ICSA Book Series in Statistics
Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Ding-Geng (Din) Chen
Jiahua Chen
Xuewen Lu
Grace Y. Yi
Hao Yu *Editors*

Advanced Statistical Methods in Data Science

Springer

ICSA Book Series in Statistics

More information about this series at http://www.springer.com/series/13402

Ding-Geng (Din) Chen • Jiahua Chen •
Xuewen Lu • Grace Y. Yi • Hao Yu
Editors

Advanced Statistical Methods in Data Science

Springer

Editors
Ding-Geng (Din) Chen
School of Social Work
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

Jiahua Chen
Department of Statistics
University of British Columbia
Vancouver, BC, Canada

Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

Grace Y. Yi
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON, Canada

Xuewen Lu
Department of Mathematics and Statistics
University of Calgary
Calgary, AB, Canada

Hao Yu
Department of Statistics and Actuarial Science
Western University
London, ON, Canada

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-981-10-2593-8 ISBN 978-981-10-2594-5 (eBook)
DOI 10.1007/978-981-10-2594-5

Library of Congress Control Number: 2016959593

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #22-06/08 Gateway East, Singapore 189721, Singapore

To my parents and parents-in-law, who value higher education and hard work; to my wife Ke, for her love, support, and patience; and to my son John D. Chen and my daughter Jenny K. Chen for their love and support.

Ding-Geng (Din) Chen, PhD

To my wife, my daughter Amy, and my son Andy, whose admiring conversations transformed into lasting enthusiasm for my research activities.

Jiahua Chen, PhD

To my wife Xiaobo, my daughter Sophia, and my son Samuel, for their support and understanding.

Xuewen Lu, PhD

To my family, Wenqing He, Morgan He, and Joy He, for being my inspiration and offering everlasting support.

Grace Y. Yi, PhD

Preface

This book is a compilation of invited presentations and lectures that were presented at the Second Symposium of the International Chinese Statistical Association–Canada Chapter (ICSA–CANADA) held at the University of Calgary, Canada, August 4–6, 2015 (http://www.ucalgary.ca/icsa-canadachapter2015). The Symposium was organized around the theme "Embracing Challenges and Opportunities of Statistics and Data Science in the Modern World" with a threefold goal: to promote advanced statistical methods in big data sciences, to create an opportunity for the exchange ideas among researchers in statistics and data science, and to embrace the opportunities inherent in the challenges of using statistics and data science in the modern world.

The Symposium encompassed diverse topics in advanced statistical analysis in big data sciences, including methods for administrative data analysis, survival data analysis, missing data analysis, high-dimensional and genetic data analysis, and longitudinal and functional data analysis; design and analysis of studies with response-dependent and multiphase designs; time series and robust statistics; and statistical inference based on likelihood, empirical likelihood, and estimating functions. This book compiles 12 research articles generated from Symposium presentations.

Our aim in creating this book was to provide a venue for timely dissemination of the research presented during the Symposium to promote further research and collaborative work in advanced statistics. In the era of big data, this collection of innovative research not only has high potential to have a substantial impact on the development of advanced statistical models across a wide spectrum of big data sciences but also has great promise for fostering more research and collaborations addressing the ever-changing challenges and opportunities of statistics and data science. The authors have made their data and computer programs publicly available so that readers can replicate the model development and data analysis presented in each chapter, enabling them to readily apply these new methods in their own research.

The 12 chapters are organized into three sections. Part I includes four chapters that present and discuss data analyses based on latent variable models in data sciences. Part II comprises four chapters that share a common focus on lifetime data analyses. Part III is composed of four chapters that address applied data analyses in big data sciences.

Part I Data Analysis Based on Latent or Dependent Variable Models (Chaps. 1, 2, 3, and 4)

Chapter 1 presents a weighted multiple testing procedure commonly used and known in clinical trials. Given this wide use, many researchers have proposed methods for making multiple testing adjustments to control family-wise error rates while accounting for the logical relations among the null hypotheses. However, most of those methods not only disregard the correlation among the endpoints within the same family but also assume the hypotheses associated with each family are equally weighted. Authors Enas Ghulam, Kesheng Wang, and Changchun Xie report on their work in which they proposed and tested a gatekeeping procedure based on Xie's weighted multiple testing correction for correlated tests. The proposed method is illustrated with an example to clearly demonstrate how it can be used in complex clinical trials.

In Chap. 2, Abbas Khalili, Jiahua Chen, and David A. Stephens consider the regime-switching Gaussian autoregressive model as an effective platform for analyzing financial and economic time series. The authors first explain the heterogeneous behavior in volatility over time and multimodality of the conditional or marginal distributions and then propose a computationally more efficient regularization method for simultaneous autoregressive-order and parameter estimation when the number of autoregressive regimes is predetermined. The authors provide a helpful demonstration by applying this method to analysis of the growth of the US gross domestic product and US unemployment rate data.

Chapter 3 deals with a practical problem of healthcare use for understanding the risk factors associated with the length of hospital stay. In this chapter, Cindy Xin Feng and Longhai Li develop hurdle and zero-inflated models to accommodate both the excess zeros and skewness of data with various configurations of spatial random effects. In addition, these models allow for the analysis of the nonlinear effect of seasonality and other fixed effect covariates. This research draws attention to considerable drawbacks regarding model misspecifications. The modeling and inference presented by Feng and Li use the fully Bayesian approach via Markov Chain Monte Carlo (MCMC) simulation techniques.

Chapter 4 discusses emerging issues in the era of precision medicine and the development of multi-agent combination therapy or polytherapy. Prior research has established that, as compared with conventional single-agent therapy (monotherapy), polytherapy often leads to a high-dimensional dose searching space, especially when a treatment combines three or more drugs. To overcome the burden of calibration of multiple design parameters, Ruitao Lin and Guosheng Yin propose a robust optimal interval (ROI) design to locate the maximum tolerated dose (MTD) in Phase I clinical trials. The optimal interval is determined by minimizing the probability of incorrect decisions under the Bayesian paradigm. To tackle high-

dimensional drug combinations, the authors develop a random-walk ROI design to identify the MTD combination in the multi-agent dose space. The authors of this chapter designed extensive simulation studies to demonstrate the finite-sample performance of the proposed methods.

Part II Lifetime Data Analysis (Chaps. 5, 6, 7, and 8)

In Chap. 5, Longlong Huang, Karen Kopciuk, and Xuewen Lu present a new method for group selection in an accelerated failure time (AFT) model with a group bridge penalty. This method is capable of simultaneously carrying out feature selection at the group and within-group individual variable levels. The authors conducted a series of simulation studies to demonstrate the capacity of this group bridge approach to identify the correct group and correct individual variable even with high censoring rates. Real data analysis illustrates the application of the proposed method to scientific problems.

Chapter 6 considers issues around Case I interval censored data, also known as current status data, commonly encountered in areas such as demography, economics, epidemiology, and medical science. In this chapter, Pooneh Pordeli and Xuewen Lu first introduce a partially linear single-index proportional odds model to analyze these types of data and then propose a method for simultaneous sieve maximum likelihood estimation. The resultant estimator of regression parameter vector is asymptotically normal, and, under some regularity conditions, this estimator can achieve the semiparametric information bound.

Chapter 7 presents a framework for general empirical likelihood inference of Type I censored multiple samples. Authors Song Cai and Jiahua Chen develop an effective empirical likelihood ratio test and efficient methods for distribution function and quantile estimation for Type I censored samples. This newly developed approach can achieve high efficiency without requiring risky model assumptions. The maximum empirical likelihood estimator is asymptotically normal. Simulation studies show that, as compared to some semiparametric competitors, the proposed empirical likelihood ratio test has superior power under a wide range of population distribution settings.

Chapter 8 provides readers with an overview of recent developments in the joint modeling of longitudinal quality of life (QoL) measurements and survival time for cancer patients that promise more efficient estimation. Authors Hui Song, Yingwei Peng, and Dongsheng Tu then propose semiparametric estimation methods to estimate the parameters in these joint models and illustrate the applications of these joint modeling procedures to analyze longitudinal QoL measurements and recurrence times using data from a clinical trial sample of women with early breast cancer.

Part III Applied Data Analysis (Chaps. 9, 10, 11, and 12)

Chapter 9 presents an interesting discussion of a confidence weighting model applied to multiple-choice tests commonly used in undergraduate mathematics and statistics courses. Michael Cavers and Joseph Ling discuss an approach to multiple-choice testing called the student-weighted model and report on findings based on the implementation of this method in two sections of a first-year calculus course at the University of Calgary (2014 and 2015).

Chapter 10 discusses parametric imputation in missing data analysis. Author Peisong Han proposes to estimate and subtract the asymptotic bias to obtain consistent estimators. Han demonstrates that the resulting estimator is consistent if any of the missingness mechanism models or the imputation model is correctly specified.

Chapter 11 considers one of the basic and important problems in statistics: the estimation of the center of a symmetric distribution. In this chapter, authors Pengfei Li and Zhaoyang Tian propose a new estimator by maximizing the smoothed likelihood. Li and Tian's simulation studies show that, as compared with the existing methods, their proposed estimator has much smaller mean square errors under uniform distribution, t-distribution with one degree of freedom, and mixtures of normal distributions on the mean parameter. Additionally, the proposed estimator is comparable to the existing methods under other symmetric distributions.

Chapter 12 presents the work of Jingjia Chu, Reg Kulperger, and Hao Yu in which they propose a new class of multivariate time series models. Specifically, the authors propose a multivariate time series model with an additive GARCH-type structure to capture the common risk among equities. The dynamic conditional covariance between series is aggregated by a common risk term, which is key to characterizing the conditional correlation.

As a general note, the references for each chapter are included immediately following the chapter text. We have organized the chapters as self-contained units so readers can more easily and readily refer to the cited sources for each chapter.

The editors are deeply grateful to many organizations and individuals for their support of the research and efforts that have gone into the creation of this collection of impressive, innovative work. First, we would like to thank the authors of each chapter for the contribution of their knowledge, time, and expertise to this book as well as to the Second Symposium of the ICSA–CANADA. Second, our sincere gratitude goes to the sponsors of the Symposium for their financial support: the Canadian Statistical Sciences Institute (CANSSI), the Pacific Institute for the Mathematical Sciences (PIMS), and the Department of Mathematics and Statistics, University of Calgary; without their support, this book would not have become a reality. We also owe big thanks to the volunteers and the staff of the University of Calgary for their assistance at the Symposium. We express our sincere thanks to the Symposium organizers: Gemai Chen, PhD, University of Calgary; Jiahua Chen, PhD, University of British Columbia; X. Joan Hu, PhD, Simon Fraser University; Wendy Lou, PhD, University of Toronto; Xuewen Lu, PhD, University of Calgary; Chao Qiu, PhD, University of Calgary; Bingrui (Cindy) Sun, PhD, University of Calgary; Jingjing Wu, PhD, University of Calgary; Grace Y. Yi, PhD, University of Waterloo; and Ying Zhang, PhD, Acadia University. The editors wish to acknowledge the professional support of Hannah Qiu (Springer/ICSA Book Series coordinator) and Wei Zhao (associate editor) from Springer Beijing that made publishing this book with Springer a reality.

We welcome readers' comments, including notes on typos or other errors, and look forward to receiving suggestions for improvements to future editions of this book. Please send comments and suggestions to any of the editors listed below.

University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
Ding-Geng (Din) Chen, MSc, PhD

University of British Columbia
Vancouver, BC, Canada
Jiahua Chen, MSc, PhD

University of Calgary
Calgary, AB, Canada
Xuewen Lu, MSc, PhD

University of Waterloo
Waterloo, ON, Canada
Grace Y. Yi, MSc, MA, PhD

Western University
West Ontario, ON, Canada
Hao Yu, MSc, PhD

July 28, 2016

Contents

Contributors

Song Cai School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

Michael Cavers Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

Jiahua Chen Big Data Research Institute of Yunnan University and Department of Statistics, University of British Columbia, Vancouver, BC, Canada

Jingjia Chu Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada

Cindy Xin Feng School of Public Health and Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, SK, Canada

Enas Ghulam Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA

Peisong Han Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Longlong Huang Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

Abbas Khalili Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

Karen Kopciuk Department of Cancer Epidemiology and Prevention Research, Alberta Health Services, Calgary, AB, Canada

Reg Kulperger Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada

Longhai Li Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, Canada

Pengfei Li Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Ruitao Lin Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

Joseph Ling Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

Xuewen Lu Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

Yingwei Peng Departments of Public Health Sciences and Mathematics and Statistics, Queens University, Kingston, ON, Canada

Pooneh Pordeli Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada

Hui Song School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, China

David A. Stephens Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

Zhaoyang Tian Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Dongsheng Tu Departments of Public Health Sciences and Mathematics and Statistics, Queens University, Kingston, ON, Canada

Kesheng Wang Department of Biostatistics and Epidemiology, East Tennessee State University, Johnson City, TN, USA

Changchun Xie Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA

Guosheng Yin Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

Hao Yu Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada

Part I
Data Analysis Based on Latent or Dependent Variable Models

Chapter 1
The Mixture Gatekeeping Procedure Based on Weighted Multiple Testing Correction for Correlated Tests

Enas Ghulam, Kesheng Wang, and Changchun Xie

Abstract Hierarchically ordered objectives often occur in clinical trials. Many multiple testing adjustment methods have been proposed to control family-wise error rates while taking into account the logical relations among the null hypotheses. However, most of them disregard the correlation among the endpoints within the same family and assume the hypotheses within each family are equally weighted. This paper proposes a gatekeeping procedure based on Xie's weighted multiple testing correction for correlated tests (Xie, Stat Med 31(4):341–352, 2012). Simulations have shown that it has power advantages compared to those non-parametric methods (which do not depend on the joint distribution of the endpoints). An example is given to illustrate the proposed method and show how it can be used in complex clinical trials.

1.1 Introduction

In order to obtain better overall knowledge of a treatment effect, the investigators in clinical trials often collect many endpoints and test the treatment effect for each endpoint. These endpoints might be hierarchically ordered and logically related. However, the problem of multiplicity arises when multiple hypotheses are tested. Ignoring this problem can cause false positive results. Currently, there are two common types of multiple testing adjustment methods. One is based on controlling

E. Ghulam • C. Xie (✉)
Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA
e-mail: xiecn@ucmail.uc.edu

K. Wang
Department of Biostatistics and Epidemiology, East Tennessee State University, Johnson City, TN, USA

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_1

family-wise error rate (FWER), which is the probability of rejecting at least one true null hypothesis, and the other is based on controlling false discovery rate (FDR), which is the expected proportion of false positives among all significant hypotheses (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001). The gatekeeping procedures we consider here belong to the type of FWER control.

Consider a clinical trial with multiple endpoints. The hypotheses associated with these endpoints can be grouped into m ordered families, $F_1, \ldots, F_m$, with $k_1, \ldots, k_m$ hypotheses respectively.

When the endpoints are hierarchically ordered with logical relations, many gatekeeping procedures have been suggested to control FWER including serial gatekeeping (Bauer et al. 1998; Maurer et al. 1995; Westfall and Krishen 2001), parallel gatekeeping (Dmitrienko et al. 2003) and their generalization called tree-structured gatekeeping (Dmitrienko et al. 2008, 2007). In serial gatekeeping procedure, the hypotheses in F_i are tested only if all hypotheses in the previously examined family, F_{i-1} are rejected. Otherwise, the hypotheses in F_i are accepted without testing. In parallel gatekeeping procedure, the hypotheses in F_i are tested only if at least one hypothesis in the previously examined family, F_{i-1} is rejected. In the tree-structured gatekeeping procedure, a hypothesis in F_i is tested only if all hypotheses in one subset (called a serial rejection set from $F_1, \ldots, F_{i-1}$) are rejected and at least one hypothesis in another subset (called a parallel rejection set from $F_1, \ldots, F_{i-1}$) is rejected. Recently, Dmitrienko and Tamhane (2011) proposed a new approach for gatekeeping, based on mixture of multiple testing procedures.

In this paper, we use the mixture method with Xie's weighted multiple testing correction, which is proposed for a single family of hypotheses, as a component procedure. We call the resulting mixture gatekeeping procedure as WMTCc-based gatekeeping procedure. Xie's WMTCc was proposed for multiple correlated tests with different weights and is more powerful than weighted Holm procedure. Thus the proposed new WMTCc-based gatekeeping procedure should have an advantage over the mixture gatekeeping procedure based on Holm procedure, including Bonferroni parallel gatekeeping multiple testing procedure.

1.1.1 WMTCc Method

Assume that the test statistics follow a multivariate normal distribution with known correlation matrix Σ. Let $p_1, \ldots, p_m$ be the observed p-values for null hypotheses $H_0^{(1)}, \ldots, H_0^{(m)}$ respectively and $w_i > 0$, $i = 1, \ldots, m$ be the weight for null hypothesis $H_0^{(i)}$. Note that we do not require that $\sum_{i=1}^{m} w_i = 1$ because it can be seen from Eqs. (1.2) or (1.3) below that the adjusted p-values only depend on the ratios of the weights. For each $i = 1, \ldots, m$, calculate $q_i = p_i / w_i$. Then the adjusted

p-value for the null hypothesis $H_0^{(i)}$ is

$$\begin{aligned} P_{adj\_i} &= P(\min_j q_j \le q_i) \\ &= 1 - P(\text{all } q_j > q_i) \\ &= 1 - P\left(\bigcap_{j=1}^{m} a_j \le X_j \le b_j \right) \\ &= 1 - P(\text{all} \quad p_j > p_i w_j / w_i), \end{aligned} \tag{1.1}$$

where $X_j, j = 1, \ldots, m$ are standardized multivariate normal with correlation matrix Σ and

$$a_j = \Phi^{-1}\left(\frac{p_i w_j}{2w_i}\right), \qquad b_j = \Phi^{-1}\left(1 - \frac{p_i w_j}{2w_i}\right) \tag{1.2}$$

for the two-sided case and

$$a_j = -\infty, \qquad b_j = \Phi^{-1}\left(1 - \frac{p_i w_j}{w_i}\right) \tag{1.3}$$

for the one-sided case.

Therefore the WMTCc is to first adjust the m observed p-values for multiple testing by computing m adjusted p-values in (1.1). If $P_{adj\_i} \le \alpha$, reject the corresponding null hypothesis $H_0^{(i)}$. Suppose k_1 null hypotheses have been rejected, we then adjust the remaining $m - k_1$ observed p-values for multiple testing after removing the rejected k_1 null hypotheses, using the corresponding correlation matrix and weights. Continue the procedures above until there is no null hypothesis left after removing the rejected null hypotheses or there is no null hypothesis which can be rejected.

1.1.2 Single-Step WMTCc Method

The single-step WMTCc is to adjust the m observed p-values for multiple testing by computing m adjusted p-values in (1.1). If $P_{adj\_i} \le \alpha$, reject the corresponding null hypothesis $H_0^{(i)}$. It does not remove the rejected null hypotheses and calculate the adjusted p-value again for the remaining observed P-values as the WMTCc does.

1.2 Mixture Gatekeeping Procedures

Following Dmitrienko and Tamhane (2011), we consider a mixture procedure P from P_1 and P_2 for testing the null hypotheses in family $F = F_1 \cup F_2$. Let $K_1 = \{1, 2, .., k_1\}$, $K_2 = \{k_1 + 1, \ldots, k\}$, $K = K_1 \cup K_2 = \{1, \ldots, k\}$ be the index sets of null hypotheses in F_1, F_2 and F, respectively. Let $H(I) = \cap_{j \in I} H_j$, where $I = I_1 \cup I_2$, in which $I_i \subseteq K_i$, $i = 1, 2$. Let m_1 and m_2 be the number of null hypotheses in I_1 and I_2, respectively. Suppose P_1 is single-step WMTCc and P_2 is the regular WMTCc procedure. The single-step WMTCc tests and rejects any intersection hypothesis $H(I_1)$ at level α if $p_1(I_1) = \min_{i=1}^{m_1} P_{adj\_i} \leq \alpha$. The regular WMTCc tests and rejects any intersection hypothesis $H(I_2)$ at level α if $p_2(I_2) = \min_{i=1}^{m_2} P_{adj\_i} \leq \alpha$.

$$\phi_I(p_1(I_1), p_2(I_2)) = \min(p_1(I_1), \frac{p_2(I_2)}{c(I_1, I_2|\alpha)}), \tag{1.4}$$

where $0 \leq c(I_1, I_2|\alpha) \leq 1$ and must satisfy the following equation:

$$P\{\phi_I(p_1(I_1), p_2(I_2)) \leq \alpha | H(I)\} = P\{p_1(I_1) \leq \alpha \quad \text{or} \quad p_2(I_2) \leq c(I_1, I_2|\alpha)\alpha | H(I)\} = \alpha. \tag{1.5}$$

The package mvtnorm (Genz et al. 2009) in the R software environment (Team 2013) can be used to calculate $c(I_1, I_2|\alpha)$. If we assume the hypotheses within family F_i, $i = 1, 2$ are correlated, but the hypotheses between families are not correlated, $c(I_1, I_2|\alpha)$ can be defined as $1 - e_1(I_1|\alpha)/\alpha$, where $e_1(I_1|\alpha) = P\{p_1(I_1) \leq \alpha | H(I_1)\}$. Note $c(I_1, I_2|\alpha)$ is independent of I_2.

1.3 Simulation Study

In this section, simulations were performed to estimate the family-wise type I error rate (FWER) and to compare the power performance of the two mixture gatekeeping procedures: Holm-based gatekeeping procedure and the proposed new WMTCc-based gatekeeping procedure. In these simulations, two families are considered. Each family has two endpoints.

We simulated a clinical trial with two correlated endpoints and 240 individuals. Each individual had probability 0.5 to receive the active treatment and probability 0.5 to receive a placebo. The two endpoints from each family were generated from a multivariate normal distribution with ρ chosen as 0.0, 0.3, 0.5, 0.7, and 0.9. The treatment effect size was assumed as (0,0,0,0), (0.4,0.1,0.4,0.1), (0.1,0.4,0.1,0.4) and (0.4,0.4,0.4,0.4), where the first two numbers are for the two endpoints in the family 1 and the last two numbers are for the two endpoints in the family 2. The corresponding weights for the four endpoints were (0.6, 0.4, 0.6, 0.4) and (0.9,

0.1, 0.9, 0.1). The observed p-values were calculated using two-sided t-tests for the coefficient of the treatment, $\beta = 0$, in linear regressions. The adjusted p-values in Holm-based gatekeeping procedure were obtained using weighted Bonferroni method for family 1 and Weighted Holm method for family 2. The adjusted p-values in the proposed WMTCc-based gatekeeping procedure were obtained using the single-step WMTCc method for family 1, and the regular WMTCc method for family 2 where the estimated correlations from simulated data were used for both families. We replicated the clinical trial 1,000,000 times independently and calculated the family-wise type I error rate, defined as the number of clinical trials where at least one true null hypothesis was rejected, divided by 1,000,000. The result is shown in Table 1.1.

From these simulations, we can conclude the following:

1. Both Holm-based gatekeeping procedure and the proposed WMTCc-based gatekeeping procedure can control the family-wise type I error rate very well. The proposed WMTCc-based gatekeeping procedure keeps the family-wise type I error rate at 5 % level when the correlation (between endpoints) increases. However, the family-wise type I error rate in Holm-based gatekeeping procedure decreases, demonstrating decreased power when the correlation increases.
2. The proposed WMTCc-based gatekeeping procedure has higher power of rejecting at least one hypothesis among the four hypotheses in the two families compared with the Holm-based gatekeeping procedure, especially when the correlation between endpoints is high.
3. The proposed WMTCc-based gatekeeping procedure has a power advantage over the Holm-based gatekeeping procedure for each individual hypothesis in family 1, especially when the correlation between endpoints is high.
4. The proposed WMTCc-based gatekeeping procedure has an advantage over the Holm-based gatekeeping procedure for each individual hypothesis in family 2, especially when the correlation between endpoints are high.

1.4 Example

Following Dmitrienko and Tamhane (2011)'s example of the schizophrenia trial. Assume that the sample size per dose group (placebo, low dose and high dose) is 300 patients and the size of the classifier-positive subpopulation is 100 patients per dose group. Further assume that the t-statistics for testing the null hypotheses of no treatment effect in the general population and classifier-positive subpopulation are given by $t_1 = 2.04$, $t_2 = 2.46$, $t_3 = 2.22$ and $t_4 = 2.66$ with 897 d.f., 897 d.f., 297 d.f. and 297 d.f., respectively. We calculate two-sided p-values for the four null hypotheses computed from these t-statistics instead of one-sided p-values considered by Dmitrienko and Tamhane (2011). The p-values are $p_1 = 0.042$, $p_2 = 0.014$, $p_3 = 0.027$ and $p_4 = 0.008$. Dmitrienko and Tamhane (2011) considered un-weighted procedures, however, for illustration purposes only, we give different

Table 1.1 Simulated family-wise error rate and power of the WMTCc-based mixture gatekeeping procedure and the Holm-based mixture gatekeeping procedure

			Holm-based gatekeeping			WMTCc-based gatekeeping		
Weight	Effect size	ρ	Family 1 (Weighted Bonferroni)	Family 2 (Weighted Holm)	FWER or Power	Family 1 (Single-step WMTCc)	Family 2 (WMTCc)	FWER or Power
(0.6,0.4, 0.6,0.4)	(0,0,0,0)	0.0	3.0, 2.0	0.1, 0.1	5.0	3.1, 2.0	0.1, 0.1	5.0
		0.3	3.0, 2.0	0.1, 0.0	4.8	3.1, 2.1	0.1, 0.1	5.0
		0.5	3.0, 2.0	0.1, 0.1	4.7	3.2, 2.2	0.1, 0.1	5.0
		0.7	3.0, 2.0	0.1, 0.1	4.3	3.5, 2.3	0.1, 0.1	5.0
		0.9	3.0, 2.0	0.1, 0.1	3.7	4.1, 2.7	0.2, 0.1	5.0
	(0.4,0.1, 0.4,0.1)	0.0	81.7, 6.0	63.2, 3.5	82.8	81.8, 6.1	63.6, 6.2	82.9
		0.3	81.7, 6.0	62.7, 3.5	82.1	82.1, 6.2	63.3, 6.6	82.5
		0.5	81.8, 6.1	62.7, 3.5	82.0	82.6, 6.4	63.9, 6.8	82.8
		0.7	81.8, 6.0	62.6, 3.5	81.9	83.4, 6.8	65.2, 7.0	83.5
		0.9	81.8, 6.1	62.6, 3.5	81.9	85.0, 7.7	68.1, 7.2	85.2
	(0.1,0.4,0.1,0.4)	0.0	8.2, 77.3	3.5, 53.1	79.1	8.3, 77.4	4.7, 53.7	79.3
		0.3	8.2, 77.3	3.5, 52.5	78.1	8.4, 77.7	5.0, 53.2	78.5
		0.5	8.2, 77.3	3.5, 52.2	77.6	8.7, 78.1	5.2, 53.4	78.5
		0.7	8.2, 77.3	3.5, 52.2	77.5	9.2, 79.1	5.3, 54.7	79.3
		0.9	8.2, 77.3	3.5, 52.2	77.4	10.4, 80.8	5.5, 57.7	81.1
	(0.4,0.4, 0.4,0.4)	0.0	81.7, 77.3	75.8, 71.3	95.8	81.8 77.4	79.7, 78.9	95.9
		0.3	81.7, 77.3	74.1, 69.8	93.2	82.1, 77.7	77.7, 76.8	93.4
		0.5	81.8, 77.3	72.7, 68.5	91.1	82.6, 78.1	76.4, 75.4	91.6
		0.7	81.8, 77.3	71.1, 67.0	88.6	83.4, 79.1	75.3, 74.3	89.8
		0.9	81.8, 77.3	68.7, 64.8	84.9	85.0, 80.8	74.7, 73.6	87.8

(0.9,0.1, 0.9,0.1)	(0,0,0,0)	0.0	4.5, 0.5	0.2, 0.0	5.0	4.5, 0.5	0.2, 0.0	5.0
		0.3	4.5, 0.5	0.2, 0.0	5.0	4.6, 0.5	0.2, 0.0	5.0
		0.5	4.5, 0.5	0.2, 0.0	5.0	4.7, 0.5	0.2, 0.1	5.0
		0.7	4.5, 0.5	0.2, 0.0	5.0	4.8, 0.5	0.2, 0.1	5.0
		0.9	4.5, 0.5	0.2, 0.0	4.5	5.0, 0.5	0.2, 0.1	5.0
	(0.4,0.1, 0.4,0.1)	0.0	85.9, 2.1	73.0, 1.7	86.1	85.9, 2.1	73.1, 8.4	86.2
		0.3	85.9, 2.1	72.9, 1.6	85.9	86.0, 2.1	73.1, 9.0	86.0
		0.5	85.9, 2.1	72.9, 1.7	85.9	86.2, 2.1	73.4, 9.3	86.2
		0.7	85.9, 2.1	72.9, 1.6	85.9	86.5, 2.2	73.9, 9.4	86.5
		0.9	85.9, 2.1	72.9, 1.7	85.9	86.8, 2.2	74.5, 9.5	86.9
	(0.1,0.4, 0.1,0.4)	0.0	11.1, 60.2	2.2, 24.6	64.6	11.2, 60.3	2.3, 25.2	64.7
		0.3	11.1, 60.3	2.2, 23.9	62.8	11.2, 60.5	2.4, 24.4	63.0
		0.5	11.1, 60.2	2.2, 23.6	61.6	11.4, 60.6	2.4, 24.2	62.0
		0.7	11.1, 60.3	2.2, 23.3	60.9	11.6, 61.1	2.4, 24.1	61.7
		0.9	11.1, 60.2	2.2, 23.2	60.5	12.0, 61.5	2.5, 24.5	61.8
	(0.4,0.4, 0.4,0.4)	0.0	85.9, 60.2	78.4, 54.1	94.4	85.9, 60.3	79.0, 75.1	94.4
		0.3	85.8, 60.3	76.8, 53.2	91.7	86.0, 60.5	77.4, 73.1	91.8
		0.5	85.9, 60.2	75.7, 52.6	89.8	86.2, 60.6	76.4, 72.0	90.0
		0.7	85.9, 60.3	74.6, 52.0	87.8	86.5, 61.1	75.7, 71.4	88.4
		0.9	85.9, 60.2	73.6, 51.5	86.1	86.8, 61.5	75.2, 72.0	87.0

Table 1.2 Adjusted p-values produced by the WMTCc-based mixture gatekeeping procedure and the Holm-based mixture gatekeeping procedure in the schizophrenia trial example with parallel gatekeeping restrictions

Family	Null hypothesis	Weight	ρ	Raw p-value	Adjusted p-value	
					Holm-based	WMTCc-based
F_1	H_1	0.8	0.5	0.042	0.052	0.049
	H_2	0.2		0.014	0.070	0.066
F_2	H_3	0.8		0.027	–	0.033
	H_4	0.2		0.008	–	0.039

weights to different tests and use the weighted Holm-based mixture gatekeeping procedure and the proposed WMTCc-based mixture gatekeeping procedure. The results are given in Table 1.2. With FWER $= 0.05$, the weighted Holm-based mixture gatekeeping procedure does not reject any of the four hypotheses while the proposed WMTCc-based mixture gatekeeping procedure rejects the 1st, 3rd and 4th hypotheses.

1.5 Concluding Remarks and Discussions

In this paper, we proposed the WMTCc-based mixture gatekeeping procedure. Simulations have shown that the proposed WMTCc-based gatekeeping procedure using estimated correlation from the data can control the family-wise type I error rate very well as summarized in Table 1.1.

The proposed WMTCc-based gatekeeping procedure has a power advantage over the Holm-based gatekeeping procedure for each individuals hypothesis in the two families, especially when the correlation ρ is high.

In conclusion, our studies show that the proposed WMTCc-based mixture gatekeeping procedure based on Xie's weighted multiple testing correction for correlated tests outperforms the non-parametric methods in multiple testing in clinical trials.

References

Bauer P, Röhmel J, Maurer W, Hothorn L (1998) Testing strategies in multi-dose experiments including active control. Stat Med 17(18):2133–2146

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodolog) 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188

Dmitrienko A, Offen WW, Westfall PH (2003) Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Stat Med 22(15):2387–2400

Dmitrienko A, Tamhane AC (2011) Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. Stat Med 30(13):1473–1488

Dmitrienko A, Tamhane AC, Wiens BL (2008) General multistage gatekeeping procedures. Biom J 50(5):667–677

Dmitrienko A, Wiens BL, Tamhane AC, Wang X (2007) Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. Stat Med 26(12):2465–2478

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2009) mvtnorm: multivariate normal and t-distributions. R package version 0.9–8. http://CRAN.R-project.org/package=mvtnorm

Maurer W, Hothorn L, Lehmacher W (1995) Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. Biometrie in der chemisch-pharmazeutischen Industrie 6:3–18

Team RC (2013) A language and environment for statistical computing. R foundation for statistical computing, Vienna

Westfall PH, Krishen A (2001) Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. J Stat Plan Inference 99(1):25–40

Xie C (2012) Weighted multiple testing correction for correlated tests. Stat Med 31(4):341–352

Chapter 2
Regularization in Regime-Switching Gaussian Autoregressive Models

Abbas Khalili, Jiahua Chen, and David A. Stephens

Abstract Regime-switching Gaussian autoregressive models form an effective platform for analyzing financial and economic time series. They explain the heterogeneous behaviour in volatility over time and multi-modality of the conditional or marginal distributions. One important task is to infer the number of regimes and regime-specific parsimonious autoregressive models. Information-theoretic criteria such as AIC or BIC are commonly used for such inference, and they typically evaluate each regime/autoregressive combination separately in order to choose the optimal model accordingly. However, the number of combinations can be so large that such an approach is computationally infeasible. In this paper, we first use a computationally efficient regularization method for simultaneous autoregressive-order and parameter estimation when the number of autoregressive regimes is pre-detertermined. We then use a regularized Bayesian information criterion (RBIC) to select the most suitable number of regimes. Finite sample performance of the proposed methods are investigated via extensive simulations. We also analyze the U.S. gross domestic product growth and the unemployment rate data to demonstrate this method.

2.1 Introduction

A standard Gaussian autoregressive (AR) model of order q postulates that

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \cdots + \theta_q Y_{t-q} + \varepsilon_t \tag{2.1}$$

for a discrete-time series $\{Y_t; t = 1, 2, \ldots\}$, where $(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-q})$ and ε_t are independent and $\varepsilon_t \sim N(0, \sigma^2)$. Under this model, the conditional variance,

A. Khalili (✉) • D.A. Stephens
Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada
e-mail: khalili@math.mcgill.ca; dstephens@math.mcgill.ca

J. Chen
Big Data Research Institute of Yunnan University and Department of Statistics, University of British Columbia, Vancouver, BC, Canada
e-mail: jhchen@stat.ubc.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_2

or volatility, of the series is $\text{var}(Y_t|Y_{t-1}, \ldots, Y_{t-q}) = \sigma^2$, which is a constant with respect to time. In some financial and econometrics applications, the conditional variance of the time series clearly changes over time. ARCH (Engle 1982) and GARCH (Bollerslev 1986) models were subsequently motivated to accommodate the volatility changes. However, the time series may also exhibit heterogeneity in conditional mean or conditional (or marginal) distribution. Such non-standard behaviours call for more flexible models beyond (2.1), ARCH and GARCH.

Wong and Li (2000) introduced finite mixture of Gaussian autoregressive (MAR) models to accommodate the above non-standard behaviour. A MAR model combines K stationary or non-stationary Gaussian AR processes to capture heterogeneity while ensuring stationarity of the overall model. Due to the presence of several AR processes, a MAR model is also termed as the regime-switching Gaussian autoregressive model. MAR models generalize the Gaussian mixture transition distributions of Le et al. (1996) which were designed to model time series with non-Gaussian characteristics such as flat stretches, bursts of activity, outliers, and change points. Wong and Li (2000) used the expectation-maximization (EM) algorithm of Dempster et al. (1977) for maximum likelihood parameter estimation in MAR models when the number of regimes is pre-determined. They examined the performance of information-theoretic criteria such AIC and BIC for selection of the number of AR regimes and the variable/model selections within each regime through simulations.

A parsimonious AR model can be obtained by setting some of $\{\theta_1, \theta_2, \ldots, \theta_q\}$ in (2.1) zero. Such variable selection is known to lead to more effective subsequent statistical inferences. Information-theoretic criteria such AIC or BIC are widely used to choose the best subset of $\{\theta_1, \theta_2, \ldots, \theta_q\}$. They typically evaluate 2^q possible AR submodels in an exhaustive calculation. When q is large, this is a formidable computational challenge. The straightforward application of AIC, BIC or other information criteria to MAR model selection poses even a greater computational challenge. To overcome the computational obstacle, Jirak (2012) proposed simultaneous confidence intervals for parameter and order estimation; Wang et al. (2007) and Nardi and Rinaldo (2011) used the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996).

Regularization techniques such as the LASSO by Tibshirani (1996), the smoothly clipped absolute deviation (SCAD) by Fan and Li (2001), the adaptive LASSO by Zou (2006) have been successfully applied in many situations. In this paper we first present a regularized likelihood approach for simultaneous AR-order and parameter estimation in MAR models when the number of AR regimes is predetermined. The new approach is computationally very efficient compared to existing methods. Extensive simulations show that the method performs well in a wide range of finite sample situations. In some applications, the data analysts must also decide on the best number of AR regimes (K) for a data set. We propose to use a regularized BIC (RBIC) for choosing K. Our simulations show that the RBIC performs well in various situations.

The rest of the paper is organized as follows. In Sect. 2.2, the MAR model and the problems of model selection and parameter estimation are introduced. In Sect. 2.3, we develop new methods for the problems of interest. Our simulation study is given in Sect. 2.4. We analyze the U.S. gross domestic product (GDP) growth and U.S. unemployment rate data in Sect. 2.5. Finally, Sect. 2.6 contains some discussion and conclusions.

2.2 Terminology and Model

Consider an observable time series $\{Y_t : t = 1, 2, \ldots\}$ with corresponding realized values $\{y_t : t = 1, 2, \ldots\}$, and a latent stochastic process $\{S_t : t = 1, 2, \ldots\}$ taking values in $\{1, 2, \ldots, K\}$ with K being the number of regimes underlying the time series. In a mixture of Gaussian autoregressive (MAR) model, $\{S_t : t = 1, 2, \ldots\}$ is an iid process, and the conditional distribution of $Y_t|(S_t = k, y_{t-1}, \ldots, y_{t-q})$ is presumed normal with variance σ_k^2 and mean

$$\mu_{t,k} = \theta_{k0} + \theta_{k1} y_{t-1} + \ldots + \theta_{kq} y_{t-q} \; ; \; k = 1, 2, \ldots, K. \tag{2.2}$$

Here q is the maximum order that is thought to be reasonable across all K AR regimes. Let $\boldsymbol{\Phi}_K = (\pi_1, \pi_2, \ldots, \pi_K, \sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K)$ denote the vector of all parameters, where $\boldsymbol{\theta}_k = (\theta_{k0}, \theta_{k1}, \ldots, \theta_{kq})^\top$ is the coefficient vector of the kth AR regime. As in the usual finite mixture formulation, in a MAR model the conditional distribution of $Y_t|(y_{t-1}, \ldots, y_{t-q})$ is a Gaussian mixture with density

$$f(y_t|y_{t-1}, \ldots, y_{t-q}, \boldsymbol{\Phi}_K) = \sum_{k=1}^{K} \pi_k \, \phi(y_t; \mu_{t,k}, \sigma_k^2) \tag{2.3}$$

where $\Pr[S_t = k] = \pi_k \in (0, 1)$ are mixing proportions that sum to one, and $\phi(\cdot; \mu_{t,k}, \sigma_k^2)$ is the density function of $N(\mu_{t,k}, \sigma_k^2)$.

We assume that in the true MAR model underlying the data some elements of the vectors $\boldsymbol{\theta}_k$ (except the intercepts θ_{k0}) are zero, which is referred to as a MAR submodel as formally defined below.

Subset-AR models – where, in formulation (2.1), parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_q)^\top$ contains zeros – are often used in time series literature (Jirak 2012). For each subset $\Im \subseteq \{1, 2, \ldots, q\}$, we denote its cardinality by $|\Im|$, introduce column vector $\boldsymbol{y}_{t-\Im} = \{1, y_{t-j} : j \in \Im\}^\top$ and coefficient sub-vector $\boldsymbol{\theta}[\Im] = \{\theta_0, \theta_j, j \in \Im\}^\top$. We denote it as $\boldsymbol{\theta}_k[\Im_k]$ when applied to the kth regime. The regime-specific conditional mean is then $(\boldsymbol{\theta}_k[\Im_k])^\top \boldsymbol{y}_{t-\Im}$. Each combination of $\Im_1, \Im_2, \ldots \Im_K$ specifies a MAR submodel with the conditional density function

$$f_{[\Im_1, \Im_2, \ldots, \Im_K]}(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-q}, \boldsymbol{\Phi}_K) = \sum_{k=1}^{K} \pi_k \, \phi(y_t; \mu_{t,k}(\Im_k), \sigma_k^2) \tag{2.4}$$

where

$$\mu_{t,k}(\Im_k) = (\boldsymbol{\theta}_k[\Im_k])^\top \boldsymbol{y}_{t-\Im_k} = \theta_{k0} + \sum_{j\in\Im_k} \theta_{kj} y_{t-j}.$$

Let $(Y_1, Y_2, \ldots, Y_n) \equiv Y_{1:n}$ be a random sample from a MAR model (2.3), with a joint density function that may be factorized as $f_1(y_1, y_2, \ldots, y_q) f_2(y_{q+1}, \ldots, y_n | y_1, y_2, \ldots, y_q)$. As a standard approach in time series, we work with conditional density $f_2(\cdot)$, and the (conditional) likelihood function in a MAR model is given by

$$\begin{aligned} l_n(\boldsymbol{\Phi}_K) &= \log\{f_2(y_{q+1}, \ldots, y_n | y_1, y_2, \ldots, y_q)\} \\ &= \sum_{t=q+1}^{n} \log\{f(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-q}, \boldsymbol{\Phi}_K)\} \\ &= \sum_{t=q+1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k \, \phi(y_t; \mu_{t,k}, \sigma_k^2)\right\}. \end{aligned} \tag{2.5}$$

In principle, once K is selected, we could carry out maximum (conditional) likelihood estimation of $\boldsymbol{\Phi}_K$ by maximizing $l_n(\boldsymbol{\Phi}_K)$. However, since all of the estimated AR coefficients would be non-zero, such an approach does not provide a MAR sub-model as postulated. Instead, we may use the information-theoretic approaches such as AIC and BIC, based on (2.5), to select a MAR submodel (2.4) out of $2^{K\times q}$ possible candidates that best balances the model parsimony and goodness of fit of the data. The K itself may be chosen over a set of possible values $\{1, 2, \ldots, \mathscr{K}\}$ for an upper bound $\mathscr{K}$ specified empirically. The difficulty with this strategy, however, is that the total number of MAR submodels is given by $\sum_{K=1}^{\mathscr{K}} 2^{K\times q}$. If AIC and BIC are used, one would have to compute the criterion for each separate model. The computational cost will explode even for moderate sized $\mathscr{K}$ and q. This observation motivates us to investigate the regularization methods in later sections.

2.3 Regularization in MAR Models

2.3.1 *Simultaneous AR-Order and Parameter Estimation when K is Known*

In the following sections, we investigate regularization of the conditional likelihood (2.5), and study effective optimization procedures.

A penalty on the mixture component variances: Similar to conventional Gaussian mixture models with unequal component variances σ_k^2's, the conditional log-likelihood $l_n(\boldsymbol{\Phi}_K)$ in (2.5) diverges to positive infinity when some component

variance σ_k^2 goes to 0. This can be avoided by introducing a penalty function as in Hathaway (1985) and Chen et al. (2008):

$$\tilde{l}_n(\boldsymbol{\Phi}_K) = l_n(\boldsymbol{\Phi}_K) - \sum_{k=1}^{K} p_n(\sigma_k^2) \tag{2.6}$$

where $p_n(\sigma_k^2)$ is a smooth penalty function of σ_k^2, such that $p_n(\sigma_k^2) \to +\infty$, as $\sigma_k \to 0$ or $+\infty$. We refer to (2.6) as the adjusted conditional log-likelihood. Specifically, we follow Chen et al. (2008) and specify

$$p_n(\sigma_k^2) = \frac{1}{\sqrt{n}}\left[\frac{V_n^2}{\sigma_k^2} + \log\left(\frac{\sigma_k^2}{V_n^2}\right)\right] \tag{2.7}$$

where $V_n^2 = (n-q)^{-1}\sum_{t=q+1}^{n}(y_t - \bar{y})^2$ is the sample variance of the observed series, and $\bar{y} = (n-q)^{-1}\sum_{t=q+1}^{n} y_t$. From a Bayesian point of view, the use of penalty corresponds to a data-dependent Gamma prior on $1/\sigma_k^2$ with its mode at $1/V_n^2$. In what follows, we will work with $\tilde{l}_n(\boldsymbol{\Phi}_K)$.

AR-order selection and parameter estimation via regularization: If we directly maximize the adjusted conditional log-likelihood $\tilde{l}_n(\boldsymbol{\Phi}_K)$, the estimates of some of the AR coefficients θ_{kj} may be close but not equal to zero. The resulting full model will not be as parsimonious as required in applications. We achieve model selection by maximizing the regularized (or penalized) conditional log-likelihood

$$pl_n(\boldsymbol{\Phi}_K) = \tilde{l}_n(\boldsymbol{\Phi}_K) - \boldsymbol{r}_n(\boldsymbol{\Phi}_K) \tag{2.8}$$

with some regularization function

$$\boldsymbol{r}_n(\boldsymbol{\Phi}_K) = \sum_{k=1}^{K} \pi_k \left\{\sum_{j=1}^{q} r_n(\theta_{kj}; \lambda_{nk})\right\} \tag{2.9}$$

for pre-specified pair K and q. The penalty function $r_n(\theta;\lambda)$ will be chosen positive, continuous in θ and having a spike at $\theta = 0$; $\lambda \geq 0$ is a tuning parameter controlling the severity of the penalty. When $r_n(\theta;\lambda)$ is appropriately chosen, maximizing (2.8) will lead to some θ_{kj}s having fitted values exactly zero. Furthermore, increasing the size of λ_{nk} generally forces more of fitted values of θ_{kj}s to be zero. Consequently, such a procedure leads to a method that performs simultaneous AR-order and parameter estimation. This procedure does not evaluate every possible MAR submodel and thereby is computationally feasible.

Example of penalties: Forms of $r_n(\theta;\lambda)$ with the desired properties are the LASSO penalty

$$r_n(\theta;\lambda) = n\lambda|\theta|,$$

and the SCAD penalty which is most often characterized by its first derivative:

$$r'_n(\theta;\lambda) = n\lambda\left\{I(|\theta| \le \lambda) + \frac{(a\lambda - |\theta|)_+}{\lambda(a-1)}\, I(|\theta| > \lambda)\right\} \times \mathrm{sgn}(\theta)$$

for some constant $a > 2$; where $I(\cdot)$ and $\mathrm{sgn}(\cdot)$ are the indicator and sign functions, and $(\cdot)_+$ is the positive part of the input, respectively. Fan and Li (2001) showed that the value $a = 3.7$ minimizes a Bayes risk criterion for θ. This choice of a has since become the standard in various model selection problems. We used this value in our simulations and data analysis.

For the MAR model (2.3) where the S_t are iid samples, the theoretical proportion of $S_t = k$ is given by π_k. Thus, we choose the penalties in (2.9) to be proportional to the mixing probabilities π_k to control the level of regime-specific penalty on θ_{kj}s. This improves the finite sample performance of the method, and the influence vanishes as the sample size n increases.

2.3.2 *Numerical Computations*

The maximization of the penalized conditional log-likelihood for a K regime MAR model with maximal AR order q is an optimization over a space of dimension $(K-1) + K(q+2)$; for example, with $K = 5$ and $q = 10$, the number of parameters is 64; this number is large, but direct optimization using Nelder-Mead or quasi-Newton methods (via `optim` in `R`) is still possible when a local quadratic approximation to the penalty is adopted: following Fan and Li (2001), the approximation

$$r_n(\theta_{kj};\lambda_{nk}) \simeq r_n(\theta_0;\lambda_{nk}) + \frac{r'_n(\theta_0;\lambda_{nk})}{2\theta_0}(\theta_{kj}^2 - \theta_0^2) \tag{2.10}$$

holds in a neighbourhood of a current value θ_0, and may be used. Coordinate-based methods operating on the incomplete data likelihood may also be useful.

In this paper, however, we use a modified EM algorithm for maximization of the penalized log-likelihood $pl_n(\boldsymbol{\Phi}_K)$ in (2.8). Let $\hat{\boldsymbol{\Phi}}_{n,K} = \arg\max\{pl_n(\boldsymbol{\Phi}_K)\}$ denote the maximum penalized conditional likelihood estimator (MPCLE) of $\boldsymbol{\Phi}_K$. By tuning the level of penalty λ_{nk}, this estimator has its $\hat{\boldsymbol{\theta}}_k$ components containing various number of zero-fitted values and parsimony is induced.

2.3.2.1 EM-Algorithm

For observation y_t, let Z_{tk}, with realization z_{tk}, equal 1 if $S_t = k$, and equal 0 otherwise. The complete conditional adjusted log-likelihood function under the MAR model is given by

$$\tilde{l}_n^c(\boldsymbol{\Phi}_K) = \sum_{k=1}^{K} \sum_{t=q+1}^{n} z_{tk} \Big\{ \log \pi_k + \log \phi(y_t; \mu_{t,k}, \sigma_k^2) \Big\} - \sum_{k=1}^{K} p_n(\sigma_k^2)$$

and thus the penalized complete conditional log-likelihood is $pl_n^c(\boldsymbol{\Phi}_K) = \tilde{l}_n^c(\boldsymbol{\Phi}_K) - \boldsymbol{r}_n(\boldsymbol{\Phi}_K)$.

Let $\boldsymbol{x}_t^\top = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-q})$, $\boldsymbol{X} = (\boldsymbol{x}_{q+1}, \boldsymbol{x}_{q+2}, \ldots, \boldsymbol{x}_n)^\top$, $\boldsymbol{y} = (y_{q+1}, y_{q+2}, \ldots, y_n)^\top$. Given the current value of the parameter vector $\boldsymbol{\Phi}_K^{(m)}$, the EM algorithm iterates as follows:

E-step: We compute the expectation of the latent Z_{tk} variables conditional on the other parameters and the data. At $(m+1)$-th iteration, the EM objective function is

$$Q(\boldsymbol{\Phi}_K; \boldsymbol{\Phi}_K^{(m)}) = \sum_{k=1}^{K} \sum_{t=q+1}^{n} \omega_{tk}^{(m)} \Big\{ \log \pi_k + \log \phi(y_t; \mu_{t,k}, \sigma_k^2) \Big\}$$
$$- \sum_{k=1}^{K} p_n(\sigma_k^2) - \sum_{k=1}^{K} \pi_k \sum_{j=1}^{q} r_n(\theta_{kj}; \lambda_{nk})$$

with weights

$$\omega_{tk}^{(m)} = E(Z_{tk} | \mathbf{y}; \boldsymbol{\Phi}_K^{(m)}) = \frac{\pi_k^{(m)} \phi(y_t; \mu_{t,k}^{(m)}, \sigma_k^{2(m)})}{\sum_{l=1}^{K} \pi_l^{(m)} \phi(y_t; \mu_{t,l}^{(m)}, \sigma_l^{2(m)})}.$$

M-step: By using the penalty $p_n(\sigma_k^2)$ in (2.7) and the quadratic approximation (2.10) for $r_n(\theta_{kj}; \lambda_{nk})$, we maximize the (approximated) $Q(\boldsymbol{\Phi}_K; \boldsymbol{\Phi}_K^{(m)})$ with respect to $\boldsymbol{\Phi}_K$. Note that the quadratic approximation to the penalty provides a minorization of the exact objective function, ensuring that the iterative algorithm still converges to the true maximum of the penalized likelihood function.

The EM updates of the regime-specific AR-coefficients and variances, for $k = 1, \ldots, K$, are given by

$$\boldsymbol{\theta}_k^{(m+1)} = (\boldsymbol{X}^\top \boldsymbol{W}_k^{(m)} \boldsymbol{X} + \boldsymbol{\Sigma}_k^{(m)})^{-1} \boldsymbol{X}^\top \boldsymbol{W}_k^{(m)} \boldsymbol{Y}$$

$$\sigma_k^{2(m+1)} = \frac{\sum_{t=q+1}^{n} \omega_{tk}^{(m)} (y_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_k^{(m+1)})^2 + 2V_n^2/\sqrt{n}}{\sum_{t=q+1}^{n} \omega_{tk}^{(m)} + 2/\sqrt{n}}$$

with diagonal matrices $\boldsymbol{W}_k^{(m)} = \text{diag}\{\omega_{tk}^{(m)}; t = q+1, \ldots, n\}$ and $\boldsymbol{\Sigma}_k^{(m)} = \text{diag}\{0, \pi_k^{(m)}\sigma_k^{2(m)} r_n'(\theta_{kj}^{(m)}; \lambda_{nk})/\theta_{kj}^{(m)}, j = 1, \ldots, q\}$.

For the mixing probabilities, the updates are

$$\pi_k^{(m+1)} = \frac{1}{n-q} \sum_{t=q+1}^{n} \omega_{tk}^{(m)} \quad , \quad k = 1, 2, \ldots, K$$

that maximize the leading term in $Q(\boldsymbol{\Phi}_K; \boldsymbol{\Phi}_K^{(m)})$, and it worked well in our simulation study.

Starting from an initial value $\boldsymbol{\Phi}_K^{(0)}$, the EM algorithm continues until some convergence criterion is met. We used the stopping rule $\|\boldsymbol{\Phi}_K^{(m+1)} - \boldsymbol{\Phi}_K^{(m)}\| \leq \epsilon$, for a pre-specified small value ε, taken 10^{-5} in our simulations and data analysis. Due to the sparse selection penalty $r_n(\theta_{kj}; \lambda_{nk})$ some of the estimates $\hat{\theta}_{kj}$ will be very close to zero at convergence; these estimates are set to zero. Thus we achieve simultaneous AR-order and parameter estimation.

2.3.2.2 Tuning of λ in $r_n(\boldsymbol{\theta}, \lambda)$

One remaining issue in the implementation of the regularization method is the choice of tuning parameter λ for each regime. We recommend a regime-specific information criterion together with a grid search scheme as follows.

We first fit the full MAR model by finding the maximum point of $\tilde{l}_n(\boldsymbol{\Phi}_K)$ defined by (2.6) using the same EM algorithm above. Once the maximum point of $\tilde{l}_n(\boldsymbol{\Phi}_K)$, $\widetilde{\boldsymbol{\Phi}}_K$, is obtained, we compute

$$\tilde{\omega}_{tk} = \frac{\tilde{\pi}_k \, \phi(y_t; \tilde{\mu}_{t,k}, \tilde{\sigma}_k^2)}{\sum_{l=1}^{K} \tilde{\pi}_l \, \phi(y_t; \tilde{\mu}_{t,l}, \tilde{\sigma}_l^2)}.$$

This is the fitted probability of $S_t = k$, conditional on $\boldsymbol{y}$, and based on the fitted full model.

Next, we pre-choose a grid of λ-values $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ for some M, say $M = 10$ or 15. For each λ_i and regime k, we define a regime specific regularized likelihood function

$$\tilde{l}_k(\boldsymbol{\theta}, \sigma^2) = \sum_{t=q+1}^{n} \tilde{\omega}_{tk} \log \phi(y_t; \mu_t, \sigma^2) - p_n(\sigma^2) - \tilde{\pi}_k \sum_{j=1}^{q} r_n(\theta_j; \lambda_i)$$

with $\mu_t = \boldsymbol{x}_t^\top \boldsymbol{\theta}$. We then search for its maximum point with respect to $\boldsymbol{\theta}$ and σ^2, $\overline{\boldsymbol{\theta}}_k(\lambda_i)$ and $\overline{\sigma}_k^2$; compute $\bar{\mu}_{t,k,i} = \boldsymbol{x}_t^\top \, \overline{\boldsymbol{\theta}}_k(\lambda_i)$, and the residual sum of squares (RSS)

$$\text{RSS}_k(\lambda_i) = \sum_{t=q+1}^{n} \tilde{\omega}_{tk} \, (y_t - \bar{\mu}_{t,k,i})^2.$$

The weights $\tilde{\omega}_{tk}$ are included because observations y_t may not be from regime k. The regime-specific information criterion is computed as

$$\text{IC}(\lambda_i; k) = n_k \log[\text{RSS}_k(\lambda_i)] + \text{DF}(\lambda_i)(\log n_k) \tag{2.11}$$

where $\text{DF}(\lambda_i) = \sum_{j=1}^{q} I(\overline{\theta}_{kj} \neq 0)$ and $n_k = \sum_{t=q+1}^{n} \tilde{\omega}_{tk}$. This information criterion mimics the one used in linear regression by Zhang et al. (2010). We choose the value of the tuning parameter for regime k as

$$\tilde{\lambda}_k = \underset{1 \leq i \leq M}{\text{argmin}} \ \text{IC}(\lambda_i; k).$$

2.3.3 Choice of the Mixture-Order or Number of AR Regimes K

The procedure presented in the last subsection is used when the number of AR regimes K is pre-specified. However, a data-adaptive choice of K is needed in most applications. We now propose a regularized BIC (RBIC) for choosing K.

Consider the situation where placing an upper bound $\mathscr{K}$ on K is possible. For each $K \in \{1, 2, \ldots, \mathscr{K}\}$, we fit a MAR model as above with resulting estimates denoted by $\hat{\boldsymbol{\Phi}}_{n,K}$. Let $\mathscr{N}_K = \sum_{k=1}^{K} \sum_{j=1}^{q} I(\hat{\theta}_{kj} \neq 0)$ be the total number of non-zero $\hat{\theta}_{kj}$s, and

$$\text{RBIC}(\hat{\boldsymbol{\Phi}}_{n,K}) = l_n(\hat{\boldsymbol{\Phi}}_{n,K}) - 0.5(\mathscr{N}_K + 3K - 1) \times \log(n - q) \tag{2.12}$$

where $3K - 1$ counts the number of parameters $(\pi_k, \sigma_k^2, \theta_{k0})$, $l_n(\cdot)$ is the conditional log-likelihood given in (2.5). We then select the estimated number of AR regimes $\hat{K}_n$ as

$$\hat{K}_n = \arg \max_{1 \leq K \leq \mathscr{K}} \text{RBIC}(\hat{\boldsymbol{\Phi}}_{n,K}). \tag{2.13}$$

We now have an estimated MAR model with $\hat{K}_n$ AR regimes, and the regime-specific AR model characterized by the corresponding $\hat{\boldsymbol{\theta}}_k$, $k = 1, 2, \ldots, \hat{K}_n$.

Extensive simulation studies show that the RBIC performs well in various situations. It is noteworthy that, for each K, RBIC is computed based on the outcome $\hat{\boldsymbol{\Phi}}_{n,K}$ from the regularization method outlined in (2.8) and (2.9), which is obtained after examining a grid of, say, $M = 10$ or 15, possible values for the tuning parameters λ_{nk}. In comparison, the standard BIC* adopted in Wong and Li (2000) must examine $2^{K \times q}$ possible MAR submodels. The RBIC thus offers a substantial computational advantage unless $\mathscr{K}$ and q are both very small.

2.4 Simulations

In this section we study the performance of the proposed regularization method for AR-order and parameter estimation, and the RBIC for selection of the number of AR regimes (mixture-order) via simulations. We generated times series data from five MAR models. The parameter settings for the first four models are:

Model	(K^*, q)	(π_1, π_2)	(σ_1, σ_2)	$\mu_{t,1}$	$\mu_{t,2}$
1	$(2, 5)$	$(.75, .25)$	$(5, 1)$	$.50y_{t-1}$	$1.3y_{t-1}$
2	$(2, 5)$	$(.75, .25)$	$(5, 1)$	$.70y_{t-1} - .65y_{t-2}$	$.45y_{t-1} - 1.2y_{t-3}$
3	$(2, 6)$	$(.75, .25)$	$(5, 1)$	$.67y_{t-1} - .55y_{t-2}$	$.45y_{t-1} + .35y_{t-3} - .65y_{t-6}$
4	$(2, 15)$	$(.65, .35)$	$(3, 1)$	$.58y_{t-1} - .45y_{t-6}$	$.56y_{t-1} - .40y_{t-3} + .44y_{t-12}$

Note that q in the above table is the pre-specified maximum AR-order, and K^* is the true number of AR-regimes in a MAR model. By Theorem 1 of Wong and Li (2000), in a MAR model, a necessary and sufficient condition for a MAR to be first-order stationary is that all roots of $1 - \sum_{j=1}^{q} \sum_{k=1}^{K} \pi_k \theta_{kj} z^{-j} = 0$ are inside the unit circle. The parameter values π_k and θ_{kj} in Models 1–4 are chosen to ensure, at least, this condition is satisfied.

The fifth MAR model under our consideration is a three-regime MAR with parameter values:

Model	(K^*, q)	(π_1, π_2, π_3)	$(\sigma_1, \sigma_2, \sigma_3)$	$\mu_{t,1}$	$\mu_{t,2}$	$\mu_{t,3}$
5	$(3, 5)$	$(.4, .3, .3)$	$(1, 1, 5)$	$.9y_{t-1} - .6y_{t-2}$	$-.5y_{t-1}$	$1.5y_{t-1} - .75y_{t-2}$

The maximum AR-order specified in the regularization method for this model is also $q = 5$, and the values of π_k and θ_{kj} are chosen such that the MAR model is, at least, first-order stationary as defined above.

2.4.1 Simulation when K is Specified*

In this section we assess the performance of the the regularization method for AR-order and parameter estimation when the number of AR regimes (K^*) of the MAR is specified. We use the EM-algorithm outlined in Sect. 2.3.2 to maximize the regularized likelihood defined in (2.8). The regime-specific tuning parameters λ_{nk} in $r_n(\theta; \lambda_{nk})$ are chosen by the criterion IC in (2.11). The computations are done in C++ and on a Mac OS X machine with 2×2.26 GHz Quad-Core Intel Xeon processor.

Table 2.1 Correct (Cor) and incorrect (Incor) number of estimated zero θ_{kj}'s in Models 1, 2 and 3. The numbers inside $[\cdot, \cdot]$ are the true Cor in regimes Reg_1 and Reg_2 of each model

		MAR	Model 1		Model 2		Model 3	
Method	n	Regimes	Cor[4,4]	Incor	Cor[3,3]	Incor	Cor[4,3]	Incor
BIC*	150	Reg_1	3.83	.024	2.92	.001	3.87	.014
		Reg_2	3.56	.025	2.81	.005	2.76	.023
	250	Reg_1	3.92	.000	2.95	.000	3.93	.001
		Reg_2	3.88	.000	2.90	.000	2.92	.002
	400	Reg_1	3.95	.000	2.96	.000	3.93	.000
		Reg_2	3.93	.000	2.95	.000	2.95	.000
SCAD	150	Reg_1	3.93	.022	2.97	.000	3.89	.013
		Reg_2	3.85	.002	2.91	.004	2.88	.013
	250	Reg_1	3.99	.001	3.00	.000	3.98	.003
		Reg_2	3.98	.000	2.99	.000	2.98	.003
	400	Reg_1	4.00	.000	3.00	.000	4.00	.000
		Reg_2	4.00	.000	3.00	.000	3.00	.000

The simulation results are based on 1000 randomly generated time series of a given size from each of the five models, and they are reported in the form of regime-specific: correct (Cor) and incorrect (Incor) number of estimated zero AR-coefficients θ_{kj}, and the empirical mean squared errors (MSE_k) of the estimators $\hat{\boldsymbol{\theta}}_k$ of the vector of AR-coefficients $\boldsymbol{\theta}_k$. The Reg_k in the tables represent AR regimes of each MAR model.

For Models 1–3, it is computationally feasible to implement the standard BIC* of Wong and Li (2000). Therefore, we also reported the results based on BIC* together with their computational costs. For Models 4 and 5, the amount of computation of BIC* is infeasible. Thus, BIC* is not included in our simulation.

Tables 2.1 and 2.2 contain the simulation results based on the SCAD regularization method and standard BIC* for Models 1, 2 and 3. From Table 2.1, the regularization method clearly outperforms BIC* by having higher rates of correctly (Cor) estimated zero AR-coefficients and lower rates of incorrectly (Incor) estimated zero AR-coefficients, in both regimes Reg_1 and Reg_2 of the three MAR models. Both methods improve as the sample size increases. Table 2.2 provides the regime-specific empirical mean square errors (MSE_k) of the estimators $\hat{\boldsymbol{\theta}}_k$. For $n = 150$, SCAD outperforms BIC* in all three models, especially with respect to the MSE_2. When the sample size increases, the two methods have similar performances. Regarding the computational time, the regularization method took about 6–16 s for $n = 150$ and 400, respectively, to complete 1000 replications for each of the three models. The BIC* took about 2.78 and 7.83 h for Models 1 and 2, and it took 22.9 and 143.8 h for Model 3 when $n = 150$ and 400, respectively.

Table 2.2 Regime-specific empirical mean squared errors (MSE) in Models 1, 2 and 3

Method	Sample size	Model 1		Model 2		Model 3	
	n	MSE_1	MSE_2	MSE_1	MSE_2	MSE_1	MSE_2
BIC*	150	.027	.121	.013	.014	.034	.029
	250	.008	.002	.006	.002	.015	.004
	400	.004	.001	.004	.001	.008	.002
SCAD	150	.024	.010	.014	.011	.035	.015
	250	.010	.001	.005	.002	.015	.005
	400	.006	.001	.003	.001	.007	.002

Table 2.3 Correct (Cor) and incorrect (Incor) number of estimated zero θ_{kj}'s, and regime-specific empirical mean squares errors (MSE) in Model 4. The numbers inside $[\cdot, \cdot]$ are the true Cor in regimes Reg_1 and Reg_2 of the model

		MAR	Cor[13,12]		MSE	
Method	n	Regimes	Reg_1	Reg_2	MSE_1	MSE_2
SCAD	250	Cor	12.8	11.3	.066	.089
		Incor	.027	.100		
	400	Cor	12.9	11.9	.014	.031
		Incor	.002	.058		
	600	Cor	13.0	11.9	.007	.016
		Incor	.000	.037		
	800	Cor	13.0	12.0	.005	.008
		Incor	.000	.014		
	1000	Cor	13.0	12.00	.004	.003
		Incor	.000	.000		
LASSO	250	Cor	12.8	11.3	.056	.076
		Incor	.019	.096		
	400	Cor	12.9	11.8	.021	.027
		Incor	.004	.045		
	600	Cor	12.9	11.9	.013	.014
		Incor	.000	.021		
	800	Cor	13.0	12.0	.011	.008
		Incor	.000	.004		
	1000	Cor	13.0	12.0	.009	.005
		Incor	.000	.000		

Table 2.3 contains the simulation results for the MAR Model 4 which has higher AR-orders. Overall, the new method based on either SCAD or LASSO performed very well. It took SCAD about 118 and 787 s, corresponding to $n = 250$ and 1000, respectively, to complete the 1000 replications. The LASSO had similar computational times. The BIC* is computationally infeasible.

Table 2.4 Correct (Cor) and incorrect (Incor) number of estimated zero θ_{kj}'s, and regime-specific empirical mean squares errors (MSE) in Model 5. The numbers inside $[\cdot, \cdot, \cdot]$ are the true Cor in regimes Reg_1, Reg_2 and Reg_3 of the model

		MAR	Cor[3,4,3]			MSE		
Method	n	Regimes	Reg_1	Reg_2	Reg_3	MSE_1	MSE_2	MSE_3
SCAD	150	Cor	2.97	3.96	2.70	.006	.002	.285
		Incor	.003	.000	.090			
	250	Cor	2.99	4.00	2.89	.001	.001	.067
		Incor	.000	.000	.025			
	400	Cor	3.00	4.00	2.97	.001	.000	.019
		Incor	.000	.000	.003			
LASSO	150	Cor	2.96	3.95	2.51	.008	.002	.368
		Incor	.004	.000	.165			
	250	Cor	2.99	4.00	2.67	.002	.000	.229
		Incor	.000	.000	.077			
	400	Cor	3.00	4.00	2.80	.002	.000	.145
		Incor	.002	.000	.023			

Model 5 has $K^* = 3$, and its simulation results are in Table 2.4. The regularization method performs reasonably well in both AR-order selection and parameter estimation. Comparatively to regimes 1 and 2, the method has lower rates of correctly estimated zero AR-coefficients, higher rates of incorrectly estimated zero AR-coefficients and also larger mean square errors for regime 3. This is more evident for LASSO. This is expected because the noise level ($\sigma_3 = 5$) in Reg_3 is much higher. Consequently, it is harder to maintain the same level of accuracy for AR-order selection and parameter estimation. When the sample size increases, the regularization method has improved precision, either when SCAD or LASSO is used. The regularization method took about 13 and 34 s, for $n = 150$ and 400, respectively, to complete the simulations.

2.4.2 Selection of K

In this section we examine the performance of the estimator $\hat{K}_n$ in (2.13). We report the observed distribution of $\hat{K}_n$ based on 1000 replications. The results for Models 1–3 and also Model 5 are reported in Table 2.5. Model 4 has more complex regime structures. Thus, it is more closely examined with additional sample sizes and the results are singled out in Table 2.6. For each model, $\hat{K}_n$ is calculated based on the RBIC. Note that if we replace the factor $\log(n-q)$ in (2.12) by number 2, we create an AIC motivated RAIC selection method. We also obtained the simulation results based on RAIC to serve as a potential yardstick. We present the results corresponding

Table 2.5 Simulated distribution of the mixture-order estimator $\hat{K}_n$. Results for the true order K^* are in **bold**. Values in [·] are the proportion of concurrently correct estimation of the regime-specific AR-orders

	Models:	1 ($K^* = 2$)		2 ($K^* = 2$)		3 ($K^* = 2$)		5 ($K^* = 3$)	
n	K	RBIC	RAIC	RBIC	RAIC	RBIC	RAIC	RBIC	RAIC
150	1	.131	.000	.031	.004	.023	.000	.000	.000
	2	**.855**$_{[.838]}$	**.207**	**.965**$_{[.962]}$	**.468**	**.957**$_{[.936]}$	**.215**	.016	.000
	3	.014	.150	.004	.177	.019	.169	**.945**$_{[.938]}$	**.589**
	4 or 5	.000	.643	.000	.351	.001	.616	.039	.411
250	1	.008	.000	.002	.002	.000	.000	.002	.002
	2	**.978**$_{[.973]}$	**.312**	**.995**$_{[.995]}$	**.572**	**.996**$_{[.994]}$	**.217**	.000	.000
	3	.014	.190	.003	.171	.004	.185	**.986**$_{[.985]}$	**.768**
	4 or 5	.000	.498	.000	.255	.000	.598	.012	.230
400	1	.002	.000	.004	.004	.000	.000	.001	.001
	2	**.998**$_{[.998]}$	**.327**	**.996**$_{[.996]}$	**.632**	**.999**$_{[.999]}$	**.220**	.001	.001
	3	.000	.194	.000	.158	.001	.229	**.996**$_{[.996]}$	**.836**
	4 or 5	.000	.479	.000	.206	.000	.551	.002	.162

Table 2.6 Simulated distribution of the mixture-order estimator $\hat{K}_n$. Results for the true order K^* are in **bold**. Values in [·] are the proportion of concurrently correct estimation of the regime-specific AR-orders. Model 4 ($K^* = 2$)

	$n = 250$		$n = 400$		$n = 600$		$n = 800$		$n = 1000$	
K	RBIC	RAIC	RBIC	RAIC	RBIC	RAIC	RBIC	RAIC	RBIC	RAIC
1	.145	.002	.012	.000	.003	.000	.001	.000	.000	.000
2	**.814**$_{[.808]}$	**.327**	**.949**$_{[.949]}$	**.431**	**.967**$_{[.967]}$	**.492**	**.983**$_{[.983]}$	**.530**	**.999**$_{[.999]}$	**.551**
3	.038	.264	.035	.244	.024	.217	.011	.190	.000	.187
4 or 5	.003	.407	.004	.325	.006	.291	.005	.280	.001	.262

to the true order K^* in bold. The subscripts inside [·] are the proportion of times that both the mixture-order and the regime-specific AR-orders are selected correctly.

From Table 2.5, when the sample size is $n = 150$, the success rates of RBIC are 85.5 %, 96.5 %, and 95.7 % for Models 1–3 respectively. The success rate of RBIC is 94.5 % for Model 5. As the sample size increases to $n = 400$, all success rates exceed 99 %. Overall, the RBIC performs well. As expected, the RAIC tends to select higher orders in all cases.

We now focus on the results in Table 2.6 for Model 4 ($K^* = 2$). The success rate of RBIC is 81.4 % when $n = 250$, and it improves to 94.9 % and 99.9 % when $n = 400$ and $n = 1000$. Note that RAIC severely over-estimates the order even when $n = 1000$.

2.5 Real Data Examples

2.5.1 U.S. Gross Domestic Product (GDP) Growth

We analyze the data comprising the quarterly GDP growth rate (Y_t) of the U.S. over the period from the first quarter of 1947 to the first quarter of 2011. The data is obtained from the US Bureau of Economic Affairs website http://www.bea.gov. Figure 2.1 contains the time series plot, the histogram and the sample autocorrelation function (ACF) of 256 observations of Y_t. The time series plot shows that the variation in the series changes over time, and the histogram of the series is multimodal. This motivates us to consider fitting a MAR model to this data. The ACF plot indicates that the sample autocorrelation function at the first two lags are significant. Thus, we let the maximum $q = 5$ and applied our method in Sect. 2.3 and fitted MAR models with $K = 1, 2, 3, 4$, to this data set. The RBIC values for $k = 1, 2, 3, 4$ are: -351.66, $\mathbf{-343.01}$, -344.14, -345.36. Thus, we select $\hat{K} = 2$.

