ABSTRACT

ZHANG, XIANG. Contributions to Statistical Methods for High Dimensional and Dependent Data. (Under the direction of Alyson Wilson and Lexin Li.)

In this thesis, we develop three new statistical methods for high dimensional and dependent data. The three methods are motivated by three independent projects.

In the first project, we investigate variable selection for support vector machines for high dimensional data. A general class of non-convex penalized support vector machines is proposed. We show that one of the local solutions to the non-convex penalized support vector machines is the oracle estimator. This is the first variable selection consistency result for support vector machines in high dimensions. We also present an algorithm with provable global convergence to the oracle estimator. Our proof techniques are novel and do not require the differentiability of the loss function, which extend the existing results in the literature where the loss function is restricted to be a smooth function.

In the second project, we study system reliability and component importance for dependent systems. We establish a unified and general framework to characterize the influence of a dependence structure on system reliability and component importance. Our results are based on recent developments for copula theory with discrete marginal distributions. We reveal the connections of system reliability and component importance under dependence to some well-known principles under independence assumption. We also extend our results to multi-state system. Our derived results are further demonstrated using a Gaussian copula, and we show that the effects of dependence under a Gaussian copula have simple interpretations.

In the third project, we propose a new method to conduct regression on longitudinal imaging data. The proposed method integrates tensor decomposition with generalized estimating equations. We exploit a low-rank tensor decomposition to reduce the high dimensionality of image covariates and use generalized estimating equations to capture the temporal dependence in the longitudinal data. This new approach is shown to possess desirable theoretical properties.

We also provide the first rank selection consistency result in the literature under the framework of tensor regression.

Contributions to Statistical Methods for High Dimensional and Dependent Data

by
Xiang Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

Alyson Wilson Co-chair of Advisory Committee	Lexin Li Co-chair of Advisory Committee
Yichao Wu	Ralph Smith

Gregory Hicks

DEDICATION

To my family and friends.

BIOGRAPHY

The author received his double Bachelor's degrees in Economics and Statistics from Peking University in 2011. After that, he moved to Raleigh from Beijing, China to continue his graduate studies. In 2013, he received his Master's degree in Statistics from North Carolina State University. He is currently a Ph.D. candidate in Statistics at North Carolina State University.

ACKNOWLEDGEMENTS

My five years at North Carolina State University has been one of the most pleasant times in my life. I want to thank all of the people who have made this happen.

I want to thank my advisor Alyson Wilson for her constant support and help for my research in reliability modeling. From her I learned to start the problem with simple ideas and keep building up step by step. She is patient when I make mistakes, and her words are always encouraging. I am not only grateful for her funding to support my research, but also her continuous guidance that has to lead to this thesis.

I owe my special thanks to my co-advisor Lexin Li for leading me into the world of imaging and network data. He is always willing to walk through the details with me, and his insights have reshaped my understandings of many statistical problems. The time we spent reading through the literature together is one of the most enjoyable times I had in my research.

I want to thank Yichao Wu for opening the door of the first research project in my life for me. Without his faith in me I couldn't have made it this far.

I owe my thanks to Ralph Smith and Gregory Hicks for serving on my committee. Their feedback is important to my research. I also want to thank Hua Zhou and Arnab Maity for their helpful questions and comments during my oral exam.

The Department of Statistics is a great place. I want to thank Donald Martin, Howard Bondell, Eric Laber, and Eric Chi for their helpful advice in my study. I also want to thank all the department staff for always being there to make thinks work.

I learned many things from my fellow students. I owe Yichi Zhang for answering so many questions for me when I was stuck. I want to thank especially Weining Shen for all the invaluable advice about my PhD studies. I want to thank Shikai Luo, Zhou Li, Zhongkai Liu, Ailin Fan, Yan Zhang, Hao Hu, and Teng Zhang for making the study at here a nice thing throughout the years.

I want to thank the Western Wake Bible Chapel, for it has been a tremendous blessing

in my life. I am grateful for Damon Amato, Richard Forth, Ted Davis, Mark Ferguson, Anil Samuel, Jonathan Peck, Leslie Ratliff, Dan Crompton, Lijoy Samuel and Abdallah Shammas, for showing their love and shedding the light in my life.

Much credit goes to my parents. They may know nothing about my research projects, but they have contributed the most to this thesis. Without their support for me to study abroad, nothing could have been achieved.

Lastly, I owe my greatest thanks to Alison Sihan Wu. She makes every day the best day of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 High Dimensional and Dependent Data Overview

The central theme of this thesis is the development of new statistical methods for certain high dimensional and dependent data problems. The proposed methods in this thesis can be viewed as extensions of some traditional statistical approaches to specifically meet the challenges from high dimensional and dependent data. We first present an overview of the data that are considered as high dimensional and dependent in this thesis.

By high dimensional, we mean that the number of predictors, or variables, is large, while the number of instances, or samples, is only moderate. As a result, the number of predictors is much larger than the number of instances. The direct challenge of high dimensional data is that we do not have enough information in the data to estimate a full model constructed using all the available predictors. The estimated model without variable selection would be either ill-conditioned or subject to severe overfitting. Another challenge is that a model directly built from high dimensional data is complex and difficult to interpret in general, which limits its applications in practice.

By dependent data, we mean that the samples in the data are not independently generated. Independence is one of the most common assumptions in traditional statistical approaches.

However, there are some scenarios when the independence assumption may not be realistic. A naive model that simply ignores the dependency within the data is essentially allowing model misspecification, which can result in considerable biases in the estimators.

In Section 1.2, we introduce three real-world examples. The first is an example of high dimensional data, the second is an example of dependent data, and the third is an example that contains both high dimensional and dependent data simultaneously. These three examples motivate the proposed statistical methods discussed in later chapters of this thesis.

1.2 Motivating Examples

1. MAQC-II Data

   This data is part of the MicroArray Quality Control (MAQC)-II project. The complete data is available at the GEO database with accession number GSE20194 at `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20194`. It contains 278 patient samples. Each sample is described by the expression values of 22,283 genes. For each sample, the estrogen receptor (ER) status, positive or negative, is also available. The goal of the analysis is to build a binary classifier to predict the ER status from the information contained in the genes. For the 278 patient samples, 164 patients have positive ER status and 114 patients have negative ER status.

   This is an example of high dimensional data. The number of genes is much larger than the number of instances in this data. To predict the ER status, one can consider building a classifier using all of the genes. However, given the small sample size, the full model is likely to be overfitted and thus making noisy predictions. The challenge of this data is to select a small subset of genes from all the available variables and build the classifier only using those selected genes.

2. Stockpile Test Data

   We obtained two data sets from Los Alamos National Laboratory, both based on stockpiles

of simple military systems. The two data sets represent different variants of a system with slightly different functionality. Test results on components and the full system are available. For both data, the system is constructed to be a series system, and a pass for the overall system means that all of the components performed as required. We are interested in the estimating the system reliability and identifying the most critical component to the system. The first data set consists of 169 tests on a four-component system. The second data set consists of 181 tests on a four-component system.

This is an example of dependent data. The test results of components within a system are likely to be dependent in practice. There are several possible sources of the dependence among components. The components may be subject to some common stress so that they are more likely to function or fail at the same time. The components may share loads in the system so that if one component fails, the remaining components have to share more loads and thus are more likely to fail as well. In Chapter 3, we provide more evidence to show that the independence assumption is unlikely to hold for the stockpile test data. The challenge of this data is to take into account the dependence among components in modeling the system reliability and component importance.

3. Longitudinal Imaging Data

   This data is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). It consists of 88 mild cognitive impairment (MCI) subjects with longitudinal magnetic resonance imaging (MRI) images of white matter. These MRI images are taken at baseline, 6-months, 12-months, 18-months and 24-months for each subject. After some standard preprocessing, each MRI image is a three dimensional array with size $32 \times 32 \times 32$. For each subject, the Mini Mental State Examination (MMSE) score is also recored at each time. The MMSE measures the orientation to time and place, the immediate and delayed recall of three words, attention and calculations, language, and visuoconstructional functions. The goal is to predict the MMSE scores based on the imaging data, which is potentially

useful for monitoring disease progression.

This is an example of both high dimensional and dependent data. If all the pixels in the MRI image are used as predictors, the total number of predictors is $32^3 = 32,768$, which is much larger than the available sample size $88 \times 5 = 440$. Therefore, this data is high dimensional by our definitions. This data is also longitudinal, meaning that there are repeated measures of the same subject across different time points. In this example, the MMSE scores of the same subject across five time points are dependent. Two challenges arise in the analysis of this data. The first challenge of this data is to find a parsimonious model for the high dimensional image covariates. Another challenge is to capture the temporal dependence in the data.

1.3 Plan of Dissertation

This thesis consists of three projects, motivated by the data examples in Section 1.2. The first project provides a new statistical approach to address the high dimensionality in the MAQC-II data. This project is presented in Chapter 2. The second project proposes a copula approach to capture the dependence in the stockpile test data. This project is discussed in detail in Chapter 3. The third project presents a new method to analyze high dimensional and dependent longitudinal imaging data. We investigate this method in depth in Chapter 4. Summary of the contributions and extensions of the three projects are discussed in Chapter 5.

1. First Project (Chapter 2)

   This project is motivated by the MAQC-II data. We establish a unified theory for a general class of non-convex penalized support vector machines. We study its theoretical properties using tools in asymptotic statistics. We show that our proposed method possesses desired oracle properties even when the dimensionality of the variables is much larger than the sample size. Simulation studies and an analysis of the MAQC-II data are implemented and provide supportive evidence.

2. Second Project (Chapter 3)

   This project is motivated by the stockpile test data. We characterize the influence of dependence structures on system reliability and component importance in coherent systems with discrete marginal distributions. The main tool we use is copula theory. We also extend our results to more general coherent multi-state system. We demonstrate the applications of our derived results using Gaussian copulas, which yield simple interpretations. We conduct simulations and analyze two real-world examples based on the stockpile test data to demonstrate the advantages of the copula model over a naive independence model in estimating system reliability and component importance.

3. Third Project (Chapter 4)

   This project is motivated by the longitudinal imaging data. We propose a new approach for longitudinal imaging analysis. The main tools we use are generalized estimating equations and tensor regression. This approach accounts for both the intra-subject correlation and the high dimensional image covariates. We provide a scalable estimation algorithm and establish asymptotic properties. We demonstrate the proposed method on both simulated real data from ADNI to predict the MMSE score from the MRI image covariates.

Chapter 2

Variable Selection for Support Vector Machines

This chapter is organized as follows. Section 2.1 gives the introduction and the setup of the problem. In Section 2.2, we introduce the proposed method. Section 2.3 contains the theoretical studies, followed by simulation studies in Section 2.4 and results on real world MAQC-II data in Section 2.5. The technical proofs are presented in Section 2.6.

2.1 Introduction

We consider the Support Vector Machine (SVM, Vapnik, 1996) for the MAQC-II data. SVM is a powerful binary classification tool with high accuracy and great flexibility. It has achieved success in many applications. In binary classification, we are typically given a random sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ from an unknown population distribution $P(\boldsymbol{X}, Y)$. Here $Y_i \in \{1, -1\}$ denotes the categorical label and $\boldsymbol{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^\intercal = (X_{i0}, (\boldsymbol{X}_i^*)^\intercal)^\intercal$ denotes the input covariates with $X_{i0} = 1$ corresponding to the intercept term. The goal is to estimate a classification rule that can be used to predict output labels for future observations with input covariates only. With potentially varying misclassification cost specified by weight $W_i = w$ if $Y_i = 1$ and $W_i = 1 - w$

if $Y_i = -1$ for some $0 < w < 1$, the linear weighted support vector machine (WSVM, Lin et al., 2002) estimates the classification boundary by solving

$$\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\beta})_+ + \lambda (\boldsymbol{\beta}^*)^{\mathsf{T}} \boldsymbol{\beta}^*,$$

where $(1-u)_+ = \max\{1-u, 0\}$ denotes the hinge loss, $\lambda > 0$ is a regularization parameter, and $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}^*)^{\mathsf{T}})^{\mathsf{T}}$ with $\boldsymbol{\beta}^* = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathsf{T}}$. The standard SVM is a special case of the WSVM with weight parameter $w = 0.5$. In this thesis, we consider the WSVM for more generality.

One drawback of the standard SVM is that its performance can be adversely affected if many redundant variables are included in building the decision rule (Friedman et al., 2001). In general, the corresponding decision rule, $\text{sign}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta})$, uses all covariates and is not capable of selecting relevant covariates. Classification using all features has been shown to be as poor as random guessing due to noise accumulation in high dimensional space (Fan and Fan, 2008). For the MAQC-II data, it is known that the only a subset of genes are informative to the ER status. Simply using all the available genes without variable selection does not give the optimal prediction accuracy, see the evidence in Section 2.5.

