tilde{X}^{(1*)}(t)-X^{(1*)}(t)\right|\right]<\infty$, and one verifies that $\sup_{t\geq 0}\mathrm{E}\left[\left|\tilde{X}^{(2*)}(t)-X^{(2*)}(t)\right|\right]<\infty$. Further, by Proposition 22.4, $\sup_{t\geq 0}\mathrm{E}[X^m(t)]\leq K_m$. Then, we have:

$$\sup_{t\geq 0}\mathrm{E}\left[\left|\tilde{X}(t)\right|\right]\leq\left(\sup_{t\geq 0}\mathrm{E}\left[\left|\tilde{X}(t)-X(t)\right|\right]+\sup_{t\geq 0}\mathrm{E}\left[|X(t)|\right]\right)<\infty.$$

This completes the proof. □

In practical terms, this can be interpreted as the process is "well-behaved" over time in terms of its expected magnitude. To leverage the stationarity and ergodicity of the auxiliary process, it is imperative to verify the validity of the following proposition.

**Proposition 22.14** *If Assumptions 22.1–22.3 hold, then* $0<\sup_{t\geqslant 0}\mathrm{E}\left[1\big/\tilde{X}(t)\right]<\infty$.

***Proof*** We have $\tilde{X}_1(0)=\tilde{X}_1^{(1*)}(0)+\tilde{X}_1^{(2*)}(0)\sim(\sigma^2/4\alpha_1)\chi^2(3,\Delta^*)$. This implies that $\mathrm{E}\left[1/\tilde{X}_1(0)\right]<4\alpha_1/\sigma^2<\infty$. Further, for $0<t\leqslant 1$, by Fatou's lemma, we have that

$$\begin{aligned}\mathrm{E}\left[1/\tilde{X}(t)\right]&=\mathrm{E}\left[\lim_{k\to\infty}1/X(t+k-1)\right]\\&\leqslant\liminf_{k}\mathrm{E}\left[1/X(t+k-1)\right]\\&\leqslant\sup_{t\geqslant 0}\mathrm{E}\left[1/X(t)\right].\end{aligned}$$

From Assumption 22.3, $m^*=2\min\left\{\alpha_1\beta_1/\sigma^2,\alpha_2\beta_2/\sigma^2\right\}$, and $m^*\geqslant 2$. By Proposition 4.1 in Chernova et al. (2024), we have $\sup_{t\geqslant 0}\mathrm{E}\left[x^{-\Lambda}(t)\right]<\infty$ for $-\Lambda\in\left[-m^*,\infty\right)$. By Proposition 22.27 and Assumption 22.3, we have that $x(t)\leqslant X(t)$ a.s. which implies that

$$\sup_{t\geqslant 0}\mathrm{E}\left[X^{-\Lambda}(t)\right]\leqslant\sup_{t\geqslant 0}\mathrm{E}\left[x^{-\Lambda}(t)\right]<\infty,\quad\text{for}\quad-\Lambda\in\left[-m^*,\infty\right),\tag{22.20}$$

and this implies that $\sup_{t\geqslant 0}\mathrm{E}\left[1/X(t)\right]<\infty$. Then, we get $\sup_{t\geqslant 0}\mathrm{E}\left[1/\tilde{X}(t)\right]<\infty$. This completes the proof. □

The importance of this proposition consists in the fact that the processes that get too close to zero can lead to numerical instability or unrealistic behavior. This proposition ensures that $\tilde{X}(t)$ behaves in a stable and controlled manner, which is crucial for reliable modeling and analysis.

**Proposition 22.15** *Suppose that Assumptions 22.1–22.3 hold, then* $\tilde{X}^{-1}(t) - X^{-1}(t) \xrightarrow[t\to\infty]{a.s.} 0$.

***Proof*** The proof follows from Proposition 22.28. We also refer to Lyu (2023) for further details. □

This result is useful when comparing two related processes. It ensures that, in the long run, the two processes behave in a nearly identical way. Proposition 22.14 guarantees that the first negative moment of the process $\left\{\tilde{X}(t+k-1), 0 < t \leqslant 1\right\}_{k\in\mathbb{N}_+}$ exists. Further, from Proposition 22.11, $\left\{\tilde{X}(t+k-1), 0 < t \leqslant 1\right\}_{k=1}^{\infty}$ is ergodic, then, by *Birkhoff Ergodic theorem*, for $t \in (0, 1]$,

$$\lim_{L\to\infty} \frac{1}{L} \sum_{k=1}^{L} 1/\tilde{X}(t+k-1) = \mathrm{E}\left[1/\tilde{X}(t)\right] \text{ a.s.}$$

From Proposition 22.14, for $s \geqslant 0$, $\sup_{s\geqslant 0} \mathrm{E}\left[1/\tilde{X}(s)\right] < \infty$. Further, due to the convexity of the function $y = 1/x$ on the interval $(0, \infty)$, by Jensen's inequality, $\mathrm{E}\left[1/\tilde{X}(t)\right] \geqslant 1/\mathrm{E}\left[\tilde{X}(t)\right]$. The equality "=" holds if and only if $\tilde{X}(t)$ is a constant with probability 1. Since $\tilde{X}(t)$ is a continuous random variable for all $t \geqslant 0$, we get $\mathrm{E}\left[1/\tilde{X}(t)\right] > 1/\mathrm{E}\left[\tilde{X}(t)\right] > 0$. Finally,

$$0 < 1/\mathrm{E}\left[\tilde{X}(t)\right] < \mathrm{E}\left[1/\tilde{X}(t)\right] \leqslant \sup_{t\geqslant 0} \mathrm{E}\left[1/\tilde{X}(t)\right] < \infty, \quad \forall t \geqslant 0. \tag{22.21}$$

The relation (22.21) plays a crucial role in proving the asymptotic normality of the estimators. In the following, we introduce an auxiliary process, which is stationary and ergodic.

### 22.3.3 Joint Asymptotic Normality of the UMLE and RMLE

In the following, we derive the joint asymptotic normality of UMLE, $\hat{\theta}_T(s)$, and RMLE, $\tilde{\theta}_T(s)$. This result is used in deriving the asymptotic distribution of a test statistic for the hypothesis testing problem defined in Eq. (22.5).

As in Lyu and Nkurunziza (2023), from Proposition 22.9, we have $\sqrt{T}(\hat{\theta}_T(s) - \theta) = \sigma T Q^{-1}(s, T) W(s, T)/\sqrt{T}$. Thus, as in Lyu and Nkurunziza (2023), the asymptotic properties of these estimators rely on the asymptotic behavior of

$Q(s,T)/\sqrt{T}$ and $W(s,T)/\sqrt{T}$ as well as on the non-singularity of the matrices. Let

$$\Sigma = \begin{bmatrix} s\Sigma_1 & 0 \\ 0 & (1-s)\Sigma_2 \end{bmatrix},$$

$$\Sigma_k = \begin{bmatrix} \int_0^1 \varphi^\top(t)\varphi(t)\,\mathrm{E}\left[\frac{1}{\tilde{X}_k(t)}\right]\mathrm{d}t & -\int_0^1 \varphi^\top(t)\mathrm{d}t \\ -\int_0^1 \varphi(t)\mathrm{d}t & \int_0^1 \mathrm{E}\left[\tilde{X}_k(t)\right]\mathrm{d}t \end{bmatrix}_{(p+1)\times(p+1)}, k = 1, 2. \tag{22.22}$$

**Proposition 22.16** *The matrix $\Sigma$ is a positive definite matrix.*

***Proof*** The proof is similar to Proposition 3.3.8 in Lyu (2023). □

Furthermore, it is essential to underscore the significance of the proposed proposition within the framework of deriving the limiting distribution. The establishment of this proposition is paramount for the deduction of the limiting distribution associated with the presented estimators. Serving as a pivotal element, this proposition facilitates an understanding of the behavior of the estimators in the context of an infinitely increasing sample size. Additionally, the proposition assumes a fundamental role in elucidating the evolving behavior of the estimators as time tends toward infinity. It provides valuable insights into the behavior and properties of the estimators as they are applied in scenarios where the time $T$ is large enough.

**Proposition 22.17** *If Assumptions 22.1–22.3 hold, then*

$$\frac{1}{T}Q_{[0,sT]} \xrightarrow[T\to\infty]{a.s.\ and\ L^{\min\{m^*/2,m/2\}}} s\Sigma_1,$$

$$\frac{1}{T}Q_{[sT,T]} \xrightarrow[T\to\infty]{a.s.\ and\ L^{\min\{m^*/2,m/2\}}} (1-s)\Sigma_2,$$

$$TQ^{-1}_{[0,sT]} \xrightarrow[T\to\infty]{a.s.} \frac{1}{s}\Sigma_1^{-1}$$

$$TQ^{-1}_{[sT,T]} \xrightarrow[T\to\infty]{a.s.} \frac{1}{1-s}\Sigma_2^{-1}$$

$$\frac{1}{T}(Q(s,T)) \xrightarrow[T\to\infty]{a.s.\ and\ L^{\min\{m^*/2,m/2\}}} \Sigma,$$

$$T(Q^{-1}(s,T)) \xrightarrow[T\to\infty]{a.s.} \Sigma^{-1}$$

$$\frac{1}{\sqrt{T}}W(s,T) \xrightarrow[T\to\infty]{\mathrm{D}} \tilde{W}^* \sim \mathcal{N}_{2(p+1)}(0,\Sigma)\ .$$

The proof is similar to the one given for Proposition 3.3.9 of Lyu (2023).

This property is crucial in deriving the limiting distributions of estimators. By providing precise asymptotic behaviors of the processes $Q_{[0,sT]}$, $Q_{[sT,T]}$, and related terms, it ensures the consistency and asymptotic normality of the estimators.

**Proposition 22.18** *If Assumptions 22.1–22.3 hold, then:*

*(1)* $\frac{1}{\sqrt{T}} W(s,T)$ *is* $L^2$*-bounded;*

*(2)* $\frac{1}{T} W(s,T) \xrightarrow[T\to\infty]{a.s.} 0.$

The proof of this result is similar to the one given for Proposition 3.3.10 in Lyu (2023).

This property indicates that $W(s,T)$ becomes asymptotically small in appropriately scaled terms. Practically, this is crucial in statistical modeling and estimation because it ensures that any randomness introduced by $W(s,T)$ diminishes as the sample size $T$ increases. This leads to more stable and reliable estimates, strongly consistent estimates, as the influence of stochastic noise vanishes over time.

Further, by the definition of $R(s,T)$ in Eq. (22.19) and the fact that the process $\{X(t), t \geqslant 0\}$ satisfies SDE in Eq. (22.2), we have:

$$\begin{aligned}\hat{\theta}_T(s) &= Q^{-1}(s,T)R(s,T) = Q^{-1}(s,T)\left(Q(s,T)\theta + \sigma W(s,T)\right)\\ &= \theta + \sigma Q^{-1}(s,T)W(s,T).\end{aligned} \tag{22.23}$$

Let $\boldsymbol{\rho}_T(s) = \sqrt{T}(\hat{\theta}_T(s) - \theta)$. From Proposition 22.17, we derive the following result which shows that $\hat{\theta}_T(s)$ is strongly consistent and asymptotically normal.

**Proposition 22.19** *Suppose that Assumptions 22.1–22.3 hold. Then,* $\hat{\theta}_T(s)$ *is a strongly consistent estimator of* $\theta$*. Furthermore,* $\hat{\theta}_T(s)$ *is asymptotically normal, i.e.,*

$$\boldsymbol{\rho}_T(s) \xrightarrow[T\to\infty]{\mathrm{D}} \boldsymbol{\rho}(s) \sim \mathcal{N}_{2(p+1)}(0, \sigma^2\Sigma^{-1}).$$

***Proof*** The proof of this proposition is similar to that of Proposition 3.5 in Lyu and Nkurunziza (2023). □

This is an important result, which is the foundation for the established method. The strong consistency ensures that $\hat{\theta}_T(s)$ provides accurate estimates of $\theta$ as the sample size increases, giving confidence in long-term predictions. For large samples, the distribution of the estimator's error becomes predictable and follows a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\Sigma^{-1}$.

From Proposition 22.9, we get:

$$\sqrt{T}(\tilde{\theta}_T(s) - \theta) = \left(I_{(p+1)} - G(s,T)M\right)\sqrt{T}(\hat{\theta}_T(s) - \theta) - \sqrt{T}G(s,T)\left(M\theta - r\right). \tag{22.24}$$

By Proposition 22.17 and continuous mapping theorem, we get:

$$G(s,T) \xrightarrow[T\to\infty]{\mathrm{P}} G^* = \Sigma^{-1}M^\top(M\Sigma^{-1}M^\top)^{-1} \tag{22.25}$$

and

$$I_{p+1} - G(s,T)M \xrightarrow[T\to\infty]{\mathrm{P}} I_{p+1} - G^*M.$$

In the following, we derive a proposition that gives the joint asymptotic normality of UMLE and RMLE. To this end, let

$$\begin{aligned}&(\boldsymbol{\rho}_T(s)^\top, \boldsymbol{\varrho}_T(s)^\top, \boldsymbol{\varsigma}_T(s)^\top)^\top\\ &\quad= \sqrt{T}\left((\hat{\theta}_T(s)-\theta)^\top, (\tilde{\theta}_T(s)-\theta)^\top, (\hat{\theta}_T(s)-\tilde{\theta}_T(s))^\top\right)^\top\end{aligned}$$

and $G^* = \Sigma^{-1}M^\top(M\Sigma^{-1}M^\top)^{-1}$. Also, consider the following set of local alternative restrictions:

$$\mathcal{H}_{a,T} : M\theta - r = r_0/\sqrt{T}, \quad T > 0 \tag{22.26}$$

where $r_0$ is a fixed $q$-column vector. Under the set of local alternatives restriction given by Eq. (22.26), we get:

$$\sqrt{T}G(s,T)(M\theta - r) = \sqrt{T}G(s,T)\frac{r_0}{\sqrt{T}} = G(s,T)r_0 \xrightarrow[T\to\infty]{\mathrm{P}} G^*r_0. \tag{22.27}$$

Before deriving an asymptotic test statistic, let

$$\begin{aligned}\Delta &= \frac{1}{\sigma^2}r_0^\top\left(M\Sigma^{-1}M^\top\right)^{-1}r_0,\\ \Gamma &= \frac{1}{\sigma^2}M^\top\left(M\Sigma^{-1}M^\top\right)^{-1}M,\\ \hat{\Gamma} &= \frac{1}{\hat{\sigma^2}}M^\top\left(MTQ^{-1}(s,T)M^\top\right)^{-1}M.\end{aligned} \tag{22.28}$$

Further, let

$$\begin{aligned}\psi_T(s) &= \boldsymbol{\varsigma}_T(s)^\top\hat{\Gamma}\boldsymbol{\varsigma}_T(s)\,,\\ \psi(s) &= \boldsymbol{\varsigma}^\top(s)\Gamma\boldsymbol{\varsigma}(s)\,,\\ \psi_0(s) &= \boldsymbol{\varsigma}_0^\top(s)\Gamma\boldsymbol{\varsigma}_0(s),\end{aligned} \tag{22.29}$$

where $\varsigma_0(s) \sim \mathcal{N}_{2(p+1)}(0, \sigma^2 G^* M \Sigma^{-1})$ and $G^* = \Sigma^{-1} M^\top \left(M \Sigma^{-1} M^\top\right)^{-1}$. We have the following proposition which is important in developing a suitable test for testing the hypothesis problem given by Eq. (22.5).

**Proposition 22.20** *If Assumptions 22.1–22.3 hold along with the set of local alternative restrictions in Eq. (22.26); then,*

$$(\rho_T^\top(s), \varrho_T^\top(s), \varsigma_T^\top(s))^\top \xrightarrow[T\to\infty]{\mathrm{D}} (\rho^\top(s), \varrho^\top(s), \varsigma^\top(s))^\top,$$

*where*

$$\begin{bmatrix} \rho(s) \\ \varrho(s) \\ \varsigma(s) \end{bmatrix} \sim \mathcal{N}_{6(p+1)} \left[ \begin{bmatrix} 0 \\ -G^* r_0 \\ G^* r_0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \Sigma^{-1} & \Sigma^{-1} - G^* M \Sigma^{-1} & G^* M \Sigma^{-1} \\ \Sigma^{-1} - G^* M \Sigma^{-1} & \Sigma^{-1} - G^* M \Sigma^{-1} & 0 \\ G^* M \Sigma^{-1} & 0 & G^* M \Sigma^{-1} \end{bmatrix} \right].$$

The proof is provided in Appendix "Proof of Proposition 22.20".

This proposition shows that for large samples, the estimator's error becomes predictable, following a multivariate normal distribution with a mean of 0. The established joint asymptotic normality plays a crucial role in constructing shrinkage estimators and in deriving their relative efficiency.

## 22.4 Shrinkage Estimators in the Case of Unknown Change-Point

In the preceding section, we assumed knowing the location of the change-point. However, in practice, we are typically uncertain about the presence of a structural change. Consequently, we must estimate the change-point based on the data at hand.

In this section, we consider the inference problem about the target parameter $\theta$ in the context of an unknown change-point and also the case where the structural changes are unknown. To determine the existence of the change-point, we slightly modify the *SIC information criterion* in Lyu and Nkurunziza (2023). Based on this criterion, we derive the estimator $\hat{s}$ of the parameter $s$. We consider a discretized version of the GCIR process in Eq. (22.2). In particular, we use the Euler-Maruyama discretization with the partition $0 = t_0 < t_1 < \cdots < t_N = T$ on a given period $[0, T]$ with $\delta_N = \max_{1 \leqslant i \leqslant N}(t_{i+1} - t_i)$. Let $\mathbb{N}[a, b] = \{i \in \{0, 1, \ldots, N-1\} : t_i \in [a, b]\}$. The chosen discretization scheme aligns with the context of dealing with high-frequency data and an extensive observation horizon. In particular, we require the following assumption.

**Assumption 22.4** *For the partition* $0 = t_0 < t_1 < \cdots < t_N = T$ *on a given period* $[0, T]$, *there exists a constant* $\boldsymbol{a}$, $\boldsymbol{a} > 1$, *such that* $N = O(T^{\boldsymbol{a}})$, *and we suppose that all the sub-intervals* $[t_{i+1} - t_i]$ *are equal, for* $i \in \mathbb{N}[0, T]$, *i.e.* $\Delta_N = T/N$.

Let

$$\mathrm{IC}(c) = -2\log\mathcal{L}_N(\hat{t^*}, [0, T], \hat{\theta}_T(s)) + (c+1)h(p)\Phi(T) \tag{22.30}$$

where $p$ is the dimension of $\varphi(t)$. $h(p) = p + 1$ and $\Phi(T)$ is a nondecreasing function of $T$, $c = 0$ or 1 is the number of change-point in the process defined in Eq. (22.2), the function $\log \mathcal{L}_N(\hat{t^*}, [0, T], \hat{\theta}_T)$ is the Riemann sum approximation of the log-likelihood function $\log\mathcal{L}(\hat{t^*}, [0, T], \hat{\theta}_T)$ which is defined in Eq. (22.51), and $\hat{t^*}$ is given by

$$\hat{t^*} = \underset{\phi}{\mathrm{argmax}}(\log \mathcal{L}_N(t^*, [0, T], \hat{\theta}_T(s))). \tag{22.31}$$

Let $c^0 = 0$ or 1 represent the exact number of change-point. The primary result of this section is that SIC given by Eq. (22.30) reaches its minimum at the value $c = c^0$. Let $s^0$ be the exact location of the change-point, and let $\hat{s}$ be an estimator of $s^0$.

As in Lyu and Nkurunziza (2023), based on the previous results and assumptions, we have that $\hat{s}$ is a consistent estimator of $s$.

In the following, we introduce shrinkage estimators (SEs) that optimally blend the Unrestricted Maximum Likelihood Estimator (UMLE) and the Restricted Maximum Likelihood Estimator (RMLE). As commonly observed in constrained inference, when the restriction is incorrect, the UMLE outperforms the RMLE. Conversely, when the restriction holds, the RMLE surpasses the UMLE. However, in many cases, it is challenging to be completely certain about the validity of the restriction. Therefore, it is crucial to develop a statistical method that remains robust regardless of the restriction's validity. The SEs offer the advantage of maintaining excellent performance regardless of whether the restriction is valid or not. Nevertheless, due to the structural change in the process, the derivation of shrinkage estimators and their relative efficiency do not align with the findings in classical literature. In particular, as in Lyu (2023), we consider the following shrinkage estimators (SEs):

$$\hat{\theta}_T^{\mathrm{sh}}(\hat{s}) = \tilde{\theta}_T(\hat{s}) + \left(1 - (q-2)\psi_T^{-1}(\hat{s})\right)\left(\hat{\theta}_T(\hat{s}) - \tilde{\theta}_T(\hat{s})\right), \tag{22.32}$$

$$\hat{\theta}_T^{\mathrm{sh}+}(\hat{s}) = \tilde{\theta}_T(\hat{s}) + \left[1 - (q-2)\psi_T^{-1}(\hat{s})\right]^+\left(\hat{\theta}_T(\hat{s}) - \tilde{\theta}_T(\hat{s})\right). \tag{22.33}$$

where $\psi_T(\hat{s})$ is defined in Eq. (22.29), $x^+ = \max(0, x)$, $3 \leqslant q = \mathrm{rank}(M) < 2(p+1)$. As in Lyu and Nkurunziza (2023), we use asymptotic distributional risk (ADR) to study the relative efficiency of the proposed estimators.

We derive the following proposition which shows that, near the restriction, the RMLE dominates the UMLE while the UMLE is better than the RMLE as one moves far away from the restriction.

Let $\Lambda^* = \Sigma^{-1} - G^* M \Sigma^{-1}$, where the matrix $G^*$ is defined in Eq. (22.25).Also, $\lambda_{\min}$ and $\lambda_{\max}$ be the smallest and largest eigenvalues of the matrix $\left(G^{*\top}\Gamma G^*\right)^{-1} G^{*\top}\Omega G^*$, respectively.

**Proposition 22.21** *If Assumptions 22.1–22.4 along with the set of local alternatives in Eq. (22.5) hold, then:*

$$\text{If } \Delta \leqslant \sigma^2 \operatorname{tr}(\Omega G^* M \Sigma^{-1})/\lambda_{\max}, \text{ then } \text{ADR}\left(\tilde{\theta}_T(\hat{s}), \theta, \Omega\right) \leqslant \text{ADR}\left(\hat{\theta}_T(\hat{s}), \theta, \Omega\right);$$

$$\text{If } \Delta \geqslant \sigma^2 \operatorname{tr}(\Omega G^* M \Sigma^{-1})/\lambda_{\min}, \text{ then } \text{ADR}\left(\tilde{\theta}_T(\hat{s}), \theta, \Omega\right) \geqslant \text{ADR}\left(\hat{\theta}_T(\hat{s}), \theta, \Omega\right).$$

The proof of this proposition is similar to that of Proposition 5.1 in Lyu and Nkurunziza (2023).

This proposition confirms that when the restriction is nearly correct, the RMLE dominates the UMLE while the performance of the RMLE becomes poor as one moves far away from the restriction. The following proposition indicates that the SEs dominate the UMLE.

**Proposition 22.22** *If Assumptions 22.1–22.4 along with the local alternative restriction defined in Eq. (22.5) hold; then, if* $2\sigma^2 \operatorname{tr}\left(\Omega G^* M \Sigma^{-1}\right) \geqslant (q+2)\lambda_{\max}$,

$$\text{ADR}\left(\hat{\theta}_T^{\text{sh}+}(\hat{s}), \theta, \Omega\right) \leqslant \text{ADR}\left(\hat{\theta}_T^{\text{sh}}(\hat{s}), \theta, \Omega\right) \leqslant \text{ADR}\left(\hat{\theta}_T(\hat{s}), \theta, \Omega\right), \quad \forall \Delta \geqslant 0.$$

The proof of this proposition is similar to that of Proposition 5.2 in Lyu and Nkurunziza (2023).

This means that under the given condition, the positive shrinkage estimator $\hat{\theta}_T^{\text{sh}+}(\hat{s})$ and shrinkage estimator $\hat{\theta}_T^{\text{sh}}(\hat{s})$ dominate the UMLE $\hat{\theta}_T(\hat{s})$, with the positive-part shrinkage estimator $\hat{\theta}_T^{\text{sh}+}(\hat{s})$ being the best. This shows, unless the practitioner is sure about the correctness of the restriction, the shrinkage estimators are preferable and the top of being robust.

## 22.5 Numerical Evaluations and Real Data Analysis

### *22.5.1 Parameter Estimation*

In this section, we aim to underscore the effectiveness of the proposed method by presenting simulation results that demonstrate its robust performance, particularly in small and medium time horizons. The evaluation employs Monte Carlo simulation techniques, coupled with the Euler-Maruyama discretization approach, to generate observations aligned with the generalized Cox-Ingersoll-Ross (GCIR) process

outlined in Eq. (22.2). Notice that the diffusion parameter $\sigma^2$ can be estimated by $(1/T)\sum_{i=1}^{n}\left(2\sqrt{X_{t_i}}-2\sqrt{X_{t_{i-1}}}\right)^2$.

In the simulations, we generate the GCIR process with a trigonometric orthogonal function system:

$$\left\{1, \sqrt{2}\cos(\omega t), \sqrt{2}\sin(\omega t), \sqrt{2}\cos(2\omega t)\right\}$$

where $\omega = 2\pi$. Hence, the simulated GCIR process is given by

$$\begin{aligned} \mathrm{d}X(t) = \Big( & \big(\mu_{11} + \mu_{12}\sqrt{2}\cos(2\pi t) + \mu_{13}\sqrt{2}\sin(2\pi t) + \mu_{14}\sqrt{2}\cos(4\pi t) \\ & -\alpha_1 X(t)\big)\mathbb{I}_{\{t\leqslant t^*\}} + \big(\mu_{21} + \mu_{22}\sqrt{2}\cos(2\pi t) + \mu_{23}\sqrt{2}\sin(2\pi t) \\ & +\mu_{24}\sqrt{2}\cos(4\pi t) - \alpha_2 X(t)\big)\mathbb{I}(t > t^*)\Big)\mathrm{d}t + \sigma\sqrt{X(t)}\mathrm{d}B_t, \quad 0 \leqslant t \leqslant T \end{aligned} \tag{22.34}$$

where the pre-assigned parameter

$$\begin{aligned} \theta &= (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \alpha_1, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24}, \alpha_2)^\top \\ &= (5, 1, 2, 1, 1, 6, 2, 2, 1, 2)^\top \end{aligned} \tag{22.35}$$

and $X(0) = 15$. We take $\sigma = 0.3$. To apply the proposed method, we consider the restriction $M\theta = 0$ with $M = (1, -1) \otimes I_5$. In each case, 500 iterations are performed. In each iteration, the position of the change-point is estimated. We use $(1/T)\sum_{i=1}^{n}(2\sqrt{X(t_i)} - 2\sqrt{X(t_{i-1})})^2$ to estimate the diffusion parameter $\sigma^2$.

We consider the case of $t^* = 0.5T$. To evaluate the effect of the period $T$, we generate the GCIR process with $T = 20, 35, 50, 80$, with the time increment of $\delta = 0.002$.

The histogram of the estimators when $T = 80$ is given in Fig. 22.1.

Through 500 iterations, we utilize the LSSE method described in Eq. (22.31) to estimate the positions of the change point. Specifically, for instances involving one change-point, the histogram of the estimates is reported in Fig. 22.2. Analysis of Fig. 22.2 reveals a noteworthy trend: as the sample size $T$ expands, the estimates of the change point rates converge toward the pre-assigned values. In simpler terms, the method's accuracy improves with an increase in the sample size, denoting a trend of enhanced precision with larger datasets. This observation underscores the method's reliability and suitability for scenarios with extensive data, reinforcing its potential for accurate inference in cases where a substantial amount of information is available.

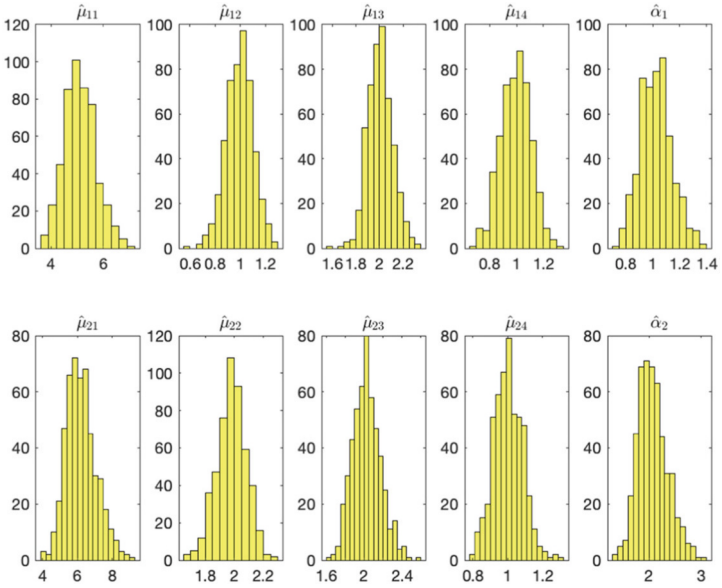



**Fig. 22.1** Histogram of estimates of $\theta$ (T=80)

### *22.5.2 Relative Mean Squared Efficiency*

In this subsection, we conduct an assessment of the performance of the UMLE, RMLE, and SEs through simulations. The approach involves employing Monte Carlo simulation techniques to generate a stochastic process that adheres to the SDE outlined in Eq. (22.2). To assess the performance of the proposed estimators over brief and moderate observation periods, we examine the empirical power's variation relative to the noncentrality parameter $\Delta$, defined as $\Delta = (1/\sigma^2)\, r_0^\top (M\Sigma^{-1}M^\top)^{-1} r_0$, where $r_0 = 0.35\, n\, r$ and $n$ takes values 1 through 6.

Following the approach in Nkurunziza and Zhang (2018), we compute the relative mean squared efficiency (RMSE) as

$$\mathrm{RMSE}(\hat{\theta}_0) = \frac{\mathrm{ADR}(\hat{\theta}(\hat{s}), \theta; \Omega)}{\mathrm{ADR}(\hat{\theta}_0, \theta; \Omega)} \tag{22.36}$$

where $\hat{\theta}_0$ represents an estimator such as $\hat{\theta}_T^{\mathrm{sh}}(\hat{s})$, $\hat{\theta}_T^{\mathrm{sh}+}(\hat{s})$, $\hat{\theta}_T(\hat{s})$, and $\tilde{\theta}_T(\hat{s})$.

From Fig. 22.3a–d, near $\Delta = 0$, RMSE of RMLE is the highest. This indicates that, in proximity to the constraint, the RMLE exhibits greater efficiency compared to the other three estimators. Notably, as one moves farther from the restriction, the efficiency of the RMLE diminishes. Additionally, it is observed that the PSE and SEs consistently outperform the UMLE. The figures confirm that RMLE is

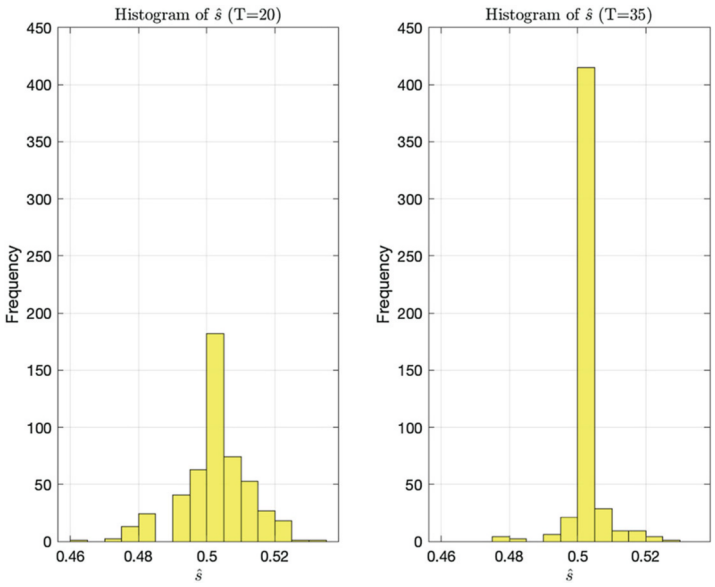



**Fig. 22.2** Histogram of estimates of $s$ $(T = 20, 35)$

particularly advantageous under the linear restriction, emphasizing the nuanced efficiency dynamics across the different estimators. The results underscore the importance of considering proximity to the null hypothesis when evaluating estimator performance. This convergence between empirical and theoretical outcomes enhances the confidence in the validity and applicability of the established theoretical framework. Overall, the study provides empirical support for the theoretical propositions, reinforcing the understanding of estimator behaviors under specified conditions and offering valuable insights for practical applications in statistical estimation.

### *22.5.3 Real Data Analysis*

In this section, we present some numerical evaluations. To save the space of this chapter, we only present in this section the analysis of real datasets. Specifically, we illustrate the application of the proposed methods to the historical corn prices. The dataset under consideration comprises historical corn prices over a specific period from Nov 18, 1991, to Nov 24, 2021, presented in US dollars per bushel. Daily observations of corn prices are essential for capturing the stochastic nature of commodity markets, and the dataset selected for this simulation section reflects these characteristics. The historical corn prices can be found at https://www.macrotrends.net/2532/corn-prices-historical-chart-data.

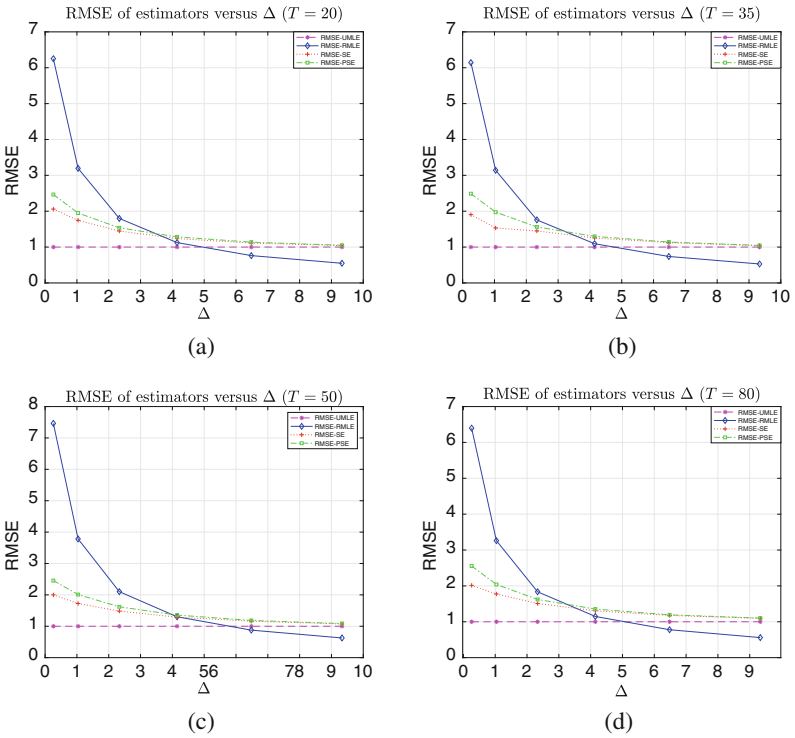


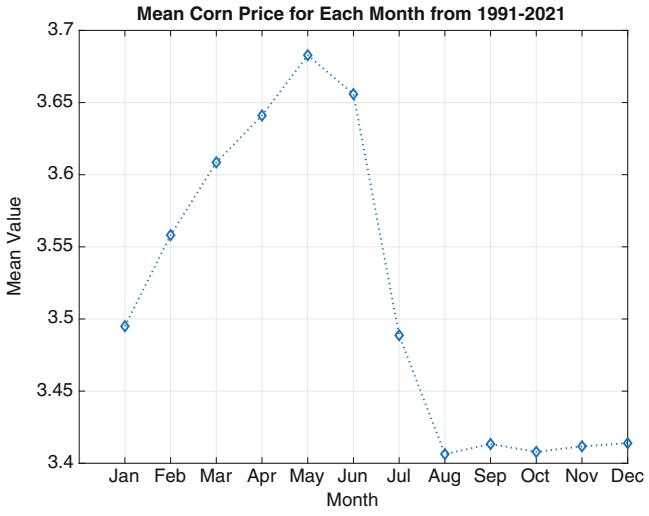

**Fig. 22.3** RMSE of UMLE, RMLE, SEs, PSE versus $\Delta$ ($T = 20, 35, 50, 80$). (**a**) RMSE $T = 20$. (**b**) RMSE T=35. (**c**) RMSE $T = 5$. (**d**) RMSE $T = 80$

To provide a visual representation, Fig. 22.4a illustrates the monthly average corn prices, capturing the periodic characteristic.

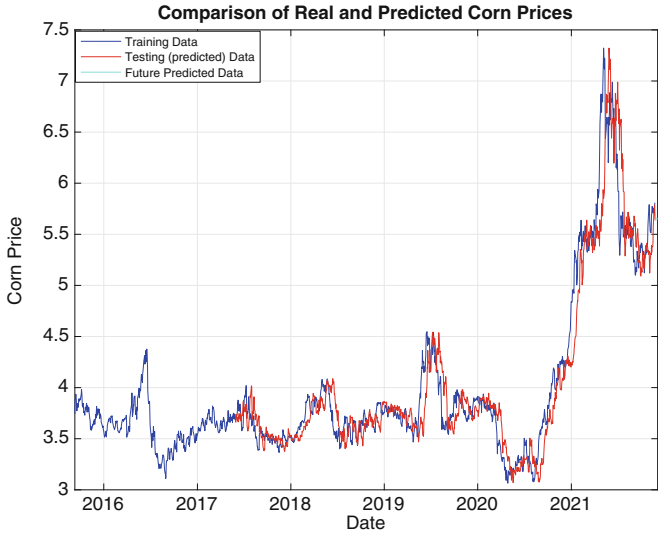
The observed trend in Fig. 22.4a unveils a gradual increase in corn prices from January onward, reaching its pinnacle by May. Subsequently, in June, a slight decline is noticeable. Post the new corn harvest in July, there is a marked and sharp reduction in corn prices. Following September and October, there are no big ups and downs in corn prices.

To apply the method, we consider that the observations have been generated by the GCIR process given by Eq. (22.34), where $T = 30$, which represents the period of the data, and $N = 7{,}567$ is the total trading days during 30 years. So the increment of time is $\delta_N = T/N = 0.004$. For the diffusion parameter $\sigma^2$, its estimate is $\hat{\sigma}^2 = (1/T)\sum_{i=1}^{N}(2\sqrt{X(t_i)} - 2\sqrt{X(t_{i-1})})^2 = (0.5044)^2$. We use the LSSE method stated in Eq. (22.31) to estimate the location parameter of the change-point. In particular, we obtain $\hat{s} = 0.4993$.

To validate the performance of the estimators derived from the GCIR model, we adopt an out-of-sample validation methodology. This approach involves dividing the historical dataset into two distinct periods: a training period (85% of the




(a)




(b)

**Fig. 22.4** Real data plots. (**a**) Plots of the average price of corn from 1991 to 2021. (**b**) Comparison of real and predicted corn prices

data), where the model parameters are estimated, and a testing period (15% of the data) where the model is applied to predict prices that were not used in the parameter estimation process. To evaluate the effectiveness of our estimators, we visually present a comprehensive comparison between the actual corn prices and the corresponding predictions in Fig. 22.4b with predicting 15 steps. Intriguingly, our analysis extends beyond historical data, delving into future predictions. To provide a glimpse into what lies ahead, we plot 8 day's anticipated price trajectories at the rightmost end of the Fig. 22.4b. This forward-looking perspective contributes to a holistic understanding of the estimators' predictive capabilities. To provide a detailed illustration, we focus on the specific time frame from 2016 to 2021. This focused analysis allows for a closer inspection of the predicted corn prices, shedding light on the nuances and trends that may unfold in the immediate future.

## 22.6 Conclusion

This chapter contributes to the existing literature by addressing the issue of change in the generalized CIR model. Our methodology provides a robust and efficient method for estimating the drift parameter, allowing for more accurate modeling and improved risk management in financial applications. By considering the presence of structural break, we enhance the reliability of the generalized CIR model, enabling practitioners to make more informed decisions in an evolving financial landscape.

We introduce shrinkage estimation method for the drift parameter of GCIR processes with an unknown change-point, when the drift parameter adheres to uncertain constraints. A notable aspect of our work lies in addressing scenarios where both the existence and location of the change-point are unknown. We embark on a comprehensive approach. Subsequently, we derive the UMLE and the RMLE.

In cases where the change-point is known, we establish the joint asymptotic normality between the UMLE and the RMLE under a set of local alternative restrictions. Further, enhancing our methodology, we introduce a class of shrinkage estimators (SEs) designed to exhibit robustness in the face of imposed constraints. To deepen our understanding, we delve into the relative risk dominance of the proposed estimators. Our findings indicate that SEs dominate the UMLE, while the RMLE excels in proximity of the restriction but falters when confronted with significant violations of the imposed restriction.

This multifaceted approach contributes to the robustness and flexibility of our methodology, making it applicable in scenarios where uncertainty and unknown change-point complicate traditional modeling. Our work not only provides practical tools for statistical inference in GCIR processes but also sheds light on the nuanced performance of estimators under varying conditions, offering valuable insights for both theoretical and practical considerations in financial modeling and analysis.