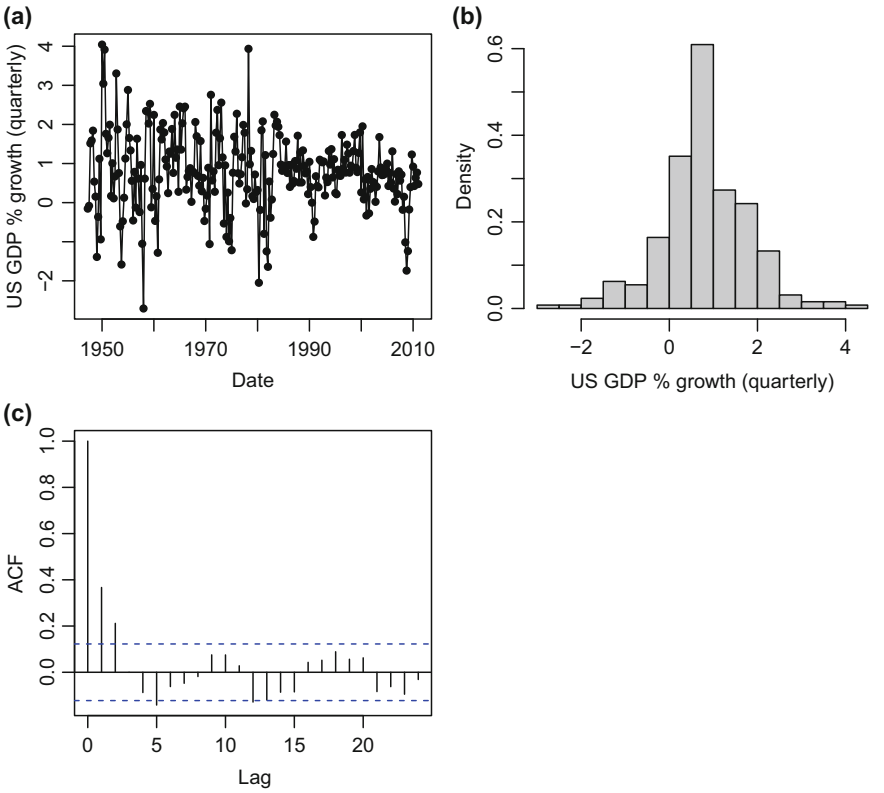

Fig. 2.1 (**a**) and (**b**) The time series plot and histogram of the U.S. GDP data. (**c**) The ACF of the U.S. GDP data

The fitted conditional density function of the model is given by

$$f(y_t|y_{t-1}) = .276\,\phi(y_t; .702, .286^2) + .724\,\phi(y_t; .497 + .401 y_{t-1}, 1.06^2).$$

The standard errors of the estimators $(\hat{\theta}_{10}, \hat{\theta}_{20}, \hat{\theta}_{21})$ are $(.060, .092, .080)$. The estimated conditional variance of Y_t is:

$$\widehat{\mathrm{Var}}(Y_t|y_{t-1}) = \sum_{k=1}^{2} \hat{\pi}_k \hat{\sigma}_k^2 + \sum_{k=1}^{2} \hat{\pi}_k \hat{\mu}_{k,t}^2 - \left(\sum_{k=1}^{2} \hat{\pi}_k \hat{\mu}_{k,t} \right)^2 = .838 - .033\, y_{t-1} + .032\, y_{t-1}^2.$$

We have the conditional variance plotted with respect to time in Fig. 2.2. It is seen that up to the year 1980, the time series has high volatility compared to the years after 1980.

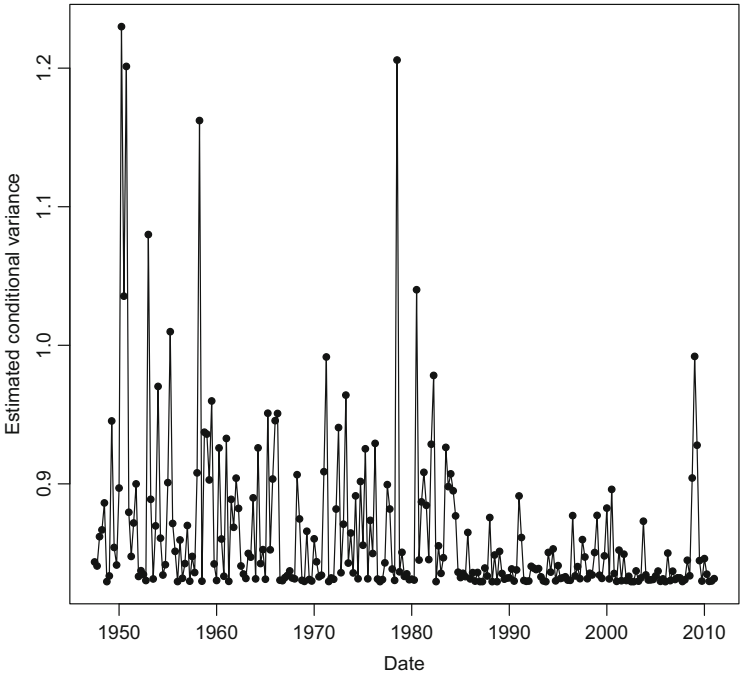

Fig. 2.2 The conditional variance of the fitted MAR model to the U.S. GDP data

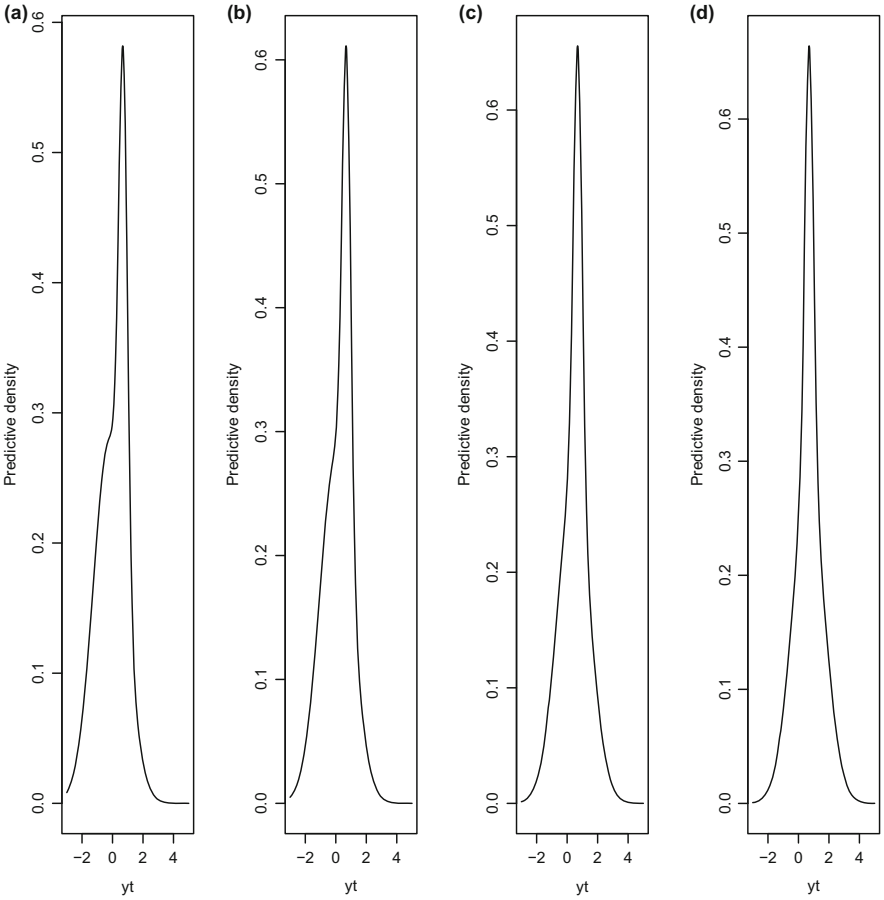

Fig. 2.3 One-step predictive density at 4 quarters of year 2009

Figures 2.3 and 2.4 give a number of first-step predictive (conditional) density functions. The time points correspond to 4 quarters in year 2009 and 8 quarters in years 1949 and 1950. Over two periods, the conditional density function changes from bimodal to unimodal or from unimodal to bimodal. It is interesting to find these changes occur when the time series experiences high volatility. The fitted MAR model has successfully captured such behaviours.

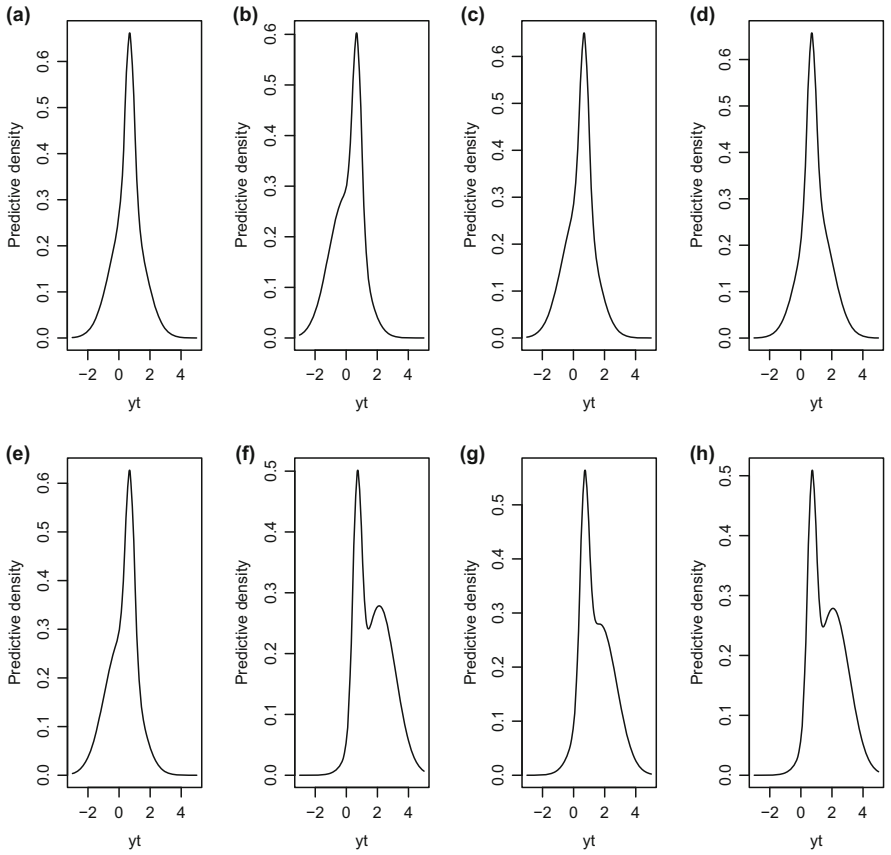

Fig. 2.4 One-step predictive density at 8 quarters in years 1949–1950

2.5.2 U.S. Unemployment Rate

The data are monthly U.S. unemployment rates over the period of 1948–2010, obtained from http://www.bea.gov. The time series plot and histogram of the observed series of length 755 is given in Fig. 2.5. The time series plot shows an increasing and decreasing trend in the series and also high volatility over time. The histogram of the series is clearly multimodal indicating that a MAR model may be appropriate. As is commonly done in time series we use the first difference transformation of the series in order to remove the increasing and decreasing trend in the series. The time series plot and also the ACF of the first difference $z_t = y_t - y_{t-1}$ are given in Fig. 2.5. The trend in the mean series has been

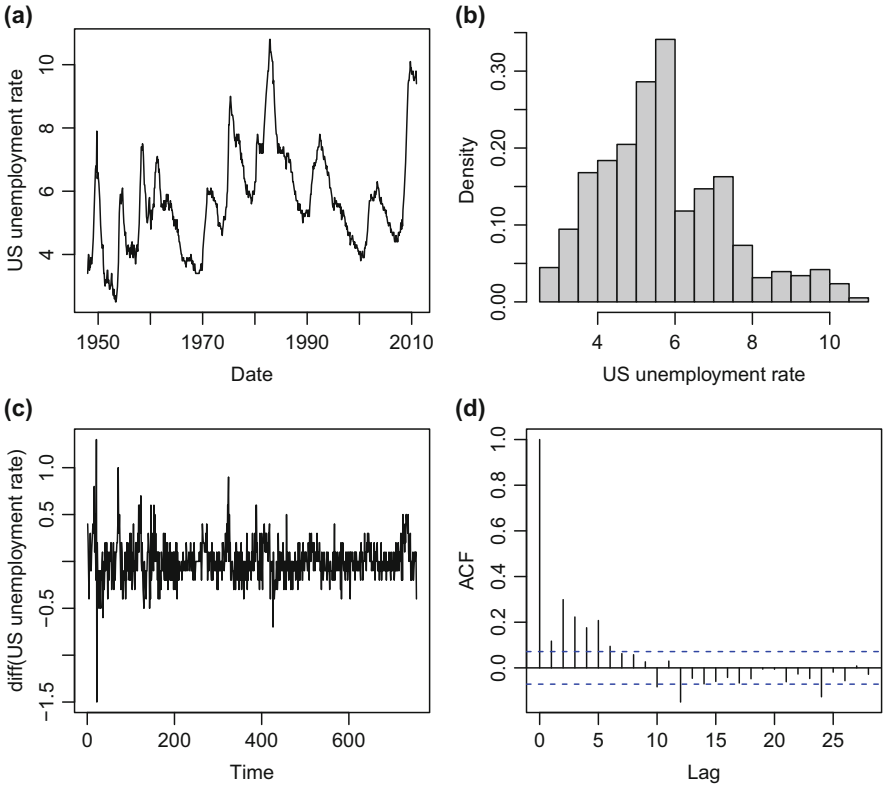

Fig. 2.5 (**a**) and (**b**) The time series plot and histogram of the monthly U.S. unemployment rate. (**c**) The time series plot of the first difference of the U.S. unemployment rate. (**d**) The ACF of the first difference

successfully removed but the variance still changes over time. In what follows we fit a MAR model to the difference z_t. Based on the ACF of the z_t in Fig. 2.5, the autocorrelation at around lag five seem significant. Thus we let $q = 5$ and applied our method in Sect. 2.3 and fitted MAR models with $K = 1, 2, 3, 4$, to z_t. The RBIC values for $K = 1, 2, 3, 4$ are: $-396.28, -359.25, -\mathbf{345.25}, -353.49$. Thus, we select $\hat{K} = 3$. The parameter estimates of the corresponding fitted MAR model are $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3; \hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3; \hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{15}, \hat{\theta}_{22}, \hat{\theta}_{23}, \hat{\theta}_{31}, \hat{\theta}_{35}) = (.184, .742, .074; .152, .148, .256; .631, .659, .298, .143, .182, -1.15, .862)$. The standard errors of the AR-coefficient estimators are $(.128, .113, .126, .036, .033, .486, .217)$.

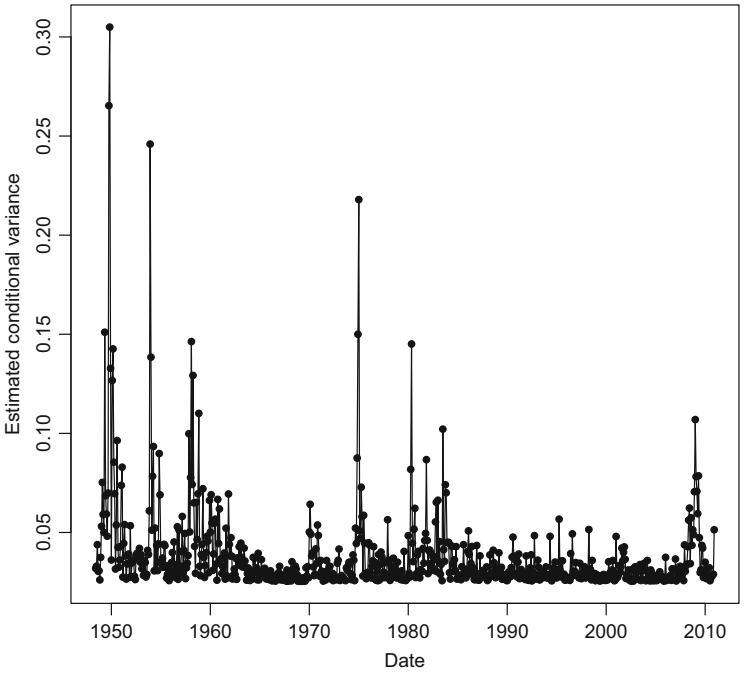

Fig. 2.6 Fitted conditional variance of MAR model to the U.S. unemployment rate

The fitted MAR model to the original series y_t has the conditional density function

$$\begin{aligned} f(y_t|y_{t-1},\ldots,y_{t-6}) &= .184\,\phi(y_t; 1.63y_{t-1} + .028y_{t-2} - .659y_{t-3} \\ &\quad + .298y_{t-5} - .298y_{t-6}, .152^2) + .742\,\phi(y_t; y_{t-1} + .143y_{t-2} \\ &\quad + .039y_{t-3} - .182y_{t-4}, .148^2) + .074\,\phi(y_t; -.150y_{t-1} \\ &\quad + 1.15y_{t-2} + .862y_{t-5} - .862y_{t-6}, .256^2). \end{aligned}$$

The estimated conditional variance of Y_t, i.e. $\widehat{Var}(Y_t|y_{t-1},\ldots,y_{t-6})$, is plotted against time in Fig. 2.6. It is seen that the unemployment rate has high volatility over different time periods.

Figure 2.7 shows the one-step predictive density function of the series y_t for the period of November 1974 to April 1975. The shape of the predictive density changes over this period where the unemployment rate y_t has experienced a dramatic change from 6.6 to 8.8. We have observed similar behaviours of the one-step predictive density over different time periods where the series has high volatility. For example, in year 1983, the unemployment rate decreases from 10.4 in January to 8.3 in December. To save space, the related plots are not reported here.

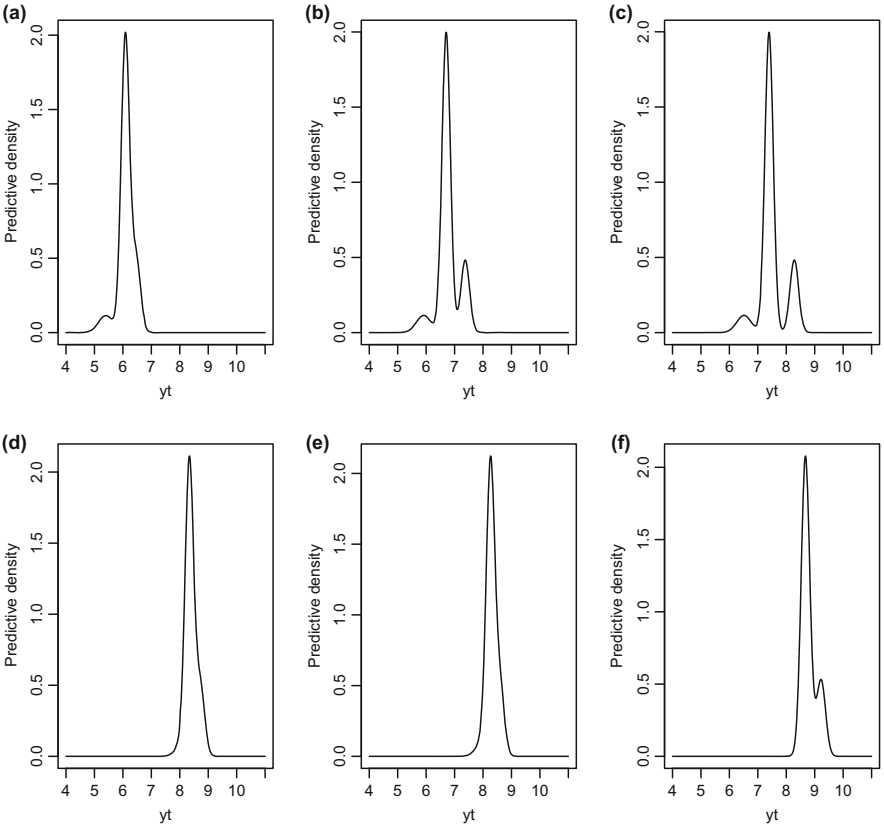

Fig. 2.7 One-step predictive density for the period of November 1974 to April 1975

2.6 Summary and Discussion

Regime-switching Gaussian autoregressive (AR) models provide a rich class of statistical models for time series data. We have developed a computationally efficient regularization method for the selection of the regime-specific AR-orders and the number of AR regimes. We evaluated finite sample performance of the proposed methods through extensive simulations. The proposed RBIC for selecting the number AR-regimes performs well in various situations considered in our simulation studies. It represents a substantial computational advantage compared to the standard BIC*. The proposed methodologies could be extended to the situations where there are exogenous variables $\boldsymbol{x}_t$ affecting the time series y_t. Large sample properties such as selection consistency and oracle properties of the proposed regularization methods are the subject of future research.

References

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31(3):307–327

Chen J, Tan X, and Zhang R (2008) Inference for normal mixtures in mean and variance. Stat Sin 18:443–465

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38

Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50(4):987–1007

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Hathaway RJ (1985) A constraint formulation of maximum-likelihood estimation for normal mixture distributions. Ann Stat 13:795–800

Jirak M (2012) Simultaneous confidence bands for Yule-Walker estimators and order selection. Ann Stat 40(1):494–528

Le ND, Martin RD, Raftery AE (1996) Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. J Am Stat Assoc 91:1504–1514

Nardi Y, Rinaldo A (2011) Autoregressive process modeling via the Lasso procedure. J Multivar Anal 102(3):528–549

Tibshirani R (1996) Regression shrinkage and selection via Lasso. J R Stat Soc B 58:267–288

Wang H, Li G, Tsai C-L (2007) Regression coefficient and autoregressive order shrinkage and selection via the Lasso. J R Stat Soc B 69(1):63–78

Wong CS, Li WK (2000) On a mixture autoregressive model. J R Stat Soc B 62:95–115

Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429

Chapter 3
Modeling Zero Inflation and Overdispersion in the Length of Hospital Stay for Patients with Ischaemic Heart Disease

Cindy Xin Feng and Longhai Li

Abstract Ischaemic heart disease is the top one cause of death in the world; however, quantifying its burden in a population is a challenge. Hospitalization data provide a proxy for measuring the severity of ischaemic heart disease. Length of stay (LOS) in hospital is often used as an indicator of hospital efficiency and a proxy of resource consumption, which may be characterized as zero-inflated if there is an over-abundance of zeros, or zero-deflated if there are fewer zeros than expected under a standard count model. Such data may also have a highly right-skewed distribution for the nonzero values. Hurdle models and zero inflated models were developed to accommodate both the excess zeros and skewness of the data with various configuration of spatial random effects, as well as allowing for analysis of nonlinear effect of seasonality and other fixed effect covariates. We draw attention to considerable drawbacks with regards to model misspecifications. Modeling and inference use the fully Bayesian approach via Markov Chain Monte Carlo (MCMC) simulation techniques. Our results indicate that both hurdle and zero inflated models accounting for clustering at the residential neighborhood level outperforms the models without counterpart models, and modeling the count component as a negative binomial distribution is significantly superior to ones with a Poisson distribution. Additionally, hurdle models provide a better fit compared to the counterpart zero-inflated models in our application.

C.X. Feng (✉)
School of Public Health, University of Saskatchewan, 104 Clinic Place, Saskatoon, SK, S7N5E5, Canada
e-mail: cindy.feng@usask.ca

L. Li
Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road, Saskatoon, SK, S7N 5E6, Canada
e-mail: longhai.li@usask.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_3

3.1 Introduction

Ischaemic heart disease (IHD), also known as coronary heart disease (CHD) (Bhatia 2010), is a disease characterized by reduced blood supply to the heart due to buildup of plaque along the inner walls of the coronary arteries. IHD is the top one cause of death in the world (Mathers and Loncar 2006; Murray 1997). In 2004, the number of IHD-related deaths was 7.2 million, accounting for 12.2 % of all deaths and 5.8 % of all years of life lost, and 23.2 million people experienced moderate or severe disability due to IHD (Mathers et al. 2008). As the most common type of heart disease, the projected total costs of IHD will increase from 46.8 billion dollars in 2015 to 106.4 billion dollars in 2030 (Go et al. 2013). In 2008, cardiovascular disease accounted for 29 % of all deaths in Canada with 28 % of all male deaths and 29.7 % of all female deaths according to the mortality, summary list of causes released by Statistics Canada in 2011. Among those, 54 % were due to ischemic heart disease. In 2005/06 there were 160,323 hospitalizations for ischemic heart disease (Tracking Heart Disease and Stroke in Canada 2009), which caused huge burden on medical care services.

Many studies on improving admission outcomes for ischemic heart disease patients tended to focus on reducing duration of in-patient care, which is often measured as length of stay (LOS), i.e., duration of a hospital admission (i.e., the difference in days between the date of admission and the date of discharge). A shorter LOS often results in reduced costs of health resources. Therefore, LOS is often used as an indicator of hospital efficiency and a proxy of resource consumption, which is a also a measure of crucial recovery time for in-patient treatment. The number of heart disease patients in need of surgery is increasing due to the aging population and prevalence (Chassot et al. 2002). To prepare for the increasing demand of inpatient treatment from a service management perspective as well as with the advances in pharmaceutical, medical technologies and clinical practice, health services provided same-day surgery, also known as ambulatory surgery or outpatient surgery. This type of surgery does not require an overnight hospital stay, so that surgery patients may go home and do not need an overnight hospital bed, leading to a decline in LOS. The purpose of outpatient surgery is to create a cost reduction to the health system, as well as saving the patient time.

Analysis of LOS data may assist monitoring and planning rescouses allocation and designing appropriate interventions. The potential risk factors for LOS can be at both patient or group level, some observed and others unobserved and maybe spatially correlated. From a statistical point of view, several important features of the LOS data must be considered. First, the data are potentially zero-inflated or deflated depending on the proportion of patients with day surgery (LOS $= 0$). Second, because the patients are clustered within neighborhood, within cluster correlation need to be addressed. Within the context of health services research, the study of regional variation in LOS can help suggesting regional health care inequalities, which can motivate further study by examining the nature of these inequalities. This type of geographical differences may be driven by socio-economic

determinants, availability and access to health care and health seeking behavior. Third, for ischaemic heart disease, temporal variation may be a factor of access to care; therefore, needs to be modeled flexibly to capture the non-linear trend. Finally, it is unclear if the probability of having day surgery for patients from certain areas are correlated with the mean of the LOS for patients from the same geographic areas reflecting the needs of health care in geographic areas. If correlation is absent, that implies the health care seeking behaviors for outpatients and inpatients are not geographically correlated. If correlation is present, failing to account for such dependence may produce biased parameter estimates. Therefore, adequate statistical modeling and analysis taking into all those features in this data is needed.

The empirical investigation indicated that the number of zeros (LOS $=$ 0) is greater than expected under a standard count distribution, so the data is zero inflated. The excess zeros often comes from two sources. Some may be from subjects who choose not to stay in hospital overnight and thereby contributing to 'sampling zeros' while some who are genuine non-user of inpatient service and are hence considered as 'structured zeros'. Standard count distributions, such as Poisson and negative binomial, may fail to provide an adequate fit, since they can not account for excess zeros arising from two data generating process simultaneously. Fitting these models may lead to inflated variance estimates of model parameters. If the data exhibit only an excess of structural zeros, a hurdle model (Heilbron 1994; Mullahy 1986) would be more appropriate. Hurdle models have been exploited in many disciplines as well, such as drug and alcohol use study (Buu et al. 2012) and health-care utilization (Neelon et al. 2013; Neelon and O'Malley 2010). The model consists of two components: a Bernoulli component modeling a point mass at zero and a zero-truncated count distribution for the positive observations. Alternatively, if excess zeros are comprised of both structural and sampling zeros, zero-inflated model (Lambert 1992) can be used, which combines the untruncated count distribution and a degenerate distribution at zero. This type of model has been extensively used in many fields, such as environmental and ecological science (Ver Hoef and Jansen 2007), substance abuse (Buu et al. 2012), dentistry (Mwalili et al. 2008) and public health (Musal and Aktekin 2013), etc.

The distinction between zero-inflated and hurdle models may be subtle, but one may be more appropriate than another depending how the zeros arise. The different models can yield different results with different interpretations. If zeros arise in only one way, then a hurdle model may be more appropriate. In the context of our study, if patients either decline or have never been referred to same day surgery, in which case the zero observations only come from 'structural' source. In contrast, if zeros arises in two sources: among those who are not at risk of being hospitalized over night or those who are at risk but nevertheless choose not to use services. In such case, a zero-inflated model would be more desirable.

Model fitting can be carried out either using EM algorithm or Bayesian approach. For each component, patient or neighborhood level fixed effect covariates, as well as random effects at neighborhood level accounting for clustering within neighbors can be included. The random effect terms for each of two model components can be modeled as an independent and identically distributed normal distribution (IID).

To provide spatial smoothing and borrowing information across neighborhoods, spatially correlated random effect – conditional autoregressive model (CAR) (Besag et al. 1991) can be applied to the count or both Bernoulli and count components (Agarwal et al. 2002; Rathbun 2006; Ver Hoef and Jansen 2007). Furthermore, (Neelon et al. 2013) developed a spatial hurdle model for exploring geographic variation in emergency department utilization by linking the Bernoulli and Poisson components of a hurdle model using a bivariate conditional CAR prior (Gelfand and Vounatsou 2003; Mardia 1988). In our study, we seek to investigate zero-inflated and hurdle models with various random effect structures, accommodating potential overdispersion by considering different parametric specifications of count distribution. We draw attention to considerable drawbacks with regards to model misspecifications.

The rest of the paper is organized as follows. We first describe the data. Next, we specify the models and outline the Bayesian approach used for model estimation. This is followed by the application of the model to the data, and then results are presented. Discussion on the results and limitations of the study conclude this Chapter.

3.2 Methods

3.2.1 Data

We used hospital discharge administrative database with the admission date ranging between January 1 and December 31 in 2011 due to IHD, which was provided by the Saskatchewan Ministry of Health. These administrative databases produced by every acute care hospital in the province of Saskatchewan, provide the following information from every single admission: age, gender, admission and discharge dates, the patient's areas of residence, and diagnosis and procedure codes [International Classification of disease (ICD) 10th revision Clinical Modification Code – I20 – I25 for IHD]. The patients' areas of residence, i.e., postal code, is confidential; therefore, each case was matched to one of the 33 health districts in Saskatchewan.

Figure 3.1 presents the histogram of the LOS for IHD patients from Saskatchewan in 2011. Of the 5777 hospitalized cases due to IHD, 1408(24 %) had same-day surgery, which constitutes the zero counts in LOS. Among those inpatients who stayed in hospital overnight, the number of days ranged from 1 to 156, with 75 % having fewer than a week of stay. Suppose that the data were generated under an independent and identically distributed Poisson regression with mean parameter as the mean 4.5 days, which is the mean of the LOS in our data. Under such model, we would expect about 1 % of zeros, which is far fewer 0s than observed. The proportion of zeros and the right-skewed non-zero counts suggest the potential zero inflation relative to the conventional Poisson distribution and overdispersion. Hence, special distributions are needed to provide an adequate fit to the data.

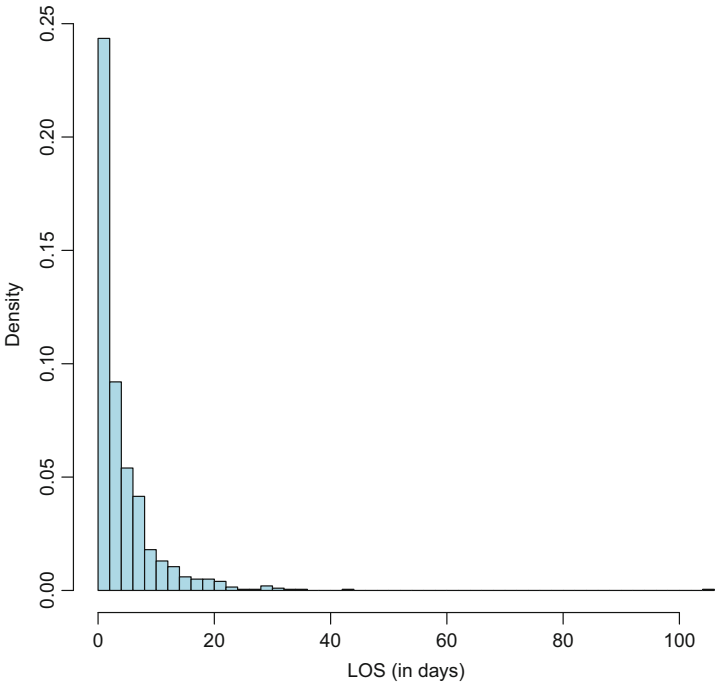

Fig. 3.1 Empirical distribution of LOS in days

Table 3.1 provides summary statistics on patient characteristics. Of the 5777 hospitalized cases due to ischaemic heart disease, 3931(68 %) are males, 391(6.8 %) are Aboriginal. During the study period, the number of IHD hospitalized cases tends to slightly vary over time, with the median age at 69 years old (Interquartile range (IQR): 59–79). Those characteristics of the data are more or less the same for those who had day-surgery and those who stayed in hospital overnight. For those who stayed in hospital over night, the median LOS is around 4 days with IQR ranging from 2 days to about a week.

The left panel in Fig. 3.2 displays the percentage of patients accessing same day surgery from each health district, which indicates higher values appeared in the south and in the middle of the province, but generally lower in the north territories. The right panel of the Fig. 3.2 presents the average number of LOS per patient from each health district, which shows that one of the health districts in the north west had higher mean LOS and a cluster of health district on the south east had a higher mean LOS compared with the rest of the health districts.

Table 3.1 Summary statistics of the data

Variable	Total	LOS = 0	LOS > 0	
	n(%)	*n*(%)	*n*(%)	Median(IQR)
Gender				
Male	3931(68.0)	1006(25.6)	2925(74.4)	3(0–5)
Female	1846(21.8)	402(21.8)	1444(78.2)	3(1–6)
Ethnicity				
Aboriginal	391(6.8)	88(22.5)	303(77.5)	4(2–6)
Non-Aboriginal	5386(93.2)	1320(24.5)	4066(75.5)	4(2–7)
Month				
Jan	529(9.2)	112(21.2)	417(78.8)	4(2–8)
Feb	469(8.1)	103(22.0)	366(78.0)	4(2–7)
Mar	520(9.0)	147(28.3)	373(71.7)	4(2–8)
Apr	483(8.4)	112(23.2)	371(76.8)	4(2–8)
May	490(8.5)	119(24.3)	371(75.7)	4(2–8)
Jun	504(8.7)	136(27.0)	368(73.0)	4(2–6)
Jul	445(7.7)	106(23.8)	339(76.2)	4(2–7)
Aug	404(7.0)	110(27.2)	294(72.8)	4(2–6)
Sep	460(8.0)	109(23.6)	351(76.3)	4(2–7)
Oct	505(8.7)	119(23.6)	386(76.4)	4(2–7)
Nov	503(8.7)	114(22.7)	389(77.3)	4(2–7)
Dec	465(8.0)	121(26.0)	344(74.0)	4(2–7)
Age				
[18, 40)	54(0.93)	12(22.2)	42(77.8)	2(1–4)
[40, 50)	368(6.4)	87(23.6)	281(76.4)	2.5(1–5)
[50, 60)	1074(18.6)	268(25.0)	806(75.0)	2(1–5)
[60, 70)	1457(25.2)	420(28.8)	1037(71.2)	2(0–5)
[70, 80)	1434(24.8)	405(28.2)	1029(71.8)	2(0–6)
80+	1390(24.1)	216(15.5)	1174(84.5)	4(1–8)

3.2.2 The Statistical Models

3.2.2.1 The Hurdle Model

The hurdle model (Heilbron 1994; Mullahy 1986) is a two-component mixture model consisting of a zero mass and the non-zero observations component following a conventional count distribution, such as Poisson or negative binomial.

Let Y_{ij} denote the LOS in days for ith, $i = 1, \cdots, n$, patient from health district j, $j = 1, \cdots, J$. The general structure of a hurdle model is given by

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & y_{ij} = 0, \\ (1 - \pi_{ij}) \frac{p(y_{ij}; \boldsymbol{\theta}_{ij})}{1 - p(0; \boldsymbol{\theta}_{ij})} & y_{ij} > 0, \end{cases} \tag{3.1}$$

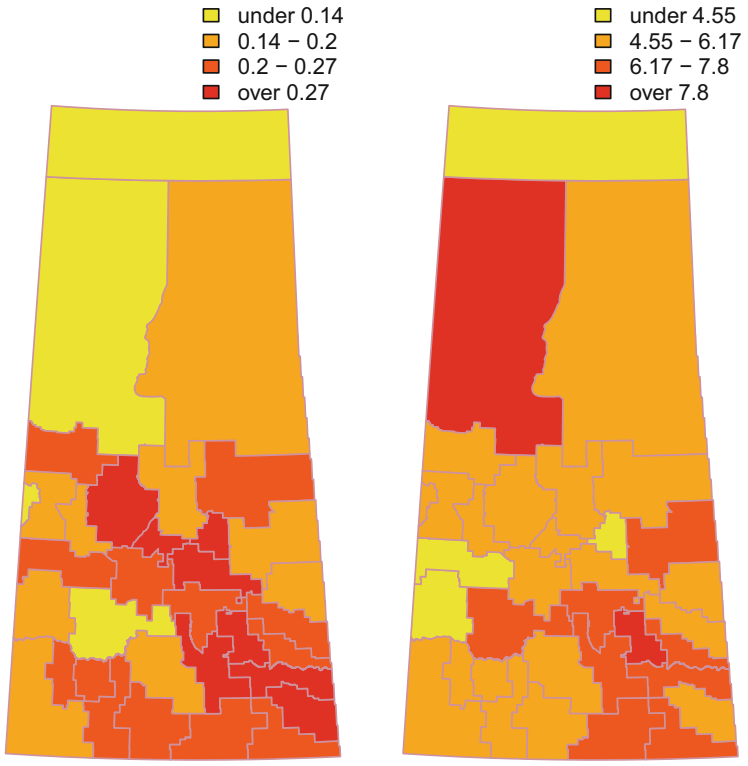

Fig. 3.2 The panel on the *left*: percentage of patients with day surgery in each of the health districts from Saskatchewan in 2011; the panel on the *right*: mean number of LOS among those inpatients with at least one day stay in hospital. The *darker color* represents higher values

where $\pi_{ij} = P(Y_{ij} = 0)$ is the probability of a subject belonging to the zero component; $p(y_{ij}; \boldsymbol{\theta}_{ij})$ represents a probability distribution for a regular count distribution with a vector of parameters $\boldsymbol{\theta}_{ij}$ and $p(0; \boldsymbol{\theta}_{ij})$ is the distribution evaluated at zero. If the count distribution follows a Poisson distribution, the probability distribution for the *hurdle Poisson model* is written as:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & \text{if } y_{ij} = 0, \\ (1 - \pi_{ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} / y_{ij}!}{1 - e^{-\mu_{ij}}} & \text{if } y_{ij} > 0 \end{cases}. \tag{3.2}$$

Alternatively, the non-zero count component can follow other distributions to account for overdispersion and negative binomial is the most commonly used. The *hurdle negative binomial model (hurdle NB)* is given by:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} & \text{if } y_{ij} = 0, \\ \frac{1 - \pi_{ij}}{1 - \left(\frac{r}{\mu_{ij} + r}\right)^r} \frac{\Gamma(y_{ij} + r)}{\Gamma(r) y_{ij}!} \left(\frac{\mu_{ij}}{\mu_{ij} + r}\right)^{y_{ij}} \left(\frac{r}{\mu_{ij} + r}\right)^r & \text{if } y_{ij} > 0 \end{cases}, \tag{3.3}$$

where $(1 + \mu_{ij}/r)$ is a measure of overdispersion. As $r \to \infty$, the negative binomial converges to a Poisson distribution. To model the association between a set of predictors and the zero-modified response, both hurdle Poisson or hurdle NB models can be extended to a regression setting by modeling each component as a function of covariates. The covariates appearing in the two components are not necessarily the same. Let $\boldsymbol{w}'_{ij}$ be the set of factors contributing to the out-patient with LOS $= 0$ and $\boldsymbol{x}'_{ij}$ be the set of factors contributing to the in-patient with non-zero LOS. The parameter π_{ij} represents the probability of using day surgery. When $\pi_{ij} = 1$, no patients received day surgery and the data follows a truncated count distribution, whereas, when $\pi_{ij} = 0$, no patients stayed in hospital overnight. π_{ij} ranges between 0 and 1. The parameter μ_{ij} measures the expected mean counts of LOS (in days) for those patients who stayed in hospital overnight, so as μ_{ij} increases, the average LOS increases. Both $\text{logit}(\pi_{ij})$ and $\log(\mu_{ij})$ are assumed to depend on a function of covariates. In addition, the random effects at the health district level are introduced in the model to account for possible correlation between the two components. The random components also control the variation at the health district level. The model can be written as:

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \boldsymbol{w}'_{ij}\boldsymbol{\alpha} + f_1(\text{month}_{ij}) + b_{1j} \\ \log(\mu_{ij}) &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + f_2(\text{month}_{ij}) + b_{2j}, \end{aligned} \tag{3.4}$$

where $\boldsymbol{w}'_{ij}$ and $\boldsymbol{x}'_{ij}$ are patient level fixed effect covariates for the logistic and Poisson components, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding vectors of regression coefficients. In our study context,

$$\begin{aligned} \boldsymbol{w}'_{ij}\boldsymbol{\alpha} = {} & \alpha_0 + \alpha_1 \text{abor}_{ij} + \alpha_2 \text{male}_{ij} + \alpha_3 I(\text{age}_{ij} \in [18, 40)) + \alpha_4 I(\text{age}_{ij} \in [40, 50)) + \\ & \alpha_5 I(\text{age}_{ij} \in [50, 60)) + \alpha_6 I(\text{age}_{ij} \in [60, 70)) + \alpha_7 I(\text{age}_{ij} \in [70, 80)) \\ \boldsymbol{x}'_{ij}\boldsymbol{\beta} = {} & \beta_0 + \beta_1 \text{abor}_{ij} + \beta_2 \text{male}_{ij} + \beta_3 I(\text{age}_{ij} \in [18, 40)) + \beta_4 I(\text{age}_{ij} \in [40, 50)) + \\ & \beta_5 I(\text{age}_{ij} \in [50, 60)) + \beta_6 I(\text{age}_{ij} \in [60, 70)) + \beta_7 I(\text{age}_{ij} \in [70, 80)), \end{aligned} \tag{3.5}$$

where α_0 and β_0 represent the intercept terms for the excess zero and random count components, respectively, and abor is the indicator of Aboriginal status for the patients with non-Aboriginal as the reference category; male denotes the male gender with female gender as the reference category. The age variable is categorized into 6 categories with the 80 years old and above as the reference category and $I()$ is 1 if the condition in the bracket is true.

Seasonal variation has been observed in mortality due to coronary heart disease, often characterized by a winter peak (Bull and Morton 1978; Rogot and Blackwelder 1979; Rose 1966). It has been postulated that temperature changes could account for practically all of the seasonal variation observed in coronary heart disease deaths, since lower environmental temperature may exert a direct effect on the heart or has an indirect effect via changes in blood pressure (Woodhouse et al. 1993). Even though the existing literature contains a vast amount of evidence on the role of

seasonal variations in the effects of IHD mortality, little is currently available on the possible effects of seasonal variations on LOS due to IHD. Hence, we considered to flexibly model the temporal effect of month of being admitted to the hospital using the smooth function with cubic B-spline basis in our study. The specification for the spline function in the logit(π_{ij}) and the log(μ_{ij}) components are:

$$f_h(\text{month}_{ij}) = \sum_{k=1}^{K} c_{kh} B_k(\text{month}_{ij}), h = 1, 2, \tag{3.6}$$

where $B_k(\text{month}_{ij}), k = 1, \cdots, K$ denote the cubic B-spline basis function with a predefined number of equidistant knots for the excess zero and the random Poisson component, respectively; and $\{c_{kh}, k = 1, \cdots, K; h = 1, 2\}$ denotes the corresponding regression coefficients for the basis functions of month. To ensure enough flexibility, we choose $K = 6$.

The parameters b_{1j} and b_{2j} in (3.4) are random effect terms to account for residual variation at the areal level unexplained by the patient level covariates, where b_{1j} is a latent areal level variable contributing to the propensity of access day surgery for patients living in health district j and b_{2j} is a latent variable contributing to the expected mean of LOS for those inpatients from health district j. As such, larger values of b_{1j} imply that inpatients living in health district j are more likely to receive day surgery compared with patients in health districts with lower b_{1j} values. Likewise, larger values of b_{2j} imply, on average, longer LOS among patients in the jth health districts compared with other health districts.

Those random effect terms can account for unmeasured characteristics at the health district level; therefore, to study their correlation is of interests, as it can reflect the association between the propensity of accessing day-surgery and the mean length of hospital stay. For example, the patients from some health districts could be more likely to be referred to day-surgery and also patients from those health districts tend to stay in hospital longer than those from other health districts. Alternatively, patients from certain health districts may be more likely to access day surgery rather than staying in hospital overnight, vise versa; or the spatial patterns for propensity of receiving day surgery and mean of LOS are not statistically related. To account for the potential association, we can assume a joint multivariate normal distribution for $\boldsymbol{b}_j = (b_{1j}, b_{2j})^T, j = 1, \cdots, J$ as $\boldsymbol{b}_j \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a 2×2 variance-covariance matrix with diagonal elements Σ_{11} and Σ_{22} representing the conditional variances of $\boldsymbol{b}_1 = (b_{1j}, \cdots, b_{1J})^T$ and $\boldsymbol{b}_2 = (b_{2j}, \cdots, b_{2J})^T$ respectively, and off-diagonal element Σ_{12} representing the within-area covariance between $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$. The correlation between the $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ is $\rho = \Sigma_{12}/\Sigma_{11}\Sigma_{22}$, which measures the strength of the association between the two process, $-1 \leq \rho \leq 1$. When $\rho = 0$, the two components of the hurdle model are uncorrelated, so the propensity of using day surgery is unrelated to the mean length of hospital stay within an area. When $\rho > 0$, health districts with a higher proportion of day surgery users tend to have higher mean of length of hospital stay and when $\rho < 0$, health districts with a higher proportion of day surgery tend to have lower mean of length of hospital stay.

To account potential spatial correlation for each component and across the two components, a bivariate intrinsic CAR prior distribution (Gelfand and Vounatsou 2003; Mardia 1988) can be used for $\boldsymbol{b}_j$ (Neelon et al. 2013):

$$\boldsymbol{b}_j|\boldsymbol{b}_{(-j)}, \boldsymbol{\Sigma} \sim MVN\left(\frac{1}{m_j}\sum_{\ell\in\delta_j}\boldsymbol{b}_\ell, \frac{1}{m_j}\boldsymbol{\Sigma}\right) \tag{3.7}$$

where δ_j and m_j denote the set of labels of the "neighbors" of area j and the number of neighbors, respectively. $\boldsymbol{\Sigma}$ is again a 2×2 variance-covariance matrix and the diagonal elements describing the spatial covariance structure characterizing each component of the hurdle model. The off-diagonal element Σ_{12} contains cross-covariance, the covariances between the two components at different areas, which allows the covariances between component of proportion of day surgery at area j and component of mean of length of hospital stay at area j' to be different from that between the proportion of day surgery at area j' and mean of length of hospital stay at area j.

3.2.2.2 The Zero-Inflated Model

A zero-inflated model assumes that the zero observations have two different origins: "structural" and "sampling". The sampling zeros are due to the usual Poisson (or negative binomial) distribution, which assumes that those zero observations happened by chance. Zero-inflated models assume that some zeros are observed due to some specific structure in the data. The general structure of a zero-inflated model is given as:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} + (1-\pi_{ij})p(0;\boldsymbol{\theta}_{ij}) & y_{ij} = 0, \\ (1-\pi_{ij})p(y_{ij};\boldsymbol{\theta}_{ij}) & y_{ij} > 0, \end{cases} \tag{3.8}$$

which consists of a degenerate distribution at zero and an untruncated count distribution with a vector of parameters $\boldsymbol{\theta}_{ij}$. If the count distribution follows a Poisson distribution, the *zero inflated Poisson model (ZIP)* is given by:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} + (1-\pi_{ij})e^{-\mu_{ij}} & \text{if } y_{ij} = 0, \\ (1-\pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!} & \text{if } y_{ij} > 0 \end{cases}, \tag{3.9}$$

where μ_{ij} is the mean of the standard Poisson distribution. As with hurdle models, overdispersion can be modeled via the negative binomial distribution. The *zero inflated negative binomial model (ZINB)* is then given by:

$$P(Y_{ij} = y_{ij}) = \begin{cases} \pi_{ij} + (1-\pi_{ij})\left[\left(\frac{r}{\mu_{ij}+r}\right)^r\right] & \text{if } y_{ij} = 0, \\ (1-\pi_{ij})\frac{\Gamma(y_{ij}+r)}{\Gamma(r)y_{ij}!}\left(\frac{\mu_{ij}}{\mu_{ij}+r}\right)^{y_{ij}}\left(\frac{r}{\mu_{ij}+r}\right)^r & \text{if } y_{ij} > 0 \end{cases}. \tag{3.10}$$

3.2.3 Bayesian Posterior Computation

Fully Bayesian inference is adopted for model estimation, which is based on the analysis of posterior distribution of the model parameters. In general, the posterior is highly dimensional and analytically intractable, which makes inference almost impossible. This problem is circumvented by using Markov chain Monte Carlo (MCMC) methods simulation techniques, where the samples are drawn from the fully conditional of parameters given the rest of the data. At convergence, the MCMC draws the Monte Carlo samples from the joint posterior distribution of the model parameter, which can be then used for parameter estimates and corresponding uncertainty intervals, thus avoiding the need for asymptotic assumptions when assessing the sampling variability of parameter estimates.

To complete the model specification, we assign uniform priors to the intercept parameters α_0 and β_0 and weakly informative proper priors $\mathrm{N}(0, 10)$ for the remaining regression coefficients, including the spline parameters. For the spatial covariance matrix, $\boldsymbol{\Sigma}$, we assume an inverse Wishart prior $\mathrm{IW}(2, \boldsymbol{I}_2)$, where $\boldsymbol{I}_2$ denotes the two-dimensional identity matrix. Updating the full conditionals of parameters is implemented in WinBUGS (Spiegelhalter et al. 2005). We ran two parallel dispersed chains for 20,000 iterations, each, discarding the first 10,000 as burn-in. Convergence of Markov chain Monte Carlo chains were assessed by using trace plots and Gelman-Rubin statistics, which indicated rapid convergence of the chains.

To compare various models, we employ the deviance information criterion (DIC), defined as $\mathrm{DIC} = \overline{D} + p_D$ where $\overline{D}$ is the posterior mean of the deviance, which measures the goodness of fit (Spiegelhalter et al. 2002). The penalty term p_D is the effective number of model parameters, which is a measure of model complexity. Models with lower $\overline{D}$ indicate good fit and lower values of pD indicate a parsimonious model. Therefore, models with smaller values of DIC are preferred as they achieve a more optimal combination of fit and parsimony.

3.3 Analysis of the LOS Data

To analyze the LOS data, we initially considered fitting the Poisson, negative binomial (NB), ZIP, ZINB, hurdle Poisson and hurdle NB regression models without any random effect terms. To assess which distribution fits the data better, various statistical tests were applied to evaluate over-dispersion and compare model fit when not including the random effect terms. Akaike's information criterion (AIC) (Akaike 1973) and Vuong statistic (Vuong 1989) were calculated.

Table 3.2 summarizes the statistics comparing the goodness of fit of the models. Positive difference in the Vuong statistic means that the model in the row fits better than the model in the column. Negative difference means that the model in the column fits better than the model in the row. The conventional Poisson

Table 3.2 Criteria for evaluating the goodness of fit and model selection of six competing models for analyzing the LOS of ischaemic heart disease patients from Saskatchewan in 2011. The second column is the Akaike's information criterion (AIC) and the rest of the columns present Vuong statistic. Negative number means that the model in the column fits better than the model in the row and positive number means that the model in the column fits better than the models in the row

	AIC	Vuong Statistic				
		NB	ZIP	ZINB	hurdleP[a]	hurdleNB[b]
Poisson	52836	−14	−22	−14	−22	−14
		<0.001	<0.001	<0.001	<0.001	<0.001
NB	29676	–	11	−3	11	−4
			<0.001	0.003	<0.001	<0.001
ZIP	45171	–	–	−11	0.9	−11
				<0.001	0.18	<0.001
ZINB	29665	–	–	–	11	−1.9
					<0.001	0.028
hurdleP[a]	45172	–	–	–	–	−11
						<0.001
hurdleNB[b]	29631	–	–	–	–	–

[a]hurdleP denotes hurdle Poisson model
[b]hurdleNB denotes hurdle Negative Binomial model

model is inferior to the other models as shown by all the negative numbers in its row; the hurdle model NB shows superior fit compared to the other models, with all the negative numbers in its column; and zero-inflated models fit better than their corresponding non-zero inflated counterparts; this suggests that the best fitting model needs to account for both zero-inflation and over-dispersion in the observed data. In addition, the hurdle Poisson and hurdle NB models fit better than their corresponding ZIP and ZINB models, which suggests that the zero counts were best modeled as being only structural zeroes.

Furthermore, we included the random effect terms in the model to compare the goodness of fit of ZIP, ZINB, hurdle Poisson and hurdle NB models with various configuration of random effect terms, ranging from the model without any random effect terms, models with random effect term in one of the two model components and models with random effect terms in both model components. The random effect term is either IID normally distributed or assigned with a CAR prior or both random effect terms are correlated through a bivariate normal distribution or MCAR prior conditional on the predictors. The results are presented in Table 3.3, which shows that the inclusion of random effects further improve the model fitness despite the model complexity. Therefore, modeling the impact of fixed effect factors alone is not sufficient to produce satisfactory fit to the data, and random effects at health district level in both the bernoulli or the counts components are needed to account for areal level heterogeneity. However, spatial correlation at the health district level is not strong in either of the components with the DIC scores for the MCAR models larger than the counterpart IID models.

Table 3.3 DIC and pD for competing models in the analysis of LOS for ischaemic heart disease patients from Saskatchewan in 2011

Model	Hurdle		ZI	
	Poisson	NB	Poisson	NB
No $(\boldsymbol{b}_1, \boldsymbol{b}_2)^T$	45172(28)	29631(29)	45172(28)	29655(9)
IID $\boldsymbol{b}_1$	45122(46)	29581(48)	45121(46)	29629(21)
IID $\boldsymbol{b}_2$	44284(46)	29534(48)	44289(46)	29606(21)
Independent CAR $\boldsymbol{b}_1$	45122(46)	29579(46)	45121(46)	29642(28)
Independent CAR $\boldsymbol{b}_2$	44292(62)	29540(53)	44296(62)	29611(31)
IID $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$	44233(76)	29484(71)	44231(76)	29543(50)
Independent CAR $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$	44240(79)	29488(71)	44236(79)	29556(56)
Bivariate IID $(\boldsymbol{b}_1, \boldsymbol{b}_2)^T$	44235(80)	29486(76)	44233(80)	29540(50)
MCAR $(\boldsymbol{b}_1, \boldsymbol{b}_2)^T$	44243(82)	29492(74)	44241(82)	29540(54)

Table 3.4 presents the posterior means and 95 % credible intervals for all parameters except for the B-spline coefficients for the hurdle NB and hurdle Poisson models with the bivariate IID random effect structure on $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$. Under the hurdle NB model, after adjusting for other predictors, males are more likely to access day surgery with posterior mean (95 % CI) as $0.143(0.008, 0.276)$, whereas male gender is not significantly associated with means of LOS with posterior mean (95 % CI) as $0.011(-0.073, 0.093)$. Aboriginal status had no impact on either propensity of receiving day surgery or LOS. In general, as age decreases, the likelihood of accessing day surgery increases, but not for the age under 40 years old. As a contrast, as age increases, the LOS increases, which is intuitively sensible, as elder patients needs longer time to recover.

The variance components of the random effect terms in the model indicate that the two components are not statistically associated with each other at the residential neighborhood with the posterior mean (95 % CI) of ρ being estimated as $0.108(-0.294, 0.489)$ under the hurdle NB model. This suggests that the probability of accessing day surgery is not correlated with the mean LOS among users at the health district level after adjusting for various patient-level covariates. In comparison with the hurdle NB model, the hurdle Poisson model yields relatively smaller variance component estimates.

Figure 3.3 displays the temporal trends of month being admitted to hospital due to IHD on the linear predictor scale for the two model components under the hurdle NB model with the bivariate IID random effect structure on $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$. The horizontal line at zero corresponds to no month effect. The log-odds of day surgery use do not vary over time with the point-wise credible interval covering zero. For the log of mean of LOS, although a bimodal pattern appears in the early spring and late fall, the effect is not significant, shown in Fig. 3.3. Under the counterpart hurdle Poisson model, the log-odds of day surgery use is consistent with the hurdle NB model; however, the temporal effect on the mean of LOS became more pronounced, shown in Fig. 3.4. Therefore, hurdle Poisson model yields greater temporal effect

Table 3.4 Posterior mean estimates and 95 % credible intervals (in parentheses) for the parameters from the hurdle NB and hurdle Poisson models with the random effect terms $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$ following a bivariate normal distribution

Variable	Parameter	Hurdle NB	Hurdle Poisson
logit(π_{ij})			
Intercept	α_0	−2.083(−2.365, −1.810)	−2.085(−2.385, −1.806)
Aboriginal	α_1	−0.072(−0.345, 0.187)	−0.069(−0.352, 0.189)
Male	α_2	0.143(0.008, 0.276)*	0.138(0.007, 0.284)*
Age 18–40	α_3	0.403(−0.261, 1.044)	0.431(−0.261, 1.103)
Age 40–50	α_4	0.474(0.189, 0.750)*	0.478(0.185, 0.754)*
Age 50–60	α_5	0.549(0.334, 0.750)*	0.551(0.348, 0.765)*
Age 60–70	α_6	0.762(0.575, 0.950)*	0.765(0.564, 0.966)*
Age 70–80	α_7	0.746(0.548, 0.928)*	0.753(0.555, 0.947)*
log(μ_{ij})			
Intercept	β_0	1.782(1.599, 1.961)	2.034(1.919, 2.167)
Aboriginal	β_1	0.000(−0.160, 0.176)	−0.031(−0.091, 0.031)
Male	β_2	0.011(−0.073, 0.093)	−0.004(−0.031, 0.023)
Age 18–40	β_3	−1.035(−1.438, −0.596)*	−0.842(−1.010, −0.681)*
Age 40–50	β_4	−0.815(−0.983, −0.646)*	−0.643(−0.704, −0.581)*
Age 50–60	β_5	−0.615(−0.730, −0.490)*	−0.484(−0.522, −0.444)*
Age 60–70	β_6	−0.499(−0.607, −0.390)*	−0.388(−0.421, −0.353)*
Age 70–80	β_7	−0.233(−0.339, −0.131)*	−0.177(−0.209, −0.145)*
Variance component			
var(b_{1j})	Σ_{11}	0.117(0.061, 0.212)	0.119(0.057, 0.211)
var(b_{2j})	Σ_{22}	0.101(0.057, 0.170)	0.085(0.050, 0.144)
cov(b_{1j}, b_{2j})	Σ_{12}	0.012(−0.037,0.062)	0.009(−0.034, 0.053)
corr(b_{1j}, b_{2j})	ρ	0.108(−0.294, 0.489)	0.085(−0.321, 0.459)

compared to the corresponding hurdle NB model, similarly for the ZIP compared with the ZINB model (not presented here), suggesting that failure to account for overdispersion leads to over-estimation of the temporal effect. The seasonality pattern is in contrast with the findings for coronary heart disease mortality in the literature, which often reported higher hospital mortality rates in winter than other seasons. Nevertheless, inpatients undergoing surgery who environmental condition may be under control. Such difference in seasonality pattern between mortality and LOS warrants further investigation.

Figure 3.5 presents the posterior mean estimates of the random effects $\boldsymbol{b_1}$ (left panel) and $\boldsymbol{b_2}$ (right panel) when $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$ following bivariate IID based on the hurdle NB model. The left panel indicates that those health districts in red have increased propensity of accessing day surgery, which were distributed mainly in the middle of the province stretching towards the south east of the province. The right panel shows that a health district in the north west and some regions in the south middle east have higher expected mean counts of LOS in days. The different residual spatial patterns imply that the spatial distribution for the two components

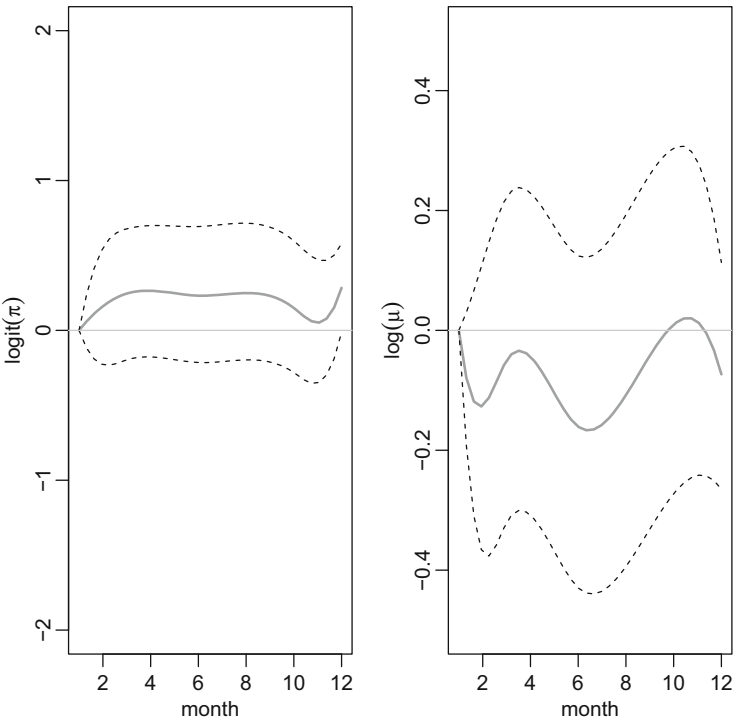

Fig. 3.3 Temporal effect on the linear predictor scale for the binary component (*left panel*) and the NB component (*right panel*) for the hurdle NB model with the random effect terms $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$ following a bivariate normal distribution. *Dashed lines* denote 95 % credible intervals

after accounting for the individual level covariates are not sharing similar spatial patterns.

3.4 Discussion

In this article, hurdle models and zero inflated models were considered to model the LOS for IHD hospitalizations. The models accommodate both excess zeros and skewness of the data with various configuration of fixed and random effects allowing for analysis of nonlinear effect of seasonality and spatial pattern. The initial inspection of the observed data, as well as fit statistics, suggested that the distribution of the LOS was both overdispersed and zero-inflated. Our results indicate that both hurdle and zero inflated models including random effects at areal level for both model components outperform the models without those terms, and modeling the count component as a negative binomial distribution is significantly superior to modeling the count component as a Poisson distribution.

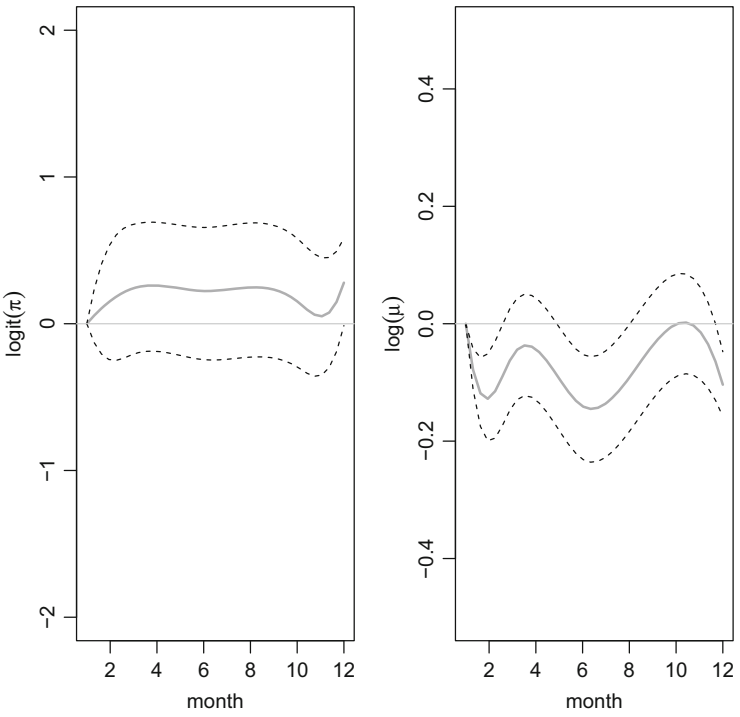

Fig. 3.4 Temporal effect on linear predictor scale for the binary component (*left panel*) and the Poisson component (*right panel*) for the hurdle Poisson model with the random effect terms $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$ following a bivariate normal distribution. *Dashed lines* denote 95 % credible intervals

Hurdle models outperform the corresponding zero-inflated models in our application. Min and Agresti (2005) suggested that hurdle models might provide better fit if there is evidence of zero deflation among subgroups of the population. Zero-inflated models imply zero inflation at all the levels of the covariates. Min and Agresti (2005) also revealed unstable nature of the zero inflated formulation, primarily because there is no distinct selection process leading to zero or non-zero value. On the contrary, the hurdle model has a very stable behavior and performance. Neelon and O'Malley (2010) gives a detailed discussion comparing the zero inflated and hurdle models in health service research setting. The importance of accounting for zero inflation and overdispersion clearly deserves further attention in the health care utilization literature.

Our results highlight some important policy implications for management of utilization of hospital services. By investigating the spatial pattern of propensity of accessing day surgery and the means of LOS, policy makers can target communities with greater needs for services such as day surgery centers to reduce the burden to

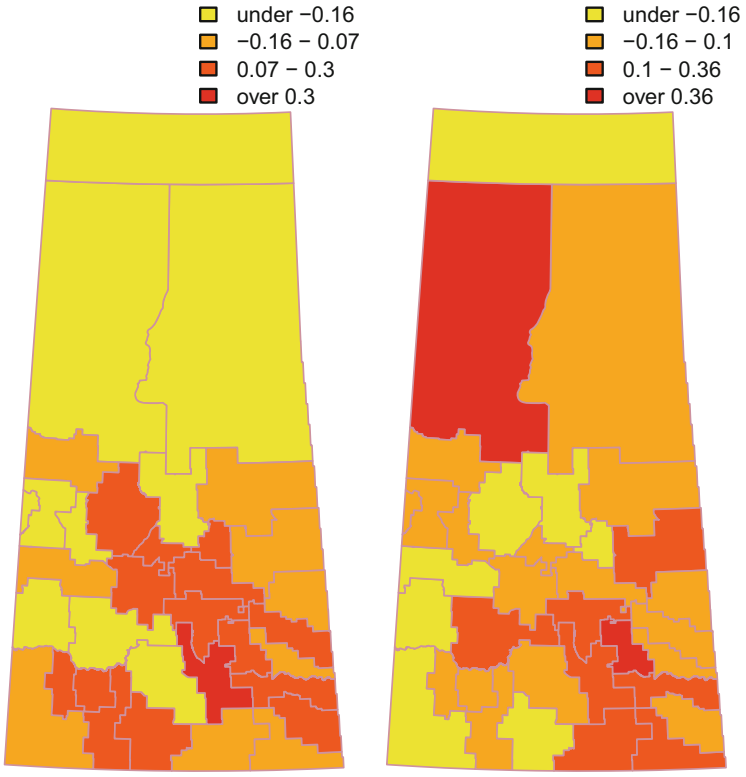

Fig. 3.5 Posterior mean estimates of the random effects $\boldsymbol{b_1}$ (*left panel*) and $\boldsymbol{b_2}$ (*right panel*)based on the hurdle NB model with the random effect terms $(\boldsymbol{b_1}, \boldsymbol{b_2})^T$ following a bivariate normal distribution

the primary health care facilities, which may be in great need at the remote or rural communities.

The estimates of the effect of covariates differed in magnitude between models. Of particular note, the Poisson models (ZIP or hurdle Poisson) estimated significant temporal effect of month of being admitted to hospitals for the means of LOS component, whereas under the negative binomial models (ZINB or hurdle NB), the temporal effect was not detected to be significant, suggesting it is important to account for overdispersion in the model, since ignoring greater dispersion in the data will result in underestimation of the variance of the estimators. This illustrates the risk of falsely identifying a significant effect if the model chosen does not model the spread of the data correctly. Although in our application the time was not detected significantly impact on either of the components, the models presented in this article can be adapted to analyze other health indicator of similar structure and in like settings.

A major limitation of our analysis is that the data used comes from hospital registers. In Saskatchewan, registered first nation patients is regulated by the Canadian federal government, so the results may be biased towards urban areas that are well covered by health facilities. Moreover, socio-demographic variables are not contained in a hospital registration data; therefore, a more representative data is to link the hospital data with a cross-sectional household surveys data, which will provide additional patient or health district level covariates reflecting patients' deprivation level; however, such data are often carried out every several years and the personal identifier is generally not being released due to confidentiality. The geographic unit in our application is only restricted at the health district level due to confidentiality of releasing postal code. Stronger spatial autocorrelation may emerge if a finer level geographic unit, such as census block, would be available for this study.

References

Agarwal DK, Gelfand A, Citron-Pousty S (2002) Zero-inflated models with application to spatial count data. Environ Ecol Stat 9:341–355

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov BV, Csaki BF (eds) Second international symposium on information theory. Academiai Kiado, Budapest

Besag J, York J, Mollie A (1991) Bayesian image restoration with two applications in spatial statistics. Ann Inst Stat Math 43:1–21

Bhatia S (2010) Biomaterials for clinical applications. Springer, New York

Bull GM, Morton J (1978) Environment, temperature and death rates. Age Ageing 7:210–230

Buu A, Li R, Tan X, Zucker R (2012) Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Stat Med 31:4074–4086

Chassot P, Delabays A, Spahn DR (2002) Preoperative evaluation of patients with, or at risk of, coronary artery disease undergoing non-cardiac surgery. Br J Anaesth 89:747–759

Gelfand AE, Vounatsou P (2003) Proper multivariate conditional autoregressive models for spatial data analysis. Biostat 4:11–25

Go A, Mozaffarian D, Roger V, Benjamin E, Berry J, Borden W, Bravata D, Dai S, Ford E, Fox C (2013) Heart disease and stroke statistics – 2013 update: a report from the American heart association. Circulation 127:e6–e245

Heilbron DC (1994) Zero-altered and other regression models for count data with added zeros. Biom J 36:531–547

Lambert D (1992) Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 34:1–14

Mardia KV (1988) Multi-dimensional multivariate gaussian Markov random fields with application to image processing. J Multivar Anal 24:265–284

Mathers C, Fat D, Boerma J (2008) The global burden of disease: 2004 update. World Health Organization, Geneva

Mathers C, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 3:2011–2030

Min Y, Agresti A (2005) Random effect models for repeated measures of zero-inflated count data. Stat Modell 5:1–19

Mullahy J (1986) Specification and testing of some modified count data models. J Econom 33:341–365

Murray CJ, Lopez AD (1997) Alternative projections of mortality and disability by cause 1990–2020: global burden of disease study. Lancet 349:1498–1504

Musal M, Aktekin T (2013) Bayesian spatial modeling of HIV mortality via zero-inflated Poisson models. Stat Med 32:267–281

Mwalili S, Lesaffre E, Declerck D (2008) The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Stat Methods Med Res 17:123–139

Neelon B, Ghosh P, Loebs P (2013) A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. J R Stat Soc Ser A 176:389–413

Neelon B, O'Malley A, Normand S (2010) A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Stat Modelling 10:421–439

Rathbun S, Fei SL (2006) A spatial zero-inflated Poisson regression model for oak regeneration. Envrion Ecol Stat 13:409–426

Rogot E, Blackwelder WC (1979) Associations of cardiovascular mortality with weather in Memphis, Tennessee. Public Health Rep 85:25–39

Rose G (1966) Cold weather and ischaemic heart disease. Br J Prev Soc Med 20:97–100

Spiegelhalter D, Thomas A, Best N, Lunn D (2005) Winbugs user manual, version 1.4. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology & Public Health, Imperial College School of Medicine. http://www.mrc-bsu.cam.ac.uk/bugs

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J R Stat Soc Ser B 64:583–640

Tracking Heart Disease and Stroke in Canada (2009) Public Health Agency of Canada. http://www.phac-aspc.gc.ca/publicat/2009/cvd-avc/pdf/cvd-avs-2009-eng.pdf

Ver Hoef JM, Jansen J (2007) Space-time zero-inflated count models of harbor seals. Environ 18:697–712

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econom 57:307–333

Woodhouse P, Khaw K, Plummer M (1993) Seasonal variation of blood pressure and its relationship to ambient temperature in an elderly population. J Hypertens 85:1267–1274

Chapter 4
Robust Optimal Interval Design for High-Dimensional Dose Finding in Multi-agent Combination Trials

Ruitao Lin and Guosheng Yin

Abstract In the era of precision medicine, combination therapy is playing a more and more important role in drug development. However, drug combinations often lead to a high-dimensional dose searching space compared to conventional single-agent dose finding, especially when three or more drugs are combined for treatment. To overcome the burden of calibration of multiple design parameters, which often intertwine with each other, we propose a robust optimal interval (ROI) design to locate the maximum tolerated dose (MTD) in phase I clinical trials. The optimal interval is determined by minimizing the probability of incorrect decisions under the Bayesian paradigm. Our method only requires specification of the target toxicity rate, which is the minimal design parameter. Neither does ROI impose any parametric assumption on the underlying distribution of the toxicity curve, nor it needs to calibrate any other design parameters. To tackle high-dimensional drug combinations, we develop a random-walk ROI design to identify the MTD combination in the multi-agent dose space. Both the single- and multi-agent ROI designs enjoy convergence properties with a large sample size. We conduct simulation studies to demonstrate the finite-sample performance of the proposed methods under various scenarios. The proposed ROI designs are simple and easy to implement, while their performances are competitive and robust.

4.1 Introduction

The primary objective of phase I dose-finding trials is to determine the maximum tolerated dose (MTD), which is typically defined as the dose with the dose-limiting toxicity (DLT) probability closest to the target toxicity rate. Nowadays, combination therapy is playing a more and more important role in drug development. After demonstrating the clinical effectiveness of two agents separately, a natural follow-up

R. Lin • G. Yin (✉)
Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China
e-mail: ruitaolin@gmail.com; gyin@hku.hk

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_4

step is to evaluate their joint effects when used in combination, especially if they target different disease pathways. In general, dose finding in two-drug combination trials is much more complicated since the joint toxicity order of the combined doses is only partially known. Due to the enormous data from the historical trials and the emergence of precision medicine, there is a trend to combine three or more drugs for the sake of improved efficacy as well as reduced side effects. However, multi-agent combination brings new challenges to the phase I dose-finding design: the dimension of the dose searching space expands multiplicatively with respect to the number of drugs in the combination. For a three-drug combination trial, a usual logistic model may need eight parameters to quantify the joint effect of the combined therapy by including the main effects, and two- and three-way interactions. More importantly, these parameters should satisfy several conditions under the partial order constraints, which in fact become very challenging to set. As the sample size of a phase I trial is typically small, it is difficult to estimate a large number of unknown parameters accurately, needless to say identifying the true MTDs in multi-agent dose finding.