Many methods have been proposed to remedy this problem, such as the recursive feature elimination suggested by Guyon et al. (2002). In particular, superior performance can be achieved with a unified method, namely achieving variable selection and prediction simultaneously (Fan and Li, 2001) by using an appropriate sparsity penalty. It is well known that the standard SVM can fit in the regularization framework of *loss* + *penalty* using the hinge loss and L_2 penalty. Based on this, several attempts have been made to achieve variable selection for the SVM by replacing the L_2 penalty with other forms of penalty. Bradley and Mangasarian (1998), Zhu et al. (2004), and Wegkamp and Yuan (2011) considered the L_1-penalized SVM; Zou and Yuan (2008) proposed to use the F_∞-norm SVM to select groups of predictors; Wang et al. (2006) and Wang et al. (2007) suggested the elastic net penalty for the SVM; Zou (2007) proposed to penalize the SVM with the adaptive LASSO penalty; Zhang et al. (2006), Becker

et al. (2011) and Park et al. (2012) studied the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001)-penalized SVM. Recently Park et al. (2012) studied the oracle property of the SCAD-penalized SVM with a fixed number of predictors. Yet, to the best of our knowledge, the theory of variable selection consistency of sparse SVMs in high dimensions or ultra-high dimensions (Fan and Lv, 2008) has not been studied so far.

2.2 Methodology

In this section we propose a general class of non-convex penalized SVMs that can achieve variable selection and prediction simultaneously. As we will show, our proposed method is the first one that possesses variable selection consistency in high dimensions.

We begin with the basic setup and notation. Consider the population linear weighted hinge loss $\mathbb{E}\{W(1-Y\boldsymbol{X}^\top\boldsymbol{\beta})_+\}$. Let $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \ldots, \beta_{0p})^\top = (\beta_{00}, (\boldsymbol{\beta}_0^*)^\top)^\top$ denote the true parameter value, which is defined as the minimizer of the population weighted hinge loss. Namely

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathbb{E}\{W(1 - Y\boldsymbol{X}^\top\boldsymbol{\beta})_+\}. \tag{2.1}$$

The number of covariates $p = p_n$ is allowed to increase with the sample size n. It is even possible that p_n is much larger than n. In this project we assume the true parameter $\boldsymbol{\beta}_0$ to be sparse. Let $A = \{1 \leq j \leq p_n; \beta_{0j} \neq 0\}$ be the index set of the nonzero coefficients. Let $q = q_n = |A|$ be the cardinality of set A, which is also allowed to increase with n. Without loss of generality, we assume that the last $p_n - q_n$ components of $\boldsymbol{\beta}_0$ are zero. That is, $\boldsymbol{\beta}_0^\top = (\boldsymbol{\beta}_{01}^\top, \mathbf{0}^\top)$. Correspondingly, we write $\boldsymbol{X}_i^\top = (\boldsymbol{Z}_i^\top, \boldsymbol{R}_i^\top)$, where $\boldsymbol{Z}_i = (X_{i0}, X_{i1}, \ldots, X_{iq})^\top = (1, (\boldsymbol{Z}_i^*)^\top)^\top$ and $\boldsymbol{R}_i = (X_{i[q+1]}, \ldots, X_{ip})^\top$. Further we denote π_+ (resp. π_-) to be the marginal probability of the label $Y = +1$ (resp. -1).

We propose the non-convex penalized hinge loss objective function:

$$Q(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} W_i(1 - Y_i\boldsymbol{X}_i^\top\boldsymbol{\beta})_+ + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \tag{2.2}$$

where $p_{\lambda_n}(\cdot)$ is a symmetric penalty function with tuning parameter λ_n. Let $p'_{\lambda_n}(t)$ be the derivative of $p_{\lambda_n}(t)$ with respect to t. We consider a general class of non-convex penalties that satisfy the following conditions.

(Condition 1) The symmetric penalty $p_{\lambda_n}(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $p'_{\lambda_n}(t)$ on $(0, +\infty)$ and $p_{\lambda_n}(0) = 0$.

(Condition 2) There exists $a > 1$ such that $\lim_{t \to 0+} p'_{\lambda_n}(t) = \lambda_n$, $p'_{\lambda_n}(t) \geq \lambda_n - t/a$ for $0 < t < a\lambda$ and $p'_{\lambda_n}(t) = 0$ for $t \geq a\lambda$.

The motivation for such a non-convex penalty is that the convex L_1 penalty lacks the oracle property due to the overpenalization of large coefficients in the selected model. Consequently it is undesirable to use the L_1 penalty when the purpose of the data analysis is to select the relevant covariates among potentially high dimensional candidates in classification. Note that p, q, λ and other related quantities are allowed to depend on n, and we suppress the subscript n whenever there is no confusion.

Two commonly used non-convex penalties that satisfy Conditions 1 and 2 are the SCAD and MCP penalties. The SCAD penalty (Fan and Li, 2001) is defined by

$$p_\lambda(|\beta|) = \lambda|\beta| I(0 \leq |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \leq |\beta| \leq a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda)$$

for some $a > 2$. The MCP (Zhang, 2010) is defined by

$$p_\lambda(|\beta|) = \lambda(|\beta| - \frac{\beta^2}{2a\lambda}) I(0 \leq |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \geq a\lambda) \text{ for some } a > 1.$$

By acting as if the true sparsity structure is known in advance, the oracle estimator is defined as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^\top, \mathbf{0}^\top)^\top$, where

$$\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} n^{-1} \sum_{i=1}^n W_i (1 - Y_i \boldsymbol{Z}_i^\top \boldsymbol{\beta}_1)_+. \tag{2.3}$$

In Section 2.3, we will present the main results that under some regularity conditions, our

proposed method can find an estimator converge to the desirable oracle estimator with high probability.

2.3 Theory

2.3.1 Regularity Conditions

To facilitate our theoretical analysis, we introduce the gradient vector and Hessian matrix of the population linear weighted hinge loss. Let $L(\boldsymbol{\beta}_1) = \mathbb{E}\{W(1 - Y\boldsymbol{Z}^\top\boldsymbol{\beta}_1)_+\}$ be the population linear weighted hinge loss using only relevant covariates. Define $S(\boldsymbol{\beta}_1) = (S(\boldsymbol{\beta}_1)_j)$ to be the $(q_n + 1)$-dimension vector given by

$$S(\boldsymbol{\beta}_1) = -\mathbb{E}\{I(1 - Y\boldsymbol{Z}^\top\boldsymbol{\beta}_1 \geq 0)WY\boldsymbol{Z}\},$$

where $I(\cdot)$ denotes the indicator function. Also define $H(\boldsymbol{\beta}_1) = (H(\boldsymbol{\beta}_1)_{jk})$ to be the $(q_n + 1) \times (q_n + 1)$ matrix given by

$$H(\boldsymbol{\beta}_1) = \mathbb{E}\{\delta(1 - Y\boldsymbol{Z}^\top\boldsymbol{\beta}_1)W\boldsymbol{Z}\boldsymbol{Z}^\top\},$$

where $\delta(\cdot)$ denotes the Dirac delta function. It can be shown that if well-defined, $S(\boldsymbol{\beta}_1)$ and $H(\boldsymbol{\beta}_1)$ can be considered to be the gradient vector and Hessian matrix of $L(\boldsymbol{\beta}_1)$, respectively. See Lemma 2 of Koo et al. (2008) for details.

We impose the following regularity conditions:

(A1) The densities of $\boldsymbol{Z}^*$ given $Y = +1$ and $Y = -1$ are continuous and have common support in $\mathbb{R}^q$.

(A2) $\mathbb{E}[X_j^2] < \infty$ for $1 \leq j \leq q$.

(A3) The true parameter $\boldsymbol{\beta}_0$ is unique and a nonzero vector.

(A4) $q_n = O(n^{c_1})$, namely $\lim_{n\to\infty} q_n/n^{c_1} < \infty$, for some $0 \leq c_1 < 1/2$.

(A5) There exists a constant $M_1 > 0$ such that $\lambda_{max}(n^{-1}\boldsymbol{X}_A^\top\boldsymbol{X}_A) \leq M_1$, where $\boldsymbol{X}_A$ is the first $q_n + 1$ columns of the design matrix and λ_{max} denotes the largest eigenvalue. It is further

assumed that $\max_{1\le i\le n} ||\boldsymbol{Z}_i|| = O_p(\sqrt{q_n}\log(n))$, $(\boldsymbol{Z}_i, Y_i)$ are in general position (Koenker, 2005, sect. 2.2), X_{ij} are sub-Gaussian random variables for $1 \le i \le n, q_n + 1 \le j \le p_n$.

(A6) $\lambda_{min}(H(\boldsymbol{\beta}_{01})) \ge M_2$ for some constant $M_2 > 0$, where λ_{min} denotes the smallest eigenvalue.

(A7) $n^{(1-c_2)/2} \min_{1\le j\le q_n} |\beta_{0j}| \ge M_3$ for some constant $M_3 > 0$ and $2c_1 < c_2 \le 1$.

(A8) Denote the conditional density of $\boldsymbol{Z}^\intercal\boldsymbol{\beta}_{01}$ given $Y = +1$ and $Y = -1$ as f and g, respectively. It is assumed that f is uniformly bounded away from 0 and ∞ in a neighborhood of 1 and g is uniformly bounded away from 0 and ∞ in a neighborhood of -1.

Conditions (A1)-(A3) and (A6) are also assumed for fixed p in Koo et al. (2008). We need these assumptions to ensure that the oracle estimator is consistent in the scenario of diverging p. Condition (A3) states that the optimal classification decision function is not constant, which is required to ensure $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ are well-defined gradient vector and Hessian matrix of the hinge loss, see Lemma 2 and Lemma 3 of Koo et al. (2008). The conditions (A4) and (A7) are common in the literature on high dimensional inference (Kim et al., 2008). More specifically, (A4) states that the divergence rate of the number of nonzero coefficients cannot be faster than root-n and (A7) simply states that the signals cannot decay too quickly. The condition on the largest eigenvalues of the design matrix in (A5) is similar to the sparse Riesz condition and also assumed in Zhang and Huang (2008), Yuan (2010) and Zhang (2010). Note that the bound on the smallest eigenvalue is not specified. The condition on the maximum norm in (A5) holds when $\boldsymbol{Z}^*$ given Y follows multivariate normal distribution. $(\boldsymbol{Z}_i, Y_i)$ are in general position if with probability one there are exactly $(q_n + 1)$ elements in $\boldsymbol{D} = \{i : 1 - Y_i\boldsymbol{Z}_i^\intercal\widehat{\boldsymbol{\beta}}_1 = 0\}$ (Koenker, 2005, sect. 2.2). The condition for general position is true with probability one w.r.t. Lebesgue measure. Condition (A8) requires that there is enough information around the nondifferentiable point of the hinge loss, similar to condition (C5) in Wang et al. (2012) for quantile regression.

For illustrative examples that satisfy all the above conditions, assume $0 < \pi_+ = 1 - \pi_- < 1$ and let the number of signals be fixed. The first example is that the conditional distributions of $\boldsymbol{X}^*$ given Y have unbounded support $\mathbb{R}^p$ with sub-Gaussian tails. It can be easily seen that the

Fisher's discriminant analysis is one special case when $\boldsymbol{X}^*$ given Y are Gaussian. Conditions (A1)-(A4) and (A7) are trivial. Condition (A5) holds by the properties of sub-Gaussian random variable. Koo et al. (2008) showed that Condition (A6) holds if the supports of the conditional densities of $\boldsymbol{Z}^*$ given Y are convex, which are naturally satisfied for $\mathbb{R}^q$. Condition (A8) is trivially satisfied by the unbounded support of the conditional distribution of $\boldsymbol{Z}^*$ given Y.

Another example is the Probit model that $\boldsymbol{X}^*$ has unbounded support $\mathbb{R}^p$ with sub-Gaussian tails and $\Pr(Y = +1|\boldsymbol{X}^*) = \Phi(\boldsymbol{X}^\top\boldsymbol{\beta})$ for some $\boldsymbol{\beta} \neq \mathbf{0}$. It can be easily checked that the conditional distributions of $\boldsymbol{X}^*$ given Y also have unbounded supports $\mathbb{R}^p$ and hence all the conditions are satisfied.

2.3.2 Oracle Property

In this subsection, we establish the theory of the oracle property for non-convex penalized SVMs; namely, the oracle estimator is one of the local minimizers of the objective function $Q(\boldsymbol{\beta})$ defined in (2.2). We start with the following lemma on the consistency of the oracle estimator, which can be viewed as an extension of the consistency result in Koo et al. (2008) to the diverging p scenario.

Lemma 2.1. *Assume that Conditions (A1)-(A7) are satisfied. The oracle estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^\top, \mathbf{0}^\top)^\top$ *satisfies* $||\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}|| = O_p(\sqrt{q_n/n})$ *when* $n \to \infty$.