# Appendix

## *Some Technical Results and Proofs*

**Lemma 22.1** *Let $\hat{s}$ be $\mathscr{F}_T-$measurable and a consistent estimator of $s$, with $0 \leqslant \hat{s} \leqslant 1$ a.s. and let $\{\varphi(t), t \geqslant 0\}$ be deterministic and bounded function. Then,*

$$\frac{1}{\sqrt{T}} \int_0^{\hat{s}T} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t - \frac{1}{\sqrt{T}} \int_0^{sT} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t \xrightarrow[T\to\infty]{L^2} 0.$$

***Proof*** Let

$$G(T) = \frac{1}{T} \mathrm{E}\left[ \left| \int_0^{\hat{s}T} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t - \int_0^{sT} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t \right|^2 \right].$$

We need to prove that $\lim\limits_{T\to\infty} G(T) = 0$. Also, let $||\varphi(t)|| \leqslant K_\varphi$ and $\epsilon > 0$, we have:

$$\lim_{T\to\infty} \mathrm{P}\left( |\hat{s} - s| \geqslant \frac{\epsilon}{4K_\varphi^2 \sup\limits_{t\geqslant 0} \mathrm{E}[1/X(t)]} \right) = 0.$$

Further, $G(T) = G_{11}(T) + G_{12}(T) + G_{21}(T) + G_{22}(T)$, where

$$G_{11}(T) = \frac{1}{\sqrt{T}} \mathrm{E}\left[ \left| \int_0^{\hat{s}T} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t - \int_0^{sT} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t \right|^2 \times \mathbb{I}(\hat{s} > s)\, \mathbb{I}\left( |\hat{s} - s| \leqslant \frac{\epsilon}{4K_\varphi^2 \sup\limits_{t\geqslant 0} \mathrm{E}[1/X(t)]} \right) \right]$$

$$G_{12}(T) = \frac{1}{\sqrt{T}} \mathrm{E}\left[ \left| \int_0^{\hat{s}T} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t - \int_0^{sT} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t \right|^2 \mathbb{I}(\hat{s} > s)\, \mathbb{I}\left( |\hat{s} - s| \geqslant \frac{\epsilon}{4K_\varphi^2 \sup\limits_{t\geqslant 0} \mathrm{E}[1/X(t)]} \right) \right]$$

$$G_{21}(T) = \frac{1}{\sqrt{T}} \mathrm{E}\left[ \left| \int_0^{\hat{s}T} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t - \int_0^{sT} \frac{\varphi(t)}{\sqrt{X(t)}} \mathrm{d}B_t \right|^2 \mathbb{I}(\hat{s} < s)\, \mathbb{I}\left( |\hat{s} - s| \leqslant \frac{\epsilon}{4K_\varphi^2 \sup\limits_{t\geqslant 0} \mathrm{E}[1/X(t)]} \right) \right]$$

$$G_{22}(T) = \frac{1}{\sqrt{T}}\mathrm{E}\Bigg[\Bigg|\int_0^{\hat{s}T}\frac{\varphi(t)}{\sqrt{X(t)}}\mathrm{d}B_t - \int_0^{sT}\frac{\varphi(t)}{\sqrt{X(t)}}\mathrm{d}B_t\Bigg|^2 \mathbb{I}(\hat{s} < s)\mathbb{I}\Bigg(|\hat{s}-s| \geqslant \frac{\epsilon}{4K_\varphi^2 \sup_{t\geqslant 0}\mathrm{E}[1/X(t)]}\Bigg)\Bigg].$$

By using Cauchy-Schwarz's inequality and Burkholder-Davis-Gundy's inequality, we prove that

$$\lim_{T\to\infty} G_{11}(T) = \lim_{T\to\infty} G_{12}(T) = \lim_{T\to\infty} G_{21}(T) = \lim_{T\to\infty} G_{22}(T) = 0.$$

This completes the proof. □

**Lemma 22.2** *Suppose that Assumptions 22.1–22.4 hold and let $\hat{s}$ be a consistent estimator of $s$, with $0 < \hat{s} < 1$ (a.s.). Then,*

$$\frac{1}{\sqrt{T}}\int_0^{\hat{s}T}\sqrt{X(t)}\mathrm{d}B_t - \frac{1}{\sqrt{T}}\int_0^{sT}\sqrt{X(t)}\mathrm{d}B_t \xrightarrow[T\to\infty]{L^m} 0.$$

***Proof*** Since $\hat{s}$ is $\mathscr{F}_T$-measurable and a consistent estimator of $s$, let $\epsilon > 0$, for every $0 < \delta < \min\{s, 1-s\}$, we have $\mathrm{P}(|\hat{s}-s| \geqslant \delta) < \epsilon$, for sufficiently large $T$. Then,

$$\mathrm{E}\Bigg[\Bigg|\frac{1}{\sqrt{T}}\int_0^{\hat{s}T}\sqrt{X(t)}\mathrm{d}B_t - \frac{1}{\sqrt{T}}\int_0^{sT}\sqrt{X(t)}\mathrm{d}B_t\Bigg|^m\Bigg]$$

$$= \mathrm{E}\Bigg[\Bigg|\frac{1}{\sqrt{T}}\int_0^{\hat{s}T}\sqrt{X(t)}\mathrm{d}B_t - \frac{1}{\sqrt{T}}\int_0^{sT}\sqrt{X(t)}\mathrm{d}B_t\Bigg|^m \mathbb{I}(|\hat{s}-s| \geqslant \delta)\Bigg] \tag{22.37}$$

$$+\mathrm{E}\Bigg[\Bigg|\frac{1}{\sqrt{T}}\int_0^{\hat{s}T}\sqrt{X(t)}\mathrm{d}B_t - \frac{1}{\sqrt{T}}\int_0^{sT}\sqrt{X(t)}\mathrm{d}B_t\Bigg|^m \mathbb{I}_{\{|\hat{s}-s|<\delta\}}\Bigg]. \tag{22.38}$$

Further, one proves that

$$(22.37) \leqslant T^{-m/2}\mathrm{E}\Bigg[\sup_{0\leqslant u\leqslant sT}\Bigg|\int_u^{sT}\sqrt{X(t)}\mathrm{d}B_t\Bigg|^m \mathbb{I}(\hat{s} \leqslant s-\delta)\Bigg]$$

$$+T^{-m/2}\mathrm{E}\Bigg[\Bigg|\sup_{sT\leqslant u\leqslant T}\int_{sT}^{u}\sqrt{X(t)}\mathrm{d}B_t\Bigg|^m \mathbb{I}(\hat{s} \geqslant s+\delta)\Bigg]$$

and

$$(22.38) \leqslant T^{-m/2}\,\mathrm{E}\left[\sup_{(s-\delta)T<u<sT}\left|\int_u^{sT}\sqrt{X(t)}\mathrm{d}B_t\right|^m \mathbb{I}(s-\delta<\hat{s}\leqslant s)\right]$$

$$+T^{-m/2}\,\mathrm{E}\left[\sup_{sT<u<(s+\delta)T}\left|\int_{sT}^{u}\sqrt{X(t)}\mathrm{d}B_t\right|^m \mathbb{I}(s<\hat{s}\leqslant s+\delta)\right].$$

The proof is completed by combining Proposition 22.4, Cauchy-Schwartz's inequality, Burkholder-Davis-Gundy's inequality, and Jensen's inequality. □

The following proposition shows that

$$\frac{1}{T}Q(\hat{s},T)-\frac{1}{T}Q(s,T)\xrightarrow[T\to\infty]{L^{\min\{m^*/2,m/2\}}}0.$$

From Assumption 22.3, $m^*\geqslant 2$.

**Proposition 22.23** *Suppose that Assumptions 22.1–22.4 hold and $\hat{s}$ is a consistent estimator of $s$, with $0<\hat{s}<1$ a.s. Then,*

$$\frac{1}{T}\int_0^{\hat{s}T}\varphi^\top(t)\varphi(t)\frac{1}{X(t)}\mathrm{d}t-\frac{1}{T}\int_0^{sT}\varphi^\top(t)\varphi(t)\frac{1}{X(t)}\mathrm{d}t\xrightarrow[T\to\infty]{L^{m^*/2}}0.$$

***Proof*** Since $\hat{s}$ is $\mathscr{F}_T$-measurable and a consistent estimator of $s$, let $\epsilon>0$, for every $0<\delta<\min\{s,1-s\}$, we have $\mathrm{P}(|\hat{s}-s|\geqslant\delta)<\epsilon$, for sufficiently large $T$. Let

$$\mathscr{A}_0=\mathrm{E}\left[\left\|\frac{1}{T}\int_0^{\hat{s}T}\varphi^\top(t)\varphi(t)\frac{1}{X(t)}\mathrm{d}t-\frac{1}{T}\int_0^{sT}\varphi^\top(t)\varphi(t)\frac{1}{X(t)}\mathrm{d}t\right\|^{m^*/2}\right].$$

Then,

$$\mathscr{A}_0\leqslant K_\varphi\,\mathrm{E}\left[\left|\frac{1}{T}\int_{(s\wedge\hat{s})T}^{(s\vee\hat{s})T}\frac{1}{X(t)}\mathrm{d}t\right|^{m^*/2}\right]$$

$$=K_\varphi\,\mathrm{E}\left[\left|\frac{1}{T}\int_0^{\hat{s}T}\frac{1}{X(t)}\mathrm{d}t-\frac{1}{T}\int_0^{sT}\frac{1}{X(t)}\mathrm{d}t\right|^{m^*/2}\right]$$

$$=K_\varphi\,\mathrm{E}\left[\left|\frac{1}{T}\int_0^{T}\frac{1}{X(t)}\left|\mathbb{I}(0<t\leqslant\hat{s}T)-\mathbb{I}(0<t\leqslant sT)\right|\mathrm{d}t\right|^{m^*/2}\right]. \quad (22.39)$$

Further, by using Jensen's inequality

$$\mathcal{A}_0 \leqslant K_\varphi \, \mathrm{E}\left[\frac{1}{T}\int_0^T \left(\frac{1}{X(t)}\right)^{m^*/2} \left|\mathbb{I}(0 < t \leqslant \hat{s}T) - \mathbb{I}(0 < t \leqslant sT)\right| \mathrm{d}t\right].$$

And then, applying the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned}\mathcal{A}_0 &\leqslant K_\varphi \frac{1}{T}\int_0^T \left(\mathrm{E}\left[\left(\frac{1}{X(t)}\right)^{m^*}\right]\right)^{1/2} \\ &\qquad \times \left(\mathrm{E}\left[\left|\mathbb{I}(0 < t \leqslant \hat{s}T) - \mathbb{I}(0 < t \leqslant sT)\right|^2\right]\right)^{1/2} \mathrm{d}t. \\ &\leqslant K_\varphi \left(\frac{1}{T}\int_0^T \mathrm{E}\left[\left(\frac{1}{X(t)}\right)^{m^*}\right] \mathrm{d}t\right)^{1/2} \\ &\qquad \times \left(\frac{1}{T}\int_0^T \mathrm{E}\left[\left|\mathbb{I}(0 < t \leqslant \hat{s}T) - \mathbb{I}(0 < t \leqslant sT)\right|^2\right] \mathrm{d}t\right)^{1/2}.\end{aligned}$$

From Eq. (22.20)

$$\begin{aligned}\mathcal{A}_0 &\leqslant K_\varphi \left(\sup_{t\geqslant 0} \mathrm{E}\left[\left(\frac{1}{X(t)}\right)^{m^*}\right]\right)^{1/2} \\ &\qquad \left(\frac{1}{T}\int_0^T \mathrm{E}\left[\left|\mathbb{I}(0 < t \leqslant \hat{s}T) - \mathbb{I}(0 < t \leqslant sT)\right|^2\right] \mathrm{d}t\right)^{1/2} \\ &\leqslant K_\varphi \left(\sup_{t\geqslant 0} \mathrm{E}\left[\left(\frac{1}{X(t)}\right)^{m^*}\right]\right)^{1/2} \left(\mathrm{E}\left[\left|s - \hat{s}\right|\right]\right)^{1/2}.\end{aligned}$$

In addition, $\left|s - \hat{s}\right| \leqslant 2$, by Lebesgue-dominated convergence theorem, we have: $\lim_{T\to\infty} \mathrm{E}\left[\left|s - \hat{s}\right|\right] = 0$. This completes the proof. □

**Proposition 22.24** *If Assumptions 22.1–22.4 hold, then:*

*(i)* $||W(s,T)|| = O_p(\sqrt{T}\log^{a^*}(T))$
*(ii) for some* $0 < a^* < \boldsymbol{a}/2$,

$$\begin{aligned}&\left(\left\|\sum_{i\in\mathbb{N}[0,sT]} \varepsilon_i Z_{1i}/(t_{i+1} - t_i)\right\|, \left\|\sum_{i\in\mathbb{N}[sT,T]} \varepsilon_i Z_{2i}/(t_{i+1} - t_i)\right\|\right) \\ &\quad = O_p(\sqrt{T}\log^{a^*}(T)).\end{aligned} \tag{22.40}$$

***Proof*** Since SDE given by Eq. (22.2) admits a strong and unique solution that is uniformly bounded in $L^m$, i.e., $\exists\ K_w > 0$, such that $\sup_{t\geq 0} \mathrm{E}[X(t)] \leq K_w$, we have:

$$\int_0^T \mathrm{E}[(X(t))]\mathrm{d}t \leq K_w T.$$

Further, from Eq. (22.7), $\exists\ K_b > 0$, such that $\sup_{t\geq 0} \mathrm{E}[(X(t))^{-1}] \leq K_b$,

$$\int_0^T \mathrm{E}[(X(t))^{-1}]\mathrm{d}t \leq K_b T.$$

The proof of Part (i) is similar to that Proposition 2.4.3 in Lyu (2023), and the proof of Part (ii) is similar to that of Proposition 2.4.4 in Lyu (2023). □

**Proposition 22.25** *If Assumptions 22.1–22.4 hold, then, for some* $0 < a^* < \boldsymbol{a}/2$,

$$\left( \left\| \sum_{i\in\mathbb{N}[0,\hat{s}T]} \varepsilon_i \mathcal{Z}_{1i}/(t_{i+1}-t_i) \right\|, \left\| \sum_{i\in\mathbb{N}[\hat{s}T,T]} \varepsilon_i \mathcal{Z}_{2i}/(t_{i+1}-t_i) \right\| \right) = O_p(\sqrt{T}\log^{a^*}(T)). \tag{22.41}$$

***Proof*** The proof follows from the triangle inequality and Proposition 22.24. □

**Proposition 22.26** *Suppose that Assumptions 22.1–22.4 hold, we have:*

$$\frac{1}{T}\left( \sum_{i\in\mathbb{N}[0,\hat{s}T]} (Z_{1i}^\top Z_{1i})/(t_{i+1}-t_i), \sum_{i\in\mathbb{N}[\hat{s}T,T]} Z_{2i}^\top Z_{2i}/(t_{i+1}-t_i) \right) \xrightarrow[\Delta_N\to 0]{L^{m^*/2}} (s\Sigma_1, (1-s)\Sigma_2)\,. \tag{22.42}$$

***Proof*** The proof follows from the triangle inequality and Proposition 22.17 along with Proposition 4.1 in Lyu and Nkurunziza (2023). This completes the proof. □

**Lemma 22.3** *If Assumptions 22.1–22.4 hold, then*

$$\gamma_1(s,T) \xrightarrow[\Delta_N\to 0]{P} s\gamma_1 \quad \textit{and} \quad \gamma_2(s,T) \xrightarrow[\Delta_N\to 0]{P} (1-s)\gamma_2.$$

***Proof*** From Proposition A.3 of Lyu (2023) and Proposition 22.26 along with the fact that $\Sigma_1$ and $\Sigma_2$ are positive definite matrices, we get for $\gamma_1 > 0$:

$$\gamma_1(s, T) \xrightarrow[\substack{T\to\infty \\ \Delta_N\to 0}]{P} \gamma_1(s) = s\gamma_1$$

and for $\gamma_2 > 0$,

$$\gamma_2(s, T) \xrightarrow[\substack{T\to\infty \\ \Delta_N\to 0}]{P} (1-s)\gamma_2.$$

□

**Proposition 22.27 (Theorem 1.1 in Geißand Manthey (1994))** *Suppose that on a certain probability space $(\Omega^*, \mathcal{F}, \mathbb{P})$ equipped with a filtration $\{\mathcal{F}_t\}$ which satisfies the usual conditions, we have a standard, one-dimensional Brownian motion $\{B_t, \mathcal{F}_t, 0 \leqslant t < \infty\}$ and two adapted processes $X_t^{(j)}$, $j = 1, 2$, such that for $j = 1, 2$,*

$$dX_t^{(j)} = X_0^{(j)} + \int_0^t b_j(s, X_s^{(j)})\mathrm{d}s + \int_0^t \sigma(s, X_s^{(j)})\mathrm{d}B_s;\ 0 \leqslant t < \infty \quad a.s.$$

*We assume that:*

1. *The coefficients $\sigma(t, x)$ and $b_j(t, x)$ are continuous, real-valued functions on $[0, +\infty) \times \mathbb{R}$,*
2. *The dispersion matrix $\sigma(t, x)$ satisfies $|\sigma(t, x) - \sigma(t, y)| \leqslant h(|x - y|)$ where $h : [0, +\infty) \to [0, +\infty)$ is a strictly increasing function with $h(0) = 0$ and*

$$\int_0^{\varepsilon} h^{-2}(u)du = \infty;\ \ \forall\, \varepsilon > 0,$$

3. *$X_0^{(1)} \leqslant X_0^{(2)}$ a.s.*
4. *$b_1(t, x) \leqslant b_2(t, x),\ \forall\, 0 \leqslant t < \infty,\ x \in \mathbb{R}$,*
5. *Both $b_1(t, x)$ and $b_2(t, x)$ satisfy the Lipschitz condition in $x$.*

*Then, we have* $\mathrm{P}\left(X_t^{(1)} \leqslant X_t^{(2)},\ \forall\, 0 \leqslant t < \infty\right) = 1.$

**Theorem 22.1 (Lévy's Characterization of Brownian Motion)** *(Theorem 1.4.4, p. 17 in Mao, 2007) Let $\{M(t),\ t \geqslant 0\}$ be a martingale relative to a filtration $\mathcal{F}_t$, $t \geqslant 0$. Assume that $M(0) = 0$, $M(t)$ has continuous paths and $\langle M\rangle_t = t$ for all $t \geqslant 0$. Then, $\{M(t),\ t \geqslant 0\}$ is a standard Brownian motion.*

## *Important Proofs*

**Proof of Proposition 22.2**

Consider the SDE without change-point. Let for $t \in [0, T]$, $G(t, X(t)) = \exp\{\alpha t\}X(t)$, by Itô's lemma, we have:

$$\mathrm{d}G(t, X(t)) = \frac{\partial G}{\partial t}(t, X(t))\mathrm{d}t + \frac{\partial G}{\partial X(t)}(t, X(t))\mathrm{d}X(t)$$

$$+\frac{1}{2}\frac{\partial^2 G}{\partial X(t)^2}(t, X(t))\mathrm{d}\langle X(t), X(t)\rangle, \tag{22.43}$$

where $\langle\cdot,\cdot\rangle$ denote the variation. Since $\mathrm{d}X(t) = (L(t) - \alpha X(t))\mathrm{d}t + \sigma\sqrt{X(t)}\mathrm{d}B_t$, then

$$\begin{aligned}\mathrm{d}G(t, X(t)) &= \alpha_1 \exp\{\alpha\, t\}X(t)\mathrm{d}t + \exp\{\alpha t\}\mathrm{d}X(t)\\ &= \alpha_1 \exp\{\alpha_1 t\}\, X(t)\mathrm{d}t + \exp\{\alpha t\}\left((L(t) - \alpha X(t))\mathrm{d}t + \sigma\sqrt{X(t)}\mathrm{d}B_t\right),\end{aligned}$$

which implies that $\mathrm{d}(\exp\{\alpha\, t\}X(t)) = \exp\{\alpha t\}L(t)\mathrm{d}t + \sigma\exp\{\alpha t\}\sqrt{X(t)}\mathrm{d}B_t$. Integrating both sides from 0 to $t$, we derive that

$$\exp\{\alpha t\}X(t) = X_0 + \int_0^t \exp\{\alpha s\}L(s)\,\mathrm{d}s + \int_0^t \sigma\exp\{\alpha_1 s\}\sqrt{X_s}\mathrm{d}B_s.$$

Then,

$$\begin{aligned}X(t) = \exp\{-\alpha\, t\}X_0&\\ + \exp\{-\alpha t\}\int_0^t \exp\{\alpha s\}L(s)\mathrm{d}s + \sigma\exp\{-\alpha t\}&\int_0^t \exp\{\alpha s\}\sqrt{X_s}\mathrm{d}B_s.\end{aligned}$$

This shows that for $t \in [0, sT]$, $X_1(t) = \exp\{-\alpha_1 t\}X_0 + r_1(t) + \tau_1(t)$ and similarly, for the case $t \in [t^*, T]$, $X_2(t) = \exp\{-\alpha_2(t - t^*)\}X^{t^*}(0) + r_2^{t^*}(t - t^*) + \tau_2^{t^*}(t - t^*)$, with the notation given in this proposition. This completes the proof. □

**Proof of Proposition 22.4**

The case where $m = 2$ is proved in Proposition 22.3. Thus, we only consider the case where $m > 2$. The model given by Eq. (22.2) is under consideration. Let $Y_j(t, X(t)) = \exp\{\alpha_j mt\}\, X_j^m(t)$ if $0 < t \leqslant t^*$ while $j = 1$, and $t^* < t \leqslant T$ while $j = 2$, by Itô's formula, we get:

$$dY_1(t, X_1(t)) = \alpha_1 m\, \exp\{\alpha_1 mt\}X_1^m(t)\mathrm{d}t + m\exp\{\alpha_1 mt\}X^{m-1}(t)dX_1(t)$$

$$
\begin{aligned}
&+\frac{1}{2}m(m-1)\exp\{\alpha_1 mt\}X_1^{m-2}(t)d\langle X_1(t),X_1(t)\rangle\\
&=\alpha_1 m\exp\{\alpha_1 mt\}X_1^m(t)\mathrm{d}t\\
&\quad+m\exp\{\alpha_1 mt\}X_1^{m-1}(t)\left(L_1(t)-\alpha_1 X_1(t)\right)\mathrm{d}t\\
&\quad+m\exp\{\alpha_1 mt\}X_1^{m-1}(t)\sigma\sqrt{X_1(t)}\mathrm{d}B_t\\
&\quad+\frac{1}{2}m(m-1)\exp\{\alpha_1 mt\}X_1^{m-2}(t)d\langle X_1(t),X_1(t)\rangle\\
&=m\exp\{\alpha_1 mt\}X_1^{m-1}(t)\left(L_1(t)+\sigma^2\frac{1}{2}(m-1)\right)\mathrm{d}t\\
&\quad+\sigma m\exp\{\alpha_1 mt\}X_1^{m-1/2}(t)\mathrm{d}B_t.
\end{aligned}
$$

Then,

$$
\begin{aligned}
\exp\{\alpha_1 mt\}X_1^m(t)&=X_0^m+\int_0^t m\exp\{\alpha_1 ms\}X_1^{m-1}(s)\left(L_1(s)+\sigma^2\frac{1}{2}(m-1)\right)\mathrm{d}s\\
&\quad+\int_0^t\sigma m\exp\{\alpha_1 ms\}X_1^{m-1/2}(s)\mathrm{d}B_s,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
X_1^m(t)&=\exp\{-\alpha_1 mt\}X_0^m\\
&\quad+m\int_0^t\exp\{-\alpha_1 m(t-s)\}X_1^{m-1}(s)\left(L_1(s)+\sigma^2\frac{1}{2}(m-1)\right)\mathrm{d}s\\
&\quad+\sigma m\int_0^t\exp\{-\alpha_1 m(t-s)\}\,X_1^{m-1/2}(s)\,\mathrm{d}B_s. \qquad (22.44)
\end{aligned}
$$

By using the stopping time $T_n$ defined in Proposition 22.3, we have:

$$
\begin{aligned}
X_1^m(t\wedge T_n)&=\exp\{-\alpha_1 m(t\wedge T_n)\}X_0^m\\
&\quad+m\int_0^{t\wedge T_n}\exp\{-\alpha_1 m(t\wedge T_n-s)\}X_1^{m-1}(s)\\
&\qquad\qquad\times\left(L_1(s)+\sigma^2\frac{1}{2}(m-1)\right)\mathrm{d}s\\
&\quad+\sigma m\int_0^{t\wedge T_n}\exp\{-\alpha_1 m(t\wedge T_n-s)\}X_1^{m-1/2}(s)\mathrm{d}B_s. \qquad (22.45)
\end{aligned}
$$

Since $M_1(t) = \int_0^t \exp\{\alpha_1 ms\} X_1^{m-1/2}(s)\mathrm{d}B_s$ is an Itô integral, $M_1(t)$ is a local martingale. The stopping time $T_n$ is a localizing sequence, so that

$$M_1(t \wedge T_n) := \int_0^{t\wedge T_n} \exp\{\alpha_1 ms\} X_1^{m-1/2}(s)\mathrm{d}B_s$$

is a martingale and $\mathrm{E}[M_1(t \wedge T_n)] = \mathrm{E}[M_1(0)] = 0$. Since $X_1(t)$ is nonnegative, $\int_0^{t\wedge T_n(\omega)} X_1^{m-1}(s,\omega)\mathrm{d}s$ is increasing and for all $\omega \in \Omega^*$,

$$\lim_{n\to\infty} \int_0^{t\wedge T_n(\omega)} X_1^{m-1}(s,\omega)\mathrm{d}s = \int_0^t X_1^{m-1}(s,\omega)\mathrm{d}s;$$

we have $\int_0^{t\wedge T_n} X_1^{m-1}(s)\mathrm{d}s \xrightarrow[n\to\infty]{a.s.} \int_0^t X_1^{m-1}(s)\mathrm{d}s$.

Therefore, by the monotone convergence theorem

$$\int_0^{t\wedge T_n} X_1^{m-1}(s)\mathrm{d}s \xrightarrow[n\to\infty]{L^1} \int_0^t X_1^{m-1}(s)\mathrm{d}s,$$

which implies that

$$\lim_{n\to\infty} \mathrm{E}\left[\int_0^{t\wedge T_n} X_1^{m-1}(s)\mathrm{d}s\right] = \mathrm{E}\left[\int_0^t X_1^{m-1}(s)\mathrm{d}s\right].$$

Further, since $X_1^m(t \wedge T_n) \xrightarrow[n\to\infty]{a.s.} X_1^m(t)$, taking expectation on both sides of Eq. (22.45) and by Fatou's lemma, we get:

$$\mathrm{E}[X_1^m(t)] = \mathrm{E}[\liminf_{n\to\infty} X_1^m(t \wedge T_n)] \leqslant \lim_{n\to\infty} \mathrm{E}[X_1^m(t \wedge T_n)].$$

Then,

$$\mathrm{E}[X_1^m(t)] \leqslant \exp\{-\alpha_1 m\, t\}\, \mathrm{E}[X_0^m] \tag{22.46}$$
$$+m\big(K_\mu K_\varphi + \frac{1}{2}\sigma^2(m-1)\big) \exp\{-\alpha_1 m\, t\} \int_0^t \exp\{\alpha_1 m\, s\}\, \mathrm{E}[X_1^{m-1}(t)]\mathrm{d}s,$$

which gives

$$\mathrm{E}[X_1^m(t)] \leqslant \mathrm{E}[X_0^m] + C_m \exp\{-\alpha_1 m\, t\} \int_0^t \exp\{\alpha_1 m\, s\}\, \mathrm{E}[X_1^{m-1}(s)]\mathrm{d}s, \tag{22.47}$$

where $C_m = m(K_\mu K_\varphi + 0.5\sigma^2(m-1))$. Then, by recursion of Eq. (22.47), we have:

$$\begin{aligned}\mathrm{E}[X_1^m(t)] \leqslant \mathrm{E}[X_0^m] + C_m \exp\{-\alpha_1 m t\} \\ \times \int_0^t \exp\{\alpha_1 m s_1\}\Big( \mathrm{E}[X_0^{m-1}] + C_{m-1}\exp\{-\alpha_1(m-1)s_1\} \\ \times \int_0^{s_1} \exp\{\alpha_1(m-1)s_2\}\,\mathrm{E}[X_1^{m-2}(s_2)]\mathrm{d}s_2\Big)\mathrm{d}s_1.\end{aligned}$$

Then,

$$\begin{aligned}\mathrm{E}[X_1^m(t)] \leqslant \mathrm{E}[X_0^m] + C_m\,\mathrm{E}[X_0^{m-1}]\frac{1}{\alpha_1 m} + C_m C_{m-1}\exp\{-\alpha_1 m t\} \\ \times \int_0^t \exp\{\alpha_1 s_1\}\left(\int_0^{s_1}\exp\{\alpha_1(m-1)s_2\}\,\mathrm{E}[X_1^{m-2}(s_2)]\mathrm{d}s_2\right)\mathrm{d}s_1.\end{aligned}$$

This gives:

$$\begin{aligned}\mathrm{E}[X_1^m(t)] \leqslant \mathrm{E}[X_0^m] + C_m\,\mathrm{E}[X_0^{m-1}]\frac{1}{\alpha_1 m} \\ + C_m C_{m-1}\,\mathrm{E}[X_0^{m-2}]\frac{1}{\alpha_1(m-1)}\frac{1}{\alpha_1 m} + \cdots \\ + C_m C_{m-1}\cdots C_{m-(k-1)}\,\mathrm{E}\left[X_0^{m-(k-1)}\right]\exp\{-\alpha_1 m t\} \\ \times \int_0^t\int_0^{s_1}\cdots\int_0^{s_{k-1}} \exp\{\alpha_1 s_1\}\exp\{\alpha_1 s_2\}\cdots \\ \times \exp\{\alpha_1(m-(k-1))s_k\}\,\mathrm{E}[X_1^{m-k}(s_k)]\mathrm{d}s_k\cdots\mathrm{d}s_2\mathrm{d}s_1,\end{aligned}$$

for $0 < m - k \leqslant 1$. By Jensen's inequality

$$\mathrm{E}[X_1^{m-k}(t)] \leqslant \big(\mathrm{E}[X_1(t)]\big)^{m-k} \leqslant \begin{cases}1, & \text{if } \mathrm{E}[X_1(t)] \leqslant 1 \\ \mathrm{E}[X_1(t)], & \text{else.}\end{cases}$$

This implies that for $t \in [0, sT]$,

$$\begin{aligned}\mathrm{E}[X_1^m(t)] \leqslant \mathrm{E}[X_0^m] + \max\{1, \sqrt{K_1 + K_2}\} \\ \times \sum_{j=1}^{m}\Big(\prod_{k=1}^{j} C_{m-(k-1)}\,\mathrm{E}[X_0^{m-(k-1)}]\frac{1}{\alpha_1(m-(k-1))}\Big).\end{aligned} \tag{22.48}$$

Further, let

$$K_{m1} = \mathrm{E}[X_0^m] + \max\{1, \sqrt{K_1 + K_2}\} \sum_{j=1}^{m} \left( \prod_{k=1}^{j} C_{m-(k-1)} \, \mathrm{E}[X_0^{m-(k-1)}] \frac{1}{\alpha_1(m-(k-1))} \right).$$

For $t \in [0, sT]$, $\mathrm{E}[X_1(t)] \leqslant \sqrt{K_1 + K_2}$ implies that $\mathrm{E}[X_1^m(t)] \leqslant K_{m1}$, which indicates that $\mathrm{E}[(X_0^{t^*})^m] \leqslant K_{m1}$. Further, by Jensen's inequality

$$\mathrm{E}[(X_0^{t^*})^m] \geqslant \left( \mathrm{E}[(X_0^{t^*})^{m-k}] \right)^{\frac{m}{m-k}}.$$

This indicates that

$$\mathrm{E}[(X_0^{t^*})^{m-k}] \leqslant \left( \mathrm{E}[(X_0^{t^*})^m] \right)^{\frac{m-k}{m}} \leqslant K_{m1}^{\frac{m-k}{m}}.$$

Similarly, we can prove that $\mathrm{E}[X_2^m(t)] \leqslant K_{m2}$, provided that $t \in [sT, T]$, $K_{m2}$ is defined in Eq. (22.12):

$$\mathrm{E}[X^m(t)] \leqslant \begin{cases} K_{m1} & \text{if } t \in [0, sT] \\ K_{m2} & \text{if } t \in [sT, T] \end{cases}$$

implies that Eq. (22.13) holds where $K_m = \max\{K_{m1}, K_{m2}\}$. This completes the proof. □

**Proof of Proposition 22.5**

As an intermediate result, before Proposition 22.5, we derive first the following lemma which is similar to Lemma 2.1 in Maghsoodi (1996).

**Lemma 22.4** *Given a one-dimensional Brownian motion $B$ defined on the probability space $(\Omega^*, \mathscr{F}, \mathrm{P})$ with $(\mathscr{F}_t)_{t \geqslant 0}$ as usual, there exists a two-dimensional Brownian motion $\tilde{W} = (\tilde{w}_1, \tilde{w}_2)$ on the same probability space, such that for $k = 1, 2$:*

$$B_t = \sum_{j=1}^{2} \int_0^t \frac{X^{(j*)}(u)}{||X^*(u)||} d\tilde{w}_j(u) \quad a.s.$$

*Furthermore, it is possible to choose $\tilde{W} = B^{(*)}$ almost surely.*

**Proof of Proposition 22.5**

By Theorem 7.1 of Ikeda and Watanabe (1989, p. 84), for the given one-dimensional Brownian motion $B$ defined on the probability space $(\Omega^*, \mathscr{F}, \mathrm{P})$, there exists a two-dimensional $(\mathscr{F}_t)-$Brownian motion $\tilde{W} = (\tilde{w}_1, \tilde{w}_2)$ with $\tilde{W}(0) = 0$, such that

$$B_t = \sum_{j=1}^{2} \int_0^t \frac{X^{(j*)}(s)}{\sqrt{X^{(1*)}(s) + X^{(2*)}(s)}} d\tilde{w}_j(s), \quad a.s. \tag{22.49}$$

Furthermore, it is possible to choose $\tilde{W} = B^{(*)}$ almost surely. It completes the proof. □

**Proof of Proposition 22.8**

Let $a = [a_1^\top, a_2]$ with $a_1$ a $p-$column vector, and $a_2$ a scalar. Then,

$$a\, Q_{[0,sT]}\, a^\top = \left( a_1^\top \int_0^{sT} \frac{\varphi^\top(t)\varphi(t)}{X(t)} \mathrm{d}t - a_2 \int_0^{sT} \varphi(t) \mathrm{d}t, -a_1^\top \int_0^{sT} \varphi^\top(t) \mathrm{d}t + a_2 \int_0^{sT} X(t) \mathrm{d}t \right) (a_1^\top, a_2)^\top$$

and

$$a\, Q_{[0,sT]}\, a^\top = \int_0^{sT} \left( a_1^\top \frac{\varphi^\top(t)}{\sqrt{X(t)}} - a_2 \sqrt{X(t)} \right)^2 \mathrm{d}t \geqslant 0\,.$$

and the equality hold if and only if

$$\left( a_1^\top \frac{\varphi^\top(t)}{\sqrt{X(t)}} - a_2 \sqrt{X(t)} \right)^2 = 0, \quad \text{almost everywhere (a.e)} \ \ \forall\, t \in [c, d],$$

which is $a_1^\top \left( \varphi^\top(t) \,/\, \sqrt{X(t)} \right) - a_2 \sqrt{X(t)} = 0$, a.e for $t \in [0, sT]$ and then,

$$\mathrm{P}\left( \left\{ \omega : a_1^\top \frac{\varphi^\top(t)}{\sqrt{X(t,\omega)}} - a_2 \sqrt{X(t,\omega)} = 0, \forall\, t \in [0, sT] \right\} \right) = 1,$$

which is equivalent to the following

$$\mathrm{P}\left( \left\{ \omega : a_1^\top \varphi^\top(t) - a_2 X(t,\omega) = 0, \forall\, t \in [0, sT] \right\} \right) = 1, \tag{22.50}$$

and since the process $\{X(t), t \geqslant 0\}$ has absolutely continuous distribution, for $a_2 \neq 0$,

$$\mathrm{P}\left(\left\{\omega : a_1^\top \varphi^\top(t) - a_2 X(t,\omega) = 0, \forall\, t \in [0, sT]\right\}\right) = 0.$$

This is a contradiction with Eq. (22.50); hence, $a_2 = 0$. Therefore, putting $a_2 = 0$ into $a_1^\top\left(\varphi^\top(t), \sqrt{X(t)}\right) - a_2\sqrt{X(t)} = 0$, provided that $[0, 1] \subset [0, sT]$, we get that $a_1^\top\left(\varphi^\top(t)/\sqrt{X(t)}\right) = 0$, which is equivalent to $a_1^\top\varphi^\top(t) = 0$. By Assumption 22.3, the base function $\varphi(t)$ is linearly independent, which means $a_1^\top\varphi^\top(t) = 0$ implies that $a_1 = \mathbf{0}_{(p+1)\times 1}$. This completes the proof of the fact that matrix $Q_{[0,sT]}$ is a positive definite matrix. Further, we can prove that the matrix $Q_{[sT,T]}$ is a positive definite matrix, since we have $[0, 1] \subset [0, (1-s)T]$. □

**Proof of Proposition 22.9**

The likelihood function of the SDE in Eq. (22.2) is given by

$$\begin{aligned}
\mathcal{L}(\theta, s, X^T) :&= \frac{\mathrm{d}P_X}{\mathrm{d}P_B}(X^T) \\
&= \exp\left\{\int_0^T \frac{S(\theta_1, t, X(t))}{\sigma^2 X(t)}\mathbb{I}(0 \leqslant t \leqslant t^*)X(t)\right. \\
&\quad + \int_0^T \frac{S(\theta_2, t, X(t))}{\sigma^2 X(t)}\mathbb{I}(t^* < t \leqslant T)\mathrm{d}X(t) \\
&\quad - \frac{1}{2}\int_0^T \frac{S^2(\theta_1, t, X(t))}{\sigma^2 X(t)}\mathbb{I}(0 \leqslant t \leqslant t^*)\mathrm{d}t \\
&\quad \left. - \frac{1}{2}\int_0^T \frac{S^2(\theta_2, t, X(t))}{\sigma^2 X(t)}\mathbb{I}(t^* < t \leqslant T)\,\mathrm{d}t\right\}. \qquad (22.51)
\end{aligned}$$

Together with some algebraic computations, the log-likelihood function is

$$\log\mathcal{L}(\theta, s, X^T) = \frac{1}{\sigma^2}\theta^\top R(s,T) - \frac{1}{2\sigma^2}\theta^\top Q(s,T)\theta.$$

Then, by combining Proposition 22.8 and some algebraic computations, we obtain that the UMLE is $\hat{\theta}_T(s) = Q^{-1}(s,T)R(s,T)$. Further, by the Lagrange multiplier method, one maximizes $\log\mathcal{L}(\theta, s, X^T)$ under the restriction defined in Eq. (22.4), which yields the RMLE as stated. □

**Proof of Proposition 22.20**

From (22.24), we get:

$$\sqrt{T}(\hat{\theta}_T(s) - \tilde{\theta}_T(s)) = \sqrt{T}(\hat{\theta}_T(s) - \theta) - \sqrt{T}(\tilde{\theta}_T(s) - \theta)$$

$$= G(s,T)M\sqrt{T}(\hat{\theta}_T(s) - \theta) + \sqrt{T}G(s,T)\,(M\theta - r)\,.$$

Hence, we have:

$$\left(\sqrt{T}(\hat{\theta}_T(s) - \theta)^\top, \sqrt{T}(\tilde{\theta}_T(s) - \theta)^\top, \sqrt{T}(\hat{\theta}_T(s) - \tilde{\theta}_T)^\top\right)^\top$$

$$= \left(I_{p+1}, I_{p+1} - M^\top G_{[0,T]}^\top, M^\top G_{[0,T]}^\top\right)^\top \sqrt{T}(\hat{\theta}_T(s) - \theta)$$

$$+ \left(0, -\sqrt{T}\,(M\theta - r)^\top G_{[0,T]}^\top, \sqrt{T}\,(M\theta - r)^\top G_{[0,T]}^\top\right)^\top.$$

Then, by combining Eqs. (22.25) and (22.27) as well as Proposition 22.19 along with Slutsky's theorem

$$\left(\sqrt{T}(\hat{\theta}_T(s) - \theta)^\top, \sqrt{T}(\tilde{\theta}_T(s) - \theta)^\top, \sqrt{T}(\hat{\theta}_T(s) - \tilde{\theta}_T(s))^\top\right)^\top \xrightarrow[T\to\infty]{D}$$

$$\left(I_{p+1}, I_{p+1} - M^\top G^{*\top}, M^\top G^{*\top}\right)^\top \rho + \left(0, -r_0^\top G^{*\top}, r_0^\top G^{*\top}\right)^\top$$

$$= \left(\rho^\top, \varrho^\top, \varsigma^\top\right)^\top,$$

where $\rho \sim \mathcal{N}_{2(p+1)}\left(\mathbf{0}, \sigma^2\Sigma^{-1}\right)$. Therefore, the proof follows from classical properties of multivariate normal distribution along with algebraic computations. This completes the proof of the joint asymptotic normality. □

## *Some Important Results with Auxiliary Processes*

### The Case Where the Dimensions Are Positive Integers

In this subsection, we consider observing the process $\{\mathscr{V}(t), t \geqslant 0\}$ which is a solution of SDE:

$$\mathrm{d}\mathscr{V}(t) = (\mathscr{L}(t) - \alpha\mathscr{V}(t))\,\mathrm{d}t + \sigma\sqrt{\mathscr{V}(t)}\mathrm{d}B_t, \quad \mathscr{V}(0) = \mathscr{V}_0, \tag{22.52}$$

where $4\mathscr{L}(t)/\sigma^2 = d$ is integer. Under this condition, the process $\{\mathscr{V}(t), t \geqslant 0\}$ is a sum of independent squared Ornstein-Uhlenbeck processes. Let $W_1, W_2, \cdots, W_d$ be independent Brownian motions. For $j = 1, 2, \cdots, d$, let $Y_j(t)$ be the Ornstein-Uhlenbeck process defined as

$$Y_j(t) = \exp\Big\{-\frac{\alpha}{2}t\Big\}Y_j(0) + \frac{1}{2}\sigma \exp\Big\{-\frac{\alpha}{2}t\Big\}\int_0^t \exp\Big\{\frac{\alpha}{2}s\Big\}\mathrm{d}W_j(s),$$

where

$$\mathrm{d}Y_j(t) = -\frac{\alpha}{2}Y_j(t)\mathrm{d}t + \frac{\sigma}{2}\mathrm{d}W_j(s)\,.$$

Set $\mathscr{V}(t) = \sum_{j=1}^d \big(Y_j(t)\big)^2$, with $\mathscr{V}(0) = \sum_{j=1}^d \big(Y_j(0)\big)^2$.