Numerous statistical methods have been proposed for phase I single-agent dose-finding trials, which can generally be classified as algorithm- and model-based designs (Yin 2012). The algorithm-based methods, such as the well-known 3 + 3 design (Storer 1989), usually proceed based on a set of prespecified rules without imposing any model assumption on the unknown toxicity curve. Despite simplicity and dominance in practice, the 3 + 3 design has been criticized for its poor performance (Ahn 1998). Alternatives to the 3 + 3 design include the accelerated titration design (Simon et al. 1997), the biased coin design (Durham et al. 1997), the group up-and-down design (Gezmu and Flournoy 2006), and so on. For a comprehensive review on the algorithm-based methods, see Liu et al. (2013). By contrast, model-based dose-finding methods typically aim to find the MTD by estimating the toxicity curve based on an imposed parametric model. The most prominent model-based method is the continual reassessment method (CRM) by O'Quigley et al. (1990), which dynamically determines the possible MTD based on the observed data. For various extensions of the CRM, see Heyd and Carlin (1999), Leung and Wang (2002), and Yuan et al. (2007). Although the model-based designs tend to have superior operating characteristics over the algorithm-based, Rogatko et al. (2007) reported that only 1.6 % of the phase I cancer trials (20 of 1235 trials) published between 1991 and 2006 used model-based designs such as the modified CRM, while the remainder used variations of the 3 + 3 design.

Interval designs, which belong to the algorithm-based class, have recently attracted enormous attention due to their simplicity and desirable properties. The entire procedure of an interval design is guided by comparing the observed toxicity rate (or the number of DLTs) with a prespecified toxicity tolerance interval. Yuan and Chappell (2004) suggested utilizing an interval to determine dose escalation or de-escalation. Ivanova et al. (2007) proposed a cumulative cohort design by modifying the group up-and-down design. Ji et al. (2007) proposed a toxicity probability interval method using penalties to determine the dose assignment, and Ji et al. (2010) made a further modification based on the unit probability mass. However,

the specification of the tolerance interval is critical for the design performance. To solve this problem, Liu and Yuan (2015) developed a Bayesian optimal interval (BOIN) design by minimizing the probability of incorrect dose allocation under a Bayesian decision-making framework. From a theoretical perspective, Oron et al. (2011) showed that the MTD identified by an interval design converges almost surely to one of the doses in the tolerance interval. Lin and Yin (2016) extended BOIN to two-dimensional dose finding by comparing the posterior probability of each dose combination falling inside the predetermined interval.

Most of the aforementioned methods require certain degrees of prespecfication of design parameters, which is crucial for the trial performance. However, very limited literature is devoted to addressing the issues on parameter calibration. For example, the CRM requires the prespecification of the toxicity rates (or the skeleton) for the dose levels under consideration. Such prespecification can be arbitrary and subjective and, as a result, the operating characteristics are sensitive to various toxicity scenarios. To overcome the arbitrariness in the prespecification of toxicity rates, Yin and Yuan (2009) proposed a Bayesian model averaging CRM approach, which is robust to the misspecification of the skeleton and thus leads to competitive trial performance. Besides the skeleton, other design specifications in the CRM include the working model and the prior distributions of the unknown parameters, which also affect the design properties. The BOIN design requires to prespecify two parameters ϕ_1 and ϕ_2, which denote respectively the toxicity rates for underdosing and overdosing. However, the performance of BOIN is sensitive to these two tuning parameters, as they uniquely determine the optimal toxicity probability interval.

Similar to single-agent trial designs, two-agent dose finding in drug combination trials can also be classified as either algorithm- or model-based. Conaway et al. (2004) proposed to use the pool-adjacent-violators algorithm to determine dose allocation in drug combination trials. Ivanova and Wang (2004) applied the Narayana design to find the MTD based on partial orders. Huang et al. (2007) developed a two-agent $3 + 3$ method by partitioning the dose space into separate zones along the diagonal direction. Fan et al. (2009) proposed a three-stage $2+1+3$ design. However, the dose-escalation rules in the existing algorithm-based two-dimensional designs are rather ad-hoc and typically lack a theoretical support. Thus, the performances of these methods are well below the satisfactory level.

Most of the model-based designs are developed under the CRM framework, which continuously update the unknown parameters by assuming a certain model for the joint toxicity surface. For example, Thall et al. (2003) considered a six-parameter model for the joint toxicity rate of two drugs. Wang and Ivanova (2005) proposed a log-linear working model for the dose-toxicity relationship. Yuan and Yin (2008) applied the CRM to subtrials in a sequential order so that overly toxic or overly safe doses can be eliminated in an efficient way. Yin and Yuan (2009) utilized a copula-type regression method to characterize the interactive effects of the two agents in combination. In a more general framework of 2×2 tables (Yin and Yuan 2009; Yin and Lin 2014), many other copulae and bivariate binary models can be applied to two-drug combination designs. Shi and Yin (2013) developed a two-dimensional approach of escalation with overdose control on the basis of a

four-parameter logistic regression model. However, the number of unknown parameters in a two- or multi-agent combination trial is relatively large in comparison to the small sample size, such that the estimation may be unstable and the trial results are sensitive to some prior specifications. The situation becomes worse when three or more drugs are combined, as more unknown parameters need to be estimated in order to characterize two-way, three-way, and four-way interactions. However, there are very limited statistical methods for dose finding with three or more drugs in combination.

Our research is motivated by a phase I dose-finding study of combined treatment with mitoxantrone and genasense in patients with metastatic hormone-refractory prostate cancer (Chi et al. 2001). One of the major goals of the prostate cancer trial was to find the MTD of the combination therapy. To broaden application of interval designs as well as to overcome the arbitrary specification of design parameters ϕ_1 and ϕ_2 in the BOIN design, we propose a robust optimal interval (ROI) design that only requires the specification of the target toxicity rate (the minimal design specification for a trial). As a result, with fewer parameters to calibrate, the proposed method is more robust to various design parameters and unknown toxicity curves. In addition to the single-agent ROI design, we also develop a multi-agent random-walk ROI (RW-ROI) design, which is applicable to dose finding with two or more combined drugs. The proposed RW-ROI design adaptively searches for the MTD using the accrued information, and it can be easily extended to high-dimensional dose finding. We compare the RW-ROI method with existing approaches and demonstrate its comparative and stable operating characteristics.

The rest of the paper is organized as follows. In Sect. 4.2, we propose the single-agent ROI design, and in Sect. 4.3 we make an extension to multi-dimensional dose-finding trials with RW-ROI. Simulation studies are conducted in Sect. 4.4 to examine the operating characteristics of the new design as well as comparisons with existing methods. Section 4.5 illustrates the proposed RW-ROI method with a trial example, and Sect. 4.6 provides some concluding remarks.

4.2 Single-Agent Robust Optimal Interval Design

Consider a phase I dose-finding trial with J prespecified dose levels, whose toxicity rates monotonically increase; that is, $p_1 < \cdots < p_J$, where p_j is the true toxicity rate at dose level j, $j = 1, \ldots, J$. Let ϕ be the target toxicity rate specified by the investigator. The trial starts with treating the first cohort of patients at the lowest dose level. Suppose the current dose level is j and the total number of patients treated at dose level j is n_j. The interval design proceeds by comparing y_j, the cumulative number of DLTs at level j, with the prespecified toxicity lower and upper boundaries $\Delta_L(n_j)$ and $\Delta_U(n_j)$:

- If $y_j \leq \Delta_L(n_j)$, the dose for the next cohort is escalated to level $j + 1$.
- If $y_j \geq \Delta_U(n_j)$, the dose for the next cohort is de-escalated to level $j - 1$.

- If $\Delta_L(n_j) < y_j < \Delta_U(n_j)$ or the next dose assignment falls outside of the prespecified dose range, the next cohort is treated at the same dose level j.

For safety, overly toxic dose levels that satisfy $\Pr(p_j > \phi | y_j) \geq \lambda$ and $n_j \geq 3$ are excluded from the trial, where λ is the prespecified threshold probability. Based on the safety constraint, we can obtain the dose elimination cutoffs $\Delta_T(n_j)$: if $y_j \geq \Delta_T(n_j)$, dose level j and all the higher levels are eliminated from the trial.

To avoid arbitrary prespecifications of $\Delta_L(n_j)$ and $\Delta_U(n_j)$, Liu and Yuan (2015) derived the lower and upper boundaries by casting the dose-finding problem in a Bayesian hypothesis testing framework for each j,

$$H_0 : p_j = \phi, \quad H_1 : p_j = \phi_1, \quad H_2 : p_j = \phi_2.$$

Their method requires prespecification of two design parameters, ϕ_1 and ϕ_2, which are viewed as the highest toxicity rate (but subtherapeutic) such that the dose should be escalated and the lowest toxicity rate (while still overly toxic) such that the dose should be de-escalated, respectively. To enhance the robustness of the design as well as to circumvent calibration of redundant parameters, we consider a hypothesis setting with a single target rate parameter ϕ,

$$H_0 : p_j = \phi, \quad H_1 : p_j < \phi, \quad H_2 : p_j > \phi,$$

where H_0, H_1 and H_2 indicate that the current dose level j is the MTD, below and above the MTD, respectively. Under the Bayesian paradigm, we assume the three hypotheses are a priori equally probable, i.e., $\Pr(H_0) = \Pr(H_1) = \Pr(H_2) = 1/3$. Under the composite alternatives H_1 and H_2, we assign noninformative uniform prior distributions for p_j,

$$p_j | H_1 \sim \text{Unif}(0, \phi) \quad \text{and} \quad p_j | H_2 \sim \text{Unif}(\phi, 1),$$

while the prior distribution under H_0 is a point mass on ϕ. Based on the accumulated data at dose level j, the posterior probability of each hypothesis π_{kj} is given by

$$\pi_{kj} \equiv \Pr(H_k | y_j) = \frac{\Pr(H_k) \Pr(y_j | H_k)}{\sum_{k'=0}^{2} \Pr(H_{k'}) \Pr(y_j | H_{k'})}, \quad k = 0, 1, 2,$$

where the marginal likelihood $\Pr(y_j | H_k)$ can be obtained by integrating out the parameter p_j with respect to its prior $f(p_j | H_k)$,

$$\Pr(y_j | H_k) \propto \int p_j^{y_j} (1 - p_j)^{(n_j - y_j)} f(p_j | H_k) \mathrm{d}p_j, \quad k = 0, 1, 2.$$

Given the accumulated data y_j, the posterior probability of making incorrect decisions is formulated as

$$\begin{aligned}\Pr(\text{Incorrect}|y_j) &= \pi_{0j}\Pr(\text{E or D}|H_0) + \pi_{1j}\Pr(\text{S or D}|H_1) + \pi_{2j}\Pr(\text{S or E}|H_2)\\ &= \pi_{0j}\Pr(y_j \le \Delta_L(n_j) \text{ or } y_j \ge \Delta_U(n_j)|H_0)\\ &\quad +\pi_{1j}\Pr(y_j > \Delta_L(n_j)|H_1) + \pi_{2j}\Pr(y_j < \Delta_U(n_j)|H_2), \qquad (4.1)\end{aligned}$$

where E, D and S stand for "Escalation", "De-escalation" and "Stay", respectively. We can also take a weighting scheme to penalize more for de-escalation under H_1 than staying at the same dose and for escalation under H_2 than staying. If it is further assumed that escalation is more dangerous than de-escalation, we can assign asymmetric weights or penalties for escalation and de-escalation under H_0. The ROI design aims to minimize the probability of making incorrect decisions at each step, and the optimal interval boundaries for y_j can be derived as

$$\begin{aligned}\Delta_L(n_j) &= \max\left\{m : \frac{\phi^m(1-\phi)^{n_j-m}}{\int_0^\phi p^m(1-p)^{n_j-m}f(p \mid H_1)\mathrm{d}p} \le 1\right\},\\ \Delta_U(n_j) &= \max\left\{m : \frac{\int_\phi^1 p^m(1-p)^{n_j-m}f(p \mid H_2)\mathrm{d}p}{\phi^m(1-\phi)^{n_j-m}} \le 1\right\}, \qquad (4.2)\end{aligned}$$

which in fact do not depend on y_j. We can see from (4.2) that the escalation rules of ROI are equivalent to escalating the dose if $\pi_{1j} > \pi_{0j}$, and de-escalating the dose if $\pi_{2j} > \pi_{0j}$. Let $\phi_L(n_j) = \Delta_L(n_j)/n_j$ and $\phi_U(n_j) = \Delta_U(n_j)/n_j$, which are the boundaries for the toxicity rate.

Theorem 1 *The values of $\phi_L(n_j)$ and $\phi_U(n_j)$ converge to ϕ almost surely, as $n_j \to \infty$.*

The proof of Theorem 1 is based on the consistency of the posterior probability of H_k as $n_j \to \infty$, which is straightforward and thus omitted. It indicates that the optimal interval would shrink to the target toxicity rate as the sample size increases.

Several remarks are in place for comparisons between the ROI and BOIN designs. First, π_{kj} in (4.1) of BOIN is the prior probability of each hypothesis H_k and thus BOIN is developed based on the prior information at the trial planning stage, while ROI aims to control the incorrect decisions based on the posterior distribution using the accrued information. Second, BOIN requires prespecification of ϕ_1 and ϕ_2, while there is no theoretical guidance for selection of the two values. In addition, the interpretations of ϕ_1 and ϕ_2 as well as the optimal intervals produced by BOIN are somewhat counterintuitive: ϕ_1 and ϕ_2 are claimed to be the highest and lowest toxicity rates corresponding to escalation and de-escalation respectively, but the trial is conducted using the derived optimal boundaries, which lie inside (ϕ_1, ϕ_2), instead of using ϕ_1 and ϕ_2 directly. By contrast, there is no ambiguity for ROI, since it only requires the specification of ϕ, the target toxicity rate. Last but most importantly, the

limiting interval of BOIN depends on the values of ϕ_1 and ϕ_2 and does not shrink to the target toxicity rate ϕ with an increasing sample size. As a result, BOIN would randomly locate one of the dose levels that lie inside the limiting optimal interval, while ROI converges almost surely to the true MTD because its optimal interval indeed shrinks to the target.

4.3 Multi-agent Robust Optimal Interval Design

4.3.1 Combining Two Drugs

The decisions of dose escalation, de-escalation or retention based on the ROI design only depend on the accumulative information at the current dose level, and thus can be applied to a multi-agent combination trial in a straightforward way. However, there are up to eight adjacent dose levels at a typical location in the two-dimensional dosing space and the toxicity orders are partially known. To determine an appropriate dose assignment, we propose a random walk rule to assign each new cohort of patients to the level that has the maximum posterior probability of being the MTD. More specifically, we consider combining J dose levels of drug A and K levels of drug B in a two-dimensional dose-finding study. Let p_{jk} denote the toxicity probability of the two agents at dose level $(j,k), j = 1,\ldots,J; k = 1,\ldots,K$. Suppose the current dose combination level is (j,k), and we define an admissible escalation set as

$$\mathscr{A}_E = \{(j+1,k),(j,k+1)\},$$

and an admissible de-escalation set as

$$\mathscr{A}_D = \{(j-1,k),(j,k-1)\},$$

as shown in Fig. 4.1. The admissible dose escalation/de-escalation set only contains the dose levels by upgrading or downgrading one drug by one dose level while fixing the level of the other drug. We exclude the dose levels that are out of the dose range from the admissible dose escalation/de-escalation set. For example, when $j = 1$, the dose level $(j-1,k)$ should be excluded from the dose de-escalation set. The random-walk robust optimal interval (RW-ROI) design begins with treating the first cohort at the lowest dose combination $(1,1)$. Based on the cumulative number of DLTs observed at dose level (j,k), y_{jk}, the dose level for the next cohort of patients is determined as follows:

1. If $y_{jk} \leq \Delta_L(n_{jk})$, escalate to the dose level in the admissible escalation set, which has the largest posterior probability $\Pr(H_0|y_{j'k'})$, $(j',k') \in \mathscr{A}_E$. If the admissible escalation set contains untried dose levels (i.e., $n_{j'k'} = 0$), we set $\Pr(H_0|y_{j'k'}) = 1$, which thus facilitates exploring the untried dose levels as well as preventing the trial from being trapped in some suboptimal doses.

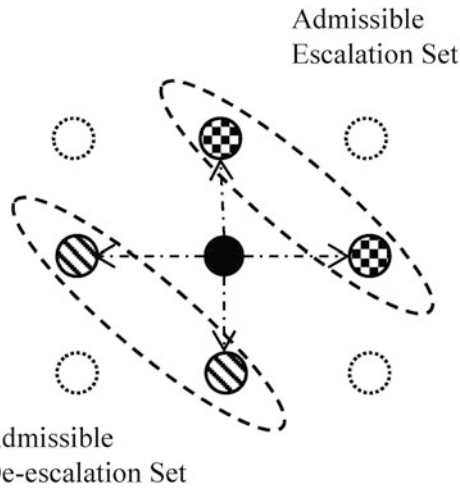

Fig. 4.1 Admissible sets for dose escalation or de-escalation in the RW-ROI design

2. If $y_{jk} \geq \Delta_U(n_{jk})$, de-escalate to the dose level in the admissible de-escalation set, which has the largest posterior probability $\Pr(H_0|y_{j'k'})$, $(j', k') \in \mathscr{A}_D$. Similarly, we take $\Pr(H_0|y_{j'k'}) = 1$ for the untried admissible dose levels.
3. Otherwise, if $\Delta_L(n_{jk}) < y_{jk} < \Delta_U(n_{jk})$, the doses stay at the same level (j, k).

During the process of dose escalation and de-escalation, if there exist multiple optimal dose levels, we randomly choose one with equal probability. The trial continues until the total sample size is exhausted. Additionally, if the most recent patients are treated at the lowest dose level $(1, 1)$ and $y_{11} \geq \Delta_U(n_{11})$, the next dose retains at the same dose level. Symmetrically, if the current dose level is the highest dose level (J, K) and $y_{JK} \leq \Delta_L(n_{JK})$, we still treat the next cohort of patients at the same dose level.

4.3.2 Combining Three or More Drugs

Existing two-agent dose-finding methods can hardly be extended to the cases with three or more drugs combined. By contrast, the proposed random walk rule is suitable for any arbitrary number of drugs. For illustration, we consider a three-dimensional dose-finding study that combines J dose levels of drug A, K levels of drug B and L levels of drug C.

Suppose that y_{jkl} out of n_{jkl} patients have experienced the DLT at the current dose level (j, k, l), whose true toxicity probability is p_{jkl}. As before, we define an admissible escalation set by increasing one dose level of one drug while fixing the other two,

$$\mathscr{A}_E = \{(j + 1, k, l), (j, k + 1, l), (j, k, l + 1)\}.$$

Similarly, the admissible de-escalation set is defined by decreasing one dose level of one drug while fixing the other two,

$$\mathscr{A}_D = \{(j-1,k,l),(j,k-1,l),(j,k,l-1)\}.$$

Following similar rules as the double-agent design, the RW-ROI for a triple-agent trial begins with treating the first cohort of patients at the lowest dose combination $(1,1,1)$. Based on the cumulative number of DLTs observed at dose level (j,k,l), y_{jkl}, the dose level for the next cohort is determined as follows:

1. If $y_{jkl} \leq \Delta_L(n_{jkl})$, escalate to the dose level in the admissible escalation set, which has the largest posterior probability $\Pr(H_0|y_{j'k'l'})$, $(j',k',l') \in \mathscr{A}_E$. If the admissible escalation set contains untried dose levels (i.e., $n_{j'k'l'} = 0$), we set $\Pr(H_0|y_{j'k'l'}) = 1$, which thus facilitates exploring the untried dose levels as well as preventing the trial from being trapped in some suboptimal doses.
2. If $y_{jkl} \geq \Delta_U(n_{jkl})$, we de-escalate to the dose level that lies inside the admissible de-escalation set and also has the largest posterior probability $\Pr(H_0|y_{j'k'l'})$, $(j',k',l') \in \mathscr{A}_D$. Similarly, we take $\Pr(H_0|y_{j'k'l'}) = 1$ for the untried admissible dose levels.
3. Otherwise, if $\Delta_L(n_{jkl}) < y_{jkl} < \Delta_U(n_{jkl})$, the doses stay at the same level (j,k,l).

The trial continues until the total sample size is exhausted. During the process of dose escalation and de-escalation, if there exist multiple optimal dose levels, we randomly choose one with equal probability.

Such an algorithm can be straightforwardly extended to the drug-combination trial with more than three drugs. Suppose the current dose level is $(j,k,l,\ldots,)$, and then the admissible escalation set is

$$\mathscr{A}_E = \{(j+1,k,l,\ldots),(j,k+1,l,\ldots),(j,k,l+1,\ldots),\ldots\},$$

and the admissible de-escalation set is

$$\mathscr{A}_D = \{(j-1,k,l,\ldots),(j,k-1,l,\ldots),(j,k,l-1,\ldots),\ldots\}.$$

The dose-finding rules remain unchanged.

After the trial by RW-ROI is completed, we perform the isotonic regression so that the estimated toxicity rates satisfy partial ordering of the toxicity rates when allowing only the dose of one drug to change and fixing the other drugs at certain levels. Specifically, in a three-agent trial, we perform three-dimensional isotonic regression (Dykstra and Robertson 1982) to the estimated toxicity rate $\hat{p}_{jkl}$, and let $\tilde{p}_{jkl}$ denote the trivariate isotonic regression estimator. The MTD $(j^\dagger,k^\dagger,l^\dagger)$ is finally selected as the dose level whose toxicity rate $\tilde{p}_{j^\dagger k^\dagger l^\dagger}$ is closest to the target ϕ:

$$(j^\dagger,k^\dagger,l^\dagger) = \arg\min_{(j,k,l)\in\mathscr{N}} |\tilde{p}_{jkl} - \phi|,$$

where the set $\mathcal{N} = \{(j, k, l) : n_{jkl} > 0\}$ contains all the tested dose levels.

When there are ties for $\tilde{p}_{j^\dagger k^\dagger l^\dagger}$ on the same row, the same column, or the same layer, the highest dose combination satisfying $\tilde{p}_{j^\dagger k^\dagger l^\dagger} < \phi$, or the lowest dose combination satisfying $\tilde{p}_{j^\dagger k^\dagger l^\dagger} > \phi$, is finally selected as the MTD. However, it is difficult to distinguish the ties when they lie on different rows, columns, or layers, e.g., $(j + 1, k - 1, l)$ and $(j - 1, k + 1, l)$. In this case, we select the one that has the largest value of $\Pr(H_0|y_{j^\dagger k^\dagger l^\dagger})$, which is approximately equivalent to the dose combination that has been tested with more patients.

Similar to the single-agent ROI design, the RW-ROI design has desirable large-sample properties. Based on the accrued information in the trial, it can be shown that the estimates of the posterior probabilities $\Pr(H_0|y_{jkl})$ in the RW-ROI design would converge to their true values (either 1 or 0). Thus, RW-ROI would adaptively assign patients to the dose level that is closer to the MTD instead of being trapped in a local neighborhood, and the dose assignment converges to the MTD.

4.4 Simulation Study

4.4.1 Single-Agent ROI Versus BOIN

First, we conduct a simulation study of the single-agent ROI design with a comparison to the BOIN design in terms of the operating characteristics. The trial under consideration consists of eight dose levels with a target toxicity rate $\phi = 0.3$. The total sample size planned is 30 and patients are assigned in cohorts of size 3. For the BOIN method, we consider three paired values for (ϕ_1, ϕ_2): the default interval $(\phi_1, \phi_2) = (0.6\phi, 1.4\phi)$ recommended by Liu and Yuan (2015), the narrow interval $(\phi_1, \phi_2) = (0.8\phi, 1.2\phi)$, and the wide interval $(\phi_1, \phi_2) = (0.5\phi, 1.5\phi)$. In addition, we impose a safety constraint by setting $\lambda = 0.95$ for the two methods.

Table 4.1 shows the simulation results under three toxicity scenarios and each scenario is replicated for 1000 times. In scenario 1, the seventh dose is the MTD, and the MTD selection probabilities of BOIN using the three intervals are very different. In particular, the BOIN design with the narrow interval has the lowest selection percentage, while that based on the default interval is the best. Under scenario 2, all three BOIN designs perform similarly, while the default one behaves slightly poor. For scenario 3, the default and wide interval BOIN designs perform much worse than the narrow interval BOIN. By contrast, the proposed ROI design does not depend on any extra design parameters, such as ϕ_1 and ϕ_2, and it tends to perform comparably with the best of the three BOIN designs under each scenario. These findings suggest that the prespecified interval indeed plays a critical role in the BOIN design, and the performance could be much compromised if the interval is chosen inappropriately. However, it is difficult, if not impossible, to justify which interval is more sensible in the trial planning stage.

Table 4.1 Comparison between BOIN and ROI for single-agent trials under three toxicity scenarios with a target toxicity rate of 0.3

	Recommendation percentage at dose level								Average	Average
Design	1	2	3	4	5	6	7	8	# DLTs	# patients
Scenario 1	0.01	0.02	0.03	0.05	0.08	0.13	**0.30**	0.50		
BOIN(0.6,1.4)	0.0	0.0	0.0	0.3	2.0	25.3	**53.0**	19.4	3.9	30.0
# patients	3.1	3.2	3.3	3.7	4.1	5.3	5.3	2.1		
BOIN(0.8,1.2)	0.0	0.0	0.1	1.3	5.5	38.5	**40.5**	14.1	3.1	30.0
# patients	3.3	3.5	3.8	4.4	4.8	5.7	3.1	1.4		
BOIN(0.5,1.5)	0.0	0.0	0.0	0.6	3.3	27.9	**50.8**	17.4	3.8	30.0
# patients	3.1	3.2	3.4	3.8	4.2	5.3	5.1	2.0		
ROI	0.0	0.0	0.0	0.8	3.9	25.1	**52.8**	17.4	3.8	30.0
# patients	3.1	3.2	3.4	3.8	4.1	5.1	5.3	2.1		
Scenario 2	0.15	**0.30**	0.42	0.55	0.65	0.68	0.70	0.80		
BOIN(0.6,1.4)	23.9	**51.8**	21.0	2.5	0.0	0.0	0.0	0.0	8.4	29.8
# patients	10.1	12.8	5.7	1.1	0.1	0.0	0.0	0.0		
BOIN(0.8,1.2)	20.9	**55.0**	20.5	2.8	0.2	0.0	0.0	0.0	7.8	29.9
# patients	12.6	11.9	4.5	0.9	0.0	0.0	0.0	0.0		
BOIN(0.5,1.5)	25.3	**55.1**	16.6	2.2	0.0	0.0	0.0	0.0	8.1	29.8
# patients	10.7	13.2	5.0	0.9	0.0	0.0	0.0	0.0		
ROI	23.1	**54.6**	18.8	2.7	0.0	0.0	0.0	0.0	8.4	29.8
# patients	9.6	13.5	5.6	1.0	0.1	0.0	0.0	0.0		
Scenario 3	0.10	0.15	0.22	**0.30**	0.38	0.46	0.55	0.60		
BOIN(0.6,1.4)	1.6	10.5	31.0	**32.4**	17.6	5.8	0.9	0.0	6.6	30.0
# patients	4.8	6.7	8.3	6.3	2.8	0.9	0.1	0.0		
BOIN(0.8,1.2)	1.5	13.5	29.1	**38.8**	13.4	2.9	0.4	0.0	5.8	29.9
# patients	6.8	8.5	7.6	5.1	1.5	0.4	0.0	0.0		
BOIN(0.5,1.5)	2.5	12.0	33.3	**31.8**	15.3	4.3	0.6	0.0	6.4	30.0
# patients	5.2	7.3	8.4	5.8	2.4	0.7	0.1	0.0		
ROI	2.8	11.6	32.1	**33.5**	14.5	4.6	0.7	0.0	6.5	30.0
# patients	5.1	7.0	8.4	6.1	2.5	0.7	0.1	0.0		

BOIN stands for the Bayesian optimal interval design, and the values in the parentheses are the prespecified design parameters ϕ_1 and ϕ_2 in BOIN; ROI is the proposed robust optimal interval design

4.4.2 Double-Agent RW-ROI Versus Model-Based Designs

For dose finding with two drugs in combination, we investigate the performance of the proposed RW-ROI design with comparisons to four existing model-based methods that are described as follows:

(1) Two-dimensional escalation with overdose control (TEWOC): Shi and Yin (2013) proposed a TEWOC design for dose finding on the basis of a four-parameter logistic regression model, under which the joint toxicity probability

at dose level (j, k) is given by

$$p_{jk} = \frac{\exp(\beta_0 + \beta_1 d_j^A + \beta_2 d_k^B + \beta_3 d_j^A d_k^B)}{1 + \exp(\beta_0 + \beta_1 d_j^A + \beta_2 d_k^B + \beta_3 d_j^A d_k^B)}, \tag{4.3}$$

where d_j^A and d_k^B are the dosages of the two agents in combination. The assignment of the next dose level is based on the estimated MTD distribution with respect to a prespecified quantile level a, which is set as $a = 0.25$.

In the simulation study, we consider $(d_1^A, d_2^A, d_3^A) = (0.1, 0.2, 0.3)$ and $(d_1^B, \ldots, d_5^B) = (0.1, 0.2, 0.3, 0.4, 0.5)$ and assign noninformative priors to the unknown parameters: $\beta_0 \sim \mathrm{N}(0, 2)$, $\beta_1, \beta_2, \beta_3 \sim \mathrm{Gamma}(4, 0.8)$.

(2) Copula-type method: To model the toxicity surface, Yin and Yuan (2009) proposed a copula-type regression method to capture drug–drug interactions. Specifically, they used a Clayton copula regression function, which is given by

$$p_{jk} = 1 - \{(1 - a_j^\alpha)^{-\gamma} + (1 - b_k^\beta)^{-\gamma} - 1\}^{-1/\gamma},$$

where $\alpha, \beta, \gamma > 0$ are unknown parameters, and a_j $(j = 1, \ldots, J)$ and b_k $(k = 1, \ldots, K)$ are the prespecified toxicity probabilities for each dose level of drug A and drug B, respectively. The dose escalation decision is based on the posterior probability given the cumulative data D, $\Pr(p_{jk} < \phi | D)$, and two prespecified cutoffs c_e and c_d: if $\Pr(p_{jk} < \phi | D) > c_e$, the dose is escalated to an adjacent dose combination with its toxicity rate higher than the current value as well as closest to the target rate; similarly, if $\Pr(p_{jk} < \phi | D) < c_d$, the dose is de-escalated for the next cohort of patients; otherwise, the current dose combination stays at the same level. We set $c_e = 0.7$ and $c_d = 0.55$ to direct dose escalation and de-escalation, respectively.

We take an even partition from 0 to 0.3 for both a_j's and b_k's: $(a_1, a_2, a_3) = (0.1, 0.2, 0.3)$, $(b_1, \ldots, b_5) = (0.06, 0.12, 0.18, 0.24, 0.3)$. We specify $\mathrm{Gamma}(2, 2)$ as the prior distribution for α and β, and a relatively noninformative $\mathrm{Gamma}(0.1, 0.2)$ as the prior distribution for γ.

(3) Log-linear model: Wang and Ivanova (2005) utilized a log-linear working model for the dose–toxicity relationship in drug-combination trials:

$$p_{jk} = 1 - (1 - a_j^\alpha)(1 - b_k^\beta) \exp\{-\gamma \log(1 - a_j) \log(1 - b_k)\},$$

where $\alpha, \beta, \gamma > 0$. For comparison, all the trial specifications under the log-linear model are identical to those in the copula-type method.

(4) Logistic model: We make a further comparison of the proposed method with the logistic model in (4.3) while keeping the dose allocation rule the same as the copula-type method.

Table 4.2 Ten toxicity scenarios for two-drug combinations with a target toxicity probability of 30 %. The MTDs are in boldface

Drug B											
	Dose level	1	2	3	4	5	1	2	3	4	5
Durg A		Scenario 1					Scenario 2				
	3	0.15	**0.30**	0.45	0.50	0.65	**0.30**	0.50	0.60	0.65	0.75
	2	0.10	0.15	**0.30**	0.45	0.50	0.15	**0.30**	0.45	0.52	0.60
	1	0.05	0.10	0.15	**0.30**	0.45	0.05	0.10	0.12	0.15	**0.30**
		Scenario 3					Scenario 4				
	3	0.10	0.15	**0.30**	0.45	0.55	0.12	0.15	0.17	**0.30**	0.50
	2	0.06	0.10	0.15	**0.30**	0.45	0.06	0.08	0.15	0.20	0.45
	1	0.04	0.06	0.10	0.15	**0.30**	0.02	0.06	0.10	0.15	**0.30**
		Scenario 5					Scenario 6				
	3	0.40	0.42	0.48	0.55	0.60	0.15	**0.30**	0.45	0.50	0.60
	2	**0.30**	0.40	0.43	0.48	0.55	0.08	0.12	0.15	**0.30**	0.45
	1	0.15	**0.30**	0.40	0.45	0.50	0.04	0.06	0.10	0.12	0.15
		Scenario 7					Scenario 8				
	3	0.50	0.60	0.70	0.75	0.80	0.08	0.15	0.45	0.60	0.70
	2	0.10	**0.30**	0.45	0.60	0.70	0.05	0.20	**0.30**	0.45	0.70
	1	0.06	0.10	0.15	**0.30**	0.40	0.02	0.10	0.15	0.40	0.50
		Scenario 9					Scenario 10				
	3	0.15	**0.30**	0.40	0.60	0.70	0.70	0.75	0.80	0.85	0.90
	2	0.02	0.05	0.08	0.12	0.15	0.45	0.55	0.60	0.65	0.70
	1	0.01	0.02	0.03	0.04	0.10	0.05	0.08	0.20	**0.30**	0.40

We compare the RW-ROI design with the four model-based methods in terms of the operating characteristics under the 10 scenarios in Table 4.2, which involves various numbers and locations of the MTDs. We take the maximum sample size to be 60 with a cohort size of 3, and the target toxicity probability is set at 0.3. To ensure comparability across different methods, we do not impose any early stopping rule so that we run the entire trial till the exhaustion of the maximum sample size. We simulate 1000 replications for each scenario.

Table 4.3 presents the simulation results of our proposed RW-ROI design in conjunction with those of existing model-based methods, which include three performance statistics: the percentage of MTD selection, the percentage of patients allocated at the MTDs, and the average number of DLTs. Among the four model-based designs considered, the logistic method performs the best, with their MTD selection and patient allocation percentages substantially greater than those of the other model-based designs under scenarios 1, 3, 6, and 8. The log-linear method and Clayton method have similar performance under the 10 scenarios. The performance of the model-based methods is sensitive to the MTD locations. For example,

Table 4.3 Comparisons of the proposed two-dimensional RW-ROI design with the model-based methods under ten scenarios with a target toxicity rate of 0.3. The best performance statistics are in boldface

Method	Scenarios									
	1	2	3	4	5	6	7	8	9	10
	Percentage of MTD selections									
TEWOC	64.4	44.4	73.5	45.8	60.2	38.8	**59.4**	31.5	12.8	**39.3**
Logistic	**74.7**	52.7	**76.2**	43.0	63.8	**56.4**	57.1	**43.8**	5.1	22.6
Log-linear	59.1	56.8	59.1	32.7	**64.2**	47.0	52.1	27.3	15.8	32.5
Clayton	58.9	50.3	59.6	32.8	63.1	40.0	47.0	24.1	17.4	28.2
RW-ROI	63.4	**66.2**	67.9	**52.5**	55.3	51.7	58.2	31.5	**35.2**	33.5
	Percentage of patients allocated at the MTDs									
TEWOC	47.8	36.1	42.8	26.1	**57.7**	31.3	36.7	24.6	7.8	**23.4**
Logistic	**48.2**	32.4	**47.3**	25.4	43.2	**32.6**	35.4	**27.2**	4.6	12.9
Log-linear	44.7	36.1	42.6	24.0	50.4	30.3	**41.8**	16.7	13.1	17.4
Clayton	40.1	37.6	40.7	25.4	43.3	28.8	34.4	16.5	14.2	9.6
RW-ROI	42.5	**42.6**	40.2	**26.4**	43.6	31.9	37.3	19.2	**21.8**	15.2
	Average number of DLTs									
TEWOC	**15.6**	17.4	**14.2**	**13.2**	18.5	15.4	18.1	16.0	16.1	19.8
Logistic	17.3	17.4	16.4	15.7	19.0	16.5	17.8	17.3	16.3	17.8
Log-linear	16.2	16.3	15.2	14.6	19.4	**14.9**	**16.5**	16.4	**14.7**	17.3
Clayton	16.1	**15.9**	15.4	15.0	**17.2**	15.2	**16.5**	**15.2**	15.3	**16.6**
RW-ROI	16.6	17.4	15.2	14.8	18.7	15.4	18.2	16.9	14.9	18.9

TEWOC stands for the two-dimensional escalation with overdose control method, and RW-ROI represents the random-walk robust optimal interval design

under scenarios 1–3 where three MTDs exist, the MTD selection percentages under TEWOC and logistic methods have an over 20 % range of variations due to different MTD locations. Similarly, the selection percentages of the four model-based methods vary from 32 % to 64 % under scenarios 4–7 which have two MTDs. These findings demonstrate that the mode-based methods are not robust. In addition, we find that the TEWOC and logistic model are also sensitive to the design calibration parameters. By contrast, the MTD selection percentages based on the RW-ROI design is more stable with respect to various toxicity scenarios. For the first seven scenarios, RW-ROI design has an average selection percentage of 60 %, with improvement between 5 % to 20 % over the log-linear and Clayton methods. Scenarios 2, 4, 7 are difficult ones because their toxicity surfaces are quite irregular, for which the proposed RW-ROI design has the best performance among all the methods. When the MTD is unique in scenarios 8–10, the proposed design is also comparable with the model-based methods. Similar conclusions can be made with respect to the percentage of patients allocated at the MTDs. The five designs have similar operating characteristics in terms of the average number of DLTs, .

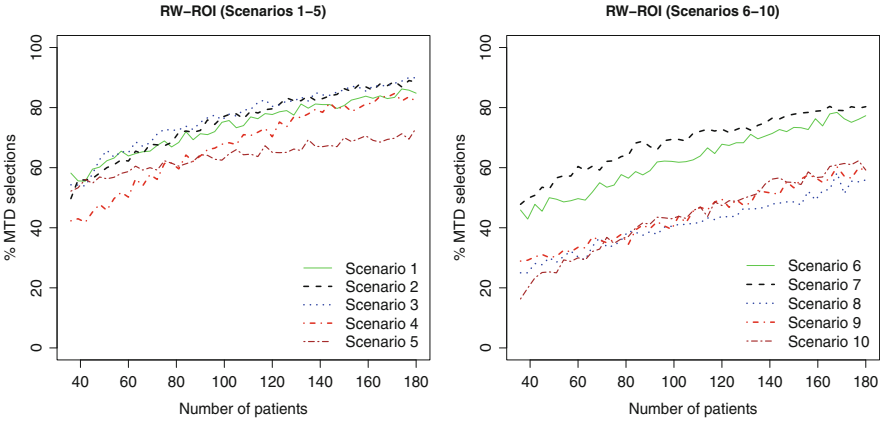

Fig. 4.2 Relationship between the sample size and the percentage of MTD selection under the 10 scenarios in Table 4.2

To examine the limiting performance of the proposed methods, we increase the maximum sample size of the simulated trials. Figure 4.2 presents the trends of the percentages of MTD selection with respect to an increasing sample size under the ten scenarios in Table 4.2. Clearly, the performance of the proposed design continuously improves by accumulating more data and would not be trapped in any suboptimal doses. In general, the more MTDs in the two-dimensional space, the higher the selection percentage. In scenarios 2 and 3, the percentages of MTD selection increase from 40 % to 80 % as the sample size is enlarged from 40 to 180. In scenarios 8 and 10, where only one MTD exists, the MTD selection percentages using RW-ROI can still improve substantially as the sample size increases.

4.4.3 Triple-Agent RW-ROI

To investigate the operating characteristics of RW-ROI in multi-agent combination trials, we expand the dosing space to three dimensions. Specifically, we consider four dose levels of drug A, three levels of drug B, and two levels of drug C. Therefore, there are 24 combination dose levels in total. Two scenarios with a target toxicity rate of 0.3 are provided in Table 4.4, where four MTDs exist under each scenario. The sample size is 90 patients with 3 patients in a cohort. Based on 5000 replications, the percentage of MTD selection for scenario 1 is 62.2 %, and on average 41.4 % patients are allocated to the MTDs by RW-ROI. For scenario 2, RW-ROI also achieves a 62.4 % correct selection percentage, and allocates 38.1 % of the patients to the MTDs. The average numbers of DLTs are 27.8 and 26.3 under scenarios 1 and 2, respectively, which are very close to the expectation as if all the 90

Table 4.4 Two toxicity scenarios for three-drug combinations with a target toxicity probability of 30 %. The MTD is in boldface

		Drug A							
	Dose level	1	2	3	4	1	2	3	4
Durg B		Scenario 1				Scenario 2			
		Drug C: Level 1				Drug C: Level 1			
	3	0.15	**0.30**	0.45	0.60	0.12	0.15	**0.30**	0.50
	2	0.05	0.15	**0.30**	0.45	0.06	0.08	0.20	0.45
	1	0.01	0.05	0.10	0.15	0.02	0.06	0.15	**0.30**
		Drug C: Level 2				Drug C: Level 2			
	3	0.45	0.55	0.65	0.80	0.17	**0.30**	0.45	0.65
	2	0.15	**0.30**	0.45	0.65	0.15	0.18	**0.30**	0.55
	1	0.05	0.15	**0.30**	0.45	0.10	0.15	0.18	0.45

patients were allocated to the MTDs (with the target toxicity rate of 0.3). The triple-agent simulation results demonstrate that RW-ROI also has a desirable and robust performance in multi-agent combination trials. With an even higher dimension, RW-ROI is expected to still perform well and its implementation is simple and straightforward.

4.5 Illustrative Example

4.5.1 Prostate Cancer Trial

For patients with metastatic hormone-refractory prostate cancer, mitoxantrone has been demonstrated to be an active agent, but its prostate-specific antigen response rate is low. Genasense is a phosphorothioate antisense oligonucleotide complementary to the bcl-2 mRNA open reading frame, which contributes to inhibiting expression of bcl-2, delaying androgen independence as well as enhancing chemosensitivity in prostate and other cancer models. As a result, a phase I dose-finding study of combined treatment with mitoxantrone and genasense is considered to meet the need for more effective treatment of the prostate cancer (Chi et al. 2001). The goals of the trial were to evaluate the safety and biological effect of the combination of genasense and mitoxantrone, and to determine the preliminary antitumor activity. Specifically, three doses (4, 8, and 12 mg/m^2) of mitoxantone and five doses (0.6, 1.2, 2.0, 3.1, 5.0 mg/kg) of genasense were investigated in this trial. To identify the MTD combination, the trial selected seven combination doses: (mitoxantone, genasense) $=$ $(4, 0.6)$, $(4, 1.2)$, $(4, 2.0)$, $(4, 3.1)$, $(8, 3.1)$, $(12, 3.1)$, $(12, 5.0)$, and applied the modified $3 + 3$ dose escalation scheme. However, the chosen dose pairs from the two-dimensional space are arbitrary, so that the true MTD might have been excluded. Due to the limitation of the $3 + 3$ design, only

one MTD can be identified in the trial, even though multiple MTDs may exist in the drug-combination space. In addition, the $3 + 3$ design does not even guarantee the recommended MTD is correct. This example demonstrates the need for a more effective dose-finding design in drug-combination trials.

4.5.2 Trial Illustration

For illustration, we apply the proposed RW-ROI design to the aforementioned prostate cancer trial. As described previously, the trial examined 3 dose levels of mitoxantrone and 5 dose levels of genasense, which results in a 3×5 drug-combination space. The target toxicity rate is $\phi = 0.3$, the cohort size is set as 3 and 20 cohorts are planned for the trial. Based on the formulae in (4.2), the optimal boundaries for the RW-ROI design are given in Table 4.5.

In addition, we impose a safety rule by setting the threshold $\lambda = 0.95$. The first cohort of patients is treated at the lowest dose level $(1, 1)$. Figure 4.3 shows the path of the dose assignments for the subsequent cohorts, from which we can see that the RW-ROI design can search the MTD adaptively and treat most of the patients at the right dose level. Specially, three DLTs are observed for the 8th cohort at dose level $(3, 3)$, which is beyond the dose elimination cutoff. Therefore, the dose level $(3, 3)$ and the higher dose combinations are eliminated from the trial, and dose de-escalation should be made for the next cohort. Note that the admissible de-escalation set is $\{(3, 2), (2, 3)\}$ while dose level $(3, 2)$ has never been administrated before, so the RW-ROI design selects dose level $(3, 2)$ for the next assignment. In addition, dose-escalation for the 14th cohort is based on comparison between the posterior probabilities $\Pr(H_0|y_{23})$ and $\Pr(H_0|y_{32})$, and finally chooses dose level $(2, 3)$. At the

Table 4.5 Interval boundaries and dose elimination cutoffs for the number of DLTs in the robust optimal interval design with a target toxicity rate $\phi = 0.3$

n_j	3	6	9	12	15	18	21	24	27	30
$\Delta_L(n_j)$	0	1	1	2	2	3	4	4	5	6
$\Delta_U(n_j)$	2	4	5	6	7	9	10	11	12	14
$\Delta_T(n_j)$	3	4	5	7	8	9	10	11	12	14
n_j	33	36	39	42	45	48	51	54	57	60
$\Delta_L(n_j)$	6	7	8	8	9	10	11	11	12	13
$\Delta_U(n_j)$	15	16	17	18	19	20	21	22	23	24
$\Delta_T(n_j)$	15	16	17	18	19	20	21	22	23	24

Note: n_j is the cumulative number of patients at dose level j

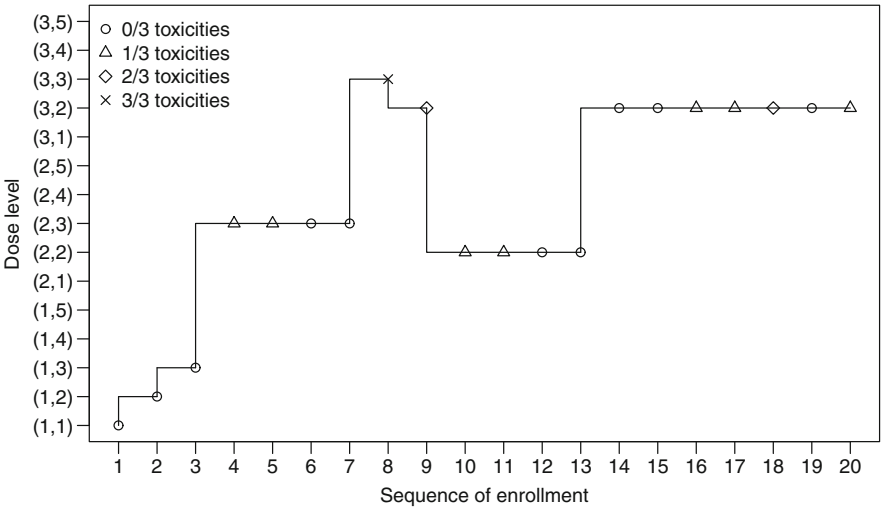

Fig. 4.3 Illustration of RW-ROI for the prostate cancer trial with a target toxicity rate of 0.3. *Circle* indicates patients without toxicity, *triangle* and *diamond* respectively denote one and two toxicities, and *cross* represents three toxicities

end of the trial, the estimated toxicity probability matrix after implementing the two-dimensional pool adjacent violators algorithm is given by

$$\begin{bmatrix} - & 0.68 & 1 & - & - \\ - & 0.15 & 0.15 & \mathbf{0.28} & - \\ 0 & 0 & 0 & - & - \end{bmatrix},$$

where "–" represents the dose levels that have not been administered in the trial. Thus, dose level $(2, 4)$, which is $8\,\text{mg/m}^2$ mitoxantone combined with $3.1\,\text{mg/kg}$ genasense, would be selected as the MTD.

4.6 Concluding Remarks

To simplify dose-finding procedure while still maintaining the trial performance, a robust optimal interval design is developed. Its extension to double-, triple-, or higher-dimensional drug combinations is straightforward, which greatly simplifies the current practice of dose finding in combination treatment. The proposed ROI and RW-ROI methods can substantially outperform the BOIN design, if the interval parameters (ϕ_1, ϕ_2) of BOIN is poorly specified. The ROI design only requires the prespecification of the target toxicity rate of the trial and thus it dramatically improves the robustness of the existing dose-finding designs. In addition, we have

demonstrated the good performance and operating characteristics of all the single-, double- and triple-agent ROI designs by conducting extensive simulation studies. An arguable point is that the allocation decisions by the ROI designs are solely determined by the information at the current dose level, while the sequential dose-escalation procedure as well as the isotonic regression at the end of the trial implicitly account for the majority of information from other doses.

Acknowledgements We are grateful to a co-editor for many helpful suggestions that have improved this chapter immensely. The research was supported in part by a grant (17125814) from the Research Grants Council of Hong Kong.

References

Ahn C (1998) An evaluation of phase I cancer clinical trial designs. Stat Med 17:1537–1549

Chi KN, Gleave ME, Klasa R, Murray N, Bryce C, de Menezes DEL, D′Aloisio S, Tolcher AW (2001) A phase I dose-finding study of combined treatment with an antisense bcl-2 oligonucleotide (genasense) and mitoxantrone in patients with metastatic hormone-refractory prostate cancer. Clin Cancer Res 7:3920–3927

Conaway MR, Dunbar S, Peddada SD (2004) Designs for single- or multiple-agent phase I trials. Biometrics 60:661–669

Durham SD, Flournoy N, Rosenberger WF (1997) A random walk rule for phase I clinical trials. Biometrics 53:745–760

Dykstra RL, Robertson T (1982) An algorithm for isotonic regression for two or more independent variables. Ann Stat 10:708–716

Fan SK, Venook AP, Lu Y (2009) Design issues in dose-finding phase I trials for combinations of two agents. J Biopharm Stat 19:509–523

Gezmu M, Flournoy N (2006) Group up-and-down designs for dose-finding. J Stat Plan Infer 136:1749–1764

Heyd JM, Carlin PB (1999) Adaptive design improvements in the continual reassessment method for phase I studies. Stat Med 18:1307–1321

Huang X, Biswas S, Oki Y, Issa JP, Berry DA (2007) A parallel phase I/II clinical trial design for combination therapies. Biometrics 63:429–436

Ivanova A, Wang K (2004) A non-parametric approach to the design and analysis of two-dimensional dose-finding trials. Stat Med 23:1861–1870

Ivanova A, Flournoy N, Chung Y (2007) Cumulative cohort design for dose finding. J Stat Plan Infer 137:2316–2327

Ji Y, Li Y, and Yin G (2007) Bayesian dose finding in phase I clinical trials based on a new statistical framework. Stat Sinica 17:531–547

Ji Y, Liu P, Li Y, Bekele BN (2010) A modified toxicity probability interval method for dose-finding trials. Clin Trials 7:653–663

Leung DHY, Wang YG (2002) An extension of the continual reassessment method using decision theory. Stat Med 21:51–63

Lin R, Yin G (2016, in press) Bayesian optimal interval design for dose finding in drug-combination trials. Stat Methods Med Res. doi: 10.1177/0962280215594494

Liu S, Cai C, Ning J (2013) Up-and-down designs for phase I clinical trials. Contemp Clin Trials 36:218–227

Liu S, Yuan Y (2015) Bayesian optimal interval designs for phase I clinical trials. J R Stat Soc Ser C Appl Stat 64:507–523

Oron A, Azriel D, Hoff P (2011) Dose-finding designs: the role of convergence properties. Int J Biostat 7, Article 39

O'Quigley J, Pepe M, Fisher L (1990) Continual reassessment method: a practical design for phase 1 clinical trials in cancer. Biometrics 46:33–48

Rogatko A, Schoeneck D, Jonas W, Tighiouart M, Khuri FR, Porter A (2007) Translation of innovative designs into phase I trials. J Clin Oncol Res 25:4982–4986

Shi Y, Yin G (2013) Escalation with overdose control for phase I drug-combination trials. Stat Med 32:4400–4412

Simon R, Rubinstein L, Arbuck SG, Christian MC, Freidlin B, Collins J (1997) Accelerated titration designs for phase I clinical trials in oncology. J Natl Cancer Inst 89:1138–1147

Storer BE (1989) Design and analysis of phase I clinical trials. Biometrics 45:925–937

Thall PF, Millikan RE, Müller P, Lee SJ (2003) Dose-finding with two agents in phase I oncology trials. Biometrics 59:487–496

Wang K, Ivanova A (2005) Two-dimensional dose finding in discrete dose space. Biometrics 61:217–222

Yuan Z, Chappell R (2004) Isotonic designs for phase I cancer clinical trials with multiple risk groups. Clin Trials 1:499–508

Yin G (2012) Clinical trial design: bayesian and frequentist adaptive methods. John Wiley & Sons, Hoboken

Yin G, Yuan Y (2009) Bayesian model averaging continual reassessment method in phase I clinical trials. J Am Stat Assoc 104:954–968

Yin G, Yuan Y (2009) Bayesian dose finding in oncology for drug combinations by copula regression. J R Stat Soc Ser C Appl Stat 61:217–222

Yin G, Yuan Y (2009) A latent contingency table approach to dose finding for combinations of two agents. Biometrics 65:866–875

Yin G, Lin R (2014) Comments on "Competing designs for drug combination in phase I dose-finding clinical trials" In: Riviere M-K, Dubois F, Zohar S (eds) Stat Med 34:13–17

Yuan Z, Chappell R, Bailey H (2007) The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. Biometrics 63:173–179

Yuan Y, Yin G (2008) Sequential continual reassessment method for two-dimensional dose finding. Stat Med 27:5664–5678

Part II
Life Time Data Analysis

Chapter 5
Group Selection in Semiparametric Accelerated Failure Time Model

Longlong Huang, Karen Kopciuk, and Xuewen Lu

Abstract In survival analysis, a number of regression models can be used to estimate the effects of covariates on the censored survival outcome. When covariates can be naturally grouped, group selection is important in these models. Motivated by the group bridge approach for variable selection in a multiple linear regression model, we consider group selection in a semiparametric accelerated failure time (AFT) model using Stute's weighted least squares and a group bridge penalty. This method is able to simultaneously carry out feature selection at both the group and within-group individual variable levels, and enjoys the powerful oracle group selection property. Simulation studies indicate that the group bridge approach for the AFT model can correctly identify important groups and variables even with high censoring rate. A real data analysis is provided to illustrate the application of the proposed method.

5.1 Introduction

Variable selection, an important objective of survival analysis, is to choose a minimum number of important variables to model the relationship between a lifetime response and potential risk factors. In an attempt to select significant variables and estimate regression coefficients automatically and simultaneously, a family of penalized or regularized approaches is proposed. Variable selection is conducted by minimizing a penalized objective function by adding a penalty

L. Huang (✉) • X. Lu
Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, T2N 1N4, Calgary, AB, Canada
e-mail: lohuang@ucalgary.ca; lux@math.ucalgary.ca

K. Kopciuk
Department of Cancer Epidemiology and Prevention Research, Alberta Health Services, 5th Floor, Holy Cross Centre Box ACB, 2210 2 St. SW, T2S 3C3, Calgary, AB, Canada
e-mail: karen.kopciuk@albertahealthservices.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_5

function with the following form

$$\min \{\text{Loss function} + \text{Penalty}\}.$$

The popular choices of loss functions are least squares and negative log-likelihood. Many different penalty functions have been used for penalized regression, such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), the bridge penalty (Fu 1998), the smoothly clipped absolute deviation (SCAD) method (Fan and Li 2001), the elastic-net method (Zou and Hastie 2005), the minimax concave penalty (MCP) (Zhang 2010) and the smooth integration of counting and absolute deviation (SICA) method (Lv and Fan 2009). These methods are designed for individual variables selection.

In many applications, covariates in $\boldsymbol{X}$ are grouped. For example, in multi-factor analysis of variance (ANOVA) problem, in which each factor may have several levels and can be expressed through a group of dummy variables, such as for response Z with two factors α and β, the intercept μ and the random error ε,

$$Z = \mu + \alpha_j + \beta_k + \varepsilon, \qquad j = 1, \ldots, J, \;\; k = 1, \ldots, K,$$

where $\{\alpha_j\}_{j=1}^{J}$ and $\{\beta_k\}_{k=1}^{K}$ can be considered as two groups. Another example is the additive model with polynomial or nonparametric components, where each component in the additive model may be expressed as a linear combination of a number of basis functions of the original measured variable, for example,

$$Z = \mu + \phi_1(W_1) + \cdots + \phi_J(W_J) + \varepsilon,$$

where each function $\phi_j(W_j) = \sum_{l=1}^{m} \gamma_{lj} B_l(W_j)$, here $\{B_l(W_j)\}_l^m$ are basis functions, and considered as a group.

Ma and Huang (2007) pointed that complex diseases such as cancer are often caused by mutations in gene pathways, it would be reasonable to select groups of related genes rather than individual genes. Bakin (1999) proposed the group LASSO and a computational algorithm. Later Yuan and Lin (2006) developed this method and related group selection methods, such as group least angle regression and group nonnegative garret methods. The group LASSO is a natural extension of the LASSO, in which an L_2 norm of the coefficients associated with a group of variables is used as a component of the penalty function. Meier et al. (2008) studied the group LASSO for logistic regression. Motivated by identifying transcriptional factors that can explain the observed variation of microarray time course gene expression over time during a given biological process, Wang et al. (2007) introduced a group SCAD penalized estimation procedure for selecting variables with time-varying coefficients in the context of functional response models. Zhao et al. (2009)

introduced the Composite Absolute Penalties (CAP) family, which allows given grouping and hierarchical relationships between the predictors to be expressed. These studies only considered group selection, but did not take the individual variable selection within groups into account. Huang et al. (2009) proposed the group bridge method in a multiple linear regression model with data uncensored, which is capable of carrying out variable selection at the group and within-group individual variable levels simultaneously. Huang et al. (2014) studied the group bridge for the Cox model. Breheny and Huang (2009) developed the group MCP approach in the linear regression model with uncensored data to select important groups as well as identifying important members of these groups. They refer to this as bi-level selection.

In this paper, we consider the group bridge method for the AFT model with right censored data. The Stute's weighted least squares estimator in AFT models is introduced in Sect. 5.2. In Sect. 5.3 we describe the group bridge method for the AFT model and present the computation steps and tuning parameters selection methods. The asymptotic properties are stated in Sect. 5.4. In Sect. 5.5 simulation studies are produced to evaluate the proposed method comparing to the group LASSO method. In Sect. 5.6 we apply the proposed methods to the primary biliary cirrhosis data set. Summary and discussion are reported in Sect. 5.7.

5.2 Stute's Weighted Least Squares Estimation in The AFT Model

For $i = 1, \ldots, n$, let T_i represent the logarithm of the survival time for the i^{th} subject, $\mathbf{X}_i$ be the associated d-dimensional vector of covariates, C_i denote the logarithm of the censoring time and δ_i denote the event indicator, i.e., $\delta_i = I(T_i \leq C_i)$, which takes value 1 if the event time is observed, or 0 if the event time is censored. Define Y_i as the minimum of the logarithm of the survival time and the censoring time, i.e., $Y_i = \min(T_i, C_i)$. Then, the observed data are in the form $(Y_i, \delta_i, \mathbf{X}_i)$, $i = 1, 2, \ldots, n$, which are assumed to be an independent and identically distributed (i.i.d.) sample from $(Y, \delta, \mathbf{X})$. Survival analysis focuses on the distribution of survival times and the association between survival time and risk factors or covariates. The AFT model directly relates the logarithm of the failure time linearly to the covariates, and resembles a conventional linear model.

The AFT model is defined as

$$T_i = \alpha + \boldsymbol{X}_i^{\top}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{5.1}$$

where α is the intercept, $\boldsymbol{\beta}$ is a $d \times 1$ regression parameter vector to be estimated, and the ε_i's are independent identically distributed random errors with a common distribution function. If the distribution function of the error term is known, this

model is a parametric model. If the distribution function of error term is unspecified, this model is considered as a semiparametric model.

In order to estimate the coefficients $(\alpha, \boldsymbol{\beta})$ in the AFT model, there are three popular approaches. One is the Buckley and James estimator (1979) that adjusts censored observations using the Kaplan-Meier estimator. This original Buckley-James approach has no theoretical justification and does not provide a reliable numerical method for implementation. Later, Ritov (1990) studied the asymptotic properties of the Buckley-James estimator. The second one is the rank based estimators (Fygenson and Ritov 1994; Heller 2007; Tsiatis 1990; Ying 1993) that are motivated by the score function of the partial likelihood. The existing rank based methods are computationally intensive for semiparametric estimators.

Stute (1993) proposed a weighted least squares estimator in the semiparametric AFT model for right censored data, which uses the Kaplan-Meier weights to account for censoring in the least squares criterion in the AFT model. The weights in the equation are the jumps in the Kaplan-Meier estimator, which is computationally more feasible than the Buckley-James and rank based estimators.

Let $\hat{F}_n$ be the Kaplan-Meier estimator of the distribution function F of T and assume $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of Y_i's; $\delta_{(1)}, \ldots, \delta_{(n)}$ and $X_{(1)}, \ldots, X_{(n)}$ are the associated censoring indicators and covariates of the ordered Y_i's, respectively. According to Stute and Wang (1993) and Stute (1996), $\hat{F}_n$ can be written as $\hat{F}_n(y) = \sum_{i=1}^{n} w_{ni} 1(Y_{(i)} \leq y)$, where the w_{ni}'s are the jumps in the Kaplan-Meier estimator and can be expressed as

$$w_{n1} = \frac{\delta_{(1)}}{n},$$

$$w_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \quad i = 2, \ldots, n.$$

The w_{ni}'s are also called the Kaplan-Meier weights. Then the Stute's weighted least squares objective function is

$$\frac{1}{2} \sum_{i=1}^{n} w_{ni} \left(Y_{(i)} - \alpha - \boldsymbol{X}_{(i)}^{\top} \boldsymbol{\beta} \right)^2.$$

By centering $\boldsymbol{X}_{(i)}$ and $Y_{(i)}$ with their w_{ni}-weighted means, the intercept becomes 0. Denote $\widetilde{\boldsymbol{X}}_{(i)} = (n w_{ni})^{1/2} (\boldsymbol{X}_{(i)} - \overline{X}_w)$ and $\widetilde{Y}_{(i)} = (n w_{ni})^{1/2} (Y_{(i)} - \overline{Y}_w)$, where $\overline{X}_w = \sum_{i=1}^{n} w_{ni} \boldsymbol{X}_{(i)} / \sum_{i=1}^{n} w_{ni}$ and $\overline{Y}_w = \sum_{i=1}^{n} w_{ni} Y_{(i)} / \sum_{i=1}^{n} w_{ni}$. We can rewrite the Stute's weighted least squares objective function as

$$L(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} \left(\widetilde{Y}_{(i)} - \widetilde{\boldsymbol{X}}_{(i)}^{\top} \boldsymbol{\beta} \right)^2. \tag{5.2}$$

The Stute's weighted least squares estimator of $\boldsymbol{\beta}$ can be obtained by minimizing the objective function (5.2). Since this objective function uses the least squares method, it is easy to solve. Assuming that T and C are independent, Stute (1993, 1996) showed that the estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal as $n \to \infty$.

Stute's weighted least squares method can be used to construct a loss function, and then by combining with penalty terms, variable selection in the AFT model will be performed.

5.3 The Group Bridge Estimator in The AFT Model

When covariates are grouped, instead of individual variable selection, we should treat the related covariates as a group. Let $\boldsymbol{X}_k = (X_{1k}, \ldots, X_{nk})^\top$, $k = 1, \ldots, d$, be the design vectors and $\boldsymbol{T} = (T_1, \ldots, T_n)^\top$ be the response vector in (5.1), then the regression model can be written as

$$\boldsymbol{T} = \alpha + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_d \boldsymbol{X}_d + \varepsilon$$

with an error vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Let $A_1, \cdots, A_J$ be subsets of $\{1, \cdots, d\}$ representing known groupings of the design vectors and denote the regression coefficients in the j^{th} group by $\boldsymbol{\beta}_{A_j} = (\beta_k, k \subseteq A_j)^\top$, $j = 1, \ldots, J$. For any $k \times 1$ vector a, $\|a\|_1$ denotes the L_1 norm: $\|a\|_1 = |a_1| + \cdots + |a_k|$.

Let $\boldsymbol{\beta}$ be the parameters of interest in the AFT model (5.1). After adding the group bridge penalty function proposed by Huang et al. (2009) to the Stute's weighted least squares loss function (5.2), the group bridge penalized Stute's weighted least squares objective function is

$$\begin{aligned} L_{\lambda_n}(\boldsymbol{\beta}) &= \frac{1}{2} \sum_{i=1}^{n} \left(\widetilde{Y}_{(i)} - \widetilde{\boldsymbol{X}}_{(i)}^\top \boldsymbol{\beta} \right)^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^\gamma \\ &= \frac{1}{2} \sum_{i=1}^{n} \left(\widetilde{Y}_{(i)} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right)^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^\gamma \\ &= \frac{1}{2} \left\| \widetilde{\boldsymbol{Y}} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^\gamma, \end{aligned} \tag{5.3}$$

where $\lambda_n > 0$ is the penalty tuning parameter and c_j's are constants for the adjustment of the different dimensions of A_j. In the case of uncensored data, Huang et al. (2009) suggested a simple choice of c_j is $c_j \propto |A_j|^{1-\gamma}$, where $|A_j|$

is the cardinality of A_j, and they also showed that when $0 < \gamma < 1$, the group bridge penalty is able to carry out variable selection at the group and individual variable levels simultaneously. For simplicity, we use $\widetilde{\boldsymbol{X}}_k = (\widetilde{X}_{(1)k}, \ldots, \widetilde{X}_{(n)k})^\top$ and $\widetilde{\boldsymbol{Y}} = (\widetilde{Y}_{(1)}, \widetilde{Y}_{(2)}, \cdots, \widetilde{Y}_{(n)})^\top$. Then we can obtain the penalized estimator of $\boldsymbol{\beta}$ by minimizing $L_{\lambda_n}(\boldsymbol{\beta})$.

5.3.1 Computation

Since the group bridge penalty is not a convex function for $0 < \gamma < 1$, direct minimization of $L_{\lambda_n}(\boldsymbol{\beta})$ is difficult. Following Huang et al. (2009), we formulate an equivalent minimization problem that is easier to solve computationally. For $0 < \gamma < 1$, define

$$L_{1n}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{2} \left\| \widetilde{\boldsymbol{Y}} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\boldsymbol{\beta}_{A_j}\|_1 + \tau \sum_{j=1}^{J} \theta_j,$$

where τ is a penalty parameter.

Proposition 1 *Suppose* $0 < \gamma < 1$. *If* $\lambda_n = \tau^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, *then* $\hat{\boldsymbol{\beta}}$ *minimizes* $L_{\lambda_n}(\boldsymbol{\beta})$ *if and only if* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ *minimizes* $L_{1n}(\boldsymbol{\beta}, \boldsymbol{\theta})$ *subject to* $\theta_j \geq 0$, *for* $j = 1, \ldots, J$.

When data are uncensored, i.e., $w_{ni} = 1/n, i = 1, \ldots, n$, Huang et al. (2009) pointed out that minimizing $L_{1n}(\boldsymbol{\beta}, \boldsymbol{\theta})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta})$ yields sparse solutions at the group and individual variable levels, that is, the penalty is an adaptively weighted L_1 penalty, which conduct the sparsity in $\boldsymbol{\beta}$, and when $0 < \gamma < 1$, small θ_j will force $\boldsymbol{\beta}_{A_j} = 0$ and leads to group selection.

Based on Proposition 1, for $s = 1, 2, \ldots$, we have the iterative computation algorithm as the following:

Step 1. Obtain an initial value $\boldsymbol{\beta}^{(0)}$.
Step 2. Compute

$$\theta_j^{(s)} = c_j \left(\frac{1-\gamma}{\tau\gamma} \right)^{\gamma} \|\boldsymbol{\beta}_{A_j}^{(s-1)}\|_1^{\gamma}, \quad j = 1, \cdots, J; \tag{5.4}$$

Step 3. Compute

$$\boldsymbol{\beta}^{(s+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left\| \widetilde{\boldsymbol{Y}} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2 + \sum_{j=1}^{J} \left(\theta_j^{(s)} \right)^{1-1/\gamma} c_j^{1/\gamma} \|\boldsymbol{\beta}_{A_j}\|_1 \right\};$$

Step 4. Repeat steps 2 and 3 until convergence.

The original value in Step 1 could be obtained by least squares method or LASSO approach. The main computation task is Step 3. Let $\omega_{A_j} = \left(\theta_j^{(s)}\right)^{1-1/\gamma} c_j^{1/\gamma}$, $\omega_k = \omega_{A_j} : \exists j$, such that $k \in A_j$, $\widetilde{\boldsymbol{X}}_{\omega_k} = \widetilde{\boldsymbol{X}}_k/\omega_k$, $\beta_{\omega_k} = \omega_k \beta_k$, $\boldsymbol{\beta}_{\omega A_j} = \omega_{A_j} \boldsymbol{\beta}_{A_j}$. Rewrite $\boldsymbol{\beta}^{(s+1)}$ in Step 3 as

$$\boldsymbol{\beta}^{(s+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left\| \widetilde{\boldsymbol{Y}} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_{\omega_k} \beta_{\omega_k} \right\|_2^2 + 1 \times \sum_{j=1}^{J} \|\boldsymbol{\beta}_{\omega A_j}\|_1 \right\}. \tag{5.5}$$

In Eq. (5.5), $\widetilde{\boldsymbol{X}}_{\omega_k}$ is the weighted covariate matrix for k^{th} covariate, β_{ω_k} is the weighted coefficient for the covariate and $\boldsymbol{\beta}_{\omega A_j}$ is the weighted coefficient for each group. Whenever ω_{A_j} is 0 or very small, we set $\beta_{\omega_k} = 0$ and $\beta_k = 0$, and remove the associated $\boldsymbol{X}_k$. Now the objective function (5.5) becomes a LASSO problem with tuning parameter fixed at 1, and this can be solved using the existing **R** function "predict.lars" by setting "s = 1". After the value of $\hat{\beta}_{\omega_k}$ has been estimated, we could calculate $\hat{\beta}_k = \hat{\beta}_{\omega_k}/\omega_k$, $k = 1, \ldots, d$.

5.3.2 Tuning Parameter Selection

Following the procedure in Huang et al. (2009), for a fixed λ_n, let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda_n)$ be the group bridge estimator of $\boldsymbol{\beta}$. Let $\hat{\theta}_j$, $j = 1, \ldots, J$, be the jth component of $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}(\lambda_n))$ as defined in Step 2. Let $\mathscr{X} = (\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_d)$ be the $n \times d$ covariate matrix. The Karush-Kuhn-Tucker condition for Step 3 is

$$\begin{cases} -\frac{1}{2} \frac{\partial \left\| \widetilde{\boldsymbol{Y}} - \sum_{l=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2}{\partial \beta_k} + \frac{\sum_{j:A_j \ni k} \left(\theta_j\right)^{1-1/\gamma} c_j^{1/\gamma} \beta_k}{|\beta_k|} = 0 \ \forall \beta_k \neq 0 \\ \left| \frac{1}{2} \frac{\partial \left\| \widetilde{\boldsymbol{Y}} - \sum_{l=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2}{\partial \beta_k} \right| \leq \sum_{j:A_j \ni k} \left(\theta_j\right)^{1-1/\gamma} c_j^{1/\gamma} \quad \forall \beta_k = 0, \end{cases}$$

which implies that

$$(\widetilde{\boldsymbol{Y}} - \mathscr{X} \hat{\boldsymbol{\beta}})^{\top} \widetilde{\boldsymbol{X}}_k = \sum_{j:A_j \ni k} \hat{\theta}_j^{1-1/\gamma} c_j^{1/\gamma} \operatorname{sgn}(\hat{\beta}_k), \quad \forall \hat{\beta}_k \neq 0. \tag{5.6}$$

Since $\operatorname{sgn}(\beta_k) = \beta_k/|\beta_k|$, then the fitted response vector is

$$\widehat{\widetilde{\boldsymbol{Y}}} = \mathscr{X} \hat{\boldsymbol{\beta}} = \mathscr{X}_{\lambda_n} [\mathscr{X}_{\lambda_n}^{\top} \mathscr{X}_{\lambda_n} + \mathscr{W}_{\lambda_n}]^{-1} \mathscr{X}_{\lambda_n}^{\top} \widetilde{\boldsymbol{Y}},$$

where $\mathscr{X}_{\lambda_n}$ is the sub matrix of $\mathscr{X}$ whose columns correspond to covariates with nonzero estimated coefficients for the given λ_n and $\mathscr{W}_{\lambda_n}$ is the diagonal matrix with diagonal elements

$$\sum_{k \in A_j} \hat{\theta}_j^{1-1/\gamma} c_j^{1/\gamma} / |\hat{\beta}_k|, \quad \hat{\beta}_k \neq 0.$$

Therefore, the number of effective parameters with a given λ_n can be approximated by

$$d(\lambda_n) = \text{trace} \left\{ \mathscr{X}_{\lambda_n} \left(\mathscr{X}_{\lambda_n}^\top \mathscr{X}_{\lambda_n} + \mathscr{W}_{\lambda_n} \right)^{-1} \mathscr{X}_{\lambda_n}^\top \right\}.$$

An AIC-type criterion for choosing λ_n is

$$\text{AIC}(\lambda_n) = \ln \left\{ \left\| \widetilde{\boldsymbol{Y}} - \mathscr{X} \hat{\boldsymbol{\beta}}(\lambda_n) \right\|_2^2 \Big/ n \right\} + 2d(\lambda_n)/n.$$

A BIC-type criterion for choosing λ_n is

$$\text{BIC}(\lambda_n) = \ln \left\{ \left\| \widetilde{\boldsymbol{Y}} - \mathscr{X} \hat{\boldsymbol{\beta}}(\lambda_n) \right\|_2^2 \Big/ n \right\} + \ln(n) d(\lambda_n)/n.$$

A generalized cross-validation (GCV)-type criterion for choosing λ_n is

$$\text{GCV}(\lambda_n) = \left\| \widetilde{\boldsymbol{Y}} - \mathscr{X} \hat{\boldsymbol{\beta}}(\lambda_n) \right\|_2^2 \Big/ \left\{ n \left(1 - d(\lambda_n)/n\right)^2 \right\}.$$

The tuning parameter λ_n is selected by minimizing the criteria $\text{AIC}(\lambda_n)$, $\text{BIC}(\lambda_n)$, or $\text{GCV}(\lambda_n)$.