Though the convexity of the non-convex penalized hinge loss objective function $Q(\boldsymbol{\beta})$ is not guaranteed, it can be written as the difference of two convex functions:

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta}), \tag{2.4}$$

where $g(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n W_i(1-Y_i\boldsymbol{X}_i^\top\boldsymbol{\beta})_+ + \lambda_n \sum_{j=1}^p |\beta_j|$ and $h(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^p |\beta_j| - \sum_{j=1}^p p_{\lambda_n}(|\beta_j|) = \sum_{j=1}^p H_{\lambda_n}(\beta_j)$. The form of $H_\lambda(\beta_j)$ depends on the penalty function. For the SCAD penalty,

we have

$$H_\lambda(\beta_j) = [(\beta_j^2 - 2\lambda|\beta_j| + \lambda^2|)/\{2(a-1)\}]I(\lambda \le |\beta_j| \le a\lambda) + \{\lambda|\beta_j| - (a+1)\lambda^2/2\}I(|\beta_j| > a\lambda),$$

while for MCP, we have $H_\lambda(\beta_j) = \{\beta_j^2/(2a)\}I(0 \le |\beta_j| < a\lambda) + (\lambda|\beta_j| - a\lambda^2/2)I(|\beta_j| \ge a\lambda)$. This decomposition is useful, as it naturally satisfies the form of the difference of convex functions (DC) algorithm (An and Tao, 2005).

To prove the oracle property of the non-convex penalized SVMs, we will use a sufficient local optimality condition for the difference convex programming first presented in Tao and An (1997). This sufficient condition is based on subgradient calculus. The subgradient can be viewed as an extension of the gradient of the smooth convex function to the non-smooth convex function. Let dom(g)=$\{\boldsymbol{x} : g(\boldsymbol{x}) < \infty\}$ be the effective domain of a convex function g. The subgradient of $g(\boldsymbol{x})$ at a point $\boldsymbol{x}_0$ is defined as $\partial g(\boldsymbol{x}_0) = \{\mathbf{t} : g(\boldsymbol{x}) \ge g(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^\intercal \mathbf{t}\}$. Note that at the non-differentiable point, the subgradient contains a collection of vectors. One can easily check that the subgradient of the hinge loss function at the oracle estimator is the collection of vectors $s(\widehat{\boldsymbol{\beta}}) = (s_0(\widehat{\boldsymbol{\beta}}), \ldots, s_p(\widehat{\boldsymbol{\beta}}))^\intercal$ with

$$s_j(\widehat{\boldsymbol{\beta}}) = -n^{-1}\sum_{i=1}^n W_i Y_i \boldsymbol{X}_{ij} I(1 - Y_i \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\beta}} > 0) - n^{-1}\sum_{i=1}^n W_i Y_i \boldsymbol{X}_{ij} v_i, \tag{2.5}$$

where $-1 \le v_i \le 0$ if $1 - Y_i X_i^\intercal \widehat{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise, $j = 0, \ldots, p$. Under some regularity conditions, we can study the asymptotic behaviors of the subgradient at the oracle estimator. The results are summarized in the following Theorem.

Theorem 2.1. *Suppose that Conditions (A1)-(A8) hold, and the tuning parameter satisfies* $\lambda = o(n^{-(1-c_2)/2})$ *and* $\log(p)q\log(n)n^{-\frac{1}{2}} = o(\lambda)$. *For the oracle estimator* $\widehat{\boldsymbol{\beta}}$, *there exists* v_i^* *which satisfies* $v_i^* = 0$ *if* $1 - Y_i X_i^\intercal \widehat{\boldsymbol{\beta}} \ne 0$ *and* $v_i^* \in [-1, 0]$ *if* $1 - Y_i X_i^\intercal \widehat{\boldsymbol{\beta}} = 0$, *such that for* $s_j(\widehat{\boldsymbol{\beta}})$

with $v_i = v_i^*$, with probability approaching one, we have

$$
\begin{aligned}
s_j(\widehat{\boldsymbol{\beta}}) = 0,\ \ j = 0, 1, \ldots, q, \\
|\hat{\beta}_j| \geq (a + \frac{1}{2})\lambda,\ \ j = 1, \ldots, q, \\
|s_j(\widehat{\boldsymbol{\beta}})| \leq \lambda, j = q+1, \ldots, p, \\
|\hat{\beta}_j| = 0, j = q+1, \ldots, p.
\end{aligned}
$$

Theorem 2.1 characterizes the subgradients of the hinge loss at the oracle estimator. It basically says that in a regular setting, with probability arbitrarily close to one, those components of the subgradients corresponding to the relevant covariates are exactly zero and those corresponding to irrelevant covariates are not far away zero.

We now present the sufficient optimality condition based on subgradient calculations. Corollary 1 of Tao and An (1997) states that if there exists a neighborhood U around the point $\boldsymbol{x}^*$ such that $\partial h(\boldsymbol{x}) \cap \partial g(\boldsymbol{x}^*) \neq \emptyset$, $\forall \boldsymbol{x} \in U \cap \mathrm{dom}(g)$, then $\boldsymbol{x}^*$ is a local minimizer of $g(\boldsymbol{x}) - h(\boldsymbol{x})$. To verify this local sufficient condition, we study the asymptotic behaviors of subgradients of the two convex functions in the aforementioned decomposition (2.4) of $Q(\boldsymbol{\beta})$. Note that, based on (2.5), the subgradient function of $g(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$ can be shown to be the following collection of vectors:

$$
\begin{aligned}
\partial g(\boldsymbol{\beta}) = & \Big\{ \boldsymbol{\xi} = (\xi_0, \ldots, \xi_p)^\intercal \in \mathcal{R}^{p+1} : \\
& \xi_j = -n^{-1} \sum_{i=1}^n W_i Y_i \boldsymbol{X}_{ij} I(1 - Y_i \boldsymbol{X}_i^\intercal \widehat{\boldsymbol{\beta}} > 0) - n^{-1} \sum_{i=1}^n W_i Y_i \boldsymbol{X}_{ij} v_i + \lambda l_j, j = 0, \ldots, p \Big\},
\end{aligned}
$$

where $l_0 = 0$, $l_j = \mathrm{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise for $1 \leq j \leq p$, and $-1 \leq v_i \leq 0$ if $1 - Y_i X_i^\intercal \widehat{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise for $1 \leq i \leq n$. Furthermore, by Condition 2 of the class of non-convex penalty functions, $\lim_{t \to 0+} H'_\lambda(t) = \lim_{t \to 0-} H'_\lambda(t) = \lambda \mathrm{sgn}(t) - \lambda \mathrm{sgn}(t) = 0$. Thus $h(\boldsymbol{\beta})$ is differentiable everywhere. Consequently the subgradient of $h(\boldsymbol{\beta})$ at point $\boldsymbol{\beta}$ is a

singleton:

$$\partial h(\boldsymbol{\beta}) = \{\boldsymbol{\mu} = (\mu_0, \ldots, \mu_p) \in \mathcal{R}^{p+1} : \mu_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, j = 0, \ldots, p\}.$$

For the class of non-convex penalty functions under consideration, $\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = 0$ for $j = 0$. For $1 \leq j \leq p$,

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = [\{\beta_j - \lambda \mathrm{sgn}(\beta_j)\}/(a-1)]I(\lambda \leq |\beta_j| \leq a\lambda) + \lambda \mathrm{sgn}(\beta_j) I(|\beta_j| > a\lambda)$$

for the SCAD penalty, and

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = (\beta_j/a) I(0 \leq |\beta_j| < a\lambda) + \lambda \mathrm{sgn}(\beta_j) I(|\beta_j| \geq a\lambda)$$

for the MCP.

Combining this with Theorem 2.1, we will prove that with probability tending to one, for any $\boldsymbol{\beta}$ in a ball in $\mathbb{R}^{p+1}$ with the center $\widehat{\boldsymbol{\beta}}$ and radius $\frac{\lambda}{2}$, there exists a subgradient $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_p)^\top \in \partial g(\widehat{\boldsymbol{\beta}})$ such that $\frac{h(\boldsymbol{\beta})}{\partial \beta_j} = \xi_j, \quad j = 0, 1, \ldots, p$. Consequently the oracle estimator $\widehat{\boldsymbol{\beta}}$ is itself a local minimizer of (2.2). This is summarized in the following theorem.

Theorem 2.2. *Assume that Conditions (A1)-(A8) hold. Let $B_n(\lambda)$ be the set of local minimizers of the objective function $Q(\boldsymbol{\beta})$ with regularization parameter λ. The oracle estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^\top, \mathbf{0}^\top)^\top$ satisfies*

$$\Pr\{\widehat{\boldsymbol{\beta}} \in B_n(\lambda)\} \to 1$$

as $n \to \infty$, if $\lambda = o(n^{-(1-c_2)/2})$, and $\log(p) q \log(n) n^{-\frac{1}{2}} = o(\lambda)$.

It can be shown that if we take $\lambda = n^{-1/2+\delta}$ for some $c_1 < \delta < c_2/2$, then the oracle property holds even for $p = o(\exp(n^{(\delta - c_1)/2}))$. Therefore, the oracle property holds for the non-convex penalized SVM even when the number of covariates grows exponentially with the sample size.

2.3.3 An Algorithm with Provable Convergence to the Oracle Estimator

Note that Theorem 2.2 indicates that one of the local minimizers possesses the oracle property. However, there can potentially be multiple local minimizers and it remains challenging to identify the oracle estimator. In the high dimensional setting, assuming that the local minimizer is unique would not be realistic.

Instead of assuming the uniqueness of solutions, we work directly on the conditions under which the oracle estimator can be identified by some numerical algorithms that solve the non-convex penalized SVM objective function. One possible algorithm is the local linear approximation (LLA) algorithm proposed by Zou and Li (2008). In LLA algorithm, we start with an initial value $\{\tilde{\boldsymbol{\beta}}^{(0)} : \tilde{\beta}_j^{(0)} = 0, j = 1, 2, \ldots, p\}$. At each step $t \geq 1$, we update by solving

$$\min_{\boldsymbol{\beta}} \{ n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \boldsymbol{X}_i^{\intercal} \boldsymbol{\beta})_+ + \sum_{j=1}^{p} p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)|\beta_j| \}, \tag{2.6}$$

where $p'_\lambda(\cdot)$ denotes the derivative of $p_\lambda(\cdot)$. Following the literature, when $\tilde{\beta}_j^{(t-1)} = 0$, we take $p'_\lambda(0)$ as $p'_\lambda(0+) = \lambda$. The LLA algorithm is an instance of the majorize-minimize (MM) algorithm and converges to a local minimizer of the non-convex objective function.

Recently LLA has been shown to be capable of identifying the oracle estimator in the setup of folded concave penalized estimation with a differentiable loss function (Wang et al., 2013; Fan et al., 2014). We generalize their results to non-differentiable loss functions, so that it can fit in the framework of the non-convex penalized SVMs. Similar to their work, the main condition required is the existence of an appropriate initial estimator inputed in the iterations of the LLA algorithm. Denote the initial estimator as $\tilde{\boldsymbol{\beta}}^{(0)}$. Intuitively, if the initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ lies in a small neighborhood of the true value $\boldsymbol{\beta}_0$, the algorithm should converge to the good local minimizer around $\boldsymbol{\beta}_0$. This localizability will be formalized in terms of L_∞ distance later. With such an appropriate initial estimator, under the aforementioned regularity conditions, one can prove that the LLA algorithm converges to the oracle estimator with probability tending to one even in ultra-high dimensions.

Let $\tilde{\boldsymbol{\beta}}^{(0)} = (\tilde{\beta}_0^{(0)}, \ldots, \tilde{\beta}_p^{(0)})^\intercal$. Consider the following events:

- $F_{n1} = \{|\tilde{\beta}_j^{(0)} - \beta_{0j}| > \lambda, \text{ for some } 1 \le j \le p\}$,
- $F_{n2} = \{|\beta_{0j}| < (a+1)\lambda, \text{ for some } 1 \le j \le q\}$,
- $F_{n3} = \{$for all subgradients $s(\widehat{\boldsymbol{\beta}}), |s_j(\widehat{\boldsymbol{\beta}})| > (1 - \frac{1}{a})\lambda$ for some $q+1 \le j \le p$ or $|s_j(\widehat{\boldsymbol{\beta}})| \neq 0$ for some $0 \le j \le q\}$,
- $F_{n4} = \{|\hat{\beta}_j| < a\lambda, \text{ for some } 1 \le j \le q\}$.

Denote the corresponding probability as $P_{ni} = \Pr(F_{ni}), i = 1, 2, 3, 4$. P_{n1} represents the localizability of the problem. When we have an appropriate initial estimator, we expect P_{n1} to converge to 0 as $n \to \infty$. P_{n2} is the probability that the true signal is too small to be detected by any method. P_{n3} describes the behavior of the subgradients at the oracle estimator. As stated in Theorem 2.1, there exists a subgradient such that its components corresponding to irrelevant variables are near 0 and those corresponding to relevant variables are exactly 0, so P_{n3} cannot be too large. P_{n4} has to do with the magnitude of the oracle estimator on relevant variables. Under regularity conditions, the oracle estimator will detect the ture signals and hence P_{n4} will be very small.