In the following, we prove that $\mathscr{V}(t)$ is the solution to the GCIR model (22.2) without a change-point. Indeed, by Itô's formula

$$\mathrm{d}Y_j^2(t) = 2Y_j(t)\,\mathrm{d}Y_j(t) + \frac{2}{2!}\,\mathrm{d}\langle Y_j(t)\rangle,$$

then

$$\mathrm{d}Y_j^2(t) = 2Y_j(t)\Big(-\frac{\alpha}{2}Y_j(t)\mathrm{d}t + \frac{\sigma}{2}\mathrm{d}W_j(t)\Big) + \mathrm{d}\langle Y_j(t)\rangle$$
$$\Big(-\alpha Y_j^2(t) + \frac{\sigma^2}{4}\Big)\mathrm{d}t + \sigma Y_j(t)\mathrm{d}W_j(t),$$

and then,

$$\mathrm{d}\mathscr{V}(t) = \mathrm{d}\left(\sum_{j=1}^d \big(Y_j(t)\big)^2\right) = \sum_{j=1}^d \mathrm{d}\big(Y_j(t)\big)^2$$

$$= \sum_{j=1}^d \left(\Big(-\alpha Y_j^2(t) + \frac{\sigma^2}{4}\Big)\mathrm{d}t + \sigma Y_j(t)\mathrm{d}W_j(t)\right).$$

This gives:

$$\mathrm{d}\mathscr{V}(t) = \left(-\alpha\sum_{j=1}^d \big(Y_j(t)\big)^2 + \frac{\sigma^2}{4}d\right)\mathrm{d}t + \sigma\sum_{j=1}^d Y_j(t)\mathrm{d}W_j(t)$$

$$= \big(\mathscr{L}(t) - \alpha\mathscr{V}(t)\big)\,\mathrm{d}t + \sigma\sqrt{\mathscr{V}(t)}\sum_{j=1}^d \frac{Y_j(t)}{\sqrt{\mathscr{V}(t)}}\mathrm{d}W_j(t).$$

Since that $W_j$, $j = 1, 2, \cdots, d$ are independent Brownian motions, we have $\langle W_{j_1}, W_{j_2}\rangle_t = t\,\mathbb{I}(\{j_1 = j_2\})$ where $j_1, j_2 = 1, \ldots, d$. Hence, the quadratic variation of $B$ is given by

$$\langle B\rangle_t = \left\langle \sum_{j=1}^{d} \int_0^{\cdot} \frac{Y_j(s)}{\sqrt{\mathscr{V}(s)}} \mathrm{d}W_j(s), \sum_{j=1}^{d} \int_0^{\cdot} \frac{Y_j(s)}{\sqrt{\mathscr{V}(s)}} \mathrm{d}W_j(s) \right\rangle_t = t.$$

It follows from the Lévy characterization for Brownian motion (Pitman & Yor, 2018, Theorem 22.1) that $B$ is a standard Brownian motion. Then, Eq. (22.52) holds, which means that $\mathscr{V}(t) = \sum_{j=1}^{d} \left(Y_j(t)\right)^2$ is the solution of the GCIR model defined in Eq. (22.2). Under the context of one change-point, consider the following SDE:

$$\mathrm{d}\mathscr{V}(t) = \left((\mathscr{L}_1(t) - \alpha_1 \mathscr{V}(t))\,\mathbb{I}(0 < t \leqslant t^*) + (\mathscr{L}_2(t) - \alpha_2 \mathscr{V}(t))\,\mathbb{I}(t^* < t \leqslant T)\right) \mathrm{d}t$$

$$+\sigma\sqrt{\mathscr{V}(t)}\, dB_t, \quad \mathscr{V}(0) = \mathscr{V}_0, \tag{22.53}$$

where the dimensions $4\mathscr{L}_i(t)/\sigma^2 = d_i$, $i = 1, 2$ are integers. Set

$$\mathscr{V}(t) = \sum_{j=1}^{d_1} \left(Y_{1j}(t)\mathbb{I}(0 < t \leqslant t^*)\right)^2 + \sum_{j=1}^{d_2} \left(Y_{2j}(t)\mathbb{I}(t^* < t \leqslant T)\right)^2,$$

with $\mathscr{V}(0) = \sum_{j=1}^{d_1} \left(Y_{1j}(0)\right)^2$. From the previous discussion, we conclude that $\mathscr{V}(t)$ is the solution to the GCIR model (22.2), which shows that

$$\mathscr{V}(t) = \sum_{j=1}^{d_1} \left(Y_{1j}(t)\mathbb{I}(0 < t \leqslant t^*)\right)^2 + \sum_{j=1}^{d_2} \left(Y_{2j}(t)\mathbb{I}(t^* < t \leqslant T)\right)^2$$

with $\mathscr{V}(0) = \sum_{j=1}^{d_1} \left(Y_{1j}(0)\right)^2$ is a GCIR process.

### Approximate Stationary Process in the Case Where the Dimension Is Integer

To establish the auxiliary process, for $i = 1, 2$, and $j = 1, \ldots, d_i$, let $\{B_{ij}(s), s \geqslant 0\}$ and $\{\bar{B}_{ij}(s), s \geqslant 0\}$ be two independent Brownian motions and $\tilde{B}_{ij}(s)$ be a bilateral Brownian motion with

$$\tilde{B}_{ij}(s) = B_{ij}(s)\mathbb{I}(s > 0) + \bar{B}_{ij}(-s)\mathbb{I}(s \leqslant 0)\,.$$

Let

$$\tilde{Y}_{1j}(t) = \frac{1}{2}\sigma \exp\left\{-\frac{\alpha_1}{2}t\right\} \int_{-\infty}^{t} \exp\left\{\frac{\alpha_1}{2}s\right\} \mathrm{d}\tilde{B}_{ij}(s)$$

and

$$\tilde{Y}_{2j}(t) = \frac{1}{2}\sigma \exp\left\{-\frac{\alpha_2}{2}t\right\} \int_{-\infty}^{t} \exp\left\{\frac{\alpha_2}{2}s\right\} \mathrm{d}\tilde{B}_{ij}(s).$$

For $s \in (-\infty, 0)$, $\tilde{B}_{ij}(s) = \bar{B}_{ij}(-s)$, we rewrite:

$$\tilde{Y}_{1j}(t) = \frac{1}{2}\sigma \exp\left\{-\frac{\alpha_1}{2}t\right\} \left(\int_{-\infty}^{0} \exp\left\{\frac{\alpha_1}{2}s\right\} \mathrm{d}\bar{B}_{ij}(-s) + \int_{0}^{t} \exp\left\{\frac{\alpha_1}{2}s\right\} \mathrm{d}B_{ij}(s)\right) \tag{22.54}$$

and

$$\tilde{Y}_{2j}(t) = \frac{1}{2}\sigma \exp\left\{-\frac{\alpha_2}{2}t\right\} \left(\int_{-\infty}^{0} \exp\left\{\frac{\alpha_2}{2}s\right\} \mathrm{d}\bar{B}_{ij}(-s) + \int_{0}^{t} \exp\left\{\frac{\alpha_2}{2}s\right\} \mathrm{d}B_{ij}(s)\right). \tag{22.55}$$

First, we establish the distribution of $\int_{-\infty}^{0} \exp\{(\alpha_i/2)s\} \mathrm{d}\bar{B}_{ij}(-s)$ for $i = 1, 2$, and $j = 1, \ldots, d_i$. Let $u = -s$,

$$\int_{-\infty}^{0} \exp\{(\alpha_i/2)s\} \mathrm{d}\bar{B}_{ij}(-s) = -\int_{0}^{\infty} \exp -(\alpha_i/2)u \mathrm{d}\bar{B}_{ij}(u).$$

We denote $\mathcal{I}_{ij}(U) = \int_{0}^{U} \exp\{-(\alpha_i/2)u\} \mathrm{d}\bar{B}_{ij}(u)$. Since the integrator $\exp\{-(\alpha_i/2)u\}$ is deterministic, the Itô integral $\mathcal{I}_{ij}(U)$ follows a normal distribution with mean 0 and variance $\int_{0}^{U} \exp\{-\alpha_i u\} du$. One can verify that

$$\int_{-\infty}^{0} \exp\{\alpha_i s\} \mathrm{d}\bar{B}_{-s} \sim \mathcal{N}\left(0, \frac{1}{\alpha_i}\right), \tag{22.56}$$

so, by the independence between $\int_{-\infty}^{0} \exp\left\{\frac{\alpha_i}{2}s\right\} \mathrm{d}\bar{B}_{ij}(-s)$ and $\int_{0}^{t} \exp\left\{\frac{\alpha_i}{2}s\right\} \mathrm{d}B_{ij}(s)$, we have:

$$\begin{aligned}\tilde{Y}_{1j}(t) = {} & \frac{1}{2}\sigma \exp\left\{-\frac{\alpha_1}{2}t\right\} \\ & \times \left(\int_{-\infty}^{0} \exp\left\{-\frac{\alpha_1}{2}t\right\} \mathrm{d}\bar{B}_{ij}(-s) + \int_{0}^{t} \exp\left\{\frac{\alpha_1}{2}s\right\} \mathrm{d}B_{ij}(s)\right)\end{aligned}$$

$$\sim \mathcal{N}\left(0, \frac{\sigma^2}{4}\frac{1}{\alpha_1}\right),$$

$$\tilde{Y}_{2j}(t) = \frac{1}{2}\sigma \exp\left\{-\frac{\alpha_2}{2}t\right\} \times \left(\int_{-\infty}^{0} \exp\left\{-\frac{\alpha_2}{2}t\right\} \mathrm{d}\bar{B}_{ij}(-s) + \int_{0}^{t} \exp\left\{-\frac{\alpha_2}{2}t\right\} \mathrm{d}B_{ij}(s)\right)$$

$$\sim \mathcal{N}\left(0, \frac{\sigma^2}{4}\frac{1}{\alpha_2}\right).$$

Let

$$\tilde{X}(t) = \sum_{j=1}^{d_1} \left(\tilde{Y}_{1j}(t)\mathbb{I}(0 < t \leqslant t^*)\right)^2 + \sum_{j=1}^{d_2} \left(\tilde{Y}_{2j}(t)\mathbb{I}(t^* < t \leqslant T)\right)^2$$

and note that $\tilde{Y}_{1j}(t)$, $\tilde{Y}_{2j}(t)$ are independent. This implies that

$$\tilde{X}(t) \sim \frac{\sigma^2}{4\alpha_1}\chi^2_{d_1} + \frac{\sigma^2}{4\alpha_2}\chi^2_{d_2}$$

and the process $\{\tilde{X}(t), t \geqslant 0\}$ is strictly stationary and ergodic.

### Convergence Between the Solution and the Auxiliary Process

In our analysis, we first derive the following proposition, elucidating the convergence behavior of the distance between $X(t)$ and $\tilde{X}(t)$. This convergence is observed in both the $L^1$ sense and almost surely. Furthermore, our analysis extends to establish the almost sure convergence to 0 of the distance between $1 / X(t)$ and $1 / \tilde{X}(t)$. The proof of the following proposition is similar to the proof of Proposition 3.4 in Lyu (2023) and Proposition 22.14.

**Proposition 22.28** *If Assumptions 22.1–22.3 hold, then:*

*(1)* $\tilde{X}(t) - X(t) \xrightarrow[t\to\infty]{a.s.} 0$;

*(2)* $\tilde{X}(t) - X(t) \xrightarrow[t\to\infty]{L^1} 0$;

*(3)* $0 < \sup_{t\geqslant 0} \mathrm{E}\left[1/\tilde{X}_t\right] < \infty$;

*(4)* $\frac{1}{\tilde{X}(t)} - \frac{1}{X(t)} \xrightarrow[t\to\infty]{a.s.} 0.$

***Proof*** The proof of Part (1) follows from Itô's isometry, $L^2$-bounded martingale convergence theorem, Markov's inequality, and Borel-Cantelli lemma along with some algebraic computations. Part (2) is established by proving that

$$\mathrm{E}\left[|\tilde{X}(t) - X(t)|^m\right] \leqslant \exp\Big\{-\frac{\alpha_i}{2}mt\Big\}d^{m/2}\sqrt{K_{2*}K_{3*}} \xrightarrow[t\to\infty]{} 0. \tag{22.57}$$

This finishes the proof of $\tilde{X}(t) - X(t) \xrightarrow[t\to\infty]{L^m} 0$. Further, the proof of Parts (3)-(4) follows from Part (1) along with the fact that $\lim_{N\to\infty} \mathrm{P}\big(\tilde{X}(t)X(t) \leqslant (1/N)\big) = 0$. This completes the proof. □

**Acknowledgments** The authors would like to thank the referees for helpful comments and useful insights. Further, Dr. S. Nkurunziza would like to acknowledge the financial support received from Discovery Grant (RGPIN-2019-04464) of the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica, 66*(1), 47–78.

Chernova, O., Dehtiar, O., Mishura, Y., & Ralchenko, K. (2024). Rate of convergence of discretized drift parameters estimators in the Cox–Ingersoll–Ross model. *Communications in Statistics—Theory and Methods, 53*(13), 4857–4879.

Cox, J. C., Ingersoll, J. E., Jr., & Ross, S. A. (1985a). An international general equilibrium model of asset prices. *Econometrica, 53*(2), 363–384.

Cox, J. C., Ingersoll, J. E., Jr., & Ross, S. A. (1985b). A theory of the term structure of interest rates. *Econometrica, 53*(2), 385–408.

Dehling, H., Franke, B., Kott, T., & Kulperger, R. (2014). Change point testing for the drift parameters of a periodic mean reversion process. *Statistical Inference for Stochastic Processes, 17*, 1–18.

Duffie, D., & Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance, 6*(4), 379–406.

Geiß, C., & Manthey, R. (1994). Comparison theorems for stochastic differential equations in finite and infinite dimensions. *Stochastic Processes and Their Applications, 53*(1), 23–35.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society, 57*(2), 357–384.

Hull, J., & White, A. (1990). Pricing interest-rate-derivative securities. *The Review of Financial Studies, 3*(4), 573–592.

Ikeda, N., & Watanabe, S. (1989). *Stochastic Differential Equations and Diffusion Processes*. North-Holland.

Lyu, Y. (2023). *Inference in Generalized Mean Reverting Processes*. PhD thesis, University of Windsor (Canada).

Lyu, Y., & Nkurunziza, S. (2023). Inference in generalized exponential O-U processes. *Statistical Inference for Stochastic Processes, 26*(3), 581–618.

Panik, M. J. (2017). *Stochastic Differential Equations: An Introduction with Applications in Population Dynamics Modeling*. Wiley.

Pitman, J., & Yor, M. (2018). A guide to Brownian motion and related stochastic processes. *arXiv:1802.09679*.

Maghsoodi, Y. (1996). Solution of the extended CIR term structure and bond option valuation. *Mathematical Finance, 6*(1), 89–109.

Mao, X. (2007). *Stochastic Differential Equations and Applications*. Elsevier.

Nkurunziza, S., & Zhang, P. P. (2018). Estimation and testing in generalized mean-reverting processes with change-point. *Statistical Inference for Stochastic Processes, 21*(1), 191–215.

Röckner, M., Schmuland, B., & Zhang, X. (2008). Yamada-Watanabe theorem for stochastic evolution equations in infinite dimensions. *Condensed Matter Physics, 11*(2), 247–259.

# Chapter 23
# Estimating and Pretesting in Additive Censored Models




**Sévérien Nkurunziza**

## 23.1 Introduction

The Cox proportional hazards (PH) model is one of the most celebrated regression models in the analysis of event history data. Also known as the multiplicative intensity model (MI), Andersen and Gill (1982) used a counting processes formulation, in the early 1980s, to introduce an extension of the Cox PH model. In short, the PH model and its variants assume that the intensity function of the counting process defined by the events of interest is composed of the product of a baseline nonparametric intensity function and a parametric part consisting of a function of linear combination of the independent variables. In this case, the effects of the covariates are measured through the unknown coefficients of the linear combination, which do not depend on time. This property is well-known as the proportional hazards assumption which gives an attractive interpretation in terms of risk ratio and has the advantageous of being mathematically tractable. Nevertheless, this property is often not satisfied. To overcome this problem, Aalen (1993) introduced an additive regression model whose intensity function depends on covariates as well as the past events. In particular, the corresponding statistical model is a counting process whose variation is a linear regression with time-varying coefficients. As shown in Martinussen and Scheike (2006), Hussein et al. (2014), and references therein, the unrestricted estimator of Aalen's nonparametric time-varying regression coefficients can be obtained via weighted least squares.

In contexts where the target vector parameter is suspected to satisfy some constraints, preliminary test estimators have been studied in the literature. In particular, when some or all components of the parameter are suspected to be insignificant, the

S. Nkurunziza (✉)
Mathematics and Statistics Department, University of Windsor, Windsor, ON, Canada
e-mail: severien@uwindsor.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_23

constraint can be represented as a linear hypothesis on the components. Often, if the restriction is nearly correct, the restricted estimator (RE) is more efficient than the unrestricted estimator (UE). In contrast, if the restriction is completely false, the UE dominates the RE. The so-called shrinkage estimators (SEs) genuinely combine the restriction with the sample information. For more information about the advantages of SEs, we refer to Saleh (2006), Chen and Nkurunziza (2015), Ghannam et al. (2022) and Lyu and Nkurunziza (2024), and references therein.

In this chapter, we propose preliminary test estimator under a general linear hypothesis. As an intermediate step, we present the asymptotic joint distribution of the RE and UE. Moreover, to overcome the difficulty due to the time-varying fact, we derive the integrated asymptotic distributional risk. Using this result, we study the relative efficiency of the proposed preliminary test estimators compared to the UE.

The remainder of this chapter is organized as follows. In Sect. 23.2, we define the objective of the chapter. In Sect. 23.3, we present the proposed preliminary test estimators. Finally, in Sect. 23.4, we present some concluding remarks and for the convenience of readers, we also give the proofs of technical results in the Appendix.

## 23.2 Statistical Model and Main Assumptions

In this section, we present the statistical model and the main assumptions for deriving the results of this chapter. The section is organized into three subsections. In Sect. 23.2.1, we present the statistical model and the notations used in this chapter. In Sect. 23.2.2, we present the main assumptions used for establishing the results of this chapter. Section 23.2.3 gives the unrestricted estimator (UE) and the restricted estimator (RE) that constitute the preliminary results.

### *23.2.1 Statistical Model and Notation*

In this chapter, we are interested in the estimation problem concerning the cumulative regression coefficients of Aalen's additive model. In particular, we consider a family of counting processes $\{N_i(t), 0 \leqslant t \leqslant \tau\}_{i=1,\ldots,n}$ with unknown intensity functions $(\lambda_i(t))_{i=1,\ldots,n}$, where for every $i = 1, \ldots, n$, $N_i(t)$ describes the number of events up to time $t$. For mathematical conveniences, let $\{\mathfrak{F}_{it}, t \geqslant 0\}$ be a family of $\mathfrak{S}$-fields generated by $\{N_i(t); t \geqslant 0\}$, $i = 1, \ldots, n$, let $\mathfrak{F}_t$ be the $\mathfrak{S}$-field generated by $\bigcup_{i=1}^{n} \mathfrak{F}_{it}$. We consider that, for $0 \leqslant t \leqslant \tau$, $\lambda_i(t)$ is a function of an unknown time-varying $k$-column vector denoted by $\beta(t)$. In particular, let $g_i(\cdot)$ be a known and nonnegative real-valued function such that $\int_0^t g_i(\beta(s))\mathrm{d}s < \infty$ almost surely

(a.s.) for $i = 1, \ldots, n$ and for all $0 \leqslant t \leqslant \tau$, set $\lambda_i(t) = g_i(\beta(t))$, $i = 1, \ldots, n$. We also set:

$$N_i(t) = \Lambda_i(t) + M_i(t), \quad 0 \leqslant t \leqslant \tau, \tag{23.1}$$

for $i = 1, \ldots, n$, where $\Lambda_i(t)$ and $M_i(t)$ represent the compensator and the martingale parts, respectively. In the sequel, we only deal with the so-called absolute continuous case for which

$$\Lambda_i(t) = \int_0^t \lambda_i(s)\mathrm{d}s = \int_0^t g_i(\beta(s))\mathrm{d}s, \quad 0 \leqslant t \leqslant \tau, \tag{23.2}$$

for $i = 1, \ldots, n$. Further, as in most survival models, we consider that an event history data occurs in the form of a triplets $(N_i(t), Y_i(t), X_i(t))$ for a sample of individuals $i = 1, \ldots, n$, where $X_i(t)$ is a $k$-column vector of covariates, and $Y_i(t)$ is the risk indicator function. For the sake of simplicity, we assume that $\{(X_i(t), Y_i(t)); t \geqslant 0\}_{i=1,\ldots,n}$ is a family of processes adapted to the filtration $\{\mathfrak{F}_t; t \geqslant 0\}$, such that the functions $g_i(t)$, $i = 1, \ldots, n$ satisfy the conditions given above. Thereafter, as in Martinussen and Scheike (2006), we focus on the case where the function $g_i(z)$ corresponds to the well-known Aalen's additive regression model. Namely, we assume that the covariates $X_i(t)$, $i = 1, \ldots, n$ are $\{\mathfrak{F}_t; t \geqslant 0\}$-predictable and bounded, and let $g_i(z) = Y_i(t)X_i^\top(t)z$, for $i = 1, \ldots, n$. Furthermore, to simplify some mathematical expressions, let $N(t) = (N_1(t), \ldots, N_n(t))^\top$, $\lambda(t) = (\lambda_1(t), \ldots, \lambda_n(t))^\top$, $M(t) = (M_1(t), \ldots, M_n(t))^\top$, $\Lambda(t) = (\Lambda_1(t), \ldots, \Lambda_n(t))^\top$, and let $X(t) = (Y_1(t)X_1(t), \ldots, Y_n(t)X_n(t))^\top$. By combining Eqs. (23.1) and (23.2), we obtain:

$$\mathrm{d}N(t) = X(t)\beta(t)\mathrm{d}t + \mathrm{d}M(t), \quad 0 \leqslant t \leqslant \tau. \tag{23.3}$$

The main objective of this chapter is to develop an efficient statistical method that genuinely combines the statistical model defined in Eq. (23.3) with the solution of the testing problem underlying the restriction on $\beta(\cdot)$ in order to estimate $B(t) = \int_0^t \beta(s)\mathrm{d}s$, $0 \leqslant t \leqslant \tau$.

The proposed method involves operations such as supremums and infimums of an uncountable collection of random vectors. Thus, to guarantee that the resulting quantities are random vectors, we assume, for technical considerations, that all processes under consideration are separable. Nevertheless, every stochastic process is equivalent to a separable process (see Billingsley, 1995), and because of that, this restriction is without loss of generality. Let $\mathcal{D}\left([0,\ \tau], \mathbb{R}^{\mathrm{P}}\right)$ be the space that consists of càdlàg functions on $[0, \tau]$ into $\mathbb{R}^{\mathrm{P}}$ and endowed with the sup-norm Skorohod topology (see Billingsley, 1995). For the sake of simplicity, we let $\mathcal{D}([0, \tau])$ stand for $\mathcal{D}\left([0, \tau], \mathbb{R}^{\mathrm{P}}\right)$.

### 23.2.2 The Hypothesized Restriction and Main Assumptions

As mentioned above, we consider an estimation problem concerning the cumulative regression coefficients defined as $B_j(t) = \int_0^t \beta_j(u)du$, for $j = 1, \ldots, k$. We also consider the scenario where prior knowledge indicates that some covariates may not be statistically significant. In particular, we consider the scenario where the cumulative coefficients are suspected to satisfy the restriction: $\{\beta(t) : R\beta(t) = r_1(t);\ 0 \leqslant t \leqslant \tau\}$ where $R$ is a known $q \times k$-matrix of full rank with $q < k$, and $r_1(t)$ is a known $q$-column vector Riemann integrable on a compact subset of $\mathbb{R}^q$. Equivalently, one can consider the following constraint:

$$\left\{B(t) : RB(t) = r_2(t) = \int_0^t r_1(s)\mathrm{d}s;\ 0 \leqslant t \leqslant \tau\right\}. \tag{23.4}$$

The constraint defined in Eq. (23.4) yields the testing problem:

$$\begin{aligned} &H_0 : RB(t) = r_2(t) = \int_0^t r_1(s)\mathrm{d}s;\ 0 \leqslant t \leqslant \tau \quad \text{versus} \\ &H_1 : RB(t) \neq r_2(t);\ \text{ for some } t. \end{aligned} \tag{23.5}$$

In this chapter, we propose a statistical method that combines the information from the observations with the solution of the testing problem defined in Eq. (23.5). To simplify the presentation of this paper, let $X(t) = (X_1(t), \ldots, X_n(t))^\top$, where $X_i(t)$ is a $k$-column vector and $W(t)$ be a $k \times k$ diagonal matrix whose diagonal components are $W_1(t), \ldots, W_k(t)$, where for $i = 1, \ldots, k$, $W_i(t)$ is positive and $\mathfrak{F}_t$-measurable. We set $W(t) = \mathrm{diag}\,(W_1(t), \ldots, W_k(t))$. For establishing the main results, we suppose that the following conditions hold. Thereafter, for simplicity, they will be referred as "Condition $\mathcal{A}$."

**Condition $\mathcal{A}$:**

(i) $\{(X_i(t), N_i(t)),\ 0 \leqslant t \leqslant \tau\}_{i=1}^n$ are independent and identically distributed (i.i.d);
(ii) for $j, s, l = 1, \ldots, k$, $\mathrm{E}\left[\sup_{0\leqslant t\leqslant\tau} \left|W_1^2(t)X_{1j}(t)X_{1s}(t)X_{1l}(t)\right|\right] < \infty$;
(iii) for $j, s = 1, \ldots, k$, $\mathrm{E}\left[\sup_{0\leqslant t\leqslant\tau} \left|W_1(t)X_{1j}(t)X_{1s}(t)\right|\right] < \infty$;
(iv) $\Gamma_n(t) = X^\top(t)W(t)X(t)$ is a.s. a positive-definite matrix, $\forall\, t \in [0, \tau]$;
(v) $\mathrm{E}[W_1(t)X_1(t)X_1^\top(t)]$ is non-singular $\forall\, t \in [0, \tau]$;
(vi) $\int_0^t |\beta(s)|\mathrm{d}s < \infty,\ \forall\, t \in [0, \tau]$.

The above Condition $\mathcal{A}$ is sufficient for the efficiency of the proposed strategy. Nevertheless, to simplify the derivation of some established results, we also require Condition $\mathcal{B}$ given below. To this end, let $\Omega_0(t) = \left(\mathrm{E}[W_1(t)X_1(t)X_1^\top(t)]\right)^{-1}$, and set $\Omega(s) = \Omega_0(s)\mathrm{E}[W_1^2(s)X_1(s)X_1^\top(s)X_1^\top(s)\beta(s)]\Omega_0(s)$, $0 \leqslant s \leqslant \tau$.

**Condition $\mathcal{B}$**

$W_i(t) = 1/\left(X_i^\top(t)\beta(t)\right),\quad i = 1, 2, \ldots, n, 0 \leqslant t \leqslant \tau.$

It should be noted that under Condition $\mathcal{B}$, $\Omega(s) = \Omega_0(s)$ for all $0 \leqslant s \leqslant \tau$. As pointed out one of the reviewers, Conditions $\mathcal{A}$–$\mathcal{B}$ seem strong, and thus, it will be interesting to investigate if the proposed method may work under weaken assumptions. At this point, our investigations do not allow us to weaken the assumptions in Condition $\mathcal{A}$. Nevertheless, Condition $\mathcal{A}$ is similar to the assumptions in Martinussen and Scheike (2006) and Hussein et al. (2014). Further, although the derivation is more challenging, similar results can be established without Condition $\mathcal{B}$. Indeed, without Condition $\mathcal{B}$, the shrinking factor and the restricted estimator will be asymptotically correlated, and in this case, the derivation of the risk of shrinkage and preliminary test estimators is not straightforward. To overcome such difficulties, one may use the identities in Chen and Nkurunziza (2015).

### 23.2.3 *Preliminary Results: The Unrestricted and Restricted Estimators*

Following Aalen (1993), Hussein et al. (2014), and references therein, the unrestricted estimator (UE) of $B(t)$ is the least squares estimator given by

$$\hat{B}(t) = \int_0^t (\Gamma_n(s))^{-1} X^\top(s) W(s) \mathrm{d}N(s). \tag{23.6}$$

Further, as in Hussein et al. (2014) and references therein, by Lagrange multiplier, it can be verified that the *restricted estimator (RE)* of $B(t)$ is given by

$$\tilde{B}(t) = \hat{B}(t) - \int_0^t A_n(s)(R\hat{b}(s) - r_1(s))\mathrm{d}s \tag{23.7}$$

where $\hat{b}(t) = (\Gamma_n(t))^{-1} X(t) W(t) \mathrm{d}N(t)$ and

$$A_n(t) = (\Gamma_n(t))^{-1} R^\top \left[ R \left(\Gamma_n(t)\right)^{-1} R^\top \right]^{-1}.$$

As established in Martinussen and Scheike (2006), Hussein et al. (2014), and references therein, the UE is asymptotically normal. More precisely, in Martinussen and Scheike (2006), it is established that

$$n^{1/2}(\hat{B} - B) \xrightarrow[n\to\infty]{D} U$$

on $\mathcal{D}([0, \tau])$, where $U$ is a Gaussian martingale with mean 0 and covariance-variance $\Sigma_{11}(t) = \int_0^t \Omega(s)\mathrm{d}s$, $0 \leqslant t \leqslant \tau$. By using the preliminary results of this subsection, we derive, in the next section, the asymptotic joint normality of the RE and UE.

## 23.3 The Main Result

In this section, we derive the main results of this chapter. The section is organized into three subsections. Namely, in Sect. 23.3.1, we derive the joint asymptotic normality of the UE and the RE. The established result generalizes the one in Hussein et al. (2014). Further, in Sect. 23.3.2, we present the shrinkage preliminary test estimator as well as the preliminary test estimator. Finally, in Sect. 23.3.3, we study the performance of the proposed estimators.

### *23.3.1 Joint Asymptotic Distribution of the Unrestricted and Restricted Estimators*

In this subsection, we derive the joint asymptotic normality of RE and UE. To this end, note that, under global alternative hypotheses, $\sqrt{n}\left(\tilde{B}(t) - B(t)\right)$ may converge in probability to infinity. To overcome this problem, in the sequel, we consider the following sequence of *local alternative hypotheses*:

$$H_{1,n} : Rb(t) = r_1(t) + \delta_1(t)/\sqrt{n},\ n = 1, 2, \ldots \tag{23.8}$$

where $R$ is a known $q \times k$ full-rank matrix and $r_1(t)$ and $\delta_1(t)$ are known $q \times 1$-vectors, Riemann-integrable on every compact subset of $\mathbb{R}^q$. From Eq. (23.8), one can also consider:

$$H_{2,n} : RB(t) = r_2(t) + \delta_2(t)/\sqrt{n}, \tag{23.9}$$

with $r_2(t) = \int_0^t r_1(s)ds$ and $\delta_2(t) = \int_0^t \delta_1(s)ds$. To simplify the notation, let

$$\rho_n(t) = \sqrt{n}(\hat{B}(t) - B(t)),$$

$$\eta_n(t) = \sqrt{n}(\tilde{B}(t) - B(t)),$$

$$\xi_n(t) = \sqrt{n}(\tilde{B}(t) - \hat{B}(t)),$$

and

$$\delta^*(t) = \int_0^t A(s)\delta_1(s)\mathrm{d}s,$$
$$A(t) = \Omega_0(t)R^\top[R\Omega_0(t)R^\top]^{-1}.$$

In addition, for $0 \leqslant t \leqslant \tau$

$$\begin{aligned}
\Sigma_{11}(t) &= \int_0^t \Omega(s)\mathrm{d}s \\
\Sigma_{21}(t) &= -\int_0^t (A(s)R\Omega(s))^\top \,\mathrm{d}s, \qquad \Sigma_{12}(t) = \Sigma_{21}^\top(t), \\
\Sigma_{31}(t) &= \Sigma_{11}(t) + \Sigma_{21}(t), \\
\Sigma_{22}(t) &= \int_0^t A(s)R\Omega(s)R^\top A^\top(s)\mathrm{d}s, \\
\Sigma_{23}(t) &= \Sigma_{32}^\top(t) \\
\Sigma_{32}(t) &= \Sigma_{22}(t) + \Sigma_{21}(t), \\
\Sigma_{33}(t) &= \Sigma_{11}(t) + \Sigma_{22}(t) + \Sigma_{21}(t) + \Sigma_{21}^\top(t).
\end{aligned}$$

**Proposition 23.1** *If Eq. (23.9) and Condition $\mathcal{A}$ hold, then*

$$((\rho_n(t))^\top, (\xi_n(t))^\top, (\eta_n(t))^\top)^\top = P_{1,n}(t) + P_{2,n}(t) - P_{3,n}(t);$$

*where*

$$\begin{aligned}
P_{1,n}(t) &= n^{-1/2}\int_0^t \Upsilon_n(s)\Omega_0(s)X^\top(s)W(s)\mathrm{d}M(s), \\
P_{2,n}(t) &= n^{-1/2}\int_0^t \Upsilon_n(s)\left\{(n^{-1}\Gamma_n(s))^{-1} - \Omega_0(s)\right\}X^\top(s)W(s)\mathrm{d}M(s), \\
P_{3,n}(t) &= -\int_0^t (0, I_k, I_k)^\top A_n(s)\delta_1(s)\mathrm{d}s;
\end{aligned}$$

*with* $\Upsilon_n(s) = \left(I_k, -R^\top(A_n(s))^\top, I_k - R^\top(A_n(s))^\top\right)^\top$.

The proof is outlined in Appendix Proof of Proposition 23.1.

Following Proposition 23.1, we derive the weak convergence of the vector $((\rho_n(t))^\top, (\xi_n(t))^\top, (\eta_n(t))^\top)^\top$ which implies the joint asymptotic normality of UE and RE. As an intermediate step, we derive below two propositions that give the convergence of $P_{1,n}(t)$, $P_{2,n}(t)$ and $P_{3,n}(t)$ on $\mathcal{D}([0,\tau])$.

**Proposition 23.2** *Assume that Condition $\mathcal{A}$ and Eq. (23.9) hold. Then, $\{P^*_{1,n}(t), t \geqslant 0\}$ converges in distribution to a Gaussian martingale on $\mathcal{D}([0,\tau])$, with mean 0 and covariance function:*

$$\Phi(t) = \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{21}^\top(t) & \Sigma_{31}^\top(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) & \Sigma_{32}^\top(t) \\ \Sigma_{31}(t) & \Sigma_{32}(t) & \Sigma_{33}(t) \end{pmatrix}. \tag{23.10}$$

The proof of this result is given in Appendix Proof of Proposition 23.2. This proposition plays a crucial role in deriving the joint asymptotic normality of the UE and the RE.

**Proposition 23.3** *Suppose that Condition $\mathcal{A}$ and the sequence of local alternative hypotheses in Eq. (23.9) hold. Then,*

$$P_{2,n}(t) \xrightarrow[n\to\infty]{P} 0 \quad \text{and} \quad P_{3,n}(t) \xrightarrow[n\to\infty]{P} -\left(0, (\delta^*(t))^\top, (\delta^*(t))^\top\right)^\top,$$

*uniformly over* $[0,\tau]$.

The proof of this proposition is given by Appendix Proof of Proposition 23.3.

By combining Propositions 23.2 and 23.3, we derive the following result that gives the limiting distribution of $((\rho_n(t))^\top, (\xi_n(t))^\top, (\eta_n(t))^\top)^\top$ on $\mathcal{D}([0,\tau])$.

**Corollary 23.1** *If Condition $\mathcal{A}$ holds along with the sequence of local alternative hypotheses in Eq. (23.9); then, on $\mathcal{D}([0,\tau])$,*

$$((\rho_n(t))^\top, (\xi_n(t))^\top, (\eta_n(t))^\top)^\top \xrightarrow[n\to\infty]{D} ((\rho(t))^\top, (\xi(t))^\top, (\eta(t))^\top)^\top,$$

*where* $\left\{((\rho(t))^\top, (\xi(t))^\top, (\eta(t))^\top)^\top, t \geqslant 0\right\}$ *is a Gaussian martingale with*

$$((\rho(t))^\top, (\xi(t))^\top, (\eta(t))^\top)^\top \sim \mathcal{N}_{3k}\left(\left(0, -(\delta^*(t))^\top, -(\delta^*(t))^\top\right)^\top, \Phi(t)\right),$$

*and $\Phi(t)$ is given in Eq. (23.10).*

The proof follows by combining Propositions 23.1–23.3 along with Slutsky's theorem. Note that Corollary 23.1 generalizes Part 1 of Proposition 2.1 in Hussein et al. (2014). In the special case where both Condition $\mathcal{A}$ and Condition $\mathcal{B}$ hold, we establish the following corollary that is useful in deriving the asymptotic optimality of the proposed estimators. To simplify some notations, for all $0 \leqslant t \leqslant \tau$, let

$$\Sigma^*_{11}(t) = \int_0^t \Omega_0(s)\mathrm{d}s,$$

$$\Sigma_{22}^{*}(t) = \int_0^t A(s) R \Omega_0(s) \mathrm{d}s = -\Sigma_{21}^{*}(t) = \Sigma_{12}^{*}(t),$$
$$\Sigma_{31}^{*}(t) = \Sigma_{11}^{*}(t) + \Sigma_{21}^{*}(t),$$
$$\Sigma_{32}^{*}(t) = \Sigma_{23}^{*^{\top}}(t) = 0,$$
$$\Sigma_{33}^{*}(t) = \Sigma_{11}^{*}(t) - \Sigma_{22}^{*}(t),$$

and thus,

$$\Phi^{*}(t) = \begin{pmatrix} \Sigma_{11}^{*}(t) & -\Sigma_{22}^{*^{\top}}(t) & \Sigma_{11}^{*}(t) + \Sigma_{21}^{*}(t) \\ -\Sigma_{22}^{*}(t) & \Sigma_{22}^{*}(t) & 0 \\ \Sigma_{11}^{*}(t) + \Sigma_{21}^{*}(t) & 0 & \Sigma_{11}^{*}(t) - \Sigma_{22}^{*}(t) \end{pmatrix}. \tag{23.11}$$

**Corollary 23.2** *Under Eq. (23.9) along with Conditions $\mathcal{A}$ and $\mathcal{B}$*

$$((\rho_n(t))^{\top}, (\xi_n(t))^{\top}, (\eta_n(t))^{\top})^{\top} \xrightarrow[n\to\infty]{D} ((\rho(t))^{\top}, (\xi(t))^{\top}, (\eta(t))^{\top})^{\top},$$

*on $\mathcal{D}([0,\ \tau])$, where $\left\{((\rho(t))^{\top}, (\xi(t))^{\top}, (\eta(t))^{\top})^{\top},\ t \geqslant 0\right\}$ is a Gaussian martingale with $((\rho(t))^{\top}, (\xi(t))^{\top}, (\eta(t))^{\top})^{\top} \sim \mathcal{N}_{3k}\left(\left(0, -(\delta^{*}(t))^{\top}, -(\delta^{*}(t))^{\top}\right)^{\top}, \Phi^{*}(t)\right)$, and $\Phi^{*}(t)$ is given in Eq. (23.11).*

***Proof*** The proof follows directly from Corollary 23.1 along with the fact that, under Condition $\mathcal{B}$, $\Omega(t) = \Omega_0(t)$ for all $0 \leqslant t \leqslant \tau$. Then, $\Phi = \Phi^{*}$, this completes the proof. □

### 23.3.2 Shrinkage Preliminary Test Estimator

In this subsection, we present the proposed shrinkage preliminary test estimator (SPE) as well as the preliminary test estimator (PE) for the cumulative regression coefficients. Let

$$J(t) = R^{\top}\left(\int_0^t R \Omega_0(s) R^{\top}\, \mathrm{d}s\right)^{-1} R,$$
$$\Delta(t) = (\delta^{*}(t))^{\top} J(t) \delta^{*}(t),$$
$$\varphi_n(t) = (\xi_n(t))^{\top} \widehat{J}(t) (\xi_n(t))^{\top}, \quad 0 \leqslant t \leqslant \tau,$$

where $\widehat{J}(t)$ is a consistent estimator for $J(t)$, uniformly on $[0, \tau]$. Note that $\varphi_n \xrightarrow[n\to\infty]{D} \xi^\top J\, \xi$ on $\mathcal{D}([0,\ \tau])$ where $(\xi(t))^\top J(t)\xi(t) \sim \chi^2_q(\Delta(t))$, with $\Delta(t) = (\delta^*(t))^\top J(t)\delta^*(t)$ and $q = \text{rank}(\Sigma^*_{11}(t)J(t)\Sigma^*_{11}(t))$. Let $\chi^2_{q;\alpha}$ be a nonnegative real number such that $\text{P}\left(\chi^2_q > \chi^2_{q;\alpha}\right) = \alpha$ and let $\mathbb{I}(A)$ be the indicator function of the event $A$. From the limiting distribution of Eq. (23.12), one derives the solution to the testing problem in Eq. (23.5). Namely, we suggest the following test:

$$\psi = \mathbb{I}\left(\varphi_n(t) > \chi^2_{q;\alpha}\right).$$

Following Saleh (2006), SPE can be defined as

$$\hat{B}^{\text{SP}}(t) = \hat{B}(t)\psi + (1-\psi)\,\hat{B}^{\text{S}}(t) \tag{23.12}$$

where

$$\hat{B}^{\text{S}}(t) = \gamma\,\tilde{B}(t) + (1-\gamma)\hat{B}(t); \quad \text{with} \quad 0 \leqslant \gamma \leqslant 1. \tag{23.13}$$

The constant $\gamma \in [0, 1]$ denotes the shrinkage intensity. It should be noted that for the special case where $\gamma = 0$, we have $\hat{B}^{\text{S}}(t) = 0\tilde{B}(t) + (1-0)\hat{B}(t) = \hat{B}(t)$ and thus, if $\gamma = 0$, the estimator $\hat{B}^{\text{S}}(t)$ becomes the UE. Further, for $\gamma = 1$, from Eq. (23.13), we have $\hat{B}^{\text{S}}(t) = 1 \times \tilde{B}(t) + (1-1)\hat{B}(t) = \tilde{B}(t)$. Thus, for $\gamma = 1$, the estimator $\hat{B}^{\text{S}}(t)$ yields the RE, and in this case, the estimator in Eq. (23.12) becomes:

$$\hat{B}^{\text{P}}(t) = \hat{B}(t)\psi + (1-\psi)\,\tilde{B}(t). \tag{23.14}$$

As mentioned in Saleh (2006), Nkurunziza (2013), and references therein, the estimator $\hat{B}^{\text{P}}$ is known as the preliminary test estimator (PE). In general, the estimator in Eq. (23.13) outperforms the UE in important part of the parameter space. Nevertheless, the main challenge consists in the method for selecting an optimal value of $\gamma$. In practice, one of the methods involves giving an arbitrary value based on the degree of trust in the null hypothesis in Eq. (23.5). For more details, we refer to Judge and Bock (1978), Bickel and Doksum (2001), Saleh (2006), Chiou and Miao (2007), Nkurunziza and Ahmed (2011), and references therein. As a second method, one can choose $\gamma$ in a data-driven fashion by minimizing a suitable risk function. As mentioned in Friedman (1989) among others, another common method that is also computationally intensive approach consists in estimating $\gamma$ by using cross-validation. From a Bayesian perspective, one can also use the empirical Bayes method to estimate $\gamma$. In this scenario, $\gamma$ is considered as a hyper-parameter that can be estimated from the data by optimizing the marginal likelihood.

### 23.3.3 Integrated Asymptotic Distributional Risk (IADR)

In this subsection, we introduce a criterion which is used in order to evaluate the performance of the proposed estimators. Often, in large-sample point estimation, it is usual to compare estimators in terms of their asymptotic distributional risk (ADR) functions. In order to accommodate the time-varying coefficient, we use the integrated quadratic loss function whose expected value is called the integrated quadratic risk. Since it is impossible to obtain the finite sample distribution, the estimators will be compared by using the quantity called integrated asymptotic distributional risk (IADR). Formally, the IADR of an estimator $\hat{B}^*$ is defined as

$$\mathrm{IADR}(\hat{B}^*, B; W^*) = \mathrm{E}[\Psi],$$

where $\Psi$ is the distributional limit of the loss function

$$L(\hat{B}^*, B; W^*) = n \int_0^\tau (\hat{B}^*(s) - B(s))^\top W^*(s)(\hat{B}^*(s) - B(s))\, \mu(\mathrm{d}s),$$

in which $W^*(s)$ is a known weight matrix continuous on $[0, \tau]$ and $\mu$ is a finite measure of $[0, \tau]$. Below, we establish the IADR of the proposed estimators. To simplify the notation, let $\boldsymbol{x}$ be a column vector and

$$\|\boldsymbol{x}\|^2_{\Xi_1} = \boldsymbol{x}^\top \boldsymbol{\Xi}_1 \boldsymbol{x}$$

with $\boldsymbol{\Xi}_1$ a known nonnegative definite matrix. Further, let $h$ be known Borel-measurable and real-valued square-integrable function, and for each $0 \leqslant t \leqslant \tau$, consider the following class (in $h$) of quantity:

$$\boldsymbol{\rho}^*(t, h) = \boldsymbol{\eta}(t) + h\big(\|\boldsymbol{\xi}(t)\|^2_{J(t)}\big)\, \boldsymbol{\xi}(t),$$

where $\boldsymbol{\eta}(t)$ and $\boldsymbol{\xi}(t)$ are defined in Proposition 23.1.