5.3.3 Comparison With the Group LASSO

Yuan and Lin (2006) introduced the group LASSO method to select grouped variables in the linear regression model with uncensored data, which uses an L_2 norm of the coefficients associated with a group of variables in the penalty function.

We propose the group LASSO estimator for the AFT model with censored data to be

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \widetilde{\boldsymbol{Y}} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_{K_j,2}, \tag{5.7}$$

where $\lambda_n > 0$ is the tuning parameter and K_j is a positive definite matrix and $\|\boldsymbol{\beta}_{A_j}\|_{K_j,2} = (\boldsymbol{\beta}_{A_j}^\top K_j \boldsymbol{\beta}_{A_j})^{1/2}$. Yuan and Lin (2006) suggested the choice of K_j is $K_j = |A_j| I_j$ with I_j being the $|A_j| \times |A_j|$ identity matrix.

Similar to the group bridge approach, let τ be a penalty parameter, then

$$L_{2n}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{2} \left\| \widetilde{\boldsymbol{Y}} - \sum_{k=1}^{d} \widetilde{\boldsymbol{X}}_k \beta_k \right\|_2^2 + \sum_{j=1}^{J} \theta_j^{-1} \|\boldsymbol{\beta}_{A_j}\|_{K_j,2}^2 + \tau \sum_{j=1}^{J} \theta_j. \tag{5.8}$$

Proposition 2 *If $\lambda_n = 2\tau^{1/2}$, then $\tilde{\boldsymbol{\beta}}$ satisfies (5.7) if and only if $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$ minimizes $L_{2n}(\boldsymbol{\beta}, \boldsymbol{\theta})$ subject to $\boldsymbol{\theta} \geq 0$, for some $\tilde{\boldsymbol{\theta}} \geq 0$.*

From the penalized objective function (5.8) we see that the sum of the squared coefficients in group j is penalized by θ_j, and the sum of θ_j's is penalized by τ. The large θ_j tends to keep all of the elements of $\boldsymbol{\beta}_{A_j}$. Therefore, in order to minimize (5.8), either $\boldsymbol{\beta}_{A_j} = 0$, that is, the group is dropped from the model, otherwise, with large θ_j, $\boldsymbol{\beta}_{A_j} \neq 0$, which means all the elements of $\boldsymbol{\beta}_{A_j}$ are nonzero and all the variables in group j are retained in the model. So the group LASSO selects either the group with all the variables inside or deletes the whole group. This is the reason why the group LASSO can conduct group selection, but it cannot select individual variables within groups. Our simulation studies in Sect. 5.5 also reexamine this property.

5.4 Asymptotic Properties of the Group Bridge Stute's Weighted Least Squares Estimator

Stute (1993, 1996) proved consistency and asymptotic normality of the weighted least squares estimator with the Kaplan-Meier weights under some conditions. Huang et al. (2009) derived the symptomatic properties of the group bridge estimators with uncensored data. Combining the methods of these authors, we derive the asymptotic distribution of the Stute's weighted estimator under the group bridge penalty. We can show that, for $0 < \gamma < 1$, the group bridge estimators correctly select nonzero groups with probability converging to one under reasonable conditions.

According to Huang and Ma (2006)'s regularization estimation in the AFT model for ungrouped variables, let H denote the distribution function of Y. By the independence between T and C, $1 - H(y) = (1 - F(y))(1 - G(y))$, where F and G are the distribution functions of T and C, respectively. Let τ_Y, τ_T and τ_C be the endpoints of the support of Y, T and C, respectively. Let F^0 be the joint distribution of $(\boldsymbol{X}, T)$. Denote

$$\tilde{F}^0(x,t) = \begin{cases} F^0(x,t), & t < \tau_Y \\ F^0(x,\tau_{Y-}) + F^0(x,\tau_Y)1\{\tau_Y \in A\} & t \geq \tau_Y, \end{cases}$$

with A denoting the set of atoms of H. Define two sub distribution functions:

$$\tilde{H}^{11}(x,y) = P(X \leq x, Y \leq y, \delta = 1),$$

$$\tilde{H}^{0}(y) = P(Y \leq y, \delta = 0).$$

For $j = 0, \ldots, d$, let

$$\gamma_0(y) = \exp\left\{ \int_0^{y^-} \frac{\tilde{H}^0(dw)}{1 - H(w)} \right\},$$

$$\gamma_{1,j}(y;\beta) = \frac{1}{1 - H(y)} \int 1_{w>y}(w - x^\top \beta) x_j \gamma_0(w) \tilde{H}^{11}(dx, dw),$$

$$\gamma_{2,j}(y,\beta) = \iint \frac{1_{v<y, v<w}(w - x^\top \beta) x_j \gamma_0(w)}{[1 - H(v)]^2} \tilde{H}^0(dv) \tilde{H}^{11}(dx, dw),$$

$$\gamma_l(y,\beta) = (\gamma_{l,0}(y,\beta), \gamma_{l,1}(y,\beta), \ldots, \gamma_{l,d}(y,\beta))^\top, \quad l = 1, 2.$$

Denote the true regression coefficients by $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0\mathscr{A}}^\top, \boldsymbol{\beta}_{0\mathscr{B}}^\top, \boldsymbol{\beta}_{0\mathscr{C}}^\top)^\top$. For $j = 1, ..., J$, let $\mathscr{A} = \{k \in A_j : \beta_{0k} \neq 0\}$, $\mathscr{B} = \left\{k \in A_j : \beta_{0k} = 0, \boldsymbol{\beta}_{0A_j} \neq 0\right\}$, $\mathscr{C} = \left\{k \in A_j : \beta_{0k} = 0, \ \boldsymbol{\beta}_{0A_j} = 0\right\}$. So $\mathscr{A}$ contains the indices of nonzero coefficients, $\mathscr{B}$ contains the indices of zero coefficients that belong to nonzero groups, and $\mathscr{C}$ contains the indices of zero coefficients that belong to zero groups. We write $\mathscr{D} = \mathscr{B} \cup \mathscr{C}$, which contains the indices of all zero coefficients. Since $\boldsymbol{\beta}_{0\mathscr{D}} = 0$, the true model is fully explained by the first $\mathscr{A}$ subset. Then $\hat{\boldsymbol{\beta}}_{\mathscr{A}}$ and $\hat{\boldsymbol{\beta}}_{\mathscr{D}}$ are the estimates of $\boldsymbol{\beta}_{\mathscr{A}}$ and $\boldsymbol{\beta}_{\mathscr{D}}$ from the group bridge estimator $\hat{\boldsymbol{\beta}}$, respectively.

Let $W = \text{diag}(nw_1, \ldots, nw_n)$ be the diagonal matrix of the Kaplan-Meier weights. Let $\mathscr{X}_{A_j} = (\widetilde{\boldsymbol{X}}_k, k \in A_j)$ be the matrix with columns $\widetilde{\boldsymbol{X}}_k$'s for $k \in A_j$, and denote $\Sigma_{A_j} = \mathscr{X}_{A_j}^\top W \mathscr{X}_{A_j}/n$. For $i = 1, \ldots, n$, let $e_i = \widetilde{Y}_{(i)} - \widetilde{\boldsymbol{X}}_{(i)}^\top \boldsymbol{\beta}_0$ and $\xi_k = \sum_{i=1}^n \widetilde{X}_{ik} e_i, 1 \le k \le d$. Define $\Sigma_n = \mathscr{X}^\top \mathscr{X}/n$ and let ρ_n and ρ_n^* be the smallest and largest eigenvalues of Σ_n. We assume the following.

(A1) The number of nonzero coefficients q is finite;

(A2) (a) The observations $(Y_i, \boldsymbol{X}_i, \delta_i), i = 1, \ldots, n$ are independent and identically distributed; (b) The random errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with mean 0 and finite variance σ^2, and furthermore, there exist $K_1, K_2 > 0$ such that the tail probabilities of ε_i satisfy $P(|\varepsilon_i > u|) \le K_2 \exp(-K_1 u^2)$ for all $u \ge 0$ and all i;

(A3) (a) The distribution of ξ_i's are subgaussian; (b) The covariates are bounded, that is, there exists a constant $M > 0$ such that $|X_{ik}| \le M, 1 \le i \le n, 1 \le k \le d$;

(A4) The covariate matrix $\mathscr{X}$ satisfies the sparse Riesz condition (SRC) with rank q^*: there exist constants $0 < c_* < c^* < \infty$, such that for $q^* = (3+4C)q$, with probability converging to 1, $c_* \le \boldsymbol{v}^\top \Sigma_{A_j} \boldsymbol{v}/\|\boldsymbol{v}\|_2^2 \le c^*$, $\forall A_j$ with $|A_j| = q^*$ and $\boldsymbol{v} \in \mathbb{R}^{q^*}$;

(A5) The maximum multiplicity $C_n^* = \max_k \sum_{j=1}^J I\{k \in A_j\}$ is bounded and

$$\frac{\lambda_n^2}{n\rho_n} \sum_{j=1}^{J_1} c_j^2 \|\boldsymbol{\beta}_{0A_j}\|_1^{2\gamma-2} |A_j| \le M_n \ln(d), \quad M_n = O(1);$$

(A6) The constants c_j are scaled so that $\min_{1 \le j \le J} c_j \ge 1$ and

$$\frac{\lambda_n (\rho_n^2)^{1-\gamma/2}}{\{\ln(d)\}^{1-\gamma/2} (q^* + \rho_n)^{1-\gamma/2} \rho_n^* n^{\gamma/2}} \to \infty.$$

The model is sparse by assumption (A1) (Huang and Ma 2010; Ma and Du 2012), which is reasonable in genomic studies, that is, although the total number of covariates may be large, the number of covariates with nonzero coefficients is still small. The subgaussian assumption (A2) has been made in high dimensional linear regression models (Zhang and Huang 2008). Assumption (A3) proposed by Ma and Du (2012) shows the subguassian tail property still holds under censoring, and it is required for Theorem 1. The SRC condition proposed by Zhang and Huang (2008) in assumption (A4) implies that all the eigenvalues of any $p \times p$ submatrix of $\mathscr{X}^\top W \mathscr{X}/n$ with $p \le q^*$ lie between c_* and c^*. It ensures that any model with dimension no greater than q^* is identifiable. Similar assumptions (A5) and (A6) were used under uncensored data in Huang et al. (2009). We allow ln(d) = o(n)

or d = exp(o(n)), so our work is more general than that of Huang et al. (2009). Also (A5) and (A6) put restrictions on the magnitude of the penalty parameter, which is $0 < \gamma < 1$.

Theorem 1 (Group Bridge) *Suppose that* $0 < \gamma < 1$, *conditions (A1)–(A6) hold and* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$. *Let* $\boldsymbol{X}_1 = (\boldsymbol{X}_k, k \in \mathscr{A})$ *and* $\Sigma_1 = \mathrm{E}(\boldsymbol{X}_1 \boldsymbol{X}_1^\top)$. *Then*

(i) (Zero Group Selection Consistency)

$$\Pr\{\hat{\boldsymbol{\beta}}_{n\mathscr{C}} = 0\} \to 1.$$

(ii) (Asymptotic Distribution of Nonzero Parameter Estimators in Nonzero Groups)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{n\mathscr{A}} - \boldsymbol{\beta}_{0\mathscr{A}}) \to_D \arg\min\{U_1(\boldsymbol{b}) : \boldsymbol{b} \in \mathbb{R}^{|\mathscr{A}|}\},$$

where

$$U_1(\boldsymbol{b}) = -\boldsymbol{b}^\top \boldsymbol{V}_1 + \frac{1}{2}\boldsymbol{b}^\top \Sigma_1 \boldsymbol{b} + \gamma\lambda_0 \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{0A_j}\|_1^{\gamma-1} \sum_{k\in\mathscr{A}} \{b_k sgn(\beta_{0k})\}.$$

Here

$$\boldsymbol{V}_1 \sim N(0, \Omega_1),$$

$$\Omega_1 = \mathrm{Var}\{\delta\gamma_0(Y)(Y - \boldsymbol{X}_1^\top \boldsymbol{\beta}_{0\mathscr{A}})\boldsymbol{X}_1 + (1-\delta)\gamma_1(Y; \boldsymbol{\beta}_{0\mathscr{A}}) - \gamma_2(Y; \boldsymbol{\beta}_{0\mathscr{A}})\}.$$

Note that if $\lambda_0 = 0$, the penalty part is negligible, the asymptotic distribution of nonzero parameters estimators in both zero and nonzero groups becomes that of the Stute's estimator. Part (i) of Theorem 1 states that the group bridge estimates of the coefficients of the zero groups exactly equal 0 with probability converging to one; part (ii) shows the normality property of the group bridge estimates of the coefficients of the nonzero parameters in both zero and nonzero groups. Part (i) of Theorem 1 implies that the group bridge estimator can distinguish nonzero groups from zero groups correctly, but it does not address the zero coefficients in nonzero groups, so the proposed method possesses group selection consistency but lacks individual selection consistency. In order to archive individual selection consistency, following Wang et al. (2009), an adaptive group bridge penalty is needed, this issue will be explored in our future research. Our part (ii) of Theorem 1 is also different from that in Theorem 1 (b) of Huang et al. (2009), where $\mathscr{A}$ is replaced by $B_1 = \mathscr{A} \cup \mathscr{B}$, which shows the asymptotic distribution of nonzero group estimators.

However, we found that their asymptotic distribution results were hard to get under the given conditions, and further investigation is needed to solve the problem.

We also present the Theorem 2 for the group LASSO estimator to compare the asymptotic properties of the group bridge and group LASSO estimators.

Theorem 2 (Group LASSO) *Suppose* $\{\boldsymbol{\beta}, d, A_j, c_j, K_j, j \le J\}$ *are all fixed.* ε_i*'s are iid errors with* $\mathrm{E}(\varepsilon_i) = 0$ *and* $\mathrm{Var}(\varepsilon_i) = \sigma^2 \in (0, \infty)$. *Let* $\Sigma_2 = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top)$ *and suppose* $n^{-1/2}\lambda_n \to \lambda_0 > 0$ *when* $n \to \infty$. *Then*

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_D \arg\min\{U_2(\boldsymbol{b}) : \boldsymbol{b} \in \mathbb{R}^d\},$$

where

$$U_2(\boldsymbol{b}) = -\boldsymbol{b}^\top \boldsymbol{V}_2 + \frac{1}{2}\boldsymbol{b}^\top \Sigma_2 \boldsymbol{b} + \lambda_0 \sum_{j=1}^{J} c_j \left\{ \frac{\boldsymbol{b}_{A_j}^\top K_j \boldsymbol{\beta}_{0A_j}}{\|\boldsymbol{\beta}_{0A_j}\|_{K_j,2}} I(\boldsymbol{\beta}_{A_j} \neq 0) + \|\boldsymbol{b}_{A_j}\|_{K_j,2} I(\boldsymbol{\beta}_{A_j} = 0) \right\}.$$

Here

$$\boldsymbol{V}_2 \sim N(0, \Omega_2),$$

$$\Omega_2 = \mathrm{Var}\{\delta\gamma_0(Y)(Y - \boldsymbol{X}^\top\boldsymbol{\beta}_0)\boldsymbol{X} + (1-\delta)\gamma_1(Y; \boldsymbol{\beta}_0) - \gamma_2(Y; \boldsymbol{\beta}_0)\}.$$

From Theorem 2 we notice that, when $\lambda_0 = 0$, the group LASSO estimator is the same as the Stute's estimator. When $\lambda_0 > 0$, the asymptotic distribution of $\tilde{\boldsymbol{\beta}}_n$ puts positive probability at 0 when $\boldsymbol{\beta}_{A_j} = 0$. Since this positive probability is less than one in general, which results in the non-consistency property in selecting the nonzero groups.

5.5 Simulation Studies

In this section, simulations are conducted to compare the bi-level (group and within-group individual variable levels) performance of the group bridge estimator and the group LASSO estimator. Following Huang et al. (2009)'s simulations set up for uncensored data, two scenarios are considered. Since the proposed method can deal with right censored data, the logarithm of censoring times, C, are generated by the logarithm of random variables from the exponential distribution with a rate parameter ν, where ν is chosen to obtain 20 %, 50 % and 70 % censoring rates for both scenarios. In Scenario 1, the number of groups is moderately large, the group

sizes are equal and relatively large, and within each group the coefficients are either all nonzero or all zero. In Scenario 2, the group sizes vary and there are coefficients equal to zero in a nonzero group. We use $\gamma = 0.5$ in the group bridge estimator. The sample size $n = 200$ in each scenario. The simulation results are based on 400 replications.

We calculate the average number of groups selected (No.Grp), the average number of variables selected (No.Var), the percentage of occasions on which the model produced contains the same groups as the true model (%Corr.Grp), the percentage of occasions on which the model produced contains the same variables as the true model (%Corr.Var) and the model error (Model Error), which is computed as $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top E(\boldsymbol{X}^\top \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}_0$ is the true coefficient value. Enclosed in parentheses are the corresponding standard deviations. And the last line in each table gives the true values used in the generation model. For example, in Scenario 1, there are 2 nonzero groups and 16 nonzero coefficients in the generation model.

For both of the group bridge estimator and group LASSO estimator, AIC, BIC and GCV tuning parameter selection methods were used to evaluate performance. The variable selection and coefficient estimation results based on GCV are similar to those using AIC. A comparison of different tuning parameter selection methods indicates that tuning based on BIC in general does better than that based on AIC and GCV in terms of selection at the group and individual variable levels. We therefore focus on the comparisons of the methods with BIC tuning parameter.

5.5.1 Scenario 1

In this experiment, there are five groups and each group consists of eight covariates. The covariate vector is $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_5)$ and, for any j in $1, \ldots, 5$, the subvector of covariates that belong to the same group is $\boldsymbol{X}_j = (\boldsymbol{x}_{8(j-1)+1}, \ldots, \boldsymbol{x}_{8(j-1)+8})$. To generate the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{40}$, we first simulate 40 random variables $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_{40}$ independently from the standard normal distribution. Then $\boldsymbol{Z}_j$ ($j = 1, \ldots, 5$) are simulated with a normal distribution and an AR(1) structure such that $cov(\boldsymbol{Z}_{j_1}, \boldsymbol{Z}_{j_2}) = 0.4^{|j_1 - j_2|}$, for $j_1, j_2 = 1, \ldots, 5$. The covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{40}$ are generated as

$$\boldsymbol{x}_j = (\boldsymbol{Z}_{g_j} + \boldsymbol{R}_j)/\sqrt{2}, \quad j = 1, \ldots, 40,$$

where g_j is the smallest integer greater than $(j-1)/8$ and the x_js with the same value of g_j belong to the same group. The logarithm of failure times are generated from the log-Normal model, $T = \sum_{j=1}^{40} \boldsymbol{x}_j \beta_j + \varepsilon$, where the random error is $\varepsilon \sim N(0, 2^2)$,

and

$$(\beta_1, \ldots, \beta_8) = (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4),$$

$$(\beta_9, \ldots, \beta_{16}) = (2, 2, \ldots, 2),$$

$$(\beta_{17}, \ldots, \beta_{24}) = (\beta_{25}, \ldots, \beta_{32}) = (\beta_{33}, \ldots, \beta_{40}) = (0, 0, \ldots, 0).$$

Thus, the coefficients in each group are either all nonzero or all zero.

Table 5.1 summarizes the simulation results for Scenario 1. From these results we notice that as the censoring rate increases, the model error increases and the percentage of correct variables selected decreases for both of the group bridge and group LASSO methods. But comparing the group bridge approach with the group LASSO approach, the group bridge method tends to more accurately select correct groups as well as the variables in each group, even when the censoring rate is high as 70 %. While the group LASSO method tends to select more groups and variables than the true models, and when the censoring rate is high, the group LASSO method performs poorer than the group bridge method. So in terms of the number of groups selected, the number of variables selected, the percentage of correct models selected and the correct variable selected, the group bridge considerably outperforms the group LASSO, and the group bridge incurs smaller model error than the group LASSO.

Table 5.1 Simulation results for Scenario 1

CR%	Method	No.Grp	No.Var	%Corr.Grp	%Corr.Var	Model error
20	GBridge	2.00(0.000)	15.98(0.156)	100.00(0.000)	97.50(0.156)	0.596(0.253)
	GLASSO	2.23(0.508)	17.84(4.061)	80.50(0.397)	80.50(0.397)	2.025(0.269)
50	GBridge	2.00(0.000)	15.93(0.251)	100.00(0.000)	93.25(0.251)	0.900(0.369)
	GLASSO	2.57(0.723)	20.52(5.780)	56.25(0.497)	56.25(0.497)	2.328(0.673)
70	GBridge	2.00(0.000)	15.86(0.352)	100.00(0.000)	86.50(0.342)	1.442(0.683)
	GLASSO	2.30(0.544)	18.40(4.351)	73.75(0.441)	73.75(0.441)	2.710(1.014)
True		2	16	100	100	0

GBridge the group bridge method, *GLASSO* the group LASSO method, *CR* censoring rate, *No.Grp* the average number of groups selected, *No.Var* the average number of variables selected, *%Corr.Grp* the percentage of occasions on which the model produced contains the same groups as the true model, *%Corr.Var* the percentage of occasions on which the model produced contains the same variables as the true model, *Model Error* $= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top E(\boldsymbol{X}^\top \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. Empirical standard deviations are in the parentheses

5.5.2 Scenario 2

In this experiment, the group size differs across groups. There are six groups made up of three groups each of size 10 and three groups each of size 4. The covariate vector is $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_6)$, where the six subvectors of covariates are $\boldsymbol{X}_j = (\boldsymbol{x}_{10(j-1)+1}, \ldots, \boldsymbol{x}_{10(j-1)+10})$, for $j = 1, 2, 3$, and $\boldsymbol{X}_j = (\boldsymbol{x}_{4(j-4)+31}, \ldots, \boldsymbol{x}_{4(j-4)+34})$, for $j = 4, 5, 6$. To generate the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{42}$, we first simulate $\boldsymbol{Z}_i (i = 1, \ldots, 6)$ and $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_{42}$ independently from the standard normal distribution. For $j = 1, \ldots, 30$, let g_j be the largest integer less than $j/10 + 1$, then $g_j = 1, 2, 3$, and for $j = 31, \ldots, 42$, let g_j be the largest integer less than $(j - 30)/4 + 4$, then $g_j = 4, 5, 6$. The covariates $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{42})$ are obtained as

$$\boldsymbol{x}_j = (\boldsymbol{Z}_{g_j} + \boldsymbol{R}_j)/\sqrt{2}, \quad j = 1, \ldots, 42.$$

The logarithm of failure times are generated from $T = \sum_{j=1}^{42} \boldsymbol{x}_j \beta_j + \varepsilon$, where the random error is $\boldsymbol{\varepsilon} \sim N(0, 2^2)$, and

$$(\beta_1, \ldots, \beta_{10}) = (0.5, -2, 0.5, 2, -1, 1, 2, -1.5, 2, -2),$$

$$(\beta_{11}, \ldots, \beta_{20}) = (-1.5, 2, 1, -2, 1.5, 0, 0, 0, 0, 0),$$

$$(\beta_{21}, \ldots, \beta_{30}) = (0, \ldots, 0), \ (\beta_{31}, \ldots, \beta_{34}) = (2, -2, 1, 1.5),$$

$$(\beta_{35}, \ldots, \beta_{38}) = (-1.5, 1.5, 0, 0), \ (\beta_{39}, \ldots, \beta_{42}) = (0, \ldots, 0).$$

Thus we consider the situation that the group size differs across groups and the coefficients in a group can be either all zero, all nonzero or partly zero.

Table 5.2 summarizes the simulation results for Scenario 2. From Table 5.2 we can see that when the censoring rate is low, the group bridge method choses more

Table 5.2 Simulation results for Scenario 2

CR%	Method	No.Grp	No.Var	%Corr.Grp	Model error
20	GBridge	4.00(0.279)	24.50(1.460)	96.0(0.196)	1.569(0.583)
	GLASSO	5.12(0.994)	35.38(5.491)	44.0(0.497)	3.045(0.882)
50	GBridge	4.08(0.473)	24.86(1.962)	90.5(0.293)	2.241(0.907)
	GLASSO	5.42(0.909)	37.20(5.075)	29.0(0.454)	3.487(1.015)
70	GBridge	4.02(0.453)	22.51(2.212)	89.3(0.310)	4.113(1.823)
	GLASSO	4.97(1.001)	34.35(5.393)	51.8(0.500)	3.775(1.122)
True		4	21	100	0

GBridge the group bridge method, *GLASSO* the group LASSO method, *CR* censoring rate, *No.Grp* the average number of groups selected, *No.Var* the average number of variables selected, *%Corr.Grp* the percentage of occasions on which the model produced contains the same groups as the true model, *Model Error* = $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top E(\boldsymbol{X}^\top \boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. Empirical standard deviations are in the parentheses

Table 5.3 Simulation results in each group for Scenario 2

CR%	Method	No.Var					
		G1	G2	G3	G4	G5	G6
20	GBridge	9.8	7.9	0.0	4.0	2.7	0.0
	GLASSO	10.0	10.0	5.5	4.0	4.0	1.9
50	GBridge	9.8	8.2	0.2	4.0	2.7	0.0
	GLASSO	10.0	10.0	7.0	4.0	4.0	2.2
70	GBridge	9.3	7.5	0.1	3.9	1.6	0.0
	GLASSO	10.0	10.0	4.7	4.0	4.0	1.6
True		10	5	0	4	2	0

GBridge the group bridge method, *GLASSO* the group LASSO method, *CR* censoring rate, *No.Var* the average number of variables selected, *G1, . . . , G6* the six groups

accurate groups and variables than the group LASSO method. When the censoring rate goes up, both of the group bridge and group LASSO approaches result in higher model errors, but the group bridge method still performs better than the group LASSO in terms of the number of groups selected, the number of variables selected and the percentage of correct models selected. Table 5.3 gives the average variable selected in each group. The group bridge estimator is closer to the true value while the group LASSO method tends to choose more variables than the true variables in each group.

5.6 Real Data Analysis

The PBC data can be found in Fleming and Harrington (2011), and is obtained by using attach(pbc) inside the **R** {SMPractials} package. The data set is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. In this study, 312 out of 424 patients who agreed to participate in the randomized trial are eligible for the analysis. Among the 312 patients, 152 were assigned to the drug D-penicillanmine, while the others were assigned to a control group with placebo drug. Some covariates, such as age, gender and albumin level, were recorded. The primary interest was to investigate the effectiveness of D-penicillanmine in curing PBC disease. To compare with the analysis of PBC data in Huang et al. (2014), we restrict our attention to the 276 observations without missing covariate values. The censoring rate is 60 %. All of the 17 risk factors are naturally clustered into 9 different categories, measuring different aspects, such as liver reserve function and demographics, etc. The definitions of the 10 continuous and 7 categorical variables are given in the accompanying study Dictionary table Table 5.4.

We fitted the PBC data in the AFT model, then used the group bridge and group LASSO methods both with BIC to select the tuning parameter λ_n. Huang et al. (2014) fitted this data set in the Cox proportional hazards model with the group bridge penalty under the BIC tuning parameter selection method. For comparison, Table 5.5 includes these different penalties and different estimation results. For the

Table 5.4 Dictionary of PBC data covariates

Group	Variable	Type	Definition
Age(G1)	X1	C	Age(years)
Gender(G2)	X2	D	Gender(0 male; 1 female)
Phynotype(G3)	X3	D	Ascites(0 absence)
	X4	D	Hepatomegaly(0 absence; 1 presence)
	X5	D	Spiders(0 absence; 1 presence)
	X6	D	Edemaoed(0 no edema; 0.5 untreated/successfully treated)
Liver function damage(G4)	X7	C	Alkaline phosphatase(units/litre)
	X8	C	Sgot(liver enzyme in units/ml)
Excretory function of the liver(G5)	X9	C	Serum bilirubin(mg/dl)
	X10	C	Serum cholesterol(mg/dl)
	X11	C	Triglyserides(mg/dl)
Liver reserve function(G6)	X12	C	Albumin(g/dl)
	X13	C	Prothrombin time(seconds)
Treatment(G7)	X14	D	Penicillamine v.s. placebo(1 control; 2 treatment)
Reflection(G8)	X15	D	Stage(histological stage of disease, graded 1,2,3 or 4)
	X16	C	Urine copper(ug/day)
Haematology(G9)	X17	C	Platelets(per cubic ml/1000)

Note: **Type**: type of variable (C: continuous; D: discrete)

Table 5.5 Estimation results of PBC data

		AFT-BIC		Cox-BIC
Group	Covariate	GroupLASSO	GroupBridge	GroupBridge
G1	Age	0.002	0	0
G2	Gender	0.454	0.25	−0.945
G3	asc	−0.455	−0.481	0.136
	hep	0.136	0.071	0.146
	spid	−0.404	−0.36	0.102
	oed	−0.310	−0.4	0.566
G4	alk	0	0	0
	sgot	0	0	0
G5	bil	−0.059	−0.047	0.060
	chol	0.001	0	0
	trig	0.001	0	0
G6	alb	1.134	1.12	−1.289
	prot	0.268	0.322	0.124
G7	trt	0	0	−0.237
G8	stage	0.033	0	0
	cop	−0.002	−0.002	0
G9	plat	0	0	0

Note: The results for Cox-BIC are from Huang et al. (2014)

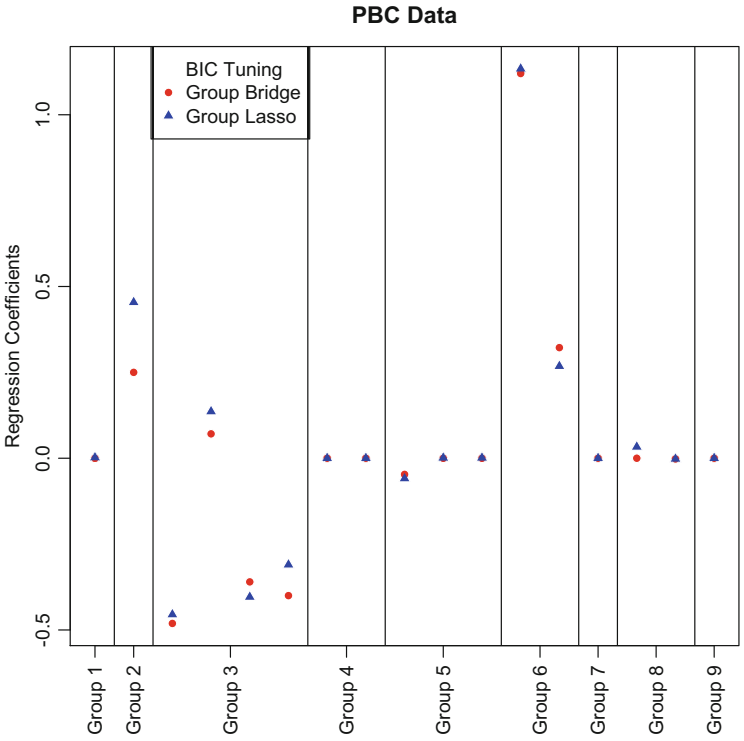

Fig. 5.1 Group bridge vs. Group LASSO estimation results of PBC data based on AFT model

AFT model, the group LASSO and group bridge methods obtain similar estimation coefficients values, except that comparing with the group bridge method, the group LASSO method selected one variable age in group 1, two variables chol and trig in group 5 and the variable stage in group 8, while the group bridge method did not select these variables. Using the group bridge with BIC under different models, the AFT and Cox models selected almost the same groups and variables, except the AFT model chose the variable cop in group 8, while the Cox model did not. The Cox model selected the variable trt in group 7, while the AFT model did not.

In order to have a clear visual comparison, we plotted the coefficients based on different models and different penalty functions. Figure 5.1 shows the estimated coefficients based on the AFT model with the group LASSO penalty (blue triangles) and the group bridge penalty (red circles), respectively. We could see that except group 8, the group bridge and group LASSO choose the same groups and the

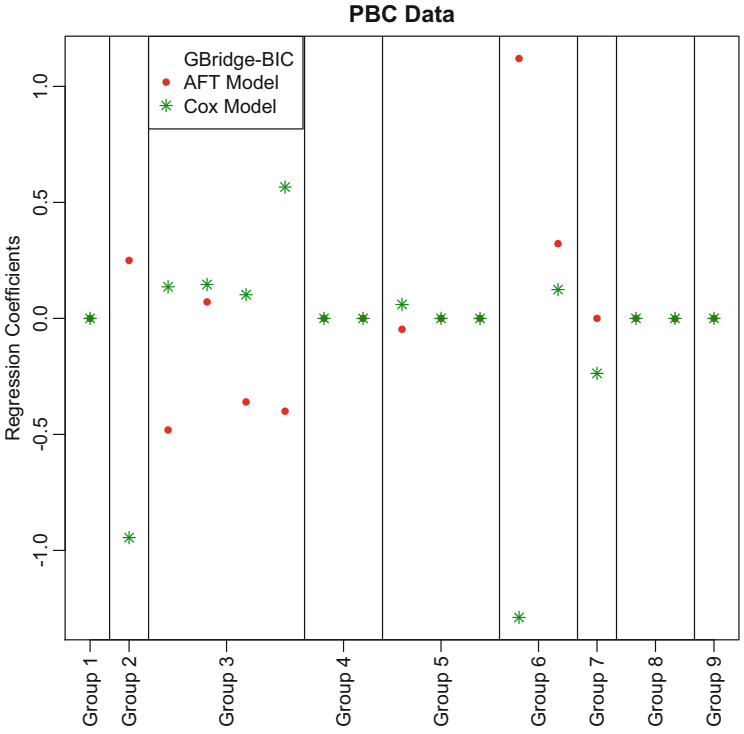

Fig. 5.2 Group bridge estimation results of PBC data based on AFT vs. Cox model

estimated coefficients for each variable are very similar. Figure 5.2 contains the estimated coefficients under the group bridge method in the AFT model and the Cox proportional hazards model. The coefficients in the AFT model indicate the relationship between the covariates and the logarithm of survival time and the coefficients in the Cox model represent the relationship between the covariates and the logarithm of hazard, their signs are opposite. From Fig. 5.2 we also see that except for group 7, these two models select the same groups and variables based on the group bridge method. Figure 5.3 combines Figs. 5.1 and 5.2 for a better visual comparison.

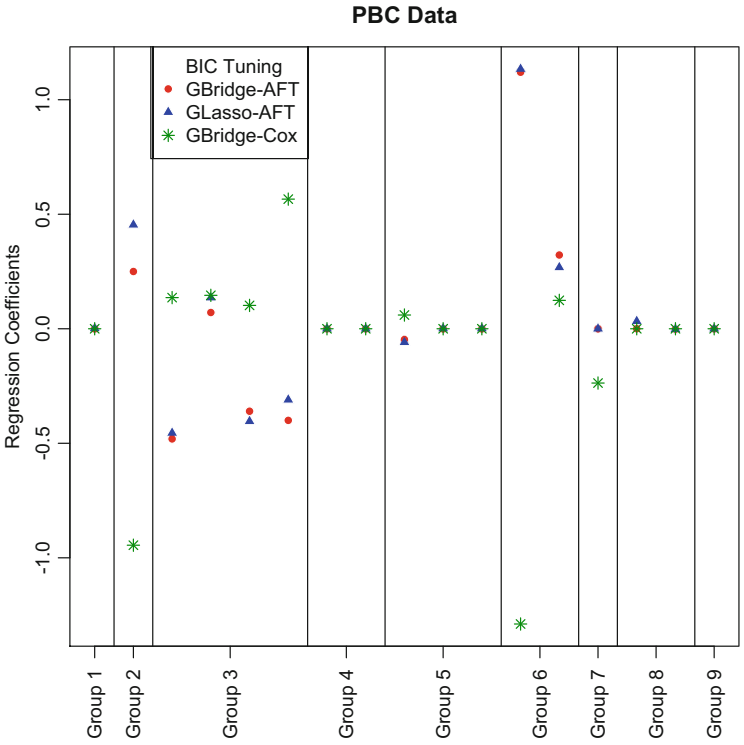

Fig. 5.3 Comparison of the estimation results of PBC data based on three methods

5.7 Summary and Discussion

We have considered an extension of the group LASSO and group bridge regression to the AFT model with right censored data. Stute's weighted least squares estimator with the group bridge penalty in AFT model is comparable to that in the Cox regression model with group bridge penalty. The group bridge approach performs better in selecting both the correct groups and individual variables than the group LASSO method even when censoring rates are high. We have established the asymptotic properties of the group bridge penalized Stute's weighted least squares estimators and allow the dimension of covariates to be larger than the sample size, which is applicable for the high-dimensional genomic data.

We focused on the group bridge penalty for the group and within-group variable selection, and only compared it to the group LASSO penalty. It is possible to consider the Stute's weighted least squares estimators with the group SCAD penalty (Wang et al. 2007), or with the group MCP penalty (Breheny and Huang 2009), although the asymptotic properties of each penalized method need to be studied.

On the other hand, in many real life survival data sets, covariates have nonparametric effects on the survival time, so the nonparametric or partial linear regressions are of interest. In order to distinguish the nonzero components from the zero components, the group bridge approach could also be applied in the nonparametric and partial linear regressions. We are working on these projects now and the detailed information will be reported elsewhere.

References

Bakin S (1999) Adaptive regression and model selection in data mining problems. PhD dissertation, The Australian National University

Breheny P, Huang J (2009) Penalized methods for bi-level variable selection. Stat Interface 2:269–380

Buckley J, James I (1979) Linear regression with censored data. Biometrika 66:429–436

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Fleming TR, Harrington DP (2011) Counting processes and survival analysis, vol 169. Wiley, New York

Fu WJ (1998) Penalized regressions: the bridge versus the lasso. J Comput Graphic Stat 7:397–416

Fygenson M, Ritov Y (1994) Monotone estimating equations for censored data. The Ann Stat 22:732–746

Heller G (2007) Smoothed rank regression with censored data. J Am Stat Assoc 102(478):552–559

Huang J, Ma S, Xie H (2006) Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics 62:813–820

Huang J, Ma S, Xie H, Zhang CH (2009) A group bridge approach for variable selection. Biometrika 96:339–355

Huang J, Ma S (2010) Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal 16:176–195

Huang J, Liu L, Liu Y, Zhao X (2014) Group selection in the Cox model with a diverging number of covariates. Stat Sinica 24:1787–1810

Lv J, Fan Y (2009) A unified approach to model selection and sparse recovery using regularized least squares. Ann Stat 37:3498–3528

Ma S, Huang J (2007) Clustering threshold gradient descent regularization: with applications to microarray studies. Bioinformatics 23:466–472

Ma S, Du P (2012) Variable selection in partly linear regression model with diverging dimensions for right censored data. Stat Sinica 22:1003–1020

Meier L, Van De Geer S, Bühlmann P (2008) The group lasso for logistic regression. J Royal Stat Soc: Ser B (Stat Methodol) 70:53–71

Ritov Y (1990) Estimation in a linear regression model with censored data. Ann Stat 18:303–328

Stute W (1993) Almost sure representations of the product-limit estimator for truncated data. Ann Stat 21:146–156

Stute W (1996) Distributional convergence under random censorship when covariables are present. Scand J Stat 23:461–471

Stute W, Wang JL (1993) The strong law under random censorship. Ann Stat 9:1591–1607

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc. Ser B (Methodol) 58:267–288

Tsiatis AA (1990) Estimating regression parameters using linear rank tests for censored data. Ann Stat 18:354–372

Wang L, Chen G, Li H (2007) Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics 23:1486–1494

Wang S, Nan B, Zhu N, Zhu J (2009) Hierarchically penalized Cox regression with grouped variables. Biometrika 96:307–322

Ying Z (1993) A large sample study of rank estimation for censored regression data. Ann Stat 21:76–99

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J Royal Stat Soc: Ser B (Stat Methodol) 68:49–67

Zhang CH, Huang J (2008) The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann Stat 36:1567–1594

Zhao P, Rocha G, Yu B (2009) The composite absolute penalties family for grouped and hierarchical variable selection. Ann Stat 37:3468–3497

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc: Ser B (Stat Methodol) 67:301–320

Chapter 6
A Proportional Odds Model for Regression Analysis of Case I Interval-Censored Data

Pooneh Pordeli and Xuewen Lu

Abstract Case I interval censored or current status data arise in many areas such as demography, economics, epidemiology and medical science. We introduce a partially linear single-index proportional odds model to analyze these types of data. Polynomial smoothing spline method is applied to estimate the nuisance parameters of our model including the baseline log-odds function and the nonparametric link function with and without monotonicity constraint, respectively. Then, we propose a simultaneous sieve maximum likelihood estimation (SMLE). It is also shown that the resultant estimator of regression parameter vector is asymptotically normal and achieves the semiparametric information bound, considering that the nonparametric link function is truly a spline. A simulation experiment presents the finite sample performance of the proposed estimation method, and an analysis of renal function recovery data is performed for the illustration.

6.1 Introduction

The proportional odds (PO) model has been used widely as a major model for analyzing survival data which is particularly practical in the analysis of categorical data and is possibly the most popular model in the case of ordinal outcomes related to survival data. Ordinal responses are very common in the medical, epidemiological, and social sciences. The PO model was first proposed by McCullagh (1980) where he extended the idea of constant odds ratio to more than two samples by means of the PO model. Pettitt (1982) and Bennett (1983) generalized this model to the survival analysis context and subsequently much effort and research has gone into proposing reasonable estimators for the regression coefficients for this model. Although the proportional hazards (PH) model is the most common approach used for studying the relationship of event times and covariates, alternate models are needed for occasions when it does not fit the data. As mentioned in Hedeker and Mermelstein (2011), in analysis of failure time data, when subjects are measured

P. Pordeli (✉) • X. Lu
Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, T2N 1N4, Calgary, AB, Canada
e-mail: ppordeli@ucalgary.ca; lux@math.ucalgary.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_6

repeatedly at fixed intervals in terms of the occurrence of some event, or when determination of the exact time of the event is only known within grouped intervals of time, the PO model is a rather elegant and popular choice considering its ordered categorical nature without any substantial increase in the difficulty of interpretation. The regression parameter estimates have a nice interpretation as the additive change in the log-odds (multiplicative effect on the odds) of survival associated with a one unit change in covariate values.

Suppose T is the failure time of some event of interest and C is a random censoring time. The observations of failure time, T, are from current status data type where the only information that we have about them is that if the failure has happened before or after the examination time C instead of being observed exactly. Let $V = (V_1, \ldots, V_q)^\top$ is a q-dimensional linear covariate vector which is time independent. The linear PO model is defined as

$$\frac{1 - S(t|V)}{S(t|V)} = \frac{1 - S_0(t)}{S_0(t)} \exp(\alpha^\top V).$$

Since $\text{logit}(u) = \ln\{u/(1-u)\}$, by taking natural logarithm of both sides, we can write the model as follows

$$\text{logit}\{1 - S(t|V)\} = \text{logit}\{1 - S_0(t)\} + \alpha^\top V, \tag{6.1}$$

where $S(t|V)$ is the survival function of T conditional on covariate V, $\alpha = (\alpha_1, \ldots, \alpha_q)$ is a q-dimensional regression coefficient vector, and $S_0(t)$ is the baseline survival function corresponding to $V = 0$. Thus, $\text{logit}\{1 - S_0(t)\}$ is the baseline log-odds function. It is a monotone increasing function since $1 - S_0(t) = F_0(t)$ and $\text{logit}(\cdot)$ are increasing. In this model, α_j, $j = 1, \ldots, q$, is the increase in log-odds of falling into or below any category of the response variable, associated with the one unit increase in V_j holding all other $V_{j'}$'s ($j' \neq j$) constant. Therefore, a positive slope indicates a tendency for the response level to increase as the covariate increases. In other words, the PO model considers the effect that changes in the explanatory variables V_1 to V_q have on the log-odds of T being in a lower rather than a higher category. A key advantage of this model is that it uses a logit link yielding constant odds ratios; hence the name is proportional odds model.

One important property of the PO model is that the hazard ratio converges from $\exp\{\alpha^\top V\}$ to unity as the time changes from zero to ∞. From (6.1) we can write

$$S(t|V) = \frac{1}{1 + \left(\frac{1-S_0(t)}{S_0(t)}\right)\exp(\alpha^\top V)},$$

and since $S(t|V) = e^{-\Lambda(t|V)}$ and $S_0(t) = e^{-\Lambda_0(t)}$ we have

$$\Lambda(t|V) = \ln\left\{1 + (e^{-\Lambda_0(t)} - 1)\exp\left(\alpha^\top V\right)\right\},$$

and then considering $\lambda(t|V) = \partial \Lambda(t|V)/\partial t$, it follows that

$$\frac{\lambda(t|V)}{\lambda_0(t)} = \frac{1}{1 + \left\{\exp(-\alpha^\top V) - 1\right\} S_0(t)}.$$

Thus, when $t = 0$, $S_0(t)$ is 1 and $\lambda(t|V)/\lambda_0(t) = \exp\{\alpha^\top V\}$, and when $t = \infty$, $S_0(t)$ is 0 and $\lambda(t|V)/\lambda_0(t) = 1$. This is different from the PH model where the hazard ratio remains constant as time passes, such that $\lambda(t|V)/\lambda_0(t) = \exp(\alpha^\top V)$, which could be unreasonable in some applications where the initial effects such as differences in the stage of disease or in treatment can disappear over the time. In this case, the property of the PO model that the hazard ratio converges to 1 as t increases to infinity makes more sense.

To analyze interval-censored data, a number of articles considered the PO model. Dinse and Lagakos (1983) focused on score tests derived under model (6.1) which expressed tumour prevalence as a function of time and treatment. Huang and Rossini (1997) used sieve maximum likelihood estimation (SMLE) to estimate the finite-dimensional regression parameter. They showed that the estimators are asymptotically normal with $\sqrt{n}$ convergence rate and achieve the information bound. Shen (1998) also developed an estimation procedure for the baseline function and the regression parameters based on a random sieve maximum likelihood method for linear regression with an unspecified error distribution, taking the PH and PO models as special cases. Their procedures used monotone splines to approximate the baseline survival function. They implemented the proposed procedures for right-censored and case II interval-censored data. The estimated regression parameters are shown to be asymptotically normal and efficient.

For PO models with current status data, Huang (1995) used maximum likelihood estimation (MLE). Rossini and Tsiatis (1996) treated the baseline log-odds of failure time as the infinite-dimensional nuisance parameter of their model and approximated it with a uniformly spaced non-decreasing step function, and then proceeded with a maximum likelihood procedure. In Rabinowitz et al. (2000) the basis for estimation of the regression coefficients of the linear PO model is maximum likelihood estimation based on the conditional likelihood. Their approach is applicable to both current status and more generally, interval-censored data. Wang and Dunson (2011) used monotone splines for approximating the baseline log-odds function, and McMahan et al. (2013) proposed new EM algorithms to analyze current status data under two popular semiparametric regression models including the PO model. They used monotone splines to model the baseline odds function and provided variance estimation in a closed form.

However, many predictors in a regression model do not necessarily affect the response linearly. In order to consider the non-linear covariate effects, we have to think about a more flexible model. There are not many articles for this in the literature, but as a special case of the transformation model for this, Ma and Kosorok (2005) presented a partly linear proportional odds (PL-PO) model which is defined

as follows

$$\text{logit}\{1 - S(t|V, X_1)\} = \text{logit}\{1 - S_0(t)\} + \alpha^\top V + \phi(X_1), \tag{6.2}$$

where ϕ is an unknown function relating to the one-dimensional non-linear covariate X_1. They used penalized MLE to estimate parameters of their model and made inference based on a block jackknife method. To analyze right-censored data, Lu and Zhang (2010) studied a PL transformation model where (6.2) is a special case. They applied a martingale based estimating equation approach, consisting of both global and kernel-weighted local estimation equations to estimate parameters of their model and presented asymptotic properties of their estimators. They also used a resampling method to estimate the asymptotic variance-covariance matrix of the estimates. In these models they could handle just one non-linear covariate $X_1 \in \mathbb{R}$.

Since in many real applications we may face with more than one non-linear covariate, we need to think about a model to incorporate high dimensionality. In this chapter, we propose a partially linear single-index proportional odds (PLSI-PO) model to deal with high dimensionality in analyzing current status data. This model reduces the dimension of data through a single-index term and it involves the log-odds of the baseline survival function which is unknown and has to be estimated. We use B-splines to approximate log-odds of baseline survival function, $\ln\{S_0(\cdot)/(1 - S_0(\cdot))\}$, and also to approximate the link function of single-index term, $\psi(\cdot)$. Asymptotic properties of the estimators are derived using the theory of counting processes.

6.2 Model Assumptions and Methods

In many situations there is limited information for a single observation about the event of interest and the only information that we have is that it has occurred before or after the examination time. The failure time is either left- or right-censored instead of being observed exactly and we only observe whether or not the event time, T, has occurred before some monitoring time C. In this case, we are dealing with current status data. Suppose $Z = (V^\top, X^\top)^\top$ is a covariate vector. In terms of the odds ratio of $S(t|Z)$ which is the survival function of T conditional on Z, we define the PLSI-PO model as follows

$$\frac{S(t|Z)}{1 - S(t|Z)} = \frac{S_0(t)}{1 - S_0(t)} \exp\{-\alpha^\top V - \psi(\beta^\top X)\}, \tag{6.3}$$

where $\alpha = (\alpha_1, \ldots, \alpha_q)^\top$ and $\beta = (\beta_1, \ldots, \beta_p)^\top$ are q- and p-dimensional regression coefficient vectors, respectively, $S_0(t)$ is the baseline survival function corresponding to $V = 0$, $X = 0$; and $\psi(\cdot)$ is the unknown link function for the single-index term. Following Huang and Liu (2006) and Sun et al. (2008), for identifiability of the model, we consider some constraints. For this respect,

we assume $\beta_1 > 0$ in order to the sign of β be identifiable, and because any constant scale can be absorbed in $\psi(\cdot)$ we can only estimate the direction of β and the scale of it is not identifiable, so it is required that $\|\beta\| = 1$, where $\|a\| = (a^\top a)^{1/2}$ is the Euclidean norm for any vector a. On the other hand, since there are two nonparametric functions and thus any constant in one of them can be assimilated in the other one, for identifiability of the model, we assume $\mathrm{E}(V) = 0$ and $\mathrm{E}\{\psi(\beta^\top X)\} = 0$.

By taking natural logarithm of both sides of (6.3), we have the model as follows

$$\ln\left\{\frac{1-S(t|Z)}{S(t|Z)}\right\} = \ln\left\{\frac{1-S_0(t)}{S_0(t)}\right\} + \alpha^\top V + \psi(\beta^\top X),$$

that is

$$\mathrm{logit}\{1-S(t|Z)\} = \mathrm{logit}\{1-S_0(t)\} + \alpha^\top V + \psi(\beta^\top X), \tag{6.4}$$

where $\mathrm{logit}(u) = \ln\{u/(1-u)\}$ for $0 < u < 1$ and $S_0(\cdot) = e^{-\Lambda_0(\cdot)}$ is the baseline survival function.

In the setting of current status data, we do not observe the values of T directly, thus the observations are in the form of independent samples of $\{C_i, \delta_i, V_i, X_i\}_{i=1}^n$, drawn from the population $\{C, \delta, V, X\}$, where censoring time C is continuous on the interval $[a_c, b_c]$ with the hazard function $\lambda_c(t|Z) = \nu(t|Z)$; $\delta = I(C \le T)$ is the censoring indicator where $\delta = 1$ if the event of interest has not occurred by time C and otherwise $\delta = 0$.

Since $H(\cdot) = \mathrm{logit}\{1 - S_0(\cdot)\} = -\mathrm{logit}\{S_0(\cdot)\}$ and $\psi(\cdot)$ are two unknown functions of the model (6.4), we need to estimate them. We use the B-spline method to approximate the two unknown functions $H(\cdot) = -\mathrm{logit}\{S_0(\cdot)\}$ and $\psi(\cdot)$, and then we use a maximum likelihood approach to estimate parameters of the model.

Suppose $\mathbb{L}_n$ is the collection of nonnegative and nondecreasing functions $\Lambda_0(\cdot)$ on $[a_c, b_c]$ and $\mathscr{L}_n$ be the space of polynomial splines of order $\rho_L \ge 1$, where each functional element of this space is defined on a sub interval of a partition. To get faster convergence rate in B-spline approximation, we assume for $\rho_L \ge 2$ and $0 \le r \le \rho_L - 2$, each functional element of this space is r times continuously differentiable on $[a_c, b_c]$. For each $\Lambda_0 \in \mathbb{L}_n$, we have

$$\begin{aligned}\mathrm{logit}\{1 - e^{-\Lambda_0(t)}\} &= \mathrm{logit}\{1 - S_0(t)\} \\ &= -\mathrm{logit}\{S_0(t)\} = \sum_{k=1}^{df_L} \tau_k L_k(C) = \tau^\top L(C),\end{aligned} \tag{6.5}$$

where $L(C) = (L_1(C), \ldots, L_{df_L}(C))^\top$ is the vector of B-spline basis functions with $L_k(C) \in \mathscr{L}_n$ for each $k = 1, \ldots, df_L$ and $\tau = (\tau_1, \ldots, \tau_{df_L})^\top$ is the vector of B-spline coefficients. We have $df_L = K_L + \rho_L$ is the degree of freedom (number of basis functions) with K_L interior knots and B-splines of order ρ_L.

Let Ψ_n be the collection of $\psi(\cdot)$ functions on $[a_{xb}, b_{xb}]$ and $\mathscr{B}_n$ be the space of polynomial splines of order $\rho_B \geq 1$, with the same properties as $\mathscr{L}_n$. To satisfy the identifiability centering constraint, $\mathrm{E}\{\psi(\beta_0 X)\} = 0$, we focus on a subspace of spline functions $S^0 = \{s : s(x) = \sum_{\ell=1}^{df_B} \gamma_\ell B_\ell(\beta^\top x),\ \sum_{i=1}^n s(X_i) = 0\}$ with basis $\{B_1(\beta^\top X), \ldots, B_{df_B}(\beta^\top X)\}$ where $B_\ell(\beta^\top x) = B_\ell(\beta^\top x) - (\sum_{i=1}^n B_\ell(\beta^\top X_i)/n)$ for $\ell = 1, \ldots, df_B$, and mention that as to the empirical version of the constraint, this subspace is $df_{B1} = df_B - 1$ dimensional. So $\psi \in \Psi_n$ can be approximated at $\beta^\top X$ as follows

$$\psi(\beta^\top X) = \sum_{\ell=1}^{df_{B1}} \gamma_\ell B_\ell(\beta^\top X) = \gamma^\top B(\beta^\top X), \tag{6.6}$$

where $B(\beta^\top X) = (B_1(\beta^\top X), \ldots, B_{df_{B1}}(\beta^\top X))^\top$ is the vector of local normalized B-spline basis functions, $B_\ell(\beta^\top X) \in \mathscr{B}_n$, $\gamma = (\gamma_1, \ldots, \gamma_{df_{B1}})^\top$ is the vector of B-spline coefficients. Having K_B interior knots and B-splines of order ρ_B, the degree of freedom of the B-spline would be $df_B = K_B + \rho_B$.

For identifiability purpose we assume $\|\beta\| = 1$ and perform the delete-one-component method by defining $\tilde{\beta} = (\beta_2, \ldots, \beta_p)$ as a $(p-1)$-dimensional vector deleting the first component β_1. We also assume $\beta_1 > 0$ which could be implemented by considering $\beta_1 = \sqrt{1 - \|\tilde{\beta}\|^2} = \sqrt{1 - \sum_{k=2}^p \beta_k^2}$. Then, we have $\beta = (\sqrt{1 - \sum_{k=2}^p \beta_k^2}, \beta_2, \ldots, \beta_p)^\top$ where the true parameter $\|\tilde{\beta}_0\| < 1$. Therefore, β is infinitely differentiable in a neighborhood of the true parameter $\tilde{\beta}_0$. Since we use B-spline to estimate $H(\cdot) = \mathrm{logit}\{1 - e^{-\Lambda_0(\cdot)}\}$, the baseline hazard function is $\Lambda_0(\cdot) = \ln\left\{1 + e^{\tau^\top L(\cdot)}\right\}$ which has to be positive and nondecreasing. The positivity is guaranteed by property of a logarithmic function and we just need to satisfy the condition of being nondecreasing by putting a constraint on the coefficients of the basis functions, that is $\tau_1 \leq \cdots \leq \tau_{df_L}$.

Under suitable smoothness assumptions $\mathrm{logit}\{1 - S_0(\cdot)\}$ and $\psi(\cdot)$ can be well approximated by functions in $\mathscr{L}_n$ and $\mathscr{B}_n$, respectively. Therefore, we have to find members of $\mathscr{L}_n$ and $\mathscr{B}_n$ along with values for α and β that maximize the semiparametric log-likelihood function.

Now in Eq. (6.4) we replace B-spline approximations for $\mathrm{logit}\{1 - S_0(t)\}$ and $\psi(\beta^\top X)$ from (6.5) and (6.6), respectively. Considering $\mathrm{logit}\{1 - S(t|Z)\} = -\mathrm{logit}\{S(t|Z)\}$, (6.4) is equivalent to

$$\mathrm{logit}(p) = -\tau^\top L(C) - \alpha^\top V - \gamma^\top B(\beta^\top X), \tag{6.7}$$

where $p = p(t) = S(t|Z)$. Since for subject i, $i = 1, \ldots, n$, we have $S(C_i|Z_i) = \Pr(C_i \leq T_i|Z_i) = \mathrm{E}\{I(C_i \leq T_i)|Z_i\} = \mathrm{E}(\delta_i|Z_i)$, by assuming δ_i as a binary response, we can consider model (6.7) as a generalized linear model (GLM) with linear predictor $-\xi = -\tau^\top L(C) - \alpha^\top V - \gamma^\top B(\beta^\top X)$ and "logit" link. Then we use the GLM methods, available in many computer software packages, to estimate

parameters α, β, τ and γ. We use "`glm`" function in the `R` package to do that. The constraints set up for the model are not used for the estimated values obtained in this step, and we consider them as the initial values of the parameters of PLSI-PO model for the next step which is maximizing the semiparametric log-likelihood function subject to the mentioned constraints.

For current status data, the likelihood function at observed censoring times, C_i, given covariates, Z_i, is proportional to $\prod_{i=1}^{n}\{S(C_i|Z_i)\}^{\delta_i}\{1-S(C_i|Z_i)\}^{1-\delta_i}$ where for each $i = 1, \ldots, n$, $S(C_i|Z_i) = \exp\{-\Lambda(C_i|Z_i)\} = p_i(C_i) = p_i$. Thus, we can write the semiparametric log-likelihood function for PLSI-PO model as follows

$$\begin{aligned}\ell_O = \ell_O(\alpha, \beta, S_0, \psi) &= \sum_{i=1}^{n}\{\delta_i \ln(p_i) + (1-\delta_i)\ln(1-p_i)\} \\ &= \sum_{i=1}^{n}\left[\delta_i \ln\left\{\frac{1}{1+\exp\{-\text{logit}\{S_0(C_i)\} + \alpha^\top V_i + \psi(\beta^\top X_i)\}}\right\}\right. \\ &\quad \left. +(1-\delta_i)\ln\left\{\frac{\exp\{-\text{logit}\{S_0(C_i)\} + \alpha^\top V_i + \psi(\beta^\top X_i)\}}{1+\exp\{-\text{logit}\{S_0(C_i)\} + \alpha^\top V_i + \psi(\beta^\top X_i)\}}\right\}\right].\end{aligned} \tag{6.8}$$

Then, we plug in the B-spline approximations of $-\text{logit}\,(S_0(\cdot))$ and $\psi(\cdot)$ obtained from (6.5) and (6.6) into the semiparametric log-likelihood function (6.8). So we have the log-likelihood function as follows

$$\begin{aligned}\ell_O = \ell_O(\alpha, \beta, \tau, \gamma) &= \sum_{i=1}^{n}\left[\delta_i \ln\left\{\frac{1}{1+\exp\{\tau^\top L(C_i) + \alpha^\top V_i + \gamma^\top B(\beta^\top X_i)\}}\right\}\right. \\ &\quad \left. +(1-\delta_i)\ln\left\{\frac{\exp\{\tau^\top L(C_i) + \alpha^\top V_i + \gamma^\top B(\beta^\top X_i)\}}{1+\exp\{\tau^\top L(C_i) + \alpha^\top V_i + \gamma^\top B(\beta^\top X_i)\}}\right\}\right].\end{aligned} \tag{6.9}$$

Now we can estimate the parameters of our regression model, $(\alpha, \beta, \tau, \gamma)$, by maximizing the log-likelihood function (6.9) which has a parametric form after using the B-spline approximated values of the infinite-dimensional nuisance parameters. To maximize (6.9), we use a sieve method through an iterative algorithm subject to the mentioned constraints on coefficients τ and β; i.e., $\tau_1 \leq \cdots \leq \tau_{df_L}$ for monotonicity of $\Lambda_0(C)$ and $\beta_1 > 0$ and $\|\beta\| = 1$ for the purpose of identifiability in $\psi(\beta^\top X)$.

Some regularity conditions have to be satisfied in order to establish large sample properties of the estimators in PLSI-PO model. These conditions are as follows:

(C1) (i) T and C are independent given the covariate history Z. (ii) Censoring time, C, has an absolutely continuous distribution on $[a_c, b_c]$ where $0 < a_c < b_c < \infty$, with hazard function $\nu(t) = \nu(t, Z) = \lambda_c(t, Z)$ conditional on covariate vector Z.

(C2) Assume for any integer $s \geq 1$, there exist continuous and positive sth derivatives $\Lambda_0^{(s)}$ and $\psi^{(s)}$. Then let the finite-dimensional parameter spaces Θ_1 for α and Θ_2 for $\tilde{\beta}$ are bounded subsets of $\mathbb{R}^q$ and $\mathbb{R}^{p-1}$, respectively. Assume the true regression parameter values $\alpha_0 \in \Theta_1$ and $\tilde{\beta}_0 \in \Theta_2$ are interior points of the true functions $\Lambda_0 \in \mathbb{L}_n$, and $\psi \in \Psi_n$.

(C3) (i) For any $\alpha_0 \neq \alpha$ and $\beta_0 \neq \beta$ we have $\Pr(\alpha_0^\top V \neq \alpha^\top V) > 0$ and $\Pr(\beta_0^\top X \neq \beta^\top X) > 0$. (ii) Assume $\mathrm{E}(V) = 0$ and for the true parameter $\tilde{\beta}_0 \in \Theta_2$ and the true function $\psi(\cdot)$ assume $\mathrm{E}\{\psi(\beta_0^\top X)\} = 0$.

(C4) (i) Covariates V and X have bounded supports which are subsets of $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively. That is, there exist v_0 and x_0 such that $\|V\| \leq v_0$ and $\|X\| \leq x_0$ with probability 1. (ii) If we denote the distribution of T by F_0 such that $F_0(0) = 0$, then the support of C is strictly contained in the support of F_0, that is for $t_{F_0} = \inf\{t : F_0(t) = 1\}$ we have $0 < a_c < b_c < t_{F_0}$.

(C5) For a small $\varepsilon > 0$ we have $\Pr(T < a_c | C, V, X) > \varepsilon$ and $\Pr(T > b_c | C, V, X) > \varepsilon$ with probability one.

(C6) The baseline cumulative hazard function Λ_0 has strictly positive derivative on $[a_c, b_c]$, and the joint distribution function $G(c, v, x)$ of (C, V, X) has bounded second-order partial derivative with respect to c.

Condition (C1) is to satisfy non-informative censoring condition. (C2) is to ensure identifiability of the parameters, (C3) implies certain characteristics in order to apply spline smoothing techniques. (C4) bounds likelihood and score functions away from infinity at the boundaries of the support of the observed event time. (C5) is required to ensure that the probability of being either left-censored or right-censored is positive and bounded away from zero regardless of the covariate values. (C6) requires for the partial score functions (or partial derivatives) of the nonparametric components in the least favorable direction to be close to zero, so that the $\sqrt{n}$ convergence rate and asymptotic normality of the finite-dimensional estimator can be obtained. Similar conditions in a linear PH model for current status data are discussed in Huang (1996).

6.3 Theory of Estimation

We first assume that (6.6) holds, i.e., the link function ψ is a B-spline function with fixed knots and order, then model (6.3) contains only one nonparametric function of $S_0(t)$, we can calculate the information matrix of the estimators of the regression

parameters. When ψ is a smooth nonparametric function instead of a B-spline function, the derived theory works as an approximation. Our simulation results indicate the approximation is quite accurate and provides a practical solution for real data analysis.

6.3.1 Information Calculation

Considering observations $\{C_i, \delta_i, V_i, X_i\}_{i=1}^n$, we define two counting processes $N_{1i}(t) = \delta_i I(C_i \leq t)$ and $N_{2i}(t) = (1-\delta_i)I(C_i \leq t)$. Then let $N_i(t) = N_{1i}(t) + N_{2i}(t) = \delta_i I(C_i \leq t) + (1-\delta_i)I(C_i \leq t) = I(C_i \leq t)$ and $Y_i(t) = I(t \leq C_i)$ be at risk process for time point t. Having $N_{1i}(t)$ and $N_{2i}(t)$ with intensity processes $f_{N_{1i}}(t) = Y_i(t)\nu_i(t)p_i(t) = Y_i\nu_i p_i$ and $f_{N_{2i}}(t) = Y_i(t)\nu_i(t)(1-p_i(t)) = Y_i\nu_i(1-p_i)$, we define $M_{1i}(t)$ and $M_{2i}(t)$ as their corresponding compensated counting processes as follows

$$M_{1i}(t) = N_{1i}(t) - \int_0^t Y_i(s)\nu_i(s)p_i(s)ds,$$

$$M_{2i}(t) = N_{2i}(t) - \int_0^t Y_i(s)\nu_i(s)(1-p_i(s))ds,$$

which are martingales as shown in Martinussen and Scheike (2002). In the following, we shall drop t from $N_1(t)$, $N_2(t)$, $M_1(t)$, $M_2(t)$, $\nu(t)$, $Y(t)$ and $p(t)$ unless it needs to be specified. We can write the log-likelihood function given in (6.8), as follows

$$\ell_O = \ell_O(\alpha, \beta, \Lambda, \psi) = \sum_{i=1}^n \left\{ \int (\ln p_i)\, dN_{1i} + \int (\ln(1-p_i))\, dN_{2i} \right\}, \tag{6.10}$$

where $p_i = p(C_i) = S(C_i|Z_i)$ for each $i = 1, \ldots, n$ with

$$S(C_i|Z_i) = \frac{1}{1 + \exp[-\text{logit}\{S_0(C_i)\} + \alpha^\top V_i + \psi(\beta^\top X_i)]}.$$

Since we have an unknown parameter in one of the two unknown functions of our model, in the procedure of obtaining the efficient information bound, it is difficult to use projection onto a sumspace of two non-orthogonal L_2 spaces. Thus, we replace the B-spline approximated value of $\psi(\beta^\top X)$ from (6.7) and consider $H_0(\cdot) = -\text{logit}\{S_0(\cdot)\} = -\text{logit}\{e^{-\Lambda_0(\cdot)}\}$ as the only infinite-dimensional nuisance parameter of the model. We consider a simpler finite-dimensional parametric submodels contained within this semiparametric model. Assuming parametric submodel for the nuisance parameter $H(\cdot) = H_0(\cdot)$ as a mapping of the form $\eta \to H_{(\eta)}$ in $\{H_{(\eta)} : \eta \in \mathbb{R}^{q+(p-1)+df_{B1}}\}$. Then, characterize $H(\cdot)$ by a finite-dimensional

parameter η such that

$$\frac{\partial H_{(\eta)}}{\partial \eta}(t) = a(t) = a. \tag{6.11}$$

As we re-parametrized β as $\beta = (\beta_1, \tilde{\beta}^\top)^\top$, the marginal score vector for $\tilde{\theta} = (\tilde{\beta}^\top, \alpha^\top, \gamma^\top)^\top$ is obtained by partially differentiating $\ell_O(\tilde{\theta}, H_{(\eta)})$ given in (6.10), for one observation, with respect to $\tilde{\theta}$ such that

$$S_{\tilde{\theta}} = \frac{\partial \ell_O}{\partial \tilde{\theta}} = \left(S_{\tilde{\beta}}^\top, S_\alpha^\top, S_\gamma^\top\right)^\top,$$

where by defining $\tilde{X} = (X_2, \ldots, X_p)^\top$, with X_1 as the first element of X, and $\beta_1 = \sqrt{1 - \sum_{k=2}^{p} \beta_k^2}$, we have

$$S_{\tilde{\beta}} = \frac{\partial \ell_O}{\partial \tilde{\beta}} = \int \gamma^\top B'(\beta^\top X) \left(\tilde{X} - X_1 \frac{\tilde{\beta}}{\beta_1}\right) \{p\, dN_2 - (1-p) dN_1\},$$

$$S_\alpha = \frac{\partial \ell_O}{\partial \alpha} = \int V \{p\, dN_2 - (1-p) dN_1\},$$

and

$$S_\gamma = \frac{\partial \ell_O}{\partial \gamma} = \int B(\beta^\top X) \{p\, dN_2 - (1-p) dN_1\},$$

with

$$p = p(C) = S(C|Z) = \frac{1}{1 + \exp\left\{H(C) + \alpha^\top V + \gamma^\top B(\beta^\top X)\right\}}.$$

Knowing that $dN_1 = dM_1 + Y\nu p dt$ and $dN_2 = dM_2 + Y\nu(1-p)dt$ we can write

$$S_{\tilde{\theta}} = \int \tilde{U}^* \{p\, dM_2 - (1-p) dM_1\},$$

where $\tilde{U}^*$ is a $(q + (p-1) + df_{B1}) \times 1$ vector defined as follows

$$\tilde{U}^* = \left[\left\{\gamma^\top B'(\beta^\top X) \left(\tilde{X} - X_1 \frac{\tilde{\beta}}{\beta_1}\right)\right\}^\top, V^\top, \left\{B(\beta^\top X)\right\}^\top\right]^\top.$$

Let $e_1 = \exp\left\{\alpha^\top V + \gamma^\top B(\beta^\top X)\right\}$, and ℓ_o be the expression inside the integral in (6.10) such that $\ell_o = (\ln p)\, dN_1 + (\ln(1-p))\, dN_2$. For $H(\cdot) = -\text{logit}\{S_0(\cdot)\}$ we have $S_H(a) = \int S_1(a)$ where $S_1(a) = \frac{\partial \ell_o}{\partial H} \times \frac{\partial H(\eta)}{\partial \eta}$ and

$$\begin{aligned}
\frac{\partial \ell_o}{\partial H} &= \frac{\partial}{\partial H}\left\{\ln\left(\frac{1}{1+e^H e_1}\right) dM_1 + \ln\left(\frac{e^H e_1}{1+e^H e_1}\right) dM_2\right\} \\
&= \frac{\partial}{\partial H}\left\{-\ln\left(1+e^H e_1\right) dM_1 - \ln\left(\frac{1+e^H e_1}{e^H e_1}\right) dM_2\right\} \\
&= -\frac{e^H e_1}{1+e^H e_1} dM_1 + \frac{1}{1+e^H e_1} dM_2 \\
&= p\, dM_2 - (1-p) dM_1.
\end{aligned}$$

Thus, the score operator associated with H is as follows

$$S_H(a) = \int a(t)\left\{p\, dM_2 - (1-p) dM_1\right\}.$$

Under conditions (C1) to (C6), the efficient score for the finite-dimensional parameter $\tilde{\theta}$ is the difference between its score vector, $S_{\tilde{\theta}}$, and the score for a particular submodel of the nuisance parameter, $S_H(a)$. The particular submodel is the one with the property that the difference is uncorrelated with the scores for all other submodels of the nuisance parameters. Thus, the score operator for $\tilde{\theta}$ is as follows

$$S^*_{\tilde{\theta}} = S_{\tilde{\theta}} - S_H(a).$$

We have to find $a^* = a^*(t)$ such that

$$S^*_{\tilde{\theta}} = S_{\tilde{\theta}} - S_H(a^*) = \int \left(\tilde{U}^* - a^*\right)\left\{p\, dM_2 - (1-p) dM_1\right\} \tag{6.12}$$

be orthogonal to any other $S_H(a) \in A_H$ where $A_H = \{S_H(a) : a \in L_2(P_C)\}$ and $L_2(P_C) = \{a : \text{E}[\|a(C)\|^2 p(C)(1-p(C))] < \infty\}$ (i.e., $\text{E}(S^*_{\tilde{\theta}} S_H) = 0$). Thus, we have

$$\text{E}(S^*_{\tilde{\theta}} S_H) = E[\{S_{\tilde{\theta}} - S_H(a^*)\} S_H(a)] = 0, \tag{6.13}$$

for any $a \in L_2(P_C)$. The orthogonality Eq. (6.13) is equivalent to

$$\text{E}\left[\int (\tilde{U}^* - a^*)\left\{p\, dM_2 - (1-p) dM_1\right\} \times \int a\left\{p\, dM_2 - (1-p) dM_1\right\}\right]$$

$$= \mathrm{E}\left[\int (\tilde{U}^* - a^*)\{p\, dM_2 - (1-p)dM_1\} \times \int a\{p\, dM_2 - (1-p)dM_1\}\right]$$

$$= \mathrm{E}\left[\int (\tilde{U}^* - a^*)ap^2 d\langle M_2\rangle + \int (\tilde{U}^* - a^*)a(1-p)^2 d\langle M_1\rangle\right] = 0. \qquad (6.14)$$

Then, since $d\langle M_1\rangle = p\nu Y dt$ and $d\langle M_2\rangle = (1-p)\nu Y dt$, it is equivalent to

$$\int \left[-a\mathrm{E}\left\{(\tilde{U}^* - a^*)p(1-p)Y\nu dt\right\}\right] = 0,$$

and then

$$\int \left[-a\left[\mathrm{E}\left\{\tilde{U}^* p(1-p)Y\nu\right\} + a^*\mathrm{E}\left\{p(1-p)Y\nu\right\}\right]\right] dt = 0.$$

So considering $a \neq 0$, we have

$$a^* = \frac{\mathrm{E}\left\{\tilde{U}^* p(1-p)Y\nu\right\}}{\mathrm{E}\left\{p(1-p)Y\nu\right\}}.$$

By plugging in a^* into (6.12) we have the efficient score for $\tilde{\theta}$ as follows

$$S^*_{\tilde{\theta}} = \int \left[\tilde{U}^* - \frac{\mathrm{E}\{\tilde{U}^* p(1-p)Y\nu\}}{\mathrm{E}\{p(1-p)Y\nu\}}\right]\{p\, dM_2 - (1-p)dM_1\}. \qquad (6.15)$$

The empirical version of the efficient score for $\tilde{\theta}$ is

$$S(\tilde{\theta}, \Lambda) = \sum_{i=1}^{n} \int \left(\tilde{U}^*_i - \frac{S_1^{(\tilde{\theta})}}{S_0^{(\tilde{\theta})}}\right)\{p_i\, dM_{2i} - (1-p_i)dM_{1i}\}, \qquad (6.16)$$

where

$$S_u^{(\tilde{\theta})} = S_u^{(\tilde{\theta})}(t) = \sum_i (\tilde{U}^*_i)^{\otimes u}\, p_i\, (1-p_i) Y_i \nu_i, \quad \text{for } u = 0, 1,$$

with $\otimes$ denoting the Kronecker operation defined as $b^{\otimes 0} = 1$, $b^{\otimes 1} = b$ and $b^{\otimes 2} = bb^\top$. Since $\nu_i = \nu(t|Z_i)$ is an unknown function of the covariate vector, in order to obtain an estimated value for $S_u^{(\tilde{\theta})}$, we need to estimate ν_i. Following the idea of Martinussen and Scheike (2002), we use a simple kernel estimator for estimating value of $S_u^{(\tilde{\theta})}$, where $\nu_i(t)dt$ is replaced by the convolution of the kernel estimator $K_b(\cdot)$ and $dN_i(s)$ such that $\hat{\nu}(t|Z_i)dt = K_b(s-t)dN_i(s)$. The kernel function satisfies $K_b(\cdot) = (1/b)K(\cdot/b)$, and $b > 0$ is the bandwidth of the kernel estimator. We also assume that $\int K_b(u)du = 1$, $\int uK_b(u)du = 0$ and the kernel has compact support.