Now we provide conditions for the LLA algorithm to find the oracle estimator $\widehat{\boldsymbol{\beta}}$ in the non-convex penalized SVMs based on P_{n1}, P_{n2}, P_{n3} and P_{n4}.

Theorem 2.3. *With probability at least* $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$*, the LLA algorithm initiated by* $\tilde{\boldsymbol{\beta}}^{(0)}$ *finds the oracle estimator* $\widehat{\boldsymbol{\beta}}$ *after two iterations. Furthermore, if (A1)-(A8) hold,* $\lambda = o(n^{-(1-c_2)/2})$ *and* $\log(p)q\log(n)n^{-\frac{1}{2}} = o(\lambda)$*, then* $P_{n2} \to 0, P_{n3} \to 0$ *and* $P_{n4} \to 0$ *as* $n \to \infty$.

The first part of Theorem 2.3 provides a non-asymptotic lower bound on the probability that the LLA algorithm converges to the oracle estimator. As we will show in the proofs, if none of the events F_{ni} happen, the LLA algorithm initiated with $\tilde{\boldsymbol{\beta}}^{(0)}$ will find the oracle estimator in the first iteration, and in the second iteration it will find the oracle estimator again and thus claim convergence. Note that only a single correction is required in the first iteration and the

second iteration is needed to stop the algorithm. Therefore, the LLA algorithm can identify the oracle estimator after two iterations and this result holds generally without the Conditions (A1)-(A8).

The second part of Theorem 2.3 indicates that under Conditions (A1)-(A8), the lower bound is determined only by the limiting behavior of the initial estimator. As long as an appropriate initial estimator is available, the problem of selecting the oracle estimator from potential multiple local minimizers is addressed. Let $\widehat{\boldsymbol{\beta}}^{L_1}$ be the solution to the L_1-penalized SVM. When the initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ is taken to be $\widehat{\boldsymbol{\beta}}^{L_1}$ and the following Condition (A9) holds, by Theorem 2.3 the oracle estimator can be identified even in the ultra-high dimensional setting. The result is summarized in the following Corollary.

(A9) $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{ for some } 1 \leq j \leq p) \to 0$ as $n \to \infty$.

Corollary 2.1. *Let $\widehat{\boldsymbol{\beta}}(\lambda)$ be the solution found by the LLA algorithm initiated by $\widehat{\boldsymbol{\beta}}^{L_1}$ after two iterations. Assume the same conditions in Theorem 2.3 and (A9) hold, then*

$$\Pr\{\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}\} \to 1 \text{ as } n \to \infty.$$

In the ultra-high dimensional case, one may require more stringent conditions to guarantee (A9). For the non-convex penalized least square regression, one can use the LASSO solution (Tibshirani, 1996) as the initial estimator, and (A9) holds if one can further assume the restricted eigenvalue condition of the design matrix (Bickel et al., 2009). However, it is still largely unknown whether this conclusion also applies to the setting where both the loss and the penalty are non-differentiable. Without imposing any new regularity conditions, we next prove that in the moderately high dimensions $p = o(\sqrt{n})$, the solution to the L_1-penalized SVM satisfies (A9) under conditions quite similar to (A1)-(A8).

The following regularity conditions are modified from (A1)-(A8). Conditions (A3) (A7), and (A8) do not change.

(A1*) The densities of $\boldsymbol{X}^*$ given $Y = +1$ and $Y = -1$ are continuous and have a common

support in $\mathbb{R}^p$.

(A2*) $\mathbb{E}[X_j^2] < \infty$ for $1 \leq j \leq p$.

(A4*) $p_n = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$.

(A5*) There exists a constant $M_1 > 0$ such that $\lambda_{max}(n^{-1}\boldsymbol{X}^\intercal \boldsymbol{X}) \leq M_1$. It is further assumed that $\max_{1\leq i\leq n} ||\boldsymbol{X}_i|| = O_p(\sqrt{p_n}\log n)$, $(\boldsymbol{X}_i, Y_i)$ are in general position (Koenker, 2005, sect. 2.2), X_{ij} are sub-Gaussian random variables for $1 \leq i \leq n, q_n + 1 \leq j \leq p_n$.

(A6*) $\lambda_{min}(H(\boldsymbol{\beta}_0)) \geq M_3$ for some constant $M_3 > 0$.

Under the new regularity conditions, we can conclude that the solution to the L_1-penalized SVM is an appropriate initial estimator. Combined with Theorem 2.3, the LLA algorithm initiated with a zero vector can identify the oracle estimator with one more iteration. The results are summarized in the following Theorem.

Theorem 2.4. *Assume $\widehat{\boldsymbol{\beta}}^{L_1}$ is the solution to the L_1-penalized SVM with tuning parameter c_n. If the modified conditions hold, $\lambda = o(n^{-(1-c_2)/2})$, $p\log(n)n^{-1/2} = o(\lambda)$ and $c_n = o(n^{-1/2})$, then we have $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda$, for some $1 \leq j \leq p) \to 0$ as $n \to \infty$. Further, the LLA algorithm initiated by $\widehat{\boldsymbol{\beta}}^{L_1}$ finds the oracle estimator in two iterations with probability tending to one. That is, $\Pr\{\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}\} \to 1$ as $n \to \infty$.*

Note that Theorem 2.4 can guarantee that the LLA algorithm initialized by the $\widehat{\boldsymbol{\beta}}^{L_1}$ identifies the oracle estimator with high probability only when $p = o(\sqrt{n})$. However, our empirical studies suggest that even for cases with p much larger than n, the LLA algorithm initiated by $\widehat{\boldsymbol{\beta}}^{L_1}$ usually converges within two iterations and the identified local minimizer has acceptable performance.

2.4 Simulations

We carry out simulation studies to evaluate the finite-sample performance of the non-convex penalized SVMs. We compare the performance of SCAD-penalized SVM, MCP-penalized SVM, standard L_2 SVM, L_1-penalized SVM, adaptively weighted L_1-penalized SVM (Zou, 2007) and

hybrid Huberized SVM (Wang et al., 2007) (denoted by SCAD-svm, MCP-svm, L_2-svm, L_1-svm, Adap L_1-svm, and Hybrid-svm, respectively) with weight parameter $w = 0.5$. The main interest here is the ability to identify the relevant covariates and the control of test error when $p > n$.

For the choice of tuning parameter λ, Claeskens et al. (2008) suggested the SVM information criterion (SVMIC). For a subset S of $\{1, 2, \ldots, p\}$, the SVMIC is defined as

$$\text{SVMIC}(S) = \sum_{i=1}^{n} \xi_i + \log(n)|S|,$$

where $|S|$ is the cardinality of S and ξ_i, $i = 1, 2, \cdots, n$ denote the corresponding optimal slack variables. This criterion directly follows the spirit of the Bayesian information criterion (BIC) by Schwarz (1978). Chen and Chen (2008) showed that BIC can be too liberal when the model space is large and proposed the extended BIC (EBIC):

$$\text{EBIC}_\gamma(S) = -2 \log \text{Likelihood} + \log(n)|S| + 2\gamma \binom{p}{|S|}, \ 0 \le \gamma \le 1.$$

By combining these ideas, we suggest the SVM-extend BIC (SVMIC_γ)

$$\text{SVMI}C_\gamma(S) = \sum_{i=1}^{n} 2W_i\xi_i + \log(n)|S| + 2\gamma \binom{p}{|S|}, \ 0 \le \gamma \le 1.$$

Note that SVMIC_γ reduces to SVMIC when $\gamma = 0$ and $w = 0.5$. We use $\gamma = 0.5$ as suggested by Chen and Chen (2008) and choose the λ that minimizes SVMIC_γ.

We consider two data generation processes. The first, adapted from Park et al. (2012), is essentially a standard linear discriminant analysis (LDA) setting. The second is related to probit regression.

- Model 1: $\Pr(Y = 1) = \Pr(Y = -1) = 0.5$, $\boldsymbol{X}^*|(Y = 1) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{X}^*|(Y = -1) \sim MN(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $q = 5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^\intercal \in \mathbb{R}^p$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ii} = 1$ for $i = 1, 2, \cdots, p$ and $\sigma_{ij} = \rho = -0.2$ for $1 \le i \ne j \le q$. The Bayes rule

is sign($2.67X_1$+$2.83X_2$+$3X_3$+$3.17X_4$+$3.33X_5$) with Bayes error: 6.3%.

- Model 2: $\boldsymbol{X}^* \sim MN(\mathbf{0}_p, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ii} = 1$ for $i = 1, 2, \cdots, p$ and $\sigma_{ij} = 0.4^{|i-j|}$ for $1 \leq i \neq j \leq p$, $\Pr(Y = 1|\boldsymbol{X}^*) = \Phi((\boldsymbol{X}^*)^\top \boldsymbol{\beta}^*)$ where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $\boldsymbol{\beta}^* = (1.1, 1.1, 1.1, 1.1, 0, \ldots, 0)^\top$, $q = 4$. The Bayes rule is sign(X_1+X_2+X_3+X_4) with Bayes error 10.4%.

We consider different (n, p) settings for each data generation process with p much larger than n. Similarly to Mazumder et al. (2011), an independent tuning dataset of size $10n$ is generated to tune any regularization parameter for all methods by minimizing the estimated prediction error calculated over the tuning dataset. We also report the performance of the SCAD- and MCP-penalized SVMs using SVMIC$_\gamma$ to select the tuning parameter λ. Notice that tuning by a large independent tuning dataset of $10n$ approximates the ideal "population tuning", which is usually not available in practice. By giving all the other methods the best possible tuning, we are controlling the effect of tuning parameter selection and conservative about the performance of the non-convex penalized SVMs tuned by SVMIC$_\gamma$. As we will see later, the results of SCAD- and MCP-penalized SVMs using the independent tuning dataset are slightly better than the corresponding results using SVMIC$_\gamma$ tuning; and all other methods have no ability to select the correct model exactly, even with an unrealistically good tuning parameter. The range of λ is $\{2^{-6}, \ldots, 2^3\}$. We use a=3.7 for the SCAD penalty and $a = 3$ for the MCP as suggested in the literature. We generate an independent test dataset of size n to report the estimated test error. The columns "Signal" and "Noise" summarize the average number of selected relevant and irrelevant covariates, respectively. The numbers in the "Correct" column summarize the percentages of selecting the exactly true model over replications.

Table 2.1 shows the results of Model 1 for different (n, p) settings. The numbers in parentheses are the corresponding standard errors based on 100 replications. When tuned by using an independent tuning set of size $10n$, both SCAD- and MCP-penalized SVMs identify more relevant variables than any other methods and they also reduce the number of falsely selected variables dramatically. When tuned by SVMIC$_\gamma$, SCAD- and MCP-penalized SVMs select slightly fewer

Table 2.1: Simulation results for Model 1

Method	n	p	Signal	Noise	Correct	Test Error
SCAD-svm	100	400	4.94(0.03)	0.89(0.19)	64%	8.71%(0.4%)
	100	800	4.93(0.03)	0.93(0.14)	51%	9.39%(0.4%)
	200	800	5.00(0.00)	0.09(0.05)	96%	7.20%(0.2%)
	200	1600	5.00(0.00)	0.07(0.04)	96%	7.24%(0.2%)
MCP-svm	100	400	4.90(0.04)	0.88(0.17)	53%	8.96%(0.4%)
	100	800	4.92(0.03)	1.37(0.20)	40%	10.59%(0.5%)
	200	800	5.00(0.00)	0.06(0.04)	97%	7.30%(0.2%)
	200	1600	5.00(0.00)	0.09(0.03)	92%	6.79%(0.2%)
SCAD-svm$^{(\mathrm{SVMIC}_\gamma)}$	100	400	4.64(0.08)	0.48(0.11)	64%	10.32%(0.6%)
	100	800	4.63(0.09)	0.57(0.09)	52%	11.68%(0.7%)
	200	800	5.00(0.00)	0.03(0.02)	97%	7.24%(0.2%)
	200	1600	4.99(0.01)	0.05(0.03)	95%	7.23%(0.2%)
MCP-svm$^{(\mathrm{SVMIC}_\gamma)}$	100	400	4.46(0.10)	0.44(0.08)	45%	11.81%(0.6%)
	100	800	4.34(0.11)	0.68(0.11)	38%	13.13%(0.7%)
	200	800	5.00(0.00)	0.09(0.03)	92%	7.34%(0.2%)
	200	1600	5.00(0.00)	0.06(0.03)	95%	7.19%(0.2%)
L_1-svm	100	400	4.87(0.05)	32.97(1.47)	0%	16.08%(0.5%)
	100	800	4.63(0.07)	44.34(2.18)	0%	19.71%(0.6%)
	200	800	5.00(0.00)	21.33(1.70)	0%	9.59%(0.3%)
	200	1600	4.99(0.01)	33.37(0.96)	0%	10.88%(0.3%)
Hybrid-svm	100	400	4.78(0.05)	24.74(1.37)	0%	16.34%(0.5%)
	100	800	4.62(0.06)	27.16(1.30)	0%	19.93%(0.6%)
	200	800	5.00(0.00)	12.86(0.99)	0%	9.93%(0.2%)
	200	1600	4.99(0.01)	10.85(0.98)	0%	10.53%(0.3%)
Adap L_1-svm	100	400	4.39(0.08)	13.14(0.90)	0%	16.76%(0.5%)
	100	800	3.99(0.08)	12.50(0.69)	0%	20.19%(0.6%)
	200	800	4.86(0.04)	3.93(0.25)	1%	10.04%(0.3%)
	200	1600	4.49(0.06)	1.01(0.09)	4%	13.43%(0.4%)
L_2-svm	100	400	5.00(0.00)	395.00(0.00)	0%	39.23%(0.5%)
	100	800	5.00(0.00)	795.00(0.00)	0%	42.99%(0.5%)
	200	800	5.00(0.00)	795.00(0.00)	0%	39.22%(0.3%)
	200	1600	5.00(0.00)	1595.00(0.00)	0%	42.50%(0.4%)