**Lemma 23.1** *Suppose that the conditions $\mathcal{A}$ and $\mathcal{B}$ hold; then,*

$$\begin{aligned}
&\mathrm{E}\left[\int_0^\tau \left((\boldsymbol{\rho}^*(t, h))^\top W^*(t) \boldsymbol{\rho}^*(t, h)\right) \mu(\mathrm{d}t)\right] \\
&\quad = \int_0^\tau \mathrm{tr}\left(\boldsymbol{W}^*(t)\left(\Sigma^*_{11}(t) - \Sigma^*_{22}(t)\right)\right) \mu(\mathrm{d}t) \\
&\quad + \int_0^\tau \left(\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)}\right) \mu(\mathrm{d}t) \\
&\quad - 2\int_0^\tau \mathrm{E}\left[h\left(\chi^2_{q+2}\left((\Delta(t))\right)\right)\right]\left(\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)}\right) \mu(\mathrm{d}t)
\end{aligned}$$

$$+\int_0^\tau \mathrm{E}\left[h^2\left(\chi^2_{q+2}\left((\Delta(t))\right)\right)\right]\mathrm{tr}\left(\boldsymbol{W}^*(t)\Sigma^*_{22}(t)\right)\mu(\mathrm{d}t)$$

$$+\int_0^\tau \mathrm{E}\left[h^2\left(\chi^2_{q+4}\left((\Delta(t))\right)\right)\right]\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)}\mu(\mathrm{d}t).$$

***Proof*** By Theorem 2.7.1 of Lehmann and Romano (2005), we conclude that $\mathrm{E}\left[(\boldsymbol{\rho}^*(t,h))^\top W^*(t)\boldsymbol{\rho}^*(t,h)\right]$ is integrable on $[0,\ \tau]$. Therefore, the proof follows by combining Lemma 2.1 in Nkurunziza (2011) and Fubini Theorem. □

Let for $x \geqslant 0$, $\mathrm{H}_\nu(x\ ;\Delta) = \mathrm{P}(\chi^2_\nu(\Delta) \leq x)$.

**Theorem 23.1** *Suppose that the conditions $\mathcal{A}$ and $\mathcal{B}$ hold, along with the sequence of local alternative hypotheses given by Eq. (23.9). Then,*

$$\mathrm{IADR}(\hat{B}, B; W^*) = \int_0^\tau \mathrm{tr}\left(W^*(t)\Sigma^*_{11}(t)\right)\mu(\mathrm{d}t);$$

$$\mathrm{IADR}(\tilde{B}, B; W^*) = \mathrm{IADR}\left(\hat{B}, B; W^*\right) - \int_0^\tau \mathrm{tr}\left(\Sigma^*_{22}(t)W^*(t)\right)\mu(\mathrm{d}t) + \int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)}\mu(\mathrm{d}t);$$

$$\mathrm{IADR}\left(\hat{B}^S, B; W^*\right) = \mathrm{IADR}(\hat{B}, B; W^*) + \gamma^2\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)}\mu(\mathrm{d}t) - \gamma(2-\gamma)\int_0^\tau \mathrm{tr}\left(\Sigma^*_{22}(t)W^*(t)\right)\mu(\mathrm{d}t);$$

$$\mathrm{IADR}\left(\hat{B}^{\mathrm{P}}, B; W^*\right) = \mathrm{IADR}\left(\hat{B}, B; W^*\right) + 2\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} H_{q+2}\left(\chi^2_{\alpha,q}; \Delta(t)\right)\mu(\mathrm{d}t) - \int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} H_{q+4}\left(\chi^2_{\alpha,q}; \Delta(t)\right)\mu(\mathrm{d}t) - \int_0^\tau \mathrm{tr}\left(\boldsymbol{W}^*(t)\boldsymbol{\Sigma}^*_{22}(t)\right) H_{q+2}(\chi^2_{\alpha,q}; \Delta(t))\mu(\mathrm{d}t);$$

$$\mathrm{IADR}\left(\hat{B}^{\mathrm{SP}}, B; W^*\right) = \mathrm{IADR}\left(\hat{B}, B; W^*\right) + 2\gamma\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} H_{q+2}\left(\chi^2_{\alpha,q}; \Delta(t)\right)\mu(\mathrm{d}t)$$

$$-\gamma(2-\gamma)\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} H_{q+4}\left(\chi^2_{\alpha,q};\,\Delta(t)\right)\mu(\mathrm{d}t)$$
$$-\gamma(2-\gamma)\int_0^\tau \mathrm{tr}\left(\boldsymbol{W}^*(t)\boldsymbol{\Sigma}^*_{22}(t)\right) H_{q+2}\left(\chi^2_{\alpha,q};\,\Delta(t)\right)\mu(\mathrm{d}t).$$

***Proof*** The IADR of the UE and the RE are trivial. Also, the IADR of $\hat{B}^{\mathrm{S}}$ follows directly from Lemma 23.1 by taking $h(x) = (1-\gamma)$, $x \geqslant 0$. Further, by taking

$$h(x) = (1-\gamma) + \gamma\mathbb{I}\left(x > \chi^2_{\alpha,q}\right), \qquad x \geqslant 0,$$

and applying Lemma 23.1, we get the IADR of $\hat{B}^{\mathrm{SP}}$. Finally, the IADR of $\hat{B}^{\mathrm{P}}$ is the IADR of $\hat{B}^{\mathrm{SP}}$ with $\gamma = 1$ and this completes the proof. □

By using Theorem 23.1, we derive below a corollary, which gives the performance of the proposed shrinkage and preliminary test estimators. To this end, for $0 \leqslant t \leqslant \tau$, we set:

$$f_1(\gamma, \Delta(t)) = \frac{H_{q+2}\left(\chi^2_{q,\alpha};\,\Delta(t)\right)}{2H_{q+2}\left(\chi^2_{q,\alpha};\,\Delta(t)\right) - (2-\gamma)H_{q+4}\left(\chi^2_{q,\alpha};\,\Delta(t)\right)}$$

and

$$f_2(\gamma, \Delta(t)) = \frac{H_{q+2}\left(\chi^2_{q,\alpha};\,\Delta(t)\right)}{2H_{q+2}\left(\chi^2_{q,\alpha};\,\Delta(t)\right) - (1-\gamma)\,H_{q+4}\left(\chi^2_{q,\alpha};\,\Delta(t)\right)}.$$

**Corollary 23.3** *Assume that $\gamma > 0$. If the Assumptions of 23.1 hold, then*

*(i)* $\mathrm{IADR}\left(\hat{B}^{\mathrm{S}}, B; W^*\right) \leqslant \mathrm{IADR}\left(\hat{B}, B; W^*\right)$ *whenever for* $0 \leqslant t \leqslant \tau$,

$$\int_0^\tau \|\delta(t)\|^2_{W^*(t)}\,\mu(\mathrm{d}t) \leqslant \frac{2-\gamma}{\gamma}\int_0^\tau \mathrm{tr}\left(\Sigma_{22}(t)W^*\right)\mu(\mathrm{d}t),$$

*(ii)* $\mathrm{IADR}\left(\hat{B}^{\mathrm{SP}}, B; W^*\right) \leqslant \mathrm{IADR}\left(\hat{B}, B; W^*\right)$ *whenever for* $0 \leqslant t \leqslant \tau$,

$$\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \leqslant (2-\gamma)\,\mathrm{tr}\left(W^*(t)\Sigma^*_{22}(4)\right) f_1(\gamma;\,\Delta(t)),$$

*(iii)* $\mathrm{IADR}\left(\hat{B}^{\mathrm{P}}, B; W^*\right) \leqslant \mathrm{IADR}\left(\hat{B}, B; W^*\right)$ *whenever for* $0 \leqslant t \leqslant \tau$,

$$\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \leqslant \frac{\mathrm{tr}\left(\boldsymbol{W}\boldsymbol{\Sigma}^*_{22}(t)\right) H_{q+2}(\chi^2_{\alpha,q};\,\Delta(t))}{2H_{q+2}\left(\chi^2_{\alpha,q};\,\Delta(t)\right) - H_{q+4}\left(\chi^2_{\alpha,q};\,\Delta(t)\right)};$$

*(iv)* $\text{IADR}\left(\hat{B}^{\text{P}}, B; W^*\right) \leqslant \text{IADR}\left(\hat{B}^{\text{SP}}, B; W^*\right)$ *whenever for* $0 \leqslant t \leqslant \tau$,

$$\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \leqslant (1-\gamma)\, f_2(\gamma, \Delta(t)),$$

*and* $\text{IADR}\left(\hat{B}^{\text{P}}, B; W^*\right) \geqslant \text{IADR}\left(\hat{B}^{\text{SP}}, B; W^*\right)$ *whenever for* $0 \leqslant t \leqslant \tau$, $\left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \geqslant (1-\gamma)\, f_2(\gamma, \Delta(t))$.

The proof follows from Theorem 23.1 along with some algebraic computations.

**Remark 23.1** Note that the IADR of $\hat{B}$ is a constant function of $\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \mu(\mathrm{d}t)$, while the IADR $\left(\hat{B}^{\text{S}}, B; W^*\right)$ is linear in $\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \mu(\mathrm{d}t)$, and for $0 < \gamma \leqslant 1$, it intersects the IADR $\left(\hat{B}, B; W^*\right)$ if and only if

$$\int_0^\tau \left\|\boldsymbol{\delta}^*(t)\right\|^2_{\boldsymbol{W}^*(t)} \mu(\mathrm{d}t) = \frac{2-\gamma}{\gamma} \int_0^\tau \text{tr}\left(\boldsymbol{W}^*(t)\boldsymbol{\Sigma}_{22}(t)\right) \mu(\mathrm{d}t).$$

Thus, near the null hypothesis, $\hat{B}^{\text{S}}$ dominates $\hat{B}$. It is also noticed that if $\gamma$ is negligible, $\hat{B}^{\text{S}}$ still dominates $\hat{B}$, and as the significance level $\alpha$ tends to 0, $\text{IADR}\left(\hat{B}^{\text{SP}}, B; W^*\right)$ tends to $\text{IADR}\left(\hat{B}^{\text{S}}, B; W^*\right)$. Thus, $\hat{B}^{\text{S}}$ dominates $\hat{B}$ for small values of $\gamma$ or near the null hypothesis. As $\alpha$ approaches 1, $\text{IADR}\left(\hat{B}^{\text{SP}}, B; W^*\right)$ tends to $\text{IADR}\left(\hat{B}, B; W^*\right)$, and the same result holds if $\Delta(t)$ tends to infinity. Further, from Part (iv) of Corollary 23.3 and as in Nkurunziza (2013), it should be noted that under the null hypothesis, $\hat{B}^{\text{P}}$ dominates $\hat{B}^{\text{SP}}$. However, as mentioned in Nkurunziza (2013), as one moves away from the null hypothesis, $\hat{B}^{\text{SP}}$ performs better than $\hat{B}^{\text{P}}$. Overall, as pointed out in Nkurunziza (2013), $\hat{B}^{\text{SP}}$ has a good control over IADR whether or not the constraint is violated.

## 23.4 Conclusion

In this chapter, we studied an estimation problem concerning the regression coefficients of Aalen's additive hazards regression model. We proposed robust estimators, which genuinely combine the information from the sample and a consistent test of the regression coefficient. Further, we established the joint asymptotic normality of the UE and RE of the cumulative hazards regression coefficients under the sequence of local alternative hypotheses. In order to accommodate the time-varying nature, we derived the IADR of the proposed estimators, and thanks to the established results, we established the risk dominance of the proposed estimators. As pointed out one of the reviewers, some of the computations may be challenging. Nevertheless, the complexity of the computation is similar to the one in Hussein

et al. (2014), and in their conclusion, Hussein et al. (2014) mention that their proposed method is computationally inexpensive.

**Acknowledgments** The author would like to thank the referees and editor for helpful comments and useful insights. Further, the author would like to acknowledge the financial support received from Discovery Grant (RGPIN-2019-04464) of the Natural Sciences and Engineering Research Council of Canada (NSERC).

## Appendix

## Proof of Proposition 23.1

By some computations, we have:

$$\rho_n(t) = \eta_n(t) - \xi_n(t)$$

$$\xi_n(t) = -\sqrt{n}\int_0^t A_n(s)(Rb(s) - r_1(s))\mathrm{d}s - \int_0^t A_n(s)R[\sqrt{n}(\hat{b}(s) - b(s))]\mathrm{d}s$$

$$\eta_n(t) = \sqrt{n}(\hat{B}(t) - B(t)) - \sqrt{n}\int_0^t A_n(s)(Rb(s) - r_1(s))\mathrm{d}s$$
$$- \int_0^t A_n(s)R[\sqrt{n}(\hat{b}(s) - b(s))]\mathrm{d}s.$$

Then, under Eq. (23.8), and as in Martinussen and Scheike (2006), we have:

$$\xi_n(t) = -\int_0^t A_n(s)\delta_1(s)\mathrm{d}s - \int_0^t A_n(s)R \times n^{-\frac{1}{2}}\Omega_0(s)X(s)W(s)\mathrm{d}M(s)$$
$$- \int_0^t A_n(s)R \times n^{-\frac{1}{2}}\left\{(n^{-1}\Gamma_n(s))^{-1} - \Omega_0(s)\right\}X(s)W(s)\mathrm{d}M(s),$$

$$\eta_n(t) = n^{-\frac{1}{2}}\int_0^t \Omega_0(s)X(s)W(s)\mathrm{d}M(s) + \xi_n(t)$$
$$+n^{-\frac{1}{2}}\int_0^t \left\{(n^{-1}\Gamma_n(s))^{-1} - \Omega_0(s)\right\}X(s)W(s)\mathrm{d}M(s).$$

This gives:

$$\xi_n(t) = -\int_0^t A_n(s)\delta_1(s)\mathrm{d}s - \int_0^t A_n(s)R \times n^{-\frac{1}{2}}\Omega_0(s)X(s)W(s)\mathrm{d}M(s)$$
$$- \int_0^t A_n(s)R \times n^{-\frac{1}{2}}\left\{(n^{-1}\Gamma_n(s))^{-1} - \Omega_0(s)\right\}X(s)W(s)\mathrm{d}M(s),$$

$$\eta_n(t) = n^{-\frac{1}{2}} \int_0^t \Omega_0(s) X(s) W(s) \mathrm{d}M(s) + \xi_n(t)$$

$$+n^{-\frac{1}{2}} \int_0^t \left\{ (n^{-1}\Gamma_n(s))^{-1} - \Omega_0(s) \right\} X(s) W(s) \mathrm{d}M(s),$$

$$\rho_n(t) = n^{-\frac{1}{2}} \int_0^t \Omega_0(s) X(s) W(s) \mathrm{d}M(s)$$

$$+n^{-\frac{1}{2}} \int_0^t \left\{ (n^{-1}\Gamma_n(s))^{-1} - \Omega_0(s) \right\} X(s) W(s) \mathrm{d}M(s).$$

The rest of the proof follows from some algebraic computations. □

## Proof of Proposition 23.2

In proof of Proposition 23.2, we use the following lemma.

**Lemma 23.2** *Let* $\widetilde{Q}_{ji}(t) = \sum_{l=1}^{k} \{\Upsilon_n(t)\Omega_0(t)\}_{jl} W_i(t) X_{il}(t)$ *and assume that Condition* $\mathcal{A}$ *holds. Then,* $\sup_{0\leqslant t\leqslant\tau} |n^{-1/2}\widetilde{Q}_{ji}(t)| \leqslant \widetilde{H}_n$ *where* $\widetilde{H}_n$ *converges in probability to* 0 *as n tends to* $\infty$. *In particular,*

$$\sup_{0\leqslant t\leqslant\tau} \left| n^{-\frac{1}{2}} \widetilde{Q}_{ji}(t) \right| \xrightarrow[n\to\infty]{P} 0.$$

***Proof*** We have $\widetilde{Q}_{ji}(t) = \sum_{h=1}^{k} \{\Upsilon_n(t)\}_{jh} \left(\widetilde{V}_{hi}(t)\right)$ where

$$\widetilde{V}_{hi}(t) = \sum_{l=1}^{k} \{\Omega_0(t)\}_{hl} W_i(t) X_{il}(t).$$

Then,

$$\sup_{0\leqslant t\leqslant\tau} |n^{-\frac{1}{2}} \widetilde{Q}_{ji}(t)| \leqslant \sup_{0\leqslant t\leqslant\tau} \left| \sum_{h=1}^{k} \{\Upsilon_n(t)\}_{jh} \right| \sup_{0\leqslant t\leqslant\tau} \left| (n^{-\frac{1}{2}} \widetilde{V}_{hi}(t)) \right|.$$

Then, using the result in Martinussen and Scheike (2006), we prove that

$$\sup_{0\leqslant t\leqslant\tau}\left|n^{-\frac{1}{2}}\widetilde{Q}_{ji}(t)\right|\leqslant\widetilde{H}_n,\qquad\text{where}\qquad\widetilde{H}_n=\sum_{h=1}^{k}\sup_{0\leqslant t\leqslant\tau}\left|\{\Upsilon_n(t)\}_{jh}\right|\times H_n,\tag{23.15}$$

with

$$H_n\equiv k\cdot\sup_{t,h,l}|(\Omega_0^{-1}(t))_{hl}^{-1}|\,n^{-\frac{1}{2}}\sup_{t,i,l}|W_i(t)X_{il}(t)|\xrightarrow[n\to\infty]{P}0.\tag{23.16}$$

Hence, the proof follows from the fact that from Eqs. (23.15) and (23.16), we have:

$$\widetilde{H}_n\xrightarrow[n\to\infty]{P}0\quad\Longrightarrow\sup_{0\leqslant t\leqslant\tau}\left|n^{-\frac{1}{2}}\widetilde{Q}_{ji}(t)\right|\xrightarrow[n\to\infty]{P}0.\tag{23.17}$$

□

Proof of Proposition 23.2

The predictable variation of $P_{1,n}(t)$ is

$$\langle P_{1,n}\rangle(t)=\int_0^t\Upsilon_n(s)\Omega_0(s)\Gamma_n^*(s)\Omega_0(s)\Upsilon_n^\top(s)\mathrm{d}s,$$

with $\Gamma_n^*(s)=\frac{1}{n}\sum_{i=1}^n W_i^2(s)[X_i(s)X_i^\top(s)]\lambda_i(s)$. Then, by some algebraic computations, we prove that

$$\sup_{0\leqslant t\leqslant\tau}|\langle P_{1,n}\rangle(t)-\Phi(t)|\xrightarrow[n\to\infty]{P}0,\qquad\text{i.e.}\qquad\langle P_{1,n}\rangle(t)\xrightarrow[n\to\infty]{P}\Phi(t),\tag{23.18}$$

uniformly on $[0,\tau]$. Now, let $P_{1j\varepsilon,n}(t)$ denote the process containing all jumps with absolute value greater than $\varepsilon$, for $\varepsilon>0$, of the $j$-th component of $P_{1,n}$. We have:

$$P_{1j\varepsilon,n}(t)=\sum_{i=1}^n\int_0^t n^{-\frac{1}{2}}\widetilde{Q}_{ji}(s)\;I\left(|n^{-\frac{1}{2}}\widetilde{Q}_{ji}(s)|>\varepsilon\right)\mathrm{d}M_i(s)$$

where $\widetilde{Q}_{ji}(s)$ is defined in Lemma 23.2 and $I(A)$ is the indicator function of $A$. The predictable variation of this process is

$$\langle P_{1j\varepsilon,n}\rangle(t)=\sum_{i=1}^n\int_0^t\left[n^{-\frac{1}{2}}\widetilde{Q}_{ji}(s)\right]^2\;I\left(|n^{-\frac{1}{2}}\widetilde{Q}_{ji}(s)|>\varepsilon\right)\lambda_i(s)\mathrm{d}s,$$

and then, by using Lemma 23.2 along with some algebraic, we prove that

$$\langle P_{1j\varepsilon,n}\rangle(t) \xrightarrow[n\to\infty]{a.s.} 0, \quad 0 \leqslant t \leqslant \tau. \tag{23.19}$$

The rest of the proof follows from Eqs. (23.18) and (23.19) along with the martingale central limit. □

## Proof of Proposition 23.3

One can verify that $P_{2,n}$ is a locally square integrable $\mathfrak{F}_t$-martingale whose predictable variation process is

$$\langle P_{2,n}\rangle(t) = n^{-1}\int_0^t \Upsilon_n(s)V(s)\,\mathrm{diag}(\lambda_i(s))(V(s))^\top(\Upsilon_n(s))^\top\,\mathrm{d}s.$$

The $j$ th component of $\langle P_{2,n}\rangle(t)$ is $\langle P_{2j,n}\rangle(t) = n^{-1}\sum_{i=1}^n\int_0^t Q_{ji}^2(s)\lambda_i(s)\mathrm{d}s$, where

$$Q_{ji}(t) = \sum_{l=1}^k\left[\sum_{h=1}^k\{\Upsilon_n(t)\}_{jh}\left\{(n^{-1}\Gamma_n(t))^{-1} - \Omega_0(t)\right\}_{hl}\right]X_{il}(t)W_i(t).$$

By Cauchy-Schwartz inequality along with some algebraic computations, we get:

$$\langle P_{2j,n}\rangle(\tau) \leqslant \sup_{0\leqslant t\leqslant\tau}\left(\sum_{h=1}^k\{\Upsilon_n(t)\}_{jh}^2\right)\left[\int_0^\tau\sum_{h=1}^k G_h(t)\mathrm{d}t\right],$$

where $G_h(t) = \sum_{l=1}^k\sum_{u=1}^k\left\{(n^{-1}\Gamma_n(t))^{-1} - \Omega_0(t)\right\}_{hu}^2\left[n^{-1}\sum_{i=1}^n(X_{il}(t)W_i(t))^2\right]\lambda_i(t)$. Then,

$$\langle P_{2j,n}\rangle(\tau) \leqslant \left[\sum_{h=1}^k\left(\sup_{0\leqslant t\leqslant\tau}|\{\Upsilon_n(t)\}_{jh}|\right)^2\right]\left[\int_0^\tau\sum_{h=1}^k G_h(t)\mathrm{d}t\right]. \tag{23.20}$$

Further, in the similar way as in Martinussen and Scheike (2006), one can verify that

$$\int_0^\tau\sum_{h=1}^k G_h(t)\mathrm{d}t \xrightarrow{P} 0. \tag{23.21}$$

Hence, using Eqs. (23.20) and (23.21), we prove that $\langle P_{2j,n}\rangle(t) \xrightarrow[n\to\infty]{P} 0$. Then, by Lenglart's inequality, $\sup_{0\leqslant t\leqslant\tau}|P_{2j}(t)| \xrightarrow[n\to\infty]{P} 0$, this proves the first statement. The rest of the proof follows from Gill's Lemma along with the fact that $A_n(t)$ converges in probability to $A(t)$, uniformly on $[0, \tau]$. □

## References


Aalen, O. O. (1993). Further results on the non-parametric linear model in survival analysis. *Statistics in Medicine, 12*, 1569–1588.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics, 10*(4), 1100–1120.

Bickel, V. P., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics* (2nd ed.). Prentice Hall.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). Wiley.

Chiou, P., & Miao, W. (2007). Shrinkage estimation for the difference between a control and treatment mean. *Journal of Statistical Computation and Simulation, 77*, 651–662.

Chen, F., & Nkurunziza, S. (2015). Optimal method in multiple regression with structural changes. *Bernoulli, 21*(4), 2217–2241.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association, 84*(405), 165–175.

Ghannam, M., & Nkurunziza, S. (2022). Improved estimation in tensor regression with multiple change-points. *Electronic Journal of Statistics, 16*(2), 4162–4221.

Hussein, A. A., Nkurunziza, S., & Tomanelli, K. (2014). Nonparametric shrinkage estimation for Aalen's additive hazards model. *Australian & New Zealand Journal of Statistics, 56*(1), 15–26.

Judge, G. G., & Bock, M. E. (1978). *The statistical implications of pre-test and Stein-rule estimators in econometrics*. Wiley.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer.

Lyu, Y., & Nkurunziza, S. (2023). Inference in generalized exponential O-U processes with change-point. *Statistical Inference for Stochastic Processes, 27*(1), 63–102.

Martinussen, T., & Scheike, T. H. (2006). *Dynamic regression models for survival data*. Springer.

Nkurunziza, S. (2011). Shrinkage strategy in stratified random sample subject to measurement error. *Statistics & Probability Letters, 81*(2), 317–325.

Nkurunziza, S. (2013). Preliminary test and estimation in some multi-factor diffusion processes. *Sankhyā A, 75*(2), 211–230.

Nkurunziza, S., & Ahmed, S. E. (2011). Estimation strategies for the regression coefficient parameter matrix in multivariate multiple regression. *Statistica Neerlandica, 65*(4), 387–406.

Saleh, A. K. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. John Wiley & Sons.

# Chapter 24
# Confidence Interval for a Univariate Normal Mean Based on a Pretest Estimator




**Nanami Taketomi, Jia-Han Shih, and Takeshi Emura**

## 24.1 Introduction

The preliminary test (pretest) estimator is an estimator that incorporates a preliminary hypothesis test (Bancroft, 1944; Saleh, 2006). The pretest estimator for a univariate normal mean (denoted as $\mu$) under a known variance (denoted as $\sigma^2$) is the simplest and most fundamental estimator among various types of pretest estimators. For the estimation of a univariate normal mean, the theoretical properties of the pretest estimator, such as the bias and mean squared error, have been studied by Khan and Saleh (2001), Magnus (2002), and Shih et al. (2023).

However, the construction of a confidence interval (CI) has rarely been investigated. The reason is that the research on the pretest estimation is focused on the risk comparison with other estimators under a decision-theoretic framework. In particular, the literature of pretest estimators has devoted itself to the evaluation of the bias and mean squared error (Aldeni et al., 2023; Hoque et al., 2009; Khan & Saleh, 2001; Kibria & Saleh, 2006; Magnus, 2000, 2002; Norouzirad & Arashi, 2019; Shih et al., 2021, 2023) while little attention has been paid to interval

N. Taketomi
Clinical Research Center in Hiroshima, Hiroshima University Hospital, Hiroshima, Japan
e-mail: nnmamikrn@gmail.com

J.-H. Shih
Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung, Taiwan
e-mail: jhshih@math.nsysu.edu.tw

T. Emura (✉)
Department of Fundamental Statistical Mathematics, The Institute of Statistical Mathematics, Tokyo, Japan

Biostatistics Center, Kurume University, Fukuoka, Japan
e-mail: takeshiemura@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_24

estimation, such as the development of the CIs. Another reason is that the sampling distribution of the pretest estimator is complex, a mixture of continuous and discrete distributions (Shih et al., 2023; Taketomi et al., 2024). This means that the traditional Wald-type CI which relies on normal approximation is invalid.

Recently, Taketomi et al. (2024) derived a valid CI for a normal mean in the context of meta-analyses (Fleiss, 1993). However, their formula for the CI in meta-analyses is limited to a sample size of one ($n = 1$). In this chapter, we extend this approach to the estimation of a univariate normal mean, where $n \geq 1$. By pivoting the cumulative distribution function (CDF) of the pretest estimator (Casella & Berger, 2002; Taketomi et al., 2024), we define an explicit formula for the CI with a closed-form coverage probability. We show that the coverage probability of the CI controls the nominal confidence level both theoretically and numerically. Furthermore, we investigate asymptotic behaviors under the large sample size with fixed $\sigma$ or the large deviation with fixed $n$. The calculation of the proposed CI is implemented in an R package called `uni.shrinkage`. Finally, a dataset is analyzed to illustrate the application of the proposed CI.

The chapter is constructed as follows: Sect. 24.2 provides mathematical backgrounds. Section 24.3 proposes a CI for a normal mean based on the pretest estimator and derives its theoretical properties. The asymptotic behavior of the proposed CI is investigated through simulation studies in Sect. 24.4. In Sect. 24.5, the proposed CI is applied to a real dataset. Finally, Sect. 24.6 concludes with a discussion of the results. Appendices provide detailed mathematical derivations.

## 24.2 Background

In this section, we first review some mathematical notation and basic definitions related to the pretest estimator. We also derive some new theoretical results that will be useful for subsequent analyses.

Let $N(\mu, \sigma^2)$ be a normal distribution with mean $\mu \in (-\infty, \infty)$ and variance $\sigma^2$, where $\sigma > 0$. Especially, $N(0, 1)$ denotes the standard normal distribution, whose CDF is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} \mathrm{d}u, \quad z \in (-\infty, \infty).$$

For $0 < p < 1$, let $z_p = \Phi^{-1}(1 - p)$ be the upper $p$-th quantile of $N(0, 1)$.

For a sample size $n \in \{1, 2, \ldots\}$ and $i \in \{1, 2, \ldots, n\}$, let $Y_i$ be a random variable that follows a normal distribution $Y_i \sim N(\mu, \sigma^2)$, where $\sigma > 0$ is known. The observed data consist of independent samples $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The target quantity to be estimated is the unknown $\mu$.

One commonly used estimator for $\mu$ is the sample mean $\bar{Y} = \sum_{i=1}^{n} Y_i / n$. However, there may be alternative estimators for $\mu$, which is generally denoted by $\delta(\boldsymbol{Y})$. Let $[L(\boldsymbol{Y}), U(\boldsymbol{Y})]$ be any random interval for $\mu$. For instance,

$$\left[\bar{Y} - z_{\gamma/2}\sigma / \sqrt{n},\ \bar{Y} + z_{\gamma/2}\sigma / \sqrt{n}\right]$$

is the Wald-type confidence interval (CI) that achieves the coverage probability (CP) of $100(1-\gamma)\%$ where $0 \le \gamma \le 1$.

In some settings, it is believed that $\mu$ is close to zero or exactly equal to zero. In such cases, the null hypothesis $\mathcal{H}_0 : \mu = 0$ is accepted by a preliminary test, then one defines $\delta(\boldsymbol{Y}) = 0$. This leads to a pretest estimator $\delta^{\mathrm{PT}}(\boldsymbol{Y}) = \bar{Y}\, I\left(\left|\sqrt{n}\,\bar{Y} / \sigma\right| > z_{\alpha/2}\right)$ for a two-sided level $\alpha \in (0, 1)$. This idea was first proposed by Bancroft (1944) and has since been studied by Emura et al. (2024), Judge and Bock (1978), Khan and Saleh (2001), Magnus (2000), Magnus et al. (2011), Saleh (2006), Shih et al. (2021), Shih et al. (2023).

A primary challenge for $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ is the construction of a CI. For $n = 1$, Taketomi et al. (2024) argued that the Wald-type CI is not suitable, as the distribution of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ is asymmetric and has a discrete component. However, for $n \ge 2$, a valid CI has not been derived in the literature.

To address this gap, we first derive the CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ and investigate its asymptotic properties. Additionally, we derive the mean and variance of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$. These preliminary results are given in the following subsections (Sects. 24.2.1 and 24.2.2).

### *24.2.1 The CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$*

After some calculus, we derive the following result.

**Theorem 24.1** *For $n \ge 1$, the CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ is given by*

$$F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \mathrm{P}(\delta^{\mathrm{PT}}(\boldsymbol{Y}) \le d) \tag{24.1}$$

$$= \begin{cases} \Phi\left(\dfrac{d-\mu}{\sigma / \sqrt{n}}\right) & \text{if} \quad d \le -z_{\alpha/2}\sigma / \sqrt{n} \text{ or } z_{\alpha/2}\sigma / \sqrt{n} \le d, \\ \Phi\left(\dfrac{-\mu}{\sigma / \sqrt{n}} - z_{\alpha/2}\right) & \text{if} \quad -z_{\alpha/2}\sigma / \sqrt{n} < d < 0, \\ \Phi\left(\dfrac{-\mu}{\sigma / \sqrt{n}} + z_{\alpha/2}\right) & \text{if} \quad 0 \le d < z_{\alpha/2}\sigma / \sqrt{n}. \end{cases}$$

The proof is outlined in Appendix The Proof of Theorem 24.1: The CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$.

Shih et al. (2023) previously obtained an alternative form of the CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ as

$$\mathrm{P}(\delta^{\mathrm{PT}}(\boldsymbol{Y}) \leq d) = p_A \, \mathrm{P}(\bar{Y} \leq d \mid A) + (1 - p_A) I(0 \leq d), \tag{24.2}$$

where $A = \left\{ \boldsymbol{Y} : \left| \sqrt{n} \bar{Y} / \sigma \right| > z_{\alpha/2} \right\}$ and

$$p_A = \mathrm{P}(A) = 1 - \left[ \Phi(-\sqrt{n}\mu / \sigma + z_{\alpha/2}) - \Phi(-\sqrt{n}\mu / \sigma - z_{\alpha/2}) \right].$$

Although the two formulas for the CDF are equivalent, the proposed formula in Theorem 24.1 is more concrete than Eq. (24.2) and convenient for subsequent analyses. Theorem 24.1 is a more general result than that of Taketomi et al. (2024), as it allows for the study of the CDF for the sample size $n \geq 1$ and asymptotic $n \to \infty$.

Figure 24.1 shows the graph of $F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d)$ against $d \in (-10, 10)$ for $\mu \in \{-4, -2, 0, 2\}$, $\sigma = 4$, $n \in \{10, 20, 100\}$. We observe that $F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d)$ is nondecreasing and has a jump at $d = 0$. The distribution is asymmetric unless $\mu = 0$.

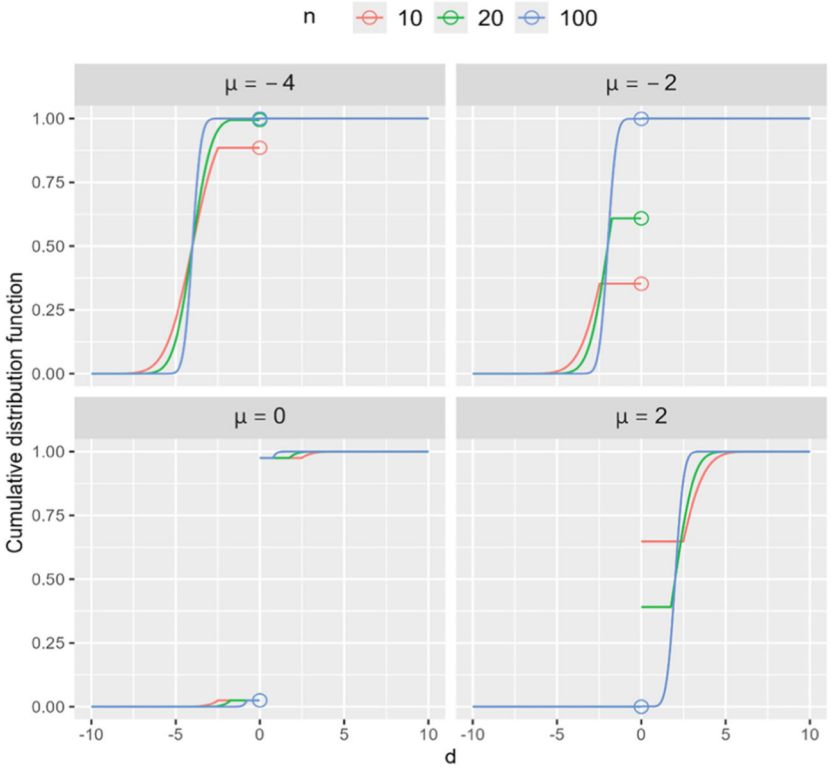



**Fig. 24.1** The graph of $F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d)$ under $d \in (-10, 10)$ for $\mu \in \{-4, -2, 0, 2\}$, $\sigma = 4$, $n \in \{10, 20, 100\}$

These observations suggest that $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ does not adapt to the asymptotic normal approximation, even for the case of $n \to \infty$. We also observe that $F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d)$ steeply increases or jumps near $\mu$, implying that $F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d)$ is distributed around the true value $\mu$ or exactly equal to $\mu$.

#### 24.2.1.1 Mean and Variance of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$

For the subsequent discussions, let $\lambda = \sqrt{n}\mu / \sigma$ be the departure parameter, which represents a standardized difference between $\mu$ and 0. The following lemma gives the mean and variance of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$.

**Lemma 24.1** *The mean and variance of* $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ *are given by*

$$\begin{aligned}
\mathrm{E}\left[\delta^{\mathrm{PT}}(\boldsymbol{Y})\right] &= \frac{\sigma}{\sqrt{n}}\left[\phi(z_{\alpha/2}-\lambda) - \phi(z_{\alpha/2}+\lambda) + \lambda p_A\right], \\
\mathrm{var}\left[\delta^{\mathrm{PT}}(\boldsymbol{Y})\right] &= \frac{\sigma^2}{\sqrt{n}} C_{\lambda,\alpha} - \mathrm{E}\left[\delta^{\mathrm{PT}}(\boldsymbol{Y})\right]^2,
\end{aligned}$$

*where* $\phi(x) = d\Phi(x)/dx$ *is the probability density function of* $N(0,1)$ *and*

$$C_{\lambda,\alpha} = \left(z_{\alpha/2}+\lambda\right)\phi(z_{\alpha/2}-\lambda) + \left(z_{\alpha/2}-\lambda\right)\phi(z_{\alpha/2}+\lambda) + (1+\lambda^2)p_A.$$

***Proof*** The proof of Lemma 24.1 is based on Corollary 1 of Shih et al. (2023). Additionally, Magnus (2000) provides the mean squared error of the pretest estimator. □

### *24.2.2 Asymptotic Properties of* $\delta^{\mathrm{PT}}(\boldsymbol{Y})$

Theorem 24.1 is useful for studying the asymptotic behavior of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ for $n \to \infty$.

**Corollary 24.1** *If* $\mu = 0$,

$$\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases} 0 & \text{if} \quad d < 0, \\ 1 - \dfrac{\alpha}{2} & \text{if} \quad d = 0, \\ 1 & \text{if} \quad d > 0. \end{cases}$$

*If* $\mu \neq 0$,

$$\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases} 0 & \text{if} \quad d < \mu, \\ \frac{1}{2} & \text{if} \quad d = \mu, \\ 1 & \text{if} \quad d > \mu. \end{cases}$$

The proof is provided in Appendix The Proof of Corollary 24.1: The Limit of the CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ as $n \to \infty$.

Corollary 24.1 implies that $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ converges "in distribution" to a random variable with a single point mass at $\mu$ (see Definition 5.5.10 of Casella and Berger, 2002, for the convergence in distribution).

Corollary 24.1 shows that the CDF converges at all continuity points of $I(d \geq \mu)$, namely,

$$\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = I(d \geq \mu) \qquad \forall d \neq \mu.$$

This also means that $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ converges "in probability" to $\mu$ according to Theorem 5.5.12 of Casella and Berger (2002). Therefore, $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ is a consistent estimator for $\mu$.

We also investigate the limit of the CDF as $\sigma \to \infty$ for fixed $n$. To do so, we consider a countable sequence $\{\sigma_m = m, m = 1, 2, \ldots\}$ such that $\sigma_m \uparrow \infty$ as $m \to \infty$. Then, the limiting set of the four ranges of $d$ in Theorem 24.1 is derived as

$$\bigcap_{m=1}^{\infty} \left(-\infty, -z_{\alpha/2}\,\sigma_m/\sqrt{n}\right] = \varnothing, \qquad \bigcup_{m=1}^{\infty} \left(-z_{\alpha/2}\,\sigma_m/\sqrt{n}, 0\right) = (-\infty, 0),$$

$$\bigcup_{m=1}^{\infty} \left[0, z_{\alpha/2}\,\sigma_m/\sqrt{n}\right) = [0, \infty), \qquad \bigcap_{m=1}^{\infty} \left[z_{\alpha/2}\,\sigma_m/\sqrt{n}, \infty\right) = \varnothing,$$

where $\varnothing$ denotes the empty set. Thus, only two ranges, $d \in (-\infty, 0)$ and $d \in [0, \infty)$, need to be considered for the limit. Therefore, for any $\mu$, we have:

$$\lim_{\sigma\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases} \lim_{\sigma\to\infty} \Phi\left(-\dfrac{\mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) & \text{if} \quad d < 0 \\ \lim_{\sigma\to\infty} \Phi\left(-\dfrac{\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) & \text{if} \quad d \geq 0 \end{cases}$$

$$= \begin{cases} \dfrac{\alpha}{2} & \text{if} \quad d < 0, \\ 1 - \dfrac{\alpha}{2} & \text{if} \quad d \geq 0. \end{cases}$$

This result implies that $\delta^{\mathrm{PT}}(\boldsymbol{Y}) = 0$ with probability $1 - \alpha$, $\delta^{\mathrm{PT}}(\boldsymbol{Y}) = -\infty$ with probability $\alpha/2$, and $\delta^{\mathrm{PT}}(\boldsymbol{Y}) = \infty$ with probability $\alpha/2$ for $\sigma \to \infty$. When $\sigma \to \infty$, the value of $\bar{Y}$ takes either $-\infty$ or $\infty$ and never takes any finite value. On the other hand, $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ can control the probability of taking a finite value of zero by adjusting $\alpha$, which may be an advantage of reducing the variability in large deviation settings.

The following Theorems 24.2 and 24.3 state that the asymptotic normality may not hold.

**Theorem 24.2** *If* $\mu = 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y})}{\sigma/\sqrt{n}} \le d\right) = \begin{cases} \Phi(d) & \text{if} \quad d \le -z_{\alpha/2} \text{ or } z_{\alpha/2} \le d, \\ \dfrac{\alpha}{2} & \text{if} \quad -z_{\alpha/2} < d < 0, \\ 1-\dfrac{\alpha}{2} & \text{if} \quad 0 \le d < z_{\alpha/2}. \end{cases}$$

*If* $\mu \neq 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu}{\sigma/\sqrt{n}} \le d\right) = \Phi(d), \qquad -\infty < d < \infty.$$

The proof of Theorem 24.2 can be found in Appendix The Proof of Theorem 24.2: The CDF of $\sqrt{n}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu\right)/\sigma$.

Theorem 24.2 indicates that if $\mu = 0$, the asymptotic distribution of $\sqrt{n}\delta^{\mathrm{PT}}(\boldsymbol{Y})/\sigma$ is a mixture of $N(0,1)$ and a point mass at zero. If $\mu \neq 0$, $\sqrt{n}(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu)/\sigma$ is asymptotically normal.

An alternative approach to obtaining the asymptotic distribution involves standardizing $\delta^{\mathrm{PT}}(\mathbf{Y})$ using its mean and standard deviation, as

$$\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mathrm{E}\left[\delta^{\mathrm{PT}}(\boldsymbol{Y})\right]}{\mathrm{sd}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})\right)},$$

where $\mathrm{sd}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})\right) = \sqrt{\mathrm{var}[\delta^{\mathrm{PT}}(\boldsymbol{Y})]}$. By applying Lemma 24.1, the next theorem is derived similarly to Theorem 24.2.