Therefore, after obtaining the semiparametric maximum likelihood estimator $\hat{\tilde{\theta}}$, using the plug-in method, $S_u^{(\tilde{\theta})}$ is estimated as follows:

$$\hat{S}_u^{(\tilde{\theta})} = \hat{S}_u^{(\tilde{\theta})}(t) = \sum_{i=1}^{n} \int \hat{p}_i(s)\,(1-\hat{p}_i(s))\,Y_i(s)(\hat{\tilde{U}}_i^*)^{\otimes u} K_b(s-t)dN_i(s), \quad \text{for } u=0,1,$$

where

$$\begin{aligned}\hat{p}_i(s) &= \frac{1}{\left[1+\exp\left\{-\text{logit}\left\{\hat{S}_0(s)\right\}+\hat{\alpha}^\top V+\hat{\gamma}^\top B(\hat{\beta}^\top X)\right\}\right]} \\ &= \frac{1}{\left[1+\exp\left\{-\text{logit}\left\{e^{-\hat{\Lambda}(s)}\right\}+\hat{\alpha}^\top V+\hat{\gamma}^\top B(\hat{\beta}^\top X)\right\}\right]},\end{aligned}$$

with kernel function $K_b(\cdot) = (1/b)K(\cdot/b)$, and $b>0$ is the bandwidth of the kernel estimator. We see that $\hat{\nu}_i(t)dt = \hat{\nu}(t|Z_i)dt = K_b(s-t)dN_i(s) = K_b(C_i-t)I(C_i \le t)$ for any $t \in [a_c, b_c]$, so we have

$$\hat{S}_u^{(\tilde{\theta})}(t) = \sum_{i=1}^{n} \hat{p}_i(C_i)\,(1-\hat{p}_i(C_i))\,Y_i(C_i)\{\hat{\tilde{U}}_i^*(C_i)\}^{\otimes u} K_b(C_i-t), \quad \text{for } u=0,1.$$

The Epanechnikov kernel function is used here which is defined as

$$K_b(u) = (1/b)(3/4)\left(1-(u/b)^2\right) I(|u/b| \le 1).$$

Then we can write the information matrix at $\tilde{\theta}_0$ as follows:

$$I(\tilde{\theta}_0) = \mathrm{E}\left[\int \left\{\tilde{U}^* - \mathrm{E}\left(\tilde{U}^* Y\nu p(1-p)\right) E^{-1}\left(Y\nu p(1-p)\right)\right\}^{\otimes 2} p(1-p)Y\nu dt\right].$$

Using the central limit theorem for martingales, $n^{-1/2}S(\tilde{\theta}_0, \hat{\Lambda})$ converges in distribution to a normal distribution with mean zero and covariance matrix Σ_1 which can be consistently estimated by

$$\hat{\Sigma}_1 = \frac{1}{n}\sum_{i=1}^{n} \int \left(\hat{\tilde{U}}_i^* - \frac{\hat{S}_1^{(\tilde{\theta})}}{\hat{S}_0^{(\tilde{\theta})}}\right)\left(\hat{\tilde{U}}_i^* - \frac{\hat{S}_1^{(\tilde{\theta})}}{\hat{S}_0^{(\tilde{\theta})}}\right)^\top \left\{\hat{p}_i^2\, dN_{2i} + (1-\hat{p}_i)^2 dN_{1i}\right\},$$

and $\hat{\Sigma}_1$ converges in probability to Σ_1 and therefore, we have $n^{1/2}(\hat{\tilde{\theta}} - \tilde{\theta}_0)$ converges in distribution to a mean zero normal distribution with covariance matrix $\Sigma = I^{-1}(\tilde{\theta}_0)\Sigma_1 I^{-1}(\tilde{\theta}_0)$. The robust sandwich estimator of the variance is given by $\hat{\Sigma} = \hat{I}^{-1}(\hat{\tilde{\theta}})\hat{\Sigma}_1\hat{I}^{-1}(\hat{\tilde{\theta}})$. With the consistent estimator of Λ we can conclude

that $\hat{\Sigma}_1 = \hat{I}(\hat{\tilde{\theta}}) + o_p(1)$, and thus, $n^{1/2}(\hat{\tilde{\theta}} - \tilde{\theta}_0)$ converges in distribution to a mean zero random vector with covariance matrix $I^{-1}(\tilde{\theta}_0)$ estimated by $\hat{I}^{-1}(\hat{\tilde{\theta}})$. Therefore, our obtained estimators are efficient. A rigorous proof of the consistency and asymptotic normality of the semimparametric estimator $\hat{\tilde{\theta}}$ can be obtained by using the theory developed by Huang (1996) for empirical processes with current status data, assuming that (6.6) holds, i.e., the link function ψ is a B-spline function with fixed knots and order.

6.3.2 *Inference*

To obtain the variance-covariance matrix of $\hat{\theta} = (\hat{\beta}^\top, \hat{\alpha}^\top, \hat{\gamma}^\top)^\top$, we define a map $G : (\tilde{\beta}, \alpha, \gamma) \to (\beta, \alpha, \gamma)$. Then after obtaining the variance-covariance matrix of $\hat{\tilde{\theta}} = (\hat{\tilde{\beta}}^\top, \hat{\alpha}^\top, \hat{\gamma}^\top)^\top$ as $\hat{I}^{-1}(\hat{\tilde{\theta}}) = \mathrm{Var}(\hat{\tilde{\beta}}, \hat{\alpha}, \hat{\gamma})$, by the delta method we have

$$\mathrm{Var}(\hat{\beta}, \hat{\alpha}, \hat{\gamma}) = \mathrm{Var}(G(\hat{\tilde{\beta}}, \hat{\alpha}, \hat{\gamma})) = G'(\hat{\tilde{\beta}}, \hat{\alpha}, \hat{\gamma})\mathrm{Var}(\hat{\tilde{\beta}}, \hat{\alpha}, \hat{\gamma})G'^{\top}(\hat{\tilde{\beta}}, \hat{\alpha}, \hat{\gamma}), \tag{6.17}$$

where G' is a $(p + q + df_{B1}) \times ((p - 1) + q + df_{B1})$ matrix and can be calculated as follows:

$$\begin{aligned} G'(\tilde{\beta}, \alpha, \gamma) &= \frac{\partial G(\tilde{\beta}, \alpha, \gamma)}{\partial(\tilde{\beta}, \alpha, \gamma)} = \frac{\partial(\beta, \alpha, \gamma)}{\partial(\tilde{\beta}, \alpha, \gamma)} \\ &= \begin{pmatrix} -\frac{\beta_2}{\beta_1} \cdots & -\frac{\beta_p}{\beta_1} & 0 \cdots 0 \\ & I_{(p-1)+q+df_{B1}} & \end{pmatrix}. \end{aligned}$$

Then, $\mathrm{Var}(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$ can be estimated using (6.17). $\mathrm{Var}(\hat{\beta})$, $\mathrm{Var}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\alpha})$ are estimated by the corresponding block matrices in $\mathrm{Var}(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$. Inferences, such as Wald-type tests and confidence intervals for the parameters, can be made using these estimated variances. Moreover, for any $s \in$ the support of $\beta_0^\top X$, an approximate $(1 - \vartheta)100\,\%$ confidence band for $\psi(s) = \gamma^\top B(s)$ is obtained as follows

$$\hat{\psi}(s) \pm Z_{\vartheta/2} \times \mathrm{SE}(\hat{\psi}(s)),$$

where $\hat{\psi}(s) = \hat{\gamma}^\top B(s)$, $\mathrm{SE}(\hat{\psi}(s)) = \{\mathrm{Var}(\hat{\psi}(s))\}^{1/2} = \{B^\top(s)\mathrm{Var}(\hat{\gamma})B(s)\}^{1/2}$ and $Z_{\vartheta/2}$ is the upper quantile of the standard normal distribution.

6.3.3 Implementation

Because the unknown functions of this model are approximated using B-spline approximation, it may cause some numerical instability in the functional estimation when available data are sparse. The monitoring time C tends to be sparse in the right tail of the distribution, and the estimator of $H(t) = \text{logit}\{1 - S_0(t)\} = -\text{logit}\{S_0(t)\}$ may deviate from the true curve there. To overcome this problem, as suggested by Gray (1992), we add a penalty term in the estimation, such that the penalty functions in the penalized approach pull the estimates away from very extreme values and the estimates from the penalized maximization are less biased and better behaved. Since the distribution of $\beta^\top X$ is less sparse in the tails than that of C, we don't impose a penalty on $\psi(\cdot)$. From (6.10), the penalized likelihood function becomes

$$\ell_P(\alpha, \beta, \tau, \gamma) = \ell_O(\alpha, \beta, \tau, \gamma) - \frac{1}{2}\zeta \int \{H'(t)\}^2 \, dt, \tag{6.18}$$

where $H(t)$ is a quadratic B-spline defined in (6.5), and ζ is a penalty tuning parameter which controls the smoothing. When ζ is close to zero it means that there is no penalty and when it goes to ∞ it forces $H(t)$ to become a constant. Because the penalty term in (6.18) is a quadratic function of τ, it can be written as follows

$$\frac{1}{2}\zeta \tau^\top \mathbf{P} \tau,$$

where $\mathbf{P} = (\int_{a_c}^{b_c} L_s'(t) L_r'(t) dt)_{1 \le s \le df_L, 1 \le r \le df_L}$ is a $df_L \times df_L$ nonnegative definite matrix, which can be approximated by a Monte Carlo integration method.

We use the sieve method for maximum penalized log-likelihood estimation of the parameters and functions in the model. We maximize the penalized log-likelihood function $\ell_P(\theta)$ in (6.18) as the objective function subject to some constraints indicated as $g_u(\theta)$. Then, we apply the adaptive barrier algorithm for solving a constrained optimization problem through the function `constrOptim.nl` in the `R` package `alabama`.

The iterative algorithm we use to maximize ℓ_P follows these steps:

- **Step 0**: Considering that the direction of β can be correctly estimated, $\alpha^{(0)}$, $\tau^{(0)}$, $\gamma^{(0)}$ can be obtained from the GLM method by fixing β at an initial value $\beta^{(0)}$.
- **Step 1**. In step k, given current values $\alpha^{(k)}$, $\tau^{(k)}$, $\gamma^{(k)}$, update the value of $\beta^{(k)}$ by maximizing the log-likelihood function given in (6.10) subject to the constraint $1 - \sum_{\ell=2}^{p} \beta_\ell^2 > 0$ which ensures the constraints $\beta_1 > 0$ and $\|\beta\| = \sqrt{\sum_{\ell=1}^{p} \beta_\ell^2} = 1$. In this respect, we use the barrier method which is implemented by the `constrOptim.nl` function in the `R` package `alabama`. Denote the updated value by $\beta^{(k+1)}$.

- **Step 2**. Having $\beta^{(k+1)}$, update the values of $\alpha^{(k)}$, $\tau^{(k)}$, $\gamma^{(k)}$ simultaneously through GLM with the binary response δ and linear predictor $\xi = \tau^\top L(C) + \alpha^\top V + \gamma^\top B(\beta^\top X)$ with logit link. Then by letting $\omega = (\alpha^\top, \tau^\top, \gamma^\top)^\top$, we use the Newton-Raphson method to obtain $\omega^{(k+1)} = ({\alpha^{(k+1)}}^\top, {\tau^{(k+1)}}^\top, {\gamma^{(k+1)}}^\top)^\top$ through maximizing the log-likelihood function ℓ_O given in (6.10) without considering the constraint on τ and ξ. This is implemented by the `nlminb` function in `R` which uses a quasi-Newton algorithm.
- **Step 3**. Using the same procedure as **Step 1**, considering $\tau_1^{(k+1)} \leq \cdots \leq \tau_{df_L}^{(k+1)}$ as the constraints on τ, we further update the value of $\tau^{(k+1)}$ using `constrOptim.nl`.
- **Step 4**. Further update $\alpha^{(k+1)}$ to obtain $\alpha^{(k+2)}$ by fixing other parameters in the $(k+1)$ step and maximizing the likelihood function. For doing this, we use `nlminb` in `R`.
- **Step 5**. Repeat Step 1 to 4 until a certain convergence criterion is met. Note: the two further updates in **Step 3** and **Step 4** considering the constraints on τ produce better results than skipping these two steps.

Finally, after m iterations where the algorithm converges, we use $\beta^{(m)}$, $\tau^{(m)}$, $\alpha^{(m)}$ and $\gamma^{(m)}$ as the estimated values for β, τ, α and γ, respectively. Variance estimation is presented using the above variance estimators.

6.4 Simulation Studies

To evaluate the finite-sample performance of our estimators, we conduct a simulation study with current status data for the PLSI-PO model. The failure time, T, is generated from model (6.3) through the inverse transform sampling method and considering a Weibull distribution for the baseline survival function with scale parameter 1 and shape parameter $w = 2$. Thus T has the following form

$$T = \left[-\ln \frac{U \exp\{\alpha^\top V + \psi(\beta^\top X)\}}{(1-U) + U \exp\{\alpha^\top V + \psi(\beta^\top X)\}} \right]^{1/w},$$

where $U \sim \text{Uniform}(0,1)$, and the single-index function is defined as $\psi(\beta_0^\top X) = \sin(\beta_0^\top X)$. Two covariate vectors are considered, one is $q = 2$ dimensional linear covariate vector $V = (V_1, V_2)^\top$ and the other is $p = 3$ dimensional non-linear covariate vector $X = (X_1, X_2, X_3)^\top$. We assume $\alpha_0 = (0.5, -1)^\top$ and $\beta_0 = (2, -1, -1)^\top/\sqrt{6}$ and generate X_1, X_2, X_3 from continuous uniform distribution on interval $(-4, 4)$ and let $V_1 \sim \text{Uniform}(1, 4) - 2.5$ and $V_2 \sim \text{Bernoulli}(0.5) - 0.5$. The covariates satisfy condition (C3) such that $\text{E}(V_1) = \text{E}(V_2) = \text{E}\{\psi(\beta_0^\top X)\} = 0$. In order to satisfy the identifiability condition on the link function, we center $\psi(\beta^\top X)$ as $\psi(\beta^\top X_i) - (1/n)\sum_j \psi(\beta^\top X_j)$ for $i = 1, \ldots, n$.

The censoring time, C, is confined to interval $[a_c, b_c] = [0.01, 3.00]$ and generated from a truncated exponential distribution, i.e.,

$$C = (-1/\lambda_c) \ln \left[\exp(-\lambda_c a_c) - U \left\{\exp(-\lambda_c a_c) - \exp(-\lambda_c b_c)\right\}\right],$$

where $\lambda_c = \lambda_{c0} + (0.5)(V_1 + V_2) + (0.1)(X1 + X_2 + X_3)$ and $\lambda_{c0} = 1$. In kernel estimation of the covariance, we use bandwidth $b = bf \times n^{(-1/5)} \times sd(C)$ computed from the data, where $bf = 1/15$.

We use cubic B-spline basis functions (order=4) for $\psi(\beta^\top X)$ and quadratic basis functions (order=3) for $H(C) = -\text{logit}\{S_0(C)\}$ to approximate the two unknown curves. The BIC method is applied to find the optimized value for the number of B-spline basis functions indicated by degree of freedom, df, and the number of interior knots, $K = df$ – (order of B-spline basis functions). That is, we choose the value of (df_L, df_B) such that it locally minimizes the BIC objective function given as follows

$$BIC(df_L, df_B) = -2\ell_O + \ln(n)\{(p-1) + q + df_L + (df_B - 1)\},$$

where ℓ_O is the log-likelihood function given in (6.9). A large value of BIC implies lack of fit. Various forms of BIC have been proposed in the literature and tested for knots selection in semiparametric models (For example see He et al. (2002)).

Table 6.1 summarizes the simulation results based on 1000 replications with sample sizes 200, 400 and 800. In this table, we can see that the bias for the estimated values of $\beta = (\beta_1, \beta_2, \beta_3)$ and $\alpha = (\alpha_1, \alpha_2)$ are reasonably small. The Monte Carlo standard deviations of the estimates which are shown as StDev$(\hat{\cdot})$ are very close to the estimated average standard errors of the estimates indicated

Table 6.1 (PLSI-PO) Simulation results for estimation of β and α using the sieve MLE

Sample size (n)	Summary statistics	True β			True α	
		$\beta_1 = \frac{2}{\sqrt{6}}$	$\beta_2 = \frac{-1}{\sqrt{6}}$	$\beta_3 = \frac{-1}{\sqrt{6}}$	$\alpha_1 = 0.5$	$\alpha_2 = -1.0$
200	Bias	−0.0272	0.0608	0.0896	0.0222	−0.1227
	StDev$(\hat{\cdot})$	0.1530	0.2438	0.2687	0.3968	0.6117
	Avg.{SE$(\hat{\cdot})$}	0.1638	0.2360	0.2364	0.4827	0.7811
	Cov. prob.	0.9327	0.9328	0.9267	0.9754	0.9769
400	Bias	−0.0103	0.0029	0.0050	0.0125	−0.0438
	StDev$(\hat{\cdot})$	0.0614	0.0967	0.1006	0.2206	0.3735
	Avg.{SE$(\hat{\cdot})$}	0.0658	0.1028	0.1013	0.2619	0.4291
	Cov. prob.	0.9585	0.9351	0.9347	0.9704	0.9621
800	Bias	−0.0035	0.0013	0.0012	0.0084	−0.0304
	StDev$(\hat{\cdot})$	0.0410	0.0641	0.0635	0.1381	0.2374
	Avg.{SE$(\hat{\cdot})$}	0.0416	0.0660	0.0687	0.1608	0.2647
	Cov. prob.	0.9593	0.9438	0.9385	0.9687	0.9643

BIAS empirical bias, *STDEV* sample empirical standard deviation, *AVG.{SE}* estimated average standard error, *COV. PROB.* empirical coverage probability of the 95 % confidence interval

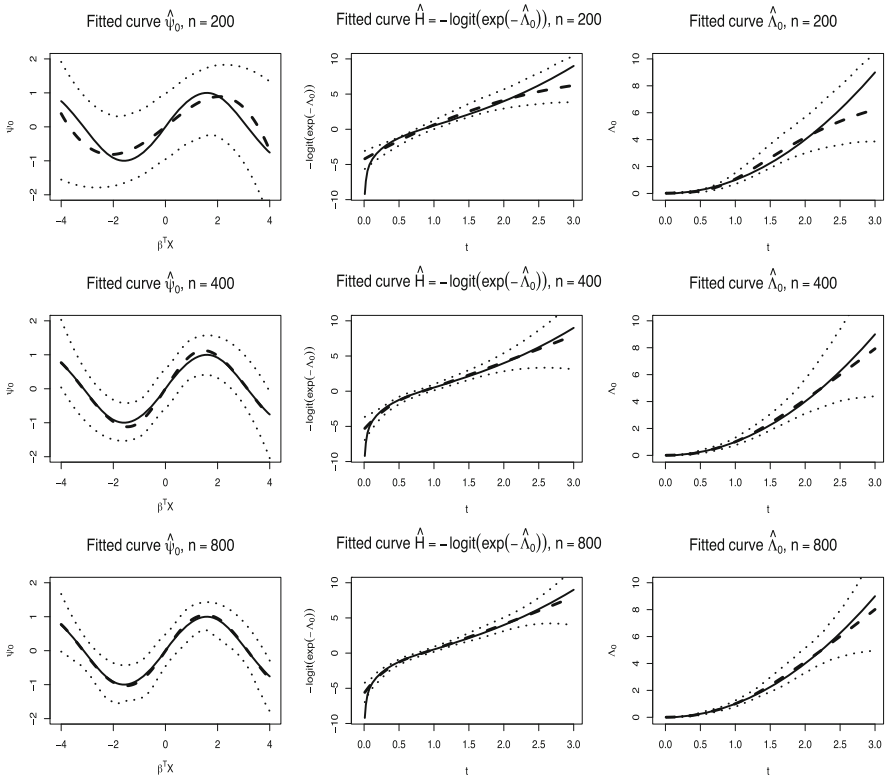

Fig. 6.1 (PLSI-PO) True and estimated curves: *Left*: The estimated ψ curves; *Middle*: The estimated curves for $-\text{logit}\{S_0\}$; *Right*: The estimated Λ_0 curves; corresponding to the sample sizes $n = 200, 400, 800$. *Solid lines* show the true curves, *dashed lines* the estimated curves and *dotted lines* illustrate the 95 % point-wise confidence bands based on Monte Carlo results

by Avg.$\{\text{SE}(\hat{\cdot})\}$. The Monte Carlo coverage probabilities of the 95 % confidence intervals are shown as Cov. prob. which are very close to the nominal level specifically for larger sample size.

Plots in Fig. 6.1 show the curves indicating the estimated nuisance parameters $\psi(\cdot)$, $H(\cdot) = -\text{logit}\{\exp(-\Lambda_0(\cdot))\} = -\text{logit}\{S_0(t)\}$ and $\Lambda_0(\cdot)$. It is seen that the fitted curves match the true functions closely, indicating good performance of the proposed method. Figures 6.2 and 6.3 illustrate the histograms for the estimated values of β and α, respectively, which are close to normal probability density curves.

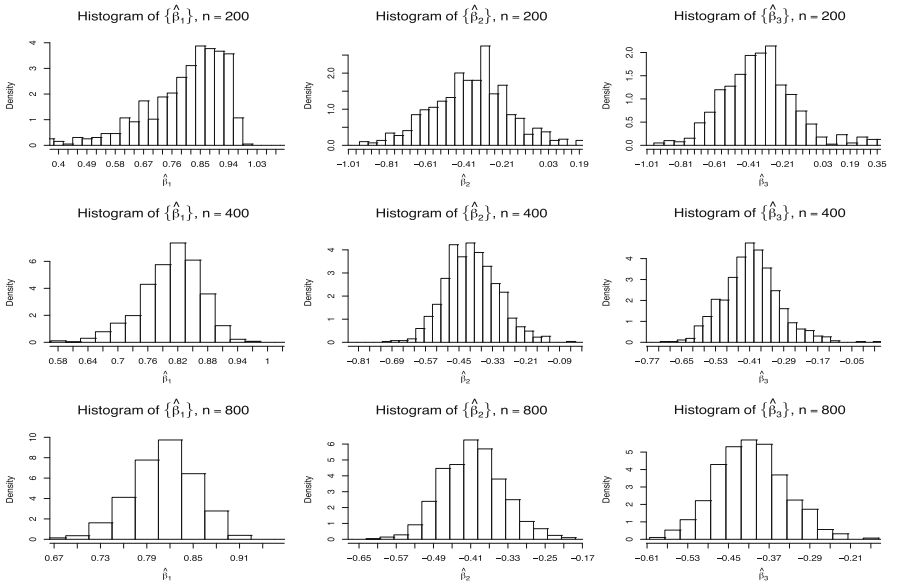

Fig. 6.2 (PLSI-PO) Histogram of estimated β values including $\hat{\beta}_1$ (on the *left*), $\hat{\beta}_2$ (in the *middle*) and $\hat{\beta}_3$ (on the *right*) corresponding to the sample sizes n= 200, 400, 800

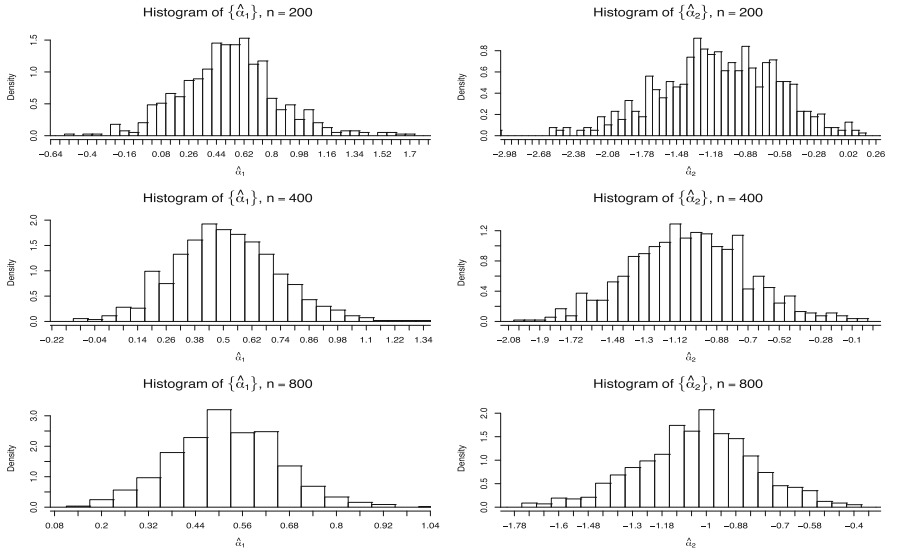

Fig. 6.3 (PLSI-PO) Histogram of estimated α values including $\hat{\alpha}_1$ (on the *left*) and $\hat{\alpha}_2$ (on the *right*) corresponding to the sample sizes n= 200, 400, 800

6.5 Real Data Analysis

In this section, we apply PLSI-PO model to the Acute kidney injury (AKI) dataset which is a typical kidney disease syndrome with substantial impact on both short and long-term clinical outcomes. A study was conducted at the University of Michigan Hospital on 170 hospitalized adult patients with AKI to identify risk factors associated with renal recovery in those who required renal replacement therapy (RRT). This study is conducted in order to help clinicians in developing strategies to prevent non-recovery and improve patient's quality of life. Data collection included patient characteristics, laboratory data, details of hospital course and degree of fluid overload at RRT initiation. For each of the patients, his/her time of the inception of dialysis was recorded along with time of hospital discharge, which may be regarded as a monitoring time. In this study, the investigators only observed patient's current status of renal recovery at discharge time but did not know exactly when renal function recovery occurred. More details about the study background and preliminary findings can be found in Heung et al. (2012) and Lu and Song (2015).

To assess the relationship between the hazard of occurrence of renal recovery and the clinical factors, let T be the failure time indicating the number of days from the time of starting dialysis to the date of renal recovery, and C be the monitoring time which is the time of hospital discharge. Two nonlinear covariates are baseline serum creatinine (BScr) and use of vasopressor (VP), and the linear ones are age (Age) and gender (Gender). VP is coded as 1 for Yes and 0 for No. Gender is coded as 1 for male and 0 for female. Let $V = (V_1 = VP, V_2 = Gender)^\top$ is the linear covariate vector where V_1 is use of vasopressor and V_2 is gender, and $X = (X_1 = BScr, X_2 = Age)^\top$ represents the non-linear covariate vector where X_1 is baseline serum creatinine level and X_2 is age. For ith observation, $i = 1, \ldots, n$, we standardize X_{pi} as $(X_{pi} - \min_i(X_{pi}))/(\max_i(X_{pi}) - \min_i(X_{pi}))$ for $p = 1, 2$. Then, the support of X_{pi} is $[0, 1]$. Suppose $\delta = I(C \leq T)$ is the indicator of renal recovery situation at the time of discharge where $\delta = 0$ means recovered and $\delta = 1$ means not recovered at the time of hospital discharge. Let $\lambda(t; V, X)$ be the hazard function of recovery time, T. We apply PLSI-PO model to establish a relationship between these four covariates and the survival odds and consequently the hazard function of T. The estimated values for the parameters of the model fitted to the data are $\hat{\alpha}_1 = -1.3873$, $\hat{\alpha}_2 = -0.0375$, $\hat{\beta}_1 = 0.6314$ and $\hat{\beta}_2 = 0.7754$ with estimated standard errors equal to 0.5909, 0.5803, 0.0953 and 0.0776, respectively. The Z-test statistic values for α_1 and α_2 equal to -2.348 and -0.0647 , and the p-values equal to 0.019 and 0.948, implying that the use of VP has an effect on the survival log-odds of renal recovery but Gender is not significant in the PLSI-PO model.

The negative log-odds of the baseline survival function, the baseline cumulative hazard function and the link function of the single-index term are fitted using B-spline approximation with respectively 5 and 7 knots which are chosen by the BIC. Figure 6.4 shows the estimated curves for $-\text{logit}\{S_0(C)\}$ and $\psi(\beta^\top X)$. It can be seen from the curve for $\psi(\cdot)$ that its effect on the log-odds of death or the hazard is mostly deceasing specifically for values of the single-index less than 0.2 and greater

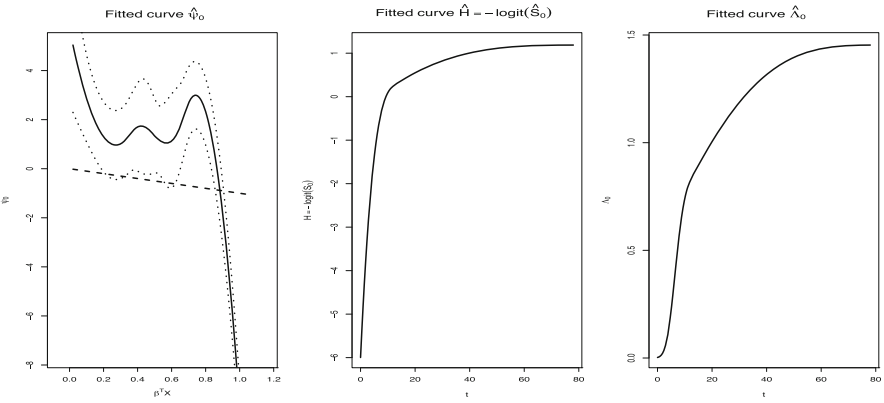

Fig. 6.4 (PLSI-PO) *Left*: The estimated nonparametric function $\psi(\cdot)$ with 95 % pointwise confidence intervals. The *solid line* is the estimated function, the *dashed line* is the identity function and the *dotted lines* are the 95 % pointwise confidence intervals; *Middle*: The estimated $-\text{logit}(S_0(\cdot))$ function; *Right*: The estimated cumulative hazard function $\Lambda_0(\cdot)$

than 0.8 and there is a mild fluctuation in between. the 95 % point-wise confidence bands are wider before the index value of 0.8. The identity link function lies outside the confidence band which indicates that the linear PO model is not appropriate for this data set. The negative log-odds of the unknown baseline survival function and baseline cumulative hazard function have an increasing nonlinear trend.

6.6 Concluding Remarks

In this chapter, we establish an efficient estimation method for the PLSI-PO model with current status data. This model can handle high dimensional nonparametric covariate effects in predicting the survival odds of failure time. This partially linear model is more practical in the analysis of current status data than models with only a single linear term of covariates or just a few nonlinear covariates. We use *B*-splines to approximate the link function of the single-index term and negative logit of the baseline survival function. The splines for the negative logit of the baseline survival as a function of the cumulative hazard function should be restricted to monotone polynomial splines. By maximizing the log-likelihood function over the splines spanned sieve spaces, we estimate the unspecified negative logit of the baseline survival function, single-index link function, orientation parameter and the parametric vector of the regression coefficients. Under the assumption that the true nonparametric link function ψ is a smoothing splines function, we show that the estimators for the parameter vector of the regression coefficients and the orientation parameter vector of the single-index term are semiparametrically efficient by applying theory of counting processes, martingales and empirical processes. Utilizing martingale theory is a new approach in the analysis of current

status data through this model. To show the efficacy of the proposed model and the estimation algorithm, we present a simulation study and apply the model to a real clinical data set.

References

Bennett S (1983) Analysis of survival data by the proportional odds model. Stat Med 2:273–277
Dinse GE, Lagakos SW (1983) Regression analysis of tumour prevalence data. Appl Stat 32:236–248
Gray RJ (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 87:942–951
He X, Zhu ZY, Fung WK (2002) Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika 89:579–590
Hedeker D, Mermelstein RJ (2011) Multilevel analysis of ordinal outcomes related to survival data. In: Hox JJ, Roberts KJ (eds) Handbook of advanced multilevel analysis. Taylor & Francis Group, New York, pp 115–136
Heung M, Wolfgram DF, Kommareddi M, Hu Y, Song PX-K, Ojo AO (2012) Fluid overload at initiation of renal replacement therapy is associated with lack of renal recovery in patients with acute kidney injury. Nephrol Dial Transpl 27:956–961
Huang J (1995) Maximum likelihood estimation for proportional odds regression model with current status data. Lect Notes Monogr Ser 27:129–145
Huang J (1996) Efficient estimation for the proportional hazards model with interval censoring. Ann Stat 24:540–568
Huang J, Rossini AJ (1997) Sieve estimation for the proportional-odds failure-time regression model with interval censoring. J Am Stat Assoc 92:960–967
Huang JZ, Liu L (2006) Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. Biometrics 62:793–802
Lu W, Zhang HH (2010) On estimation of partially linear transformation models. J Am Stat Assoc 105:683–691
Lu X, Song PX-K (2015) Efficient estimation of the partly linear additive hazards model with current status data. Scand J Stat 42:306–328
Ma S, Kosorok MR (2005) Penalized log-likelihood estimation for partly linear transformation models with current status data. Ann Stat 33:2256–2290
Martinussen T, Scheike TH (2002) Efficient estimation in additive hazards regression with current status data. Biometrika 89:649–658
McCullagh P (1980) Regression models for ordinal data. J R Stat Soc Ser B Stat Methodol 42:109–142
McMahan CS, Wang L, Tebbs JM (2013) Regression analysis for current status data using the EM algorithm. Stat Med 32:4452–4466
Pettitt AN (1982) Inference for the linear model using a likelihood based on ranks. J R Stat Soc Ser B Methodol 44:234–243
Rabinowitz D, Betensky RA, Tsiatis AA (2000) Using conditional logistic regression to fit proportional odds models to interval censored data. Biometrics 56:511–518
Rossini AJ, Tsiatis AA (1996) A semiparametric proportional odds regression model for the analysis of current status data. J Am Stat Assoc 91:713–721
Shen X (1998) Propotional odds regression and sieve maximum likelihood estimation. Biometrika 85:165–177
Sun J, Kopciuk KA, Lu X (2008) Polynomial spline estimation of partially linear single-index proportional hazards regression models. Comput Stat Data Anal 53:176–188
Wang L, Dunson DB (2011) Semiparametric Bayes' proportional odds models for current status data with underreporting. Biometrics 67:1111–1118

Chapter 7
Empirical Likelihood Inference Under Density Ratio Models Based on Type I Censored Samples: Hypothesis Testing and Quantile Estimation

Song Cai and Jiahua Chen

Abstract We present a general empirical likelihood inference framework for Type I censored multiple samples. Based on this framework, we develop an effective empirical likelihood ratio test and efficient distribution function and quantile estimation methods for Type I censored samples. In particular, we pool information across multiple samples through a semiparametric density ratio model and propose an empirical likelihood approach to data analysis. This approach achieves high efficiency without making risky model assumptions. The maximum empirical likelihood estimator is found to be asymptotically normal. The corresponding empirical likelihood ratio is shown to have a simple chi-square limiting distribution under the null model of a composite hypothesis about the DRM parameters. The power of the EL ratio test is also derived under a class of local alternative models. Distribution function and quantile estimators based on this framework are developed and are shown to be more efficient than the empirical estimators based on single samples. Our approach also permits consistent estimations of distribution functions and quantiles over a broader range than would otherwise be possible. Simulation studies suggest that the proposed distribution function and quantile estimators are more efficient than the classical empirical estimators, and are robust to outliers and misspecification of density ratio functions. Simulations also show that the proposed EL ratio test has superior power compared to some semiparametric competitors under a wide range of population distribution settings.

S. Cai (✉)
School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada
e-mail: scai@math.carleton.ca

J. Chen
Big Data Research Institute of Yunnan University and Department of Statistics, University of British Columbia, Vancouver, BC, Canada
e-mail: jhchen@stat.ubc.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_7

7.1 Introduction

Type I censored observations are often encountered in reliability engineering and medical studies. In a research project on long-term monitoring of lumber quality, lumber strength data have been collected from mills across Canada over a period of years. The strength-testing machines were set to prefixed tension levels. Those pieces of lumber that are not broken in the test yield Type I right-censored observations. A primary task of the project is to monitor the quality index based on lower quantiles, such as the 5 % quantile, as the years pass. Another important task is to detect the possible change in the overall quality of lumber over time. The statistical nature of the first task is quantile estimation, and that of the second is testing for difference among distribution functions of different populations. We are hence motivated to develop effective quantile estimation and hypothesis testing methods based on multiple Type I censored samples. To achieve high efficiency without restrictive model assumptions, we pool information in multiple samples via a semiparametric density ratio model (DRM) and propose an empirical likelihood (EL) approach to data analysis.

Suppose we have Type I censored samples from $m + 1$ populations with cumulative distribution functions (CDFs) F_k, $k = 0, 1, \ldots, m$. Particularly for our target application, it is reasonable to assume that these CDFs satisfy the relationship

$$dF_k(x) = \exp\left\{\alpha_k + \boldsymbol{\beta}_k^{\mathsf{T}} \boldsymbol{q}(x)\right\} dF_0(x) \tag{7.1}$$

for a pre-specified d-dimensional *basis function* $\boldsymbol{q}(x)$ and model parameter $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^{\mathsf{T}})^{\mathsf{T}}$. The baseline distribution $F_0(x)$ in the above model is left unspecified. Due to symmetry, the role of F_0 is equivalent to that of any of the F_k. When the data are subject to Type I censoring, it is best to choose the population with the largest censoring point as the baseline.

The DRM (7.1) has some advantages. First, it pools the information across different samples through a link between the population distributions without a restrictive parametric model assumption. Highly statistical efficient data analysis methods are therefore possible compared with methods based on models without such links. Second, the DRM is semiparametric and hence very flexible. For example, every exponential family of distributions is a special case of this DRM. In life-data modeling, the moment-parameter family and Laplace-transform-parameter family of distributions (Marshall and Olkin 2007, 7. H & I) both satisfy the DRM. The commonly used logistic regression model under case-control study is also closely related to the DRM (Qin and Zhang 1997). This flexibility makes DRM resistant to misspecification of $\boldsymbol{q}(x)$. Last but not the least, standard nonparametric methods would permit sensible inferences of $F_k(x)$ only for $x \leq c_k$, where c_k is the censoring cutting point of the kth sample. The proposed EL method based on the DRM, however, extends this range to $x \leq \max\{c_k\}$.

Data analysis under the DRM (7.1) based on full observations has attracted a lot of attention. Qin (1998) introduced the EL-based inference under a two-sample

setting; Zhang (2002) subsequently developed a goodness-of-fit test; Fokianos (2004) studied the corresponding density estimation; quantile estimation was investigated by Chen and Liu (2013); in Cai et al. (2016), a dual EL ratio test was developed for testing composite hypotheses about DRM parameters. Research on DRM based on Type I censored data has been scarce. Under a two-sample setting with equal Type I censoring points, Wang et al. (2011) studied the properties of the parameter estimations based on full empirical likelihood. However, without building a link between the EL and its dual, as we do in this paper, both numerical solution and analytical properties of the quantile estimation were found technically challenging.

In this chapter, we first establish a general EL inference framework for Type I censored multiple samples under the DRM. Based on this framework, we then develop an effective EL ratio test and efficient CDF and quantile estimation methods. Instead of a direct employment of the full EL function, we develop a *dual partial empirical likelihood function* (DPEL). The DPEL is equivalent to the full EL function for most inference purposes. However, unlike EL, it is concave and therefore allows simple numerical solutions as well as facilitating deeper analytical investigations of the resulting statistical methods. Using the DPEL, we show that the maximum EL estimators of the $\boldsymbol{\theta}_k$ is asymptotically normal. We also show that the corresponding EL ratio has a chi-square and a non-central chi-square limiting distribution under the null model and local alternative model of a composite hypothesis about the DRM parameters, respectively. We further construct CDF and quantile estimators based on DPEL, and show that they have high efficiency and nice asymptotic properties. The DPEL-based approach is readily extended to address other inference problems, such as EL density estimation and goodness-of-fit test.

The chapter is organized as follows. In Sect. 7.2, we work out the EL function for the DRM based on Type I censored data, introduce the maximum EL estimators, and study their properties. Section 7.3 presents the theory of EL ratio test under the DRM. Sections 7.4 and 7.5 study the EL CDF and quantile estimations. Section 7.6 provides numerical solutions. Simulation results are reported in Sects. 7.7 and 7.8. Section 7.9 illustrates the use of the proposed methods with real lumber quality data.

7.2 Empirical Likelihood Based on Type I Censored Observations

Consider the case where n_k sample units are drawn from the kth population out of which $n_k - \tilde{n}_k$ units are right-censored at c_k, with $\tilde{n}_k$ being the number of uncensored observations. Without loss of generality, we assume $c_0 \geq c_k$ for all k. Denote the uncensored observations by x_{kj} for $j = 1, \ldots, \tilde{n}_k$. Write $dF_k(x) = F_k(x) - F_k(x^-)$. Based on the principle of empirical likelihood of Owen (2001), the

EL is defined to be

$$L_n(\{F_k\}) = \prod_{k=0}^{m} \Big[\prod_{j=1}^{\tilde{n}_k} dF_k(x_{kj}) \{1 - F_k(c_k)\}^{n_k - \tilde{n}_k} \Big].$$

Under the DRM assumption (7.1), the above EL can be further written as

$$L_n(\{F_k\}) = \Big[\prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} dF_0(x_{kj}) \Big] \Big[\prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} \exp\{\boldsymbol{\theta}_k^{\mathsf{T}} \boldsymbol{Q}(x_{kj})\} \Big] \Big[\prod_{k=0}^{m} \{1 - F_k(c_k)\}^{n_k - \tilde{n}_k} \Big], \tag{7.2}$$

with $\boldsymbol{Q}(x) = (1, \boldsymbol{q}(x)^{\mathsf{T}})^{\mathsf{T}}$ and $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^{\mathsf{T}})^{\mathsf{T}}$.

Denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^{\mathsf{T}}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_m^{\mathsf{T}})^{\mathsf{T}}$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathsf{T}}, \boldsymbol{\beta}^{\mathsf{T}})^{\mathsf{T}}$. For convenience, we set $\alpha_0 = 0$ and $\boldsymbol{\beta}_0 = \mathbf{0}$. Further, let $p_{kj} = dF_0(x_{kj})$, $\boldsymbol{p} = \{p_{kj}\}$, $\varsigma_k = F_k(c_k)$, and $\boldsymbol{\varsigma} = \{\varsigma_k\}$. Finally, introduce the new notation

$$\varphi_k(\boldsymbol{\theta}, x, c) = \exp\{\boldsymbol{\theta}_k^{\mathsf{T}} \boldsymbol{Q}(x)\} \mathbb{1}(x \le c).$$

The EL is seen to be a function of $\boldsymbol{\theta}, \boldsymbol{p}$, and $\boldsymbol{\varsigma}$, and we will denote it $L_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma})$.

The maximum EL estimators (MELEs) are now defined to be

$$\begin{aligned} (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\varsigma}}) = \underset{\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma}}{\operatorname{argmax}} \Big\{ & L_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma}) : \sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} p_{kj} \varphi_r(\boldsymbol{\theta}, x_{kj}, c_r) = \varsigma_r, \\ & p_{kj} \ge 0,\ 0 < \varsigma_r \le 1,\ r = 0, \ldots, m \Big\}. \end{aligned} \tag{7.3}$$

The constraints for (7.3) are given by a equality implied by the DRM assumption (7.1) as follows:

$$\varsigma_k = F_k(c_k) = \int \exp\{\boldsymbol{\theta}_k^{\mathsf{T}} \boldsymbol{Q}(x)\} \mathbb{1}(x \le c_k) dF_0(x) = \int \varphi_k(\boldsymbol{\theta}, x, c_k) dF_0(x)$$

for $k = 0, 1, \ldots, m$.

7.2.1 MELE and the Dual PEL Function

The MELE of ς_k is given by $\hat{\varsigma}_k = \tilde{n}_k / n_k$, a useful fact for a simple numerical solution to (7.3). To demonstrate, let us factorize the EL (7.2) as

$$L_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma}) = \Big\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} (p_{kj} / \varsigma_0) \prod_{k=1}^{m} \prod_{j=1}^{\tilde{n}_k} (\varsigma_0 / \varsigma_k) \varphi_k(\boldsymbol{\theta}, x_{kj}, c_k) \Big\} \Big\{ \prod_{k=0}^{m} \varsigma_k^{\tilde{n}_k} \{1 - \varsigma_k\}^{n_k - \tilde{n}_k} \Big\} \tag{7.4}$$

$$= PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma}) \cdot \mathbb{L}_n(\boldsymbol{\varsigma}) \tag{7.5}$$

We call $PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma})$ the *partial empirical likelihood* (PEL) function. Under the constraints specified in (7.3), $\sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma})$ is constant in $\boldsymbol{\varsigma}$ as follows.

Proposition 1 *Let $\tilde{\boldsymbol{\varsigma}}$ and $\check{\boldsymbol{\varsigma}}$ be two values for parameter $\boldsymbol{\varsigma}$ that satisfy the constraints in (7.3). Then we have*

$$\sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \tilde{\boldsymbol{\varsigma}}) = \sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \check{\boldsymbol{\varsigma}}).$$

Proof (Proof) Suppose $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{\theta}}$ form a solution to $\sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \tilde{\boldsymbol{\varsigma}})$, namely,

$$PL_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\varsigma}}) = \sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \tilde{\boldsymbol{\varsigma}}).$$

Let $\check{\boldsymbol{p}}$ and $\check{\boldsymbol{\theta}}$ be defined by

$$\check{p}_{kj} = \tilde{p}_{kj}(\check{\varsigma}_0/\tilde{\varsigma}_0), \quad \check{\alpha}_k = \tilde{\alpha}_k + \log(\tilde{\varsigma}_0/\check{\varsigma}_0) - \log(\tilde{\varsigma}_k/\check{\varsigma}_k).$$

It is easily verified that

$$PL_n(\check{\boldsymbol{\theta}}, \check{\boldsymbol{p}}, \check{\boldsymbol{\varsigma}}) = PL_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\varsigma}}).$$

Hence, we must have

$$\sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \tilde{\boldsymbol{\varsigma}}) \le \sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \check{\boldsymbol{\varsigma}}).$$

Clearly, the reverse inequality is also true, and the proposition follows.

This proposition implies that $\hat{\boldsymbol{\varsigma}} = \text{argmax}\mathbb{L}_n(\boldsymbol{\varsigma})$ and therefore that $\hat{\varsigma}_k = \tilde{n}_k/n_k$. It further implies that $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}) = \text{argmax}PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \hat{\boldsymbol{\varsigma}})$ under the same set of constraints. Because of this, we can compute $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}})$ with standard EL tools. We first get the profile function $\tilde{\ell}_n(\boldsymbol{\theta}) = \sup_{\boldsymbol{p}} \log\{PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \hat{\boldsymbol{\varsigma}})\}$ and then compute for $\hat{\boldsymbol{\theta}} = \text{argmax } \tilde{\ell}_n(\boldsymbol{\theta})$.

More specifically, let $\tilde{n} = \sum_{k=0}^{m} \tilde{n}_k$ be the total number of uncensored observations. By the method of Lagrange multipliers, for each given $\boldsymbol{\theta}$, the solution of $\sup PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \hat{\boldsymbol{\varsigma}})$ in $\boldsymbol{p}$ is given by

$$p_{kj} = \tilde{n}^{-1}\Big\{1/\hat{\varsigma}_0 + \sum_{r=1}^{m} \lambda_r\big[\varphi_r(\boldsymbol{\theta}, x_{kj}, c_r) - \hat{\varsigma}_r/\hat{\varsigma}_0\big]\Big\}^{-1}, \tag{7.6}$$

where the Lagrange multipliers $\{\lambda_k\}_{k=1}^m$ solve $\sum_{k=0}^m \sum_{j=1}^{\tilde{n}_k} p_{kj}\varphi_r(\boldsymbol{\theta}, x_{kj}, c_r) = \hat{\varsigma}_r$ for $r = 0, 1, \ldots, m$. The resulting profile log-PEL is then given by

$$\tilde{\ell}_n(\boldsymbol{\theta}) = -\sum_{k=0}^{m}\sum_{j=1}^{\tilde{n}_k} \log \tilde{n}\Big\{1/\hat{\varsigma}_0 + \sum_{r=1}^{m} \lambda_r\big[\varphi_r(\boldsymbol{\theta}, x_{kj}, c_r) - \hat{\varsigma}_r/\hat{\varsigma}_0\big]\Big\} + \sum_{k=1}^{m}\sum_{j=1}^{\tilde{n}_k} \boldsymbol{\theta}_k^{\mathsf{T}} \boldsymbol{Q}(x_{kj}). \tag{7.7}$$

At its maximum when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, the profile log-PEL satisfies $\partial\tilde{\ell}_n(\boldsymbol{\theta})/\partial\alpha_k = 0$ for $k = 1, \ldots, m$. Some simple algebra shows that consequently, when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we have the corresponding Lagrange multipliers are $\hat{\lambda}_k = n_k/\tilde{n}$.

Put $\hat{\rho}_k = n_k/n$ and recall that $\hat{\varsigma}_r = \tilde{n}_r/n_r$. We find that

$$\tilde{\ell}_n(\hat{\boldsymbol{\theta}}) = -\sum_{k=0}^{m}\sum_{j=1}^{\tilde{n}_k} \log\Big\{\sum_{r=0}^{m} \hat{\rho}_r\varphi_r(\hat{\boldsymbol{\theta}}, x_{kj}, c_r)\Big\} + \sum_{k=1}^{m}\sum_{j=1}^{\tilde{n}_k} \hat{\boldsymbol{\theta}}_r^{\mathsf{T}} \boldsymbol{Q}(x_{kj}).$$

For this reason, we define a *dual PEL* (DPEL) function

$$\ell_n(\boldsymbol{\theta}) = -\sum_{k=0}^{m}\sum_{j=1}^{\tilde{n}_k} \log\Big\{\sum_{r=0}^{m} \hat{\rho}_r\varphi_r(\boldsymbol{\theta}, x_{kj}, c_r)\Big\} + \sum_{k=1}^{m}\sum_{j=1}^{\tilde{n}_k} \boldsymbol{\theta}_r^{\mathsf{T}} \boldsymbol{Q}(x_{kj}). \tag{7.8}$$

Clearly, we have $\hat{\boldsymbol{\theta}} = \operatorname{argmax} l_n(\boldsymbol{\theta})$, and the DPEL is a concave function of $\boldsymbol{\theta}$. The relationship also implies that

$$\hat{p}_{kj} = n^{-1}\Big\{\sum_{r=0}^{m} \hat{\rho}_r\varphi_r(\hat{\boldsymbol{\theta}}, x_{kj}, c_r)\Big\}^{-1}. \tag{7.9}$$

The DPEL is analytically simple, facilitating deeper theoretical investigations and simplifying the numerical problem associated with the data analysis.

7.2.2 Asymptotic Properties of the MELE $\hat{\boldsymbol{\theta}}$

Let $\boldsymbol{\theta}^*$ be the true value of the parameter $\boldsymbol{\theta}$. Suppose that, for some constants $\rho_k \in (0, 1)$, $\hat{\rho}_k = n_k/n \to \rho_k$ as $n \to \infty$, for $k = 0, \ldots$. Define the *partial empirical information matrix* $U_n = -n^{-1}\partial^2\ell_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\mathsf{T}}$. By the strong law of large numbers, U_n converges almost surely to a matrix U because it is the average of several independent and identically distributed (i.i.d) samples. The limit U serves as a *partial information matrix*.

Define

$$\boldsymbol{h}(\boldsymbol{\theta}, x) = (\rho_1\varphi_1(\boldsymbol{\theta}, x, c_1), \ldots, \rho_m\varphi_m(\boldsymbol{\theta}, x, c_m))^{\mathsf{T}},$$

$$s(\boldsymbol{\theta}, x) = \sum_{k=0}^{m} \rho_k\varphi_k(\boldsymbol{\theta}, x, c_k),$$

$$H(\boldsymbol{\theta}, x) = \mathrm{diag}\{\boldsymbol{h}(\boldsymbol{\theta}, x)\} - \boldsymbol{h}(\boldsymbol{\theta}, x)\boldsymbol{h}^{\mathsf{T}}(\boldsymbol{\theta}, x)/s(\boldsymbol{\theta}, x).$$

We partition the entries of U in agreement with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and denote them by $U_{\boldsymbol{\alpha\alpha}}$, $U_{\boldsymbol{\alpha\beta}}$, $U_{\boldsymbol{\beta\alpha}}$, and $U_{\boldsymbol{\beta\beta}}$. The blockwise algebraic expressions of U can be written as

$$U_{\boldsymbol{\alpha\alpha}} = \mathrm{E}_0\left\{H(\boldsymbol{\theta}^*, X)\right\},$$

$$U_{\boldsymbol{\beta\beta}} = \mathrm{E}_0\left\{H(\boldsymbol{\theta}^*, X) \otimes \left(\boldsymbol{q}(X)\boldsymbol{q}^{\mathsf{T}}(X)\right)\right\},$$

$$U_{\boldsymbol{\alpha\beta}} = U_{\boldsymbol{\beta\alpha}}^{\mathsf{T}} = \mathrm{E}_0\left\{H(\boldsymbol{\theta}^*, X) \otimes \boldsymbol{q}^{\mathsf{T}}(X)\right\},$$

where $\mathrm{E}_0(\cdot)$ is the expectation operator with respect to $F_0(x)$ and $\otimes$ is the Kronecker product operator. The above blockwise expressions of U reveal that U is positive definite when $\int \boldsymbol{Q}(x)\boldsymbol{Q}^{\mathsf{T}}(x)dF_0(x) > 0$.

We found that the MELE $\hat{\boldsymbol{\theta}}$ is asymptotically normal as summarized as follows.

Theorem 1 *Suppose we have $m + 1$ Type I censored random samples with censoring cutting points c_k, $k = 0, \ldots, m$, from populations with distributions satisfying the DRM assumption (7.1) with a true parameter value $\boldsymbol{\theta}^*$ such that $\int \exp\{\boldsymbol{\beta}_k^{\mathsf{T}}\boldsymbol{q}(x)\}dF_0(x) < \infty$ for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^*$. Also, suppose $\int \boldsymbol{Q}(x)\boldsymbol{Q}^{\mathsf{T}}(x)dF_0(x) > 0$ and $\hat{\rho}_k = n_k/n \to \rho_k$ as $n \to \infty$ for some constants $\rho_k \in (0, 1)$.*

Then, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{(d)} N\big(\mathbf{0},\ U^{-1} - W\big),$$

where

$$W = \begin{pmatrix} T_{m\times m} & 0 \\ 0 & 0_{md\times md} \end{pmatrix} \quad \textit{with} \quad T = \begin{pmatrix} \rho_0^{-1} + \rho_1^{-1} & \rho_0^{-1} & \cdots & \rho_0^{-1} \\ \rho_0^{-1} & \rho_0^{-1} + \rho_2^{-1} & \cdots & \rho_0^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_0^{-1} & \rho_0^{-1} & \cdots & \rho_0^{-1} + \rho_m^{-1} \end{pmatrix}.$$

The asymptotic normality of $\hat{\boldsymbol{\theta}}$ forms the basis for developing the EL–DRM hypothesis testing and estimation methods in the sequel.

7.3 EL Ratio Test for Composite Hypotheses About $\boldsymbol{\beta}$

An appealing property of the classical EL inference for single samples is that the EL ratio has a simple chi-square limiting distribution (Owen 2001). For uncensored multiple samples satisfying the DRM assumption (7.1), the dual EL (DEL) ratio is also shown to have an asymptotic chi-square distribution by Cai et al. (2016) under the null hypothesis of a large class of composite hypothesis testing problems. Such nice property of the EL ratio also carries to the case of Type I censored multiple samples as we will show in this section.

As noted in the Introduction, a primary interest in our lumber project is to check whether the overall quality of the lumber change over time. This amounts to test the difference among the underlying distributions $\{F_k\}_{k=0}^m$ of random lumber samples collected from different years, i.e. testing $H_0 : F_0 = \ldots = F_m$ against $H_a : F_i \neq F_j$ for some $i \neq j$. When the $\{F_k\}$ satisfy the DRM assumption (7.1), this hypothesis testing problem is equivalent to a hypothesis testing problem about the DRM parameter $\boldsymbol{\beta}$: $H_0 : \boldsymbol{\beta} = 0$ against $H_a : \boldsymbol{\beta} \neq 0$. Note that, the parameter $\boldsymbol{\alpha}$ is not included because it is just a normalizing constant, and $\boldsymbol{\beta}_k = 0$ implies $\alpha_k = 0$.

Here we consider a more general composite hypothesis testing problem about $\boldsymbol{\beta}$

$$H_0 : \boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{g}(\boldsymbol{\beta}) \neq \mathbf{0} \tag{7.10}$$

for some smooth function $\boldsymbol{g} : \mathbb{R}^{md} \to \mathbb{R}^q$, with $q \leq md$, the length of $\boldsymbol{\beta}$. We will assume that $\boldsymbol{g}$ is thrice differentiable with a full rank Jacobian matrix $\partial \boldsymbol{g}/\partial \boldsymbol{\beta}$. The parameters $\{\alpha_k\}$ are usually not a part of the hypothesis, because their values are fully determined by the $\{\boldsymbol{\beta}_k\}$ and F_0 under the DRM assumption.

We propose an EL ratio test for the above hypothesis. Let $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\zeta}})$ be the MELE based on Type I censored samples under the null constraint of $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$. Define the EL ratio statistic as

$$R_n = 2\{\log L_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\zeta}}) - \log L_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\zeta}})\}.$$

Our following lemma shows that R_n is equal to the DPEL ratio, a quantity that enjoys a much simpler analytical expression.

Lemma 1 *The EL ratio statistic R_n equals the DEPL ratio statistic, i.e.*

$$R_n = 2\{\ell_n(\hat{\boldsymbol{\theta}}) - \ell_n(\tilde{\boldsymbol{\theta}})\},$$

where $\ell_n(\theta)$ is the DPEL function (7.8)*.*

Note that, except for the additional indicator terms, the expression of the DPEL is identical to that of the DEL function defined in Cai et al. (2016) for uncensored samples under the DRM. Hence the techniques for showing the asymptotic properties of the DEL ratio in Cai et al. (2016) can be readily adapted here to prove our

next two theorems (Theorems 2 and 3) about the asymptotic properties of the EL ratio R_n based on Type I censored samples.

Let χ_q^2 denote a chi-square distribution with q degrees of freedom, and $\chi_q^2(\delta^2)$ denote a non-central chi-square distribution with q degrees of freedom and non-central parameter δ^2. Partition the Jacobian matrix of $\boldsymbol{g}(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^*$, $\bigtriangledown = \partial \boldsymbol{g}(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}$, into $(\bigtriangledown_1, \bigtriangledown_2)$, with q and $md - q$ columns respectively. Without loss of generality, we assume that $\bigtriangledown_1$ has a full rank. Let I_k be an identity matrix of size $k \times k$ and $J^{\mathsf{T}} = (-(\bigtriangledown_1^{-1}\bigtriangledown_2)^{\mathsf{T}}, I_{md-q})$.

Theorem 2 *Adopt the conditions postulated in Theorem 1. Then, under the null hypothesis, $H_0 : \boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$, of (7.10), we have*

$$R_n \xrightarrow{(d)} \chi_q^2$$

as $n \to \infty$.

Theorem 2 is most useful for constructing a EL ratio test for composite hypothesis testing problem (7.10). In addition, it can be used to construct a confidence region for the true DRM parameter $\boldsymbol{\beta}^*$. Take the null hypothesis to be $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}^* = 0$ for any given $\boldsymbol{\beta}^*$, then $R_n = l_n(\hat{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}^*) \to \chi_{md}^2$. We can use this result to construct a chi-square confidence region for $\boldsymbol{\beta}^*$. The advantage of a chi-square confidence region over a normal region is extensively discussed in Owen (2001), so we do not elaborate further.

We shall focus on the hypothesis testing problem since that is our primary goal in application. The next theorem gives the limiting distribution of the EL ratio under a class of local alternatives. It can be used to approximate the power of the EL ratio test and to calculate the required sample size for achieving a certain power.

Theorem 3 *Adopt the conditions postulated in Theorem 1. Let $\{\boldsymbol{\beta}_k^*\}_{k=1}^m$ be a set of DRM parameter values that satisfy the null hypothesis of (7.10).*

Then, under the local alternative model:

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^* + n_k^{-1/2}\boldsymbol{c}_k, \quad k = 1, \ldots, m, \tag{7.11}$$

where $\{\boldsymbol{c}_k\}$ are some constants, we have

$$R_n \xrightarrow{(d)} \chi_q^2(\delta^2)$$

as $n \to \infty$. The expression of the non-central parameter δ^2 is given by

$$\delta^2 = \begin{cases} \boldsymbol{\eta}^{\mathsf{T}}\{\tilde{\Lambda} - \tilde{\Lambda}J(J^{\mathsf{T}}\tilde{\Lambda}J)^{-1}J^{\mathsf{T}}\tilde{\Lambda}\}\boldsymbol{\eta} & \text{if } q < md \\ \boldsymbol{\eta}^{\mathsf{T}}\tilde{\Lambda}\boldsymbol{\eta} & \text{if } q = md \end{cases}$$

where $\tilde{\Lambda} = U_{\boldsymbol{\beta}\boldsymbol{\beta}} - U_{\boldsymbol{\beta}\boldsymbol{\alpha}}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}U_{\boldsymbol{\alpha}\boldsymbol{\beta}}$. and $\boldsymbol{\eta} = \left(\rho_1^{-1/2}\boldsymbol{c}_1^{\mathsf{T}}, \rho_2^{-1/2}\boldsymbol{c}_2^{\mathsf{T}}, \ldots, \rho_m^{-1/2}\boldsymbol{c}_m^{\mathsf{T}}\right)^{\mathsf{T}}$. Moreover, $\delta^2 > 0$ unless $\boldsymbol{\eta}$ is in the column space of J.

In many situations, a hypothesis of interest may well focus on characteristics of just a subset of the populations $\{F_k\}_{k=0}^m$. If so, should our test be based on all the samples or only the samples of interest? An answer is found in the improved local power of the EL ratio test based on all samples over the test based on the subset of the samples. Such an answer is the same as that for the DEL ratio test under the DRM based on uncensored multiple samples. A rigorous treatment of this argument is given by Theorem 3 of for uncensored case. Again, because of the similarity in expressions of the DPEL and DEL, that theorem also holds for the proposed EL ratio test for Type I censored samples. The details can be found therein.

7.4 Estimation of F_k

We now turn to the estimation of CDFs $\{F_k\}$. The estimation of population quantiles is studied in the next section.

Due to the DRM assumption, we naturally estimate F_r at any $z \le c_0$ by

$$\hat{F}_r(z) = \sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \hat{p}_{kj} \exp\left\{\hat{\boldsymbol{\theta}}_r^{\mathsf{T}} \boldsymbol{Q}(x_{kj})\right\} \mathbb{1}(x_{kj} \le z) = \sum_{k=0}^{m} \sum_{j=1}^{n_k} \hat{p}_{kj} \varphi_k(\boldsymbol{\theta}, x_{kj}, z). \tag{7.12}$$

A few notational conventions have been and will continue to be used: $\sum_{k,j}$ will be regarded as summation over $k = 0, \ldots, m$ and $j = 1, \ldots, n_k$. When $\tilde{n}_k < j \le n_k$, the value of x_{kj} is censored and we define

$$\varphi_k(\boldsymbol{\theta}, x_{kj}, z) = \exp\left\{\hat{\boldsymbol{\theta}}_r \boldsymbol{Q}(x_{kj})\right\} \mathbb{1}(x_{kj} \le z) = 0 \text{ or } 1. \tag{7.13}$$

Whether $\varphi_k(\boldsymbol{\theta}, x_{kj}, z)$ takes value 0 or 1 when $\tilde{n}_k < j \le n_k$ depends on whether it serves as an additive or a product term, and similarly for other quantities involving x_{kj}. With this convention, we may regard $\sum_{k,j}$ as being a sum over $m + 1$ i.i.d samples.

Even though observations from the population F_r are censored at $c_r \le c_0$, the connection between F_r and F_0 through DRM makes it possible to consistently estimate $F_r(z)$ for $z \in (c_r, c_0]$ when $c_r < c_0$.

Theorem 4 *Assume the conditions of Theorem 1. For any pair of integers* $0 \le r_1, r_2 \le m$ *and pair of real values* $z_1, z_2 \le c_0$*, we have*

$$\sqrt{n}\left\{\hat{F}_{r_1}(z_1) - F_{r_1}(z_1), \hat{F}_{r_2}(z_2) - F_{r_2}(z_2)\right\}^{\mathsf{T}} \xrightarrow{(d)} N(\mathbf{0}, \Omega_{EL}).$$

The expression for the (i, j) *entry of* Ω_{EL} *is*

$$\omega_{i,j} = \mathrm{E}_0\{\varphi_{r_i}(\boldsymbol{\theta}^*, X, z_i)\varphi_{r_j}(\boldsymbol{\theta}^*, X, z_j)/s(\boldsymbol{\theta}^*, X)\} + \Psi_{r_i}^{\mathsf{T}}(z_i)U^{-1}\Psi_{r_j}(z_j)$$
$$- \rho_{r_i}^{-1}\big[\mathrm{E}_0\{\varphi_{r_i}(\boldsymbol{\theta}^*, X, z_i)\}\,\mathrm{E}_0\{\varphi_{r_j}(\boldsymbol{\theta}^*, X, z_j)\}\big]\mathbb{1}(i = j),$$

where, for $k = 0, \ldots, m$, $\Psi_k(z) = \big(\Psi_{k,1}^{\mathsf{T}}(z),\ \Psi_{k,2}^{\mathsf{T}}(z)\big)^{\mathsf{T}}$, *is a vector with*

$$\Psi_{k,1}(z) = \mathrm{E}_0\big[\{\boldsymbol{e}_k\mathbb{1}(k \neq 0) - \boldsymbol{h}(\boldsymbol{\theta}^*, X)/s(\boldsymbol{\theta}^*, X)\}\varphi_k(\boldsymbol{\theta}^*, X, z)\big],$$
$$\Psi_{k,2}(z) = \mathrm{E}_0\big[\{\boldsymbol{e}_k\mathbb{1}(k \neq 0) - \boldsymbol{h}(\boldsymbol{\theta}^*, X)/s(\boldsymbol{\theta}^*, X)\}\varphi_k(\boldsymbol{\theta}^*, X, z) \otimes \boldsymbol{q}(X)\big],$$

and $\boldsymbol{e}_k$ *is a vector of length m with the kth entry being* 1 *and the others* 0.

The result of this theorem extends to multiple $r_1, \ldots, r_J$ with their corresponding $z_1, \ldots, z_J \leq c_0$. For ease of presentation, we have given the result only for $J = 2$ above.

We have assumed (without of loss of generality) that $c_r \leq c_0$. When $c_r < c_0$, there is no direct information to estimate $F_r(z)$ for $z \in (c_r, c_0]$. The DRM assumption, however, allows us to borrow information from samples from F_0 and other F_k to sensibly estimate $F_r(z)$ for z in this range. This is an interesting result. When $z \leq c_r$, we may estimate $F_r(z)$ via its empirical distribution based only on the sample from F_r. Theorem 4 can be used to show that the EL–DRM-based estimator has lower asymptotic variance, i.e., $\Omega_{EL} \leq \Omega_{EM}$, where Ω_{EM} is the covariance matrix of the empirical distribution.

7.5 Quantile Estimation

With a well-behaved EL–DRM CDF estimator in hand, we propose to estimate the τth, $\tau \in (0, 1)$, quantile ξ_r of the population $F_r(x)$ by

$$\hat{\xi}_r = \inf\{x : \hat{F}_r(x) \geq \tau\}. \tag{7.14}$$

Our following theorem characterizes an important asymptotic property of $\hat{\xi}_r$.

Theorem 5 *Assume the conditions of Theorem 1. Assume also that the density function,* $f_r(x)$, *of* $F_r(x)$, *is positive and differentiable at* $x = \xi_r$ *for some* $\tau \in (0, F_r(c_0))$.

Then the EL quantile estimator (7.14) *admits the following representation:*

$$\hat{\xi}_r = \xi_r - \{\hat{F}_r(\xi_r) - \tau\}/f_r(\xi_r) + O_p(n^{-3/4}\{\log n\}^{1/2}).$$

A result of this nature was first obtained by Bahadur (1966) for the sample quantiles, hence the name *Bahadur representation*. It is proven to hold much more

broadly, including for EL-based quartile estimators such as those in Chen and Chen (2000) and Chen and Liu (2013). Our result is of particular interest as the representation goes beyond the range restricted due to Type I censorship: $\xi_r < c_r$. Bahadur representation sheds light on the large sample behavior of the quantile process, and it is most useful for studying the multivariate asymptotic normality of $\hat{\xi}_r$, as follows.

Theorem 6 *Let ξ_j be the τ_jth quantile of F_{r_j} with $\tau_j \in (0, F_{r_j}(c_0))$ for $j = 1, 2$. Under the conditions of Theorem 5, we have*

$$\sqrt{n}\big(\hat{\xi}_1 - \xi_1, \hat{\xi}_2 - \xi_2\big)^{\mathsf{T}} \xrightarrow{(d)} N(\mathbf{0}, A\Omega_{EL}A)$$

with $A = \text{diag}\{f_{r_1}(\xi_1), f_{r_2}(\xi_2)\}^{-1}$.

Clearly, improved precision in terms of Ω_{EL} over Ω_{EM} leads to improved efficiency of the proposed $\hat{\xi}_r$ over the sample quantile.

7.6 Other Inferences on Quantiles

Theorem 6 has laid a solid basis for constructing approximate Wald-type confidence intervals and hypothesis tests for quantiles. For this purpose, we need a consistent estimator of Ω_{EL} and $f_r(\xi_r)$, the density function at ξ_r.

The analytical expression for Ω_{EL} is a function of $\boldsymbol{\theta}^*$ and has the general form $E_0\{g(X, \boldsymbol{\theta}^*)\mathbb{1}(X \le c_0)\}$. To estimate Ω_{EL}, it is most convenient to use the method of moments with EL weights $\hat{p}_{kj}$:

$$\hat{E}_0\{g(X, \hat{\boldsymbol{\theta}})\mathbb{1}(X \le c_0)\} = \sum_{k,j} \hat{p}_{kj} g(x_{kj}, \hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is the MELE of $\boldsymbol{\theta}$ and $\hat{p}_{kj}$ is as given in (7.9).

The value of $f_r(\xi_r)$ is most effectively estimated by the kernel method. Let $K(t)$ be a positive-valued function such that $\int K(t)dt = 1$ and $\int tK(t)dt = 0$. We estimate $f_r(z)$ at $z \le c_0$ by

$$\hat{f}_r(z) = h_n^{-1} \sum_{k=0}^{m} \sum_{j=1}^{n_k} \hat{p}_{kj} \varphi_k(\boldsymbol{\theta}, x_{kj}, c_k) K\big\{(z - x_{kj})/h_n\big\}.$$

When the bandwidth $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, this kernel estimator is easily shown to be consistent. The optimal bandwidth is of order $n^{-1/5}$ in terms of the asymptotic mean integrated squared error (Silverman 1986) at z values that are not close to the boundary. In our simulation, we choose the density function of the standard normal distribution as the kernel and $n^{-1/5}$ as the bandwidth.

The covariance matrix, Σ_{EL}, of the proposed quantile estimator is then estimated by substituting the estimated Ω_{EL} and the density estimates $\hat{f}_{r_i}(\hat{\xi}_i)$ into the expression of Σ_{EL}. The procedures have been implemented in our R software package "drmdel", which is available on the Comprehensive R Archive Network (CRAN).