Table 2.2: Simulation results for Model 2 with $n = 250$ and $p = 800$

Method	Signal	Noise	Correct	Test Error
SCAD-svm	3.99(0.01)	0.26(0.08)	92.5%	11.4%(0.1%)
MCP-svm	3.99(0.01)	0.17(0.07)	93.5%	11.3%(0.1%)
SCAD-svm$^{(\text{SVMIC}_\gamma)}$	3.96(0.02)	0.05(0.02)	94%	11.5%(0.1%)
MCP-svm$^{(\text{SVMIC}_\gamma)}$	3.98(0.01)	0.07(0.02)	92.5%	11.4%(0.1%)
L_1-svm	4.00(0.00)	6.84(0.42)	7.5%	12.4%(0.1%)
Hybrid-svm	4.00(0.00)	4.03(0.41)	10.5%	11.9%(0.1%)
Adap L_1-svm	4.00(0.00)	2.90(0.28)	38%	11.8%(0.1%)
L_2-svm	4.00(0.00)	796.00(0.00)	0%	32.5%(0.2%)

signals when $n = 100$, but this is based on the fact that other methods select a much larger model without proper control of noise. A large proportion of the missed relevant covariates are from X_1 as it has the weakest signal. Notice that SVMIC$_\gamma$ performs almost the same as "population tuning" when n is relatively large. In general, the non-convex penalized SVMs have an overwhelmingly high probability to select the exact true mode as n and p increase, while other methods show very weak, if any at all, ability to recover the exact true model. This is consistent with our theory of asymptotic oracle property of non-convex penalized SVMs. The test errors of SCAD- and MCP-penalized SVMs are uniformly smaller than those of any other method in all settings, even in the settings with a small sample size $n = 100$ and tuned by SVMIC$_\gamma$, where they select slightly fewer signals. This is due to the fact that in high dimensional classification problem, a large number of falsely selected variables will greatly blur the prediction power of the relevant variables.

Table 2.2 shows the results of Model 2 for $n = 250$ and $p = 800$. The numbers in the parentheses are the corresponding standard errors based on 200 replications. We observe similar performance patterns in terms of both variable selection and prediction error. Due to the higher correlation between signal and noise, in Model 2 it is generally more difficult to select the relevant covariates. Both SCAD- and MCP-penalized SVM still have reasonable performance in identifying the underlying true model and result in more accurate prediction. Note that under this data generation process the adaptively weighted L_1-penalized SVM behaves similar

Table 2.3: Classification error of MAQC-II dataset

Method	Test error	Genes
SCAD-svm	9.8%(0.2%)	2.06(0.43)
MCP-svm	9.6%(0.2%)	1.04(0.02)
L_1-svm	10.9%(0.2%)	28.74(1.36)
Adap L_1-svm	13.1%(0.2%)	34.30(1.03)
Hybrid-svm	10.0%(0.1%)	1391.60(94.86)
L_2-svm	10.8%(0.2%)	3000.00(0.00)

to non-convex penalized SVMs, though its oracle property is largely unknown.

2.5 Real Data

We consider the MAQC-II data described in Section 1.2. The original data have been standardized for each predictor. To reduce the computational burden, only the 3000 genes with largest absolute values of the two sample t-statistics are used. Such simplification has been considered in Cai and Liu (2011). Though only 3000 genes are used, the classification result is satisfactory. We randomly split the data into an equally balanced training set with 50 samples with positive ER status and 50 samples with negative ER status. The rest are designated as the test set. As in the simulation study, we use a=3.7 for the SCAD penalty and a=3 for the MCP penalty. To get a fair comparison, a 5-fold cross validation is implemented on the training set to select a tuning parameter by a grid search over $\{2^{-15}, \ldots, 2^3\}$ for all methods and the test error is calculated on the test data. The above procedure is repeated 100 times.

Table 2.3 summarizes the average classification error and number of selected genes. The numbers in the parentheses are the corresponding standard errors based on 100 replications. Non-convex penalized SVMs achieve significantly lower test error than all the other methods except for the doubly penalized hybrid SVM. Although the doubly penalized hybrid SVM performs similar to SCAD- and MCP-penalized SVMs in terms of test error, it selects a much more complex model in general. In addition, the number of genes selected by non-convex penalized SVMs is stable, while the model size selected by hybrid SVM ranges from 102 genes

to 2576 genes across the 100 replications. Such stability is desirable, so that the procedure is robust to the random partition of the data. The numerical results confirm that SCAD- and MCP-penalized SVMs can achieve both promising prediction power and excellent gene selection ability.

2.6 Proofs

We first prove Lemma 2.1.

Proof of Lemma 2.1. Let $l(\boldsymbol{\beta}_1) = n^{-1}\sum_{i=1}^n W_i(1 - Y_i\boldsymbol{Z}_i^\intercal\boldsymbol{\beta}_1)_+$. Note that $\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\beta_1} l(\boldsymbol{\beta}_1)$. We will show that when $\forall\eta > 0$, there exists a constant $\triangle$ such that for all n sufficiently large, $\Pr\{\inf_{||\boldsymbol{u}||=\triangle} l(\boldsymbol{\beta}_{01} + \sqrt{q/n}\boldsymbol{u}) > l(\boldsymbol{\beta}_{01})\} \geq 1 - \eta$. Because $l(\boldsymbol{\beta}_1)$ is convex, with probability at least $1 - \eta$, $\widehat{\boldsymbol{\beta}}_1$ is in the ball $\{\boldsymbol{\beta}_1 : ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}|| \leq \triangle\sqrt{q/n}\}$. Denote $\Lambda_n(\boldsymbol{u}) = nq^{-1}\{l(\boldsymbol{\beta}_{01} + \sqrt{q/n}\boldsymbol{u}) - l(\boldsymbol{\beta}_{01})\}$. Observe that $\mathbb{E}\{\Lambda_n(\boldsymbol{u})\} = nq^{-1}\{L(\beta_{01} + \sqrt{q/n}\boldsymbol{u}) - L(\boldsymbol{\beta}_{01})$. Recall also that $\boldsymbol{\beta}_0 = \arg\min_\beta \mathbb{E}\{W(1 - Y\boldsymbol{X}^\intercal\boldsymbol{\beta})\}$. If we restrict the last $p - q$ elements to be 0, it can be easily seen that $\boldsymbol{\beta}_{01} = \arg\min_{\beta_1} \mathbb{E}\{W(1 - Y\boldsymbol{Z}^\intercal\boldsymbol{\beta}_1)\} = \arg\min_{\beta_1} L(\boldsymbol{\beta}_1)$, thus $S(\boldsymbol{\beta}_{01}) = 0$. By Taylor series expansion of $L(\boldsymbol{\beta}_1)$ around $\boldsymbol{\beta}_{01}$, we have $\mathbb{E}\{\Lambda_n(\boldsymbol{u})\} = \frac{1}{2}\mathbf{u}^\intercal H(\tilde{\boldsymbol{\beta}})\boldsymbol{u} + o_p(1)$, where $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_{01} + \sqrt{q/n}t\boldsymbol{u}$ for some $0 < t < 1$. As shown in Koo et al. (2008), for $0 \leq j, k \leq q$, the (j, k)-th element of the Hessian Matrix $H(\boldsymbol{\beta}_{01})$ is continuous given (A1) and (A2); thus $H(\boldsymbol{\beta})$ is continuous. By continuity of $H(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_{01}$, then $\frac{1}{2}\boldsymbol{u}^\intercal H(\tilde{\boldsymbol{\beta}})\boldsymbol{u} = \frac{1}{2}\boldsymbol{u}^\intercal H(\boldsymbol{\beta}_{01})\boldsymbol{u} + o(1)$ as $n \to \infty$. Define $\boldsymbol{W}_n = -\sum_{i=1}^n \zeta_i W_i Y_i \boldsymbol{Z}_i$ where $\zeta_i = I(1 - Y_i\boldsymbol{Z}_i^\intercal\beta_{01} \geq 0)$. Recall that $S(\boldsymbol{\beta}_{01}) = -\mathbb{E}[\zeta_i W_i Y_i \boldsymbol{Z}_i] = 0$. If we define

$$R_{i,n}(\boldsymbol{u}) = W_i(1 - Y_i\boldsymbol{Z}_i^\intercal(\beta_{01} + \frac{\sqrt{q}}{\sqrt{n}}\boldsymbol{u}))_+ - W_i(1 - Y_i\boldsymbol{Z}_i^\intercal\beta_{01})_+ + \zeta_i W_i Y_i \boldsymbol{Z}_i^\intercal\sqrt{q/n}\boldsymbol{u}$$

then we have

$$\Lambda_n(\boldsymbol{u}) = \mathbb{E}\{\Lambda_n(\boldsymbol{u})\} + \boldsymbol{W}_n^\intercal\boldsymbol{u}/\sqrt{qn} + q^{-1}\sum_{i=1}^n[R_{i,n}(\boldsymbol{u}) - \mathbb{E}\{R_{i,n}(\boldsymbol{u})\}]. \tag{2.7}$$

Then similar to Equation (28) in Koo et al. (2008) we have

$$q^{-2}\sum_{i=1}^{n} E[|R_{i,n}(\boldsymbol{u}) - E\{R_{i,n}(\boldsymbol{u})\}|^2] \le C\triangle^2 E\{q^{-1}(1+||\boldsymbol{Z}||^2)U(\sqrt{1+||\boldsymbol{Z}||^2}\triangle\sqrt{q/n})\},$$

where $U(t) = I\left(|1 - Y_i\boldsymbol{Z}_i^{\intercal}\boldsymbol{\beta}_{01}| < t\right)$. (A2) implies that $E\{q^{-1}(1+||\boldsymbol{Z}||^2)\} < \infty$. Hence, for any $\epsilon > 0$, we can choose a positive constant C such that $\mathrm{E}[q^{-1}(1+||\boldsymbol{Z}||^2)I\{q^{-1}(1+||\boldsymbol{Z}||^2) > C\}] < \epsilon/2$, then

$$\begin{aligned}
&E\{q^{-1}(1+||\boldsymbol{Z}||^2)U(\sqrt{1+||\boldsymbol{Z}||^2}\triangle\sqrt{q/n})\} \\
\le& E[q^{-1}(1+||\boldsymbol{Z}||^2)I\{q^{-1}(1+||\boldsymbol{Z}||^2) > C\}] + C\Pr(|1 - Y_i\boldsymbol{Z}_i^{\intercal}\boldsymbol{\beta}_{01}| < C\triangle\sqrt{q/n}).
\end{aligned}$$

We can take a large N such that $\Pr(|1 - Y_i\boldsymbol{Z}_i^{\intercal}\boldsymbol{\beta}_{01}| < C\triangle\sqrt{q/n}) < \frac{\epsilon}{2C}$ for all $n > N$ by (A4). This proves that $q^{-2}\sum_{i=1}^{n} E\{|R_{i,n}(\boldsymbol{u}) - E[R_{i,n}(\boldsymbol{u})]|^2\} \to 0$ as $n \to \infty$. Observe that $E(\boldsymbol{W}_n^{\intercal}\boldsymbol{u}/\sqrt{qn}) = 0$, and

$$\mathrm{Var}\left(\boldsymbol{W}_n^{\intercal}\boldsymbol{u}/\sqrt{qn}\right) \le Cn^{-1}q^{-1}\sum_{i=1}^{n}(\boldsymbol{Z}_i^{\intercal}\boldsymbol{u})^2 \le Cq^{-1}\lambda_{max}(n^{-1}\boldsymbol{X}_A^{\intercal}\boldsymbol{X}_A)||\boldsymbol{u}||^2 \to 0$$

as $n \to \infty$. Therefore, the first term of (2.7) will dominate other terms as $n \to \infty$. By (A6) we have $\frac{1}{2}\mathbf{u}^{\intercal}H(\boldsymbol{\beta}_{01})\boldsymbol{u} > 0$. Thus we can choose a sufficiently large $\triangle$ such that $\Lambda_n(\boldsymbol{u}) > 0$ with probability $1-\eta$ for $||\boldsymbol{u}|| = \triangle$ and all sufficiently large n. □

The proof of Theorem 2.1 relies on the following Lemmas.