**Theorem 24.3** *If* $\mu = 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y})}{\mathrm{sd}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})\right)} \le d\right) = \begin{cases} \Phi(\sqrt{C_{0,\alpha}}d) & \text{if} \quad |\sqrt{C_{0,\alpha}}\,d| \ge z_{\alpha/2}, \\ \dfrac{\alpha}{2} & \text{if} \quad -z_{\alpha/2} < \sqrt{C_{0,\alpha}}\,d < 0, \\ 1-\dfrac{\alpha}{2} & \text{if} \quad 0 \le \sqrt{C_{0,\alpha}}\,d < z_{\alpha/2}. \end{cases}$$

*where* $C_{0,\alpha} = 2z_{\alpha/2}\phi(z_{\alpha/2}) + \alpha$. *If* $\mu \neq 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mathrm{E}\left[\delta^{\mathrm{PT}}(\boldsymbol{Y})\right]}{\mathrm{sd}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})\right)} \le d\right) = \Phi(d), \qquad -\infty < d < \infty.$$

If $\mu = 0$, the asymptotic distribution of $\delta^{\mathrm{PT}}(\boldsymbol{Y}) / \mathrm{sd}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})\right)$ using Theorem 24.3 is a mixture of $N(0, C_{0,\alpha}^{-1})$ and a point mass at zero. This result aligns naturally with the fact that $\mathrm{var}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})\right) = \sigma^2 C_{0,\alpha} / n$ when $\mu = 0$. Note that $C_{0,\alpha} < 1$ for all $0 < \alpha < 1$.

Figure 24.2 shows the empirical CDF and asymptotic CDF of $\sqrt{n}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu\right) / \sigma$ for $\mu \in \{-2, 0, 0.1, 2\}$ with $\sigma = 4$ and $n = 100$. The empirical CDF was computed by 10,000 replicates. The results show that the empirical CDF closely matches the asymptotic CDF, validating the derived formulas in Theorem 24.2. For $\mu = -2, 0.1, 2$, the empirical CDF aligns well with the standard normal distribution.

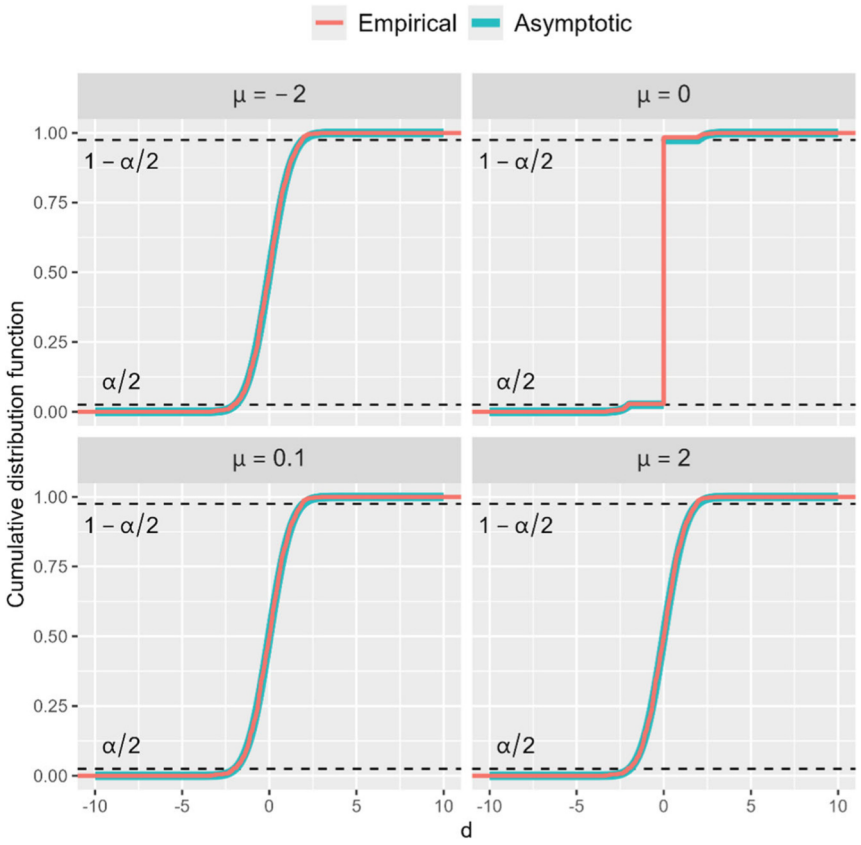



**Fig. 24.2** The graph of the empirical CDF and asymptotic CDF of $\sqrt{n}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu\right) / \sigma$ for $\mu \in \{-2, 0, 0.1, 2\}$; $\sigma = 4$; $n = 100$

## 24.3 Proposed Confidence Interval

This section proposes a $100(1-\gamma)\%$ CI based on $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ for $\gamma \in (0, 1)$ under $n \geq 1$. To achieve this, we extend the valid CI obtained by Taketomi et al. (2024) for $n = 1$ to the case of $n \geq 1$. The idea involves treating $g(\mu, \boldsymbol{Y}) = F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y})|\mu\right)$ as a "pivot," a statistic whose distribution is known and independent of $\mu$. By choosing an interval for $g(\mu; \boldsymbol{Y})$ with probability $(1-\gamma)$, and then inverting the lower and upper bounds, we obtain a desired CI.

Regarding this well-known pivot method, see Lemma 24.2 in Appendix The Derivation of the Pivot CI, reproduced from Theorem 9.2.14 of Casella and Berger (2002). By applying the pivot method, we define the following CI (derivation given in Appendix The Derivation of the Pivot CI).

**Definition 24.1 (The Pivot CI)** The $100(1-\gamma)\%$ pivot CI based on $\delta^{\mathrm{PT}} = \delta^{\mathrm{PT}}(\boldsymbol{Y})$ is defined as

$$\left[L^{\mathrm{PT}}(\boldsymbol{Y}), U^{\mathrm{PT}}(\boldsymbol{Y})\right]$$
$$= \begin{cases} \left[\delta^{\mathrm{PT}} - z_{\gamma_2}\frac{\sigma}{\sqrt{n}}, \delta^{\mathrm{PT}} + z_{\gamma_1}\frac{\sigma}{\sqrt{n}}\right] & \text{if} \quad \delta^{\mathrm{PT}} < -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \\ \left[-\left(z_{\alpha/2} + z_{\gamma_2}\right)\frac{\sigma}{\sqrt{n}}, \left(z_{\alpha/2} + z_{\gamma_1}\right)\frac{\sigma}{\sqrt{n}}\right] & \text{if} \quad \delta^{\mathrm{PT}} = 0, \\ \left[\delta^{\mathrm{PT}} - z_{\gamma_2}\frac{\sigma}{\sqrt{n}}, \delta^{\mathrm{PT}} + z_{\gamma_1}\frac{\sigma}{\sqrt{n}}\right] & \text{if} \quad \delta^{\mathrm{PT}} > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \end{cases}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ satisfy $\gamma_1 + \gamma_2 = \gamma$. If $\gamma_1 = \gamma_2 = \gamma/2$, a symmetric CI is

$$\left[L^{\mathrm{PT}}(\boldsymbol{Y}), U^{\mathrm{PT}}(\boldsymbol{Y})\right]$$
$$= \begin{cases} \left[\delta^{\mathrm{PT}} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}, \delta^{\mathrm{PT}} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\right] & \text{if} \quad \delta^{\mathrm{PT}} < -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \\ \left[-\left(z_{\alpha/2} + z_{\gamma/2}\right)\frac{\sigma}{\sqrt{n}}, \left(z_{\alpha/2} + z_{\gamma/2}\right)\frac{\sigma}{\sqrt{n}}\right] & \text{if} \quad \delta^{\mathrm{PT}} = 0, \\ \left[\delta^{\mathrm{PT}} - z_{\gamma/2}\frac{\sigma}{\sqrt{n}}, \delta^{\mathrm{PT}} + z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\right] & \text{if} \quad \delta^{\mathrm{PT}} > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \end{cases}$$

In most of statistical analyses, the confidence level is generally chosen to be 95% (i.e., $\gamma = 0.05$) with equal allocations $\gamma_1 = \gamma_2 = \gamma/2 = 0.025$ (Ben-Shachar et al., 2023; Fagerland et al., 2015; Fleiss, 1993). Accordingly, our simulations and data analyses will focus on $\gamma/2 = \gamma_1 = \gamma_2 = 0.025$. However, other confidence levels, such as 90% (i.e., $\gamma = 0.1$ with $\gamma_1 = \gamma_2 = \gamma/2 = 0.05$), will be explored in simulations for comparison.

### 24.3.1 Coverage Probabilities

Next, we compute the coverage probabilities for $\left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right] = \left[L^{\mathrm{PT}}(\boldsymbol{Y}), U^{\mathrm{PT}}(\boldsymbol{Y})\right]$. Recall that $\lambda = \sqrt{n}\mu / \sigma$. After some calculus, we obtain the following result.

**Theorem 24.4** *The CP of the* $100(1-\gamma)\%$ *CI is one of the three cases:*

*(Case 1)* *for* $2z_{\alpha/2} < z_{\gamma_1} + z_{\gamma_2}$,

$$\mathrm{P}(\mu \in [L^{\mathrm{PT}}, U^{\mathrm{PT}}]) = \begin{cases} 1-\gamma & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma_2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \gamma_1 & \text{if} \quad -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} - z_{\gamma_2}, \\ 1-\gamma & \text{if} \quad z_{\alpha/2} - z_{\gamma_2} < \lambda < -z_{\alpha/2} + z_{\gamma_1}, \\ 1-\gamma_2 - \Phi(-\lambda + z_{\alpha/2}) & \text{if} \quad -z_{\alpha/2} + z_{\gamma_1} \le \lambda \le z_{\alpha/2} + z_{\gamma_1}, \\ 1-\gamma & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma_1}, \end{cases}$$

*(Case 2)* *for* $2z_{\alpha/2} = z_{\gamma_1} + z_{\gamma_2}$,

$$\mathrm{P}(\mu \in [L^{\mathrm{PT}}, U^{\mathrm{PT}}]) = \begin{cases} 1-\gamma & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma_2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \gamma_1 & \text{if} \quad -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} - z_{\gamma_2}, \\ 1-\gamma_2 - \Phi(-\lambda + z_{\alpha/2}) & \text{if} \quad z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} + z_{\gamma_1}, \\ 1-\gamma & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma_1}, \end{cases}$$

*(Case 3)* *for* $2z_{\alpha/2} > z_{\gamma_1} + z_{\gamma_2}$,

$$\mathrm{P}(\mu \in [L^{\mathrm{PT}}, U^{\mathrm{PT}}]) = \begin{cases} 1-\gamma & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma_2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \gamma_1 & \text{if} \quad z_{\alpha/2} - z_{\gamma_2} \le \lambda \le -z_{\alpha/2} - z_{\gamma_1}, \\ \Phi(-\lambda + z_{\alpha/2}) - \Phi(-\lambda - z_{\alpha/2}) & \text{if} \quad -z_{\alpha/2} - z_{\gamma_1} \le \lambda \le z_{\alpha/2} + z_{\gamma_2}, \\ 1-\gamma_2 - \Phi(-\lambda + z_{\alpha/2}) & \text{if} \quad z_{\alpha/2} + z_{\gamma_2} \le \lambda \le z_{\alpha/2} + z_{\gamma_1}, \\ 1-\gamma & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma_1}. \end{cases}$$

The proof is given in Appendix The Proof of Theorem 24.4.

From Theorem 24.4, the pivot CI has a closed-form expression for the CP. Theorem 24.4 allows unequal allocations of $\gamma_1 > 0$ and $\gamma_2 > 0$ to achieve the confidence level $1 - \gamma = 1 - \gamma_1 - \gamma_2$. However, most statistical applications solely use the equal allocation, where $\gamma_1 = \gamma_2 = \gamma / 2$. From Theorem 24.4, we immediately obtain the next corollary.

**Corollary 24.2** *For a symmetric CI under $\gamma_1 = \gamma_2 = \gamma/2$, the CP of the $100(1-\gamma)\%$ CI is one of the three cases:*

*(Case 1)* *for $\alpha > \gamma$,*

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma/2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \frac{\gamma}{2} & \text{if} \quad -z_{\alpha/2} - z_{\gamma/2} \le \lambda \le z_{\alpha/2} - z_{\gamma/2}, \\ 1-\gamma & \text{if} \quad z_{\alpha/2} - z_{\gamma/2} < \lambda < -z_{\alpha/2} + z_{\gamma/2}, \\ 1 - \frac{\gamma}{2} - \Phi(-\lambda + z_{\alpha/2}) & \text{if} \quad -z_{\alpha/2} + z_{\gamma/2} \le \lambda \le z_{\alpha/2} + z_{\gamma/2}, \\ 1-\gamma & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma/2}, \end{cases}$$

*(Case 2)* *for $\alpha = \gamma$,*

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \lambda < -2z_{\alpha/2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \frac{\gamma}{2} & \text{if} \quad -2z_{\alpha/2} \le \lambda \le 0, \\ 1 - \frac{\gamma}{2} - \Phi(-\lambda + z_{\alpha/2}) & \text{if} \quad 0 \le \lambda \le 2z_{\alpha/2}, \\ 1-\gamma & \text{if} \quad \lambda > 2z_{\alpha/2}, \end{cases}$$

*(Case 3)* *for $\alpha < \gamma$,*

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma/2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \frac{\gamma}{2} & \text{if} \quad -z_{\alpha/2} - z_{\gamma/2} \le \lambda \le -z_{\alpha/2} + z_{\gamma/2}, \\ \Phi(-\lambda + z_{\alpha/2}) - \Phi(-\lambda - z_{\alpha/2}) & \text{if} \quad -z_{\alpha/2} + z_{\gamma/2} < \lambda < z_{\alpha/2} - z_{\gamma/2}, \\ 1 - \frac{\gamma}{2} - \Phi(-\lambda + z_{\alpha/2}) & \text{if} \quad z_{\alpha/2} - z_{\gamma/2} \le \lambda \le z_{\alpha/2} + z_{\gamma/2}, \\ 1-\gamma & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma/2}. \end{cases}$$

The next theorem combines the three cases of Theorem 24.4.

**Theorem 24.5** *For any $\gamma_1 > 0$ and $\gamma_2 > 0$ with $\gamma_1 + \gamma_2 = \gamma$, the confidence coefficient is*

$$\inf_{\mu} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = 1 - \gamma. \tag{24.3}$$

Appendix The Proof of Theorem 24.5 contains the proof.

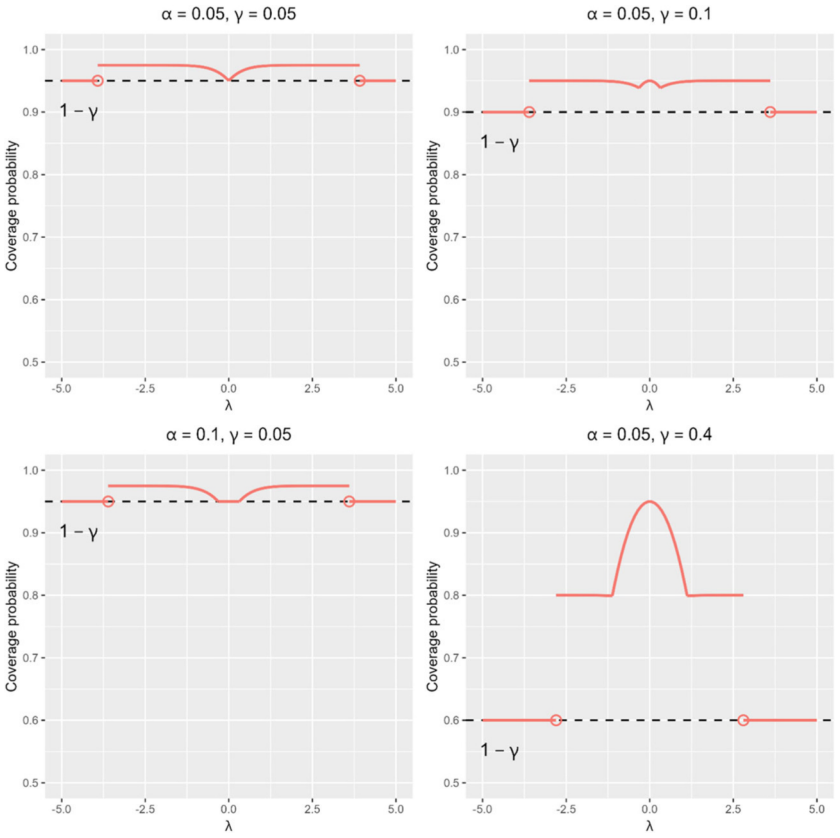



**Fig. 24.3** The CPs of the pivot CI (The nominal value $1-\gamma$ is described by the dashed line)

Figure 24.3 shows the theoretical values of the CPs of the pivot CI (computed using Theorem 24.4). We observe that the CP of the pivot CI is always greater than the nominal level $1-\gamma$, and the confidence coefficient (minimum value of the CP) is equal to $1-\gamma$. This agrees with Theorems 24.4 and 24.5.

### *24.3.2 Asymptotic Property of the CP of the Pivot CI*

We investigate the asymptotic property of the CP of the pivot CI.

**Theorem 24.6 (The Asymptotic CP)** *The limit of the CP of the pivot CI results in the following six mutually exclusive cases.*

*(Case 1)* *for* $0 \leq z_{\alpha/2} - z_{\gamma_2} \leq -z_{\alpha/2} + z_{\gamma_1}$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\gamma_1 & \text{if} \quad \mu = 0, \end{cases}$$

*(Case 2)* *for* $z_{\alpha/2} - z_{\gamma_2} \leq 0 < -z_{\alpha/2} + z_{\gamma_1}$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = 1-\gamma,$$

*(Case 3)* *for* $z_{\alpha/2} - z_{\gamma_2} \leq -z_{\alpha/2} + z_{\gamma_1} < 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\gamma_2 & \text{if} \quad \mu = 0, \end{cases}$$

*(Case 4)* *for* $0 \leq -z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2}$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\gamma_1 & \text{if} \quad \mu = 0, \end{cases}$$

*(Case 5)* *for* $-z_{\alpha/2} + z_{\gamma_1} \leq 0 < z_{\alpha/2} - z_{\gamma_2}$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\alpha & \text{if} \quad \mu = 0, \end{cases}$$

*(Case 6)* *for* $-z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2} < 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\gamma_2 & \text{if} \quad \mu = 0. \end{cases}$$

The proof is thoroughly explained in Appendix The Proof of Theorem 24.6.

**Corollary 24.3** *For any* $\gamma_1 > 0$ *and* $\gamma_2 > 0$ *such that* $\gamma_1 + \gamma_2 = \gamma$, *the confidence coefficient is*

$$\inf_{\mu} \left\{ \lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) \right\} = 1-\gamma.$$

The proof of Corollary 24.3 is given in Appendix The Proof of Theorem 24.5.

**Corollary 24.4** *For any $\alpha$ and $\gamma$, if $\gamma_1 = \gamma_2 = \gamma/2$, the asymptotic CP satisfies:*

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\dfrac{\gamma}{2} & \text{if} \quad \mu = 0. \end{cases}$$

The proof is detailed in Appendix The Proof of Corollary 24.3.

## 24.4 Simulation Studies

We conducted Monte Carlo simulations to confirm the correctness of the asymptotic CP described in Theorem 24.6. To this end, we compared the theoretical asymptotic CP to its Monte Carlo estimates under a variety of scenarios, all in a symmetric setting ($\gamma_1 = \gamma_2 = \gamma/2$). Specifically, we considered the following three scenarios:

Scenario (a): $\alpha = 0.05$, $\gamma = 0.05$ (corresponds to Case 1 in Theorem 24.6),
Scenario (b): $\alpha = 0.05$, $\gamma = 0.10$ (corresponds to Case 5 in Theorem 24.6),
Scenario (c): $\alpha = 0.10$, $\gamma = 0.05$ (corresponds to Case 2 in Theorem 24.6).

To compute the Monte Carlo approximation to the true CP, we generated $Y_1, \ldots, Y_n$ independently from $N(\mu, \sigma^2)$ where $\mu \in \{0, 0.1, 1\}$, $\sigma = 1$, and

$$n \in \{100, 200, 500, 1000, 1200, 1500, 2000, 5000, 10000\}.$$

We then computed the Monte Carlo approximation as

$$\text{Monte Carlo CP} = \frac{1}{10000} \sum_{i=1}^{10000} I\left(\mu \in \left[L_i^{\mathrm{PT}}, U_i^{\mathrm{PT}}\right]\right),$$

where $\left[L_i^{\mathrm{PT}}, U_i^{\mathrm{PT}}\right]$ is the CI in the $i$ th repetition.

Table 24.1 summarizes the results. It shows that the Monte Carlo CP becomes closer to the asymptotic CP as $n \to \infty$. For instance, in Scenario (b), if $\mu = 0.1$ and $n = 10{,}000$, the Monte Carlo CP ( $= 0.899$) is almost equivalent to the asymptotic CP ( $= 0.90$). Overall, the results confirm the validity of the asymptotic CP formula in Theorem 24.6.

**Table 24.1** The results of the Monte Carlo simulation based on 10,000 repetitions

| Scenarios | $n$ | Coverage Probability (CP) | | | | |
|---|---|---|---|---|---|---|
| | | $\mu = 0$ | | $\mu = 0.1, 1$ | $\mu = 0.1$ | $\mu = 1$ |
| | | Asymp[a] | MC[b](SE[c]) | Asymp | MC(SE) | MC(SE) |
| $\alpha = 0.05$, $\gamma = 0.05$ | 100 | 0.95 | 0.9508(0.002) | 0.95 | 0.9726(0.002) | 0.9508(0.002) |
| | 200 | 0.95 | 0.9456(0.002) | 0.95 | 0.9731(0.002) | 0.9456(0.002) |
| | 500 | 0.95 | 0.9492(0.002) | 0.95 | 0.9775(0.001) | 0.9492(0.002) |
| | 1000 | 0.95 | 0.9508(0.002) | 0.95 | 0.9765(0.002) | 0.9508(0.002) |
| | 1200 | 0.95 | 0.9469(0.002) | 0.95 | 0.9751(0.002) | 0.9469(0.002) |
| | 1500 | 0.95 | 0.9477(0.002) | 0.95 | 0.9735(0.002) | 0.9477(0.002) |
| | 2000 | 0.95 | 0.9496(0.002) | 0.95 | 0.9496(0.002) | 0.9496(0.002) |
| | 5000 | 0.95 | 0.9462(0.002) | 0.95 | 0.9462(0.002) | 0.9462(0.002) |
| | 10,000 | 0.95 | 0.9489(0.002) | 0.95 | 0.9489(0.002) | 0.9489(0.002) |
| $\alpha = 0.05$, $\gamma = 0.10$ | 100 | 0.95 | 0.9508(0.002) | 0.90 | 0.9484(0.002) | 0.9028(0.003) |
| | 200 | 0.95 | 0.9456(0.002) | 0.90 | 0.9512(0.002) | 0.8995(0.003) |
| | 500 | 0.95 | 0.9492(0.002) | 0.90 | 0.9534(0.002) | 0.9012(0.003) |
| | 1000 | 0.95 | 0.9508(0.002) | 0.90 | 0.9515(0.002) | 0.9022(0.003) |
| | 1200 | 0.95 | 0.9469(0.002) | 0.90 | 0.9492(0.002) | 0.8955(0.003) |
| | 1500 | 0.95 | 0.9477(0.002) | 0.90 | 0.8953(0.003) | 0.8953(0.003) |
| | 2000 | 0.95 | 0.9496(0.002) | 0.90 | 0.8955(0.003) | 0.8955(0.003) |
| | 5000 | 0.95 | 0.9462(0.002) | 0.90 | 0.8989(0.003) | 0.8989(0.003) |
| | 10,000 | 0.95 | 0.9489(0.002) | 0.90 | 0.8992(0.003) | 0.8992(0.003) |
| $\alpha = 0.10$, $\gamma = 0.05$ | 100 | 0.95 | 0.9508(0.002) | 0.95 | 0.9694(0.002) | 0.9508(0.002) |
| | 200 | 0.95 | 0.9456(0.002) | 0.95 | 0.9722(0.002) | 0.9456(0.002) |
| | 500 | 0.95 | 0.9492(0.002) | 0.95 | 0.9775(0.001) | 0.9492(0.002) |
| | 1000 | 0.95 | 0.9508(0.002) | 0.95 | 0.9765(0.002) | 0.9508(0.002) |
| | 1200 | 0.95 | 0.9469(0.002) | 0.95 | 0.9751(0.002) | 0.9469(0.002) |
| | 1500 | 0.95 | 0.9477(0.002) | 0.95 | 0.9477(0.002) | 0.9477(0.002) |
| | 2000 | 0.95 | 0.9496(0.002) | 0.95 | 0.9496(0.002) | 0.9496(0.002) |
| | 5000 | 0.95 | 0.9462(0.002) | 0.95 | 0.9462(0.002) | 0.9462(0.002) |
| | 10,000 | 0.95 | 0.9489(0.002) | 0.95 | 0.9489(0.002) | 0.9489(0.002) |

[a] Asymptotic CP based on Theorem 24.6
[b] MC: Monte Carlo
[c] $\text{SE} = \sqrt{\text{var}(\text{MC})} = \sqrt{\text{MC}(1 - \text{MC})/10{,}000}$

## 24.5 Data Analysis

We demonstrate the proposed method using the allergic reaction dataset of size $n = 5$, originally presented in Table 1 of Taketomi and Emura (2023). The dataset focuses on allergic reactions among medical students exposed to formaldehyde during an anatomy course in years 2015, 2016, 2018, 2019, and 2022 (Mori & Kakuma, 2022). The dataset consists of the mean difference (MD) of the severity of eye symptoms between male and female students obtained in each year ($i =$

$1, \ldots, 5$), denoted by $(Y_1, \ldots, Y_5) = (-0.17, -0.58, -0.45, -0.75, -0.56)$. That is, $Y_i \sim N(\mu, \sigma^2)$, where $Y_i$ represents the MD (male-female) for each year ($i = 1, \ldots, 5$). Since all students sampled are different across the years, we assume that $Y_i$'s are independent. We will estimate $\mu$ using $Y_1, \ldots, Y_5$ under a fixed $\sigma^2$.

Before analyzing the data, we introduce a function `uni.pt(.)` that can compute the pretest estimate and the proposed CI as defined in Definition 24.1, which is included in our original R package "`uni.shrinkage`." The `uni.pt(.)` can compute $\delta^{\mathrm{PT}}$ (the default pretest level is $\alpha = 0.05$) and CI (the default confidence level is 95%; $\gamma = 0.05$) via the command `uni.pt(y,s)` where `y` is a vector $(Y_1, \ldots, Y_n)$ and `s` is a known value for $\sigma$. The input arguments for the function are as follows:

- `alpha`: significance level $\alpha$ ($0 < \alpha < 1$). The default value is 0.05.
- `gamma`: $1 -$ confidence level ($\gamma$). The default value is 0.05 (confidence level $=$ 0.95).
- `gamma1`: an optional value $\gamma_1$ satisfying $\gamma_1 + \gamma_2 = \gamma$. Default is `gamma1 = NA`.
- `gamma2`: an optional value $\gamma_2$ satisfying $\gamma_1 + \gamma_2 = \gamma$. Default is `gamma2 = NA`.
- `conf.int`: a logical variable indicating whether the CI is shown in the output. The default is `TRUE`.

Based on the sample mean of $Y_1, \ldots, Y_5$, we have $\bar{Y} = -0.502$, which is also the maximum likelihood estimator (MLE) for $\mu$. In this example, it is reasonable to set $\sigma = 0.32$ ($\sigma^2 = 0.10$) according to Table 1 of Taketomi and Emura (2023). Using the pretest estimator with a significance level of $\alpha = 0.05$ and computing a 95% CI ($1 - \gamma = 0.95$), we obtained $\delta^{\mathrm{PT}} = -0.502$ and 95% CI $= [-0.609, -0.314]$. The corresponding R code and its output are as

```
library(uni.shrinkage)
y=c(-0.17,-0.58,-0.45,-0.75,-0.56)
s=0.32
uni.pt(y,s)
>  Sample_mean  PT    Lower.pivotCI  Upper.pivotCI
1       -0.502   -0.502     -0.7824872     -0.2215128
```

We observe that the pretest estimator is equal to the MLE and therefore not is shrunk to zero. In addition, the 95% CI does not include zero. This leads us to conclude that there is a significant difference in allergic reactions between male and female students to formaldehyde.

To demonstrate the case where the pretest estimator shrinks to zero, we set $\sigma = 1$. The R code and its output are corresponding

```
y=c(-0.17,-0.58,-0.45,-0.75,-0.56)
s=1
uni.pt(y,s)
>  Sample_mean  PT   Lower.pivotCI  Upper.pivotCI
1       -0.502     0      -1.753045       1.753045
```

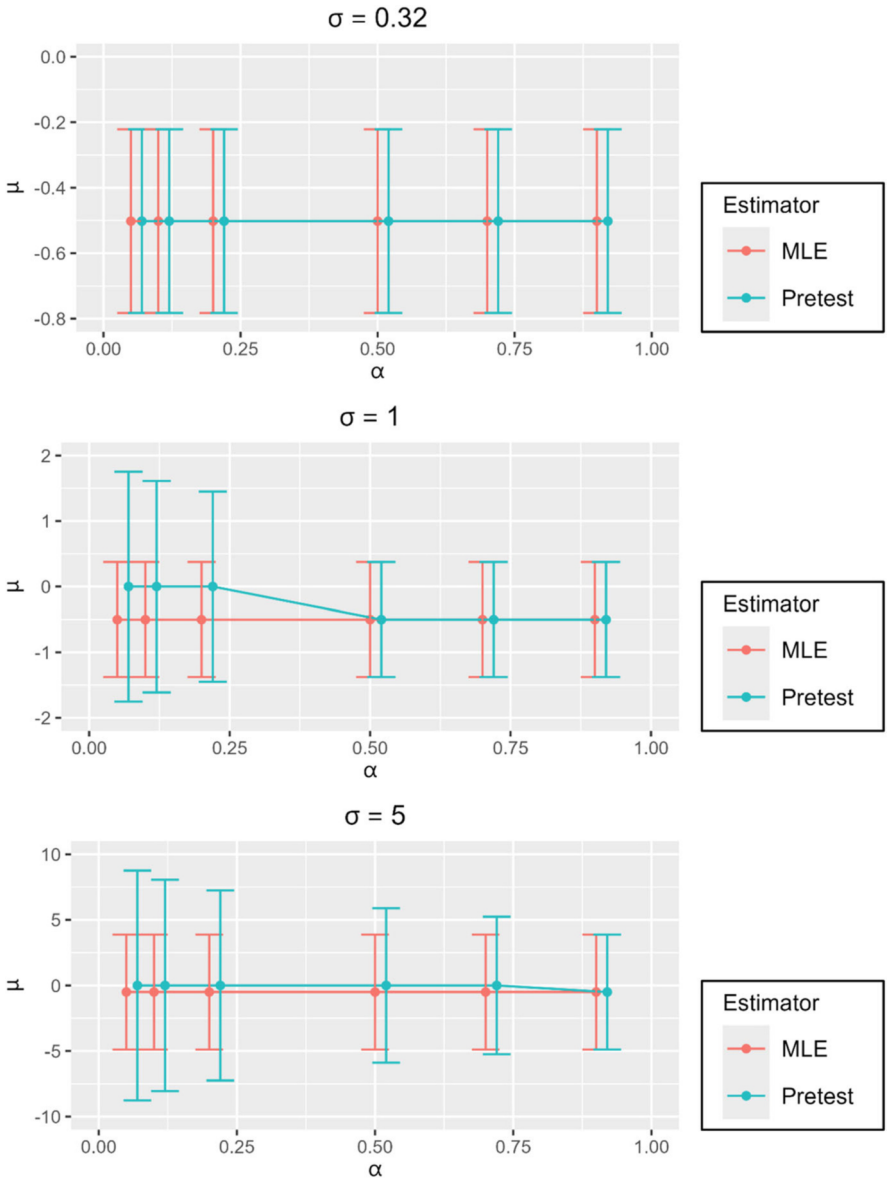



**Fig. 24.4** Estimates and 95% CIs based on the MLE $\bar{Y}$ and the pretest estimator ($\delta^{\mathrm{PT}}$) for $\alpha \in \{0.05, 0.1, 0.2, 0.5, 0.7, 0.9\}$ and $\sigma \in \{0.32, 1, 5\}$ based on the allergic reaction data

Under $\sigma = 1$, we obtain $\delta^{\mathrm{PT}} = 0$. Therefore, we conclude that male and female students do not show significant difference in their allergic reactions. However, the CI still contains a range around zero. This highlights the fact that the true MD is contained in the CI with 95% confidence, rather than concluding definitively that $\mu = 0$. Given that conclusions using the pretest estimator can depend on $\alpha$ and $\sigma$, we conducted sensitivity analyses. Figure 24.4 compares the MLE and the pretest estimator under different values of $\alpha \in \{0.05, 0.1, 0.2, 0.5, 0.7, 0.9\}$ and $\sigma \in \{0.32, 1, 5\}$.

Key observations are as follows:

- For $\sigma = 0.32$, the results of MLE and the pretest estimator are the same.
- For $\sigma = 1$, the pretest estimator shrinks $\bar{Y}$ to zero for $\alpha \in \{0.05, 0.1, 0.2\}$, but not for $\alpha \in \{0.5, 0.7, 0.9\}$.
- For $\sigma = 5$, $\bar{Y}$ is almost always shrunk to zero except for $\alpha = 0.90$.

Overall, for any $\alpha$ as $\sigma$ gets larger or $\alpha$ becomes smaller, the pretest estimator tends to be more shrunk to zero (with a wider CI). We further observe that the CI based on the pretest estimator is always wider than the one based on the MLE.

## 24.6 Conclusion and Discussion

In this chapter, we introduce a valid approach to construct a confidence interval (CI) based on the pretest estimator for a normal mean. We first derive the CDF of the pretest estimator and its asymptotic properties, which form the basis for constructing the CI. By pivoting the CDF of the pretest estimator, we obtain the explicit formula for the CI. Furthermore, we show that the coverage probability (CP) of the CI theoretically controls the nominal confidence level. Additionally, we derive the asymptotic properties of the proposed CI (Theorem 24.6) and check its validity by Monte Carlo simulations. To facilitate the implementation of the proposed CI, we have developed computational tools in the R package `uni.shrinkage`. Finally, a real dataset is analyzed to illustrate the utility of the proposed CI.

An important limitation of this study is that the standard deviation (SD) of the normal variable is identical across individuals. If the SDs of the normal variables differ across individuals, several complications arise. First, the sample mean is no longer the MLE, and the best unbiased estimator should be used instead. Some regularity conditions are required to prove the consistency, as recently shown by Taketomi and Emura (2023) in the context of meta-analysis. The setting is often referred to as the "common mean" problem (Jena & Tripathy, 2022; Pal et al., 2007; Shih et al., 2019, 2022). Similar statements were previously noticed in other contexts Emura et al. (2017). Therefore, deriving the CI under heterogeneous SDs is an important direction for future research.

Another important future work is the exploration of alternative methods to derive a CI. Beyond the proposed pivot method, there are various methods for dealing with discrete distributions (Casella & Berger, 2002; Klaschka & Reiczigel, 2021; Pradhan et al., 2016; Zhang & Nadarajah, 2018).

**Acknowledgments** We thank Mohammad Arashi and Mina Norouzirad for the invitation to contribute to a springer publication Honoring Professor A. K. Md. Ehsanes Saleh. We also thank two anonymous referees for their suggestions that greatly improved the presentation of the paper. Emura T was supported financially by JSPS KAKENHI (22K11948; 20H04147). The research of Shih JH is funded by the grant from the National Science and Technology Council of Taiwan (NSTC 112-2118-M-110-004-MY2).

# Appendix

## The Proof of Theorem 24.1: The CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$

The proof is quite similar to the one given in Taketomi et al. (2024) for the case of $n = 1$. However, it is instructive to explain the derivation of the CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ for $n \geq 1$ from scratch. One has:

$$
\begin{aligned}
F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) &= \mathrm{P}(\delta^{\mathrm{PT}}(\boldsymbol{Y}) \leq d) \\
&= \mathrm{P}\left(\bar{Y} \leq d, \left|\frac{\bar{Y}}{\sigma / \sqrt{n}}\right| > z_{\alpha/2}\right) + \mathrm{P}\left(\left|\frac{\bar{Y}}{\sigma / \sqrt{n}}\right| \leq z_{\alpha/2}\right) \cdot I(d \geq 0) \\
&= \mathrm{P}\left(\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \leq \frac{d - \mu}{\sigma / \sqrt{n}}, \left|\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} + \frac{\mu}{\sigma / \sqrt{n}}\right| > z_{\alpha/2}\right) \\
&\quad + \mathrm{P}\left(\left|\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} + \frac{\mu}{\sigma / \sqrt{n}}\right| \leq z_{\alpha/2}\right) \cdot I(d \geq 0).
\end{aligned}
$$

Putting $Z_n = (\overline{Y} - \mu) / \sigma_n$ and $\sigma_n = \sigma / \sqrt{n}$, we have:

$$
\begin{aligned}
F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) &= \mathrm{P}\left(Z_n \leq \frac{d - \mu}{\sigma_n}, \left(Z_n < -\frac{\mu}{\sigma_n} - z_{\alpha/2} \text{ or } -\frac{\mu}{\sigma_n} + z_{\alpha/2} < Z_n\right)\right) \\
&\quad + \mathrm{P}\left(-\frac{\mu}{\sigma_n} - z_{\alpha/2} \leq Z_n \leq -\frac{\mu}{\sigma_n} + z_{\alpha/2}\right) \cdot I(d \geq 0)
\end{aligned}
$$

$$
= \begin{cases}
\mathrm{P}\left(Z_n \leq \dfrac{d-\mu}{\sigma_n}\right) & \text{if} \quad d \leq -z_{\alpha/2}\sigma_n, \\
\mathrm{P}\left(Z_n < -\dfrac{\mu}{\sigma_n} - z_{\alpha/2}\right) & \text{if} \quad -z_{\alpha/2}\sigma_n < d < 0, \\
\mathrm{P}\left(Z_n < -\dfrac{\mu}{\sigma_n} - z_{\alpha/2}\right) & \text{if} \quad 0 \leq d \leq z_{\alpha/2}\sigma_n, \\
\quad +\mathrm{P}\left(-\dfrac{\mu}{\sigma_n} - z_{\alpha/2} \leq Z_n \leq -\dfrac{\mu}{\sigma_n} + z_{\alpha/2}\right) & \\
\mathrm{P}\left(Z_n < -\dfrac{\mu}{\sigma_n} - z_{\alpha/2}\right) & \text{if} \quad z_{\alpha/2}\sigma_n \leq d. \\
\quad +\mathrm{P}\left(-\dfrac{\mu}{\sigma_n} - z_{\alpha/2} < Z_n \leq \dfrac{d-\mu}{\sigma_n}\right) & \\
\quad +\mathrm{P}\left(-\dfrac{\mu}{\sigma_n} - z_{\alpha/2} \leq Z_n \leq -\dfrac{\mu}{\sigma_n} + z_{\alpha/2}\right) &
\end{cases}
$$

Since $Z_n$ is a continuous random variable, $P(Z_n < z) = P(Z_n \leq z)$ for $z \in (-\infty, \infty)$. Therefore, $F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d)$ will be equivalent to Eq. (24.1) and the proof is complete. □

## The Proof of Corollary 24.1: The Limit of the CDF of $\delta^{\mathrm{PT}}(\boldsymbol{Y})$ as $n \to \infty$

***Proof*** If $\mu = 0$, the CDF in Eq. (24.1) is

$$
F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases}
\Phi\left(\dfrac{d}{\sigma/\sqrt{n}}\right) & \text{if} \quad d \leq -z_{\alpha/2}\,\sigma/\sqrt{n}, \\
\Phi\left(-z_{\alpha/2}\right) & \text{if} \quad -z_{\alpha/2}\,\sigma/\sqrt{n} < d < 0, \\
\Phi\left(z_{\alpha/2}\right) & \text{if} \quad 0 \leq d < z_{\alpha/2}\,\sigma/\sqrt{n}, \\
\Phi\left(\dfrac{d}{\sigma/\sqrt{n}}\right) & \text{if} \quad z_{\alpha/2}\,\sigma/\sqrt{n} \leq d.
\end{cases}
$$

As $n \to \infty$, the limiting set of the four ranges of $d$ in the above equation is

$$\bigcup_{n=1}^{\infty} \left(-\infty, z_{\alpha/2}\,\sigma/\sqrt{n}, 0\right) = (-\infty, 0), \qquad \bigcap_{n=1}^{\infty} \left(-z_{\alpha/2}\,\sigma/\sqrt{n}, 0\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty} \left[0, z_{\alpha/2}\,\sigma/\sqrt{n}\right) = \{0\}, \qquad \bigcup_{n=1}^{\infty} \left[z_{\alpha/2}\,\sigma/\sqrt{n}, \infty\right) = (0, \infty).$$

Furthermore, since $d/\sigma < 0$ if $d \le -z_{\alpha/2}\,\sigma/\sqrt{n}$, and $d/\sigma > 0$ if $z_{\alpha/2}\,\sigma/\sqrt{n} \le d$, we have:

$$\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases} \lim\limits_{n\to\infty} \Phi\left(\dfrac{d}{\sigma/\sqrt{n}}\right) = 0 & \text{if} \quad d < 0, \\ \lim\limits_{n\to\infty} \Phi\left(z_{\alpha/2}\right) = 1 - \dfrac{\alpha}{2} & \text{if} \quad d = 0, \\ \lim\limits_{n\to\infty} \Phi\left(\dfrac{d}{\sigma/\sqrt{n}}\right) = 1 & \text{if} \quad 0 < d. \end{cases} \tag{24.4}$$

If $\mu < 0$, we have:

$$\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases} \lim\limits_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(-\infty) = 0 & \text{if} \quad d < \mu, \\ \lim\limits_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(0) = \frac{1}{2} & \text{if} \quad d = \mu, \\ \lim\limits_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(-\infty) = 0 & \text{if} \quad \mu < d < 0, \\ \lim\limits_{n\to\infty} \Phi\left(-\dfrac{\mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) = \Phi(\infty) = 1 & \text{if} \quad d = 0, \\ \lim\limits_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(\infty) = 1 & \text{if} \quad 0 < d. \end{cases}$$

If $\mu > 0$, we have:

$$\begin{aligned}
&\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) \\
&= \begin{cases}
\lim_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(-\infty) = 0 & \text{if} \quad d < 0, \\
\lim_{n\to\infty} \Phi\left(-\dfrac{\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) = \Phi(-\infty) = 0 & \text{if} \quad d = 0, \\
\lim_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(-\infty) = 0 & \text{if} \quad 0 < d < \mu, \\
\lim_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(0) = \frac{1}{2} & \text{if} \quad d = \mu, \\
\lim_{n\to\infty} \Phi\left(\dfrac{d-\mu}{\sigma/\sqrt{n}}\right) = \Phi(\infty) = 1 & \text{if} \quad \mu < d.
\end{cases}
\end{aligned}$$

So, for $\mu \neq 0$,

$$\lim_{n\to\infty} F_{\delta^{\mathrm{PT}}(\boldsymbol{Y})}(d) = \begin{cases} 0 & \text{if} \quad d < \mu, \\ \frac{1}{2} & \text{if} \quad d = \mu, \\ 1 & \text{if} \quad \mu < d. \end{cases}$$

The proof is complete. □

## The Proof of Theorem 24.2: The CDF of $\sqrt{n}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu\right)/\sigma$

By Theorem 24.1, one has:

$$\mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu}{\sigma/\sqrt{n}} \leq d\right) = \mathrm{P}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) \leq \mu + d\frac{\sigma}{\sqrt{n}}\right)$$

$$= \begin{cases} \Phi(d) & \text{if} \quad d \le -\sqrt{n}\mu/\sigma - z_{\alpha/2} \\ & \quad \text{or} \quad -\sqrt{n}\mu/\sigma + z_{\alpha/2} \le d, \\ \Phi\left(-\dfrac{\mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) & \text{if} \quad -\sqrt{n}\mu/\sigma - z_{\alpha/2} < d < -\sqrt{n}\mu/\sigma, \\ \Phi\left(-\dfrac{\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) & \text{if} \quad -\sqrt{n}\mu/\sigma \le d < -\sqrt{n}\mu/\sigma + z_{\alpha/2}. \end{cases}$$

If $\mu = 0$,

$$\mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y})}{\sigma/\sqrt{n}} \le d\right) = \begin{cases} \Phi(d) & \text{if} \quad d \le -z_{\alpha/2} \ \text{ or } \ z_{\alpha/2} \le d, \\ \dfrac{\alpha}{2} & \text{if} \quad -z_{\alpha/2} < d < 0, \\ 1 - \dfrac{\alpha}{2} & \text{if} \quad 0 \le d < z_{\alpha/2}. \end{cases}$$

If $\mu < 0$, the four ranges of $d$ satisfy:

$$\bigcap_{n=1}^{\infty} \left(-\infty, -\sqrt{n}\mu/\sigma - z_{\alpha/2}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty} \left(-\sqrt{n}\mu/\sigma - z_{\alpha/2}, -\sqrt{n}\mu/\sigma\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty} \left[-\sqrt{n}\mu/\sigma, -\sqrt{n}\mu/\sigma - z_{\alpha/2}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty} \left[-\sqrt{n}\mu/\sigma - s_{\alpha/2}, \infty\right) = (-\infty, \infty).$$

Therefore, the asymptotic distribution of $\sqrt{n}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu\right)/\sigma$ is

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu}{\sigma/\sqrt{n}} \le d\right) = \Phi(d), \quad -\infty < d < \infty.$$

If $\mu > 0$, the asymptotic distribution of $\sqrt{n}\left(\delta^{\mathrm{PT}}(\boldsymbol{Y}) - \mu\right)/\sigma$ is obtained by the same way as the method for $\mu < 0$. □

## The Derivation of the Pivot CI

We first review the pivoting method in a general setting. The following lemma is quoted from Theorem 9.2.14 of Casella and Berger (2002).