7.7 Simulation Studies I: CDF and Quantile Estimation

We use simulation studies to demonstrate the advantages of combining information through the use of the DRM and the EL methodology. This section presents the simulation results for the proposed CDF and quantile estimation methods are presented in this section, and those for EL ratio test are given in the next section. Our estimation approach is particularly effective at efficient estimation of lower quantiles and is resistant to mild model misspecification and the influence of large outliers. As individual samples are by nature sparse at observations at lower quantiles, it is particularly important to pool information from several samples. At the same time, the presence of Type I censorship matters little for estimating lower quantiles.

In all simulations, we put the number of samples to be $m + 1 = 4$, and the sample sizes n_k to be $(110, 90, 100, 120)$. The number of simulation repetitions is set to 10,000.

7.7.1 Populations Satisfying the DRM Assumption

The DRM encompasses a large range of statistical models as subsets. In this simulation, we consider populations from a flexible parametric family, the generalized gamma distribution (Stacy 1962), denoted $GG(a, b, p)$. It has density function

$$f(x) = \frac{a}{b^{ap}\gamma(p)} x^{ap-1} \exp\{-(x/b)^a\},\ x > 0.$$

When $p = 1$, the generalized gamma distribution becomes the Weibull distribution; when $a = 1$, it becomes the gamma distribution. Generalized gamma distributions with known shape parameter a satisfy the DRM assumption with basis function $\boldsymbol{q}(x) = (\log x, x^a)^\intercal$. In our simulations, we fix a at 2 and choose parameter values such that the shapes and the first two moments of populations closely resemble those of our lumber quality samples in real applications. We generate samples from four such populations and make these samples right-censored around their 75 % population quantiles. We also conduct simulations based on samples from Weibull, gamma, and normal populations, respectively.

We first study CDF estimation for $F_r(z)$ for all $r = 0, \ldots, m$ with z being the 5 %, 10 %, 30 %, 40 %, and 50 % quantiles of the baseline population F_0. Based on the Type I censored data, the empirical distribution $\breve{F}_r(z)$ is well defined for z values smaller than the censoring point but not for larger values of z. We purposely selected

Table 7.1 Simulation results for CDF estimation based on generalized gamma samples

	z	$F_r(z)$	$\mathrm{V}(\hat{F}_r)(\times 10^{-3})$	$\frac{\sigma^2(\hat{F}_r)}{n\mathrm{V}(\hat{F}_r)}$	$\frac{\hat{\sigma}^2(\hat{F}_r)}{n\mathrm{V}(\hat{F}_r)}$	$\frac{\mathrm{V}(\check{F}_r)}{\mathrm{V}(\hat{F}_r)}$	$\mathrm{B}(\hat{F}_r)(\%)$	$\mathrm{B}(\check{F}_r)(\%)$
$GG(2, 4.3, 2.8)$	3.64	0.05	0.34	0.99	1.00	1.24	0.92	0.45
$c_0 = 8.2$	4.26	0.10	0.67	0.96	0.98	1.21	1.36	0.40
	5.68	0.30	1.62	0.98	0.98	1.20	0.08	0.03
	6.23	0.40	1.92	0.99	0.98	1.14	−0.07	0.03
	6.76	0.50	2.10	0.99	0.98	1.09	−0.05	0.06
$GG(2, 2.6, 5)$	3.64	0.05	0.39	1.02	1.02	1.32	−0.93	0.07
$c_1 = 6.5$	4.26	0.13	1.05	0.96	0.97	1.26	−0.34	−0.16
	5.68	0.52	2.43	1.00	1.00	1.14	0.10	0.05
	6.23	0.68	2.07	1.02	1.01	1.15	−0.01	0.01
	6.76	0.80	3.27	0.95	1.02	NA	0.34	NA
$GG(2, 3.8, 3.5)$	3.64	0.03	0.23	0.96	0.98	1.38	0.78	0.20
$c_2 = 8.1$	4.26	0.07	0.54	0.94	0.95	1.27	0.21	0.22
	5.68	0.28	1.62	1.00	0.99	1.21	−0.20	0.06
	6.23	0.39	2.00	1.01	1.01	1.18	−0.01	0.11
	6.76	0.50	2.23	1.02	1.01	1.10	0.15	0.05
$GG(2, 2.5, 6)$	3.64	0.02	0.12	1.02	1.05	1.54	1.24	1.21
$c_3 = 6.8$	4.26	0.07	0.45	0.98	0.98	1.31	−0.52	0.39
	5.68	0.41	1.76	1.01	1.00	1.14	0.02	−0.05
	6.23	0.59	1.77	1.01	1.01	1.13	0.10	−0.06
	6.76	0.74	1.55	1.01	1.01	1.03	−0.03	−0.05

F_r: true CDF value; $\hat{F}_r$ and $\check{F}_r$: EL–DRM and EM estimates; V: simulated variance; σ^2: theoretical variance; $\hat{\sigma}^2$: average of variance estimates; B: relative bias in percentage; NA: $\check{F}_r(z)$ not defined

populations such that at some z values, $\check{F}_r(z)$ is not well-defined but the EL–DRM estimator $\hat{F}_r(z)$ is. We report various summary statistics from the simulation for both the empirical distribution and our proposed EL–DRM estimators.

The simulation results for data from the generalized gamma populations are reported in Table 7.1. The parameter values of the populations are listed in the leftmost column. Note that the basis function of the DRM for these populations is $\boldsymbol{q}(x) = (\log x, x^2)^{\mathsf{T}}$. The censoring points for the four samples are (8.2, 6.5, 8.1, 6.8). The z values at which the CDFs are estimated are listed in the "z" column, and the corresponding true CDF values are listed in the "$F_r(z)$" column.

The fourth column is the variance of the CDF estimator observed in the simulation (over 10,000 repetitions). The fifth (respectively, sixth) column is the ratio of the theoretical asymptotic variance (respectively, average estimated variance) to the variance observed in the simulation. All the values in these columns are close to 1, indicating that the variance estimator constructed in Sect. 7.4 and its numerical computation method given in Sect. 7.6 work very well. The seventh column is the ratio between two simulated variances, one based on the empirical distribution and the other on our proposed EL–DRM-based estimator. For small z values, our

proposed estimator gains 20 %–55 % precision in terms of simulated variances. The gain in efficiency is smaller at large z values but still noticeable.

The last two columns are relative biases. The relative biases of both the EL–DRM estimators and the empirical (EM) estimators are generally below 1.5 %, which seems rather satisfactory. Overall, compared to the variance, the bias is a very small portion of the mean squared error (MSE), so we omit the comparison of the two estimators in MSE.

When the empirical distribution $\breve{F}_r(z)$ is not defined for a particular z due to Type I censorship, we put "NA"s in the corresponding cells in the table. As indicated, $z = 6.76$ is beyond the censoring point of F_1, so $F_1(6.76)$ cannot be estimated by the empirical distribution.

We next examine the results for quantile estimation. For each population, we estimate quantiles at levels $\tau = 0.05, 0.1, 0.3, 0.5$ and 0.6. The results for data from the generalized gamma populations are given in Table 7.2. The first column lists the parameter values of the populations, the second gives the levels at which the quantiles are estimated, and the third provides the corresponding true quantile values. The fourth column is the simulated variance. The fifth and sixth columns reflect the quality of the theoretical asymptotic variance and the variance estimation

Table 7.2 Simulation results for quantile estimation based on generalized gamma samples

	τ	ξ_r	$\mathrm{V}(\hat{\xi}_r)(\times 10^{-2})$	$\frac{\sigma^2(\hat{\xi}_r)}{n\mathrm{V}(\hat{\xi}_r)}$	$\frac{\hat{\sigma}^2(\hat{\xi}_r)}{n\mathrm{V}(\hat{\xi}_r)}$	$\frac{\mathrm{V}(\breve{\xi}_r)}{\mathrm{V}(\hat{\xi}_r)}$	$\mathrm{B}(\hat{\xi}_r)(\%)$	$\mathrm{B}(\breve{\xi}_r)(\%)$
$GG(2, 4.3, 2.8)$	0.05	3.64	8.44	1.02	0.98	1.27	0.30	0.44
$c_0 = 8.2$	0.10	4.26	6.75	0.96	0.99	1.23	−0.10	−0.93
	0.30	5.68	5.38	0.97	1.02	1.17	0.07	−0.40
	0.50	6.76	5.89	1.00	1.01	1.09	0.09	−0.33
	0.60	7.31	6.44	0.99	1.04	1.06	0.03	−0.36
$GG(2, 2.6, 5)$	0.05	3.65	4.41	1.02	1.01	1.29	0.52	0.36
$c_3 = 6.5$	0.10	4.06	3.30	0.99	1.00	1.27	0.25	−0.69
	0.30	4.96	2.43	0.98	1.03	1.22	0.03	−0.38
	0.50	5.62	2.53	1.01	1.04	1.15	−0.01	−0.32
	0.60	5.95	2.72	0.99	1.06	1.12	−0.02	−0.34
$GG(2, 3.8, 3.5)$	0.05	3.96	7.94	0.95	0.94	1.31	0.37	−1.40
$c_2 = 8.1$	0.10	4.52	6.07	0.92	0.96	1.28	0.14	−0.86
	0.30	5.81	4.47	1.01	1.03	1.20	0.12	−0.40
	0.50	6.77	4.80	1.05	1.06	1.13	−0.00	−0.35
	0.60	7.25	5.34	1.02	1.06	1.08	−0.03	−0.32
$GG(2, 2.5, 6)$	0.05	4.04	3.28	0.98	0.96	1.40	0.27	−0.97
$c_1 = 6.8$	0.10	4.44	2.38	0.98	0.99	1.30	0.16	−0.54
	0.30	5.31	1.72	1.00	1.04	1.21	0.07	−0.23
	0.50	5.95	1.76	1.02	1.05	1.12	−0.02	−0.19
	0.60	6.27	1.91	0.99	1.05	1.10	−0.02	−0.19

ξ_r: true quantile; $\hat{\xi}_r$ and $\breve{\xi}_r$: EL–DRM and EM quantile estimates; V: simulated variance; σ^2: theoretical variance; $\hat{\sigma}^2$: average of variance estimates; B: relative bias in percentage

relative to the simulated variance. Again, we see that all numbers in these columns are close to 1, showing that the asymptotic variance and its estimates are good approximations to the actual variance of the estimator.

The seventh column, which gives the ratio of two simulated variances-one based on the empirical distribution and the other on the EL–DRM quantile estimator-shows that the EL–DRM quantile estimator is uniformly more efficient than the EM quantile estimator (i.e., the sample quantile). In applications, the gain is particularly substantial and helpful at lower quantiles (5 % and 10 %), being between 23 % and 40 %. The biases of the EL–DRM quantile estimator are generally smaller than those of the EM estimator in absolute value, except in one case. In general, the biases of both estimators are satisfactorily low.

The asymptotic normality shown in Theorem 6 can be used to construct confidence intervals for quantiles and quantile differences based on EL–DRM quantile estimators. The same is true for the EM quantile estimators. The simulation results for confidence intervals of quantiles and quantile differences are shown in Table 7.3.

In all cases, the EL–DRM intervals outperform the EM intervals in terms of both coverage probabilities and lengths. For single lower quantiles at $\tau = 0.05$ and 0.1, the coverage probabilities of EL–DRM intervals are between 91 % and 93.5 %. In comparison, on average the EM intervals have around 2 % less coverage. Both methods have closer to nominal coverage probabilities for the higher quantiles. The EL–DRM intervals have shorter average lengths and therefore are more efficient.

The simulation results for confidence intervals of quantile differences between F_0 and F_1 are also shown in the last two rows in Table 7.3. The coverage probabilities of the EL–DRM intervals are all between 94 % and 95 % so are very close to the nominal level. The coverage probabilities of the EM intervals are 1 % to 2 % less than those of the EL–DRM intervals, while the average lengths are somewhat greater. The simulation results for quantile differences between other populations are similar and are omitted.

Table 7.3 Simulation results for EL–DRM and EM confidence intervals of quantiles

	τ	0.05		0.10		0.30		0.50		0.60	
		EL	EM	EL	EM	EL	EM	EL	EM	EL	EM
ξ_0	Length:	1.09	1.22	0.99	1.10	0.91	0.96	0.94	0.97	0.99	1.01
	Coverage:	91.1	89.0	92.9	91.1	93.9	92.3	93.4	92.7	94.2	92.8
ξ_1	Length:	0.80	0.92	0.70	0.80	0.61	0.67	0.63	0.67	0.65	0.71
	Coverage:	91.6	90.2	93.0	91.9	94.4	93.4	94.7	93.5	94.4	94.7
ξ_2	Length:	1.04	1.21	0.93	1.05	0.83	0.91	0.87	0.91	0.91	0.94
	Coverage:	91.5	89.6	92.8	90.9	94.2	92.8	94.0	92.8	94.0	93.4
ξ_3	Length:	0.68	0.80	0.59	0.68	0.52	0.57	0.53	0.56	0.55	0.59
	Coverage:	92.3	91.8	93.3	93.1	94.8	94.1	94.9	94.6	94.8	95.6
$\xi_0 - \xi_1$	Length:	1.27	1.47	1.13	1.30	1.05	1.12	1.08	1.12	1.14	1.18
	Coverage:	94.7	92.7	94.7	93.1	94.3	94.0	94.3	93.7	94.8	94.5

Nominal level: 95 %; $\xi_0 - \xi_1$: difference between quantiles of F_0 and F_1; EL: EL–DRM intervals; EM: EM intervals

The simulation results for the Weibull, gamma and normal populations are quite similar to the reported results for the generalized gamma populations: the EL–DRM CDF and quantile point and interval estimators are generally more efficient than their empirical counterparts. The detailed results are omitted for brevity. It is worth noting that, compared to the EM quantile estimator, the efficiency gain of the EL–DRM quantile estimator for the Weibull distributions is as high as 100 % to 200 % for lower quantiles at $\tau = 0.05$ and 0.1. The efficiency gains of the EL–DRM estimators for both the normal and the gamma populations are between 10 % and 45 %.

7.7.2 Robustness to Model Misspecification

The DRM is flexible and includes a large number of distribution families as special cases. The model fits most situations, being applicable whenever the basis function $\boldsymbol{q}(x)$ is sufficiently rich. Misspecifying $\boldsymbol{q}(x)$ has an adverse effect on the estimation of DRM parameters (Fokianos and Kaimi 2006). However, in the targeted applications, and likely many others, the coefficients of $\boldsymbol{q}(x)$ in the model do not have direct interpretations. The adverse effect of model misspecification on the estimation of population distributions and quantiles is limited.

In this simulation, we choose the population distributions from among four different families: Weibull, generalized gamma, gamma, and log-normal. The parameter values are again chosen such that the shape and the first two moments of the populations approximately match real lumber data. These parameter values are listed in the first column of Table 7.4, where $W(\lambda, \kappa)$ stands for a Weibull distribution with shape λ and scale κ, $G(a, b)$ stands for a gamma distribution with shape a and rate b, and $LN(\mu, \sigma)$ stands for a log-normal distribution with mean μ and standard deviation σ on log scale. Note that the generalized gamma distribution in the current simulation is the F_1 used in the simulation in Sect. 7.7.1. These four distributions do not fit into any DRM, and therefore no true $\boldsymbol{q}(x)$ exists. In applications, these four families are often chosen to model positively distributed populations. None of them are necessarily true. Hence, taken in combination, they form a sensible example of a misspecified model.

In the simulation, the observations are censored around their 75 % population quantiles, which are 9.7, 6.5, 9.1, and 8.7. We choose $\boldsymbol{q}(x) = (\log x, x, x^2)^\intercal$, which combines the basis function that suits gamma distributions and the one that suits the generalized gamma distributions in our previous simulation. The settings for this simulation are otherwise the same as the one in the previous subsection.

The simulation results for quantile estimation are summarized in Table 7.4. As the fifth column of the table shows, the EL–DRM-based variance estimates obtained according to Theorem 6 closely match the simulated variances. The efficiency comparison again strongly favours the EL–DRM quantile estimator, showing gains in the range of 4 % to 45 %. The efficiency gain is most prominent at the 5 % and

Table 7.4 Simulation results for quantile estimation under misspecified DRM

	τ	ξ_r	$V(\hat{\xi}_r)$	$\frac{\hat{\sigma}^2(\hat{\xi}_r)}{nV(\hat{\xi}_r)}$	$\frac{V(\check{\xi}_r)}{V(\hat{\xi}_r)}$	$B(\hat{\xi}_r)(\%)$	$B(\check{\xi}_r)(\%)$
$W(4.5, 9)$	0.05	4.65	0.16	1.03	1.15	0.05	0.32
$c_0 = 9.7$	0.10	5.46	0.11	1.07	1.16	−0.06	−1.00
	0.30	7.16	0.07	1.06	1.15	0.12	−0.35
	0.50	8.30	0.06	1.00	1.06	−0.00	−0.29
	0.60	8.83	0.06	1.03	1.04	−0.01	−0.30
$GG(2, 2.6, 5)$	0.05	3.65	0.05	0.99	1.05	0.53	0.43
$c_1 = 6.5$	0.10	4.06	0.04	1.01	1.12	0.24	−0.70
	0.30	4.96	0.03	1.04	1.08	−0.05	−0.35
	0.50	5.62	0.03	1.02	1.06	0.05	−0.27
	0.60	5.95	0.03	1.03	1.06	0.06	−0.30
$\Gamma(20, 2.5)$	0.05	5.30	0.06	1.00	1.44	0.39	−1.10
$c_2 = 9.1$	0.10	5.81	0.05	1.01	1.27	0.19	−0.56
	0.30	6.97	0.04	1.02	1.13	0.00	−0.33
	0.50	7.87	0.05	1.00	1.08	0.01	−0.28
	0.60	8.32	0.05	1.01	1.06	0.01	−0.25
$LN(2, 0.25)$	0.05	4.90	0.04	0.99	1.35	0.46	−0.79
$c_3 = 8.7$	0.10	5.36	0.04	0.97	1.20	0.23	−0.46
	0.30	6.48	0.03	1.02	1.15	0.00	−0.26
	0.50	7.39	0.04	1.06	1.09	−0.00	−0.24
	0.60	7.87	0.05	1.06	1.05	0.02	−0.26

ξ_r: true quantile; $\hat{\xi}_r$ and $\check{\xi}_r$: EL–DRM and EM quantile estimates; V: simulated variance; $\hat{\sigma}^2$: average of variance estimates; B: relative bias in percentage

10 % quantiles. The relative bias of the EL–DRM estimator is smaller than that of the EM estimator in most cases, and both are reasonably small.

The results for quantile interval estimation are given in Table 7.5. The EL–DRM intervals for both quantiles and quantile differences have closer to nominal coverage probabilities in most cases compared to the EM intervals. In all cases, the EL–DRM intervals are also superior to the EM intervals in terms of average lengths.

In conclusion, the simulation results show that the EL–DRM method retains its efficiency gain even when there is a mild model misspecification.

We also conducted simulations for two-parameter Weibull populations and two-component normal mixture populations under misspecified DRMs. The results are similar to those presented and are omitted for brevity.

7.7.3 Robustness to Outliers

Often, the assumed model in an application fails to account for a small proportion of comparatively very large values in the sample. These outliers may introduce

Table 7.5 Simulation results for confidence intervals of quantiles under misspecified DRM

	τ	0.05		0.10		0.30		0.50		0.60	
		EL	EM	EL	EM	EL	EM	EL	EM	EL	EM
ξ_0	Length:	1.52	1.56	1.32	1.38	1.02	1.07	0.95	0.97	0.95	0.96
	Coverage:	89.9	87.2	93.0	90.6	94.1	92.2	93.1	92.5	93.3	92.8
ξ_1	Length:	0.87	0.91	0.75	0.80	0.65	0.67	0.65	0.67	0.67	0.70
	Coverage:	89.6	89.8	91.9	92.3	94.1	93.8	93.8	93.3	93.1	94.6
ξ_2	Length:	0.93	1.11	0.85	0.95	0.79	0.83	0.83	0.86	0.88	0.91
	Coverage:	92.2	90.1	92.6	91.8	94.1	93.1	93.8	92.4	93.1	92.9
ξ_3	Length:	0.79	0.92	0.73	0.81	0.72	0.76	0.79	0.82	0.86	0.88
	Coverage:	91.9	91.5	92.5	92.7	94.2	93.2	94.2	93.1	93.9	93.8
$\xi_0 - \xi_1$	Length:	1.65	1.70	1.46	1.51	1.16	1.20	1.09	1.12	1.11	1.14
	Coverage:	90.7	90.2	94.0	92.5	94.7	93.6	94.0	93.7	94.1	94.2

Nominal level: 95 %; $\xi_0 - \xi_1$: difference between quantiles of F_0 and F_1
EL: EL–DRM intervals; EM: EM intervals

substantial instability into the classical optimal inference methods. Therefore, specific robustified procedures are often developed to limit the influence of the potential outliers.

The proposed EL–DRM method for Type I censored data is by nature robust to large-valued outliers for lower quantile estimation. In fact, potential outliers would be censored automatically. We may purposely induce Type I censoring to full observations to achieve robust estimation of lower quantiles.

In this simulation, we form new populations by mixing the F_k in Sect. 7.7.1 with a 10 % subpopulation from a normal distribution. The mean of the normal subpopulation is set to 11, around the 95 % quantile of F_0, and its standard deviation is taken to be 1, about half of that of F_0. More precisely, the population distributions are mixtures given by

$$0.9F_k + 0.1N(11,\ 1),\quad k = 0, 1, 2, 3.$$

Accurately estimating the lower population quantiles remains our target. The observations are censored at the 85 % quantiles of the corresponding populations: 10.0, 7.8, 9.8, and 8.0. We compute the estimates for the 0.05, 0.1, and 0.15 population quantiles. The simulation results are reported in Table 7.6. The first column lists the parameter values of the populations, the second gives the levels at which the quantiles are estimated, and the third provides the corresponding true quantile values. The fourth and fifth columns are the relative biases of the EL–DRM estimators based on censored data ($\hat{\xi}_r$) and full data ($\hat{\xi}_r^{(f)}$). At the 0.05 quantiles, the relative biases of $\hat{\xi}_r$ are below 0.51 %, compared to between 4 and 8 % for $\hat{\xi}_r^{(f)}$.

The sixth and seventh columns are the simulated variances of $\hat{\xi}_r$ and $\hat{\xi}_r^{(f)}$, respectively. The variances of $\hat{\xi}_r$ are slightly larger than those of $\hat{\xi}_r^{(f)}$ in general. The eighth and ninth columns reflect the precision of the variance estimators.

Table 7.6 Simulation results for quantile estimation when samples contain large outliers

F_r	τ	ξ_r	$\mathrm{B}(\hat{\xi}_r)$	$\mathrm{B}(\hat{\xi}_r^{(f)})$	$\mathrm{V}(\hat{\xi}_r)$	$\mathrm{V}(\hat{\xi}_r^{(f)})$	$\frac{\hat{\sigma}^2(\hat{\xi}_r)}{n\mathrm{V}(\hat{\xi}_r)}$	$\frac{\hat{\sigma}^2(\hat{\xi}_r^{(f)})}{n\mathrm{V}(\hat{\xi}_r^{(f)})}$	$\mathrm{M}(\hat{\xi}_r)$	$\mathrm{M}(\hat{\xi}_r^{(f)})$	$\mathrm{M}(\check{\xi}_r)$
$0.9GG(2,\ 4.3,\ 2.8)$	0.05	3.73	0.44	7.64	0.86	0.71	0.98	0.61	0.86	1.52	1.12
$+0.1N(11,\ 1)$	0.10	4.37	−0.06	3.90	0.70	0.55	1.00	0.78	0.70	0.84	0.91
	0.15	4.83	−0.13	1.71	0.63	0.49	1.01	0.89	0.63	0.56	0.78
$0.9GG(2,\ 2.6,\ 5)$	0.05	3.71	0.51	−4.63	0.45	0.42	1.00	1.16	0.45	0.72	0.60
$+0.1N(11,\ 1)$	0.10	4.13	0.23	−2.81	0.35	0.34	1.00	1.26	0.36	0.47	0.47
	0.15	4.42	0.17	−1.74	0.31	0.31	1.01	1.30	0.31	0.37	0.39
$0.9GG(2,\ 3.8,\ 3.5)$	0.05	4.04	0.28	5.37	0.76	0.60	1.00	0.66	0.76	1.07	1.15
$+0.1N(11,\ 1)$	0.10	4.62	0.10	2.78	0.62	0.49	1.00	0.83	0.62	0.65	0.83
	0.15	5.04	0.11	1.24	0.55	0.44	1.00	0.93	0.55	0.48	0.72
$0.9GG(2,\ 3.8,\ 3.5)$	0.05	4.10	0.27	−4.05	0.32	0.27	0.99	1.14	0.32	0.55	0.47
$+0.1N(11,\ 1)$	0.10	4.51	0.18	−2.62	0.25	0.23	1.02	1.28	0.25	0.36	0.33
	0.15	4.79	0.12	−1.78	0.21	0.20	1.04	1.36	0.21	0.28	0.28

ξ_r: true quantile; $\hat{\xi}_r$ and $\hat{\xi}_r^{(f)}$: EL–DRM quantile estimates based on censored and full data; $\check{\xi}_r$: EM quantile estimates; V: simulated variance on the scale of 10^{-1}; $\hat{\sigma}^2$: average of variance estimates; B: relative bias in percentage; M: MSE on the scale of 10^{-1}

Table 7.7 Simulation results for confidence intervals of quantiles when samples contain large outliers

	τ	0.05			0.10			0.15		
		EL	EL(f)	EM	EL	EL(f)	EM	EL	EL(f)	EM
ξ_0	Length:	1.10	0.81	1.26	1.02	0.81	1.14	0.98	0.81	1.07
	Coverage:	91.8	63.9	89.1	93.3	81.6	91.0	93.6	90.8	91.5
ξ_1	Length:	0.81	0.85	0.94	0.73	0.80	0.83	0.68	0.78	0.77
	Coverage:	92.0	92.4	90.3	93.1	94.8	92.3	93.6	96.3	92.7
ξ_2	Length:	1.05	0.77	1.24	0.96	0.78	1.10	0.91	0.79	1.03
	Coverage:	92.3	72.3	89.3	93.6	85.9	91.2	93.8	92.1	92.0
ξ_3	Length:	0.69	0.69	0.83	0.61	0.66	0.71	0.58	0.65	0.65
	Coverage:	92.7	89.0	92.1	93.8	93.5	93.1	94.2	95.8	93.6
$\xi_0 - \xi_1$	Length:	1.27	0.71	1.51	1.17	0.92	1.34	1.11	1.02	1.26
	Coverage:	94.8	63.4	92.5	95.1	83.3	92.7	94.7	92.6	93.2

Nominal level: 95 %; $\xi_0 - \xi_1$: difference between quantiles of population 0 and 1 EM: EM intervals; EL: EL–DRM intervals based on censored data; EL(f): EL–DRM intervals based on full data

The entries for $\hat{\xi}_r$ are all close to 1, showing that both precision and stability of variance estimation for $\hat{\xi}_r$ are superior. The entires for $\hat{\xi}_r^{(f)}$ based on the full data fluctuate between 0.61 and 1.36, revealing non-robustness and inaccurate variance estimation.

We also report the MSEs of the estimators in the last three columns of Table 7.6. In most of the cases, $\hat{\xi}_r$ is superior to both $\hat{\xi}_r^{(f)}$ and $\check{\xi}_r$. The gains in MSE are most remarkable at the 0.05 quantile.

Table 7.7 shows the simulation results for quantile confidence intervals. Because of the outliers, the EL–DRM intervals based on the full data in most cases have much lower coverage probabilities than the nominal 95 % level for 0.05 and 0.1 quantiles; however, the performance improves much for the 0.15 quantile. The EL–DRM intervals based on censored data have much closer to nominal coverage probabilities in general.

In conclusion, for lower quantile estimation, it makes sense to collect much cheaper censored data, as doing so will result in a large gain in robustness with only a very mild loss of efficiency. If efficiency is deemed especially important, the money saved by collecting cheaper data could be used to increase the sample size.

7.8 Simulation Studies II: EL Ratio Test

We now carry out simulations to study the power of the EL ratio test under correctly specified and misspecified DRMs. As in simulation studies for CDF and quantile estimations, we set the number of populations to be $m + 1 = 4$, and consider populations with generalized gamma distributions. The parameters of these

distributions are again chosen so that their first two moments closely match those of the lumber strength samples in our application. Again, we set the number of simulation repetitions to 10,000.

Since our primary application is to detect difference among lumber populations, we test the following hypotheses in our simulations

$$H_0 : F_0 = F_1 = F_2 = F_3 \quad \text{against} \quad H_a : F_i \neq F_j \ \text{ for some } \ i, j = 0, \ldots, 3 \tag{7.15}$$

at the significance level of 0.05. This is the same as (7.10) with $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta}$.

In all simulations, the four samples are set to Type I censored at 0.9, 0.8, 0.87 and 0.83 population quantiles of the baseline distribution F_0, respectively.

When samples are Type I censored, the proposed EL ratio test does not have many non or semiparametric competitors. A straightforward competitor would be a Wald type test based on the asymptotic normality of $\boldsymbol{\beta}$ under the DRM. It uses the test statistic $n\hat{\boldsymbol{\beta}}^{\mathsf{T}} \hat{\Sigma}^{-1} \hat{\boldsymbol{\beta}}$ with $\hat{\Sigma}$ being a consistent estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. The Wald type test also has a chi-square limiting reference distribution under H_0 of (7.15). Probably the most commonly used semiparametric competitor is the partial likelihood ratio test based on the celebrated Cox Proportional Hazards (CoxPH) model (Cox 1972). The CoxPH model for multiple samples amounts to assuming

$$h_k(x) = \exp(\beta_k) h_0(x),$$

where $h_k(x)$ is the hazard function of the kth sample. Clearly, the CoxPH model impose strong restrictions on how $m + 1$ populations are connected. In comparison, the DRM is more flexible by allowing the density ratio to be a function of x. This limitation of the CoxPH approach has been shown by Cai et al. (2016) for the case of uncensored multiple samples: the power of the partial likelihood ratio test under the CoxPH model is lower than that of the DEL ratio test under the DRM when the hazards ratio of populations is not constant in x.

In our simulations, we compare the powers of the proposed EL ratio test under the DRM (ELRT), the Wald type test under the DRM (Wald-DRM), and the partial likelihood ratio test under the CoxPH model. In the CoxPH model, we use m dummy population indicators as covariates. The corresponding partial likelihood ratio has a χ^2_m limiting distribution under the null hypothesis of (7.15).

7.8.1 Populations Satisfying the DRM Assumption

We consider two settings for population distributions: generalized gamma distributions with fixed $a = 2$ in the first setting, and generalized gamma distributions with fixed $a = 1$ in the second. Note that in the second settings, $a = 1$ gives regular

Table 7.8 Parameter values for power comparison under correctly specified DRMs (Sect. 7.8.1). F_0 remains unchanged across parameter settings 0–5

$GG(a, b, p)$: generalized gamma distribution with parameters a, b and p											
Parameter settings											
		1		2		3		4		5	
F_0		b	p	b	p	b	p	b	p	b	p
$GG(2, 6, 0.8)$	F_1:	6.6	0.75	6.9	0.70	7.2	0.68	7.2	0.7	7.5	0.66
	F_2:	6.35	0.66	6.35	0.63	6.65	0.61	6.6	0.65	6.8	0.60
$a = 2$ in all settings	F_3:	5.5	0.88	5.3	0.92	5.2	1	5.2	1.2	5.1	1.3
$GG(1, 2, 2)$;	F_1:	2.2	1.8	2.3	1.75	2.4	1.72	2.5	1.62	2.6	1.40
$a = 1$ in all settings	F_2:	1.75	2.2	1.7	2.3	1.7	2.35	1.6	2.45	1.55	2.5
(gamma)	F_3:	1.8	2.3	1.78	2.55	1.78	2.7	1.75	2.8	1.7	2.9

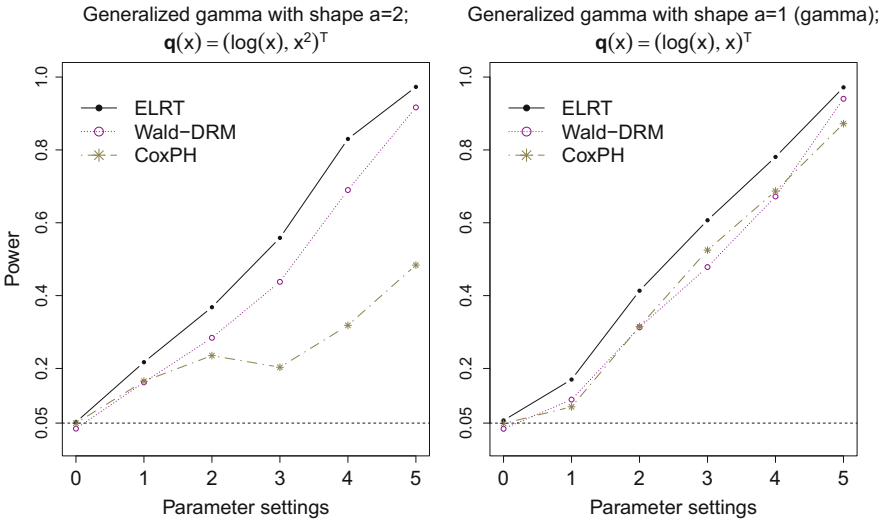

Fig. 7.1 Power curves of the ELRT, Wald-DRM, and CoxPH; the ELRT and Wald-DRM are based on correctly specified DRMs; the parameter setting 0 corresponds to the null model and the settings 1–5 correspond to alternative models

gamma distributions. Under each setting of population distributions we compare the power of the three competing tests for six parameter settings (settings 0–5), with parameter setting 0 under the null hypothesis H_0 and parameter settings 1–5 under the alternative hypothesis H_a. These parameter settings are given in Table 7.8.

The populations in the first distribution setting satisfy a DRM with basis function $\boldsymbol{q}(x) = (\log(x), x^2)^\mathsf{T}$, and those in the second distribution setting satisfy a DRM with basis function $\boldsymbol{q}(x) = (\log(x), x)^\mathsf{T}$. We fit DRMs with these basis functions to the corresponding Type I censored samples.

The power curves of the three testing methods are shown in Fig. 7.1. The type I errors of the ELRT and the CoxPH are very close to the nominal level, while

that of the Wald-DRM is a little lower than the nominal level. The ELRT has the highest power under all the settings. The Wald-DRM is also more powerful than the CoxPH under most parameter settings. In summary, under our simulation settnigs, the DRM-based tests are more powerful the CoxPH, and the proposed ELRT is the most powerful of all.

7.8.2 Robustness to Model Misspecification

We now examine the power of the ELRT under misspecified DRMs. As we have argued, since the DRM is very flexible, a mild misspecification of the basis function should not significantly reduce the power of a DRM-based test. We now demonstrate this idea by simulation studies.

We again consider two settings for population distributions: three-parameter generalized gamma distributions in the first setting, and generalized gamma distributions with fixed $p = 1$ in the second. Note that in the second settings, $p = 1$ gives the two-parameter Weibull distributions. As in the last simulation study, under each setting of population distributions we compare the power of the three competing tests for six parameter settings (settings 0–5), with parameter setting 0 under the null hypothesis H_0 and parameter settings 1–5 under the alternative hypothesis H_a. These parameter settings are given in Table 7.9.

The populations in neither distribution setting satisfy the DRM assumption (7.1). In either case, we still fit a DRM with the basis function that is suitable to gamma populations, $\boldsymbol{q}(x) = (\log(x), x)^{\mathsf{T}}$, to the censored samples. Such a basis function is chosen by the shapes of the histograms of the samples: both generalized gamma and Weibull samples have similar shapes to gamma samples, and hence are easily recognized falsely as from the gamma family.

The simulated rejection rates of the three tests are shown in Fig. 7.2. It is clear that, although the DRMs are misspecified, the ELRT still has the highest power while its type I error rates are close to the nominal. Again, the Wald-DRM is not as powerful as the ELRT, but is more powerful than the CoxPH in most cases.

7.9 Analysis of Lumber Quality Data

Two important lumber strength measures are *modulus of tension* (MOT) and *modulus of rupture* (MOR). They measure, respectively, the tension and bending strengths of lumber. Three MOT and three MOR samples were collected in labs.

The first MOT sample (MOT 0) is not subject to censoring, the second (MOT 1) is right-censored at 5.0×10^3 pounds per square inch (psi), and the third (MOT 2) at 4.0×10^3 psi. The size of each sample is 80. The number of uncensored observations for MOT 1 and MOT 2 are 52 and 38, respectively. The kernel density plots of these samples are shown in Fig. 7.3a.

Table 7.9 Parameter values for power comparison under misspecified DRMs (Sect. 7.8.2). F_0 remains unchanged across parameter settings 0–5

$GG(a, b, p)$: generalized gamma distribution with parameters a, b and p															
Parameter settings															
		1			2			3			4			5		
F_0		a	b	p	a	b	p	a	b	p	a	b	p	a	b	p
$GG(1.5, 2, 2)$	F_1:	1.47	2.1	1.85	1.45	2.2	1.85	1.4	2.3	1.8	1.4	2.3	1.75	1.35	2.35	1.7
	F_2:	1.42	2.2	1.8	1.4	2.3	1.75	1.35	2.4	1.65	1.3	2.43	1.6	1.3	2.55	1.55
	F_3:	1.55	1.9	2.2	1.6	1.8	2.45	1.65	1.75	2.5	1.7	1.8	2.5	1.75	1.78	2.65
$GG(2, 6, 1)$;	F_1:	1.75	5.7	1	1.7	5.6	1	1.65	5.55	1	1.6	5.45	1	1.5	5.4	1
$p = 1$ in all settings	F_2:	2.1	6.2	1	2.2	6.3	1	2.3	6.35	1	2.4	6.45	1	2.5	6.6	1
(Weibull)	F_3:	1.87	5.9	1	1.8	5.8	1	1.75	5.7	1	1.7	5.7	1	1.6	5.6	1

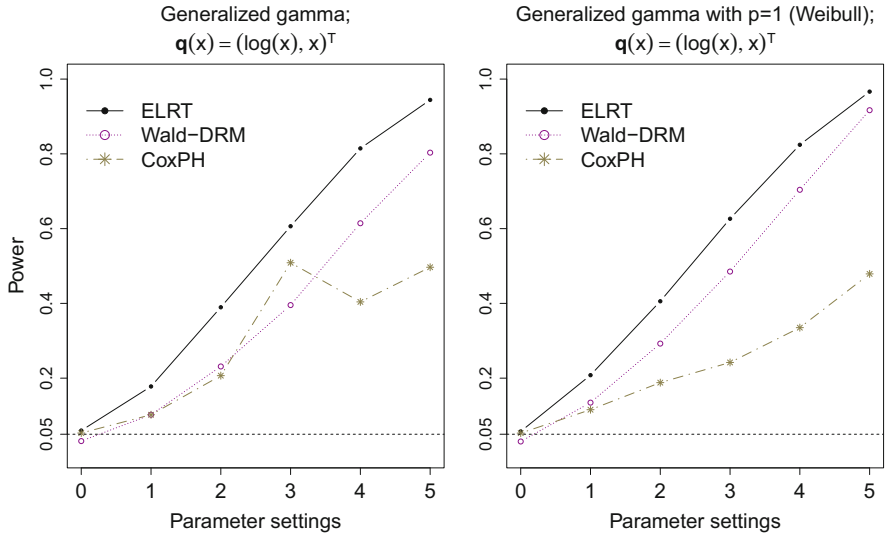

Fig. 7.2 Power curves of the ELRT, Wald-DRM, and CoxPH; the ELRT and Wald-DRM are based on misspecified DRMs; the parameter setting 0 corresponds to the null model and the settings 1–5 correspond to alternative models

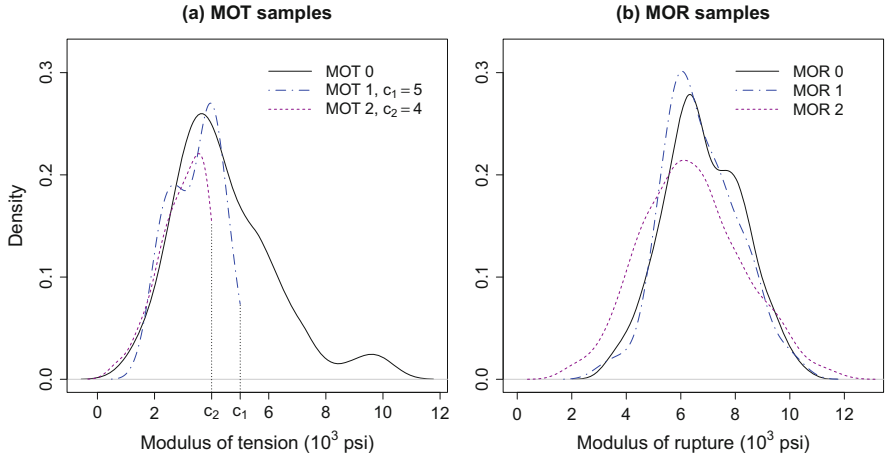

Fig. 7.3 Kernel density plots of MOT and MOR samples

We fit a DRM with basis function $\boldsymbol{q}(x) = (\log x,\ x,\ x^2)^{\mathsf{T}}$ to the samples. Quantile estimations are similar when other two- and three-dimensional basis functions are used, although the estimates for $\boldsymbol{\theta}$ are quite different.

We compute the EL–DRM and EM quantile estimates ($\hat{\xi}_r$ and $\check{\xi}_r$), along with their estimated standard deviations, at $\tau = 0.05$, 0.1, 0.3, 0.5 and 0.6. Note that

Table 7.10 Quantile estimates for MOT populations (unit: 10^3 psi)

F_r	τ	$\hat{\xi}_r$	$\hat{\sigma}(\hat{\xi}_r)$	$\check{\xi}_r$	$\hat{\sigma}(\check{\xi}_r)$
MOT 0	0.05	2.11	3.38	2.08	5.33
	0.10	2.47	3.15	2.47	3.67
	0.30	3.45	2.82	3.28	2.58
	0.50	4.11	2.68	4.11	3.12
	0.60	4.36	3.21	4.38	3.31
MOT 1	0.05	2.11	2.49	2.10	2.51
	0.10	2.45	2.50	2.30	2.85
	0.30	3.33	3.61	3.48	3.85
	0.50	4.19	3.27	4.15	3.11
	0.60	4.60	4.13	4.55	4.89
MOT 2	0.05	1.72	5.17	1.66	5.84
	0.10	2.29	3.82	2.19	3.97
	0.30	3.38	3.24	3.37	3.70
	0.50	4.09	2.85	NA	NA
	0.60	4.33	6.36	NA	NA

$\hat{\xi}_r$: EL–DRM quantile estimates; $\check{\xi}_r$: EM quantile estimates; $\hat{\sigma}$: estimated standard deviation; NA: $\check{\xi}_r$ not well-defined

only 47.5 % of observations in the third MOT sample are uncensored, invalidating the EM estimator at $\tau = 0.5$ and 0.6. The EL–DRM estimator, however, is effective at all τ values.

The results are shown in Table 7.10. The EL–DRM and EM quantile estimates are close. Except in two cases, we see that $\hat{\sigma}(\hat{\xi}_r)$, the estimated standard error of $\hat{\xi}_r$, is smaller than $\hat{\sigma}(\check{\xi}_r)$, the estimated standard error of $\check{\xi}_r$. On average, $\hat{\sigma}(\hat{\xi}_r)$ is 12 % smaller than $\hat{\sigma}(\check{\xi}_r)$. Such an efficiency gain is likely to be real, as implied by our earlier simulation studies.

We next analyze the non-censored MOR samples of sizes 282, 98, and 445, denoted MOR 0, MOR 1 and MOR 2 in the kernel density plots shown in Fig. 7.3b. There are no obvious large outliers.

In addition to being robust, the EL–DRM estimator based on Type I censored data is believed to lose little efficiency compared to that based on full data for lower quantile estimation. To illustrate this point, we induced right censoring to all the MOR samples around their 85 % quantiles at 8.4, 8.3 and 8.3 ($\times 10^3$ psi). We fitted a DRM with basis function $\boldsymbol{q}(x) = (\log x,\, x,\, x^2)^{\mathsf{T}}$ to the censored and full samples, respectively.

The quantile estimates for the MOR populations and the corresponding estimated standard deviations ($\hat{\sigma}$) are shown in Table 7.11. The three estimators give similar quantile estimates. The standard deviations of $\hat{\xi}_r$ based on censored observations are close to the ones based on the full data. The average of the ratio $\hat{\sigma}(\hat{\xi}_r^{(f)})/\hat{\sigma}(\hat{\xi}_r)$ is 0.98. Moreover, $\hat{\sigma}(\hat{\xi}_r)$ and $\hat{\sigma}(\hat{\xi}_r^{(f)})$ are 16 % and 18 % smaller than $\hat{\sigma}(\check{\xi}_r)$ on average. In conclusion, the EL–DRM estimator based on the censored data is almost as

Table 7.11 Quantile estimates for MOR populations (unit: 10^3 psi)

F_r	τ	$\hat{\xi}_r$	$\hat{\sigma}(\hat{\xi}_r)$	$\hat{\xi}_r^{(f)}$	$\hat{\sigma}(\hat{\xi}_r^{(f)})$	$\check{\xi}_r$	$\hat{\sigma}(\check{\xi}_r)$
MOR 0	0.05	4.52	3.62	4.53	3.70	4.40	5.50
	0.10	4.96	3.54	5.00	3.44	5.00	4.10
	0.30	5.99	2.69	6.02	2.55	6.05	2.82
	0.50	6.74	3.06	6.72	2.84	6.67	3.14
	0.60	7.11	3.56	7.06	3.16	7.23	4.44
MOR 1	0.05	4.57	5.86	4.55	5.61	4.70	8.01
	0.10	5.03	5.35	4.97	5.20	5.17	4.92
	0.30	5.93	3.76	5.93	3.89	5.88	3.86
	0.50	6.59	4.25	6.62	4.41	6.46	5.32
	0.60	6.88	4.71	6.95	4.79	6.92	6.46
MOR 2	0.05	3.54	4.08	3.58	3.91	3.58	4.43
	0.10	4.03	3.91	4.03	3.84	4.08	3.74
	0.30	5.36	3.25	5.34	3.18	5.32	3.58
	0.50	6.26	2.82	6.21	2.80	6.27	3.35
	0.60	6.72	3.08	6.72	3.13	6.76	3.01

$\hat{\xi}_r$ and $\hat{\xi}_r^{(f)}$: EL–DRM quantile estimates based on censored and full data; $\check{\xi}_r$: EM quantile estimates; $\hat{\sigma}$: estimated standard deviation

efficient as the one based on the full data, and both are more efficient than the EM estimator.

7.10 Concluding Remarks

We have developed EL–DRM-based statistical methods for multiple samples with Type I censored observations. The proposed EL ratio test is shown to have a simple chi-square limiting distribution under the null model of a composite hypothesis about the DRM parameters. The limiting distribution of the EL ratio under a class of local alternative models is shown to be non-central chi-square. This result is useful for approximating the power of the EL ratio test and calculating the required sample size for achieving a certain power. Simulations show that the proposed EL ratio test has a superior power compared to some semiparametric competitors under a wide range of population distribution settings.

The proposed CDF and quantile estimators are shown to be asymptotically normal, as well as to be more efficient and have a broader range of consistent estimation than their empirical counterparts. Extensive simulations support these theoretical findings. The advantages of the new methods are particularly remarkable for lower quantiles. Simulation results also suggest that the proposed method is robust to mild model misspecifications and useful when data are corrupted by large outliers.

This work is motivated by a research project on the long-term monitoring of lumber quality. The proposed methods have broad applications in reliability engineering and medical studies, where Type I censored samples are frequently encountered.

References

Bahadur RR (1966) A note on quantiles in large samples. Ann Math Stat 37(3):577–580
Cai S, Chen J, Zidek JV (2016) Hypothesis testing in the presence of multiple samples under density ratio models. In press Stat Sin
Chen H, Chen J (2000) Bahadur representation of the empirical likelihood quantile process. J Nonparametric Stat 12:645–665
Chen J, Liu Y (2013) Quantile and quantile-function estimations under density ratio model. Ann Stat 41(3):1669–1692
Cox DR (1972) Regression models and life-tables. J R Stat Soc Ser B (Stat Methodol) 34(2):187–220
Fokianos K (2004) Merging information for semiparametric density estimation. J R Stat Soc Ser B (Stat Methodol) 66(4):941–958
Fokianos K, Kaimi I (2006) On the effect of misspecifying the density ratio model. Ann Inst Stat Math 58:475–497
Marshall AW, Olkin I (2007) Life distributions – structure of nonparametric, semiparametric and parametric families. Springer, New York
Owen AB (2001) Empirical likelihood. Chapman & Hall, New York
Qin J (1998) Inferences for case-control and semiparametric two-sample density ratio models. Biometrika 85(3):619–630
Qin J, Zhang B (1997) A goodness-of-fit test for the logistic regression model based on case-control data. Biometrika 84:609–618
Silverman BW (1986) Density estimation for statistics and data analysis, 1st edn. Chapman & Hall, Boca Raton
Stacy EW (1962) A generalization of the gamma distribution. Ann Math Stat 33(3):1187–1192
Wang C, Tan Z, Louis TA (2011) Exponential tilt models for two-group comparison with censored data. J Stat Plan Inference 141:1102–1117
Zhang B (2002) Assessing goodness-of-fit of generalized logit models based on case-control data. J Multivar Anal 82:17–38

Chapter 8
Recent Development in the Joint Modeling of Longitudinal Quality of Life Measurements and Survival Data from Cancer Clinical Trials

Hui Song, Yingwei Peng, and Dongsheng Tu

Abstract In cancer clinical trials, longitudinal Quality of Life (QoL) measurements and survival time on a patient may be analyzed by joint models, which provide more efficient estimation than modeling QoL data and survival time separately, especially when there is strong association between the longitudinal measurements and the survival time. Most joint models in the literature assumed classical linear mixed model for longitudinal measurements and Cox's proportional hazards model for survival times. The linear mixed model with normal distributed random components may not be sufficient to model bounded QoL measurements. Moreover, when some patients are immune to recurrence of relapse and can be viewed as cured, the proportional hazards model is not suitable for survival times. In this paper, we review some recent developments in joint models to deal with bounded longitudinal QoL measurements and survival times with a possible cure fraction. One of such joint models assumes a linear mixed tt model for longitudinal measurements and a promotion time cure model for survival data, and the two parts are linked through a latent variable. Another joint model employs a simplex distribution to model the bounded QoL measurements and a classical proportional hazard to model survival times, and the two parts share a random effect. Semiparametric estimation methods have been proposed to estimate the parameters in the models. The models are illustrated with QoL measurements and recurrence times from a clinical trial on women with early breast cancer.

H. Song (✉)
School of Mathematical Sciences, Dalian University of Technology, 116024, Dalian, Liaoning, China
e-mail: songh@dlut.edu.cn

Y. Peng • D. Tu
Departments of Public Health Sciences and Mathematics and Statistics, Queens University, K7L 3N6, Kingston, ON, Canada
e-mail: yingwei.peng@queensu.ca; dtu@ctg.queensu.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_8

8.1 Introduction

In cancer clinical trials, the patient's quality of life (QoL) is an important subjective endpoint beyond the traditional objective endpoints such as tumor response and relapse-free or overall survival. Specially, when the improvement in survival may be limited by a new treatment for a specific type of cancer, patients' QoL is important to determine whether this new treatment is useful (Richards and Ramirez 1997). Several studies (Dancey et al. 1997; Ganz et al. 2006) have also found that QoL measurements such as overall QoL, physical well-being, mood and pain are of prognostic importance for patients with cancer, so they may help to make a treatment decision.

The quality of life of cancer patients is usually assessed by a questionnaire, which consists of a large number of questions assessing various aspects of quality of life, at different timepoints before, during, and after their cancer treatments. The score for a specific domain or scale of QoL is usually calculated as the mean of the answers from a patient to a set of the questions which define this QoL domain or scale and, therefore, can be considered as a continuous measurement. In this chapter, we only consider measurements of a specific QoL domain or scale, which can be defined as $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$, where Y_{ij} is the measurement of this QoL domain or scale from the ith subject at the jth occasion, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$. These longitudinal QoL measurements from cancer clinical trials can be analyzed by standard statistical methods for repeated measurements, such as linear mixed models (Fairclough 2010). These models provide valid statistical inference when complete longitudinal measurements from all patients are available or the missing longitudinal measurements can be assumed missing at random. In cancer clinical trials, some seriously ill patients who have worse QoL may drop out of the study because of disease recurrence or death. The QoL measurements are not available from these patients. In this case, dropping out is directly related to what is being measured, and the missing of QoL measurements caused by the dropout of these patients is informative and may not be assumed as missing at random. Application of a standard longitudinal data analysis in this context could give biased results. Tu et al. (1994) analyzed the longitudinal QoL data of a breast cancer clinical trial based on a nonparametric test taking into account of informative censoring.

Patients' poor QoL can lead to both informative dropout for QoL measurements and censoring in survival time. Let T_i' be the survival time and δ_i be the censoring indicator of the ith subject. A joint modeling framework for longitudinal QoL measurements Y_i and survival data (T_i', δ_i) allows modeling both longitudinal measurements and survival time measurements jointly to accommodate the association between them. It may provide a valid inference for data with missing measurements caused by the dropout of patients. As pointed out in an introductory review by Ibrahim et al. (2010), Tsiatis and Davidian (2004) and Diggle et al. (2008), joint analysis of longitudinal QoL data together with data on a time to event outcome, such as disease-free or overall survival would also provide more efficient estimates of the treatment effects on both QoL and time to event endpoints and also reduce bias in the estimates of overall treatment effect.

Since Schluchter (1992) proposed the analysis of the longitudinal measurements jointly with the dropout time by a multivariate normal distribution, many approaches have been proposed for jointly modeling longitudinal measurements and survival data. Some recent reviews can be found in, for example, Wu et al. (2012) and Gould et al. (2015).

In this paper, we will review some of our recent contributions in this area, which were motivated from the analysis of data from a clinical trial on early breast cancer. We developed some new joint models which take into account of features of QoL and survival data observed from the clinical trials. The models employ the Student's t distribution for the random components of the model for the longitudinal QoL data to better accommodate extreme observations. The simplex distribution is considered to accommodate longitudinal bounded QoL data. A promotion time cure model is considered for survival data to allow a possible cure fraction in the patients. To simplify the model estimation, a penalized joint likelihood generated by the Laplace approximation is attempted to estimate the parameters in the models. The details of these new models and statistical procedures for the inference of parameters in these models are reviewed in the following sections, followed by an illustration with the breast cancer data. The paper is concluded with conclusions and discussions.

8.2 QoL and Survival Data from a Breast Cancer Clinical Trial

The data that motivated our recent contributions are from a clinical trial (MA.5) on early breast cancer conducted by NCIC Clinical Trials Group which compared two chemotherapy regimens for early breast cancer, a new and intensive treatment of cyclophosphamide, epirubicin, and fluorouracil (CEF) with the standard treatment of cyclophosphamide, methotrexate, and fluorouracil (CMF) (Levine et al. 1998, 2005). In this study, 356 patients were randomly assigned to CEF and 360 patients to the CMF arm. Both CEF and CMF were administered monthly for six months. The survival time of our interest is the time to relapse or recurrence free survival time (RFS), which was the primary endpoint of this trial. The median follow-up time of all patients is 59 months, and there are 169 and 132 uncensored RFS times from patients randomized to CEF and CMF respectively in the data set updated in 2002. The difference in 10-year relapse-free survival between two treatment arm is 7 % (52 % on CEF and 45 % on CMF, respectively), which was considered as moderate and may not be able to completely determine the relative advantages of CEF over CMF. Therefore, information on QoL may be helpful in the process of decision-making (Brundage et al. 2005).

In MA.5, the quality of patients undergoing chemotherapy treatment was assessed by the self-answered Breast Cancer Questionnaire (BCQ) which consists of 30 questions measuring different dimensions of QoL and was administered at each of the clinical visits (every one during the treatment and then every 3 months

after the completion of the treatment) until the end of the second year or until recurrence or death, whichever came first. The specific QoL scale of interest is the global QoL of patients defined as the means of the answers to all 30 questions. Because of drop-out of patients due to recurrence or death, joint analysis of RFS and the global QoL would provide more robust and also efficient inference on the difference in global QoL between two treatment groups.

Joint modeling of longitudinal QoL measures and survival data was studied previously by Zeng and Cai (2005). They assumed a linear mixed effects model to the longitudinal QoL measurements Y_i and a multiplicative hazards model to the survival time T_i'. These two models share a vector of common random effects, which reflects the unobserved heterogeneity for different subjects and its coefficient in the model characterizes the correlation between the longitudinal QoL measurements and survival times. The EM algorithm was implemented by them to calculate the maximum likelihood estimates. This simultaneous joint model with shared random effect established a basic framework for the joint modeling of longitudinal QoL measurements and survival data but during the process to apply it to analyze data from MA.5, we found it necessary to extend this framework for several reasons. One reason is, for early breast cancer, with advances in the development of new cancer treatments, the existence of cured patients becomes possible and, therefore, the models for the survival times need to take this into consideration. This can also be seen from the plateau of the Kaplan-Meier survival curves of the two treatments shown in Fig. 8.1. Another reason is that QoL measurements may be restricted to an interval. For example, since each question on BCQ are on a Likert scale from 0 to 7 with the best outcome marked as 7, the minimum and maximum of the scores

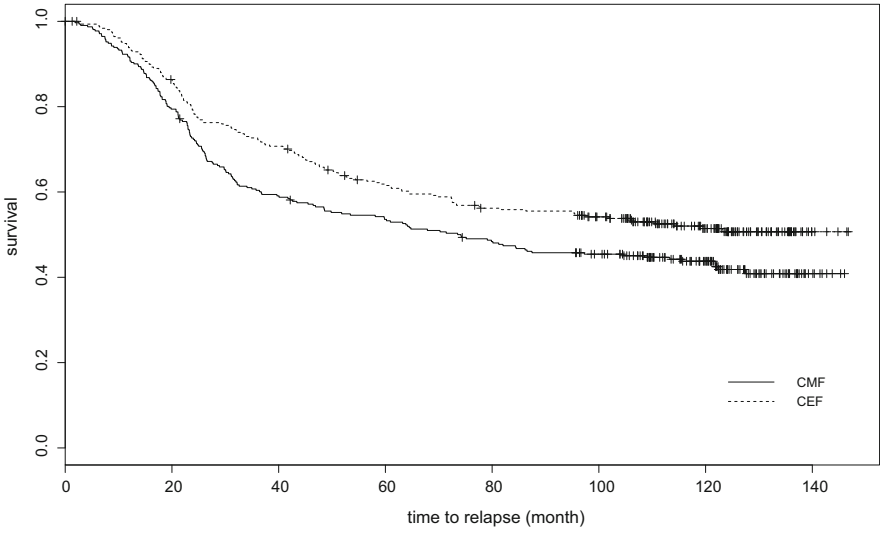

Fig. 8.1 Kaplan-Meier survival curves of patients in the two treatment groups of MA.5

for global QoL are respectively 0 and 7. Such data is often called bounded data (Aitchison 1986). There are totally 7769 QoL measurements from both arms.

8.3 Modeling QoL and Survival Data with a Cure Fraction

As mentioned above, Zeng and Cai (2005) considered the change in longitudinal QoL measures over time and the time to event as two simultaneous processes. They used a classical linear mixed model for the longitudinal QoL measurements Y_i and a Cox proportional hazards model with random effects for the survival time T_i' and assumed the random effects in these two models are the same. When there are potential cured patients, this simultaneous joint model may not be suitable to model longitudinal QoL measures and survival times. To accommodate a possible cure fraction in the data, we proposed to use a linear mixed tt model for the longitudinal QoL measurements Y_i and a promotion time cure model for T_i' (Song et al. 2012). These two models are connected through a latent gamma variate, which defines a joint model for the longitudinal QoL measurements and survival time. Let ξ_i be the latent gamma variate with the density function $P(\xi_i|\eta, \eta)$ so that $E[\xi_i] = 1$ and $var[\xi_i] = 1/\eta$. Given ξ_i, the proposed joint model specifies that the longitudinal QoL measurements Y_i in the joint model is assumed to follow the linear mixed *tt* model,

$$Y_i = X_i\beta + \tilde{X}_i\alpha_i + e_i, \tag{8.1}$$

where X_i and $\tilde{X}_i$ are the design matrices of covariates for the fixed effects, β and the random effects α_i respectively, and e_i is the random error. Given ξ_i, α_i and e_i are assumed to be independent and both are normal random variates with $\alpha_i|\xi_i \sim N(0, \sigma_\alpha^2/\xi_i)$ and $e_i|\xi_i \sim N(0, I_{n_i\times n_i}\sigma_e^2/\xi_i)$.

It can be shown that the marginal distributions of α_i and e_i are respectively $t(0, \sigma_\alpha^2, 2\eta)$ and $t_{n_i}(0, \sigma_e^2 I_{n_i\times n_i}, 2\eta)$, where $t_{n_i}(0, \sigma_e^2 I_{n_i\times n_i}, 2\eta)$ is an n_i-dimensional t distribution (Pinheiro et al. 2001) with mean 0, a positive definite scale matrix $\sigma_e^2 I_{n_i\times n_i}$ and a degree of freedom 2η, and $t(0, \sigma_\alpha^2, 2\eta)$ is that distribution for univariate variable (Cornish 1954; Lange et al. 1989). Therefore, Eq. (8.1) is often called the linear mixed *tt* model (Song 2007; Song et al. 2007; Zhang et al. 2009). Since the marginal variances of α_i and e_i are respectively $var[\alpha_i] = \frac{\eta}{\eta-1}\sigma_\alpha^2$ and $var[e_i] = \frac{\eta}{\eta-1}\sigma_e^2 I_{n_i\times n_i}$, a condition $\eta > 1$ in the joint model is required to guarantee the existence of the variances.

Given ξ_i, the survival time T_i' in the joint model is assumed to follow a conditional distribution with the following survival function,

$$S_{pop}(t|Z_i, \xi_i) = \exp[-\xi_i\theta(Z_i)F_0(t)], \tag{8.2}$$

where $F_0(t)$ is an arbitrary proper distribution function, $\theta(Z_i) = \exp[\gamma^T Z_i]$ with Z_i the vector of covariates which may have partial or complete overlap with X_i, and γ is the corresponding regression coefficients. Eq. (8.2) is often called the promotion time cure model (Chen et al. 1999; Yakovlev et al. 1993; Yin 2005). The unconditional cure probability for a subject with Z_i under this model is $\lim_{t\to\infty} \int_0^\infty S_{pop}(t|Z_i, \xi_i)P(\xi_i|\eta, \eta)d\xi_i = (\frac{\theta(Z_i)}{\eta} + 1)^{-\eta}$.

Equations (8.1) and (8.2) are connected by the latent gamma variate ξ_i, which acts multiplicatively on the hazard function of T_i' and on the variances of α_i and e_{ij} in the model for Y_i. These two equations together define a joint model for the longitudinal QoL measures and survival times, referred as JMtt.

Since conditional on α_i, ξ_i, the contribution of y_i to the likelihood under (8.1) is

$$l_{li}(y_i|\alpha_i, \xi_i) = \frac{1}{(\sqrt{2\pi\sigma_e^2/\xi_i})^{n_i}} \exp[-\frac{1}{2\sigma_e^2/\xi_i}(y_i - X_i\beta - \tilde{X}_i\alpha_i)^T(y_i - X_i\beta - \tilde{X}_i\alpha_i)]$$

and the contribution of (t_i, δ_i) to the likelihood under (8.2) and independent censoring is

$$l_{si}(t_i|\alpha_i, \xi_i) = [\xi_i\theta(Z_i)f_0(t_i)]^{\delta_i} \exp[-\xi_i\theta(Z_i)F_0(t_i)].$$

The observed likelihood for the proposed joint model based on all n subjects is

$$l = \prod_{i=1}^{n} \iint l_{li}(y_i|\alpha_i, \xi_i, \sigma_e^2, \beta)l_{si}(t_i|\alpha_i, \xi_i, \gamma, F_0(t))\varphi(\alpha_i|0, \sigma_\alpha^2/\xi_i)\psi(\xi_i|\eta, \eta)d\alpha_i d\xi_i,$$

where $\varphi(\cdot|0, \sigma_\alpha^2/\xi_i)$ is the density function of the normal distribution with mean 0 and variance σ_α^2/ξ_i, and $\psi(\xi_i|\eta, \eta)$ is the density function of the gamma distribution with mean 1 and variance $1/\eta$. The unknown parameters in this model are denoted as $\Theta = (\beta, \gamma, \eta, \sigma_\alpha, \sigma_e, F_0(t))$.

EM algorithm is used to obtain the maximum likelihood estimates of parameters (Klein 1992). If α_i and ξ_i are observed, the complete log-likelihood from the joint model is

$$L_c = L_l + L_s + L_\alpha + L_\xi,$$

where

$$\begin{aligned}
L_l &= \textstyle\sum_{i=1}^n [\frac{n_i}{2}(\log\xi_i - \log\sigma_e^2) - \frac{\xi_i}{2\sigma_e^2}(y_i - X_i\beta - \tilde{X}_i\alpha_i)^T(y_i - X_i\beta - \tilde{X}_i\alpha_i)],\\
L_s &= \textstyle\sum_{i=1}^n [\delta_i(\log f_0(t_i) + \log\xi_i + \log\theta(Z_i)) - \xi_i\theta(Z_i)F_0(t_i)],\\
L_\alpha &= \textstyle\sum_{i=1}^n [\frac{1}{2}(\log\xi_i - \log\sigma_\alpha^2) - \frac{\xi_i\alpha_i^2}{2\sigma_\alpha^2}],\\
L_\xi &= \textstyle\sum_{i=1}^n [\eta\log\eta - \log\Gamma(\eta) + (\eta - 1)\log\xi_i - \eta\xi_i].
\end{aligned}$$

Denote the entire observed data as $O = \{y_i, t_i, \delta_i, X_i, \tilde{X}_i, Z_i\}$. Let Θ_k be the estimate of Θ in the kth iteration. The E-step in the $(k+1)$th iteration of the EM algorithm computes the conditional expectation of L_c with respect to α_i and ξ_i, which is equivalent to evaluating the following four conditional expectations,

$$\begin{aligned}
E[L_l|\Theta_k, O] &= \sum_{i=1}^{n} \{\tfrac{n_i}{2}(\tilde{a}_i - \log \sigma_e^2) - \tfrac{1}{2\sigma_e^2}[(y_i - X_i\beta)^T(y_i - X_i\beta)\tilde{r}_i - \\
&\qquad 2(y_i - X_i\beta)^T \tilde{X}_i \tilde{g}_i + \tilde{X}_i^T \tilde{X}_i \tilde{b}_i]\}; \\
E[L_s|\Theta_k, O] &= \sum_{i=1}^{n} \{\delta_i[\log f_0(t_i) + \tilde{a}_i + \log \theta(Z_i)] - \theta(Z_i)F_0(t_i)\tilde{r}_i\}; \\
E[L_\alpha|\Theta_k, O] &= \sum_{i=1}^{n} \{\tfrac{1}{2}[\tilde{a}_i - \log \sigma_\alpha^2] - \tfrac{\tilde{b}_i}{2\sigma_\alpha^2}\}; \\
E[L_\xi|\Theta_k, O] &= \sum_{i=1}^{n} [\eta \log \eta - \log \Gamma(\eta) + (\eta - 1)\tilde{a}_i - \eta \tilde{r}_i],
\end{aligned}$$

where $\tilde{a}_i$, $\tilde{b}_i$, $\tilde{g}_i$ and $\tilde{r}_i$ are given in the Appendix of Song et al. (2012). In the M-step of the EM algorithm, maximizing $E[L_l|\Theta_k, O]$, $E[L_\alpha|\Theta_k, O]$, and $E[L_\xi|\Theta_k, O]$ can be accomplished through the Newton-Raphson method. To update the parameters $(\gamma, F_0(t))$ in $E[L_s|\Theta_k, O]$, it is assumed that $F_0(t)$ is a proper cumulative distribution function and only increases at the observed event times, and maximizing $E[L_s|\Theta_k, O]$ can be carried out using the semiparametric method of Tsodikov et al. (2003). The maximum likelihood estimates of the parameters are obtained after iterating the E-step and M-step until convergence. To calculate the variance estimates for the parameter estimates, the method of Louis (1982) is employed to obtain the observed information matrix. Extensive simulations studies showed that the proposed joint model and estimation method are more efficient than the estimation methods based on separate models and also the joint model proposed by Zeng and Cai (2005).

Other researchers also proposed joint models to accommodate a possible cure fraction. Brown and Ibrahim (2003) considered a Bayesian method for a joint model that joins a promotion time cure model (8.2) for the survival data and a mixture model for the longitudinal data. Law et al. (2002) and Yu et al. (2008) proposed a joint model which employed another mixture cure model for the survival data and nonlinear mixed effect model for longitudinal data and employed a Bayesian method to estimate the parameters in the model. Chen et al. (2004) investigated the joint model of multivariate longitudinal measurements and survival data through a Bayesian method, where the promotion time cure model with a log-normal frailty is assumed for survival data. Due to complexity of the joint models, the Bayesian methods used in the models can be very tedious. In contrast, the estimation method proposed by Song et al. (2012) for model (8.1) and (8.2) is simple to use and takes less time to obtain estimates due to the simplicity of the conditional distributions of the latent variables in the model, which greatly reduces the complexity of estimation.

8.4 Modeling QoL and Survival Data with the Simplex Distribution

As mentioned in the introduction, although QoL measurements can be considered as continuous, their values are usually restricted in an interval and thus are bounded data. Therefore, classical linear mixed models with normal or t-distributions may not be proper models for the QoL measurements. Recently, we have explored the use of a model which assumes that the QoL measurements follow a simplex distribution for the longitudinal QoL measurements and developed estimation procedures for a joint model with simplex distribution based models for the longitudinal QoL measurements and Cox proportional hazards models for the survival times (Song et al. 2015). Specifically, given a random effect α_i, the density function of a simplex distribution for a longitudinal QoL measurements Y_{ij} has the following form (Barndorff-Nielsen and Jørgensen 1991):

$$f(y|\mu_{ij}, \sigma^2) = \begin{cases} a(y;\sigma^2)\exp[-\frac{d(y;\mu_{ij})}{2\sigma^2}], & y \in (0,1) \\ 0 & \text{otherwise} \end{cases}, \tag{8.3}$$

where $a(y;\sigma^2) = [2\pi\sigma^2(y(1-y))^3]^{-1/2}$, $d(y;\mu_{ij}) = \frac{(y-\mu_{ij})^2}{y(1-y)\mu_{ij}^2(1-\mu_{ij})^2}$, $0 < \mu_{ij} < 1$ is the location parameter and is equal to the mean of the distribution, and $\sigma^2 > 0$ is the dispersion parameter. The simplex distribution has a support on (0, 1) which make it suitable as a model for bounded data. Further details and properties of the simplex distribution can be found in Song and Tan (2008), Qiu et al. (2008), Jørgensen (1997), and Song (2007). We assume that the mean of Y_{ij}'s satisfies the following:

$$\text{logit}(\mu_{ij}) = X_{ij}^T\beta + \alpha_i, \tag{8.4}$$

where X_{ij} is a vector of covariates with coefficient β and α_i is a random effect which satisfies that

$$\alpha_1, \ldots, \alpha_n \sim^{i.i.d.} \varphi(\alpha|0, \sigma_\alpha^2). \tag{8.5}$$

This model is called a simplex model for Y_{ij}, which allows a directly modeling effects of X_{ij} on the mean of Y_{ij}, a desirable property that the existing joint modeling approaches for bounded longitudinal data do not have. The effect of X_{ij} on the mean, measured by β, can be easily interpreted in a similar way as the log odds ratio in the logistic regression.

For the survival time T_i', we assume it follows the proportional hazards model

$$h(t|Z_i, \alpha_i) = h_0(t)\exp(Z_i^T\gamma + \phi\alpha_i), \tag{8.6}$$

where γ is the coefficient of Z_i, ϕ is the coefficient of the random effect α_i, and $h_0(t)$ is an arbitrary unspecified baseline hazard function. The assumptions for α_i's are the same as in (8.5).

Similarly, the random effect α_i in the joint model (8.4) and (8.6) reflects the unobserved heterogeneity in the mean of Y_{ij} and the hazard of T_i' for different subject, and ϕ characterizes the correlation between the longitudinal bounded measurements and survival times, which can be seen from the joint density function of T_i' and Y_{ij}

$$f(t,y) = h_0(t)[2\pi\sigma^2(y(1-y))^3]^{-1/2}e^{Z^T\gamma} \times \int_{-\infty}^{+\infty} e^{\phi\alpha - H_0(t)e^{Z^T\gamma+\phi\alpha}} \exp\left[-\frac{(y-\frac{e^{X^T\beta+\alpha}}{1+e^{X^T\beta+\alpha}})^2}{2\sigma^2 y(1-y)\frac{e^{2(X^T\beta+\alpha)}}{(1+e^{X^T\beta+\alpha})^4}}\right]\varphi(\alpha)d\alpha.$$

When $\phi = 0$, the survival times and longitudinal measurements will be independent conditional on the observed covariates since $f(t,y) = f(t)\int_{-\infty}^{+\infty} f(y|\alpha)\varphi(\alpha)d\alpha = f(t)f(y)$. The above joint model (8.3), (8.4), (8.5), and (8.6) is referred as JMSIM. Since the simplex distribution is not in an exponential family, the joint model JMSIM is an extension of Viviani and Rizopoulos (2013) who assumed the distribution of the longitudinal response is in the exponential family and used the generalized linear mixed model together with the Cox model to construct the model.