Lemma 2.2.

$$\Pr\{\max_{q+1\le j\le p} n^{-1}|\sum_{i=1}^{n} W_iY_iX_{ij}I(1 - Y_i\boldsymbol{Z}_i^{\intercal}\boldsymbol{\beta}_{01} \ge 0)| > \lambda/2\} \to 0 \textit{ as } n \to \infty.$$

Proof. Recall that $\mathbb{E}\{W_iY_iX_{ij}I(1-Y_i\boldsymbol{Z}_i^{\intercal}\boldsymbol{\beta}_{01} \ge 0)\} = 0$. By (A5) and Lemma 14.9 of Bühlmann and Van De Geer (2011), we have $\Pr\{n^{-1}|\sum_{i=1}^{n} W_iY_iX_{ij}I(1 - Y_i\boldsymbol{Z}_i^{\intercal}\boldsymbol{\beta}_{01} \ge 0)| > \lambda/2\} \le$

$\exp(-Cn\lambda^2)$. Note that

$$\Pr\{\max_{q+1\le j\le p} n^{-1}|\sum_{i=1}^n W_iY_iX_{ij}I(1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_{01}\ge 0)|>\lambda/2\}$$
$$=\Pr\{\cup_{q+1\le j\le p}\{n^{-1}|\sum_{i=1}^n W_iY_iX_{ij}I(1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_{01}\ge 0)|>\lambda/2\}\}\le p\exp(-Cn\lambda^2)\to 0$$

as $n\to\infty$ by the fact that $\log(p)=o(n\lambda^2)$.

□

Lemma 2.3. *For any* $\triangle>0$,

$$\Pr\{\max_{q+1\le j\le p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}}|\sum_{i=1}^n W_iY_iX_{ij}[I(1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_1\ge 0)-I(1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_{01}\ge 0)$$
$$-\Pr(1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_1\ge 0)+\Pr(1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_{01}\ge 0)]|>n\lambda\}\to 0$$

as $n\to\infty$.

Proof. We generalize an approach by Welsh (1989). We cover the ball $\{\boldsymbol{\beta}_1 : ||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}\}$ with a net of balls with radius $\triangle\sqrt{q/n^5}$. It can be shown that this net can be constructed with cardinality $N\le dn^{4q}$ for some $d>0$. Denote the N balls by $B(\mathbf{t_1}),\ldots,B(\mathbf{t_N})$, where $\mathbf{t_k}, k=1,\ldots,N$ are the centers. Denote $\kappa_i(\boldsymbol{\beta}_1)=1-Y_i\boldsymbol{Z}_i^\top\boldsymbol{\beta}_1$, and

$$\begin{aligned}
J_{nj1} &= \sum_{k=1}^N \Pr(|\sum_{i=1}^n W_iY_iX_{ij}[I\{\kappa_i(\mathbf{t_k})\ge 0\}-I\{\kappa_i(\boldsymbol{\beta}_{01})\ge 0\}\\
&\quad -\Pr\{\kappa_i(\mathbf{t_k})\ge 0\}+\Pr\{\kappa_i(\boldsymbol{\beta}_{01})\ge 0\}]|>n\lambda/2),\\
J_{nj2} &= \sum_{k=1}^N \Pr(\sup_{\tilde{\boldsymbol{\beta}}_1\in B(\mathbf{t_k})}|\sum_{i=1}^n W_iY_iX_{ij}[I\{\kappa_i(\tilde{\boldsymbol{\beta}}_1)\ge 0\}\\
&\quad -I\{\kappa_i(\mathbf{t_k})\ge 0\}-\Pr\{\kappa_i(\tilde{\boldsymbol{\beta}}_1)\ge 0\}+\Pr\{\kappa_i(\mathbf{t_k})\ge 0\}]|>n\lambda/2).
\end{aligned}$$

Then by (A5),

$$\Pr(\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\leq\triangle\sqrt{q/n}}|\sum_{i=1}^{n}W_iY_iX_{ij}[I\{\kappa_i(\boldsymbol{\beta}_1)\geq 0\}-I\{\kappa_i(\boldsymbol{\beta}_{01})\geq 0\}$$
$$-\Pr\{\kappa_i(\boldsymbol{\beta}_1)\geq 0\}+\Pr\{\kappa_i(\boldsymbol{\beta}_{01})\geq 0\}]|>n\lambda)\leq J_{nj1}+J_{nj2}.$$

To evaluate J_{nj1}, let $U_i = W_iY_iX_{ij}[I\{\kappa_i(\mathbf{t_k}) \geq 0\} - I\{\kappa_i(\boldsymbol{\beta}_{01}) \geq 0\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\} + \Pr\{\kappa_i(\boldsymbol{\beta}_{01}) \geq 0\}]$. The U_i are independent mean-zero random variable , and $\mathrm{Var}(U_i) = \mathbb{E}(U_i^2) = \mathbb{E}(U_i^2|Y_i = 1)\Pr(Y_i = 1) + \mathbb{E}(U_i^2|Y_i = -1)\Pr(Y_i = -1)$. Denote F and G the CDF of the conditional distribution of $\mathbf{Z}^\top\boldsymbol{\beta}_{01}$ given $Y = +1$ and $Y = -1$. Observe that

$$\begin{aligned}\mathbb{E}(U_i^2|Y_i=1)\leq& C\{F_i(1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k}))(1-F_i(1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k})))+F_i(1)(1-F_i(1))\\&-2F_i(\min(1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k}),1))+2F_i(1)F_i(1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k}))\}\\\leq& C|\boldsymbol{Z}_i^\top(\mathbf{t_k}-\boldsymbol{\beta}_{01})|,\end{aligned}$$

and it follows by (A8) that

$$\begin{aligned}&\mathbb{E}(U_i^2|Y_i=-1)\\\leq& C\{G_i(-1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k}))(1-G_i(-1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k})))+G_i(-1)(1-G_i(-1))\\&-2(1-G_i(\max(-1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k}),-1)))+2(1-G_i(-1))(1-G_i(-1+\boldsymbol{Z}_i^\top(\boldsymbol{\beta}_{01}-\mathbf{t_k})))\}\\\leq& C|\boldsymbol{Z}_i^\top(\mathbf{t_k}-\boldsymbol{\beta}_{01})|.\end{aligned}$$

Thus we have

$$\sum_{i=1}^{n}\mathrm{Var}(U_i)\leq nC\max_i||\boldsymbol{Z}_i||||\mathbf{t_k}-\boldsymbol{\beta}_{01}||=nO(\sqrt{q}\log(n))O(\sqrt{q/n})=O(\sqrt{n}q\log(n)).$$

Applying Lemma 14.9 of Bühlmann and Van De Geer (2011), for some positive constant C_1

and C_2 under the assumptions on the rate of λ,

$$J_{nj1} \leq 2N \exp(-\frac{n^2\lambda^2/4}{C_1\sqrt{n}q\log(n) + C_2 n\lambda}) \leq C\exp\{4q\log(n) - Cn\lambda\}. \tag{2.8}$$

To evaluate J_{nj2}, note that $I(x \geq s)$ is decreasing in s. Denote

$$V_i = [I\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geq 0\} - I\{\kappa_i(\mathbf{t_k}) \geq 0\} - \Pr\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geq 0\} + \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\}].$$

We have $-B_i \leq V_i \leq A_i$ for any $\tilde{\boldsymbol{\beta}}_1 \in B(\mathbf{t_k})$, where

$$A_i = [I\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t_k}) \geq 0\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq \triangle\sqrt{q/n^5}\} + \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\}],$$

$$B_i = [I\{\kappa_i(\mathbf{t_k}) \geq 0\} - I\{\kappa_i(\mathbf{t_k}) \geq \triangle\sqrt{q/n^5}\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\} + \Pr\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\}].$$

Therefore, we have

$$\begin{aligned}
&\Pr(\sup_{\tilde{\boldsymbol{\beta}}_1 \in B(\mathbf{t_k})} |\sum_{i=1}^n W_i Y_i X_{ij}[I\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geq 0\} - I\{\kappa_i(\mathbf{t_k}) \geq 0\} - \Pr\{\kappa_i(\tilde{\boldsymbol{\beta}}_1) \geq 0\} + \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\}]| > n\lambda/2) \\
\leq &\Pr(C\max_i |X_{ij}| \sup_{\tilde{\boldsymbol{\beta}}_1 \in B(\mathbf{t_k})} |\sum_{i=1}^n V_i| > n\lambda/2) \leq \Pr\{C\max_i |X_{ij}| \max(\sum_{i=1}^n A_i, \sum_{i=1}^n B_i) > n\lambda/2\}
\end{aligned}$$

by the fact that $A_i > 0, B_i > 0$. Note that

$$\begin{aligned}
\sum_{i=1}^n A_i = &\sum_{i=1}^n [I\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t_k}) \geq 0\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} + \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\}] \\
&+ \sum_{i=1}^n [\Pr\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq \triangle\sqrt{q/n^5}\}]
\end{aligned}$$

and

$$
\begin{aligned}
&\sum_{i=1}^{n}[\Pr\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq \triangle\sqrt{q/n^5}\}] \\
=&[F_i(1+\triangle\sqrt{q/n^5} - \boldsymbol{Z}_i^{\intercal}(\boldsymbol{\beta}_{01} - \mathbf{t_k})) - F_i(1-\triangle\sqrt{q/n^5} - \boldsymbol{Z}_i^{\intercal}(\boldsymbol{\beta}_{01} - \mathbf{t_k}))]\Pr(Y_i = 1) \\
&+ [G_i(-1+\triangle\sqrt{q/n^5} - \boldsymbol{Z}_i^{\intercal}(\boldsymbol{\beta}_{01} - \mathbf{t_k})) - G_i(-1-\triangle\sqrt{q/n^5} - \boldsymbol{Z}_i^{\intercal}(\boldsymbol{\beta}_{01} - \mathbf{t_k}))]\Pr(Y_i = -1) \\
\leq& Cn\log(q)\sqrt{q/n^5}\sqrt{q} = C\log(q)qn^{-3/2}
\end{aligned}
$$

by (A8). Denote

$$
O_i = [I\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t_k}) \geq 0\} - \Pr\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} + \Pr\{\kappa_i(\mathbf{t_k}) \geq 0\}].
$$

Thus for sufficiently large n by $\lambda = o(n^{-(1-c_2)/2})$ and A(7), we have

$$
\sum_{k=1}^{N}\Pr(C\sum_{i=1}^{n}A_i > n\lambda/2) \leq \sum_{k=1}^{N}\Pr(C\sum_{i=1}^{n}O_i > n\lambda/2 - C\log(q)qn^{-3/2}) \leq \sum_{k=1}^{N}\Pr(C\sum_{i=1}^{n}O_i > n\lambda/4).
$$

Notice that O_i are independent mean-zero random variables, and

$$
\mathbb{E}(O_i^2) = \mathbb{E}[I\{\kappa_i(\mathbf{t_k}) \geq -\triangle\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t_k}) \geq 0\}]^2 \leq \sqrt{q/n^5}\max_i ||\boldsymbol{Z}_i|| = Cq\log(n)n^{-5/2},
$$

using a similar idea to deriving the upper bound of $\mathbb{E}(U_i^2)$. Applying Bernstein's inequality and the fact that $\max_i |X_{ij}| = O_p(\sqrt{\log(n)})$ for sub-Gaussian random variable, for some positive constant C_1 and C_2,

$$
\sum_{k=1}^{N}\Pr(C\max_i|X_{ij}|\sum_{i=1}^{n}A_i > n\lambda/2) \leq N\exp(-\frac{n^2\lambda^2/4}{C_1qn^{-3/2}\log(n)^{3/2} + C_2n\lambda}) \leq C\exp\{4q\log(n) - Cn\lambda\}.
$$

Similarly, we can prove that $\sum_{k=1}^{N}\Pr(C\max_i|X_{ij}|\sum_{i=1}^{n}B_i > n\lambda/2) \leq C\exp\{4q\log(n) - Cn\lambda\}$. Therefore, we have

$$
J_{nj2} \leq C\exp\{4q\log(n) - Cn\lambda\}. \tag{2.9}
$$

Using (2.8) and (2.9), then the probability of Lemma 2.3 is bounded by

$$\sum_{j=q+1}^{p} (J_{nj1} + J_{nj2}) \leq C \exp\{\log(p) + 4q\log(n) - Cn\lambda\} \to 0 \tag{2.10}$$

which completes the proof. □

Now we prove Theorem 2.1.