**Lemma 24.2 (Pivoting a CDF)** *(Casella and Berger, 2002) Let $\delta$ be a discrete statistic with CDF $F_\delta(d \mid \mu) = P(\delta \leq d \mid \mu)$. Let $\mathcal{D}$ be the sample space of $\delta$. Let $\gamma_1 + \gamma_2 = \gamma$ with $0 < \gamma < 1$ be fixed values. Suppose that for each $d \in \mathcal{D}$, $\mu_{\mathrm{L}}(d)$ and $\mu_{\mathrm{U}}(d)$ can be defined as follows:*

*(i) If $F_\delta(d \mid \mu)$ is a decreasing function of $\mu$ for each d, define $\mu_{\mathrm{L}}(d)$ and $\mu_{\mathrm{U}}(d)$ by*

$$\mathrm{P}\left(\delta \leq d \mid \mu_{\mathrm{U}}(d)\right) = \gamma_1, \quad \mathrm{P}\left(\delta \geq d \mid \mu_{\mathrm{L}}(d)\right) = \gamma_2.$$

*(ii) If $F_\delta(d \mid \mu)$ is an increasing function of $\mu$ for each d, define $\mu_{\mathrm{L}}(d)$ and $\mu_{\mathrm{U}}(d)$ by*

$$\mathrm{P}\left(\delta \geq d \mid \mu_{\mathrm{U}}(d)\right) = \gamma_1, \quad \mathrm{P}\left(\delta \leq d \mid \mu_{\mathrm{L}}(d)\right) = \gamma_2.$$

*Then the random interval $[\mu_{\mathrm{L}}(\delta), \mu_U(\delta)]$ is a $100(1-\gamma)\%$ CI for $\mu$.*

***Proof of the Pivot CI in Definition 24.1*** Let $\gamma_1 + \gamma_2 = \gamma$ with $0 < \gamma < 1$ being a fixed value. From Eq. (24.1), the sample space of $\delta^{\mathrm{PT}}$ is $\delta^{\mathrm{PT}} < -z_{\alpha/2}\sigma/\sqrt{n}$ or $\delta^{\mathrm{PT}} = 0$ or $\delta^{\mathrm{PT}} > z_{\alpha/2}\sigma/\sqrt{n}$. For each of three cases, we derive:

$$\left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right] = \left[\mu_{\mathrm{L}}\left(\delta^{\mathrm{PT}}\right), \mu_{\mathrm{U}}\left(\delta^{\mathrm{PT}}\right)\right]$$

For each $d_0 < -z_{\alpha/2}\sigma/\sqrt{n}$, $F_{\delta^{\mathrm{PT}}}(d_0 \mid \mu) = \Phi\left(\sqrt{n}\left(d_0 - \mu_i\right)/\sigma\right)$ is the decreasing function of $\mu$. We define $\mu_{\mathrm{L}}(d_0)$ and $\mu_{\mathrm{U}}(d_0)$ by

$$\mathrm{P}\left(\delta^{\mathrm{PT}} \leq d_0 \mid \mu_{\mathrm{U}}(d_0)\right) = \Phi\left(\frac{d_0 - \mu_{\mathrm{U}}(d_0)}{\sigma/\sqrt{n}}\right) = \gamma_1,$$

$$\begin{aligned}\mathrm{P}\left(\delta^{\mathrm{PT}} \geq d_0 \mid \mu_{\mathrm{L}}(d_0)\right) &= 1 - \mathrm{P}\left(\delta^{\mathrm{PT}} < d_0 \mid \mu_{\mathrm{L}}(d_0)\right) \\ &= 1 - \Phi\left(\frac{d_0 - \mu_{\mathrm{L}}(d_0)}{\sigma/\sqrt{n}}\right) = \gamma_2.\end{aligned}$$

These are equivalent to

$$\Phi\left(\frac{d_0 - \mu_{\mathrm{U}}(d_0)}{\sigma/\sqrt{n}}\right) = \gamma_1, \qquad 1 - \Phi\left(\frac{d_0 - \mu_{\mathrm{L}}(d_0)}{\sigma/\sqrt{n}}\right) = \gamma_2,$$

or equivalently

$$\mu_{\mathrm{U}}(d_0) = d_0 + z_{\gamma_1}\sigma/\sqrt{n}, \qquad \mu_{\mathrm{L}}(d_0) = d_0 - z_{\gamma_2}\sigma/\sqrt{n}.$$

Therefore, we have $[\mu_{\mathrm{L}}(d_0), \mu_{\mathrm{U}}(d_0)] = \left[d_0 - z_{\gamma_2}\sigma/\sqrt{n}, d_0 + z_{\gamma_1}\sigma/\sqrt{n}\right]$.

For $d_0 = 0$, we know that $\delta^{\mathrm{PT}}(d_0 \mid \mu) = \Phi\left(-\sqrt{n}\mu/\sigma + z_{\alpha/2}\right)$ is the decreasing function of $\mu$. We define $\mu_{\mathrm{L}}(d_0)$ and $\mu_{\mathrm{U}}(d_0)$ by

$$\mathrm{P}\left(\delta^{\mathrm{PT}} \le 0 \mid \mu_{\mathrm{U}}(0)\right) = \Phi\left(-\frac{\mu_{\mathrm{U}}(0)}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) = \gamma_1,$$

$$\mathrm{P}\left(\delta^{\mathrm{PT}} \ge 0 \mid \mu_{\mathrm{L}}(0)\right) = 1 - \mathrm{P}\left(\delta^{\mathrm{PT}} < 0 \mid \mu_{\mathrm{L}}(0)\right)$$

$$= 1 - \Phi\left(-\frac{\mu_{\mathrm{L}}(0)}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) = \gamma_2.$$

Expanding these equations, we have:

$$\Phi\left(-\frac{\mu_{\mathrm{U}}(0)}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) = \gamma_1, \quad \Phi\left(-\frac{\mu_{\mathrm{L}}(0)}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) = \gamma_2.$$

They are equivalent to

$$\mu_{\mathrm{U}}(0) = \left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma/\sqrt{n}, \quad \mu_{\mathrm{L}}(0) = -\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma/\sqrt{n}.$$

Therefore, for $\delta^{\mathrm{PT}} = 0$,

$$[\mu_{\mathrm{L}}(d), \mu_{\mathrm{U}}(d)] = \left[-\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma/\sqrt{n}, \left(z_{\alpha/2} + z_{\gamma_1}\right)\sigma/\sqrt{n}\right].$$

For $d_0 > z_{\alpha_\alpha/2}\sigma/\sqrt{n}$, one can derive:

$$[\mu_{\mathrm{L}}(d_0), \mu_{\mathrm{U}}(d_0)] = \left[d_0 - z_{\gamma_2}\sigma/\sqrt{n}, d_0 + z_{\gamma_1}\sigma/\sqrt{n}\right].$$

The case of $d_0 < -z_{\alpha/2}\sigma/\sqrt{n}$ is the same. In conclusion, we have:

$$\left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]$$
$$= \left[\mu_{\mathrm{L}}(\delta^{\mathrm{PT}}), \mu_{\mathrm{U}}(\delta^{\mathrm{PT}})\right]$$

$$= \begin{cases} \left[\delta^{\mathrm{PT}} - z_{\gamma_2}\sigma/\sqrt{n}, \delta^{\mathrm{PT}} + z_{\gamma_1}\sigma/\sqrt{n}\right] & \text{if } \delta^{\mathrm{PT}} < -z_{\alpha/2}\sigma/\sqrt{n}, \\ \left[-\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma/\sqrt{n}, \left(z_{\alpha/2} + z_{\gamma_1}\right)\sigma/\sqrt{n}\right] & \text{if } \delta^{\mathrm{PT}} = 0, \\ \left[\delta^{\mathrm{PT}} - z_{\gamma_2}\sigma/\sqrt{n}, \delta^{\mathrm{PT}} + z_{\gamma_1}\sigma/\sqrt{n}\right] & \text{if } \delta^{\mathrm{PT}} > z_{\alpha/2}\sigma/\sqrt{n}. \end{cases}$$

Thus, the derivation is complete. □

## The Proof of Theorem 24.4

We have:

$$\begin{aligned}
&P\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) \\
&= \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right], \delta^{\mathrm{PT}} < -z_{\alpha/2}\sigma/\sqrt{n}\right) + \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right], \delta^{\mathrm{PT}} = 0\right) \\
&\quad + \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right], \delta^{\mathrm{PT}} > z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&= \mathrm{P}\left(\bar{Y} - z_{\gamma_2}\sigma/\sqrt{n} \le \mu \le \bar{Y} + z_{\gamma_1}\sigma/\sqrt{n}, \bar{Y} < -z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&\quad + \mathrm{P}\left(-\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma/\sqrt{n} \le \mu \le \left(z_{\alpha/2} + z_{\gamma_1}\right)\sigma/\sqrt{n}, \left|\frac{\bar{Y}}{\sigma/\sqrt{n}}\right| \le z_{\alpha/2}\right) \\
&\quad + \mathrm{P}\left(\bar{Y} - z_{\gamma_2}\sigma/\sqrt{n} \le \mu \le \bar{Y} + z_{\gamma_1}\sigma/\sqrt{n}, \bar{Y} > z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&= \mathrm{P}\left(\mu - z_{\gamma_1}\sigma/\sqrt{n} \le \bar{Y} \le \mu + z_{\gamma_2}\sigma/\sqrt{n}, \bar{Y} < -z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&\quad + \mathrm{P}\Big(-\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma/\sqrt{n} \le \mu \le \left(z_{\alpha/2} + z_{\gamma_1}\right)\sigma/\sqrt{n}, \\
&\qquad\qquad -z_{\alpha/2}\sigma/\sqrt{n} \le \bar{Y} \le z_{\alpha/2}\sigma/\sqrt{n}\Big) \\
&\quad + \mathrm{P}\left(\mu - z_{\gamma_1}\sigma/\sqrt{n} \le \bar{Y} \le \mu + z_{\gamma_2}\sigma/\sqrt{n}, \bar{Y} > z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&= \mathrm{P}\left(-z_{\gamma_1} \le \frac{\bar{\gamma} - \mu}{\sigma/\sqrt{n}} \le z_{\gamma_2}, \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < -\frac{\mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right) \\
&\quad + \mathrm{P}\Big(-\left(z_{\alpha/2} + z_{\gamma_2}\right) \le \frac{\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2} + z_{\gamma_2}, \\
&\qquad\qquad -\frac{\mu}{\sigma/\sqrt{n}} - z_{\alpha/2} \le \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \le -\frac{\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\Big)
\end{aligned}$$

$$
\begin{aligned}
&+\mathrm{P}\left(-z_{\gamma_1} \le \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \le z_{\gamma_2},\ \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} > -\frac{\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) \\
&= \mathrm{P}\left(-z_{\gamma_1} \le Z \le z_{\gamma_2}, Z < -\lambda - z_{\alpha/2}\right) \\
&\quad +\mathrm{P}\left(-\left(z_{\alpha/2}+z_{\gamma_2}\right) \le \lambda \le z_{\alpha/2}+z_{\gamma_2}, -\lambda - z_{\alpha/2} \le Z \le -\lambda + z_{\alpha/2}\right) \\
&\quad +\mathrm{P}\left(-z_{\gamma_1} \le Z \le z_{\gamma_2},\ Z > -\lambda + z_{\alpha/2}\right).
\end{aligned}
$$

Each term in the above equation can be expressed as follows. For the first term

$$
\mathrm{P}\left(-z_{\gamma_1} \le Z \le z_{\gamma_2},\ Z < -\lambda - z_{\alpha/2}\right)
= \begin{cases}
1-\gamma & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma_2} \\
\Phi\left(-\lambda - z_{\alpha/2}\right) - \gamma_1 & \text{if} \quad -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le -z_{\alpha/2} + z_{\gamma_1} \\
0 & \text{if} \quad -z_{\alpha/2} + z_{\gamma_1} \le \lambda \le z_{\alpha/2} + z_{\gamma_2} \\
0 & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma_2}
\end{cases}
$$

and for the second term

$$
\mathrm{P}\left(-\left(z_{\alpha/2}+z_{\gamma_2}\right) \le \lambda \le z_{\alpha/2}+z_{\gamma_1},\ -\lambda - z_{\alpha/2} \le Z \le -\lambda + z_{\alpha/2}\right)
$$

$$
= \begin{cases}
0 & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma_2} \\
\Phi\left(-\lambda + z_{\alpha/2}\right) - \Phi\left(-\lambda - z_{\alpha/2}\right) & \text{if} \quad -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} + z_{\gamma_1} \\
0 & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma_1}
\end{cases}
$$

and for the third term

$$
\mathrm{P}\left(-z_{\gamma_1} \le Z \le z_{\gamma_2},\ Z > -\lambda + z_{\alpha/2}\right)
= \begin{cases}
0 & \text{if} \quad \lambda < -z_{\alpha/2} - z_{\gamma_2}, \\
0 & \text{if} \quad -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} - z_{\gamma_2}, \\
1-\gamma_2-\Phi\left(-\lambda + z_{\alpha/2}\right) & \text{if} \quad z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} + z_{\gamma_1}, \\
1-\gamma & \text{if} \quad \lambda > z_{\alpha/2} + z_{\gamma_1}.
\end{cases}
$$

Summing up the three terms, we obtain the following CP.

For $z_{\alpha/2} - z_{\gamma_2} < -z_{\alpha/2} + z_{\gamma_1}$,

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if } \lambda < -z_{\alpha/2} - z_{\gamma_2} \\ \Phi\left(-\lambda + z_{\alpha/2}\right) - \gamma_1 & \text{if } -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} - z_{\gamma_2} \\ 1-\gamma & \text{if } z_{\alpha/2} - z_{\gamma_2} < \lambda < -z_{\alpha/2} - z_{\gamma_1} \\ 1-\gamma_2 - \Phi\left(-\lambda - z_{\alpha/2}\right) & \text{if } -z_{\alpha/2} + z_{\gamma_1} \le \lambda \le z_{\alpha/2} + z_{\gamma_1} \\ 1-\gamma & \text{if } \lambda > z_{\alpha/2} + z_{\gamma_1} \end{cases}$$

for $z_{\alpha/2} - z_{\gamma_2} = -z_{\alpha/2} + z_{\gamma_1}$,

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if } \lambda < -z_{\alpha/2} - z_{\gamma_2} \\ \Phi\left(-\lambda + z_{\alpha/2}\right) - \gamma_1 & \text{if } -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} - z_{\gamma_2} \\ 1-\gamma_2 - \Phi\left(-\lambda - z_{\alpha/2}\right) & \text{if } z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} + z_{\gamma_1} \\ 1-\gamma & \text{if } \lambda > z_{\alpha/2} + z_{\gamma_1} \end{cases}$$

for $z_{\alpha/2} - z_{r_2} > -z_{\alpha/2} + z_{\gamma_1}$,

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} 1-\gamma & \text{if } \lambda < -z_{\alpha/2} - z_{\gamma_2}, \\ \Phi\left(-\lambda + z_{\alpha/2}\right) - \gamma_1 & \text{if } -z_{\alpha/2} - z_{\gamma_2} \le \lambda \le -z_{\alpha/2} + z_{\gamma_1}, \\ \Phi\left(-\lambda + z_{\alpha/2}\right) - \Phi\left(-\lambda - z_{\alpha/2}\right) & \text{if } -z_{\alpha/2} + z_{\gamma_1} < \lambda < z_{\alpha/2} - z_{\gamma_2}, \\ 1-\gamma_2 - \Phi\left(-\lambda - z_{\alpha/2}\right) & \text{if } z_{\alpha/2} - z_{\gamma_2} \le \lambda \le z_{\alpha/2} + z_{\gamma_1}, \\ 1-\gamma & \text{if } \lambda > z_{\alpha/2} + z_{\gamma_1}. \end{cases}$$

□

## The Proof of Theorem 24.5

We recall that the CP takes three cases (Cases 1–3 in Theorem 24.4). The first and second cases are expressed as $z_{\alpha/2} - z_{\gamma_2} < -z_{\alpha/2} + z_{\gamma_1}$ and $z_{\alpha/2} - z_{\gamma_2} = -z_{\alpha/2} + z_{\gamma_1}$, respectively. In these cases, the CP of Theorem 24.4 immediately yields Eq. (24.3).

Now, we consider the last case expressed as $-z_{\alpha/2}+z_{\gamma_1} < z_{\alpha/2}-z_{\gamma_2}$. Note that $-\lambda+z_{\alpha/2} > z_{\gamma_2}$ is equivalent to $\Phi\left(-\lambda+z_{\alpha/2}\right) > 1-\gamma_2$ and that $-\lambda-z_{\alpha/2} < -z_{\gamma_1}$ is equivalent to $-\Phi(-\lambda-z_{\alpha/2}) > -\gamma_1$. From these inequalities, for $-z_{\alpha/2}+z_{\gamma_1} < \lambda < z_{\alpha/2}-z_{\gamma_2}$, we have:

$$\Phi\left(-\lambda+z_{\alpha/2}\right)-\Phi\left(-\lambda-z_{\alpha/2}\right) > 1-\gamma_2-\gamma_1 = 1-\gamma.$$

For $\lambda \le -z_{\alpha/2}+z_{\gamma_1}$ and $\lambda \ge z_{\alpha/2}-z_{\gamma_2}$,

$$\inf_{\lambda \le -z_{\alpha/2}+z_{\gamma_1},\, \lambda \ge z_{\alpha/2}-z_{\gamma_2}} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = 1-\gamma.$$

Therefore, Eq. (24.3) is satisfied. The proof is complete. □

## The Proof of Theorem 24.6

We consider the limit of the CP in each of three cases of Theorem 24.3:

(I) For $2z_{\alpha/2} < z_{\gamma_1}+z_{\gamma_2}$, that is, for $z_{\alpha/2}-z_{\gamma_2} < -z_{\alpha/2}+z_{\gamma_1}$, there are three cases:

(i) $0 \le z_{\alpha/2}-z_{\gamma_2} < -z_{\alpha/2}+z_{\gamma_1}$;
(ii) $z_{\alpha/2}-z_{\gamma_2} \le 0 < -z_{\alpha/2}+z_{\gamma_2}$;
(iii) $z_{\alpha/2}-z_{\gamma_2} < -z_{\alpha/2}+z_{\gamma_2} \le 0$.

(i) For $0 \le z_{\alpha/2}-z_{\gamma_2} < -z_{\alpha/2}+z_{\gamma_1}$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \{0\},$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left[\frac{\left(-z_{\alpha/2}+z_{r_1}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{r_2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore, we have:

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu < 0 \\ \lim\limits_{n\to\infty}\left\{\Phi\left(z_{\alpha/2}\right) - \gamma_1\right\} & \text{if} \quad \mu = 0 \\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2} - \gamma_1 & \text{if} \quad \mu = 0. \end{cases}$$

(ii) For $z_{\alpha/2} - z_{1/2} \leq 0 < -z_{\alpha/2} + z_{\gamma_1}$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2} - z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2} - z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2} + z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right) = \{0\},$$

$$\bigcap_{n=1}^{\infty}\left[\frac{\left(-z_{\alpha/2} + z_{\gamma_1}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore, for $\mu \in (-\infty, \infty)$

$$\mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu < 0 \\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu = 0 \\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu > 0 \end{cases} \tag{24.5}$$

(iii) For $z_{\alpha/2} - z_{\gamma_2} < -z_{\alpha/2} + z_{\gamma_1} \leq 0$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2} - z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2} - z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2} + z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left[\frac{\left(-z_{\alpha/2} + z_{\gamma_1}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \{0\},$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu < 0 \\ \lim\limits_{n\to\infty}\left\{1 - \gamma_2 - \Phi\left(-z_{\alpha/2}\right)\right\} & \text{if} \quad \mu = 0 \\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if} \quad \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1 - \dfrac{\alpha}{2} - \gamma_2 & \text{if} \quad \mu = 0. \end{cases}$$

(II) For $2z_{\alpha/2} = z_{\gamma_1} + z_{\gamma_2}$, that is, for $z_{\alpha/2} - z_{\gamma_2} = -z_{\alpha/2} + z_{\gamma_1}$, there are two cases:

(i) $0 \leq z_{\alpha/2} - z_{\gamma_2} = -z_{\alpha/2} + z_{\gamma_1}$;
(ii) $z_{\alpha/2} - z_{\gamma_2} = -z_{\alpha/2} + z_{\gamma_1} \leq 0$.

(i) For $0 \leq z_{\alpha/2} - z_{\gamma_2} = -z_{\alpha/2} + z_{\gamma_1}$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}},\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right]=\{0\}$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}},\frac{\left(z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right)=\varnothing,$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}},\infty\right)=(0,\infty).$$

Therefore, we have:

$$\lim_{n\to\infty}\mathrm{P}\left(\mu\in\left[L^{\mathrm{PT}},U^{\mathrm{PT}}\right]\right)=\begin{cases}\lim\limits_{n\to\infty}(1-\gamma) & \text{if}\quad \mu<0\\ \lim\limits_{n\to\infty}\left\{\left(z_{\alpha/2}\right)-\gamma_1\right\} & \text{if}\quad \mu=0\\ \lim_{n\to\infty}(1-\gamma) & \text{if}\quad \mu>0\end{cases}$$

$$=\begin{cases}1-\gamma & \text{if}\quad \mu\neq 0,\\ 1-\dfrac{\alpha}{2}-\gamma_1 & \text{if}\quad \mu=0.\end{cases}$$

(ii) For $z_{\alpha/2}-z_{\gamma_2}=-z_{\alpha/2}+z_{\gamma_1}\leq 0$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty,-\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right)=(-\infty,0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}},\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right]=\varnothing,$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}},\frac{\left(z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right)=\{0\},$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{r_2}\right)\sigma}{\sqrt{n}},\infty\right)=(0,\infty).$$

Therefore,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim_{n\to\infty} (1-\gamma) & \text{if} \quad \mu < 0 \\ \lim_{n\to\infty} \left\{1-\gamma_2-\Phi\left(-z_{\alpha/2}\right)\right\} & \text{if} \quad \mu = 0 \\ \lim_{n\to\infty} (1-\gamma) & \text{if} \quad \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\gamma_2 & \text{if} \quad \mu = 0. \end{cases}$$

(III) For $2z_{\alpha/2} > z_{\gamma_1} + z_{\gamma_2}$, that is, for $-z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2}$, there are three cases of the signs for the both sides:

(i) $0 \le -z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2}$;
(ii) $-z_{\alpha/2} + z_{\gamma_2} \le 0 < z_{\alpha/2} - z_{\gamma_2}$;
(iii) $-z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2} < 0$.

(i) For $0 \le -z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2}$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty} \left(-\infty, -\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty} \left[-\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right] = \{0\},$$

$$\bigcap_{n=1}^{\infty} \left(\frac{\left(-z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty} \left[\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcup_{n=1}^{\infty} \left(\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim_{n\to\infty} (1-\gamma) & \text{if} \quad \mu < 0 \\ \lim_{n\to\infty} \left\{\Phi\left(z_{\alpha/2}\right) - \gamma_1\right\} & \text{if} \quad \mu = 0 \\ \lim_{n\to\infty} (1-\gamma) & \text{if} \quad \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if} \quad \mu \neq 0, \\ 1-\dfrac{\alpha}{2} - \gamma_1 & \text{if} \quad \mu = 0. \end{cases}$$

(ii) For $-z_{\alpha/2} + z_{\gamma_1} \leq 0 < z_{\alpha/2} - z_{\gamma_2}$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty} \left(-\infty, -\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty} \left[-\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2} + z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcap_{n=1}^{\infty} \left(\frac{\left(-z_{\alpha/2} + z_{\gamma_1}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2} - s_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = \{0\},$$

$$\bigcap_{n=1}^{\infty} \left[\frac{\left(z_{\alpha/2} - z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2} + s_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcup_{n=1}^{\infty} \left(\frac{\left(z_{\alpha/2} + z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore,

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim_{n\to\infty} (1-\gamma) & \text{if } \mu < 0 \\ \lim_{n\to\infty} \left\{\Phi\left(z_{\alpha/2}\right) - \Phi\left(-z_{\alpha/2}\right)\right\} & \text{if } \mu = 0 \\ \lim_{n\to\infty} (1-\gamma) & \text{if } \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if } \mu \neq 0, \\ 1-\alpha & \text{if } \mu = 0. \end{cases}$$

(iii) For $-z_{\alpha/2}+z_{\gamma_1} < z_{\alpha/2}-z_{\gamma_2} \le 0$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(-z_{\alpha/2}+z_{\gamma_1}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left[\frac{\left(z_{\alpha/2}-z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}\right] = \{0\},$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{\gamma_2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore, we have:

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim\limits_{n\to\infty}(1-\gamma) & \text{if } \mu < 0 \\ \lim\limits_{n\to\infty}\left\{1-\gamma_2-\Phi\left(-z_{\alpha/2}\right)\right\} & \text{if } \mu = 0 \\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if } \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if } \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\gamma_2 & \text{if } \mu = 0. \end{cases}$$

Since the two pairs of cases, (I)(i) and (II)(ii) and (I)(iii) and (II)(ii) can be considered simultaneously, the proof is complete. □

## The Proof of Corollary 24.3

We compute the infimums for all the six cases.

(Case 1) For $0 \leq z_{\alpha/2} - z_{\gamma_2} \leq -z_{\alpha/2} + z_{\gamma_1}$,

$$\begin{aligned} z_{\gamma_2} \leq z_{\alpha/2} &\Leftrightarrow \alpha/2 \leq \gamma_2 \Leftrightarrow \alpha/2 + \gamma_1 \leq \gamma_1 + \gamma_2 = \gamma \\ &\Leftrightarrow 1 - \alpha/2 - \gamma_1 \geq 1 - \gamma. \end{aligned}$$

Therefore,

$$\begin{aligned} &\inf_{-\infty<\mu<\infty} \left\{ \lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right)\right\} \\ &= \min\left( \inf_{\mu\neq 0}(1-\gamma), \inf_{\mu=0}\left(1 - \frac{\alpha}{2} - \gamma_1\right)\right) \\ &= 1 - \gamma. \end{aligned}$$

(Case 2) For $z_{\alpha/2} - z_{\gamma_2} \leq 0 < -z_{\alpha/2} + z_{\gamma_1}$, we have

$$\inf_{-\infty<\mu<\infty} \left\{ \lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right)\right\} = 1 - \gamma.$$

(Case 3) For $z_{\alpha/2} - z_{\gamma_2} \leq -z_{\alpha/2} + z_{\gamma_1} < 0$,

$$\begin{aligned} z_{\gamma_1} \leq z_{\alpha/2} &\Leftrightarrow \alpha/2 \leq \gamma_1 \Leftrightarrow \alpha/2 + \gamma_2 \leq \gamma_1 + \gamma_2 = \gamma \\ &\Leftrightarrow 1 - \alpha/2 - \gamma_2 \geq 1 - \gamma. \end{aligned}$$

Therefore,

$$\begin{aligned} &\inf_{-\infty<\mu<\infty} \left\{ \lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right)\right\} \\ &= \min\left( \inf_{\mu\neq 0}(1-\gamma), \inf_{\mu=0}\left(1 - \frac{\alpha}{2} - \gamma_2\right)\right) \\ &= 1 - \gamma. \end{aligned}$$

(Case 4) For $0 \leq -z_{\alpha/2} + z_{\gamma_1} < z_{\alpha/2} - z_{\gamma_2}$,

$$\begin{aligned} z_{\gamma_2} < z_{\alpha/2} &\Leftrightarrow \alpha/2 < \gamma_2 \Leftrightarrow \alpha/2 + \gamma_1 < \gamma_1 + \gamma_2 = \gamma \\ &\Leftrightarrow 1 - \alpha/2 - \gamma_1 > 1 - \gamma. \end{aligned}$$

Therefore,

$$\begin{aligned}
&\inf_{-\infty<\mu<\infty}\left\{\lim_{n\to\infty}\mathrm{P}\left(\mu\in\left[L^{\mathrm{PT}},U^{\mathrm{PT}}\right]\right)\right\}\\
&\quad=\min\left(\inf_{\mu\neq 0}(1-\gamma),\ \inf_{\mu=0}\left(1-\frac{\alpha}{2}-\gamma_1\right)\right)\\
&\quad=1-\gamma.
\end{aligned}$$

(Case 5) For $-z_{\alpha/2}+z_{\gamma_1}\le 0<z_{\alpha/2}-z_{\gamma_2}$,

$$\begin{aligned}
z_{\gamma_1}\le z_{\alpha/2}\ \text{ and }\ z_{\gamma_2}<z_{\alpha/2}&\Leftrightarrow \alpha/2\le\gamma_1\ \text{ and }\ \alpha/2<\gamma_2\\
&\Leftrightarrow \alpha\le\gamma_1+\gamma_2=\gamma\Leftrightarrow 1-\alpha\ge 1-\gamma.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\inf_{-\infty<\mu<\infty}\left\{\lim_{n\to\infty}\mathrm{P}\left(\mu\in\left[L^{\mathrm{PT}},U^{\mathrm{PT}}\right]\right)\right\}\\
&\quad=\min\left(\inf_{\mu\neq 0}(1-\gamma),\ \inf_{\mu=0}(1-\alpha)\right)\\
&\quad=1-\gamma
\end{aligned}$$

(Case 6) For $-z_{\alpha/2}+z_{\gamma_1}<z_{\alpha/2}-z_{\gamma_2}<0$,

$$\begin{aligned}
z_{\gamma_1}<z_{\alpha/2}&\Leftrightarrow \alpha/2<\gamma_1\Leftrightarrow \alpha/2+\gamma_2<\gamma_1+\gamma_2\\
&=\gamma\Leftrightarrow 1-\alpha/2-\gamma_2<1-\gamma.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\inf_{-\infty<\mu<\infty}\left\{\lim_{n\to\infty}\mathrm{P}\left(\mu\in\left[L^{\mathrm{PT}},U^{\mathrm{PT}}\right]\right)\right\}\\
&\quad=\min\left(\inf_{\mu\neq 0}(1-\gamma),\ \inf_{\mu=0}\left(1-\frac{\alpha}{2}-\gamma_2\right)\right)\\
&\quad=1-\gamma.
\end{aligned}$$

Combining the above results, we conclude

$$\inf_{-\infty<\mu<\infty}\left\{\lim_{n\to\infty}\mathrm{P}\left(\mu\in\left[L^{\mathrm{PT}},U^{\mathrm{PT}}\right]\right)\right\}=1-\gamma.$$

The proof is complete. □

## The Proof of Corollary 24.4

We consider the limit of the CP in each of three cases.

(I) For $\alpha < \gamma$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{y/2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}-z_{z/2}\right)\sigma}{\sqrt{n}}\right] = \{0\},$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}-z_{\gamma/2}\right)\sigma}{\sqrt{n}}, \frac{\left(-z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right) = \varnothing,$$

$$\bigcap_{n=1}^{\infty}\left[\frac{\left(-z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}, \frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right] = \varnothing,$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}, \infty\right) = (0, \infty).$$

Therefore, we have:

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim\limits_{n\to\infty}(1-\gamma) & \text{if } \mu < 0 \\ \lim\limits_{n\to\infty}\left\{\Phi\left(z_{\alpha/2}\right) - \frac{\gamma}{2}\right\} & \text{if } \mu = 0 \\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if } \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if } \mu \neq 0, \\ 1-\dfrac{\alpha}{2}-\frac{\gamma}{2} & \text{if } \mu = 0. \end{cases}$$

(II) For $\alpha = \gamma$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty, -\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right) = (-\infty, 0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}},0\right]=\{0\},$$

$$\bigcap_{n=1}^{\infty}\left(0,\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right)=\varnothing,$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}},\infty\right)=(0,\infty).$$

Therefore,

$$\lim_{n\to\infty}\mathrm{P}\left(\mu\in\left[L^{\mathrm{PT}},U^{\mathrm{PT}}\right]\right)=\begin{cases}\lim\limits_{n\to\infty}(1-\gamma) & \text{if } \mu<0\\ \lim\limits_{n\to\infty}\left\{\Phi\left(z_{\alpha/2}\right)-\frac{\gamma}{2}\right\} & \text{if } \mu=0\\ \lim\limits_{n\to\infty}(1-\gamma) & \text{if } \mu>0\end{cases}$$

$$=\begin{cases}1-\gamma & \text{if } \mu\neq 0,\\ 1-\dfrac{\alpha}{2}-\dfrac{\gamma}{2} & \text{if } \mu=0.\end{cases}$$

(III) For $\alpha<\gamma$, the ranges of $\mu$ satisfy that

$$\bigcup_{n=1}^{\infty}\left(-\infty,-\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right)=(-\infty,0),$$

$$\bigcap_{n=1}^{\infty}\left[-\frac{\left(z_{\alpha/2}+z_{y/2}\right)\sigma}{\sqrt{n}},\frac{\left(-z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right]=\{0\},$$

$$\bigcap_{n=1}^{\infty}\left(\frac{\left(-z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}},\frac{\left(z_{\alpha/2}-z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right)=\varnothing,$$

$$\bigcap_{n=1}^{\infty}\left[\frac{\left(z_{\alpha/2}-z_{\gamma/2}\right)\sigma}{\sqrt{n}},\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\right]=\varnothing,$$

$$\bigcup_{n=1}^{\infty}\left(\frac{\left(z_{\alpha/2}+z_{\gamma/2}\right)\sigma}{\sqrt{n}}\infty\right)=(0,\infty).$$

Therefore, we have:

$$\lim_{n\to\infty} \mathrm{P}\left(\mu \in \left[L^{\mathrm{PT}}, U^{\mathrm{PT}}\right]\right) = \begin{cases} \lim\limits_{n\to\infty} (1-\gamma) & \text{if } \mu < 0 \\ \lim\limits_{n\to\infty} \left\{\Phi\left(z_{\alpha/2}\right) - \frac{\gamma}{2}\right\} & \text{if } \mu = 0 \\ \lim\limits_{n\to\infty} (1-\gamma) & \text{if } \mu > 0 \end{cases}$$

$$= \begin{cases} 1-\gamma & \text{if } \mu \neq 0, \\ 1 - \dfrac{\alpha}{2} - \dfrac{\gamma}{2} & \text{if } \mu = 0. \end{cases}$$

The proof is complete. □

## References


Aldeni, M., Wagaman, J., Amezziane, M., & Ahmed, S. E. (2023). Pretest and shrinkage estimators for log-normal means. *Computational Statistics, 38*, 1555–1578.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Annals of Mathematical Statistics, 15*(2), 190–204.

Ben-Shachar, M. S., Patil, I., Thériault, R., Wiernik, B. M., & Lüdecke, D. (2023). Phi, Fei, Fo, Fum: Effect sizes for categorical data that use the chi-squared statistic. *Mathematics, 11*(9), 1982.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury.

Emura, T., Hu, Y. H., & Konno, Y. (2017). Asymptotic inference for maximum likelihood estimators under the special exponential family with double-truncation. *Statistical Papers, 58*, 877–909.

Emura, T., Matsumoto, K., Uozumi, R., & Michimae, H. (2024). g.ridge: An R package for generalized ridge regression for sparse and high-dimensional linear models. *Symmetry, 16*(2), 223.

Fagerland, M. W., Lydersen, S., & Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research, 24*(2), 224–254.

Fleiss, J. L. (1993). Review papers: The statistical basis of meta-analysis. *Statistical Methods in Medical Research, 2*(2), 121–145.

Hoque, Z., Khan, S., & Wesolowski, J. (2009). Performance of preliminary test estimator under LINEX loss function. *Communications in Statistics - Theory and Methods, 38*(2), 252–261.

Jena, P., & Ranjan Tripathy, M. (2022). Estimating positive powers of the scale parameters under order restriction for two normal populations with a common mean. *Statistics, 56*(3), 598–630.

Judge, G. G., & Bock, M. E. (1978). *The statistical implications of pre-test and Stein-rule estimators in econometrics*. Elsevier.

Khan, S., & Saleh, A. K. M. E. (2001). On the comparison of the pre-test and shrinkage estimators for the univariate normal mean. *Statistical Papers, 42*(4), 451–473.

Kibria, B. G., & Saleh, A. K. M. E. (2006). Optimum critical value for pre-test estimator. *Communications in Statistics - Simulation and Computation, 35*(2), 309–319.

Klaschka, J., & Reiczigel, J. (2021). On matching confidence intervals and tests for some discrete distributions: Methodological and computational aspects. *Computational Statistics, 36*, 1775–1790.

Magnus, J. R. (2000). The traditional pretest estimator. *Theory of Probability and Its Applications, 44*(2), 293–308.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal, 5*, 225–236.

Magnus, J. R., Wan, A. T., & Zhang, X. (2011). Weighted average least squares estimation with nonspherical disturbances and an application to the Hong Kong housing market. *Computational Statistics and Data Analysis, 55*(3), 1331–1341.

Mori, M., & Kakuma, T. (2022). Development of a prediction model for subjective physical symptoms induced during systematic anatomy practice among medical students. *The Kurume Medical Journal, 69*(3,4), 209–216.

Norouzirad, M., & Arashi, M. (2019). Preliminary test and Stein-type shrinkage ridge estimators in robust regression. *Statistical Papers, 60*, 1849–1882.

Pal, N., Lin, J. J., Chang, C. H., & Kumar, S. (2007). A revisit to the common mean problem: Comparing the maximum likelihood estimator with the Graybill–Deal estimator. *Computational Statistics & Data Analysis, 51*(12), 5673–5681.

Pradhan, V., Evans, J. C., & Banerjee, T. (2016). Binomial confidence intervals for testing non-inferiority or superiority: A practitioner's dilemma. *Statistical Methods in Medical Research, 25*(4), 1707–1717.

Saleh, A. K. M. E. (2006). *Theory of preliminary test and Stein-type estimation with applications*. John Wiley & Sons.

Shih, J. H., Konno, Y., Chang, Y. T., & Emura, T. (2019). Estimation of a common mean vector in bivariate meta-analysis under the FGM copula. *Statistics, 53*(3), 673–695.

Shih, J. H., Konno, Y., Chang, Y. T., & Emura, T. (2022). Copula-based estimation methods for a common mean vector for bivariate meta-analyses.*Symmetry, 14*(2), 186.

Shih, J. H., Konno, Y., Chang, Y. T., & Emura, T. (2023). A class of general pretest estimators for the univariate normal mean. *Communications in Statistics - Theory and Methods, 25*(8), 2538–2561.

Shih, J. H., Lin, T. Y., Jimichi, M., & Emura, T. (2021). Robust ridge M-estimators with pretest and Stein-rule shrinkage for an intercept term. *Japanese Journal of Statistics and Data Science, 4*, 107–150.

Taketomi, N., Chang, Y. T., Konno, Y., Mori, M., & Emura, T. (2024). Confidence interval for normal means in meta-analysis based on a pretest estimator. *Japanese Journal of Statistics and Data Science, 7*, 537–568.

Taketomi, N., & Emura, T. (2023). Consistency of the estimator for the common mean in fixed-effect meta-analyses. *Axioms, 12*(5), 503.

Zhang, Z., & Nadarajah, S. (2018). On confidence interval estimation of normal percentiles. *Japanese Journal of Statistics and Data Science, 1*, 373–391.

# Chapter 25
# Prediction of Interruptions in Energy Supply: A Machine Learning Study with Post-Shrinkage Modeling




**Minjie Yu and Ravindra Khattree**

## 25.1 Introduction

It is practically impossible to imagine modern civilizations and societies without energy. Energy is what runs the industry, supports the agriculture, lightens our world, and powers the appliances. Energy is what charges our computers and cell phones and runs our automobiles. Energy is what helps our hospitals operate smoothly where patients are inevitably dependent on medical equipment which runs on electricity. Any interruption to the energy supply puts our lives into tailspin, and if this interruption is prolonged, then the losses to business can amount to hundreds of millions of dollars. A case in point is the big power outage of Aug 14, 2003, in the USA when a transmission fault in Ohio was caused by contact with a tree. The domino effect resulted in one of the largest outages ever in North America, affecting more than 50 million people in 8 states in the USA and the Ontario province of Canada. Economic losses were estimated to be at least 10 billions, according to the Electricity Consumers Resource Council (2004). Similarly, a winter storm in Texas in February 2021 caused a widespread blackout when the state's electric grid operator ceased to adequately supply power in the extreme cold weather. As a result millions were left without heat for days in the frigid temperatures.

These events and other similar episodes all over the world underscore the importance of the problem of predicting interruptions in the power supply. Needless

M. Yu
DTE Energy, Distribution Operation Performance Center, Detroit, MI, USA
e-mail: minjieyu@oakland.edu

R. Khattree (✉)
Department of Mathematics and Statistics, Center for Data Science and Big Data Analytics, Oakland University, Rochester, MI, USA
e-mail: Khattree@oakland.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

M. Arashi, M. Norouzirad (eds.), *Advances in Shrinkage and Penalized Estimation Strategies*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-94050-7_25

to say that any such prediction is helpful in advance planning and in taking the precautionary actions to minimize the losses, economic, of lives or otherwise.