The parameters in the joint model based on the simplex distribution, denoted as $\Theta = (\beta, \phi, \gamma, h_0(t), \sigma^2, \sigma_\alpha^2)$, can be estimated from the marginal log-likelihood of the proposed joint model:

$$l(\Theta) = \log\prod_{i=1}^{n}\int\left[\prod_{j=1}^{n_i} f(y_{ij}|\mu_{ij},\sigma^2)\right][h_0(t_i)e^{Z_i^T\gamma+\phi\alpha_i}]^{\delta_i}e^{-H_0(t_i)e^{Z_i^T\gamma+\phi\alpha_i}}\varphi(\alpha_i|0,\sigma_\alpha^2)d\alpha_i \tag{8.7}$$

where $H_0(t) = \int_0^t h_0(t)dt$. Since (8.7) involves an integration that does not have a closed form, it is difficult to maximize it directly. Instead, we considered the Laplace approximation to (8.7) and obtained the following first-order and the second-order penalized joint partial log-likelihoods (PJPL):

$$\widetilde{LL}_{PJPL}(\Theta') = -\frac{\sum_{i=1}^n n_i}{2}\log\sigma^2 - \frac{n}{2}\log\sigma_\alpha^2 + \sum_{i=1}^{n}\lambda_{PJPL}(\hat{\alpha}_i),$$

$$LL_{PJPL}(\Theta') = -\frac{\sum_{i=1}^n n_i}{2}\log\sigma^2 - \frac{n}{2}\log\sigma_\alpha^2 + \sum_{i=1}^{n}\lambda_{PJPL}(\hat{\alpha}_i) - \frac{1}{2}\sum_{i=1}^{n}\log|\lambda_{PJPL}^{(2)}(\hat{\alpha}_i)|,$$

where $\Theta' = (\beta, \phi, \gamma, \sigma^2, \sigma_\alpha^2)$,

$$\lambda_{PJPL}(\alpha_i) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n_i} d(y_{ij}; \mu_{ij}) + \delta_i[Z_i^T\gamma + \phi\alpha_i - \log(\sum_{k\in R(t_i)} e^{Z_k^T\gamma+\phi\alpha_k})] - \frac{\alpha_i^2}{2\sigma_\alpha^2},$$

$$\lambda^{(2)}_{PJPL}(\alpha_i) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n_i} \frac{\partial^2 d(y_{ij}; \mu_{ij})}{\partial {\alpha_i}^2} - \delta_i \frac{e^{Z_i^T\gamma+\phi_i\alpha_i}\phi^2}{\sum_{k\in R(t_i)} e^{Z_k^T\gamma+\phi\alpha_k}}(1 - \frac{e^{Z_i^T\gamma+\phi_i\alpha_i}}{\sum_{k\in R(t_i)} e^{Z_k^T\gamma+\phi\alpha_k}}) - \frac{1}{\sigma_\alpha^2}.$$

and $\hat{\alpha}_i = \text{argmax}_{\alpha_i}\lambda_{PJPL}(\alpha_i)$. Following Ripatti and Palmgren (2000), Ye et al. (2008), and Rondeau et al. (2003), omitting the complicated term $\log|\lambda^{(2)}(\hat{\alpha}_i)|$ in $LL_{PJPL}(\Theta')$ have a negligible effect on the parameter estimation but can make the computation faster. The parameter β, γ can be updated by maximizing $\widetilde{LL}_{PJPL}(\Theta')$, then $\sigma^2, \sigma_\alpha^2, \phi$ can be updated with current estimates of β, γ by maximizing $LL_{PJPL}(\Theta')$. The maximum likelihood estimate of Θ' is obtained by iterating these two steps until convergence, and $h_0(t)$ can be estimated by using Breslow's estimator (Breslow 1972). The variances of these parameter estimates are approximated from inverting the information matrices of $\widetilde{LL}_{PJPL}(\Theta')$ for parameters (β, γ) and of $LL_{PJPL}(\Theta')$ for parameters $(\sigma^2, \sigma_\alpha^2, \phi)$.

Another intuitive approach to deal with longitudinal bounded data is to perform logit transformation (Aitchison and Shen 1980; Lesaffre et al. 2007) of the bounded data before applying a joint model, such as Zeng and Cai (2005), to the data. Song et al. (2015) explored this approach in details. Numerical studies showed that the two approaches are comparable in performance and both are better than the simple approach that ignores the restrictive nature of the longitudinal bounded data, such as the method of Ye et al. (2008). The joint model based on logit-transformed longitudinal bounded data is more robust to model mis-specification. The approach based on simplex distribution has, however, an advantage in its simpler and easier interpretation of covariate effects on the original scale of the data. Similar to the classic generalized linear model, this approach allows the effects of covariates on the mean of the longitudinal bounded data while the dispersion of the distribution stays intact.

8.5 Application to a Data from a Breast Cancer Clinical Trial

As an illustration, the joint models and the estimation methods of JMtt and JMSIM are applied respectively to analyze the data from the MA.5 described in Sect.8.1. Preliminary examination of QoL data revealed that patterns of change in the scores are different in different time periods and in different treatment groups (Fig. 8.2). This lead to the piecewise polynomial linear predictors (8.8) and (8.9) to better interpret the change in scores over time of joint models JMtt and JMSIM,

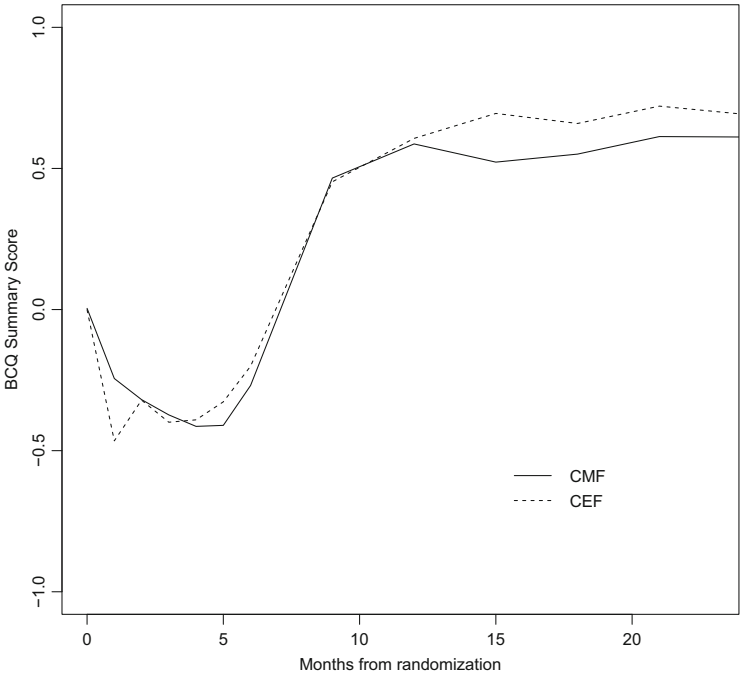

Fig. 8.2 Averages of BCQ scores for patients in the two treatment groups of MA.5

respectively:

$$\begin{aligned} JMtt : X_{ij}^T\beta + \alpha_i = &[\beta_0 + \beta_1 x_i + \beta_2 t_{ij} + \beta_3 x_i t_{ij} + \beta_4 t_{ij}^2 + \beta_5 x_i t_{ij}^2] I_{t_{ij}\in[0,2)} + \\ &[\beta_6 + \beta_7 x_i + \beta_8 t_{ij} + \beta_9 t_{ij}^2] I_{t_{ij}\in[2,9)} + \\ &[\beta_{10} + \beta_{11} x_i + \beta_{12} t_{ij}] I_{t_{ij}\geq 9} + \alpha_i, \end{aligned} \tag{8.8}$$

$$\begin{aligned} JMSIM : X_{ij}^T\beta + \alpha_i = &[\beta_0 + \beta_1 x_i + \beta_2 t_{ij} + \beta_3 x_i t_{ij} + \beta_4 t_{ij}^2 + \beta_5 x_i t_{ij}^2] I_{t_{ij}\in[0,2)} + \\ &[\beta_6 + \beta_7 x_i + \beta_8 t_{ij} + \beta_9 x_i t_{ij} + \beta_{10} t_{ij}^2 + \beta_{11} x_i t_{ij}^2] I_{t_{ij}\in[2,9)} + \\ &[\beta_{12} + \beta_{13} x_i + \beta_{14} t_{ij} + \beta_{15} x_i t_{ij}] I_{t_{ij}\geq 9} + \alpha_i, \end{aligned} \tag{8.9}$$

where x_i is the binary treatment indicator (=1 for CEF and =0 for CMF) and t_{ij} denotes the time(in month from the randomization) when the QoL of a patient was assessed. The RFS times in the joint model are modeled by (8.2) in JMtt and (8.6) in JMSIM with $Z_i = x_i$.

Table 8.1 Estimations, standard errors of parameters in joint models JMtt and JMSIM with longitudinal parts (8.8) and (8.9) for MA.5 respectively

Joint Model	Parameter	β_0	β_1	β_2	β_3	β_4	β_5
JMtt	Est.	0.0382	0.0679	−0.4426	−0.5360	0.1380	0.2662
	Std.	0.0306	0.0433	0.0597	0.0841	0.0309	0.0436
JMSIM	Est.	1.0516	0.0350	−0.2171	−0.4345	0.0393	0.1904
	Std.	0.0316	0.0447	0.0624	0.0900	0.0332	0.0469
Joint Model	Parameter	β_6	β_7	β_8	β_9	β_{10}	β_{11}
JMtt	Est.	−0.0278	0.0732	−0.2174	0.0297	0.3630	0.1192
	Std.	0.0553	0.0355	0.0207	0.0019	0.0356	0.0361
JMSIM	Est.	0.7431	−0.0152	−0.0311	−0.0105	0.0102	0.0016
	Std.	0.0787	0.1118	0.0316	0.0447	0.0028	0.0004
Joint Model	Parameter	β_{12}	β_{13}	β_{14}	β_{15}	γ_0	γ_1
JMtt	Est.	0.0062	–	–	–	−0.0454	−0.2863
	Std.	0.0015	–	–	–	0.0788	0.1163
JMSIM	Est.	1.4364	0.0110	0.0070	0.0012	–	−0.2679
	Std.	0.0397	0.0565	0.0019	0.0027	–	0.1122
Joint Model	Parameter	σ_α^2	$\eta(\phi)$				
JMtt	Est.	0.1499	3.6392				
	Std.	0.0099	0.2991				
JMSIM	Est.	0.1781	0.3444				
	Std.	0.0101	0.1425				

Table 8.1 summarizes the estimates of the parameters and their standard errors in the two joint models of JMtt and JMSIM fitted to the data respectively. The hazard of the RFS times was significantly lower for patients randomized to CEF than those to CMF(p-value 0.0138 from JMtt, p-value 0.0170 from JMSIM). In addition, both $\hat{\eta}$ in JMtt and $\hat{\phi}$ in JMSIM are significantly different from 0, which indicate a strong dependence between longitudinal QoL and RFS data from MA.5. The fitted mean curves and their confidence bands of the longitudinal data for CEF and CMF from JMtt and JMSIM are in Fig. 8.3. The plots fitted by JMtt and JMSIM have a similar trend. The mean BCQ score in the CMF group decreases to a nadir after randomization and them increases steadily over the next 7 months. In contrast, the mean BCQ score in the CEF group decreases more quickly to a nadir and gradually increases in the remaining months of chemotherapy treatment. After 6-month of the chemotherapy treatments, the scores of both arms tend to be stable. This implies that patients treated by CEF had worse QoL than those treated by CMF at very beginning but gradually recovered to a slightly better than those treated by CMF.

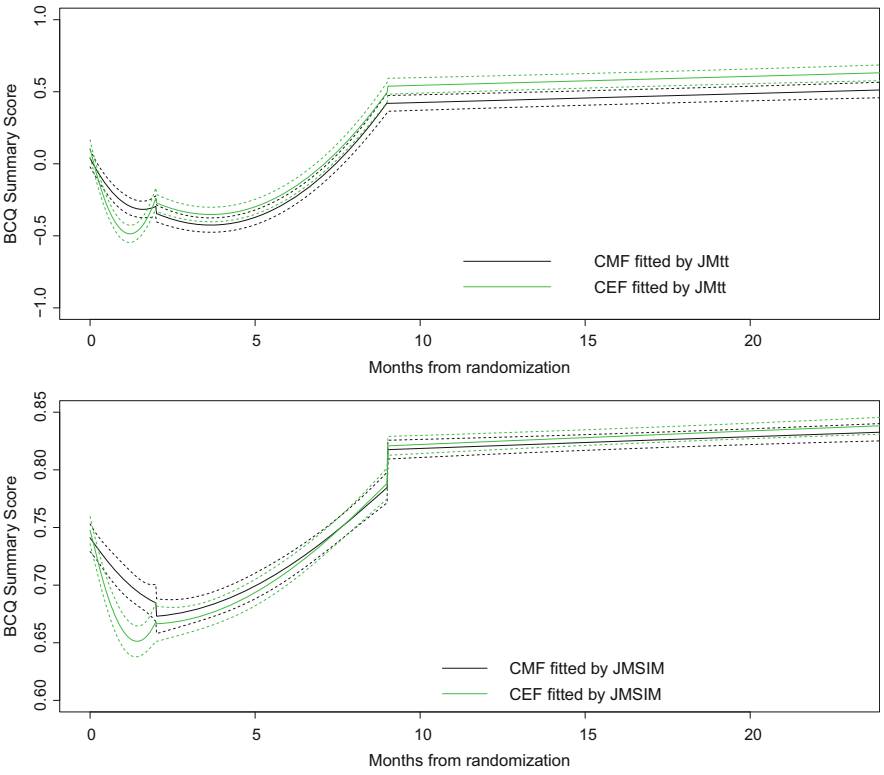

Fig. 8.3 Estimated mean curves and their confidence bands of longitudinal data from JMtt and JMSIM for CMF and CEF arms respectively

8.6 Discussions and Future Work

In this paper, we have reviewed our recent contributions in the joint modeling of longitudinal QoL and survival data to deal with bounded longitudinal QoL data and a possible cure fraction. Our work involves a linear mixed tt model and a generalized linear mixed effect model with simplex distribution for the longitudinal QoL data to better accommodate extreme and bounded QoL measurements. A promotion time cure model is considered for survival data to accommodate a possible cure fraction. Semiparametric inference procedures with an EM algorithm and a penalized joint likelihood based on the Laplace approximation were developed for the parameters in the joint models. The simulation studies show that the estimation of parameters in the joint models is more efficient than that in the separate analysis and the existing joint models. The models also enjoy intuitive interpretation. The models are illustrated with the data from a breast cancer clinical trial.

Our work is limited in the definition of new models and inference of parameters in these models. Recently, motivated also by the analysis of data from MA.5, Park

and Qiu (2014) developed statistical procedures for the selection of diagnostics for a joint model which uses a linear mixed model for longitudinal measurements and a time-varying coefficient model for the survival times. Procedures for the assessment of fit for each component of the joint model were derived recently by Zhang et al. (2014). He et al. (2015) developed some procedures which can be used to select simultaneously variables in both components of a joint model. Developments of similar procedures for the models we proposed would be an interesting but also challenging future research topic.

Other future extensions of the work reviewed in this paper include treating the QoL measurements as categorical data based on their original scales. An item response model (Wang et al. 2002) may be considered in a joint model to accommodate categorical QoL measurements. It is also an interesting topic to develop a smooth function of time for the longitudinal trajectory and the baseline hazard function via splines. Also the relapse free survival time is, by nature, interval-censored since assessment of recurrence is usually done at each clinical assessment. Extensions of the joint models reviewed to take into account of the interval censoring are of interest in the applications.

Acknowledgements The authors wish to thank the editors and associate editors for their helpful comments and suggestions. Hui Song was supported by National Natural Sciences Foundation of China (Grant No.11601060), Dalian High Level Talent Innovation Programme (No.2015R051) and Fundamental Research Funds for the Central Universities of China.

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London

Aitchison J, Shen SM (1980) Logistic-normal distributions: some properties and uses. Biometrika 67:261–272

Barndorff-Nielsen OE, Jørgensen B (1991) Some parametric models on the simplex. J Multivar Anal 39:106–116

Breslow NE (1972) Contribution to the discussion of D. R. Cox. J R Stat Soc Ser B 34:216–217

Brown ER, Ibrahim JG (2003) Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. Biometrics 59:686–693

Brundage M, Feld-Stewart D, Leis A, Bezjak A, Degner L, Velji K, Zetes-Zanatta L, Tu D, Ritvo P, Pater J (2005) Communicating quality of life information to cancer patients: a study of six presentation formats. J Clin Oncol 28:6949–6956

Chen MH, Ibrahim JG, Sinha D (1999) A new Bayesian model for survival data with a surviving fraction. J Am Stat Assoc 94:909–919

Chen MH, Ibrahim JG, Sinha D (2004) A new joint model for longitudinal and survival data with a cure fraction. J Multivar Anal 91:18–34

Cornish EA (1954) The multivariate t-distribution associated with a set of normal sample deviates. Aust J Phys 7:531–542

Dancey J, Zee B, Osoba D, Whitehead M, Lu F, Kaizer L, Latreille J, Pater JL (1997) Quality of life score: an independent prognostic variable in a general population of cancer patients receiving chemotherapy. Qualif Life Res 6:151–158

Diggle P, Sousa I, Chetwynd A (2008) Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. Stat Med 27:2981–2998

Fairclough DL (2010) Design and analysis of quality of life studies in clinical trials. Chapman and Hall/CRC, Boca Raton

Ganz PA, Lee JJ, Siau J (2006) Qualify of life assessment: an independent prognostic variable for survival in lung cancer. Cancer 67:3131–3135

Gould LA, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, Bois FY (2015), Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat Med 34:2181–2195

He Z, Tu W, Wang S, Fu H, Yu Z (2015) Simultaneous Variable Selection for Joint Models of Longitudinal and Survival Outcomes. Biometrics 71:178–187

Ibrahim JG, Chu H, Chen LM (2010) Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol 28:2796–2801

Jørgensen B (1997) The theory of dispersion models. Chapman and Hall, London

Klein JP (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 48:795–806

Lange KL, Little RJA, Taylor JMG (1989) Robust statistical modeling using the t distribution. J Am Stat Assoc 84:881–896

Law NJ, Taylor JMG, Sandler H (2002) The joint modelling of a longitudinal disease progression marker and the failure time process in the presence of cure. Biostatistics 3:547–563

Lesaffre E, Rizopoulos D, Tsonaka R (2007) The logistic transform for bounded outcome scores. Biostatistics 8:72–85

Levine MN, Bramwell VH, Pritchard KI, Norris BD, Shepherd LE, Abu-Zahra H, Findlay B, Warr D, Bowman D, Myles J, Arnold A, Vandenberg T, MacKenzie R, Robert J, Ottaway J, Burnell M, Williams CK, Tu DS (1998) Randomized trial of intensive cyclophosphamide, epirubicin, and fluorouracil chemotherapy compared with cyclophosphamide, methotrexate, and fluorouracil in premenopausal women with node-positive breast caner. J Clin Oncol 16:2651–2658

Levine MN, Pritchard KI, Bramwell VH, Shepherd LE, Tu D, Paul N (2005) Randomized trial comparing cyclophosphamide, epirubicin, and fluorouracil with cyclophosphamide, methotrexate, and fluorouracil in premenopausal women with node-positive breast cancer: Update of national cancer institute of canada clinical trials group trial MA.5. J Clin Oncol 23:5166–5170

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc Ser B 44:226–233

Park KY, Qiu P (2014) Model selection and diagnostics for joint modeling Of survival and longitudinal data with crossing hazard rate functions. Stat Med 33: 4532–4546

Pinheiro JC, Liu CH, Wu YN (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. J Comput Graph Stat 10:249–276

Qiu Z, Song PXK, Tan M (2008) Simplex mixed-effects models for longitudinal proportional data. Scand J Stat 35:577–596. http://ideas.repec.org/a/bla/scjsta/v35y2008i4p577-596.html

Richards MA, Ramirez AJ (1997) Quality of life: the main outcome measure of palliative care. Palliat Med 11:89–92

Ripatti S, Palmgren J (2000) Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56:1016–1022

Rondeau V, Commenges D, Joly P (2003) Maximum penalized likelihood estimation in a gamma-frailty model. Lifetime Data Anal 9:139–153

Schluchter MD (1992) Methods for the analysis of informatively censored longitudinal data. Stat Med 11:1861–1870

Song PXK (2007) Correlated data analysis: modeling, analytics, and applications. Springer, New York

Song PXK, Tan M (2000) Marginal models for longitudinal continuous proportional data. Biometrics 56:496–502

Song PXK, Zhang P, Qu A (2007) Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. Stat Sin 17:929–943

Song H, Peng YW, Tu DS (2012) A new approach for joint modelling of longitudinal measurements and survival times with a cure fration. Can J Stat 40:207–224

Song H, Peng YW, Tu DS (2015, Online) Jointly modeling longitudinal proportional data and survival times with an application to the quality of life data in a breast cancer trial. Lifetime Data Anal. doi: 10.1007/s10985-015-9346-8

Tsiatis AA, Davidian M (2004) Joint modelling of longitudinal and time-to-event data: an overview. Stat Sin 14:809–834

Tsodikov A, Ibrahim JG, Yakovlev AY (2003) Estimating cure rates from survival data: an alternative to two-component mixture mdoels. J Am Stat Assoc 98:1063–1078

Tu DS, Chen YQ, Song PX (2004) Analysis of quality of life data from a clinical trial on early breast cancer based on a non-parametric global test for repeated measures with informative censoring. Qual Life Res 13:1520–1530

Viviani S, Alfó M, Rizopoulos D (2013) Generalized linear mixed joint model for longitudinal and survival outcomes. Stat Comput 24:417–427

Wang C, Douglas J, Anderson S (2002) Item response models for joint analysis of quality of life and survival. Stat Med 21:129–142

Wu L, Liu W, Yi GY, Huang Y (2012) Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. J Probab Stat Article ID 640153. doi:10.1155/2012/640153

Yakovlev AH, Asselain B, Bardou VJ, Fourquet A, Hoang T, Rochefordiere A, Tsodikov AD (1993) A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. Biometriete Analysse de Donnees Spatio-Temporelles 12:66–82

Ye W, Lin XH, Taylor JMG (2008) A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. Stat Interface 1:33–45

Yin GS (2005) Bayesian cure rate frailty models with application to a root canal therapy study. Biometrics 61:552–558

Yu MG, Taylor JMG, Sandler HM (2008) Individual prediction in prostate cancer studies using a joint longitudinal survival-cure model. J Am Stat Assoc 103:178–187

Zeng D, Cai J (2005) Simultaneous modelling of survival and longitudinal data with an application to repeated quality of life measures. Lifetime Data Anal 11:151–174

Zhang P, Qiu ZG, Fu YJ, Song PXK (2009) Robust transformation mixed-effects models for longitudinal continuous proportional data. Can J Stat 37:266–281

Zhang D, Chen MH, Ibrahim JG, Boye ME, Wang P, Shen W (2014) Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. Stat Med 33:4715–4733

Part III
Applied Data Analysis

Chapter 9
Confidence Weighting Procedures for Multiple-Choice Tests

Michael Cavers and Joseph Ling

Abstract Multiple-choice tests are extensively used in the testing of mathematics and statistics in undergraduate courses. This paper discusses a confidence weighting model of multiple choice testing called the student-weighted model. In this model, students are asked to indicate an answer choice and their certainty of its correctness. This method was implemented in two first year Calculus courses at the University of Calgary in 2014 and 2015. The results of this implementation are discussed here.

9.1 Introduction

Multiple-choice exams are extensively used to test students' knowledge in mathematics and statistics. Typically a student has four or five options to select from with exactly one option being correct. Using multiple-choice exams allow the examiners to cover a broad range of topics, to score the exam quickly, and to measure various learning outcomes. However, it is difficult to write questions that test mathematical ideas instead of factual recall. Shank (2010) is an excellent resource discussing the construction of good multiple-choice questions. It is also debated as to whether multiple-choice exams truly measure a students' knowledge since in the conventional model there is no penalty for guessing (Echternacht 1972).

This paper discusses an implementation of a confidence weighting procedure known as the *student-weighted model*. In this model, each multiple-choice question has four or five options to select from with exactly one correct answer. The examinee must also indicate their certainty of the correctness of their answers, for example on a three-point scale. We first briefly discuss past studies where confidence testing was applied.

In 1932, Hevner applied a confidence testing method to music appreciation true/false tests. In 1936, Soderquist completed a similar study with true/false tests, whereas Wiley and Trimble (1936) analyzed whether or not there are personality factors in the confidence testing method. A five-point scale for confidence levels was implemented by Gritten and Johnson in 1941.

M. Cavers (✉) • J. Ling
Department of Mathematics and Statistics, University of Calgary, T2N 1N4, Calgary, AB, Canada
e-mail: mcavers@ucalgary.ca;; jling@ucalgary.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_9

A particularly interesting study, completed in 1953 by Dressel and Schmid, compares four non-conventional multiple-choice models. Five sections of a course in physical science were used in the study where each section contained about 90 students. In the first hour of the exam, each section wrote a common (conventional) multiple-choice test. In the second hour, five multiple-choice models were implemented, one for each of the five sections: a *conventional* multiple-choice test with one correct answer from five options; a *free-choice* test where each question had exactly one correct answer and examinees must mark as many answers as they felt they needed to be sure they had not omitted the correct answer; a *degree-of-certainty* test where each question had exactly one correct answer and examinees indicate how certain they are of their answer as being correct; a *multiple-answer* test where any number of the options may be correct and the examinee is to mark each correct alternative; and a *two-answer* test where exactly two of the five options is known to be correct. After comparing the above models, it was found that students completed the conventional test the fastest, followed by degree-of-certainty, free-choice, two-answer, and multiple-answer. From most to least reliable the models rank as multiple-answer, two-answer, degree-of-certainty, conventional followed by free-choice.

Relevant studies to confidence weighting methods, including those described above, are summarized by Echternacht (1972). In Echternacht (1972), it is stated that such methods can be used to discourage guessing since the expected score is maximized only if the examinee reveals the true degree of certainty in their responses. Frary (1989) reviews methods used in the 1970s and 1980s.

Ling and Cavers (2015) describe overall results of an early implementation of the student-weighted model completed in 2014. The current paper compares results from Ling and Cavers (2015) to our Spring 2015 study. This study is approved by the Conjoint Faculties Research Ethics Board (CFREB) at the University of Calgary.

In Sect. 9.2 we discuss the implementation of the student-weighted model at the University of Calgary in 2014. We then discuss differences in how the method was applied in 2015. Survey results and the effects of the method on class performance are analyzed. In Sect. 9.3 we conclude with issues for future investigation.

9.2 The Student-Weighted Model

The student-weighted model was implemented on the multiple-choice sections of the midterms and final examination in two first year Calculus courses at the University of Calgary in 2014 and 2015. In particular, we applied the method in four semesters: Winter 2014 (one section of Math 249 and one section of Math 251), Spring 2014 (Math 249), Fall 2014 (five sections of Math 249), and Spring 2015 (one section of Math 249). This paper mainly focuses on the data collected from the Fall 2014 and Spring 2015 semesters.

In what follows, we first describe the method and how scores are calculated. Second, we discuss the effect relative weights had on class average along with percentage of beneficial and detrimental cases. Third, student feedback from surveys conducted to solicit student feedback about the method are analyzed.

9.2.1 Description of the Method

Out of five options, each question had exactly one correct answer. After each question, the examinee was asked to assign a relative weight to the question (i.e., a confidence level) on a three-point scale. In an early implementation of the model in the Winter 2014 semester, students indicated their answers on a custom answer sheet. They were instructed to place an X in the box for their choice (options A, B, C, D, E) along with an X in one of three relative weight boxes for that question (options 1, 2, 3). Sample answer sheets with tallied scores were distributed in advance to explain the scoring method to the students. Later implementations made use of the standard university multiple-choice answer sheets where the odd numbered items were exam questions and the even numbered items were the relative weights. A sample exam question followed by a request for students to indicate a relative weight is shown in Fig. 9.1.

Students were told to assign weights to questions based on their confidence level for each problem. That is, when they felt very confident, "Relative Weight = 3" should be assigned, whereas, when they did not feel confident, "Relative Weight = 1" should be assigned. When the weighting part of the question is left blank, a default weight of 2 was assigned. To calculate the multiple-choice score for each

Question 1. Suppose that $y = f(x)$ is everywhere continuous and that for all x,

$$f(x) = 1 + 2\int_0^x f(t)\,dt.$$

What is the value of $f(1)$?

(a) 1.
(b) 2.
(c) e.
(d) e^2.
(e) There is insufficient information for us to determine the value of $f(1)$.

Question 2. Please assign a (relative) weight to **Question 1** above.

(a) Weight = 1.
(b) Weight = 2.
(c) Weight = 3.

Fig. 9.1 A sample exam question and follow-up asking examinees to indicate a relative weight

student, the total sum of the relative weights for all correct answers was divided by the total sum of the relative weights assigned to all questions. Thus, in such a method, each student may have a different total sum of relative weights. These scores were then multiplied by the portion of the test that is multiple-choice: in the case the exam had a 30 % written component and 70 % multiple-choice component, by 70; in the case the exam was all multiple-choice, by 100. Students were instructed they could default to a conventional multiple-choice test by assigning the same relative weight to each question, or by leaving the relative weights blank. Students who scored 100 % on the multiple-choice did not see a benefit in grade from the method since a perfect score would have been obtained regardless of the assignment of relative weights.

9.2.2 Effect on Class Average and Student Grades

Using actual data from midterms and examinations we are able to compare students' scores between a conventional (uniform-weight) scoring method and the student-weighted model. Here, we present the data collected from the Fall 2014 and Spring 2015 semesters. In the Fall 2014 semester, we implemented the method in five sections of Math 249 (Introductory Calculus) for the two midterms and the final examination. A total of 723 students wrote the first midterm, 657 students wrote the second midterm and 603 students wrote the final examination. In the Spring 2015 semester, the method was implemented in one section of Math 249 for both the midterm and final examination. A total of 87 students wrote the midterm and 81 students wrote the final examination. Note that at the University of Calgary, the Fall and Winter semesters last 13 weeks each whereas the Spring and Summer semesters range from 6 to 7 weeks each, thus, one midterm was given in the Spring 2015 semester where two midterms were given in the Fall 2014 semester.

To measure the effect of relative weights on student grades, for each student, we first computed their grade assuming a conventional scoring method where all questions are weighted the same, we then compared this to the student-weighted model using the relative weights assigned by the student. Below we will present three ways of looking at the impacts of relative weights on student grades:

- The change in the class average.
- The percentage of students that received a higher mark, a lower mark and the same mark on tests and exams.
- The percentage of students that experienced a substantial difference (5 percentage points or more and 3 percentage points or more) in test and exam marks.

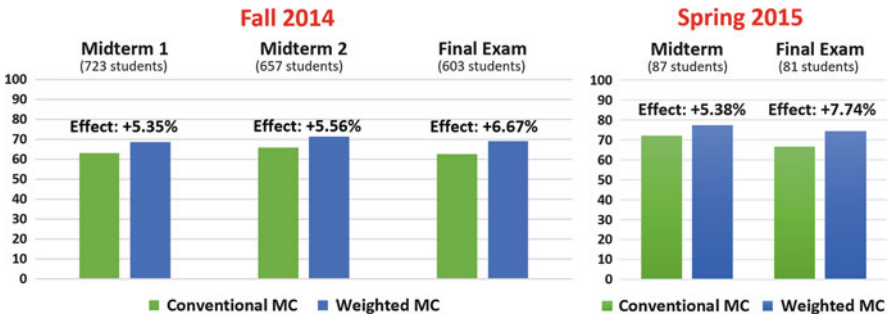

Fig. 9.2 Effect of relative weights on class average

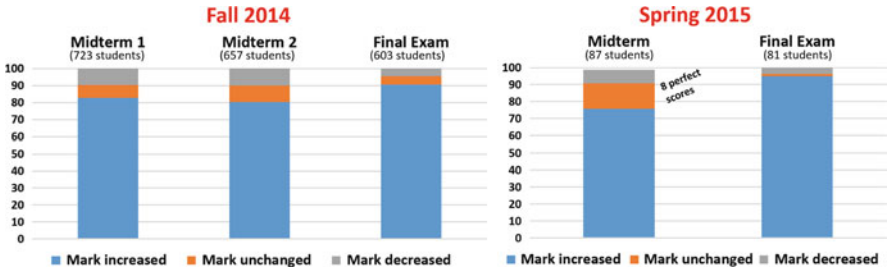

Fig. 9.3 Percentage of students whose midterm/final examination mark increased, remained unchanged or decreased after applying the student-weighted scoring method

Figure 9.2 shows the effect that relative weights had on the class average. These results indicate that course performance as a whole is better when we use the student-weighted scoring method compared to using the conventional scoring method.

Figure 9.3 shows that a majority of students had an increase in their midterm or final examination mark after applying the student-weighted model. In the Spring 2015 semester, eight students scored 100 % on the midterm and thus their mark is indicated as unchanged when compared to an equal weight scheme. Additionally, some students either left the relative weight options blank or assigned each question the same relative weight, thus, their score is also unchanged when comparing the relative weight method to the conventional. Note that some students who saw their mark decrease come from top performers who had one or two incorrect responses but assigned a high relative weight to those particular problems.

In Figs. 9.4 and 9.5, we further analyze the effect that relative weights had on student grades. Students whose multiple-choice mark is affected by at least five percentage points would likely receive a different letter grade than if the conventional scoring method were used. Here, we have separated the "beneficial" cases from the "detrimental" cases.

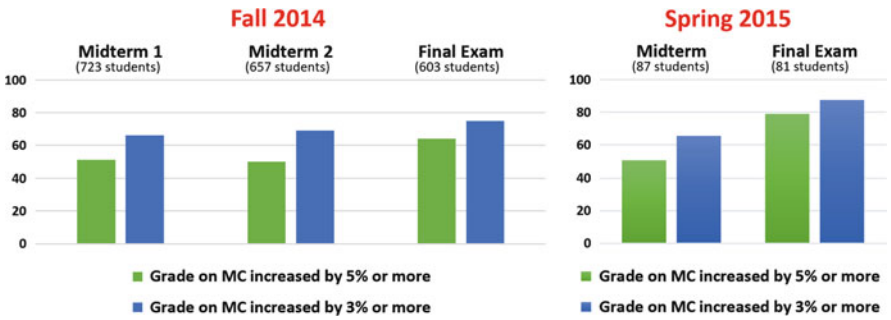

Fig. 9.4 Percentage of beneficial cases

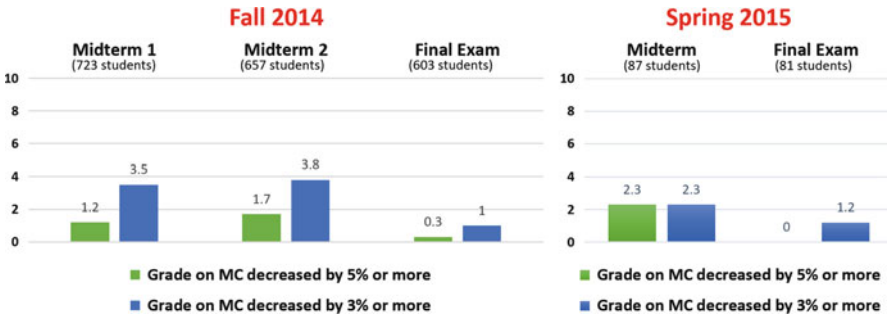

Fig. 9.5 Percentage of detrimental cases

9.2.3 Survey Results

During each course, a survey was conducted to solicit student feedback about the student-weighted method. The main purpose of each survey is for us to learn about the students' perspectives on their experience with the student-weighted method. We use student comments to help us improve both our teaching and future students' learning experience. See Ling and Cavers (2015) for comments on survey questions and responses from the Winter 2014, Spring 2014 and Fall 2014 semesters. Here, we summarize results from the Spring 2015 semester and compare it to that of the Fall 2014 semester.

In the Fall 2014 semester, one survey was conducted about two weeks after the second midterm test, but response rate was low: 73 out of 646 students completed the survey. Questions were asked to measure students' perception on learning and course performance. Note that Figs. 9.2 and 9.3 illustrate the actual effect of the method. Because of the low response rate, we hesitate to draw any definite conclusions about the general student perceptions on the scoring method.

In the Spring 2015 semester, we incorporated the use of the student-weighted format in the tutorials/labs. All lab problem sets were laid out in this format, thus practice with the method is built into the entire course. In the Fall 2014 semester,

a multi-page explanatory document was posted for students to read, but such a document was not posted in the Spring 2015 semester. Bonus effort marks for submitting individual or group solutions to the lab problem sets were awarded.

In the Spring 2015 semester, the number of surveys was increased from one to two, and the number of survey questions from 6 to 30 for the first survey and 31 for the second survey. By conducting two surveys, we are better able to track possible change of feeling/experience during the semester. In the Spring 2015 semester we also asked about impacts on study habits and exam strategies, impact on stress level, and perception of fairness of the method.

Completion of the two surveys was built into the course grade, in particular, 1 % of the course mark was for completion of the surveys. As a result, the response rate to the surveys increased when compared to the Fall 2014 semester: 72 out of 87 students (82.76 %) for survey one and 65 out of 81 students (80.25 %) for survey two. This gives us a lot more concrete information to help improve our future teaching. However, valid consent for research use of the data is low: twelve students gave consent to present their responses for both surveys. Three students consented for use of their data in one of the surveys but not the other, while another three students consented but did not complete one (or both) of the surveys. We caution the reader that information shared here is from the twelve consented respondents who completed both surveys and may or may not be representative of all respondents.

In what follows, we highlight, report and reflect on a few survey questions from consented participants. We use a vector

$$(SA, A, N, D, SD)$$

to indicate the number of students that selected Strongly Agree (*SA*), Agree (*A*), Neutral (*N*), Disagree (*D*) and Strongly Disagree (*SD*), respectively, and a vector

$$(VCon, SCon, N, SClear, VClear)$$

to indicate the number of students that selected Very Confusing (*VCon*), Somewhat Confusing (*SCon*), Neither (*N*), Somewhat Clear (*SClear*) and Very Clear (*VClear*), respectively.

The first question asked on both surveys was "*Which multiple-choice model, traditional or weighted, do you prefer?*". On survey one, one student selected the "Traditional method", ten students selected the "Weighted method", zero students selected "Neither" and one student selected "No opinion". Interestingly, on survey two all twelve students chose the "Weighted method".

The second question asked was "*How would you rate your understanding of the weighting method?*" for survey one and "*How would you rate your understanding of the weighting method after the midterm test?*" for survey two. We observe

$$(VCon, SCon, N, SClear, VClear) = (0, 1, 0, 6, 5)$$

for survey one and

$$(VCon, SCon, N, SClear, VClear) = (0, 1, 0, 5, 6)$$

for survey two. The student who chose "Somewhat Confusing" in survey one, chose "Somewhat Clear" in survey two, while the student who chose "Somewhat Confusing"in survey two, chose "Somewhat Clear" in survey one. Overall, most of the students understood the weighting.

The third question asked was "*How would you rate the calculation of your grade using the weighting method?*" for survey one and "*How would you rate the calculation of your grade using the weighting method after the midterm test?*" for survey two. We observe

$$(VCon, SCon, N, SClear, VClear) = (0, 2, 1, 7, 2)$$

for survey one and

$$(VCon, SCon, N, SClear, VClear) = (0, 1, 2, 4, 5)$$

for survey two. Five students selected the same option on both surveys while five students chose options indicating the calculation of grade became clearer after the midterm test. The remaining two students chose "Somewhat Clear" on survey one but on survey two, one chose "Neither" while the other chose "Somewhat Confusing".

The fourth item on the surveys was "*I paid attention to how I assigned weight when I did the questions in the lab problem sets*" for survey one and " *Since the midterm test, I have paid attention to how I assigned weight when I did the questions in the lab problem sets*" for survey two. We observe $(SA, A, N, D, SD) = (4, 6, 0, 1, 1)$ for survey one and $(SA, A, N, D, SD) = (1, 4, 6, 0, 1)$ for survey two. Overall, most of the students paid more attention to weight assignment in the labs before the midterm while still learning about the method.

The next set of questions asked students about double-checking their work. For the question "*How often do you typically double-check your work on midterms (respectively, examinations)?*", we observe

$$(Always, Very\ Frequently, Occasionally, Rarely, Never) = (7, 3, 0, 2, 0)$$

for survey one and

$$(Always, Very\ Frequently, Occasionally, Rarely, Never) = (5, 3, 4, 0, 0)$$

for survey two. For the statements "*While working on the problem sets, I felt an increased need to double-check my work as I was to assign weights to my answers*" for survey one and "*Since the midterm test, I felt an increased need to double-check my work as I was to assign weights to my answers while working on the lab problem*

sets" for survey two, we observe $(SA, A, N, D, SD) = (1, 2, 5, 4, 0)$ for survey one and $(SA, A, N, D, SD) = (5, 2, 4, 1, 0)$ for survey two. Finally, for the question "*During the midterm test (respectively, final examination), I felt an increased need to double-check my work as I was to assign weights to my answers*" we observe $(SA, A, N, D, SD) = (4, 5, 2, 1, 0)$ for survey one and $(SA, A, N, D, SD) = (8, 4, 0, 0, 0)$ for survey two. Overall, although students already frequently double-check their work, students perceived the assignment of weights to cause an increase in need to double-check their work.

The next question asked "*During the midterm test (respectively, final examination), I experienced an increased level of stress as I was to assign weights to my answers*" with a follow-up question requesting students to provide a reason for their choice. We observe $(SA, A, N, D, SD) = (2, 4, 2, 2, 2)$ for survey one and $(SA, A, N, D, SD) = (1, 4, 3, 3, 1)$ for survey two. The most common reasons given for those who selected (SA) or (A) are that the stake is high and weighting impacts grade. For those who selected (SD) or (D) the most common reason given is that the method can lower weight when unsure.

We also asked students their perception of the weighted method on learning: "*Assigning relative weights to multiple-choice questions was beneficial to my learning.*" We observe $(SA, A, N, D, SD) = (3, 7, 1, 1, 0)$ for survey one and $(SA, A, N, D, SD) = (6, 5, 1, 0, 0)$ for survey two. After the final exam, it appears that an increased number found the method beneficial. When asked how did they think the method was beneficial to their learning, the most common response was that it helped identify weak areas for more practice and study. We also asked students the following: "*My test mark based on the weighting method is a more accurate reflection of my knowledge of the course material than the mark based on the traditional grading method would have been.*" We observe $(SA, A, N, D, SD) = (2, 5, 4, 1, 0)$ on survey one, however, there was a mistake in two of the options on survey two so we omit that data here.

The next question asked "*Prior to writing the midterm test (respectively, final exam) I developed my own strategy to assign weights to my answers.*" We observe $(SA, A, N, D, SD) = (2, 2, 4, 2, 2)$ for survey one and $(SA, A, N, D, SD) = (2, 7, 2, 1, 0)$ for survey two. From this it appears that more students developed a strategy for the final exam than in midterm. When asked if they used the same strategy we found $(SA, A, N, D, SD) = (4, 4, 0, 4, 0)$.

The next question asked "*I personally believe that there is a positive correlation between one's knowledge of a subject and one's confidence in one's knowledge of that subject*" with a follow-up question requesting students to provide a reason for their choice. We observe $(SA, A, N, D, SD) = (9, 3, 0, 0, 0)$ for survey one and $(SA, A, N, D, SD) = (9, 2, 0, 0, 1)$ for survey two. Note that the student with the (SD) response in survey two chose (SA) in survey one for the same question. Reasons are respectively "*If you are confident you know something, you're more likely to have it stick to your brain instead of just learning it for the exam*" and "*If you think you don't know it, you probably don't.*"

When asked "*Knowing that I could assign weights affected how I studied,*" we observe $(SA, A, N, D, SD) = (0, 5, 2, 4, 1)$ for survey one and $(SA, A, N, D, SD) = (1, 4, 4, 2, 1)$ for survey two. Reasons given for (SA) and (A) are that students focus more on content they weight a one and that if there was a topic they did not understand they did not stress as much about it knowing they can weight it a one. Reasons given for (SD) and (D) mostly mention that they would study all of the material regardless and that they need to know all concepts for future courses. On the other hand, when asked "*Knowing that I could assign weights affected how I distrubuted (sic) my study time to different topics,*" we observe $(SA, A, N, D, SD) = (1, 4, 2, 4, 1)$ for survey one and $(SA, A, N, D, SD) = (2, 4, 2, 3, 1)$ for survey two. Comments given for (SA) and (A) are that they focused more time on topics they were not confident in, whereas comments for (D) and (SD) are that the weighting method does not affect grade enough to completely ignore a topic.

To conclude, consented respondents found the method beneficial to their learning, mainly by helping them to identify areas of weakness; developed strategies as the term progressed; believed that there is a positive correlation between knowledge and confidence; thought that the weighting method reflected their level of knowledge more accurately than the traditional method; and seemed to be neutral in relation to the impact on stress level in tests and exam.

9.3 Issues for Future Investigation

After implementation of the student-weighted method, feedback from students and colleagues have sparked many questions, for example, see Ling and Cavers (2015). Results indicate that course performance as a whole is better when we use the student-weighted scoring method compared to that of the conventional scoring method. However, how is one to interpret this in terms of student learning? More research targeted at specific issues is needed.

References

Dressel PL, Schmid J (1953) Some modifications of the multiple-choice item. Educ Psychol Measurement 13:574–595

Echternacht GJ (1972) The use of confidence testing in objective tests. Rev Educ Res 42(2):217–236

Frary RB (1989) Partial-credit scoring methods for multiple-choice tests. Appl Measur Educ 2(1):79–96

Gritten F, Johnson DM (1941) Individual differences in judging multiple-choice questions. J Educ Psychol 32:423–430

Hevner KA (1932) A method of correcting for guessing in true-false tests and empirical evidence in support of it. J Soc Psychol 3:359–362

Ling J, Cavers M (2015) Student-weighted multiple choice tests. In: Proceedings of the 2015 UC postsecondary conference on learning and teaching, PRISM: University of Calgary Digital Repository

Shank P (2010) Create better multiple-choice questions, vol 27, Issue 1009. The American Society for Training & Development, Alexandria

Soderquist HO (1936) A new method of weighting scores in a true-false test. J Educ Res 30: 290–292

Wiley LN, Trimble OC (1936) The ordinary objective test as a possible criterion of certain personality traits. Sch Soc 43:446–448

Chapter 10
Improving the Robustness of Parametric Imputation

Peisong Han

Abstract Parametric imputation is widely used in missing data analysis. When the imputation model is misspecified, estimators based on parametric imputation are usually inconsistent. In this case, we propose to estimate and subtract off the asymptotic bias to obtain consistent estimators. Estimation of the bias involves modeling the missingness mechanism, and we allow multiple models for it. Our method simultaneously accommodates these models. The resulting estimator is consistent if any one of the missingness mechanism models or the imputation model is correctly specified.

10.1 Introduction

Imputation is a widely used method in missing data analysis, where the missing values are filled in by imputed values and the analysis is done as if the data were completely observed. Parametric imputation (Little and Rubin 2002; Rubin 1978, 1987), which imputes the missing values based on a parametric model, is the most commonly taken form in practice, due to its simplicity and straightforwardness. However, parametric imputation is sensitive to misspecification of the imputation model. The resulting estimators are usually inconsistent when the model is misspecified. Because of this sensitivity, many researchers suggested nonparametric imputation; for example, Cheng (1994), Lipsitz et al. (1998), Aerts et al. (2002), Wang and Rao (2002), Zhou et al. (2008) and Wang and Chen (2009). Despite the robustness against model misspecification, nonparametric imputation usually suffers from the curse of dimensionality. In addition, for kernel-based techniques, bandwidth selection could be a complicated problem. In this paper, within the framework of parametric imputation, we propose a method to improve the robustness against possible model misspecifications. The idea is to estimate and subtract off the asymptotic bias of the imputation estimators when the imputation model is misspecified. Estimation of the bias involves modeling the missingness

P. Han (✉)
Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave. W., N2L 3G1, Waterloo, ON, Canada
e-mail: peisonghan@uwaterloo.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_10

mechanism, and we allow multiple models for it. Our method simultaneously accommodates these models. The resulting estimator is consistent if any one of these models or the imputation model is correctly specified. A detailed numerical study of the proposed method is presented.

10.2 Notations and Existing Methods

We consider the setting of estimating the population mean of an outcome of interest that is subject to missingness. This simple yet important setting has been studied in many recent works on missing data analysis, including Tan (2006, 2010), Kang and Schafer (2007) and its discussion, Qin and Zhang (2007), Qin et al. (2008), Cao et al. (2009), Rotnitzky et al. (2012), Han and Wang (2013), Chan and Yam (2014), and Han (2014a).

Let Y denote an outcome that is subject to missingness, and $\mathbf{X}$ a vector of auxiliary variables that are always observed. Our goal is to estimate $\mu_0 = E(Y)$, the marginal mean of Y. Let R denote the indicator of observing Y; that is, $R = 1$ if Y is observed and $R = 0$ if Y is missing. Our observed data are $(R_i, R_i Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, which are independent and identically distributed. We assume the missingness mechanism to be missing-at-random (MAR) (Little and Rubin 2002) in the sense that

$$P(R = 1|Y, \mathbf{X}) = P(R = 1|\mathbf{X}),$$

and we use $\pi(\mathbf{X})$ to denote this probability. Under the MAR assumption, the sample average based on complete cases, namely $n^{-1} \sum_{i=1}^{n} R_i Y_i$, is not a consistent estimator of μ_0.

Parametric imputation postulates a parametric model $a(\boldsymbol{\gamma}; \mathbf{X})$ for $E(Y|\mathbf{X})$, and imputes the missing Y_i by $a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i)$, where $\hat{\boldsymbol{\gamma}}$ is some estimated value of $\boldsymbol{\gamma}$. One typical way of calculating $\hat{\boldsymbol{\gamma}}$ is the complete-case analysis, as $E(Y|\mathbf{X}) = E(Y|\mathbf{X}, R = 1)$ under the MAR mechanism. When $a(\boldsymbol{\gamma}; \mathbf{X})$ is a correct model for $E(Y|\mathbf{X})$, in the sense that $a(\boldsymbol{\gamma}_0; \mathbf{X}) = E(Y|\mathbf{X})$ for some $\boldsymbol{\gamma}_0$, the two imputation estimators of μ_0,

$$\hat{\mu}_{\text{imp},1} = \frac{1}{n} \sum_{i=1}^{n} a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i),$$

$$\hat{\mu}_{\text{imp},2} = \frac{1}{n} \sum_{i=1}^{n} \{R_i Y_i + (1 - R_i) a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i)\},$$

are both consistent. When $a(\boldsymbol{\gamma}; \mathbf{X})$ is a misspecified model, we have $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_*$ for some $\boldsymbol{\gamma}_* \neq \boldsymbol{\gamma}_0$. In this case neither $\hat{\mu}_{\text{imp},1}$ nor $\hat{\mu}_{\text{imp},2}$ is consistent. Their probability limits are $E\{a(\boldsymbol{\gamma}_*; \mathbf{X})\}$ and $E\{RY + (1 - R)a(\boldsymbol{\gamma}_*; \mathbf{X})\}$, respectively, and

their asymptotic biases are

$$bias_1 = E\{a(\boldsymbol{\gamma}_*; \mathbf{X}) - Y\}, \tag{10.1}$$

$$bias_2 = E[\{1 - \pi(\mathbf{X})\}\{a(\boldsymbol{\gamma}_*; \mathbf{X}) - Y\}]. \tag{10.2}$$

Thus, if a consistent estimator of $bias_1$ (or $bias_2$) can be found, subtracting this bias estimator off from $\hat{\mu}_{\text{imp},1}$ (or $\hat{\mu}_{\text{imp},2}$) leads to a consistent estimator of μ_0.

Noticing that

$$bias_1 = E\left[\frac{R}{\pi(\mathbf{X})}\{a(\boldsymbol{\gamma}_*; \mathbf{X}) - Y\}\right],$$

$$bias_2 = E\left[\frac{R}{\pi(\mathbf{X})}\{1 - \pi(\mathbf{X})\}\{a(\boldsymbol{\gamma}_*; \mathbf{X}) - Y\}\right]$$

under the MAR mechanism, one straightforward way to obtain a consistent estimator of the asymptotic bias is to model $\pi(\mathbf{X})$. Let $\pi(\boldsymbol{\alpha}; \mathbf{X})$ denote a parametric model for $\pi(\mathbf{X})$, and $\hat{\boldsymbol{\alpha}}$ the maximizer of the Binomial likelihood

$$\prod_{i=1}^{n} \{\pi(\boldsymbol{\alpha}; \mathbf{X}_i)\}^{R_i} \{1 - \pi(\boldsymbol{\alpha}; \mathbf{X}_i)\}^{1-R_i}. \tag{10.3}$$

The two biases in (10.1) and (10.2) can be respectively estimated by

$$\widetilde{bias}_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \mathbf{X}_i)}\{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\}\right],$$

$$\widetilde{bias}_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \mathbf{X}_i)}\{1 - \pi(\hat{\boldsymbol{\alpha}}; \mathbf{X}_i)\}\{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\}\right].$$

It is easy to see that, when $\pi(\boldsymbol{\alpha}; \mathbf{X})$ is a correctly specified model in the sense that $\pi(\boldsymbol{\alpha}_0; \mathbf{X}) = \pi(\mathbf{X})$ for some $\boldsymbol{\alpha}_0$, $\widetilde{bias}_1 \xrightarrow{p} bias_1$ and $\widetilde{bias}_2 \xrightarrow{p} bias_2$, and thus both $\hat{\mu}_{\text{imp},1} - \widetilde{bias}_1$ and $\hat{\mu}_{\text{imp},2} - \widetilde{bias}_2$ are consistent estimators of μ_0. On the other hand, when $a(\boldsymbol{\gamma}; \mathbf{X})$ is a correctly specified model for $E(Y|\mathbf{X})$, we have

$$\begin{aligned}\widetilde{bias}_1 &\xrightarrow{p} E\left[\frac{R}{\pi(\boldsymbol{\alpha}_*; \mathbf{X})}\{a(\boldsymbol{\gamma}_0; \mathbf{X}) - Y\}\right] \\ &= E\left[\frac{E(R|\mathbf{X})}{\pi(\boldsymbol{\alpha}_*; \mathbf{X})}\{a(\boldsymbol{\gamma}_0; \mathbf{X}) - E(Y|\mathbf{X})\}\right] = 0,\end{aligned}$$

where $\boldsymbol{\alpha}_*$ is the probability limit of $\hat{\boldsymbol{\alpha}}$ and may not be equal to $\boldsymbol{\alpha}_0$, and the second last equality follows from $R \perp Y \mid \mathbf{X}$ under the MAR mechanism. Similarly, we have $\widetilde{bias}_2 \xrightarrow{p} 0$. So $\hat{\mu}_{\text{imp},1} - \widetilde{bias}_1$ and $\hat{\mu}_{\text{imp},2} - \widetilde{bias}_2$ are again consistent estimators

of μ_0. Therefore, $\hat{\mu}_{\text{imp},1} - \widetilde{bias}_1$ and $\hat{\mu}_{\text{imp},2} - \widetilde{bias}_2$ are more robust than $\hat{\mu}_{\text{imp},1}$ and $\hat{\mu}_{\text{imp},2}$ against possible misspecification of $a(\boldsymbol{\gamma}; \mathbf{X})$.

Actually, $\hat{\mu}_{\text{imp},1} - \widetilde{bias}_1$ and $\hat{\mu}_{\text{imp},2} - \widetilde{bias}_2$ are both equal to

$$\hat{\mu}_{\text{aipw}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi(\hat{\boldsymbol{\alpha}}; \mathbf{X}_i)} Y_i - \frac{R_i - \pi(\hat{\boldsymbol{\alpha}}; \mathbf{X}_i)}{\pi(\hat{\boldsymbol{\alpha}}; \mathbf{X}_i)} a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) \right\},$$

which is the augmented inverse probability weighted (AIPW) estimator (Robins et al. 1994). The fact that $\hat{\mu}_{\text{aipw}} \xrightarrow{p} \mu_0$ if either $\pi(\mathbf{X})$ or $E(Y|\mathbf{X})$ is correctly modeled is known as the double robustness property (Scharfstein et al. 1999; Tsiatis 2006). The improvement of robustness is achieved by introducing an extra model $\pi(\boldsymbol{\alpha}; \mathbf{X})$ in addition to $a(\boldsymbol{\gamma}; \mathbf{X})$.

10.3 The Proposed Method

In observational studies, the correct model for $\pi(\mathbf{X})$ is typically unknown. To increase the likelihood of correct specification, multiple models $\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})$, $j = 1, \ldots, J$, may be fitted instead of just one. Refer to Robins et al. (2007) for more discussion on some practical scenarios where multiple models may be fitted. Our goal is to propose a method that can simultaneously accommodates all these models, so that $bias_1$ and $bias_2$ in (10.1) and (10.2) are consistently estimated if one of $\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})$ is correct or $a(\boldsymbol{\gamma}; \mathbf{X})$ is correct.

Because of

$$\begin{aligned} 0 &= E\left(\frac{R}{\pi(\mathbf{X})} [\pi^j(\boldsymbol{\alpha}^j; \mathbf{X}) - E\{\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})\}] \right) \\ &= E\left(\frac{1}{\pi(\mathbf{X})} [\pi^j(\boldsymbol{\alpha}^j; \mathbf{X}) - E\{\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})\}] \mid R = 1 \right) P(R = 1), \end{aligned}$$

$\pi^j(\boldsymbol{\alpha}^j; \mathbf{X}), j = 1, \ldots, J$, satisfy

$$E\left(\frac{1}{\pi(\mathbf{X})} [\pi^j(\boldsymbol{\alpha}^j; \mathbf{X}) - E\{\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})\}] \mid R = 1 \right) = 0. \tag{10.4}$$

Let $m = \sum_{i=1}^{n} R_i$ be the number of subjects who have their outcome observed, and index those subjects by $i = 1, \ldots, m$ without loss of generality. Let $\hat{\boldsymbol{\alpha}}^j$ denote the maximizer of the Binomial likelihood (10.3) with $\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})$. We construct the

empirical version of (10.4) as

$$w_i \geq 0 \quad (i = 1, \ldots, m), \quad \sum_{i=1}^{m} w_i = 1,$$

$$\sum_{i=1}^{m} w_i \{\pi^j(\hat{\boldsymbol{\alpha}}^j; \mathbf{X}_i) - \hat{\theta}^j(\hat{\boldsymbol{\alpha}}^j)\} = 0 \quad (j = 1, \ldots, J), \tag{10.5}$$

where w_i, $i = 1, \ldots, m$, are positive weights on the complete cases under sum-to-one regularization, and $\hat{\theta}^j(\hat{\boldsymbol{\alpha}}^j) = n^{-1} \sum_{i=1}^{n} \pi^j(\hat{\boldsymbol{\alpha}}^j; \mathbf{X}_i)$. The w_i naturally accommodate all models $\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})$.

Being positive and sum-to-one, w_i may be viewed as an empirical likelihood on the complete cases. Applying the principle of maximum likelihood, we maximize $\prod_{i=1}^{m} w_i$ subject to the constraints in (10.5). Write $\hat{\boldsymbol{\alpha}}^{\mathrm{T}} = \{(\hat{\boldsymbol{\alpha}}^1)^{\mathrm{T}}, \ldots, (\hat{\boldsymbol{\alpha}}^J)^{\mathrm{T}}\}$ and

$$\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})^{\mathrm{T}} = \{\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i) - \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1), \ldots, \pi^J(\hat{\boldsymbol{\alpha}}^J; \mathbf{X}_i) - \hat{\theta}^J(\hat{\boldsymbol{\alpha}}^J)\}.$$

From empirical likelihood theory (Owen 1988, 2001; Qin and Lawless 1994) the maximizer is given by

$$\hat{w}_i = \frac{1}{m} \frac{1}{1 + \hat{\boldsymbol{\rho}}^{\mathrm{T}} \hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})} \quad (i = 1, \cdots, m), \tag{10.6}$$

where $\hat{\boldsymbol{\rho}}^{\mathrm{T}} = (\hat{\rho}_1, \cdots, \hat{\rho}_J)$ is the J-dimensional Lagrange multiplier solving

$$\frac{1}{m} \sum_{i=1}^{m} \frac{\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})}{1 + \boldsymbol{\rho}^{\mathrm{T}} \hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})} = \mathbf{0}. \tag{10.7}$$

We propose to estimate $bias_1$ and $bias_2$ in (10.1) and (10.2) respectively by

$$\widehat{bias}_1 = \sum_{i=1}^{m} \hat{w}_i \{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\},$$

$$\widehat{bias}_2 = \sum_{i=1}^{m} (\hat{w}_i - n^{-1}) \{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\}.$$

To discuss the large sample properties of $\widehat{bias}_1$ and $\widehat{bias}_2$, let $\boldsymbol{\alpha}_*^j$, θ_*^j and $\boldsymbol{\rho}_*$ denote the probability limits of $\hat{\boldsymbol{\alpha}}^j$, $\hat{\theta}^j(\hat{\boldsymbol{\alpha}}^j)$ and $\hat{\boldsymbol{\rho}}$, respectively. It is clear that $\theta_*^j = E\{\pi^j(\boldsymbol{\alpha}_*^j; \mathbf{X})\}$. Write $\boldsymbol{\alpha}_*^{\mathrm{T}} = \{(\boldsymbol{\alpha}_*^1)^{\mathrm{T}}, \cdots, (\boldsymbol{\alpha}_*^J)^{\mathrm{T}}\}$ and

$$\mathbf{g}(\boldsymbol{\alpha}_*)^{\mathrm{T}} = \{\pi^1(\boldsymbol{\alpha}_*^1; \mathbf{X}) - \theta_*^1, \ldots, \pi^J(\boldsymbol{\alpha}_*^J; \mathbf{X}) - \theta_*^J\}. \tag{10.8}$$

When $a(\boldsymbol{\gamma};\mathbf{X})$ is a correct model for $E(Y|\mathbf{X})$, we have $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$. Therefore,

$$\begin{aligned}
\widehat{bias}_1 &= \sum_{i=1}^{m} \hat{w}_i\{a(\hat{\boldsymbol{\gamma}};\mathbf{X}_i) - Y_i\} \\
&= \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})}\{a(\hat{\boldsymbol{\gamma}};\mathbf{X}_i) - Y_i\}\right] \\
&\xrightarrow{p} \frac{1}{P(R=1)}E\left[\frac{R}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\mathbf{g}(\boldsymbol{\alpha}_*)}\{a(\boldsymbol{\gamma}_0;\mathbf{X}) - Y\}\right] \\
&= \frac{1}{P(R=1)}E\left[\frac{E(R|\mathbf{X})}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\mathbf{g}(\boldsymbol{\alpha}_*)}\{a(\boldsymbol{\gamma}_0;\mathbf{X}) - E(Y|\mathbf{X})\}\right] = 0,
\end{aligned}$$

where the second last equality follows from $R \perp Y \mid \mathbf{X}$ under the MAR mechanism. Similarly,

$$\begin{aligned}
\widehat{bias}_2 &= \sum_{i=1}^{m}(\hat{w}_i - n^{-1})\{a(\hat{\boldsymbol{\gamma}};\mathbf{X}_i) - Y_i\} \\
&= \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n}\left[R_i\{\frac{1}{1+\hat{\boldsymbol{\rho}}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})} - \frac{m}{n}\}\{a(\hat{\boldsymbol{\gamma}};\mathbf{X}_i) - Y_i\}\right] \\
&\xrightarrow{p} \frac{1}{P(R=1)}E\left[R\{\frac{1}{1+\boldsymbol{\rho}_*^{\mathrm{T}}\mathbf{g}(\boldsymbol{\alpha}_*)} - P(R=1)\}\{a(\boldsymbol{\gamma}_0;\mathbf{X}) - Y\}\right] = 0.
\end{aligned}$$

Hence, both $\hat{\mu}_{\mathrm{imp},1} - \widehat{bias}_1$ and $\hat{\mu}_{\mathrm{imp},2} - \widehat{bias}_2$ are consistent estimators of μ_0 when $a(\boldsymbol{\gamma};\mathbf{X})$ is a correct model for $E(Y|\mathbf{X})$.

In the following, we will show that $\widehat{bias}_1$ and $\widehat{bias}_2$ are consistent estimators of $bias_1$ and $bias_2$, respectively, when any one of $\pi^j(\boldsymbol{\alpha}^j;\mathbf{X})$ is a correct model for $\pi(\mathbf{X})$. Without loss of generality, let the correct model be $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$. It is easy to see that

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})}{1+\boldsymbol{\rho}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})} \\
&= \frac{\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)}{m}\sum_{i=1}^{m}\frac{\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})}{\pi^1(\hat{\boldsymbol{\alpha}}^1;\mathbf{X}_i) + \left\{\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\rho_1 - 1, \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\rho_2, \cdots, \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\rho_J\right\}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})} \\
&= \frac{\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)}{m}\sum_{i=1}^{m}\frac{\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1;\mathbf{X}_i)}{1 + \left\{\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\rho_1 - 1, \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\rho_2, \cdots, \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\rho_J\right\}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1;\mathbf{X}_i)}.
\end{aligned}$$

Because $\hat{\boldsymbol{\rho}}$ solves (10.7), if we define $\hat{\lambda}_1 = \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\hat{\rho}_1 - 1$ and $\hat{\lambda}_t = \hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)\hat{\rho}_t$, $t = 2, \ldots, J$, then $\hat{\boldsymbol{\lambda}}^{\mathrm{T}} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_J)$ solves

$$\frac{\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)}{m} \sum_{i=1}^{m} \frac{\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)} = \mathbf{0}. \tag{10.9}$$

In terms of $\hat{\boldsymbol{\lambda}}$, $\hat{w}_i$ given by (10.6) can be re-expressed as

$$\hat{w}_i = \frac{1}{m} \frac{\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)}.$$

Since $\pi^1(\boldsymbol{\alpha}^1; \mathbf{X})$ is a correct model for $\pi(\mathbf{X})$, we have $\boldsymbol{\alpha}^1_* = \boldsymbol{\alpha}^1_0$, and thus $\boldsymbol{\lambda} = \mathbf{0}$ is the solution to

$$E\left\{ \frac{R\mathbf{g}(\boldsymbol{\alpha}_*)/\pi^1(\boldsymbol{\alpha}^1_*; \mathbf{X})}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\alpha}_*)/\pi^1(\boldsymbol{\alpha}^1_*; \mathbf{X})} \right\} = \mathbf{0},$$

where $\mathbf{g}(\boldsymbol{\alpha}_*)$ is given by (10.8). In addition, the left-hand side of (10.9) converges in probability to the left-hand side of the above equation. Therefore, $\hat{\boldsymbol{\lambda}}$ solving (10.9) has probability limit $\mathbf{0}$ and is of order $O_p(n^{-1/2})$ from the M-estimator theory (e.g., van der Varrt 1998). With this fact, we have

$$\begin{aligned}
\widehat{bias}_1 &= \sum_{i=1}^{m} \hat{w}_i \{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\} \\
&= \frac{n\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)}{m} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{R_i/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)} \{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\} \right] \\
&\xrightarrow{p} E\left[\frac{R}{\pi(\mathbf{X})} \{a(\boldsymbol{\gamma}_*; \mathbf{X}) - Y\} \right] = bias_1
\end{aligned}$$

and

$$\begin{aligned}
\widehat{bias}_2 &= \sum_{i=1}^{m} (\hat{w}_i - n^{-1}) \{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\} \\
&= \frac{n\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)}{m} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{R_i}{\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)} \left\{ \frac{1}{1 + \hat{\boldsymbol{\lambda}}^{\mathrm{T}} \hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})/\pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i)} \right. \right. \\
&\qquad \left. \left. - \frac{m}{n\hat{\theta}^1(\hat{\boldsymbol{\alpha}}^1)} \pi^1(\hat{\boldsymbol{\alpha}}^1; \mathbf{X}_i) \right\} \{a(\hat{\boldsymbol{\gamma}}; \mathbf{X}_i) - Y_i\} \right] \\
&\xrightarrow{p} E\left[\frac{R}{\pi(\mathbf{X})} \{1 - \pi(\mathbf{X})\} \{a(\boldsymbol{\gamma}_*; \mathbf{X}) - Y\} \right] = bias_2.
\end{aligned}$$

Thus, $\widehat{bias}_1$ and $\widehat{bias}_2$ are consistent estimators of $bias_1$ and $bias_2$, respectively, which makes $\hat{\mu}_{\text{imp},1} - \widehat{bias}_1$ and $\hat{\mu}_{\text{imp},2} - \widehat{bias}_2$ consistent estimators of μ_0.

As a matter of fact, simple algebra shows that $\hat{\mu}_{\text{imp},1} - \widehat{bias}_1$ is equal to $\hat{\mu}_{\text{imp},2} - \widehat{bias}_2$. Let $\hat{\mu}_{\text{mr}}$ denote this difference, where "mr" stands for multiple robustness. Based on our arguments above, $\hat{\mu}_{\text{mr}}$ is a consistent estimator of μ_0 if any one of $\pi^j(\boldsymbol{\alpha}^j; \mathbf{X})$, $j = 1, \ldots, J$, is a correct model for $\pi(\mathbf{X})$ or $a(\boldsymbol{\gamma}; \mathbf{X})$ is a correct model for $E(Y|\mathbf{X})$. Therefore, $\hat{\mu}_{\text{mr}}$ improves the robustness over the imputation estimators and the AIPW estimator. The asymptotic distribution of $\hat{\mu}_{\text{mr}}$ depends on which model is correctly specified, but such information is usually unavailable in real studies. This makes the asymptotic distribution to be of little practical use for inference. Hence, we choose not to derive the asymptotic distribution, but rather recommend the bootstrapping method to calculate the standard error of $\hat{\mu}_{\text{mr}}$. Numerical performance of the bootstrapping method will be evaluated in the next section.

A main step in the numerical implementation of the proposed method is to calculate $\hat{\boldsymbol{\rho}}$. Since $\hat{\boldsymbol{\rho}}$ solves (10.7) and $\hat{w}_i$ are positive, $\hat{\boldsymbol{\rho}}$ is actually the minimizer of

$$\mathbf{F}_n(\boldsymbol{\rho}) = -\frac{1}{n}\sum_{i=1}^{n} R_i \log\{1 + \boldsymbol{\rho}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}})\}$$

over the region $\mathscr{D}_n = \{\boldsymbol{\rho} : \ 1 + \boldsymbol{\rho}^{\mathrm{T}}\hat{\mathbf{g}}_i(\hat{\boldsymbol{\alpha}}) > 0, \ i = 1, \ldots, m\}$. Han (2014a) showed that the minimizer of $\mathbf{F}_n(\boldsymbol{\rho})$ over $\mathscr{D}_n$ indeed exists, and $\hat{\boldsymbol{\rho}}$ is the unique and global minimizer, at least when n is large. On the other hand, it is easy to verify that $\mathscr{D}_n$ is an open convex set and $\mathbf{F}_n(\boldsymbol{\rho})$ is a strictly convex function. Therefore, calculating $\hat{\boldsymbol{\rho}}$ pertains to a convex minimization problem. Refer to Chen et al. (2002) and Han (2014a) for a detailed description of the numerical implementation using Newton–Raphson algorithm.

10.4 Simulation Study

Our simulation setting follows that in Kang and Schafer (2007). The data are generated with $\mathbf{X} = \{X^{(1)}, \ldots, X^{(4)}\} \sim N(0, \mathbf{I}_4)$, $Y|\mathbf{X} \sim N\{E(Y|\mathbf{X}), 1\}$, and $R|\mathbf{X} \sim \text{Ber}\{\pi(\mathbf{X})\}$, where $\mathbf{I}_4$ is the 4×4 identity matrix,

$$\pi(\mathbf{X}) = [1 + \exp\{X^{(1)} - 0.5X^{(2)} + 0.25X^{(3)} + 0.1X^{(4)}\}]^{-1},$$
$$E(Y|\mathbf{X}) = 210 + 27.4X^{(1)} + 13.7\{X^{(2)} + \cdots + X^{(4)}\}.$$

The $\pi(\mathbf{X})$ leads to approximately 50 % of the subjects with missing Y. As in Kang and Schafer (2007), we calculate $Z^{(1)} = \exp\{X^{(1)}/2\}$, $Z^{(2)} = X^{(2)}/[1 + \exp\{X^{(1)}\}] + 10$, $Z^{(3)} = \{X^{(1)}X^{(3)}/25 + 0.6\}^3$ and $Z^{(4)} = \{X^{(2)} + X^{(4)} + 20\}^2$. The correct models

for $\pi(\mathbf{X})$ and $E(Y|\mathbf{X})$ are given by

$$\pi^1(\boldsymbol{\alpha}^1;\mathbf{X}) = [1+\exp\{\alpha_1^1+\alpha_2^1X^{(1)}+\cdots+\alpha_5^1X^{(4)}\}]^{-1},$$
$$a^1(\boldsymbol{\gamma}^1;\mathbf{X}) = \gamma_1^1+\gamma_2^1X^{(1)}+\cdots+\gamma_5^1X^{(4)},$$

respectively. The incorrect models are fitted by $\pi^2(\boldsymbol{\alpha}^2;\mathbf{Z})$ and $a^2(\boldsymbol{\gamma}^2;\mathbf{Z})$, which replace $\mathbf{X}$ in $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$ and $a^1(\boldsymbol{\gamma}^1;\mathbf{X})$ by $\mathbf{Z} = \{Z^{(1)},\ldots,Z^{(4)}\}$. As in Robins et al. (2007) and Rotnitzky et al. (2012), we also consider the scenario where Y is observed when $R = 0$ instead of when $R = 1$. The estimators under our comparison include the inverse probability weighted (IPW) (Horvitz and Thompson 1952) estimator

$$\hat{\mu}_{\text{ipw}} = \frac{\sum_{i=1}^n R_iY_i/\pi(\boldsymbol{\alpha};\mathbf{X}_i)}{\sum_{i=1}^n R_i/\pi(\boldsymbol{\alpha};\mathbf{X}_i)},$$

the imputation estimators $\hat{\mu}_{\text{imp},1}$ and $\hat{\mu}_{\text{imp},2}$, the AIPW estimator $\hat{\mu}_{\text{aipw}}$, and the estimator $\hat{\mu}_{\text{rlsr}}$ proposed by Rotnitzky et al. (2012). We use a four-digit superscript to distinguish estimators constructed using different postulated models, with each digit, from left to right, indicating if $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$, $\pi^2(\boldsymbol{\alpha}^2;\mathbf{Z})$, $a^1(\boldsymbol{\gamma}^1;\mathbf{X})$ and $a^2(\boldsymbol{\gamma}^2;\mathbf{Z})$ is used, respectively. We take the sample sizes $n = 200,800$ and conduct 2000 replications for the simulation study. The results are summarized in Table 10.1.