Proof of Theorem 2.1. The unpenalized hinge loss objective function is convex. By convex optimization theorem, there exists v_i^* such that $s_j(\widehat{\boldsymbol{\beta}}) = 0,\ j = 0, 1, ...q$, with $v_i = v_i^*$.

Note that $\min_{1\leq j\leq q} |\hat{\beta}_j| \geq \min_{1\leq j\leq q} |\beta_{0j}| - \max_{1\leq j\leq q} |\hat{\beta}_j - \beta_{0j}|$. Under Condition (A7), we have $n^{(1-c_2)/2} \min_{1\leq j\leq q_n} |\beta_{0j}| \geq M_1$, and $\max_{1\leq j\leq q} |\hat{\beta}_j - \beta_{0j}| = O_p(\sqrt{q/n})$ by Lemma 2.1. Thus we have $\min_{1\leq j\leq q} |\hat{\beta}_j| = O_p(n^{-(1-c_2)/2})$. By $\lambda = o(n^{-(1-c_2)/2})$, we have $\Pr(|\hat{\beta}_j| \geq (a+\frac{1}{2})\lambda) \to 1$, , for $j = 0, 1, \ldots, q$.

By the definition of the oracle estimator, we have $|\hat{\beta}_j| = 0,\ j = q+1, \ldots, p$. It suffices to show that $\Pr\{|s_j(\widehat{\boldsymbol{\beta}})| > \lambda, \text{for some } j = q+1, \ldots, p\} \to 0$. Let $\boldsymbol{D} = \{i : 1 - Y_i \boldsymbol{Z}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_1 = 0\}$; then for $j = q+1, \ldots, p$, we have

$$s_j(\widehat{\boldsymbol{\beta}}) = -n^{-1} \sum_{i=1}^{n} W_i Y_i X_{ij} I(1 - Y_i \boldsymbol{Z}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_1 \geq 0) - n^{-1} \sum_{i\in \boldsymbol{D}} W_i Y_i X_{ij}(v_j - 1),$$

where $-1 \leq v_i \leq 0$ if $i \in \boldsymbol{D}$ and $v_i = 0$ otherwise. By (A5) $(\boldsymbol{Z}_i, Y_i)$ are in general positions, with probability one there are exactly $(q+1)$ elements in $\boldsymbol{D}$. Then by (A4), with probability one $|n^{-1} \sum_{i\in \boldsymbol{D}} W_i Y_i X_{ij}(v_j - 1)| = O(qn^{-1}\log(q)) = o(\lambda)$. Thus we only need to show that

$\Pr\{\max_{q+1\le j\le p}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_1\ge 0)|>\lambda\}\to 0$. Observe that

$$\begin{aligned}
&\Pr\{\max_{q+1\le j\le p}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_1\ge 0)|>\lambda\}\\
\le&\Pr\{\max_{q+1\le j\le p}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}[I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_1\ge 0)-I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)]|>\frac{\lambda}{2}\}\\
&+\Pr\{\max_{q+1\le j\le p}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)|>\frac{\lambda}{2}\}. \qquad (2.11)
\end{aligned}$$

By Lemma 2.2 the second term of (2.11) is $o_p(1)$. Notice that from Lemma 2.1, the first term of (2.11) is bounded by

$$\begin{aligned}
&\Pr[\max_{q+1\le j\le p}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}\{I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_1\ge 0)-I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)\}|>\frac{\lambda}{2}]\\
\le&\Pr[\max_{q+1\le j\le p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}\{I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_1\ge 0)-I(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)\\
&-\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_1\ge 0)+\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)\}|>\frac{\lambda}{4}]\\
&+\Pr[\max_{q+1\le j\le p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}}|n^{-1}\sum_{i=1}^n W_iY_iX_{ij}\{\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_1\ge 0)\\
&-\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)\}|>\frac{\lambda}{4}]. \qquad (2.12)
\end{aligned}$$

By Lemma 2.3, the first term of (2.12) is $o_p(1)$. Thus we only need to bound the second term of (2.12). Notice that

$$\begin{aligned}
&|\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_1\ge 0)-\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)|\\
\le&|F_i(1+\boldsymbol{Z}_i^{\mathsf{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}))-F_i(1)|\Pr(Y_i=1)+|G_i(-1+\boldsymbol{Z}_i^{\mathsf{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}))-G_i(-1)|\Pr(Y_i=-1).
\end{aligned}$$

Then we have

$$\max_{q+1\le j\le p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}}|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_1\ge 0)-\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)\}|$$

$$\le C\max_{i,j}|X_{ij}|\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}}n^{-1}\sum_{i=1}^{n}||\boldsymbol{Z}_i||||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||=O_p(\sqrt{\log pn})O(\sqrt{q/n})O_p(\sqrt{q}\log(n))$$

$$=o_p(\lambda).$$

Thus

$$\Pr[\max_{q+1\le j\le p}\sup_{||\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}||\le\triangle\sqrt{q/n}}|n^{-1}\sum_{i=1}^{n}W_iY_iX_{ij}\{\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_1\ge 0)$$
$$-\Pr(1-Y_i\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_{01}\ge 0)\}|>\frac{\lambda}{4}]=o_p(1),$$

which completes the proof. □

Now we prove Theorem 2.2.

Proof of Theorem 2.2. We will show $\widehat{\boldsymbol{\beta}}$ is a local minimizer of $Q(\boldsymbol{\beta})$ by writing $Q(\boldsymbol{\beta})$ as $g(\boldsymbol{\beta})-h(\boldsymbol{\beta})$.

By Theorem 2.1, we have $\Pr\{\mathcal{G}\subseteq\partial g(\widehat{\boldsymbol{\beta}})\}\to 1$, where

$$\mathcal{G}=\{\boldsymbol{\xi}=(\xi_0,\ldots,\xi_p):\xi_0=0;\ \xi_j=\lambda\mathrm{sgn}(\widehat{\boldsymbol{\beta}})_j),j=1,\ldots,q;\xi_j=s_j(\boldsymbol{\beta})+\lambda l_j,j=q+1,\ldots,p.\},$$

where $l_j\in[-1,+1]$, $j=q+1,\ldots,p$.

Consider any $\boldsymbol{\beta}$ in the $\mathbb{R}^{p+1}$ with the center $\widehat{\boldsymbol{\beta}}$ and radius $\frac{\lambda}{2}$. It is suffices to show that there exist $\boldsymbol{\xi}^*\in\mathcal{G}$ such that $\Pr\{\xi_j^*=\frac{\partial h(\boldsymbol{\beta})}{\partial\beta_j}\}\to 1$ as $n\to\infty$.

Since $\frac{\partial h(\boldsymbol{\beta})}{\partial\beta_0}=0$, we have $\xi_0^*=\frac{\partial h(\boldsymbol{\beta})}{\partial\beta_0}$.

For $j=1,\ldots,q$, we have $\min_{1\le j\le q}|\beta_j|\ge\min_{1\le j\le q}|\hat{\beta}_j|-\max_{1\le j\le q}|\hat{\beta}_j-\beta_j|\ge(a+\frac{1}{2})\lambda-\frac{\lambda}{2}=a\lambda$ with probability one by Theorem 2.1. Therefore by Condition 2 of the class of penalties

$\Pr\{\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = \lambda \mathrm{sgn}(\beta_j)\} \to 1$ for $j = 1, \ldots, q$. For sufficently large n, $\mathrm{sgn}(\beta_j) = \mathrm{sgn}(\hat{\beta}_j)$. Thus we have $\Pr\{\xi_j^* = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\} \to 1$ as $n \to \infty$ for $j = 1, \ldots, q$.

For $j = q+1, \ldots, p$, we have $\Pr\{|\beta_j| \le |\hat{\beta}_j| + |\beta_j - \hat{\beta}_j| \le \lambda\} \to 1$ by Theorem 2.1. Therefore we have $\Pr\{\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = 0\} \to 1$ for SCAD and $\Pr\{\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{\beta_j}{a}\} \to 1$ for MCP. Observe that by Condition 2 we have $\Pr\{|\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}| \le \lambda\} \to 1$ for the class of penalties. By Lemma 1 we have $\Pr\{|s_j(\hat{\beta}_j)| \le \lambda\} \to 1$ for $j = q+1, \ldots, p$. We can always find $l_j \in [-1, +1]$ such that $\Pr\{\xi_j^* = s_j(\hat{\beta}) + \lambda l_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\} \to 1$ for $j = 1, \ldots, q$, for both penalties. This completes the proof.

□

The proof of Theorem 2.3 consists of two parts. First we wil show that LLA algorithm initiated by $\tilde{\boldsymbol{\beta}}^{(0)}$ gives the oracle estimator after one iteration. Then we will show that once LLA algorithm finds the oracle estimator $\widehat{\boldsymbol{\beta}}$, the LLA algorithm will find it again in the next iteration, that is, the LLA algorithm will converge.

Proof of Theorem 2.3. Assume that none of the events F_{ni} is true, for $i = 1, \ldots, 4$. The probability that none of these event is true is at least $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$. Then we have

$$
\begin{aligned}
&|\tilde{\beta}_j^{(0)}| = |\tilde{\beta}_j^{(0)} - \beta_{0j}| \le \lambda, q+1 \le j \le p, \\
&|\tilde{\beta}_j^{(0)}| \ge |\beta_{0j}| - |\tilde{\beta}_j^{(0)} - \beta_{0j}| \ge a\lambda, 1 \le j \le q.
\end{aligned}
$$

By Condition 2 of the class of non-convex penalties, we have $p'_\lambda(|\tilde{\beta}_j^{(0)}|) = 0$ for $1 \le j \le q$. Therefore the solution of the next iteration of $\tilde{\boldsymbol{\beta}}^{(1)}$ is the solution to the convex optimization

$$
\tilde{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} W_i (1 - Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta})_+ + \sum_{q+1 \le j \le p} p'_\lambda(|\tilde{\beta}_j^{(0)}|) \cdot |\beta_j|. \tag{2.13}
$$

By the fact the F_{n3} is not true, there exist some subgradients of oracle estimator $s(\widehat{\boldsymbol{\beta}})$ such that $s_j(\widehat{\boldsymbol{\beta}}) = 0$ for $0 \le j \le q$ and $|s_j(\widehat{\boldsymbol{\beta}})| < (1 - \frac{1}{a})\lambda$ for $q+1 \le j \le p$. Note that by the definition

of subgradient, we have

$$n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})_+ \geq n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}})_+ + \sum_{0\leq j\leq p} s_j(\widehat{\boldsymbol{\beta}})(\beta_j-\hat{\beta}_j)$$
$$= n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}})_+ + \sum_{q+1\leq j\leq p} s_j(\widehat{\boldsymbol{\beta}})(\beta_j-\hat{\beta}_j).$$

Then we have for any $\boldsymbol{\beta}$

$$\{n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})_+ + \sum_{q+1\leq j\leq p} p'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j|\} - \{n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}})_+ + \sum_{q+1\leq j\leq p} p'_\lambda(|\tilde{\beta}_j^{(0)}|)|\hat{\beta}_j|\}$$
$$\geq \sum_{q+1\leq j\leq p}\{p'_\lambda(|\tilde{\beta}_j^{(0)}|) - s_j(\widehat{\boldsymbol{\beta}})\cdot\ \mathrm{sgn}(\beta_j)\}\cdot|\beta_j| \geq \sum_{q+1\leq j\leq p}\{(1-\frac{1}{a})\lambda - s_j(\widehat{\boldsymbol{\beta}})\cdot\mathrm{sgn}(\beta_j)\}\cdot|\beta_j| \geq 0.$$

The strict inequality holds unless $\beta_j = 0$ for all $q+1 \leq j \leq p$. Since we consider the non-separable case where the oracle estimator is unique, we know the oracle estimator is the unique minimizer of (2.13) and hence $\tilde{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}$. This proves that the LLA algorithm finds the oracle estimator after one iteration.

In the case that F_{n2} is not true, we have $|\hat{\beta}_j| > a\lambda$ for all $1 \leq j \leq q$. Hence by Condition 2 of the class of penalties $p'_\lambda(|\hat{\beta}_j|) = 0$ for all $1 \leq j \leq q$ and $p'_\lambda(|\hat{\beta}_j|) = p'_\lambda(0) = \lambda$ for all $q+1 \leq j \leq p$. Once the LLA algorithm finds $\widehat{\boldsymbol{\beta}}$, the solution to the next LLA iteration $\tilde{\boldsymbol{\beta}}^{(2)}$ is the minimizer of the convex optimization problem

$$\tilde{\boldsymbol{\beta}}^{(2)} = \arg\min_{\boldsymbol{\beta}} n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})_+ + \sum_{q+1\leq j\leq p}\lambda|\beta_j|. \tag{2.14}$$

Then we have for any $\boldsymbol{\beta}$

$$\{n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})_+ + \sum_{q+1\leq j\leq p}\lambda|\beta_j|\} - \{n^{-1}\sum_{i=1}^{n} W_i(1-Y_i\boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}})_+ + \sum_{q+1\leq j\leq p}\lambda|\hat{\beta}_j|\}$$
$$\geq \sum_{q+1\leq j\leq p}\{\lambda - s_j(\widehat{\boldsymbol{\beta}})\cdot\mathrm{sgn}(\beta_j)\}\cdot|\beta_j| \geq 0.$$

and hence $\tilde{\boldsymbol{\beta}}^{(2)} = \widehat{\boldsymbol{\beta}}$ is the unique minimizer of (2.14). That is, the LLA algorithm finds the oracle estimator again and stops.