This chapter deals with one such problem for the state of Michigan which has a very unstable weather, with rains and storms all through the year, heavy snowfall, and extremely cold weather during winters. DTE Energy is one of the main energy suppliers in the state and routinely faces power interruptions throughout the year. Needless to say, a priori signals for such interruptions are weak and further these signals are many. How do we then arrive at a reasonable prediction? Furthermore, can there be a machine learning approach where one can learn from a model which suitably assembles the information in various variables which are unequal in quality?

To address the above question and to forecast daily interruptions in the supply of energy, as well as to identify the significant variables that influence outage volume using various weather variables and reliability program data, this chapter focuses on the application of various post-shrinkage strategies. We apply these strategies to select the optimal set of predictors and estimate the corresponding coefficients for accurately predicting the number of interruptions and provide insights into the effectiveness of programs in improving the service to the energy customers.

The organization of the chapter is as follows. In Sect. 25.2, we describe the basic problem and the dataset and also introduce the definitions of various variables of potential relevance in the models to be developed. Section 25.3 introduces the basic model and then various modified or improved versions thereof. Various approaches to the analysis are also briefly described in this section. This forms the basis for various extensive analyses and statistical evaluations of their merits which are done in Sect. 25.4. Such evaluations must be done on a test sample independently where models are separately trained on a training sample. Given the nature of the basic problem, emphasis is on prediction. In Sect. 25.5, we then take an entirely different approach for modeling based on copula transformation and analyze the same data with this new approach. A comparative evaluation of the above two approaches is also discussed. In Sect. 25.6, we discuss the analysis of entire data by combining the test and training samples. We finally close our discussion in Sect. 25.7 with some important concluding remarks about our analyses, interpretations, and approaches.

## 25.2 Description of the Problem, Variables, and Basic Model

The problem here is to predict the daily interruptions to the power supply as defined by the daily count of sustained power interruptions in the entire electricity service territory of the company. We will denote it by $\mathcal{Y}$. To better forecast the daily interruptions, we take into account various factors that may influence the outcome. These factors include a range of weather variables, and reliability upgrade program variables. In addition, to account for aspects such as seasonality, we also define variables for week and month. A total of 73 predictors were considered and will be

referred to as $X_1, X_2, \ldots, X_{73}$ throughout this article. These predictors are listed in Table 25.1.

The dataset utilized in this analysis is sourced from the DTE Electric Company and comprises outage data for the period from year 2020 to 2023. For the response variable $\mathcal{Y}$, the number of daily interruptions as reported by the smart meter for the whole service territory, a histogram is shown in Fig. 25.1a. Clearly the distribution of the response variable is highly skewed. Taking the logarithm of $\mathcal{Y}$ somewhat improves the skewness. The corresponding distribution shown in Fig. 25.1b looks more symmetric. Nonetheless, asymmetry still cannot be completely overlooked. The Shapiro-Wilks normality test confirms that, but we can see from Fig. 25.2 that it is close to the normal distribution after excluding a few outliers. That is however not a major problem since we intent to predict interruptions using some or all of the predictors and hence we are more concerned about the conditional distribution of the natural logarithm of the number of daily interruptions, namely, $\log(\mathcal{Y})$. Let us denote it by $\mathbf{Y}$.

## 25.3 The Statistical Model

We will first describe the basic model and its variations. With a multitude of variables, one has to be mindful of the strength of signals provided by each of these variables. For improved estimation, these in turn will determine the submodels. The idea is to not ignore moderate or weak signals but to use them to improve the model. This is where our shrinkage strategies lie. The number of variables is as such quite large, and for simplicity, we will not include any polynomial or cross-product (interaction) terms. Many variables are categorical and they will be treated as such. There may be multicollinearity among variables, but since the objective here is not to select or discard variables or to define new transformations, to save the space, we will not address these aspects of the problem here. Our analyses will address various penalized and post-shrinkage strategies only.

### *25.3.1 The Full Model Estimator*

To start with, the basic model relating $\mathbf{Y} = \log(\mathcal{Y})$ to all predictors is the standard linear model:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \text{Rank}(\mathbf{X}) = p < n. \tag{25.1}$$

With the number of predictors as defined above, the matrix $\mathbf{X}$ is a $n \times p = 1463 \times 74$ matrix. $\mathbf{Y}$ is the vector of data on dependent variable, $\boldsymbol{\beta}$ is the unknown linear parameter vector, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of random errors. We assume errors to be randomly and independently distributed as normal with $\mathrm{E}[\boldsymbol{\varepsilon}] = 0$, and

**Table 25.1** Variable definitions

| Variable | Description |
|---|---|
| $X_1$ | Maximum Air Temperature [F] |
| $X_2$ | Minimum Air Temperature [F] |
| $X_3$ | Maximum Dew Point [F] |
| $X_4$ | Minimum Dew Point [F] |
| $X_5$ | Daily Precipitation [inch] |
| $X_6$ | Average Wind Speed [knots] |
| $X_7$ | Average Wind Direction [deg] |
| $X_8$ | Minimum Relative Humidity [percentage] |
| $X_9$ | Average Relative Humidity [percentage]: computed by time averaging the available reports, it is not average of the daily max/min |
| $X_{10}$ | Maximum Relative Humidity [percentage] |
| $X_{11}$ | Reported Snowfall [inch] |
| $X_{12}$ | Reported Snow Depth [inch] |
| $X_{13}$ | Maximum sustained wind speed in knots |
| $X_{14}$ | Maximum wind gust in knots |
| $X_{15}$ | 1 when lightning occurs, 0 otherwise |
| $X_{16}$ | 1 when freezing rain occurs, 0 otherwise |
| $X_{17}$ | 1 when moderate rain occurs, 0 otherwise |
| $X_{18}$ | 1 when heavy rain occurs, 0 otherwise |
| $X_{19}$ | 1 when light rain occurs, 0 otherwise |
| $X_{20}$ | 1 when moderate thunderstorm occurs, 0 otherwise |
| $X_{21}$ | 1 when heavy thunderstorm occurs, 0 otherwise |
| $X_{22}$ | 1 when light thunderstorm occurs, 0 otherwise |
| $X_{23}$ | 1 when moderate thunderstorm with rainfall occurs, 0 otherwise |
| $X_{24}$ | 1 when light thunderstorm with rainfall occurs, 0 otherwise |
| $X_{25}$ | 1 when freezing drizzle occurs, 0 otherwise |
| $X_{26}$ | 1 when moderate drizzle occurs, 0 otherwise |
| $X_{27}$ | 1 when heavy drizzle occurs, 0 otherwise |
| $X_{28}$ | 1 when light drizzle occurs, 0 otherwise |
| $X_{29}$ | 1 when squalls occurs, 0 otherwise |
| $X_{30}$ | 1 when heavy hail occurs, 0 otherwise |
| $X_{31}$ | 1 when light hail occurs, 0 otherwise |
| $X_{32}$ | 1 when Cloud-to-cloud lightning occurs, 0 otherwise |
| $X_{33}$ | 1 when Cloud-to-air lightning occurs, 0 otherwise |
| $X_{34}$ | 1 when Cloud-to-ground lightning occurs, 0 otherwise |
| $X_{35}$ | 1 when ice pellets occur, 0 otherwise |
| $X_{36}$ | 1 when snow pellets occur, 0 otherwise |
| $X_{37} - X_{41}$ | Number of reliability upgrade program A performed in the last 3 months, 6 months, 1, 2, and 3 years |
| $X_{42} - X_{46}$ | Number of reliability upgrade program B performed in the last 3 months, 6 months, 1, 2, and 3 years |

(continued)

**Table 25.1** (continued)

| Variable | Description |
|---|---|
| $X_{47} - X_{51}$ | Number of reliability upgrade program C performed in the last 3 months, 6 months, 1, 2, and 3 years |
| $X_{52} - X_{56}$ | Number of reliability upgrade program D performed in the last 3 months, 6 months, 1, 2, and 3 years |
| $X_{57} - X_{61}$ | Number of reliability upgrade program E performed in the last 3 months, 6 months, 1, 2, and 3 years |
| $X_{62} - X_{66}$ | Number of reliability upgrade program F performed in the last 3 months, 6 months, 1, 2, and 3 years |
| $X_{67}$ | 1 when small storm occurs, 0 otherwise |
| $X_{68}$ | 1 when median storm occurs, 0 otherwise |
| $X_{69}$ | 1 when large storm occurs, 0 otherwise |
| $X_{70}$ | 1 when catastrophic level 1 storm occurs, 0 otherwise |
| $X_{71}$ | 1 when catastrophic level 2 storm occurs, 0 otherwise |
| $X_{72}$ | Periodic value for week-number of the year, calculated as week of year $\times \frac{2\pi}{53}$ for the year 2020 and 2021, while week-number of year $\times \frac{2\pi}{52}$ for other years |
| $X_{73}$ | Periodic value for month, calculated as month $\times \frac{2\pi}{12}$ |

$\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\sigma^2 I_n$, where $\sigma^2 > 0$. The matrix $\mathbf{X}$ is assumed to be of full column rank $p < n$.

Equation (25.1) gives us what we call as the *full model*. The method of maximum likelihood is used to estimate the coefficients for various predictor variables. Some of the predictors may not provide strong signals for the response variable, and thus it may be necessary to modify the model either by eliminating some of the predictors (thereby obtaining a submodel) or by de-emphasizing the roles of these weak predictors by using various specialized techniques.

Let $\mathbf{X} = \mathbf{X}_1 | \mathbf{X}_2$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. Then, Eq. (25.1) can be written as

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{25.2}$$

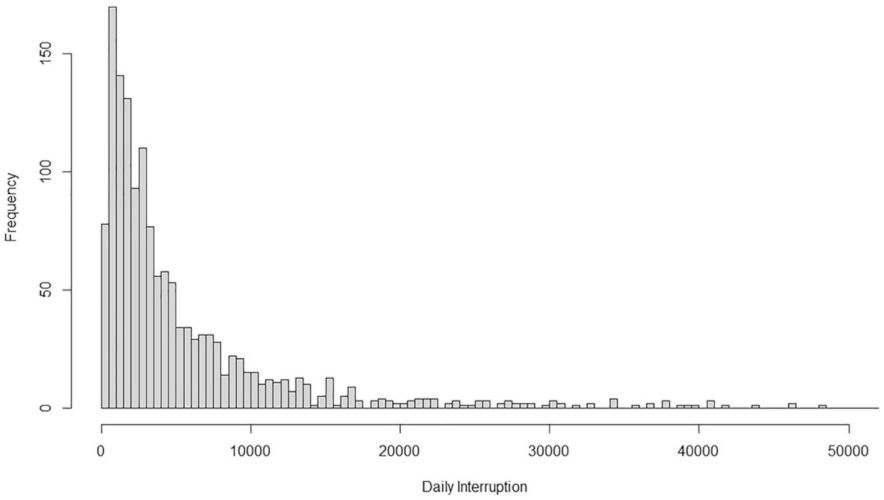
where $\boldsymbol{\beta}_1$ is the vector of parameters for variables $\mathbf{X}_1$ that may be retained in the submodel. We essentially assume that $\boldsymbol{\beta}_2 = \mathbf{0}$ and thus the reduced model is

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}. \tag{25.3}$$

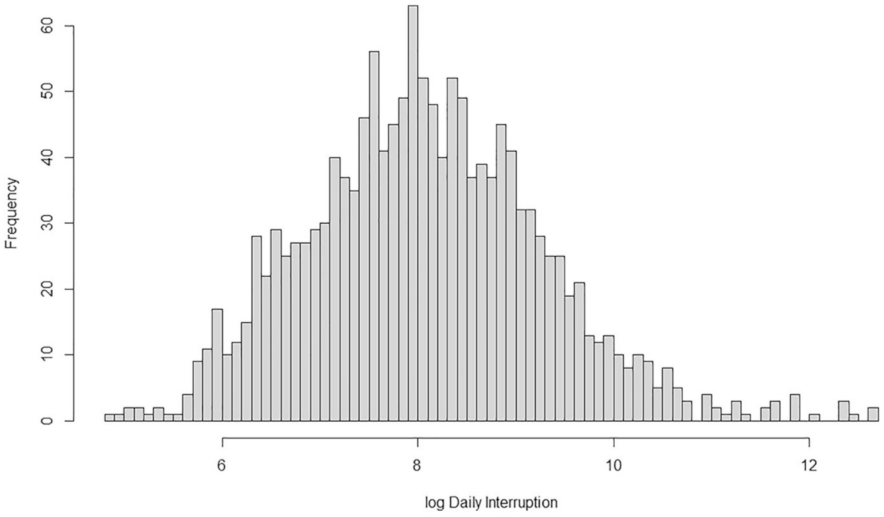
By assuming $\boldsymbol{\beta}_2 = \mathbf{0}$, this model corresponds to the assumption that predictors corresponding to matrix $\mathbf{X}_2$ are unimportant. This decision to drop these variables may be arrived at through various variable selection approaches.

Under model in Eq. (25.1), the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{\mathrm{FM}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} \\ \hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} \end{bmatrix} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}. \tag{25.4}$$




(a) Daily number ofinterruptions.




(b) Logarithm of the daily number ofinterruptions.

**Fig. 25.1** Histograms of daily number of interruptions and logarithm of the daily number of interruptions

It is easy to verify that the full model estimator of $\boldsymbol{\beta}_1$ can be written as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} = (\mathbf{X}_1^\top \mathbf{S} \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{S} \mathbf{Y} \tag{25.5}$$

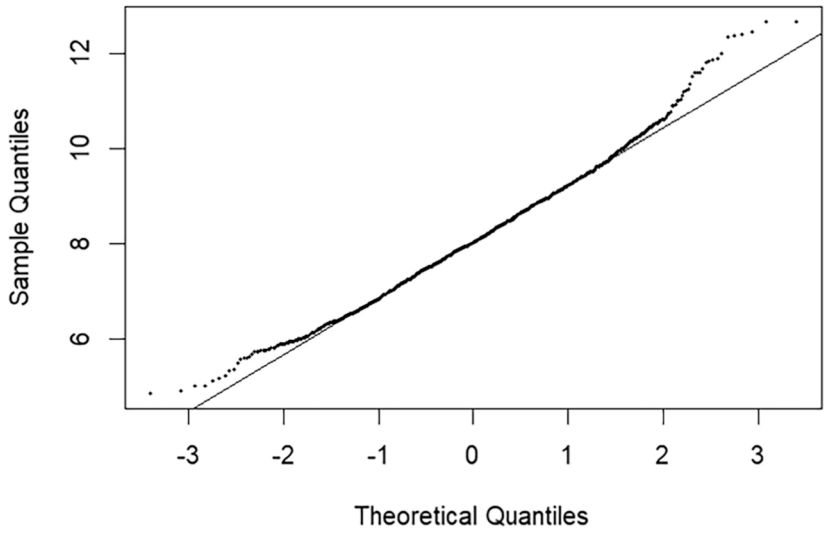



**Fig. 25.2** Normal Q-Q plot of logarithm of the daily number of interruptions

where $\mathbf{S} = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top$. Also under model in Eq. (25.3), the maximum likelihood estimator of $\boldsymbol{\beta}_1$ is given by

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \mathbf{Y}. \tag{25.6}$$

We will employ a variety of estimation techniques depending on how we regularize the parameters. Various approaches of regularization and corresponding criteria as well as various post-selection estimation techniques are briefly discussed in the next two subsections. These approaches have been widely discussed in the recent literature. Specifically, for the ridge regression-related discussion, readers are referred to Saleh et al. (2019), Yüzbaşı et al. (2020), Al-Momani and Arashi (2024), and Du Plessis et al. (2023).

For scaled LASSO and other related techniques, refer to Arashi et al. (2021) and numerous references given therein. Most importantly, in a recent book volume by Saleh et al. (2022), these authors provide a detailed discussion of the post-selection estimation with special emphasis on machine learning. Also, see Ahmed et al. (2023). Our work here borrows from ideas inherent in the abovementioned work, and we thus skip any detailed theoretical discussions of these techniques.

### 25.3.2 Submodel Selection and Penalized Methods

Numerous model selection techniques have been developed to effectively eliminate predictors with sparse signals and retain only the most significant variables, resulting in a refined submodel.

The classical variable selection methods rely on the full model estimators and employ stepwise regression with Akaike information criterion (AIC) and Bayesian information criterion (BIC) to identify a handful of highly influential predictors. Additionally, various penalized model selection methods (such as the LASSO method) may be employed to obtained stable prediction equations. We will describe some of these that we will use below. See Ahmed et al. (2023):

- The Least Absolute Shrinkage and Selection Operator (LASSO) is a widely utilized penalized method that leverages an $L_1$-norm penalty to regularize regression parameters. Mathematically, it is

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\boldsymbol{\beta}_j)^2 + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \right\}. \tag{25.7}$$

  The first term in above is the sum of squared errors and the second term is a $L_1$-norm penalty function. The parameter $\lambda$ is suitably selected through a grid search, where the value that minimizes the BIC is chosen for this analysis. The same approach is used for choosing the tuning parameter $\lambda$ (or $\lambda_1$, $\lambda_2$) for ridge, Elastic Net (ENET), and ALASSO. In the LASSO estimation, sufficiently small coefficients are shrunk toward zero.
- The ridge estimation shrinks the coefficients by using a $L_2$-norm penalty and minimizes the sum of squared error. Specifically,

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\boldsymbol{\beta}_j)^2 + \lambda \sum_{j=1}^{p} \boldsymbol{\beta}_j^2 \right\}. \tag{25.8}$$

- The Elastic Net (ENET) regularization is introduced as another method which is especially effective when some of the independent variables are highly correlated (Zou & Hastie, 2005). It combines a $L_1$ penalty term and $L_2$ penalty term in the objection function and thus,

$$\hat{\boldsymbol{\beta}}^{\text{ENET}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\boldsymbol{\beta}_j)^2 + \lambda_1 \sum_{j=1}^{p} |\boldsymbol{\beta}_j| + \lambda_2 \sum_{j=1}^{p} \boldsymbol{\beta}_j^2 \right\}. \tag{25.9}$$

- The adaptive LASSO (ALASSO) method modifies the LASSO estimator to use adaptive weights on the $L_1$ penalty. It is

$$\hat{\boldsymbol{\beta}}^{\text{ALASSO}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\boldsymbol{\beta}_j)^2 + \lambda \sum_{j=1}^{p} \hat{\omega_j} |\boldsymbol{\beta}_j| \right\}. \tag{25.10}$$

  It is known that this estimator has the property of model consistency. See Zou (2006).
- The Smoothly Clipped Absolute Deviation (SCAD) method combines the favorable characteristics of subset selection and ridge regression, and produces sparse solutions that effectively addresses variable selection (Fan & Li, 2001). The estimator is derived as

$$\hat{\boldsymbol{\beta}}^{\text{SCAD}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} X_{ij}\boldsymbol{\beta}_j)^2 + \lambda \sum_{j=1}^{p} p_{\alpha,\lambda}(|\boldsymbol{\beta}_j|) \right\}. \tag{25.11}$$

  where $p_{\alpha,\lambda}(\cdot)$ is the smoothly clipped absolute deviation penalty function. This penalty function leaves large values of $\theta$ not excessively penalized and makes the solution continuous, defined by

$$p'_{\alpha,\lambda}(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(\alpha\lambda - \theta)_+}{(\alpha - 1)\lambda} I(\theta > \lambda) \right\} \tag{25.12}$$

  where $\lambda > 0$ and $\alpha > 2$ are the tuning parameters that are optimized through the grid search aiming to minimize BIC in this analysis, $\theta > 0$, the function $f(z) = z^+$ is equal to $z$ if $z \geq 0$ and 0 otherwise.

### *25.3.3 Post-Shrinkage Estimators*

The post-shrinkage estimators are useful when a model has some strong signals, a few moderate signals, and many weak signals. It combines estimates from the full model with the estimates from a submodel to improve the estimation of the submodel.

The estimator in Eq. (25.6) ignores all information available in sub-matrix $\mathbf{X}_2$. The corresponding variables may or may not be important depending on whether or not the model is over-fitted (i.e., whether $\boldsymbol{\beta}_2 \neq \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$). To this end, we may want to consider the idea propagated by James and Stein (1961) and define a shrinkage estimator of $\boldsymbol{\beta}_1$ by utilizing a smooth function of a corresponding test

statistics for testing $\boldsymbol{\beta}_2 \neq \mathbf{0}$ and by shrinking the full model estimator toward the direction of submodel estimator as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{S}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}})\Big(1 - (p_2 - 2)\tau_n^{-1}\Big),\ p_2 > 3 \tag{25.13}$$

where the test statistics $\tau_n$ is defined as

$$\tau_n = \frac{n}{\hat{\sigma}^2}(\hat{\boldsymbol{\beta}}_2^{\mathrm{LSE}})^{\top}\mathbf{X}_2^{\top}\mathbf{S}_1\mathbf{X}_2(\hat{\boldsymbol{\beta}}_2^{\mathrm{LSE}})$$

with $\hat{\boldsymbol{\beta}}_2^{\mathrm{LSE}} = (\mathbf{X}_2^{\top}\mathbf{S}_1\mathbf{X}_2)^{-1}\mathbf{X}_2^{\top}\mathbf{S}_1\mathbf{Y}$ and $\mathbf{S}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^{\top}\mathbf{X}_1)^{-1}\mathbf{X}_1^{\top}$. Also,

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{FM}})^{\top}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{FM}})}{n}$$

is a consistent estimator of $\sigma^2$. See Ahmed (2014).

Another estimator similar to that in Eq. (25.13) is the positive shrinkage estimator. It was introduced to control the over-shrinking of the shrinkage estimator which may results in reversing the sign of the full model estimator. It is defined as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{PS}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} + \Big(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}\Big)\Big(1 - (p_2 - 2)\tau_n^{-1}\Big)^{+} \tag{25.14}$$

where $z^{+} = \max(0, z)$.

Our approach will be to apply all of the above estimation methods to our data and examine which of these result in superior prediction of the logarithm of daily interruptions $Y$. Later, in a separate section, we also adopt an approach based on a copula transformation of the data on quantitative variables.

A comparison of these estimators given by Eqs. (25.7)–(25.14) is warranted. The mean prediction error (PE) of a model using an estimator $\hat{\boldsymbol{\beta}}_1^{*}$ is defined as

$$\mathrm{PE}(\hat{\boldsymbol{\beta}}_1^{*}) = \frac{1}{n}\big(\mathbf{Y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^{*}\big)^{\top}\big(\mathbf{Y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^{*}\big). \tag{25.15}$$

It is appropriate that to avoid any possible bias, such evaluations are independent of the data used to estimate the vector $\boldsymbol{\beta}$. Thus, the data are randomly divided into training data which are used to arrive at the corresponding estimator and the test data which are used to evaluate the performance. Let them be $(\mathbf{Y}_{\mathrm{Training}} : \mathbf{X}_{\mathrm{Training}})$ and $(\mathbf{Y}_{\mathrm{Test}} : \mathbf{X}_{\mathrm{Test}})$. Accordingly, Eq. (25.15) for the test data becomes:

$$\mathrm{PE}(\hat{\boldsymbol{\beta}}_1^{*}) = \frac{1}{n_{\mathrm{Test}}}\big(\mathbf{Y}_{\mathrm{Test}} - (\mathbf{X}_1)_{\mathrm{Test}}\hat{\boldsymbol{\beta}}_1^{*}\big)^{\top}\big(\mathbf{Y}_{\mathrm{Test}} - (\mathbf{X}_1)_{\mathrm{Test}}\hat{\boldsymbol{\beta}}_1^{*}\big), \tag{25.16}$$

where $\mathbf{Y}_{\text{Test}}$ is the vector of response values in the test dataset and $\hat{\boldsymbol{\beta}}_1^*$ is a typical estimator obtained via one of Eqs. (25.7)–(25.14) with $\mathbf{X}$ matrix replaced by $\mathbf{X}_{\text{Training}}$. To ensure a robust measure, we employ bootstrap sampling to calculate the average of the mean prediction errors for each estimator given by Eq. (25.16) across all bootstrap samples. Mathematically, this can be expressed as

$$\text{Average PE}(\hat{\boldsymbol{\beta}}_1^*) = \frac{1}{B}\sum_{i=1}^{B} \text{PE}_i(\hat{\boldsymbol{\beta}}_1^*), \tag{25.17}$$

where $B$ is the number of bootstrap iterations and $\text{PE}_i(\hat{\boldsymbol{\beta}}_1^*)$ represents the mean prediction error for the $i$ th bootstrap sample.

## 25.4 The Analysis of the Data

We will utilize post-shrinkage strategies and penalized methods to fit the data and then compare the average prediction errors of the resulting prediction equations to identify the optimal choices.

To achieve this, we will employ the criteria of AIC, BIC, and LASSO and accordingly construct the shrinkage and positive shrinkage estimators using submodels obtained from these techniques. The corresponding statistics are defined below:

- $\text{AIC} = 2k - 2\log(\hat{L})$,
  where $k$ is the number of estimated parameters in the model and $\hat{L}$ is the maximized value of the likelihood function for the model. In the AIC approach, we compare different candidate models based on their AIC values, and prefer the model with the smallest value:
- $\text{BIC} = k\log(n) - 2\log(\hat{L})$,
  where $n$ is the number of observations $k$ and $\hat{L}$ are defined the same as above. The BIC criterion involves computing the BIC value for each candidate model and selecting the one with the lowest value. Compared to AIC, BIC imposes a stronger penalty for models with more parameters.
- In the LASSO solution, sufficient small coefficients are shrunk toward zero and hence are dropped from the model. The resulting submodel is the one selected by LASSO.

Equipped with all of the above machinery, we now describe our analyses. Specifically, we first start with the full model.

### 25.4.1 Full Model Estimation

We first fit the full model that incorporates all independent variables. The coefficients obtained from this model will be utilized to construct the shrinkage estimators. Table 25.2 provides the estimates of the coefficients for the full model along with various other statistics.

Upon examining the output, we identify several statistically significant variables that considerably impact the response of interest, for example, $X_3$, $X_5$, $X_{13}$. Additionally, there are variables that exhibit a moderate influence on the response, such as $X_{12}$, $X_{47}$, and $X_{49}$. However, certain variables, such as $X_1$, $X_{11}$, $X_{73}$, display a weak influence on the response variable.

Figure 25.3a provides the residual plot where we see that the points are randomly scattered around the horizontal line. This indicates that the model successfully captures the underlying relationship in the dataset. The Q-Q plot in Fig. 25.3b compares empirical quantiles of the residuals to the theoretical quantiles of a normal distribution. The points generally fall along the diagonal line. This suggests that the residuals are approximately normally distributed.

Figure 25.3c shows the square root of the absolute standardized residuals and the fitted values to assess the homoscedasticity assumption of a linear regression model. There is no trend in the plot, thereby indicating the homoscedasticity. This suggests that the variance of the residuals remains relatively constant. Figure 25.3d helps identify influential observations that may have a large impact on the model's fit and conclusions. From the plot, we can observe that there are no points on upper right corner that exhibit both high leverage and large standardized residuals.

In accordance with the post-shrinkage method, the independent variables that do not have a significant impact on the response variable are regarded as the nuisance variables. However, they are not to be entirely disregarded. The method employs an adaptive approach by combining the full model with a selected submodel, ensuring that the underlying information from all covariates is taken into consideration. See Eqs. (25.13) and (25.14).

### 25.4.2 Submodel Estimation

In order to construct our shrinkage estimators, we utilize various criteria such as AIC, BIC, and LASSO to obtain the corresponding submodels. These methods thus assist us in determining the most appropriate prediction equations.

The model that is chosen using the AIC consists of 39 independent variables. The BIC criterion selects a model with only 10 independent variables. The LASSO criterion uses a penalty term to ensure the sparsity in the model, effectively shrinking less important variables towards zero and selecting the most relevant variables. It selects a model with 20 independent variables for our data. The selected submodels are given in Tables 25.3, 25.4, and 25.5 respectively.

**Table 25.2** Estimates of the regression coefficients for the full model

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| (Intercept) | −1.4349 | 2.3624 | 0.5437 | $R^2 = 0.5891$ |
| $X_1$ | 0.0033 | 0.0072 | 0.6443 | Adjusted $R^2 = 0.5675$ |
| $X_2$ | 0.0153 | 0.0075 | 0.0408 | $F = 27.2800$ |
| $X_3$ | 0.0172 | 0.0075 | 0.0231 | RootSqErr = 0.7954 |
| $X_4$ | −0.0127 | 0.0077 | 0.1015 | P-Value $< 2.2 \times 10^{-16}$ |
| $X_5$ | 0.3988 | 0.1075 | 0.0002 | |
| $X_6$ | 0.0279 | 0.0134 | 0.0371 | |
| $X_7$ | 0.0002 | 0.0002 | 0.5103 | |
| $X_8$ | −0.0059 | 0.0050 | 0.2323 | |
| $X_9$ | 0.0132 | 0.0082 | 0.1088 | |
| $X_{10}$ | 0.0041 | 0.0052 | 0.4354 | |
| $X_{11}$ | 0.0089 | 0.0417 | 0.8310 | |
| $X_{12}$ | 0.0364 | 0.0194 | 0.0609 | |
| $X_{13}$ | 0.0264 | 0.0113 | 0.0194 | |
| $X_{14}$ | 0.0080 | 0.0061 | 0.1891 | |
| $X_{15}$ | 0.2637 | 0.0961 | 0.0062 | |
| $X_{16}$ | 0.1810 | 0.3024 | 0.5496 | |
| $X_{17}$ | 0.4022 | 0.2616 | 0.1245 | |
| $X_{18}$ | 0.1452 | 0.1026 | 0.1573 | |
| $X_{19}$ | −0.3294 | 0.2614 | 0.2078 | |
| $X_{20}$ | −0.0291 | 0.2105 | 0.8899 | |
| $X_{21}$ | 0.1768 | 0.1660 | 0.2871 | |
| $X_{22}$ | 1.5509 | 0.8866 | 0.0805 | |
| $X_{23}$ | −0.1870 | 0.2884 | 0.5168 | |
| $X_{24}$ | −1.5606 | 0.9169 | 0.0890 | |
| $X_{25}$ | 0.0950 | 0.2473 | 0.7008 | |
| $X_{26}$ | −0.0510 | 0.1685 | 0.7621 | |
| $X_{27}$ | 0.0379 | 0.5856 | 0.9484 | |
| $X_{28}$ | 0.1180 | 0.1631 | 0.4697 | |
| $X_{29}$ | 0.2636 | 0.2636 | 0.3175 | |
| $X_{30}$ | −0.0353 | 0.3763 | 0.9253 | |
| $X_{31}$ | −0.6767 | 0.6110 | 0.2683 | |
| $X_{32}$ | −0.0612 | 0.1362 | 0.6534 | |
| $X_{33}$ | −0.2504 | 0.2202 | 0.2556 | |
| $X_{34}$ | 0.1894 | 0.1870 | 0.3114 | |
| $X_{35}$ | −0.2477 | 0.2129 | 0.2448 | |
| $X_{36}$ | 1.4986 | 0.8483 | 0.0775 | |
| $X_{37}$ | 0.0338 | 0.0135 | 0.0124 | |
| $X_{38}$ | 0.0324 | 0.0193 | 0.0933 | |
| $X_{39}$ | 0.0324 | 0.0214 | 0.1292 | |
| $X_{40}$ | 0.0191 | 0.0219 | 0.3825 | |

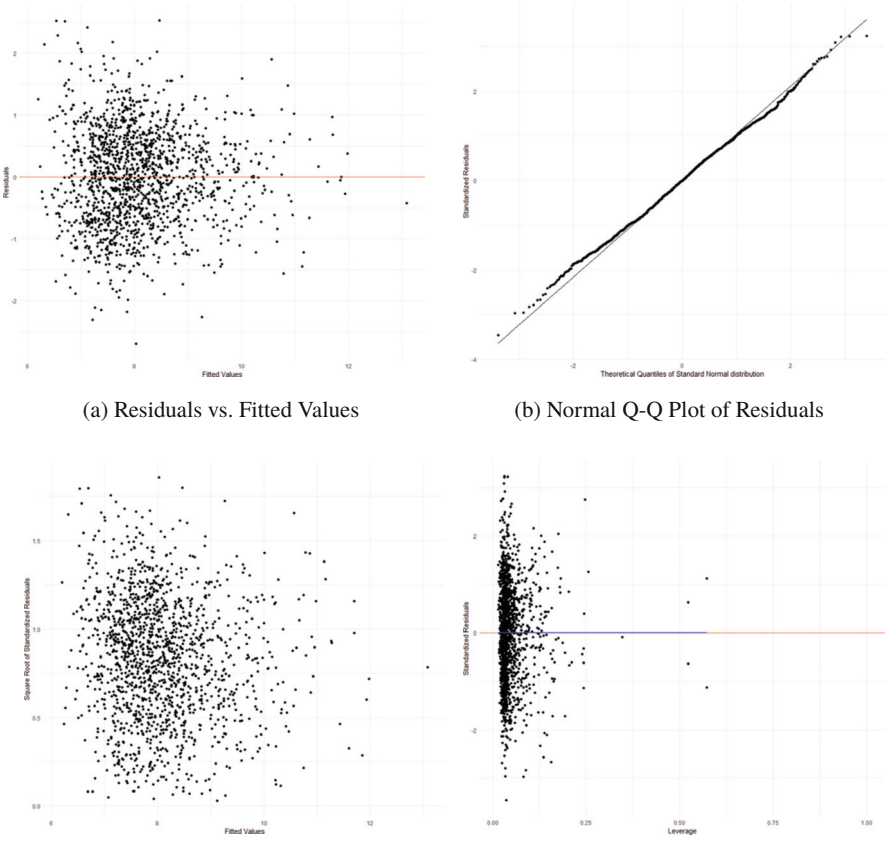
(continued)

**Table 25.2** (continued)

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| $X_{41}$ | −0.0088 | 0.0244 | 0.7169 | |
| $X_{42}$ | 0.0002 | 0.0018 | 0.9104 | |
| $X_{43}$ | −0.0008 | 0.0021 | 0.6944 | |
| $X_{44}$ | 0.0004 | 0.0019 | 0.8404 | |
| $X_{45}$ | −0.0003 | 0.0010 | 0.8085 | |
| $X_{46}$ | −0.0009 | 0.0007 | 0.2170 | |
| $X_{47}$ | 0.0072 | 0.0037 | 0.0502 | |
| $X_{48}$ | 0.0025 | 0.0061 | 0.6776 | |
| $X_{49}$ | 0.0126 | 0.0069 | 0.0703 | |
| $X_{50}$ | 0.0186 | 0.0067 | 0.0058 | |
| $X_{51}$ | 0.0103 | 0.0046 | 0.0254 | |
| $X_{52}$ | 0.0051 | 0.0034 | 0.1320 | |
| $X_{53}$ | −0.0040 | 0.0035 | 0.2548 | |
| $X_{54}$ | −0.0006 | 0.0038 | 0.8673 | |
| $X_{55}$ | 0.0019 | 0.0039 | 0.6217 | |
| $X_{56}$ | 0.0041 | 0.0033 | 0.2178 | |
| $X_{57}$ | −0.0316 | 0.0178 | 0.0757 | |
| $X_{58}$ | −0.0039 | 0.0215 | 0.8571 | |
| $X_{59}$ | −0.0263 | 0.0230 | 0.2545 | |
| $X_{60}$ | −0.0360 | 0.0212 | 0.0899 | |
| $X_{61}$ | −0.0563 | 0.0231 | 0.0148 | |
| $X_{62}$ | −0.0006 | 0.0364 | 0.9860 | |
| $X_{63}$ | −0.1390 | 0.0621 | 0.0253 | |
| $X_{64}$ | −0.2455 | 0.0929 | 0.0083 | |
| $X_{65}$ | 0.0535 | 0.0909 | 0.5566 | |
| $X_{66}$ | 0.0383 | 0.0733 | 0.6012 | |
| $X_{67}$ | 0.6252 | 0.1516 | 0.0001 | |
| $X_{68}$ | 0.9122 | 0.1488 | $1.13\times10^{-9}$ | |
| $X_{69}$ | 0.7587 | 0.2111 | 0.0003 | |
| $X_{70}$ | 1.0506 | 0.1216 | $<2\times10^{-16}$ | |
| $X_{71}$ | 1.5200 | 0.1414 | $<2\times10^{-16}$ | |
| $X_{72}$ | −0.0024 | 0.2694 | 0.9929 | |
| $X_{73}$ | −0.0620 | 0.1941 | 0.7495 | |

### *25.4.3 Development of Refined Estimators*

In the previous section, we have defined several refined estimators, either based on penalty or shrinkage. Here we obtain these estimators for our data. By applying the technique of bootstrap, we will evaluate their performances. To do so, for each bootstrap sample of same size, we construct a training sample to obtain the model and corresponding estimates of the regression coefficients. These are then evaluated




(a) Residuals vs. Fitted Values

(b) Normal Q-Q Plot of Residuals

(c) Examination of Homoscedasticity

(d) Residuals vs. Leverage

**Fig. 25.3** Full model diagnostics plots

on a test sample for their predictive performances. Test data consists of 25% of the entire sample and these observations are randomly chosen.

For each training dataset, we fit the full model and the selected submodels, followed by likelihood ratio tests for each pair of the full model and a selected submodel. We then obtain various post-shrinkage estimators based on Eqs. (25.13) and (25.14). In addition, the model is fitted using five penalization methods, including LASSO, ENET, ALASSO, SCAD, and Ridge. The performance of each model is evaluated based on average prediction error. For illustration, Table 25.6 presents the predictions for just one bootstrap sample, along with values of $\mathcal{Y}$ and $\hat{\mathcal{Y}} = \exp\{\hat{Y}\}$. See Table 25.7 for corresponding prediction errors. For brevity, only nine observations are listed in Table 25.6, three each corresponding to low, moderate, and high values of $\mathcal{Y}$.

**Table 25.3** Estimates of the regression coefficients for the AIC-based submodel

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| (Intercept) | −0.3886 | 1.0541 | 0.712450 | $R^2 = 0.5843$ |
| $X_2$ | 0.0153 | 0.0071 | 0.030666 | Adjusted $R^2 = 0.5729$ |
| $X_3$ | 0.0198 | 0.0050 | 0.000072 | $F = 51.28$ |
| $X_4$ | −0.0110 | 0.0065 | 0.090003 | RootSqErr = 0.7904 |
| $X_5$ | 0.4323 | 0.1018 | 0.000023 | P-Value $< 2\times10^{-16}$ |
| $X_6$ | 0.0255 | 0.0124 | 0.040649 | |
| $X_8$ | −0.0085 | 0.0037 | 0.021302 | |
| $X_9$ | 0.0176 | 0.0051 | 0.000552 | |
| $X_{12}$ | 0.0339 | 0.0180 | 0.059893 | |
| $X_{13}$ | 0.0394 | 0.0075 | $2.06\times10^{-7}$ | |
| $X_{15}$ | 0.2547 | 0.0823 | 0.002023 | |
| $X_{17}$ | 0.4887 | 0.2257 | 0.030544 | |
| $X_{18}$ | 0.1726 | 0.0946 | 0.068320 | |
| $X_{19}$ | −0.4180 | 0.2271 | 0.065945 | |
| $X_{22}$ | 1.2598 | 0.8120 | 0.121008 | |
| $X_{24}$ | −1.3237 | 0.8129 | 0.103664 | |
| $X_{36}$ | 1.1650 | 0.8084 | 0.149781 | |
| $X_{37}$ | 0.0325 | 0.0084 | 0.000117 | |
| $X_{38}$ | 0.0343 | 0.0098 | 0.000468 | |
| $X_{39}$ | 0.0369 | 0.0110 | 0.000836 | |
| $X_{40}$ | 0.0344 | 0.0080 | 0.000018 | |
| $X_{43}$ | −0.0013 | 0.0006 | 0.031814 | |
| $X_{46}$ | −0.0006 | 0.0003 | 0.067710 | |
| $X_{47}$ | 0.0074 | 0.0026 | 0.005190 | |
| $X_{49}$ | 0.0117 | 0.0025 | 0.0000037 | |
| $X_{50}$ | 0.0137 | 0.0032 | 0.0000179 | |
| $X_{51}$ | 0.0079 | 0.0018 | 0.0000162 | |
| $X_{52}$ | 0.0067 | 0.0014 | 0.0000024 | |
| $X_{56}$ | 0.0054 | 0.0022 | 0.017829 | |
| $X_{57}$ | −0.0302 | 0.0109 | 0.005607 | |
| $X_{59}$ | −0.0325 | 0.0102 | 0.001465 | |
| $X_{60}$ | −0.0462 | 0.0135 | 0.000665 | |
| $X_{61}$ | −0.0572 | 0.0176 | 0.001192 | |
| $X_{63}$ | −0.1442 | 0.0392 | 0.000239 | |
| $X_{64}$ | −0.2907 | 0.0611 | 0.000002 | |
| $X_{67}$ | 0.6553 | 0.1449 | 0.000007 | |
| $X_{68}$ | 0.9398 | 0.1425 | $5.99\times10^{-11}$ | |
| $X_{69}$ | 0.7742 | 0.2061 | 0.000179 | |
| $X_{70}$ | 1.0799 | 0.1170 | $<2\times10^{-16}$ | |
| $X_{71}$ | 1.5509 | 0.1333 | $<2\times10^{-16}$ | |

**Table 25.4** Estimates of the regression coefficients for the BIC-based submodel

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| (Intercept) | 5.5667 | 0.1038 | $<2\times10^{-16}$ | $R^2 = 0.5543$ |
| $X_3$ | 0.0274 | 0.0015 | $<2\times10^{-16}$ | Adjusted $R^2 = 0.5512$ |
| $X_5$ | 0.6185 | 0.0856 | $8.14\times10^{-13}$ | $F = 180.6$ |
| $X_{13}$ | 0.0506 | 0.0040 | $<2\times10^{-16}$ | RootSqErr = 0.8102 |
| $X_{15}$ | 0.2599 | 0.0707 | 0.000244 | P-Value $< 2\times10^{-16}$ |
| $X_{52}$ | 0.0023 | 0.0006 | 0.000248 | |
| $X_{67}$ | 0.7602 | 0.1457 | 0.000000208 | |
| $X_{68}$ | 0.9142 | 0.1431 | $2.28\times10^{-10}$ | |
| $X_{69}$ | 0.8507 | 0.2055 | 0.0000369 | |
| $X_{70}$ | 1.2717 | 0.1104 | $<2\times10^{-16}$ | |
| $X_{71}$ | 1.6988 | 0.1254 | $<2\times10^{-16}$ | |

**Table 25.5** Estimates of the regression coefficients for the LASSO-based submodel

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| (Intercept) | −0.3309 | 1.1600 | 0.77555 | $R^2 = 0.4413$ |
| $X_3$ | 0.0280 | 0.0028 | $< 2\times10^{-16}$ | Adjusted $R^2 = 0.4336$ |
| $X_7$ | −0.0001 | 0.0003 | 0.84591 | $F = 56.95$ |
| $X_8$ | −0.0047 | 0.0040 | 0.23586 | RootSqErr = 0.9103 |
| $X_9$ | 0.0199 | 0.0051 | 0.000098 | P-Value $< 2\times10^{-16}$ |
| $X_{13}$ | 0.0605 | 0.0100 | $1.72\times10^{-9}$ | |
| $X_{14}$ | 0.01282 | 0.0064 | 0.04834 | |
| $X_{42}$ | 0.0021 | 0.0012 | 0.08283 | |
| $X_{43}$ | 0.0031 | 0.0013 | 0.01392 | |
| $X_{44}$ | 0.0062 | 0.0012 | 0.000001 | |
| $X_{45}$ | 0.0016 | 0.0008 | 0.04551 | |
| $X_{46}$ | 0.0004 | 0.0004 | 0.44665 | |
| $X_{47}$ | 0.0103 | 0.0026 | 0.00009 | |
| $X_{49}$ | 0.0020 | 0.0027 | 0.45803 | |
| $X_{50}$ | 0.0123 | 0.0026 | 0.000004 | |
| $X_{51}$ | 0.0073 | 0.0025 | 0.00341 | |
| $X_{52}$ | −0.0028 | 0.0019 | 0.13661 | |
| $X_{53}$ | −0.0070 | 0.0017 | 0.00003 | |
| $X_{54}$ | −0.0067 | 0.0023 | 0.00317 | |
| $X_{55}$ | −0.0030 | 0.0019 | 0.11494 | |
| $X_{56}$ | 0.0033 | 0.0020 | 0.10915 | |

Our evaluation is based on 250 bootstrap samples. We outline the steps followed in our approach below:

1. Perform the bootstrap with 250 iterations.
2. Randomly split each bootstrap sample into training and testing sets in a ratio of 3/4:1/4.