Both imputation estimators $\hat{\mu}_{\text{imp},1}$ and $\hat{\mu}_{\text{imp},2}$ have large bias when $E(Y|\mathbf{X})$ is incorrectly modeled by $a^2(\boldsymbol{\gamma}^2;\mathbf{Z})$, especially in the second scenario where Y is observed if $R = 0$. Using the correct model $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$ for $\pi(\mathbf{X})$, our proposed estimators $\hat{\mu}_{\text{mr}}^{1001}$ and $\hat{\mu}_{\text{mr}}^{1101}$ are able to significantly reduce the bias. While the existing estimators $\hat{\mu}_{\text{aipw}}^{1001}$ and $\hat{\mu}_{\text{rlsr}}^{1001}$ have similar capability, they explicitly require to know which model for $\pi(\mathbf{X})$ is correct. Our estimator $\hat{\mu}_{\text{mr}}^{1101}$, on the contrary, accommodates both $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$ and $\pi^2(\boldsymbol{\alpha}^2;\mathbf{Z})$ to reduce the bias without requiring such knowledge. This is important in practice, as it is usually impossible to tell which one among the multiple fitted models is correct. The estimator $\hat{\mu}_{\text{mr}}^{1101}$ also has high efficiency, illustrated by its significantly smaller root mean square error compared to the AIPW estimator $\hat{\mu}_{\text{aipw}}^{1001}$. It is well known that this AIPW estimator using an incorrect model for $E(Y|\mathbf{X})$ could be very inefficient (e.g., Tan 2006, 2010; Cao et al. 2009), and the estimator $\hat{\mu}_{\text{rlsr}}^{1001}$ was proposed to improve the efficiency over $\hat{\mu}_{\text{aipw}}^{1001}$. Our estimator $\hat{\mu}_{\text{mr}}^{1101}$ has efficiency even comparable to $\hat{\mu}_{\text{rlsr}}^{1001}$, judging by their root mean square errors.

When only the incorrect model $\pi^2(\boldsymbol{\alpha}^2;\mathbf{Z})$ is used in addition to $a^2(\boldsymbol{\gamma}^2;\mathbf{Z})$, our estimator $\hat{\mu}_{\text{mr}}^{0101}$ is not consistent, the same as $\hat{\mu}_{\text{aipw}}^{0101}$ and $\hat{\mu}_{\text{rlsr}}^{0101}$. In this case, similar to $\hat{\mu}_{\text{rlsr}}^{0101}$, $\hat{\mu}_{\text{mr}}^{0101}$ has much more stable numerical performance than $\hat{\mu}_{\text{aipw}}^{0101}$ in the first scenario where Y is observed if $R = 1$. Here the poor performance of $\hat{\mu}_{\text{aipw}}^{0101}$ is because that, some $\pi^2(\hat{\boldsymbol{\alpha}}^2;\mathbf{Z}_i)$ for a few subjects with $R_i = 1$ are erroneously close to zero, yielding extremely large weights $R_i/\pi^2(\hat{\boldsymbol{\alpha}}^2;\mathbf{Z}_i)$ (Robins et al. 2007).

Table 10.1 Comparison of different estimators based on 2000 replications. Each digit of the four-digit superscript, from left to right, indicates if $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$, $\pi^2(\boldsymbol{\alpha}^2;\mathbf{Z})$, $a^1(\boldsymbol{\gamma}^1;\mathbf{X})$ and $a^2(\boldsymbol{\gamma}^2;\mathbf{Z})$ is used, respectively. The numbers have been multiplied by 100

	Y observed if and only if $R=1$						Y observed if and only if $R=0$					
	$n=200$			$n=800$			$n=200$			$n=800$		
	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
$\hat{\mu}_{\text{ipw}}^{1000}$	−9	388	241	5	202	126	15	427	253	8	195	125
$\hat{\mu}_{\text{ipw}}^{0100}$	154	863	315	505	1252	265	381	509	387	371	407	373
$\hat{\mu}_{\text{imp,1}}^{0010}$	4	261	180	−1	127	84	4	261	179	−1	127	85
$\hat{\mu}_{\text{imp,1}}^{0001}$	−52	338	227	−81	184	131	497	584	500	496	518	496
$\hat{\mu}_{\text{imp,2}}^{0010}$	4	261	180	−1	127	84	4	261	179	−1	127	85
$\hat{\mu}_{\text{imp,2}}^{0001}$	−52	338	227	−81	184	131	497	584	500	496	518	496
$\hat{\mu}_{\text{aipw}}^{1010}$	4	261	179	−1	127	84	4	261	179	−1	127	85
$\hat{\mu}_{\text{aipw}}^{1001}$	35	358	233	4	190	116	43	432	252	16	192	120
$\hat{\mu}_{\text{aipw}}^{0110}$	3	261	179	−10	456	86	4	261	180	−1	127	85
$\hat{\mu}_{\text{aipw}}^{0101}$	−693	5688	361	−2397	37088	510	326	462	350	308	348	311
$\hat{\mu}_{\text{rlsr}}^{1010}$	4	261	179	−1	127	84	4	261	179	−1	127	85
$\hat{\mu}_{\text{rlsr}}^{1001}$	31	297	202	7	137	91	116	312	210	34	141	94
$\hat{\mu}_{\text{rlsr}}^{0110}$	3	262	178	−2	129	84	4	261	178	−1	127	86
$\hat{\mu}_{\text{rlsr}}^{0101}$	−170	356	244	−237	345	261	254	396	287	157	221	163
$\hat{\mu}_{\text{mr}}^{1001}$	49	344	233	12	170	113	62	308	214	18	150	97
$\hat{\mu}_{\text{mr}}^{0101}$	−244	417	295	−314	432	314	300	435	328	277	318	278
$\hat{\mu}_{\text{mr}}^{1101}$	8	309	214	−1	156	107	85	308	211	27	148	94
$\hat{\mu}_{\text{mr}}^{1010}$	4	261	179	−1	127	84	4	261	179	−1	127	85
$\hat{\mu}_{\text{mr}}^{0110}$	4	261	179	−1	127	85	4	261	179	−1	127	85
$\hat{\mu}_{\text{mr}}^{1110}$	4	261	180	−1	127	84	4	261	179	−1	127	85

RMSE root mean square error, *MAE* median absolute error

This also explains the problematic performance of the corresponding IPW estimator $\hat{\mu}_{\text{ipw}}^{0100}$. Our estimator $\hat{\mu}_{\text{mr}}^{0101}$ is not affected much by the close-to-zero $\pi^2(\hat{\boldsymbol{\alpha}}^2;\mathbf{Z}_i)$ because it uses weights $\hat{w}_i$ that maximize $\prod_{i=1}^m w_i$. The maximization prevents the occurrence of extreme weights for our proposed method.

When the correct model $a^1(\boldsymbol{\gamma}^1;\mathbf{X})$ for $E(Y|\mathbf{X})$ is used, the proposed estimators $\hat{\mu}_{\text{mr}}^{1010}$, $\hat{\mu}_{\text{mr}}^{0110}$ and $\hat{\mu}_{\text{mr}}^{1110}$ have almost identical performance to the imputation estimators $\hat{\mu}_{\text{imp,1}}^{0010}$ and $\hat{\mu}_{\text{imp,2}}^{0010}$.

Table 10.2 summarizes the performance of the bootstrapping method in calculating the standard error of the proposed estimator. The re-sampling size is 200. The means of bootstrapping-based standard errors over the 2000 replications are close the corresponding empirical standard errors. In addition, except for the case where the proposed estimator is inconsistent (i.e. $\hat{\mu}_{\text{mr}}^{0101}$), the 95 % bootstrapping-based confidence intervals have coverage probabilities very close to 95 %. These observations demonstrate that the bootstrapping method provides a reliable way to make statistical inference.

Table 10.2 Bootstrapping method for the calculation of standard errors based on 2000 replications with re-sampling size 200. Each digit of the four-digit superscript, from left to right, indicates if $\pi^1(\boldsymbol{\alpha}^1;\mathbf{X})$, $\pi^2(\boldsymbol{\alpha}^2;\mathbf{Z})$, $a^1(\boldsymbol{\gamma}^1;\mathbf{X})$ and $a^2(\boldsymbol{\gamma}^2;\mathbf{Z})$ is used, respectively. Except for the percentages, the numbers have been multiplied by 100

	Y observed if and only if $R=1$						Y observed if and only if $R=0$					
	$n=200$			$n=800$			$n=200$			$n=800$		
	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER	EMP	EST	PER
$\hat{\mu}_{\text{mr}}^{1001}$	338	326	94.3 %	171	167	94.4 %	295	285	93.3 %	148	146	93.5 %
$\hat{\mu}_{\text{mr}}^{0101}$	346	334	87.4 %	226	176	54.3 %	303	298	82.3 %	152	151	55.0 %
$\hat{\mu}_{\text{mr}}^{1101}$	311	309	94.2 %	156	155	94.5 %	284	281	93.5 %	145	143	93.8 %
$\hat{\mu}_{\text{mr}}^{1010}$	253	255	95.2 %	125	128	95.2 %	253	255	95.1 %	125	128	94.9 %
$\hat{\mu}_{\text{mr}}^{0110}$	253	255	95.2 %	125	128	95.3 %	253	255	95.0 %	125	128	94.9 %
$\hat{\mu}_{\text{mr}}^{1110}$	253	255	95.2 %	125	128	95.2 %	253	255	95.1 %	125	128	94.9 %

EMP empirical standard error, *EST* mean of estimated standard error, *PER* percentage out of 2000 replications that the 95 % confidence interval based on the estimated standard error covers μ_0

10.5 Discussion

In the literature, many researchers model both $\pi(\mathbf{X})$ and $E(Y|\mathbf{X})$ to improve the robustness against model misspecification over the IPW estimator and the imputation estimator. It has been a common practice to propose and study new estimators by incorporating the imputation approach into the weighting approach (e.g., Robins et al. 1994; Tan 2006, 2010; Tsiatis 2006; Qin and Zhang 2007; Cao et al. 2009; Rotnitzky et al. 2012; Han and Wang 2013; Chan and Yam 2014; Han 2014a,b). We took an alternative view by incorporating the weighting approach into the imputation approach. This is similar to Qin et al. (2008). But the estimator proposed by Qin et al. (2008) can only take one model for $\pi(\mathbf{X})$, resulting double robustness.

Although the proposed idea was described in the setting of estimating the population mean, it can be easily extended to regression setting with missing responses and/or covariates. We leave this straightforward extension to empirical researchers who choose to apply the proposed idea.

References

Aerts M, Claeskens G, Hens N, Molenberghs G (2002) Local multiple imputation. Biometrika 89:375–388

Cao W, Tsiatis AA, Davidian M (2009) Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96:723–734

Chan KCG, Yam SCP (2014) Oracle, multiple robust and multipurpose calibration in a missing response problem. Stat Sci 29:380–396

Chen J, Sitter RR, Wu C (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. Biometrika 89:230–237

Cheng PE (1994) Nonparametric estimation of mean functionals with data missing at random. J Am Stat Assoc 89:81–87
Han P (2014a) A further study of the multiply robust estimator in missing data analysis. J Stat Plan Inference 148:101–110
Han P (2014b) Multiply robust estimation in regression analysis with missing data. J Am Stat Assoc 109:1159–1173
Han P, Wang L (2013) Estimation with missing data: beyond double robustness. Biometrika 100:417–430
Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685
Kang JDY, Schafer JL (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). Stat Sci 22:523–539
Lipsitz SR, Zhao LP, Molenberghs G (1998) A semiparametric method of multiple imputation. J R Stat Soc Ser B 60:127–144
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Owen A (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Owen A (2001) Empirical likelihood. Chapman & Hall/CRC Press, New York
Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. Ann Stat 22:300–325
Qin J, Shao J, Zhang B (2008) Efficient and doubly robust imputation for covariate-dependent missing responses. J Am Stat Assoc 103:797–810
Qin J, Zhang B (2007) Empirical-likelihood-based inference in missing response problems and its application in observational studies. J R Stat Soc Ser B 69:101–122
Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866
Robins JM, Sued M, Gomez-Lei Q, Rotnitzky A (2007) Comment: performance of double-robust estimators when “inverse probability” weights are highly variable. Stat Sci 22:544–559
Rotnitzky A, Lei Q, Sued M, Robins JM (2012) Improved double-robust estimation in missing data and causal inference models. Biometrika 99:439–456
Rubin DB (1978) Multiple imputations in sample surveys. In: Proceedings of the survey research methods section, American Statistical Association, Washington, DC, pp 20–34
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc 94:1096–1120
Tan Z (2006) A distributional approach for causal inference using propensity scores. J Am Stat Assoc 101:1619–1637
Tan Z (2010) Bounded efficient and doubly robust estimation with inverse weighting. Biometrika 97:661–682
Tsiatis AA (2006) Semiparametric theory and missing data. Springer, New York
van der Varrt AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wang D, Chen SX (2009) Empirical likelihood for estimating equations with missing values. Ann Stat 37:490–517
Wang Q, Rao JNK (2002) Empirical likelihood-based inference under imputation for missing response data. Ann Stat 30:896–924
Zhou Y, Wan ATK, Wang X (2008) Estimating equations inference with missing data. J Am Stat Assoc 103:1187–1199

Chapter 11
Maximum Smoothed Likelihood Estimation of the Centre of a Symmetric Distribution

Pengfei Li and Zhaoyang Tian

Abstract Estimating the centre of a symmetric distribution is one of the basic and important problems in statistics. Given a random sample from the symmetric distribution, natural estimators of the centre are the sample mean and sample median. However, these two estimators are either not robust or inefficient. Other estimators, such as Hodges-Lehmann estimator (Hodges and Lehmann, Ann Math Stat 34:598–611, 1963), the location M-estimator (Huber, Ann Math Stat 35:73–101, 1964) and Bondell (Commun Stat Theory Methods 37:318–327, 2008)'s estimator, were proposed to achieve high robustness and efficiency. In this paper, we propose an estimator by maximizing a smoothed likelihood. Simulation studies show that the proposed estimator has much smaller mean square errors than the existing methods under uniform distribution, t-distribution with one degree of freedom, and mixtures of normal distributions on the mean parameter, and is comparable to the existing methods under other symmetric distributions. A real example is used to illustrate the proposed method. The R code for implementing the proposed method is also provided.

11.1 Introduction

Estimating the centre of a symmetric distribution is one of the basic and important problems in statistics. When the population distribution is symmetric, the centre represents the population mean (if it exists) and population median. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. Assume that their probability distribution function is $f(x - \mu)$ with f being a probability density function and symmetric with respect to the origin. Our interest is to estimate μ.

P. Li (✉) • Z. Tian
Department of Statistics and Actuarial Science, University of Waterloo, N2L 3G1, Waterloo, ON, Canada
e-mail: pengfei.li@uwaterloo.ca; Z26Tian@uwaterloo.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_11

Symmetric population is seen in many applications. For example, Naylor and Smith (1983) and Niu et al. (2015) analyzed an dataset about biochemical measurements by a mixture of normal distributions on the scale parameter, which is a special case of symmetric distributions. More description and further analysis of this dataset will be given in Sect. 11.4. Symmetric distribution also naturally appears in paired comparison. Two data sets of the same kind, such as heights of two groups of people and the daily outputs of a factory in two different months, can be regarded as coming from the same distribution family with different location parameters. Estimating the location difference between those two data sets is one of the focuses. Assume that X and Y are two independent random variables from those two populations respectively, and let $\mu = E(Y) - E(X)$. Then X and $Z = Y - \mu$ have the same distribution. By symmetry

$$P(X - Z < t) = P(Z - X < t) = P(X - Z > -t).$$

Replacing Z with $Y - \mu$ gives that

$$P(Y - X < \mu + t) = P\,(Y - X > \mu - t).$$

That is, the distribution of $Y - X$ is symmetric with respect to μ. A natural question is whether $\mu = 0$ or not. An accurate estimator of μ can be used to construct a reliable testing procedure for $H_0 : \mu = 0$.

With the random sample $X_1, \ldots, X_n$ from the symmetric distribution $f(x - \mu)$, a simple and also traditional estimator of μ is the sample mean of the random sample. This estimator is the maximum likelihood estimator of μ under the normality assumption on $f(x)$. In such situations, sample mean has many nice properties. For example, the sample mean is normally distributed and its variance attains the Cramer-Rao lower bound. Even the underlined distribution is not normal, the central limit theorem implies that the sample mean is asymptotically normal as long as $Var(X_1) < \infty$. However, when the data exhibits heavy tails, the sample mean may have pool performance. Because of that, many other estimators are proposed in the literature. For example, sample median of $X_1, \ldots, X_n$ is another natural choice. It performs much better than the sample mean when the data appear heavy tailed. Although, sample median displays high robustness, its efficiency is disappointing sometimes. Other nonparametric methods are then proposed to improve the efficiency. Two popular choices are the Hodges-Lehmann (HL) estimator (Hodges and Lehmann 1963) and the location M-estimator of Huber (1964). Recently, Bondell (2008) proposed an estimator based on the characteristic function, which is shown to be more robust than HL estimator and M-estimator in simulation studies.

Although HL estimator, M-estimator, and Bondell's estimator are shown to be robust and efficient in simulation studies, to the best of our knowledge, none of them has a nonparametric likelihood interpretation. We feel that there is room for improvement. On one hand, empirical likelihood method (Owen 2001) seems to be

a quite natural choice. On the other hand, it is quite challenging to incorporate the symmetry information on $f(x)$. See the discussion in Sect. 10.1 of Owen (2001).

In this paper, we propose to estimate μ by maximizing a smoothed likelihood. The concept of smoothed likelihood is first proposed by Eggermont and LaRiccia (1995). This method gives a nonparametric likelihood interpretation for the classic kernel density estimation. Later on, it has been used to estimate the monotone and unimodal densities (Eggermont and Lariccia 2000), the component densities in multivariate mixtures (Levine et al. 2011), the component densities in mixture model with known mixing proportions (Yu et al. 2014). These papers demonstrated that the smoothed likelihood method can easily incorporate the constraint on the density function. This motivates us to use it to estimate the centre μ by incorporating the symmetric assumption through the density function.

The organization of the paper is as follows. In Sect. 11.2, we present the idea of smoothed likelihood and apply it to obtain the maximum smoothed likelihood estimator of μ. In Sects. 11.3 and 11.4, we present some simulation results and a real data analysis, respectively. Conclusion is given in Sect. 11.5. The R (R Development Core Team 2011) code for implementing the proposed method is given in the Appendix.

11.2 Maximum Smoothed Likelihood Estimation

In this section, we first review the idea of smoothed likelihood and then apply it to estimate the centre μ of a symmetric distribution.

11.2.1 Idea of Smoothed Likelihood

Given a sample $X_1, \ldots, X_n$ from the probability density function $g(x)$, the log-likelihood function of $g(x)$ is defined to be

$$\sum_{i=1}^{n} \log\{g(X_i)\}$$

subject to the constraint that $g(x)$ is a probability density function. Maximizing such a log-likelihood, however, does not lead to a consistent solution, since we can make the log-likelihood function arbitrarily large by setting $g(X_i) \to \infty$ for a specific i or even every $i = 1, \ldots, n$.

The unboundedness of the likelihood function can be tackled by the following smoothed log likelihood approach (see Eggermont and LaRiccia 1995 and

Eggermont and Lariccia 2001, Chapter 4). Define the nonlinear smoothing operator $\mathscr{N}_h g(x)$ of a density $g(x)$ by

$$\mathscr{N}_h g(x) = \exp\left\{ \int K_h(u-x) \log g(u) du \right\},$$

where $K_h(x) = \frac{1}{h} K(x/h)$, $K(\cdot)$ is a symmetric kernel function, and h is the bandwidth for the nonlinear smoothing operator. The smoothed likelihood of $g(x)$ is defined as

$$\sum_{i=1}^{n} \log \mathscr{N}_h g(X_i) = n \int \tilde{g}_n(x) \log g(x) dx,$$

where $\tilde{g}_n(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$ is the usual kernel density estimator of $g(x)$. Interestingly, the smoothed likelihood function is maximized at $g(x) = \tilde{g}_n(x)$. This gives a nonparametric likelihood interpretation for the usual kernel density estimator.

11.2.2 Maximum Smoothed Likelihood Estimator of the Centre

Suppose we have a random sample $X_1, \ldots, X_n$ from the population with the probability density function $f(x - \mu)$. In this subsection, we consider estimating the centre μ by using the maximum smoothed likelihood method. Following the principle of the smoothed likelihood presented in last subsection, we define the smoothed likelihood of $\{f, \mu\}$ as follows:

$$l_n(f, \mu) = \sum_{i=1}^{n} \log \mathscr{N}_h f(X_i - \mu).$$

After some calculation, $l_n(f, \mu)$ can be written as

$$l_n(f, \mu) = n \int \tilde{f}_n(x + \mu) \log f(x) dx,$$

where $\tilde{f}_n(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$ is the kernel density estimation of $f(x - \mu)$. The maximum smoothed likelihood estimator of $\{f(x), \mu\}$ is defined to be

$$\{\hat{f}_{Smo}(x), \hat{\mu}_{Smo}\} = \arg \sup_{f(x),\ \mu} l_n(f, \mu)$$

subject to the constraint that $f(x)$ is a symmetric probability density function with respect to the origin.

We proceed in two steps to get $\hat{\mu}_{Smo}$. In the first step, we fix μ and maximize $l_n(f, \mu)$ subject to the constraint that $f(x)$ is a symmetric probability density function around 0. Since $f(x)$ is symmetric around 0, we have

$$\int \tilde{f}_n(x+\mu)\log f(x)dx = \int \tilde{f}_n(-x+\mu)\log f(-x)dx = \int \tilde{f}_n(-x+\mu)\log f(x)dx.$$

Hence the smoothed likelihood function can be written as

$$l_n(f, \mu) = n\int 0.5\{\tilde{f}_n(x+\mu) + \tilde{f}_n(-x+\mu)\}\log f(x)dx.$$

Note that $0.5\{\tilde{f}_n(x+\mu) + \tilde{f}_n(-x+\mu)\}$ is a probability density function and is symmetric around 0. Hence $l_n(f, \mu)$ is maximized at

$$f(x) = 0.5\{\tilde{f}_n(x+\mu) + \tilde{f}_n(-x+\mu)\}.$$

In the second step, we plug $f(x) = 0.5\{\tilde{f}_n(x+\mu) + \tilde{f}_n(-x+\mu)\}$ to $l_n(f, \mu)$ and obtain the profile smoothed likelihood function of μ:

$$\begin{aligned} pl_n(\mu) &= n\int \tilde{f}_n(x+\mu)\log\{0.5\tilde{f}_n(x+\mu) + 0.5\tilde{f}_n(\mu-x)\}dx \\ &= n\int \tilde{f}_n(x)\log\{0.5\tilde{f}_n(x) + 0.5\tilde{f}_n(2\mu-x)\}dx. \end{aligned}$$

The maximum smoothed likelihood estimator of μ can be equivalently defined as

$$\hat{\mu}_{Smo} = \arg\sup_{\mu} pl_n(\mu).$$

To implement the above method, we need to specify the kernel density $K(x)$ and select the bandwidth h. The commonly used kernel density is the standard normal density, which is used in our implementation. Methods for choosing a bandwidth for kernel density estimation are readily available in the literature. In our implementation, we have used function *dpik()* in the R package *KernSmooth* to choose the bandwidth h. This package essentially implements the kernel methods in Wand and Jones (1995). We have written a R function to calculate $pl_n(\mu)$ and then use *optim()* to numerically calculate $\hat{\mu}_{Smo}$. These code is provided in the Appendix.

11.3 Simulation Study

We conduct simulation to test the efficiency of the maximum smoothed likelihood method and compare it with the five existing methods: the sample mean, sample median, HL estimator, M-estimator, and Bondell (2008)'s estimator. We choose

Table 11.1 Mean square errors (× sample size) for six estimates under thirteen symmetric distributions with the sample size equal to 50

Distribution	Mean	Median	HL	M-estimator	Bondell	$\hat{\mu}_{Smo}$
$N(0,1)$	1.092	1.613	1.130	1.110	1.147	1.140
DE	1.906	1.138	1.300	1.379	1.361	1.434
$U(-2,2)$	1.376	3.780	1.556	1.454	1.686	0.682
$t(1)$	5634	2.701	3.786	4.128	3.017	2.940
$t(2)$	71.13	1.986	1.925	2.102	1.834	2.010
$t(3)$	2.855	1.821	1.580	1.506	1.552	1.678
$0.5N(-1,0.5)+0.5N(1,0.5)$	1.555	5.088	1.749	1.574	1.858	1.150
$0.5N(-1,0.75)+0.5N(1,0.75)$	1.722	4.148	1.956	2.074	1.991	1.610
$0.5N(0,1)+0.5N(0,3)$	2.058	2.461	1.885	1.990	1.892	2.016
$0.5N(0,1)+0.5N(0,5)$	3.080	3.135	2.644	2.585	2.684	2.903
$0.5N(0,1)+0.5N(0,10)$	5.705	3.891	4.014	4.220	4.166	4.123
$0.9N(0,1)+0.1N(0,9)$	1.680	1.663	1.247	1.232	1.178	1.211
$0.8N(0,1)+0.2N(0,9)$	2.574	1.939	1.697	1.728	1.680	1.802

Table 11.2 Mean square errors (× sample size) for six estimates under thirteen symmetric distributions with the sample size equal to 100

Distribution	Mean	Median	HL	M-estimator	Bondell	$\hat{\mu}_{Smo}$
$N(0,1)$	0.991	1.626	1.042	1.013	1.045	1.032
DE	2.035	1.071	1.357	1.466	1.418	1.392
$U(-2,2)$	1.351	3.872	1.483	1.456	1.633	0.572
$t(1)$	20572	2.476	3.306	3.749	2.801	2.487
$t(2)$	9.517	2.050	2.045	2.132	1.944	1.980
$t(3)$	2.829	1.772	1.503	1.677	1.486	1.589
$0.5N(-1,0.5)+0.5N(1,0.5)$	1.590	5.318	1.770	1.519	1.887	1.122
$0.5N(-1,0.75)+0.5N(1,0.75)$	1.766	4.328	2.035	1.898	2.064	1.553
$0.5N(0,1)+0.5N(0,3)$	1.996	2.588	1.864	1.843	1.883	1.978
$0.5N(0,1)+0.5N(0,5)$	3.058	3.026	2.548	2.766	2.600	2.730
$0.5N(0,1)+0.5N(0,10)$	5.493	3.538	3.585	3.887	3.703	3.443
$0.9N(0,1)+0.1N(0,9)$	1.868	1.736	1.274	1.243	1.205	1.260
$0.8N(0,1)+0.2N(0,9)$	2.568	2.091	1.694	1.690	1.644	1.717

thirteen symmetric distributions with the centre zero: the standard normal distribution $N(0,1)$, double exponential (DE) distribution, uniform distribution over $(-2,2)$ (denoted by $U(-2,2)$), three t distributions (denoted by $t(v)$ with v being the degrees of freedom), two mixtures of normal distributions on the mean parameter, and five mixtures of normal distributions on the scale parameter. We consider three sample sizes: 50, 100, and 200. We calculate the mean square errors for each of six estimates based on 1000 replications. The results are summarized in Tables 11.1, 11.2 and 11.3, in which the mean square errors are multiplied by the sample size.

It is seen from Tables 11.1, 11.2 and 11.3 that mean square errors of sample means are significantly larger under some specific distributions, such as $t(1)$

Table 11.3 Mean square errors (× sample size) for six estimates under thirteen symmetric distributions with the sample size equal to 200

Distribution	Mean	Median	HL	M-estimator	Bondell	$\hat{\mu}_{Smo}$
$N(0,1)$	1.003	1.537	1.031	1.085	1.034	1.033
DE	1.847	1.069	1.256	1.403	1.316	1.260
$U(-2,2)$	1.407	4.125	1.512	1.410	1.707	0.448
$t(1)$	74611	2.653	3.608	3.881	2.908	2.575
$t(2)$	13.17	1.977	1.914	2.011	1.821	1.815
$t(3)$	3.470	1.745	1.537	1.756	1.492	1.552
$0.5N(-1,0.5)+0.5N(1,0.5)$	1.471	5.426	1.611	1.605	1.736	0.980
$0.5N(-1,0.75)+0.5N(1,0.75)$	1.758	4.521	2.007	1.945	2.041	1.483
$0.5N(0,1)+0.5N(0,3)$	1.893	2.257	1.745	1.912	1.766	1.860
$0.5N(0,1)+0.5N(0,5)$	3.224	3.115	2.650	2.480	2.691	2.738
$0.5N(0,1)+0.5N(0,10)$	5.322	3.679	3.568	4.006	3.709	3.226
$0.9N(0,1)+0.1N(0,9)$	1.790	2.036	1.377	1.370	1.340	1.375
$0.8N(0,1)+0.2N(0,9)$	2.795	2.213	1.793	1.693	1.714	1.714

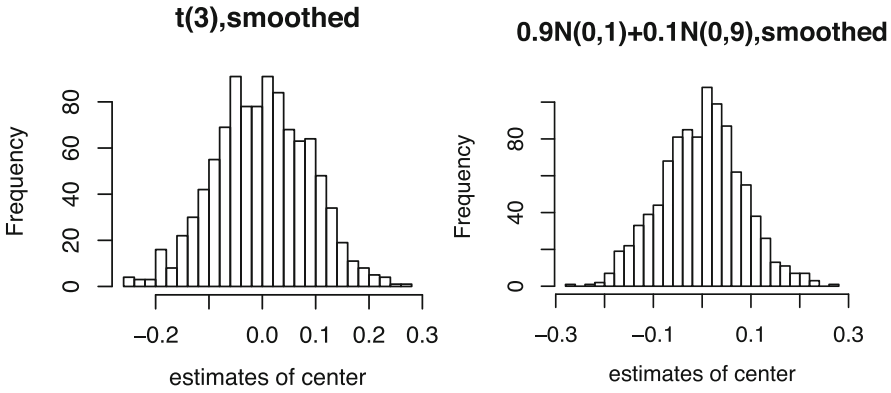

Fig. 11.1 Histograms of maximum smoothed likelihood estimates under $t(3)$ and $0.9N(0,1)+0.1N(0,9)$ distributions

and $t(2)$, as variances of those distributions do not exist or are very large. On average, efficiency of sample median is low compared to other methods except for the double exponential distribution. Compared with HL estimator, M-estimator, and Bondell (2008)'s estimator, the maximum smoothed likelihood estimator has smaller mean square errors under uniform distribution and $t(1)$, and mixtures of normal distributions on mean parameters. For other distributions, the new estimator has comparable performance as HL estimator, M-estimator, and Bondell (2008)'s estimator.

As we can see from Tables 11.1, 11.2 and 11.3, the products of sample sizes and mean square errors of $\hat{\mu}_{Smo}$ almost remain as a constant. Hence we conjecture that the mean square error of $\hat{\mu}_{Smo}$ is of the order $O\left(n^{-1}\right)$. In Fig. 11.1, we plot the

histograms of maximum smoothed likelihood estimates under $t(3)$ and $0.9N(0, 1) + 0.1N(0, 9)$. It is seen that their histograms behave similarly to normal distribution. Hence we also conjecture that the asymptotic distribution of $\hat{\mu}_{Smo}$ is normal. We leave these two conjectures for future study.

11.4 Real Application

In clinical chemistry, the clinical assessment of biochemical measurements is typically carried out by reference to a "normal range", which is the 95 % prediction interval of the mean measurement for a "healthy" population (Naylor and Smith 1983). One way of obtaining such a normal range is to first collect a large sample of biochemical measurements from a healthy population. However, in practice, it may be difficult to collect measurements only from healthy individuals. Instead, measurements from a contaminated sample, containing both healthy and unhealthy individuals, are obtained. Because of the potential heterogeneity in the contaminated sample, mixtures of normal distributions are widely used in such analyses.

Naylor and Smith (1983) and Niu et al. (2015) used a mixture of two normal distributions on the scale parameter to model a contaminated sample of 542 blood-chloride measurements collected during routine analysis by the Department of Chemical Pathology at the Derbyshire Royal Infirmary. Note that the mixture of two normal distributions on the scale parameter is a symmetric distribution. Here we illustrate our method by estimating the centre of this data set. The maximum smoothed likelihood estimate and other five existing estimates are shown in Table 11.4. As we can see, all six estimates are close to each other. According to our simulation experience on the mixtures of normal distributions on the scale parameter, we expect that the variances of the proposed method, Bondell's method, HL estimator, and M-estimator are similar and may be smaller than those of sample mean and sample median.

Table 11.4 Six estimates of the centre of 542 blood-chloride measurements

Mean	Median	HL	M-estimator	Bondell	$\hat{\mu}_{Smo}$
99.985	100.017	100.033	100.068	100.087	100.100

11.5 Conclusion

In this paper, we proposed the maximum smoothed likelihood estimator for the centre of a symmetric distribution. The proposed method performs better than those widely used estimators such as the HL estimator, M-estimator, and Bondell (2008)'s estimator under the uniform, $t(1)$, and mixtures of normal distributions on mean parameters. It has comparable performance to the HL estimator, M-estimator, and Bondell (2008)'s estimator under other symmetric distributions.

We admit that so far the proposed method lacks theoretical justification. Further work on the consistency and asymptotic distribution of $\hat{\mu}_{Smo}$ will provide solid theoretical support for its application. There is also room for improvement for its computational efficiency.

Acknowledgements Dr. Li's work is partially supported by the Natural Sciences and Engineering Research Council of Canada grant No RGPIN-2015-06592.

Appendix: R code for calculating $\hat{\mu}_{Smo}$

```
library("KernSmooth")
library("ICSNP")

norm.kern=function(x,data,h)
{
out=mean( dnorm( (x-data)/h ) )/h
out
}

dkern=function(x,data)
{
h=dpik(data,kernel="normal")
out=lapply(x,norm.kern,data=data,h=h)
as.numeric(out)
}

dfint=function(x,data,mu)
{
p1=dkern(x,data)
p2=log(0.5*dkern(x,data)+0.5*dkern(2*mu-x,data)+1e-100)
p1*p2
}

pln=function(mu,data)
{
h=dpik(data,kernel="normal")
```

```
out=integrate(dfint, lower=min(data)-10*h,
                    upper=max(data)+10*h, data=data, mu=mu)
-out$value
}


hatmu.smooth=function(data)
{
##Input: data set
##Output: maximum smoothed likelihood estimate

hl.est=hl.loc(data)
est=optim(hl.est,pln,data=data,method="BFGS")$par
est
}

##Here is an example
set.seed(1221)
data=rnorm(100,0,1)
hatmu.smooth(data)
##Result: 0.1798129
```

References

Bondell HD (2008) On robust and efficient estimation of the center of symmetry. Commun Stat Theory Methods 37:318–327

Eggermont PPB, LaRiccia VN (1995) Maximum smoothed likelihood density estimation for inverse problems. Ann Stat 23:199–220

Eggermont PPB, Lariccia VN (2000) Maximum likelihood estimation of smooth monotone and unimodal densities. Ann Stat 28:922–947

Eggermont PPB, Lariccia VN (2001) Maximum penalized likelihood estimation: volume I: density estimation. Springer, New York

Hodges JL, Lehmann EL (1963) Estimates of location based on rank tests. Ann Math Stat 34:598–611

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101

Levine M, Hunter D, Chauveau D (2011) Maximum smoothed likelihood for multivariate mixtures. Biometrika 98:403–416

Naylor JC, Smith AFM (1983) A contamination model in clinical chemistry: an illustration of a method for the efficient computation of posterior distributions. J R Stat Soc Ser D 32:82–87

Niu X, Li P, Zhang P (2016) Testing homogeneity in a scale mixture of normal distributions. Stat Pap 57:499–516

Owen AB (2001) Empirical likelihood. Chapman and Hall/CRC, New York

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Wand MP, Jones MC (1995) Kernel smoothing. Chapman and Hall, London

Yu T , Li P, Qin J (2014) Maximum smoothed likelihood component density estimation in mixture models with known mixing proportions. arXiv preprint, arXiv:1407.3152

Chapter 12
Modelling the Common Risk Among Equities: A Multivariate Time Series Model with an Additive GARCH Structure

Jingjia Chu, Reg Kulperger, and Hao Yu

Abstract The DCC GARCH models (Engle, J Bus Econ Stat 20:339–350, 2002) have been well studied to describe the conditional covariance and correlation matrices while the common risk among series cannot be captured intuitively by the existing multivariate GARCH models. A new class of multivariate time series model with an additive GARCH type structure is proposed. The dynamic conditional covariance between series are aggregated by a common risk term which has been the key to characterize the conditional correlation.

12.1 Introduction

The volatility of the log return is necessary to be estimated for the financial modelling. The Black-Scholes-Merton (BSM) model (Black and Scholes 1973; Merton 1973) assumes that the log returns follow independently and identically (i.i.d) normal distribution while the real market data shows volatility clustering and heavy tail which violates this assumption. The conditional heteroscedasticity models have been widely used in finance today to estimate the volatility to take the feature of the financial return series into consideration. Let $\mathscr{F}_t$ be the information set (σ-field) of all information up to time point t and assume $\{x_t : t \in T\}$ is the observed process, then the general form of the model is in a multiplicative structure given by

$$\begin{cases} x_t = \epsilon_t \sigma_t \\ E(x_t|\mathscr{F}_{t-1}) = 0 \\ E(x_t^2|\mathscr{F}_{t-1}) = \sigma_t^2, \end{cases} \tag{12.1}$$

where the innovations $\{\epsilon_t : t \in T\}$ are i.i.d. white noises with mean 0 and variance 1. The innovations are independent of $\mathscr{F}_{t-1}$, σ_t's are $\mathscr{F}_{t-1}$ adapted.

J. Chu (✉) • R. Kulperger • H. Yu
Department of Statistical and Actuarial Sciences, Western University, London, ON, Canada
e-mail: jchu223@uwo.ca; rjk@stats.uwo.ca; hyu@stats.uwo.ca

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5_12

Engle (1982) introduced the autoregressive conditional heteroscedasticity (ARCH) model with the unique ability of capturing volatility clustering in financial time series. The ARCH(q) model defines the conditional variance of x_t to be

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i x_{t-i}^2.$$

The lag q tends to be large when the model was applied to real data. Subsequently, GARCH model introduced by Bollerslev (1986) extends the formula for σ_t by adding autoregressive terms of σ_t^2. The conditional variance of the univariate GARCH(p,q) model was defined as

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i x_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2.$$

When there are more than one time series, it is necessary to understand the co-movements of the returns since that the volatility of stock returns are correlated with each other. In contrast to the univariate cases, the multivariate volatility models based on a GARCH dependence are much more flexible. The multivariate GARCH models are specified based on the first two conditional moments as well as the univariate cases. The first multivariate volatility model is the half-Vec ($vech$) multivariate GARCH defined by Bollerslev et al. (1988), which is also one of the most general forms of multivariate GARCH models.

$$\begin{cases} \mathbf{x_t} = H_t^{1/2} \boldsymbol{\epsilon_t}, \\ \mathbf{h_t} = \mathbf{c} + \sum_{i=1}^{q} A_i \boldsymbol{\eta}_{\mathbf{t-i}} + \sum_{j=1}^{p} B_j \mathbf{h}_{\mathbf{t-j}}, \end{cases} \tag{12.2}$$

where $\boldsymbol{\epsilon_t}$'s are white noises with mean $\mathbf{0}$ and variance I_m,

$$\mathbf{h_t} = \text{vech}(H_t),$$
$$\boldsymbol{\eta_t} = \text{vech}(\boldsymbol{\epsilon_t}\boldsymbol{\epsilon_t}^{\intercal}).$$

In this class of model, the conditional covariance matrix is modelled directly. The number of parameters in the general m-dimensional case increases at a rate proportional to m^4, which makes it difficult to get the estimations.

There are simpler forms of multivariate GARCH by specifying the H_t in different ways. The constant correlation coefficient (CCC)-GARCH model is presented by Bollerslev (1990), who assumes that the conditional correlation matrix $R = (\rho_{i,j})_{i,j=1,\cdots,m}$ would be time-invariant, and reduces the number of parameters to

$O(m^2)$. The model is defined as

$$\begin{cases} \mathbf{x_t} = H_t^{1/2}\boldsymbol{\epsilon_t} \\ H_t = S_t R S_t \\ \Delta_t = c + \sum_{i=1}^{q} A_i \mathbf{x_{t-i}}^2 + \sum_{j=1}^{p} B_j \Delta_{t-j} \end{cases} \tag{12.3}$$

where Δ_t is a vector contains the diagonal elements of H_t and S_t is the diagonal matrix of $\sqrt{\Delta_t}$. A less restrictive time-varying conditional correlation version, called the dynamic correlation coefficient (DCC) GARCH, is studied by Engle (2002) and Tse and Tsui (2002). Both CCC-GARCH and DCC-GARCH models are built by modelling the conditional variance of each series and the conditional correlation between series. However, all of these multivariate GARCH models and their extensions do not have a simple way to capture the common risk among different stocks.

Information flows around the world almost instantaneously, thus most markets (Asian, European, and American) will react to the same events (good news or bad news). Most stock prices will go up or down together following the big events (random shocks) currently. The strong positive association between the equity variance and several variables is confirmed by Christie (1982). Recently, Carr and Wu (2009) found that a common stochastic variance risk factor exists among the stocks by using the market option premiums. There are several different approaches used in finance and economy to describe the same driven process in the literature. The first approach is the asset pricing model (Treynor, Market value, time, and risk. Unpublished manuscript, pp 95–209, 1961; Toward a theory of market value of risky assets. Unpublished manuscript, p 40, 1962; Sharpe 1964; Treynor 2008) and its generalizations (Burmeister and Wall 1986; Fama and French 1993, 2015) which quantifies the process by the market indices or some macro economic factors. Another approach is the volatility jump models (Duffie et al. 2000; Tankov and Tauchen 2011) which assumes the effect of the news is a discrete process would happen by chance over the time. Different multivariate GARCH models are introduced to describe the same underlying driven process in the financial time series (Engle et al. 1990; Girardi and Tolga Ergun 2013; Santos and Moura 2014). These models either involve other market variables, modelling the underlying risk as a discrete process or characterize the common risk implicitly.

We develop a simple common risk model which keeps the GARCH structure and involves the return only. In the light of Carr and Wu (2009), the innovations are divided into two parts. A new additive GARCH type model was proposed by using a common risk term to characterize the internal relationship among series explicitly. The common risk term could be used as an indicator of the shock among series. The conditional correlations aggregated by this common risk term are changing dynamically. This model is also able to capture the conditional correlation clustering

phenomenon described in So and Yip (2012). The common risk term would show a latency after it reaching a peak since it follows a GARCH type structure, which means that a big shock would take some time to calm down.

The notation and the new common underlying risk model is introduced in Sect. 12.2. In Sect. 12.3, we discuss the model is identifiable and the estimates based on Gaussian quasi likelihood are unique under certain assumptions. In Sect. 12.4, the results of a Monte Carlo simulation study are shown and the estimated conditional volatility is compare with some other GARCH models based on a bivariate dataset.

12.2 Model Specification

Consider an $\mathbb{R}^m$-valued stochastic process $\{\mathbf{x_t}, t \in \mathbb{Z}\}$ on a probability space $(\Omega, \mathscr{A}, \mathscr{P})$ and a multidimensional parameter $\boldsymbol{\theta}$ in the parameter space $\Theta \subset \mathbb{R}^s$. We say that $\mathbf{x_t}$ is a common risk model with an additive GARCH structure if, for all $t \in \mathbb{Z}$, we have

$$\begin{cases} x_{1,t} = \epsilon_{1,t}\sigma_{1,t} + \epsilon_{0,t}\sigma_{0,t} \\ x_{2,t} = \epsilon_{2,t}\sigma_{2,t} + \epsilon_{0,t}\sigma_{0,t} \\ \quad \cdots\cdots \\ x_{m,t} = \epsilon_{m,t}\sigma_{m,t} + \epsilon_{0,t}\sigma_{0,t} \end{cases} \tag{12.4}$$

where $\sigma_{1,t}^2, \cdots, \sigma_{m,t}^2$ are following a GARCH type structure,

$$\begin{cases} \sigma_{1,t}^2 = \alpha_1 g(x_{1,t-1})^2 + \beta_1 \sigma_{1,t-1}^2 \\ \sigma_{2,t}^2 = \alpha_2 g(x_{2,t-1})^2 + \beta_2 \sigma_{2,t-1}^2 \\ \quad \cdots\cdots \\ \sigma_{m,t}^2 = \alpha_m g(x_{m,t-1})^2 + \beta_m \sigma_{m,t-1}^2 \\ \sigma_{0,t}^2 = \omega_0 + \beta_{01}\sigma_{1,t}^2 + \cdots + \beta_{0m}\sigma_{m,t}^2. \end{cases} \tag{12.5}$$

The size of the effect on $\sigma_{0,t}^2$ is linearly increasing with each observed element of $\mathbf{x_t}$. The conditional volatilities based on this model will expose to infinity and the mean reverse property will hardly hold when one of the $\sigma_{i,t}$ terms larger than 1. The function g could be a continuous bounded function to avoid this kind of situation.

The $m+1$ dimensional innovation terms $\{\boldsymbol{\epsilon_t} - \infty < t < \infty\}$ are independent and identically distributed with mean $\mathbf{0}$ and covariance Σ where Σ has the same parameterization as a correlation matrix,

$$\Sigma = \begin{pmatrix} R & 0 \\ 0 & 1 \end{pmatrix}.$$

The innovation can be divided into two parts $\boldsymbol{\epsilon}_\mathbf{t}^\mathsf{T} = (\boldsymbol{\epsilon}_{\mathbf{t,ind}}^\mathsf{T}, \epsilon_{0,t})$. The first part is a m dimensional correlated individual shocks $\boldsymbol{\epsilon}_{\mathbf{t,ind}}$ and the second part is a univariate common shock term $\epsilon_{0,t}$.

Define the following notations, $D_t = \text{diag}\{\sigma_{1,t}, \sigma_{2,t} \cdots, \sigma_{m,t}\}$, $\mathbf{1} = (1, 1, \cdots, 1)^\mathsf{T}$

$$\boldsymbol{\epsilon}_{\mathbf{t,ind}} = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{m,t} \end{pmatrix}_{m\times 1}, \boldsymbol{\epsilon}_\mathbf{t} = \begin{pmatrix} \boldsymbol{\epsilon}_{\mathbf{t,ind}} \\ \epsilon_{0,t} \end{pmatrix} = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ \vdots \\ \epsilon_{m,t} \\ \epsilon_{0,t} \end{pmatrix}_{(m+1)\times 1}.$$

Then Eq. 12.4 could be written in a matrix form:

$$\mathbf{x_t} = D_t \boldsymbol{\epsilon}_{\mathbf{t,ind}} + \sigma_{0,t}\epsilon_{0,t}\mathbf{1}. \tag{12.6}$$

So the model could be specified either by Eqs. 12.6 and 12.5 together or Eqs. 12.4 and 12.5 together. The conditional covariance matrix of $\mathbf{x_t}$ can be computed by definition $H_t = \text{cov}(\mathbf{x_t}|\mathscr{F}_{t-1})$.

$$H_t = \begin{pmatrix} \sigma_{0,t}^2 + \sigma_{1,t}^2 & \sigma_{0,t}^2 + \rho_{1,2}\sigma_{1,t}\sigma_{2,t} & \cdots & \sigma_{0,t}^2 + \rho_{1,m}\sigma_{1,t}\sigma_{m,t} \\ \sigma_{0,t}^2 + \rho_{1,2}\sigma_{1,t}\sigma_{2,t} & \sigma_{0,t}^2 + \sigma_{2,t}^2 & \cdots & \sigma_{0,t}^2 + \rho_{2,m}\sigma_{2,t}\sigma_{m,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{0,t}^2 + \rho_{1,m}\sigma_{1,t}\sigma_{m,t} & \sigma_{0,t}^2 + \rho_{2,m}\sigma_{2,t}\sigma_{m,t} & \cdots & \sigma_{0,t}^2 + \sigma_{m,t}^2 \end{pmatrix}.$$

H_t can be written as a sum of two parts: $H_t = \sigma_{0,t}^2 J + D_t R D_t$ where J is a $m \times m$ matrix with 1 as all the elements (or $J = \mathbf{1}\mathbf{1}^\mathsf{T}$).

The number of parameters is increasing at the rate $O(m^2)$ which is in the same manner as in the CCC-GARCH model. We could separate the vector of unknown parameters into two parts: the parameters in the innovations correlation matrix Σ and the coefficients in the Eq. 12.5. The number of total parameters is $s = s_1 + 3m + 1$ where $s_1 = \frac{m(m-1)}{2}$ is the number of parameters in R.

The conditional correlation between series i and series j can be represented by the elements in H_t matrix. The dynamic correlation between series i and j can be

calculated as

$$\begin{aligned}\rho_{ij,t} &= \frac{\operatorname{cov}(x_{i,t}, x_{j,t})}{\sqrt{\operatorname{var}(x_{i,t})\operatorname{var}(x_{j,t})}} \\ &= \frac{\sigma_{0,t}^2 + \rho_{i,j}\sigma_{i,t}\sigma_{j,t}}{\sqrt{(\sigma_{0,t}^2 + \sigma_{i,t}^2)(\sigma_{0,t}^2 + \sigma_{j,t}^2)}} \\ &= \frac{1 + \rho_{i,j}\left(\frac{\sigma_{i,t}}{\sigma_{0,t}}\right)\left(\frac{\sigma_{j,t}}{\sigma_{0,t}}\right)}{\sqrt{1 + \left(\frac{\sigma_{i,t}}{\sigma_{0,t}}\right)^2}\sqrt{1 + \left(\frac{\sigma_{j,t}}{\sigma_{0,t}}\right)^2}}.\end{aligned}$$

From the equations above, the conditional correlation matrix $R_t = (\rho_{ij,t})_{i,j=1,\cdots,m}$ tends to be J defined above when the common risk term $\sigma_{0,t}$ is much larger than both $\sigma_{i,t}$ and $\sigma_{j,t}$. In this case, the common risk term is dominant and all the log return series are nearly perfect correlated. On the contrary, the conditional correlation matrix will be approaching the constant correlation matrix R when the common risk term is much smaller and close to 0. Then, the conditional correlation will become time invariant which is the same as a CCC-GARCH model. Mathematically,

$$\begin{aligned}R_t &\to J \text{ when } \sigma_{0,t} \to \infty, \\ R_t &\to R \text{ when } \sigma_{0,t} \to 0.\end{aligned}$$

12.3 Gaussian QMLE

A distribution must be specified for the innovations $\boldsymbol{\epsilon_t}$ process in order to form the likelihood function. The maximum likelihood (ML) method is particularly useful in statistical inferences for models because it provides us with an estimator which has both consistency and asymptotic normality. The quasi-maximum likelihood (QML) method could draw statistical inference based on a misspecified distribution of the innovations while the ML method assumes that the true distribution of the innovation is the specified distribution. ML method essentially is a special case of the QML method with no specification error.

We can construct the Gaussian quasi likelihood function based on the density of conditional distribution $\mathbf{x_t}|\mathscr{F}_{t-1}$. The vector of the parameters

$$\boldsymbol{\theta} = (\rho_{1,2}, \cdots, \rho_{m-1,m}, \alpha_1, \cdots, \alpha_m, \beta_1, \cdots, \beta_m, \omega_0, \beta_{01}, \cdots, \beta_{0m})^{\mathsf{T}} \tag{12.7}$$

belongs to a parameter space of the form

$$\Theta \subset [0, \infty)^{\frac{m(m-1)}{2}+3m+1}. \tag{12.8}$$

The true value of the parameter is unknown, and denoted by

$$\boldsymbol{\theta}_0 = (\rho_{1,2}^{(0)}, , \cdots, \rho_{m-1,m}^{(0)}, \alpha_1^{(0)}, \cdots, \alpha_m^{(0)}, \beta_1^{(0)}, \cdots, \beta_m^{(0)}, \omega_0^{(0)}, \beta_{01}^{(0)}, \cdots, \beta_{0m}^{(0)})^{\intercal}.$$

12.3.1 The Distribution of the Observations

The observations $\mathbf{x_t}$'s are assumed following a realization of a m-dimensional common risk process and $\boldsymbol{\epsilon_t}$'s are i.i.d. normally distributed with mean $\mathbf{0}$ and covariance Σ. Equation 12.4 shows that based on the past, the observations can be written as linear combinations of normally distributed variables, then the conditional distribution of the observations $\mathbf{x_t}$'s are multivariate normal as well, e.g. $\mathbf{x_t}|\mathscr{F}_{t-1} \sim N(\mathbf{0}, H_t)$. The model in Sect. 12.2 can be revised to a different form as

$$\begin{cases} \mathbf{x_t} = H_t^{1/2}\boldsymbol{\xi_t} \\ H_t = \begin{pmatrix} \sigma_{0,t}^2 + \sigma_{1,t}^2 & \sigma_{0,t}^2 + \rho_{1,2}\sigma_{1,t}\sigma_{2,t} & \dots & \sigma_{0,t}^2 + \rho_{1,m}\sigma_{1,t}\sigma_{m,t} \\ \sigma_{0,t}^2 + \rho_{1,2}\sigma_{1,t}\sigma_{2,t} & \sigma_{0,t}^2 + \sigma_{2,t}^2 & \dots & \sigma_{0,t}^2 + \rho_{2,m}\sigma_{2,t}\sigma_{m,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{0,t}^2 + \rho_{1,m}\sigma_{1,t}\sigma_{m,t} & \sigma_{0,t}^2 + \rho_{2,m}\sigma_{2,t}\sigma_{m,t} & \dots & \sigma_{0,t}^2 + \sigma_{m,t}^2 \end{pmatrix}, \end{cases} \tag{12.9}$$

where the innovation $\boldsymbol{\xi_t}$ are a sequence of i.i.d m-dimensional standard normal variables. Then the quasi log likelihood function is given by

$$L_n(\boldsymbol{\theta}) = -\frac{1}{2n}\sum_{t=1}^{n}\{\log|H_t(\boldsymbol{\theta})| + \mathbf{x_t}^{\intercal}H_t(\boldsymbol{\theta})^{-1}\mathbf{x_t}\} = -\frac{1}{2n}\sum_{t=1}^{n} l_t(\boldsymbol{\theta}). \tag{12.10}$$

The driving noises $\boldsymbol{\epsilon_t}$'s are i.i.d $N(\mathbf{0}, \Sigma)$, so the conditional distribution of $\mathbf{x_t}$ is $N(\mathbf{0}, H_t(\boldsymbol{\theta}))$. The QML estimator is defined as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_n &= \arg\max_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{t=1}^{n}\{\log|H_t(\boldsymbol{\theta})| + \mathbf{x_t}^{\intercal}H_t(\boldsymbol{\theta})^{-1}\mathbf{x_t}\} \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{t=1}^{n} l_t(\boldsymbol{\theta}). \end{aligned} \tag{12.11}$$

12.3.2 Identifiability

We start this section with the concept of parameter identifiability.

Definition 1 Let $H_t(\boldsymbol{\theta})$ be the conditional second moment of $\mathbf{x_t}$, Θ be the parameter space, then $H_t(\boldsymbol{\theta})$ is identifiable if $\forall \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \Theta$, $H_t(\boldsymbol{\theta_1}) = H_t(\boldsymbol{\theta_2}) \quad a.s. \Rightarrow \boldsymbol{\theta_1} = \boldsymbol{\theta_2}$.

It is necessary to study the condition of parameter identification since the parameter estimates are based on the maximum of the likelihood function. The solution needs to be unique when the likelihood function reaches its maximum.

Theorem 1 *Assume that:*

Assumption 1 $\forall \boldsymbol{\theta} \in \Theta$, $\alpha_i > 0$ *and* $\beta_i \in [0, 1)$ *for* $i = 1, \cdots, m$.

Assumption 2 *The model in Eq. 12.4 is stationary and ergodic.*

Then there exists a unique solution of $\boldsymbol{\theta} \in \Theta$ *which maximizes the quasi likelihood function for n sufficiently large.*

If the Assumption 1 is satisfied, then the conditional second moment of $\mathbf{x_t}$, H_t, is identifiable in the quasi-likelihood function. Suppose that $\boldsymbol{\theta}_0$ is the true value of the parameters and H_t is identifiable, then $\mathbb{E}(L_n(\boldsymbol{\theta}_0)) > \mathbb{E}(L_n(\boldsymbol{\theta}))$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. If the time series $\mathbf{x_t}$ is ergodic and stationary, there will be a unique solution of $\boldsymbol{\theta}_0$ in the parameter space Θ which maximize the likelihood function when the sample size n is sufficient large.

12.4 Numeric Examples

12.4.1 Model Modification

To reduce the number of parameters and simplify the model, the contributions from each individual stock to the common risk indicator $\sigma^2_{0,t}$ can be assumed equal, $\beta_{01} = \beta_{02} = \cdots = \beta_{0m} = \beta_0$. In this case, the number of parameters in $\sigma^2_{0,t}$ is changed to 2 from $m + 1$ and the last line in Eq. 12.5 become

$$\sigma^2_{0,t} = \omega_0 + \beta_0(\sigma^2_{1,t-1} + \cdots + \sigma^2_{m,t-1}).$$

The g function presented in this section is chosen as a piecewise function which defined as

$$g(x) = \begin{cases} x & |x| < 0.01 \\ 0.01 & |x| >= 0.01 \end{cases}$$

The effect of observed data will be bounded within 10 % once the observed data reaches extreme large values (larger than 10 %). If the daily log return of a stock exceeds 10 % in real world, we would consider to do more research on the stock since it is unusual.

12.4.2 Real Data Analysis

A bivariate example is shown in this subsection which is based on the centered log returns of two equity series (two stocks in New York Stock Exchange: the International Bushiness Machines Corporation (IBM) and the Bank of American (BAC) from 1995 to 2007 (Fig. 12.1)). The conditions for stationarity and ergodicity are not solved yet. The ergodicity of the process could be partly verified by numeric results while the stationarity is commonly assumed in financial log returns. The default searching parameter space is chosen to be $\Theta = [-1, 1] \times [0, 1]^7$ and the numeric checks are set to verify the positive definite constraints on H_t and R matrices.

A numeric study called parametric bootstrap method (or Monte Carlo simulation) is used to test the asymptotic normality of the MLE estimator.

The histograms of the estimates in Figs. 12.2 and 12.3 were well shaped as normal distributions which verifies the asymptotic normality of the MLE in this model by using the empirical study.

The horizontal lines in Fig. 12.4 show some big events in the global stock markets over that time period. The 1997 mini crash in the global market was caused by the economic crisis in Asia on Oct 27, 1997. The time period between two solid lines was October, 1999 to October, 2001 where the Internet bubble burst. The last line was the peak before financial crisis on Oct 9, 2007. The conditional variances were significantly different in some time points. During the 1997 mini crash, the estimated conditional variances in DCC-GARCH model are different from the ones in the common risk model. The conditional variance of IBM was high while the conditional variance of BAC was relative low from DCC-GARCH model. However,

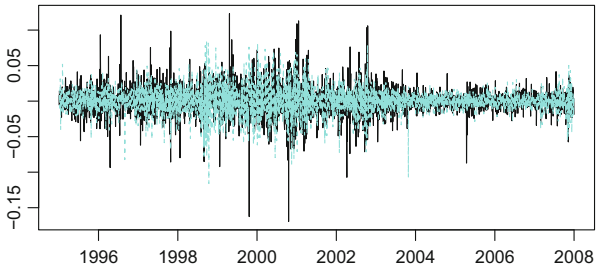

Fig. 12.1 Centered log returns of IBM and BAC from Jan. 1, 1995 to Dec. 31, 2007. The *solid black line* represents the centered log returns of IBM and the *cyan dash line* represents the centered log returns of BAC

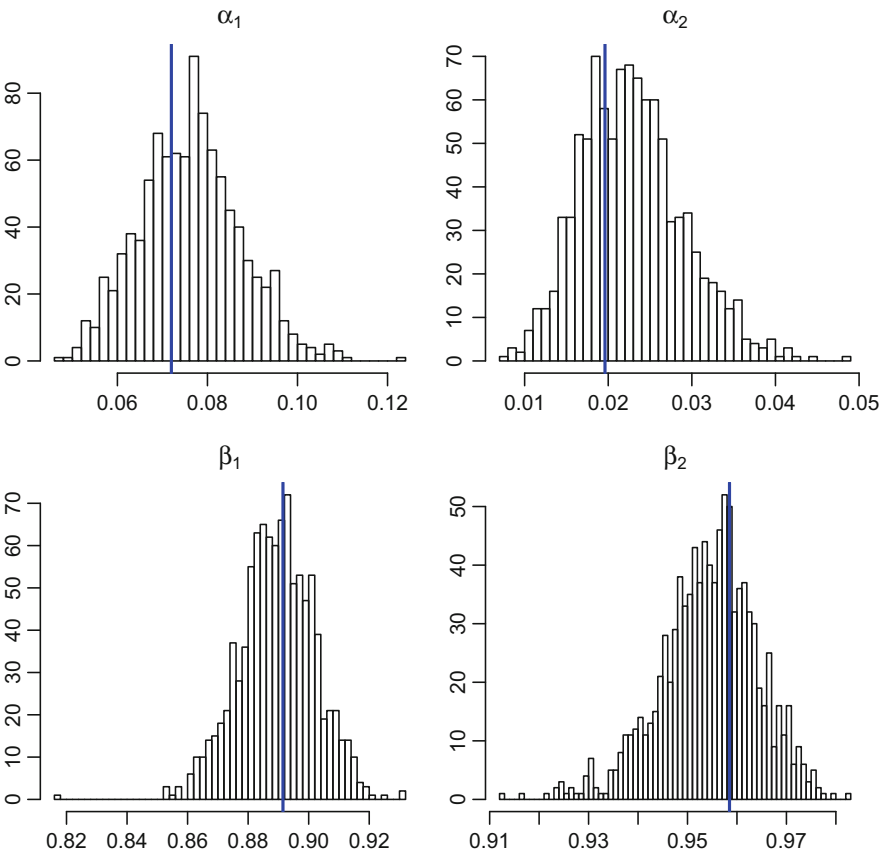

Fig. 12.2 The histogram of 1000 parameter estimates from the Monte Carlo Simulations (α_1, α_2, β_1, β_2)

the conditional covariance of both log returns were quite high from the common risk model. It is a difficult task to tell which model fits the data better since the main usages of these models are all based on the conditional volatilities or the conditional correlations (Fig. 12.5).

Denote the conditional variance estimating from model 1 by V_{mod1} in the following equation. Define a variable to measure the relative difference of the estimated conditional variance between two models.

$$\text{Relative difference} = \frac{\frac{1}{n}\sum_1^n |V_{\text{mod1}} - V_{\text{mod2}}|}{\frac{1}{n}\sum_1^n V_{\text{mod2}}}$$

Table 12.1 is not a symmetric table since the elements in the symmetric locations have different denominators according to the definition formula above. The estimated conditional variance for IBM and BAC log return series from the traditional models are really close to each other while the relative differences

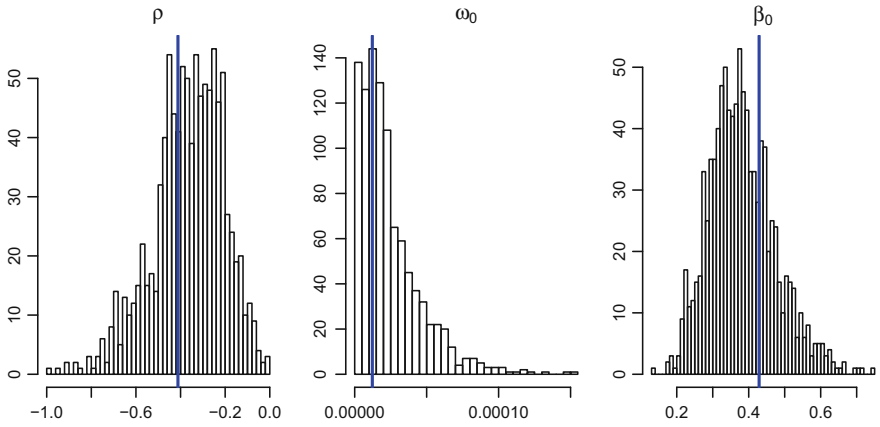

Fig. 12.3 The histogram of 1000 parameter estimates from the Monte Carlo Simulations ($\rho_{1,2}$, ω, α_0, β_0)

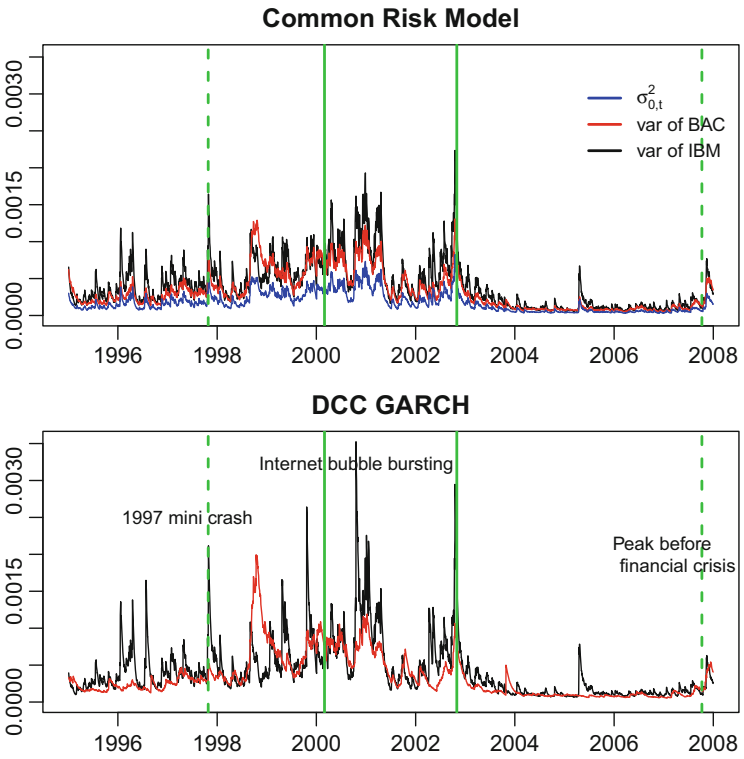

Fig. 12.4 The estimated conditional variances of IBM and BAC with $\sigma^2_{0,t}$

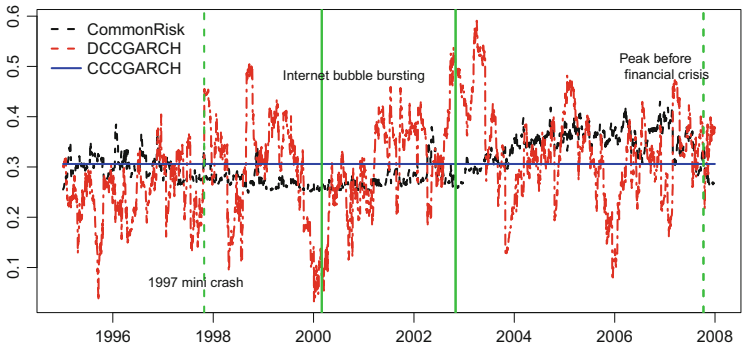

Fig. 12.5 The estimated conditional correlations between IBM and BAC

Table 12.1 The relative difference of BAC series between models

		Model2		
Model1	CommonRisk	CCCGARCH	DCCGARCH	GARCH(1,1)
CommonRisk	–	15.06 %	15.06 %	15.00 %
CCCGARCH	15.11 %	–	0.01 %	0.12 %
DCCGARCH	15.11 %	0.01 %	–	0.13 %
GARCH(1,1)	15.06 %	0.12 %	0.13 %	–

Table 12.2 The 95 % confidence interval of the estimates by using parametric bootstrap

	$\hat{\rho}_{1,2}$	$100\hat{\alpha}_1$	$100\hat{\alpha}_2$	$10\hat{\beta}_1$	$10\hat{\beta}_2$	$10^5\hat{\omega}_0$	$10\hat{\beta}_0$
'True'	−0.45	6.69	1.90	8.93	9.58	1.51	4.34
LB	−0.78	5.18	1.22	8.68	9.30	0.24	2.34
UB	−0.12	9.41	3.74	9.13	9.71	7.47	6.04

between our new model and other models are large. It is worth to build up such a complicate model since it will change the investment strategy dramatically.

12.4.3 Numeric Ergodicity Study

This example demonstrates the ergodicity and the long term behavior of the model. The data were simulated from the 'True' parameter values in Table 12.2. The plots illustrates the behavior of log returns from two common risk models (denote by M_1 and M_2) starting from different initial $\boldsymbol{\sigma_0}$'s. Denote the log return of the first simulated bivariate common risk model M_1 by (x_1, x_2) and the initial value $(\sigma_{1,0}, \sigma_{2,0}, \sigma_{0,0}, x_{1,0}, x_{2,0})$ in this model is $(0.020, 0.018, 0.013, 0.0079, 0.0076)$. The log returns simulated from M_2 were denoted by (y_1, y_2) and the initial value of M_2 is $(0.01, 0.01, 0.01, 0.009, 0.009)$.

In Figs. 12.6 and 12.7, we can see that the effect of the starting volatilities vanishes after a long enough bursting time period.

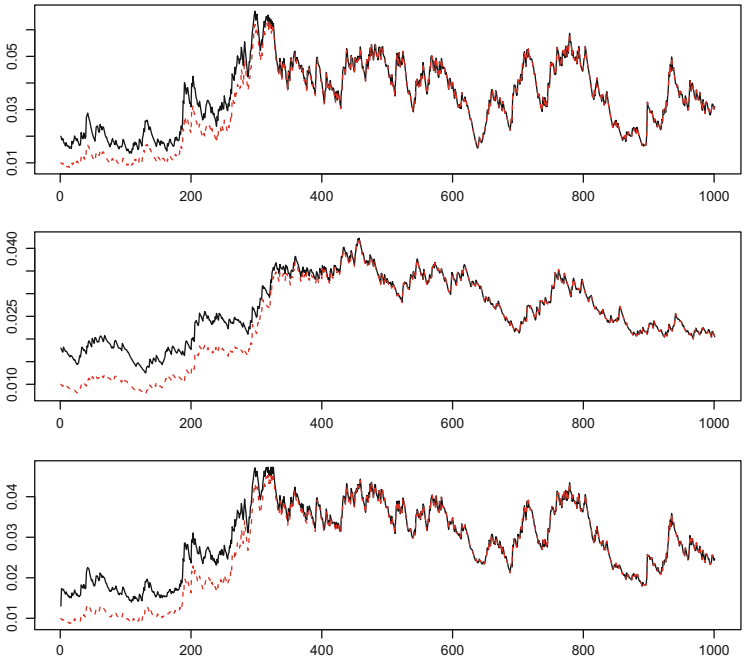

Fig. 12.6 The simulated σ's from two groups of initial values $M1$ and $M2$: the *upper plot* is $\sigma_{1,t}$, the *middle plot* is $\sigma_{2,t}$, the *bottom plot* is $\sigma_{0,t}$. The *solid black lines* represents the simulated values from $M1$ while the *red dash line* shows the simulated values from $M2$

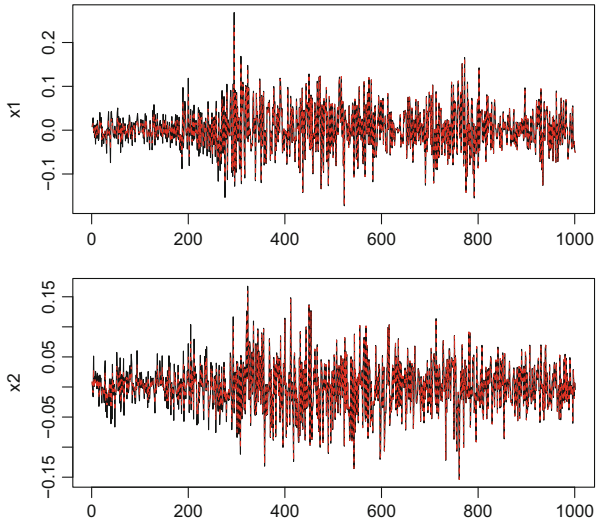

Fig. 12.7 The simulated bivariate log returns from two different initial values $M1$ and $M2$: the simulated path of $x_{1,t}$ is shown in the *upper plot* and the simulated path of $x_{2,t}$ is in the *lower plot*. The *solid black lines* represents the simulated values from $M1$ while the *red dash line* shows the simulated values from $M2$

References

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654
Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econ 31:307–327
Bollerslev T (1990) Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. Rev Econ Stat 72:498–505
Bollerslev T, Engle RF, Wooldridge JM (1988) Capital asset pricing model with time-varying covariances. J Polit Econ 96:116–131
Burmeister E, Wall KD (1986) The arbitrage pricing theory and macroeconomic factor measures. Financ Rev 21:1–20
Carr P, Wu L (2009) Variance risk premiums. Rev Financ Stud 22:1311–1341
Christie AA (1982) The stochastic behavior of common stock variances. Value, leverage and interest rate effects. J Financ Econ 10:407–432
Duffie D, Pan J, Singleton K (2000) Transform analysis and asset pricing for affine jump-diffusions. Econometrica 68:1343–1376
Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
Engle RF (2002) Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. J Bus Econ Stat 20:339–350
Engle RF, Ng VK, Rothschild M (1990) Asset pricing with a factor-arch covariance structure. Empirical estimates for treasury bills. J Econ 45:213–237
Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. J Financ Econ 33:3–56
Fama EF, French KR (2015) A five-factor asset pricing model. J Financ Econ 116:1–22
Girardi G, Tolga Ergun A (2013) Systemic risk measurement: multivariate GARCH estimation of CoVaR. J Bank Financ 37:3169–3180
Merton RC (1973) An intertemporal capital asset pricing model. Econometrica 41:867–887
Santos AAP, Moura GV (2014) Dynamic factor multivariate GARCH model. Comput Stat Data Anal 76:606–617
Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19:425–442
So MKP, Yip IWH (2012) Multivariate GARCH models with correlation clustering. J Forecast 31:443–468
Tankov P, Tauchen G (2011) Volatility jumps. J Bus Econ Stat 29:356–371
Treynor JL (2008) Treynor on institutional investing. John Wiley & Sons, Hoboken
Tse YK, Tsui AKC (2002) A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. J Bus Econ Stat 20:351–362

Index

D.-G. (Din) Chen et al. (eds.), *Advanced Statistical Methods in Data Science*,
ICSA Book Series in Statistics, DOI 10.1007/978-981-10-2594-5