**Table 25.6** Predicted results for estimators in a single bootstrap sample

| Obs. | $y$ | $\hat{y}^{\text{FM}}$ | $\hat{y}^{\text{SM.BIC}}$ | $\hat{y}^{\text{S.BIC}}$ | $\hat{y}^{\text{PS.BIC}}$ | $\hat{y}^{\text{SM.AIC}}$ | $\hat{y}^{\text{S.AIC}}$ | $\hat{y}^{\text{PS.AIC}}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 149 | 869 | 933 | 900 | 900 | 903 | 911 | 903 |
| 2 | 206 | 3246 | 2177 | 2668 | 2668 | 2864 | 2780 | 2864 |
| 3 | 214 | 746 | 1132 | 916 | 916 | 714 | 707 | 714 |
| 4 | 2959 | 3629 | 3887 | 3754 | 3754 | 3495 | 3464 | 3495 |
| 5 | 2967 | 3258 | 3273 | 3265 | 3265 | 3085 | 3045 | 3085 |
| 6 | 2999 | 3323 | 3350 | 3336 | 3336 | 3554 | 3611 | 3554 |
| 7 | 37502 | 32635 | 24261 | 28215 | 28215 | 37316 | 38522 | 37316 |
| 8 | 40506 | 36863 | 43975 | 40197 | 40197 | 44885 | 47031 | 44885 |
| 9 | 48150 | 43429 | 51916 | 47405 | 47405 | 63218 | 69108 | 63218 |

| Obs. | $\hat{y}^{\text{SM.LASSO}}$ | $\hat{y}^{\text{S.LASSO}}$ | $\hat{y}^{\text{PS.LASSO}}$ | $\hat{y}^{\text{LASSO}}$ | $\hat{y}^{\text{ENET}}$ | $\hat{y}^{\text{ALASSO}}$ | $\hat{y}^{\text{SCAD}}$ | $\hat{y}^{\text{Ridge}}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 723 | 845 | 845 | 969 | 1046 | 863 | 918 | 973 |
| 2 | 2685 | 3153 | 3153 | 2349 | 2686 | 2239 | 2309 | 3048 |
| 3 | 952 | 775 | 775 | 1313 | 1148 | 1181 | 1081 | 907 |
| 4 | 3639 | 3631 | 3631 | 3943 | 3412 | 3780 | 3662 | 4044 |
| 5 | 4529 | 3425 | 3425 | 3226 | 3179 | 3211 | 3216 | 3185 |
| 6 | 4093 | 3430 | 3430 | 3208 | 3413 | 2958 | 3309 | 3439 |
| 7 | 14881 | 28957 | 28957 | 16758 | 25602 | 26546 | 25172 | 26354 |
| 8 | 16022 | 32471 | 32471 | 32128 | 47411 | 46765 | 45116 | 49730 |
| 9 | 15251 | 37032 | 37032 | 32635 | 51490 | 47709 | 51241 | 53767 |

**Table 25.7** Prediction error of estimators in a single bootstrap sample

| Estimator | Prediction error |
|---|---|
| Full model | 0.7571333 |
| Submodel BIC | 0.7258863 |
| Shrinkage estimator BIC | 0.7247393 |
| Positive shrinkage estimator BIC | 0.7247393 |
| Submodel AIC | 0.7282156 |
| Shrinkage estimator AIC | 0.7289327 |
| Positive shrinkage estimator AIC | 0.7289327 |
| Submodel LASSO | 0.9069681 |
| Shrinkage estimator LASSO | 0.7503530 |
| Positive shrinkage estimator LASSO | 0.7503530 |
| LASSO | 0.7520862 |
| ENET | 0.7353917 |
| aLASSO | 0.7349673 |
| SCAD | 0.7318220 |
| Ridge | 0.7550037 |

3. For each bootstrap iteration:
    a. Fit the full model using the training dataset and obtain the coefficients of all independent variables.
    b. Fit the three submodels (selected previously using AIC, BIC, and LASSO) using the training dataset and obtain the coefficients of the selected variables in each submodel.
    c. Perform the likelihood ratio test on the full model with each of the above three submodels. With the test statistic, compute the shrinkage estimators and positive shrinkage estimators for each pair of full model and submodel estimator.
    d. Fit the models with other five penalized methods (LASSO, ENET, aLASSO, SCAD, and Ridge) using the training dataset and store the coefficients.
    e. Estimate the average prediction errors based on the test dataset for the following models:
        i. Full model
        ii. Three submodels
        iii. Three models with shrinkage estimators
        iv. Three models with positive shrinkage estimators
        v. Five penalized models (LASSO, ENET, aLASSO, SCAD, and Ridge).
4. Calculate the averages and standard deviations of the mean prediction errors over all bootstrap samples.

By taking the averages and standard deviations of respective mean prediction errors, we can evaluate the performance of each model and compare their predictive accuracy, as shown in the Columns 2 and 3 of Table 25.8. When analyzing each bootstrap sample, if we focus solely on the top 100 days with the highest count of outages in the testing dataset, the average prediction errors and corresponding standard deviations for different estimators are updated as Columns 4 and 5 of the same table.

### *25.4.4 Future Predictions*

Models so far obtained have attempted to relate the number of outages to various predictors in current time, and they assured us that such associations do exist and improvement can be realized by penalized approach and via shrinkage. The practical problem, however, is the prediction of outages in future, say for next day, in advance, using the past data on predictors. In order to prepare the company a priori and thus minimize the inconvenience to the customers, this is what we actually need. This necessitates modeling of response in terms of lagged variables.

We adopt the same approach as earlier, but the predictors are now lagged by 1 day. Obviously, the estimated coefficients in full model must change, and the

**Table 25.8** Average and standard deviation (s.d.) of prediction errors for estimators (all data and worst 100 days' data)

| | All data | | Worst 100 days' data | |
|---|---|---|---|---|
| | Prediction error | | Prediction error | |
| Estimator | Average | s.d. | Average | s.d. |
| Full model | 0.6478 | 0.0444 | 0.8649 | 0.1218 |
| Submodel BIC | 0.6582 | 0.0452 | 0.9213 | 0.1211 |
| Shrinkage estimator BIC | 0.6312 | 0.0422 | 0.8626 | 0.1165 |
| Positive shrinkage estimator BIC | 0.6312 | 0.0422 | 0.8626 | 0.1165 |
| Submodel AIC | 0.6330 | 0.0421 | 0.8476 | 0.1159 |
| Shrinkage estimator AIC | 0.6323 | 0.0423 | 0.8462 | 0.1166 |
| Positive shrinkage estimator AIC | 0.6322 | 0.0422 | 0.8460 | 0.1164 |
| Submodel LASSO | 0.8333 | 0.0667 | 1.2687 | 0.2114 |
| Shrinkage estimator LASSO | 0.6440 | 0.0443 | 0.8544 | 0.1230 |
| Positive shrinkage estimator LASSO | 0.6440 | 0.0443 | 0.8544 | 0.1230 |
| LASSO | 0.6832 | 0.0501 | 0.9933 | 0.1428 |
| ENET | 0.6623 | 0.0460 | 0.9458 | 0.1298 |
| ALASSO | 0.6693 | 0.0464 | 0.9467 | 0.1251 |
| SCAD | 0.6624 | 0.0466 | 0.9174 | 0.1230 |
| Ridge | 0.6418 | 0.0435 | 0.8844 | 0.1202 |

**Table 25.9** Selected variables for submodels of future predictions

| Criterion | Variables selected |
|---|---|
| BIC | $X_1$, $X_3$, $X_4$, $X_5$, $X_9$, $X_{17}$, $X_{33}$, $X_{41}$, $X_{46}$, $X_{47}$, $X_{49}$, $X_{50}$, $X_{51}$, $X_{52}$, $X_{61}$, $X_{64}$, $X_{68}$, $X_{70}$, $X_{71}$ |
| AIC | $X_2$, $X_3$, $X_4$, $X_5$, $X_7$, $X_9$, $X_{14}$, $X_{16}$, $X_{17}$, $X_{25}$, $X_{26}$, $X_{30}$, $X_{32}$, $X_{33}$, $X_{37}$, $X_{38}$, $X_{39}$, $X_{40}$, $X_{41}$, $X_{43}$, $X_{46}$, $X_{47}$, $X_{49}$, $X_{50}$, $X_{51}$, $X_{52}$, $X_{56}$, $X_{57}$, $X_{59}$, $X_{60}$, $X_{61}$, $X_{63}$, $X_{64}$, $X_{68}$, $X_{70}$, $X_{71}$ |
| LASSO | $X_3$, $X_4$, $X_7$, $X_{14}$, $X_{42}$, $X_{43}$, $X_{44}$, $X_{45}$, $X_{46}$, $X_{47}$, $X_{49}$, $X_{50}$, $X_{52}$, $X_{53}$, $X_{54}$, $X_{55}$, $X_{56}$ |

submodels selected would not be the same as earlier in terms of predictors as well as their coefficients. In addition, we must expect somewhat relatively poorer fit and higher average prediction errors. This being the most realistic situation, we present the variables chosen by three criteria in Table 25.9.

As earlier, improved estimation is desired, and this is especially welcome given that corresponding full and reduced models will have somewhat poorer fit. Table 25.10 gives us the average prediction errors, computed for the test data, for various approaches. We also give the average prediction errors for the 100 worst days of outages.

It is clear from Table 25.10 that all shrinkage-based methods using the AIC criterion and those with the BIC criterion are quite effective compared to other approaches with respect to average prediction error as well as the average prediction error computed for 100 worst days. Average prediction errors for these winning

**Table 25.10** Average and standard deviation of prediction errors for future (1-day-forward) predictive models (all data and worst 100 days' data)

| | All data | | Worst 100 days' data | |
|---|---|---|---|---|
| | Prediction error | | Prediction error | |
| Estimator | Average | s.d. | Average | s.d. |
| Full model | 0.9405 | 0.0782 | 1.5842 | 0.2621 |
| Submodel BIC | 0.9246 | 0.0769 | 1.6146 | 0.2636 |
| Shrinkage estimator BIC | 0.9121 | 0.0752 | 1.5786 | 0.2592 |
| Positive shrinkage estimator BIC | 0.9121 | 0.0752 | 1.5786 | 0.2592 |
| Submodel AIC | 0.9144 | 0.0744 | 1.5572 | 0.2582 |
| Shrinkage estimator AIC | 0.9149 | 0.0746 | 1.5563 | 0.2569 |
| Positive shrinkage estimator AIC | 0.9147 | 0.0747 | 1.5560 | 0.2570 |
| Submodel LASSO | 1.0543 | 0.0916 | 1.8823 | 0.3278 |
| Shrinkage estimator LASSO | 0.9300 | 0.0766 | 1.5870 | 0.2656 |
| Positive shrinkage estimator LASSO | 0.9300 | 0.0766 | 1.5870 | 0.2656 |
| LASSO | 0.9731 | 0.0881 | 1.7758 | 0.3086 |
| ENET | 0.9560 | 0.0849 | 1.7351 | 0.3007 |
| ALASSO | 0.9646 | 0.0847 | 1.7295 | 0.2980 |
| SCAD | 0.9597 | 0.0825 | 1.6987 | 0.2891 |
| Ridge | 0.9289 | 0.0783 | 1.6239 | 0.2777 |

models are only slightly different from each other. While the average predictor errors are now larger compared to previously obtained models, these observations are consistent with what we observed for the models which related the current response to the current values of the predictors. In Table 25.11 we present the results from one bootstrap simulation and give the actual values of the response variable and predicted values.

We adopt the same approach as earlier, but the predictors are now lagged by 1 day. Obviously, the estimated coefficients in full model must change, and the submodels selected would not be the same as earlier in terms of predictors as well as their coefficients. Further, we must expect somewhat relatively poorer fit and higher average prediction errors. This being the most realistic situation, we present the variables chosen by three criteria in Table 25.9.

As earlier, improved estimation is desired, and this is especially welcome given that corresponding full and reduced models will have somewhat poorer fit. Table 25.10 gives us the average prediction errors, computed for the test data, for various approaches. We also give the average prediction errors for 100 worst days of outages.

**Table 25.11** Predicted results for future (1-day-forward) predictive models in a single bootstrap sample

| Obs. | $y$ | $\hat{y}^{FM}$ | $\hat{y}^{SM.BIC}$ | $\hat{y}^{S.BIC}$ | $\hat{y}^{PS.BIC}$ | $\hat{y}^{SM.AIC}$ | $\hat{y}^{S.AIC}$ | $\hat{y}^{PS.AIC}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 135 | 1569 | 1719 | 1649 | 1649 | 1559 | 1562 | 1562 |
| 2 | 167 | 812 | 954 | 886 | 886 | 861 | 847 | 847 |
| 3 | 305 | 1385 | 1625 | 1510 | 1510 | 1307 | 1330 | 1330 |
| 4 | 3105 | 2264 | 2085 | 2166 | 2166 | 2673 | 2545 | 2545 |
| 5 | 3156 | 1789 | 1537 | 1647 | 1647 | 1725 | 1743 | 1743 |
| 6 | 3171 | 2745 | 2817 | 2784 | 2784 | 2503 | 2572 | 2572 |
| 7 | 40510 | 20576 | 14057 | 16741 | 16741 | 16037 | 17255 | 17255 |
| 8 | 40891 | 3491 | 4750 | 4125 | 4125 | 3989 | 3836 | 3836 |
| 9 | 41525 | 19748 | 13801 | 16266 | 16266 | 14375 | 15781 | 15781 |

| Obs. | $\hat{y}^{SM.LASSO}$ | $\hat{y}^{S.LASSO}$ | $\hat{y}^{PS.LASSO}$ | $\hat{y}^{LASSO}$ | $\hat{y}^{ENET}$ | $\hat{y}^{ALASSO}$ | $\hat{y}^{SCAD}$ | $\hat{y}^{Ridge}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2229 | 1704 | 1704 | 2026 | 1913 | 1908 | 2084 | 1629 |
| 2 | 757 | 798 | 798 | 1031 | 967 | 910 | 921 | 764 |
| 3 | 1226 | 1346 | 1346 | 1649 | 1591 | 1510 | 1645 | 1393 |
| 4 | 1751 | 2131 | 2131 | 1986 | 2004 | 1727 | 1863 | 2046 |
| 5 | 2180 | 1874 | 1874 | 2160 | 2033 | 1885 | 1881 | 1699 |
| 6 | 2987 | 2800 | 2800 | 2790 | 2864 | 2772 | 2880 | 2960 |
| 7 | 8075 | 16509 | 16509 | 11174 | 11822 | 13723 | 11452 | 17276 |
| 8 | 5882 | 3948 | 3948 | 5463 | 5373 | 6142 | 5600 | 4790 |
| 9 | 12306 | 17667 | 17667 | 13031 | 14212 | 16545 | 13532 | 19968 |

## 25.5 Copula Transformation-Based Modeling

Bahuguna and Khattree (2020) have presented an approach to induce normality in the data by copula transformation under the assumption that data on continuous variables exhibit a dependence which can be approximated by a Gaussian copula. Thus, data from any multivariate distribution with multivariate cumulative distribution function (c.d.f.) $F(\cdot)$ and a Gaussian copula $C(\cdot)$ can be transformed to multivariate normal distribution with c.d.f. $G(\cdot)$ via $C(\cdot)$ as shown in Fig. 25.4.

Thus, in essence, we make the assumption that our data have a common Gaussian copula $G(\cdot) = \Phi_{(\boldsymbol{\mu},\ \boldsymbol{\Sigma})}(\cdot)$. In principle, the choices of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are arbitrary. Since our interest is in doing the multivariate analysis of dependence, we must choose $\boldsymbol{\Sigma}$ cautiously to retain the essential dependence features of data. Since choice of $\boldsymbol{\mu}$ is often unimportant in such a situation, we will

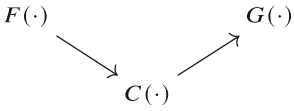



**Fig. 25.4** Transformation through common copula (here $F(\cdot)$ and $G(\cdot)$ are $cdf$s and $C(\cdot)$ is common copula)

take its value to be zero vector. As a graphical representation for a given distribution of data, say $\mathscr{D}$, our transformation works as

$$\mathscr{D} \longrightarrow \mathscr{U} \longrightarrow \mathscr{N}$$

and in a reverse direction as

$$\mathscr{N} \longrightarrow \mathscr{U} \longrightarrow \mathscr{D},$$

where $\mathscr{U}$ represents a multivariate distribution of c.d.f.s (Copula) in which each marginal is Uniform[0, 1] and $\mathscr{N}$ represents a chosen multivariate normal distribution.

Following Bahuguna and Khattree (2020), we will conduct the analysis of our data after transformation, and on the copula-transformed data, we will fit all the previous models. The predictions can be back-transformed as shown in the above diagram. The transformation is applied on the response variable and the 46 numerical independent variables. Subsequently, we conduct regression analysis on these 46 transformed variables, along with other original 27 dummy or binary variables, for the transformed response variables. As earlier, all the previous approaches will be adopted but this time on the copula-transformed data.

For transforming predictions back to the original data, we will apply the interpolation techniques suggested by Bahuguna and Khattree (2020). For this interpolation, we simulate 2,000 draws from the fitted copula distribution. These simulated draws are subsequently employed to transform the prediction results back to the original scale. Details are given in the abovementioned articles.

### *25.5.1 Full Model and Submodels for the Transformed Data*

Let the transformed variables be denoted by $W_i$, while the remaining untransformed predictors retain their original notation as $X_i$. Following the same approach as for the analysis conducted on the original data, we first fit a full regression model. The results of the regression analysis are presented in Table 25.12.

The variables chosen by BIC, AIC, and LASSO criteria are presented in Table 25.13.

### *25.5.2 Computation of Average Prediction Errors*

Since predictions are obtained for the copula-transformed responses, we must take an additional step to retransform these predictions to the original scale. This is the $\mathscr{N} \longrightarrow \mathscr{U} \longrightarrow \mathscr{D}$ step of back transformation. For every $i = 1, , \ldots, n$, $\mathscr{N}_i \longrightarrow \mathscr{U}_i$ step is straightforward, and then by simulating 2,000 observations from the

**Table 25.12** Estimates of the regression coefficients for the full model on copula-transformed data

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| (Intercept) | −0.1585 | 0.1783 | 0.37437 | $R^2 = 0.555$ |
| $W_1$ | 0.2049 | 0.0770 | 0.00788 | Adjusted $R^2 = 0.5316$ |
| $W_2$ | 0.1330 | 0.0871 | 0.12703 | $F = 23.73$ |
| $W_3$ | 0.1510 | 0.0827 | 0.06814 | RootSqErr = 0.6814 |
| $W_4$ | −0.1665 | 0.0920 | 0.07059 | P-Value $< 2 \times 10^{-16}$ |
| $W_5$ | 0.3143 | 0.0623 | $5.09 \times 10^{-7}$ | |
| $W_6$ | 0.0850 | 0.0392 | 0.03024 | |
| $W_7$ | 0.0060 | 0.0192 | 0.75998 | |
| $W_8$ | −0.0360 | 0.0574 | 0.53106 | |
| $W_9$ | 0.0539 | 0.0742 | 0.46759 | |
| $W_{10}$ | 0.0631 | 0.0308 | 0.04122 | |
| $W_{11}$ | −0.1388 | 0.0844 | 0.10041 | |
| $W_{12}$ | 0.2288 | 0.0933 | 0.01434 | |
| $W_{13}$ | 0.0948 | 0.0552 | 0.08644 | |
| $W_{14}$ | 0.0356 | 0.0465 | 0.44536 | |
| $W_{15}$ | 0.1142 | 0.0569 | 0.04512 | |
| $W_{16}$ | 0.0580 | 0.0672 | 0.38814 | |
| $W_{17}$ | −0.0230 | 0.0859 | 0.78893 | |
| $W_{18}$ | 0.1140 | 0.0938 | 0.22465 | |
| $W_{19}$ | 0.1063 | 0.0847 | 0.20973 | |
| $W_{20}$ | −0.0889 | 0.0555 | 0.10974 | |
| $W_{21}$ | −0.1209 | 0.0728 | 0.09731 | |
| $W_{22}$ | −0.0437 | 0.1000 | 0.66199 | |
| $W_{23}$ | −0.0059 | 0.0547 | 0.91444 | |
| $W_{24}$ | −0.0854 | 0.0951 | 0.36919 | |
| $W_{25}$ | 0.1333 | 0.0577 | 0.02106 | |
| $W_{26}$ | 0.0917 | 0.0724 | 0.20582 | |
| $W_{27}$ | 0.2013 | 0.1136 | 0.07678 | |
| $W_{28}$ | 0.2801 | 0.1342 | 0.03719 | |
| $W_{29}$ | 0.1842 | 0.1246 | 0.13974 | |
| $W_{30}$ | 0.0988 | 0.0617 | 0.10945 | |
| $W_{31}$ | −0.0067 | 0.0576 | 0.90720 | |
| $W_{32}$ | −0.0345 | 0.0720 | 0.63209 | |
| $W_{33}$ | 0.0627 | 0.1155 | 0.58758 | |
| $W_{34}$ | 0.0849 | 0.0917 | 0.35480 | |
| $W_{35}$ | −0.1311 | 0.0641 | 0.0412 | |
| $W_{36}$ | 0.0042 | 0.0569 | 0.94100 | |
| $W_{37}$ | −0.0582 | 0.0618 | 0.34643 | |
| $W_{38}$ | −0.0874 | 0.0765 | 0.25370 | |
| $W_{39}$ | −0.0891 | 0.0793 | 0.26121 | |
| $W_{40}$ | 0.0057 | 0.0777 | 0.94190 | |

(continued)

**Table 25.12** (continued)

| Variable | $\hat{\beta}_i$ | $\hat{S}_{\beta_i}$ | P-value | Statistics |
|---|---|---|---|---|
| $W_{41}$ | −0.1501 | 0.0760 | 0.04869 | |
| $W_{42}$ | −0.0966 | 0.0448 | 0.03140 | |
| $W_{43}$ | −0.0194 | 0.0399 | 0.62746 | |
| $W_{44}$ | −0.0632 | 0.0356 | 0.07656 | |
| $W_{45}$ | −0.0818 | 0.0666 | 0.21972 | |
| $W_{46}$ | 0.0169 | 0.0430 | 0.69505 | |
| $X_{15}$ | 0.2253 | 0.0828 | 0.00658 | |
| $X_{16}$ | 0.0496 | 0.2556 | 0.84622 | |
| $X_{17}$ | 0.1121 | 0.2227 | 0.61483 | |
| $X_{18}$ | 0.0575 | 0.0913 | 0.52895 | |
| $X_{19}$ | −0.1219 | 0.2228 | 0.58442 | |
| $X_{20}$ | −0.1270 | 0.1800 | 0.48060 | |
| $X_{21}$ | 0.1215 | 0.1411 | 0.38960 | |
| $X_{22}$ | 0.9763 | 0.7410 | 0.18790 | |
| $X_{23}$ | −0.1636 | 0.2475 | 0.50874 | |
| $X_{24}$ | −0.9457 | 0.7687 | 0.21882 | |
| $X_{25}$ | 0.1242 | 0.2110 | 0.55625 | |
| $X_{26}$ | −0.0913 | 0.1439 | 0.52571 | |
| $X_{27}$ | 0.0067 | 0.5019 | 0.98927 | |
| $X_{28}$ | 0.0672 | 0.1396 | 0.63033 | |
| $X_{29}$ | 0.0682 | 0.2211 | 0.75776 | |
| $X_{30}$ | 0.0073 | 0.3227 | 0.98187 | |
| $X_{31}$ | −0.4407 | 0.5188 | 0.39576 | |
| $X_{32}$ | −0.0522 | 0.1167 | 0.65471 | |
| $X_{33}$ | −0.2272 | 0.1889 | 0.22934 | |
| $X_{34}$ | 0.0866 | 0.1609 | 0.59045 | |
| $X_{35}$ | −0.1512 | 0.1824 | 0.40719 | |
| $X_{36}$ | 1.0634 | 0.7252 | 0.14280 | |
| $X_{67}$ | 0.6538 | 0.1283 | $3.96\times10^{-7}$ | |
| $X_{68}$ | 0.7182 | 0.1285 | $2.76\times10^{-8}$ | |
| $X_{69}$ | 0.6477 | 0.1813 | 0.00036 | |
| $X_{70}$ | 0.8481 | 0.1042 | $8.92\times10^{-16}$ | |
| $X_{71}$ | 1.1553 | 0.1189 | $<2\times10^{-16}$ | |

copula distribution, we interlace the two sets of observations in increasing order, and then for each of $\mathscr{U}_i$ we identify the immediate lower and upper values from the simulated observations. Let them be $\mathscr{U}_{i,\text{lower}}$ and $\mathscr{U}_{i,\text{upper}}$. Accordingly, by $\mathscr{U} \longrightarrow \mathscr{D}$ step, one can obtain limits on $\hat{F}^{-1}(u_i)$ as

$$\hat{F}^{-1}(u_{i,\text{lower}}) \leq \hat{F}^{-1}(u_i) \leq \hat{F}^{-1}(u_{i,\text{upper}}),$$

**Table 25.13** Selected variables for submodels on copula-transformed data

| Criterion | Variables selected |
|---|---|
| BIC | $W_1, W_5, W_6, W_{10}, W_{20}, W_{28}, W_{29}, W_{42}, W_{36}, X_{15}, X_{67}, X_{68}, X_{69}, X_{70}, X_{71}$ |
| AIC | $W_1, W_2, W_3, W_4, W_5, W_6, W_{10}, W_{11}, W_{12}, W_{13}, W_{15}, W_{16}, W_{18}, W_{20}, W_{21}, W_{25}, W_{27}, W_{28}, W_{29}, W_{30}, W_{35}, W_{38}, W_{39}, W_{41}, W_{42}, W_{44}, X_{15}, X_{23}, X_{67}, X_{68}, X_{69}, X_{70}, X_{71}$ |
| LASSO | $W_2, W_3, W_5, W_6, W_{13}, W_{22}, W_{42}, W_{44}$ |

**Table 25.14** Example of transforming copula-transformed data back to the original scale

| $\hat{y}_i^*$ | $u_i$ | $u_{i,\text{lower}}$ | $\hat{F}^{-1}(u_{i,\text{lower}})$ | $u_{i,\text{upper}}$ | $\hat{F}^{-1}(u_{i,\text{upper}})$ | $\hat{F}^{-1}(u_i) = \hat{y}_i$ |
|---|---|---|---|---|---|---|
| −1.6757 | 0.0469 | 0.0468 | 462.9167 | 0.0494 | 478.4278 | 470.6723 |
| −1.5875 | 0.0562 | 0.0559 | 512.1258 | 0.0563 | 515.768 | 513.9469 |
| −1.5649 | 0.0588 | 0.0578 | 526.6234 | 0.0593 | 536.6676 | 531.6455 |
| −1.5433 | 0.0614 | 0.0612 | 546.6164 | 0.0615 | 547.9147 | 547.2655 |
| −1.5347 | 0.0624 | 0.0621 | 550.8367 | 0.0628 | 554.7244 | 552.7805 |

where $\hat{F}^{-1}$ is the empirical distribution function of the data. Thus, we approximate the prediction $\hat{F}^{-1}(u_i)$ as

$$\hat{y}_i = \hat{F}^{-1}(u_i) = \frac{\hat{F}^{-1}(u_{i,\text{lower}}) + \hat{F}^{-1}(u_{i,\text{upper}})}{2}$$

Hence, for each set of predicted values, we convert them back to the original scale. Table 25.14 presents data for an example. The column $\hat{y}_i^*$ shows the prediction results from the full model estimator. By applying the c.d.f., we transform these values into uniformly distributed values, which are displayed in the column $u_i$. By interlacing these uniformly distributed values in ascending order with our simulated data, we identify their positions and corresponding positions in the original scale list. The column $\hat{F}^{-1}(u_{i,\text{lower}})$ and column $\hat{F}^{-1}(u_{i,\text{upper}})$ depict the immediate lower and upper values in the original scale. The column $\hat{F}^{-1}(u_i)$ represents the mean of these two values, which serves as the prediction result that has been converted back to the original scale $\hat{y}_i$.

In order to compare the average prediction errors on the log scale, we take the logarithm of both the prediction results and the actual response values. Subsequently, we calculate the prediction errors for each bootstrap sample. In Table 25.15, columns 2 and 3 present the average and standard deviation prediction errors using this approach, while columns 4 and 5 display the average and standard deviation of prediction errors for different estimators for the bootstrap sample consisting of the top 100 days with the highest count of outages.

**Table 25.15** Average and standard deviation of prediction errors for estimators on copula-transformed data (all data and worst 100 days' data)

| | All data | | Worst 100 days' data | |
|---|---|---|---|---|
| | Prediction error | | Prediction error | |
| Estimator | Average | s.d. | Average | s.d. |
| Full model | 0.6765 | 0.0516 | 0.9780 | 0.1352 |
| Submodel BIC | 1.1353 | 0.1523 | 1.4709 | 0.2561 |
| Shrinkage estimator BIC | 0.7774 | 0.0807 | 1.0353 | 0.1706 |
| Positive shrinkage estimator BIC | 0.7774 | 0.0807 | 1.0353 | 0.1706 |
| Submodel AIC | 1.4534 | 0.2154 | 1.9962 | 0.3269 |
| Shrinkage estimator AIC | 1.0851 | 0.2361 | 1.5247 | 0.3454 |
| Positive shrinkage estimator AIC | 1.0777 | 0.2168 | 1.5182 | 0.3324 |
| Submodel LASSO | 0.8125 | 0.0658 | 1.3034 | 0.1875 |
| Shrinkage estimator LASSO | 0.6675 | 0.0498 | 0.9620 | 0.1291 |
| Positive shrinkage estimator LASSO | 0.6675 | 0.0498 | 0.9620 | 0.1291 |
| LASSO | 0.7301 | 0.0587 | 1.1453 | 0.1557 |
| ENET | 0.6941 | 0.0528 | 1.0610 | 0.1386 |
| ALASSO | 0.7046 | 0.0511 | 1.0702 | 0.1372 |
| SCAD | 0.6887 | 0.0526 | 1.0268 | 0.1333 |
| Ridge | 0.6631 | 0.0476 | 0.9687 | 0.1218 |

**Table 25.16** Selected variables for submodels of future predictions on copula-transformed data

| Criterion | Variables selected |
|---|---|
| BIC | $W_1$, $W_5$, $W_7$, $W_{19}$, $W_{44}$, $X_{68}$, $X_{70}$, $X_{71}$ |
| AIC | $W_1$, $W_4$, $W_5$, $W_7$, $W_8$, $W_9$, $W_{11}$, $W_{12}$, $W_{14}$, $W_{15}$, $W_{17}$, $W_{18}$, $W_{19}$, $W_{25}$, $W_{30}$, $W_{38}$, $W_{39}$, $W_{42}$, $W_{44}$, $X_{16}$, $X_{22}$, $X_{24}$, $X_{25}$, $X_{26}$, $X_{30}$, $X_{33}$, $X_{68}$, $X_{70}$, $X_{71}$ |
| LASSO | $W_1$, $W_3$, $W_5$, $W_{13}$, $W_{42}$, $W_{43}$ |

### *25.5.3 Future Predictions Using Copula-Transformed Approach*

As earlier, we consider an analysis that utilizes the independent variables to forecast the number of outages for the following day. To achieve this, we employ the same dataset but employ the next day's outage value as the response variable. We then perform the regression using the full model estimator and apply BIC, AIC, and LASSO to select submodels, as shown in Table 25.16.

Using the same approach as earlier, we evaluate the performances of various approached here as well. The average prediction errors are listed in Table 25.17, columns 2 and 3, while the comparison result for the testing dataset of the top 100 days with the highest count of outages is columns 4 and 5. Compared to results from previous analyses on the response variable **Y** shown in Tables 25.8 and 25.10, the copula-transformed model yields slightly larger prediction errors for present model and exhibits very similar prediction errors for 1-day-forward forecasts.

**Table 25.17** Average and standard deviation of prediction errors for future (1-day-forward) predictive model on copula-transformed data (all data and worst 100 days' data)

| | All data | | Worst 100 days' data | |
|---|---|---|---|---|
| | Prediction error | | Prediction error | |
| Estimator | Average | s.d. | Average | s.d. |
| Full model | 0.9349 | 0.0749 | 1.6003 | 0.2492 |
| Submodel BIC | 0.9438 | 0.0835 | 1.6987 | 0.2721 |
| Shrinkage estimator BIC | 0.9125 | 0.0783 | 1.6138 | 0.2598 |
| Positive shrinkage estimator BIC | 0.9125 | 0.0783 | 1.6138 | 0.2598 |
| Submodel AIC | 0.9107 | 0.0762 | 1.5974 | 0.2548 |
| Shrinkage estimator AIC | 0.9085 | 0.0755 | 1.5855 | 0.2557 |
| Positive shrinkage estimator AIC | 0.9085 | 0.0754 | 1.5856 | 0.2554 |
| Submodel LASSO | 1.0182 | 0.0893 | 1.8336 | 0.2937 |
| Shrinkage estimator LASSO | 0.9217 | 0.0769 | 1.6102 | 0.2562 |
| Positive shrinkage estimator LASSO | 0.9217 | 0.0769 | 1.6102 | 0.2562 |
| LASSO | 0.9948 | 0.0916 | 1.8167 | 0.2865 |
| ENET | 0.9718 | 0.0859 | 1.7720 | 0.2809 |
| ALASSO | 0.9690 | 0.0864 | 1.6679 | 0.3048 |
| SCAD | 0.9539 | 0.0814 | 1.6713 | 0.2835 |
| Ridge | 0.9196 | 0.0755 | 1.6061 | 0.2514 |

It is also of interest to evaluate if copula transformation helps improve predictions for our data. Table 25.18 presents various test statistics for the corresponding full model where (i) data on $\mathcal{Y}$ are used (ii) data on $\mathbf{Y} = \log(\mathcal{Y})$ are used and (iii) copula-transformed data are used. The model for the present-day prediction using copula-transformed data yields a lower $R^2$ (=0.5550) compared to the other two models. For the models of 1-day-forward prediction, as we have already seen that the data on $\mathcal{Y}$ was highly skewed and thus resulted in very low $R^2$ (=0.2031). The other two models have very comparable $R^2$ values (0.4021 and 0.3941, respectively). This suggests that copula transformation has not improved the full model substantially. The same can be seen for other statistics such as the F-statistic.

## 25.6 The Analysis of Entire Data

So far, we have used the training data to arrive at various models and evaluated them on the test data. For completeness, it is also of interest to analyze the entire data and look at the prediction errors. Clearly since the models use all observations (combining training data and test data) and then compute prediction errors for the very same observations that were used to arrive at the model, the situation may look better than it actually will be when a yet-unobserved future value is to be predicted. These onetime analyses cannot be used for statistical comparisons of various approaches. Nonetheless, it is still of interest to fit these models and looks

**Table 25.18** Statistics of full models for future (1-day-forward) prediction on different response variables

| Type | Response | $R^2$ | Adjusted $R^2$ | RootSqErr | F statistic |
|---|---|---|---|---|---|
| Present day | Daily interruption | 0.5843 | 0.5625 | 14160 | 26.75 |
| | Logarithm of daily interruption | 0.5891 | 0.5675 | 0.7954 | 27.28 |
| | Copula-transformed daily interruption | 0.5550 | 0.5316 | 0.6814 | 23.73 |
| Future prediction | Daily interruption | 0.2031 | 0.1612 | 19610 | 4.846 |
| (One-day-forward) | Logarithm of daily interruption | 0.4021 | 0.3707 | 0.9588 | 12.79 |
| | Copula-transformed daily interruption | 0.3941 | 0.3622 | 0.7952 | 12.37 |

**Table 25.19** Prediction errors for two scenarios (using present-day data and 1-day in the past data) for the whole dataset (training and test combined)

| | Present-day data | | One day in the past data | |
|---|---|---|---|---|
| | Prediction error | | Prediction error | |
| Estimator | All data | Worst 100 days | All data | Worst 100 days |
| Full model | 0.6006 | 0.7835 | 0.8727 | 3.0743 |
| Submodel BIC | 0.6515 | 1.0598 | 0.9086 | 3.3402 |
| Shrinkage estimator BIC | 0.6136 | 0.8924 | 0.8997 | 3.2967 |
| Positive shrinkage estimator BIC | 0.6136 | 0.8924 | 0.8997 | 3.2967 |
| Submodel AIC | 0.6077 | 0.8051 | 0.8821 | 3.1302 |
| Shrinkage estimator AIC | 0.6237 | 0.8481 | 0.9174 | 3.2553 |
| Positive shrinkage estimator AIC | 0.6077 | 0.8051 | 0.8821 | 3.1302 |
| Submodel LASSO | 0.8167 | 2.3561 | 1.0352 | 4.4115 |
| Shrinkage estimator LASSO | 0.6033 | 0.8471 | 0.8801 | 3.2381 |
| Positive shrinkage estimator LASSO | 0.6033 | 0.8471 | 0.8801 | 3.2381 |
| LASSO | 0.6618 | 1.3216 | 0.9340 | 3.7386 |
| ENET | 0.6378 | 1.1286 | 0.9291 | 3.6963 |
| ALASSO | 0.6577 | 1.1024 | 0.9385 | 3.6036 |
| SCAD | 0.6597 | 1.1227 | 0.9432 | 3.6677 |
| Ridge | 0.6138 | 0.9327 | 0.8918 | 3.3760 |

at the prediction errors. These are given in Table 25.19. As one observes, in most cases shrinkage does help. The prediction error values are smaller, except for the last column corresponding to predictions for the next day and for 100 worst days. Smaller prediction errors are to be expected as data and models together are self-validating in that the corresponding models were optimal for the given data and thus their best performances will naturally be seen for that very data.

## 25.7 Conclusions and Closing Remarks

Based on the extensive analyses of the data, few facts come up which we summarize below. The original data was highly skewed for the response variable $\mathcal{Y}$, the daily interruptions. Thus, a transformation of data was called for. The logarithmic transformation as well as copula transformation both considerably improved the full model. However, performance of these two models was comparable. There however can be situations when transformation by copula transformation can result in superior performance.

Uniformly, the model selection method AIC appears to work favorably when it comes to smaller mean prediction error. There are instances when BIC also gives the comparable performance. However, LASSO did not seem to give that level of performance. This is true for both sets of analyses—when we use the current values of predictors as well as when lagged predictors are used. Further, this is true for the log-transformed data as well as the copula-transformed data.

Generally, shrinkage-based models outperform the penalized regression models, in terms of average prediction error. The performances of all shrinkage methods under AIC to select the variables were essentially interchangeable. The dataset is large enough here, and thus positive shrinkage was perhaps not needed because $1 - (p_2 - 2)\tau_n^{-1}$ was greater than zero, in most of the bootstrapped cases.

For the worst 100 days with highest number of interruptions, as expected, the average prediction errors were larger. Nonetheless, the patterns in the relative comparisons of various approaches did not change. The conclusions still remain the same.

One may ask, why not just use a submodel while ignoring all the statistically insignificant variables. There are two legitimate reasons for not doing so: First, there is bound to be high level of multicollinearity in such data with 73 predictors. Thus, the statistical insignificance of various variables only indicates lack of significance of these, while all other variables are retained in the model. Secondly, the dataset is also large enough that despite 73 predictors in the model, it leaves a very large degree of freedom for error. Thus, $R^2$ may be low while the $F$ statistic for any model will be highly statistically significant. That leaves much ambiguity in the interpretation of corresponding submodels. Consequently, it is better to still incorporate the information from these other variables. The shrinkage-based estimators do exactly that.

**Acknowledgments** Minjie Yu (ORCID ID: 0009-0003-7080-1065) wishes to acknowledge Yong Li of DTE Energy for his support in this research effort. The work of Ravindra Khattree (ORCID ID: 0000-0002-9305-2365) was, in part, supported by funding from the NSF Award FAIN: 2244091.

## References


Ahmed, S. E. (2014). *Penalty, shrinkage and pretest strategies: Variable selection and estimation*. Springer.

Ahmed, S. E., Ahmed, F., & Yüzbaşı, B. (2023). *Post-shrinkage strategies in statistical and machine learning for high dimensional data*. Chapman and Hall/CRC.

Al-Momani, M., & Arashi, M. (2024). Ridge-type pretest and shrinkage estimation strategies in spatial error models with an application to a real data example. *Mathematics, 12*(3), 390.

Arashi, M., Asar, Y., & Yüzbaşı, B. (2021). SLASSO: A scaled LASSO for multicollinear situations. *Journal of Statistical Computation and Simulation, 91*(15), 3170–3183.

Bahuguna, M., & Khattree, R. (2020) A generic all-purpose transformation for multivariate modeling through copulas. *International Journal of Data Science and Analytics, 10*(1), 1–23.

Council, Electricity Consumers Resource. (2004). *The economic impacts of the August 2003 blackout*. Washington.

Du Plessis, S., Arashi, M., Maribe, G., & Millard, S. M. (2023). Efficient estimation and validation of shrinkage estimators in big data analytics. *Mathematics, 11*(22), 4632.

Fan, J., & Li, R.(2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association, 96*(456), 1348–1360.

James, W., & Stein, C. (1961) Estimation with quadratic loss. In Neyman, J. (Ed.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Vol. 4(1), pp. 361–379).

Saleh, A. K. Md. E., Arashi, M., & Kibria, B. M. G. (2019). *Theory of ridge regression estimation with applications*. John Wiley & Sons.

Saleh, A. K. Md. E., Arashi, M., Saleh, R. A., & Norouzirad, M. (2022). *Rank-based methods for shrinkage and selection: With application to machine learning*. John Wiley & Sons.

Yüzbaşı, B., Arashi, M., & Ahmed, S. E. (2020). Shrinkage estimation strategies in generalised ridge regression models: Low/high-dimension regime. *International Statistical Review, 88*(1), 229–251.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H., & Hastie, T.(2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67*(2), 301–320.