As $n \to \infty$, by Theorem 2.1 we have $P_{n2} \to 0$ and $P_{n4} \to 0$. The proof for $P_{n3} \to 0$ is similar to the proof for Theorem 2.1 by changing the constant to be $(1 - \frac{1}{a})$. $\square$

Now we prove Theorem 2.4.

Proof of Theorem 2.4. Let $||\cdot||_1$ be the L_1 norm of a vector. Denote $l_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n W_i(1 - Y_i\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta})_+ + c_n||\boldsymbol{\beta}||_1$. Note that

$$\begin{aligned}
&E[np^{-1}\{l_n(\boldsymbol{\beta}_0 + \sqrt{p/n}\boldsymbol{u}) - l_n(\boldsymbol{\beta}_0)\}] \\
=&E[np^{-1}\{W(1 - Y\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{\beta}_0 + \sqrt{p/n}\boldsymbol{u}))_+ - W(1 - Y\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}_0)_+\}] \\
&+ np^{-1}c_n(||\boldsymbol{\beta}_0 + \sqrt{p/n}\boldsymbol{u}||_1 - ||\boldsymbol{\beta}_0||_1)
\end{aligned}$$

for some constant $\triangle$ that $||\boldsymbol{u}|| = \triangle$. Observe that $||\boldsymbol{\beta}_0 + \sqrt{p/n}\boldsymbol{u}||_1 - ||\boldsymbol{\beta}_0||_1 \leq ||\sqrt{p/n}\boldsymbol{u}||_1 = \sqrt{p/n}||\boldsymbol{u}||_1$. By the fact that $c_n = o(n^{-1/2})$, we have $np^{-1}c_n(||\boldsymbol{\beta}_0 + \sqrt{p/n}\boldsymbol{u}||_1 - ||\boldsymbol{\beta}_0||_1) \to 0$ as $n \to \infty$. Then similar to the proof of Lemma 2.1, we can show that the expectation is dominated by $\frac{1}{2}u^{\mathsf{T}}H(\boldsymbol{\beta}_0)\boldsymbol{u} > 0$ and $\Pr\{\inf_{||\boldsymbol{u}||=\triangle} l_n(\boldsymbol{\beta}_0 + \sqrt{p/n}\boldsymbol{u}) > l_n(\boldsymbol{\beta}_0)\} \geq 1 - \eta$. Hence $||\widehat{\boldsymbol{\beta}}^{L_1} - \boldsymbol{\beta}_0|| = O_p(\sqrt{p/n})$. Because $pn^{-\frac{1}{2}} = o(\lambda)$, $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{ for some } 1 \leq j \leq p) \to 0$ as $n \to \infty$. Then using Theorem 2.1 and Corollary 2.1 we have $\Pr\{\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}\} \to 1$, which completes the proof. $\square$

Chapter 3

Reliability Modeling for Dependent Systems

This chapter is organized as follows. We introduce the problem and notation in Section 3.1. Section 3.2 summarizes the main results on the effects of dependence on system reliability and component importance and the extensions to multi-state systems. Implementation details are discussed in Section 3.3. Simulation studies are presented in Section 3.4, followed by the analysis of the stockpile test data in Section 3.5. The technical proofs can be found in Section 3.6.

3.1 Introduction

A common goal in reliability modeling is to provide accurate estimates of system reliability and component importance. Understanding reliability allows us to predict system performance. Importance measures evaluate the relative importance of individual components and can be applied to system design. Traditionally, the statistical models for reliability and component importance depend heavily on the assumption that the components operate independently within the system. In practice, however, it is more realistic to consider dependence among the components. Barlow and Proschan (1975) summarizes several examples of dependence,

including components subjected to common stresses and components sharing a load. Models that ignore such sources of dependence are mis-specified and can result in biased estimates of system reliability and component importance. The assumption of component independence can limit the applicability of reliability modeling (Lawless, 1983).

In this chapter, we consider a system containing p components. In a binary system, each component has two states: functioning and failed. Define a binary random variable X_j as the state of the jth component, with $X_j = 1$ indicating the jth component is functioning, and $X_j = 0$ indicating the jth component has failed, $j = 1, \ldots, p$. The reliability of the component is denoted as $\pi_j = \Pr(X_j = 1)$. Write $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)^\top$. It is clear that if the components are mutually independent, the joint distribution of $\boldsymbol{X}$ is determined completely by the marginal probabilities $\boldsymbol{\pi}$.

Let X_S denote a binary random variable representing the state of the system. As with the components, the system can be either in a functioning state or a failed state. We assume the state of the system is completely determined by the states of the components. That is, there exists a non-random mapping $\phi : \{0,1\}^p \to \{0,1\}$ such that $X_S = \phi(\boldsymbol{X})$. This function $\phi(\boldsymbol{X})$ is called the *structure function* of the system. Interest focuses on *coherent* systems, which require certain regularity conditions on the structure function. A component is *relevant* to the system if the structure function $\phi(\boldsymbol{X})$ is not constant in that component. The system is coherent if each component in the system is relevant and $\phi(\boldsymbol{X})$ is non-decreasing in each argument X_j, $j = 1, \ldots, p$. In this project, we only consider coherent systems. Some examples include the series system, with $\phi(\boldsymbol{X}) = \prod_{j=1}^p X_j$, the parallel system, with $\phi(\boldsymbol{X}) = 1 - \prod_{j=1}^p (1 - X_j)$, and the k-out-of-p system, with $\phi(\boldsymbol{X}) = 1$ if $\sum_{j=1}^p X_j \geq k$ and $\phi(\boldsymbol{X}) = 0$ if $\sum_{j=1}^p X_j < k$, where the system is functioning if and only if at least k out of p components are functioning.

The system reliability, R is defined as the probability of the system being in a functioning state, with $R = \Pr(\phi(\boldsymbol{X}) = 1)$. To calculate the reliability of a system with independent components, it suffices to find all of the minimal cut sets or the minimal path sets (Barlow and Proschan, 1975). A minimal cut set is a minimal set of components such that the system fails

whenever all the components in the set fail. A minimal path set is a minimal set of components such that the system functions whenever all the components in the set are functioning. System reliability satisfies

$$R = \prod_{\boldsymbol{C} \in \mathcal{C}} (1 - \prod_{i \in \boldsymbol{C}} (1 - \pi_i)) = 1 - \prod_{\boldsymbol{P} \in \mathcal{P}} (1 - \prod_{i \in \boldsymbol{P}} \pi_i), \tag{3.1}$$

where $\mathcal{C}$ is the set of all minimal cut sets and $\mathcal{P}$ is the set of all minimal path sets. Note that Eq. (3.1) holds only when the component states are independent.

The condition of independent component states is a simplifying assumption that may not be realistic in practice. Many methods have been proposed to study the impact of dependence on system reliability. Two popular classes of models are the common cause failure model (Marshall and Olkin, 1967; Hokstad, 1988; Vaurio, 2005) and the load sharing model (Durham and Lynch, 2000; Kvam and Pena, 2005). Recently, Bayesian networks have been used to capture the conditional independence structure of components (Bobbio et al., 2001; Langseth and Portinale, 2007). All of these models make specific assumptions on the dependence structure among components. Esary and Proschan (1970) considers general dependence structures and provides bounds for dependent system reliability. However, the reliability bounds therein result from the limiting case of independence and do not capture the effects of dependence on reliability directly.

To allow dependence among components, we follow the definitions in Esary and Proschan (1970). The random variables $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$ are *positively associated* if $\mathrm{Cov}(f(\boldsymbol{X}), g(\boldsymbol{X})) \geq 0$ for any pair of increasing functions (f, g). The random variables can be arbitrary and do not have to be binary. It can be easily seen that independence satisfies the positively associated assumption. Since f and g can be any increasing functions, the condition of positively associated random variables describes a particular type of positive dependence. Throughout this chapter, we assume $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$ are positively associated. Under this assumption, the reliability of a dependent system depends on both the marginals of the components and the dependence structure.

Another important problem in reliability modeling is to rank the importance of components

in a coherent system. Among many existing importance measures, we focus on the *reliability importance*, because our primary interest is in identifying the most important components from the perspective of system reliability. For positively associated components, Barlow and Proschan (1975) defined the reliability importance of the jth component as

$$\boldsymbol{I}_j = \mathbb{E}[\phi(1_j, \boldsymbol{X}_{-j})] - \mathbb{E}[\phi(0_j, \boldsymbol{X}_{-j})],$$

where $(1_j, \boldsymbol{X}_{-j})$ denotes the state vector $(X_1, \ldots, X_{j-1}, 1, X_{j+1}, \ldots, X_p)$, and similarly for $(0_j, \boldsymbol{X}_{-j})$. The jth component is *critical* to the system if $\phi(1_j, \boldsymbol{x}_{-j}) = 1$ and $\phi(0_j, \boldsymbol{x}_{-j}) = 0$ given $\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}$ for the other components. That is, a component is critical to the system, given the states of other components, when the system fails if and only if that component fails. It can be shown that the reliability importance is equivalent to the probability of the component being critical to the system (Barlow and Proschan, 1975). A comprehensive review of reliability importance measures and their applications can be found in Kuo and Zhu (2012).

The effect of dependence on component reliability importance is much less studied in the literature. The definition of reliability importance in Birnbaum (1968) assumes component independence. Barlow and Proschan (1975) extends the definition of component reliability importance to allow the dependence among the components. Under the extended definition, both the marginal distributions and the dependence structure can change component importance; however, Barlow and Proschan (1975) discusses only the effects of the marginals. Ebrahimi et al. (2014) provides interesting results on the order of information importance under dependent components. However, we are not aware of results for reliability importance as defined in Barlow and Proschan (1975) for dependent components.

3.2 Main Results

In this section, we establish a unified framework to study the effects of dependence on system reliability and component importance. The foundation of this unified framework is copula the-

ory. We first provide a brief review of copula theory; for more details in a modern treatment of copula theory, see Nelsen (2007).

The main motivation for copulas is to represent a joint distribution through modeling the marginals and the dependence structure separately. A copula function $C = C(u_1, \ldots, u_p)$ is a mapping $C : [0,1]^p \to [0,1]$ that is a cumulative distribution function (CDF) with each marginal a uniform distribution on $[0,1]$. In other words, $C(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p)$, where $U_j \sim \text{Unif}(0,1)$, $j = 1, \ldots, p$. Let F denote the joint CDF of $\boldsymbol{X}$, and F_j the marginal CDF of X_j. If the joint distribution of $\boldsymbol{X}$ can be represented by the marginal F_j's and the copula C, the marginals and the copula satisfy

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)). \tag{3.2}$$

Sklar (1959) showed that for any joint distribution, such a copula function C satisfying Eq. (3.2) always exists. The copula function C is usually modeled parametrically, and the marginals can be constructed either parametrically or nonparametrically. Copula theory provides a flexible way to model joint distributions with arbitrary dependence structures, which forms the basis for the results in the section.

The results for system reliability are provided in Section 3.2.1, followed by the results for component importance in Section 3.2.2. Extensions to multi-state system are summarized in Section 3.2.3.

3.2.1 System Reliability with Dependence

When a system has dependent components, the system reliability is a function of both the component marginals and the dependence structure, making it challenging to calculate the exact system reliability. However, certain reliability bounds hold for coherent dependent systems. Esary and Proschan (1970) proposed the following well-known reliability bounds: for a coherent

binary system with positively associated component states, the system reliability satisfies

$$\prod_{\boldsymbol{C}\in\mathcal{C}}(1-\prod_{i\in\boldsymbol{C}}(1-\pi_i)) \leq R \leq 1-\prod_{\boldsymbol{P}\in\mathcal{P}}(1-\prod_{i\in\boldsymbol{P}}\pi_i), \tag{3.3}$$

where $\mathcal{C}$ is the set of all minimal cut sets and $\mathcal{P}$ is the set of all minimal path sets. The reliability bounds in Eq. (3.3) essentially come from the extreme scenarios of independent minimal cut sets and minimal path sets. To see this, consider a simple series system. It can be easily seen that each component is a minimal cut set. Therefore, the lower bound of the system reliability in Eq. (3.3) is $\prod_{j=1}^{p}\pi_j$. That is, the lower bound of any coherent series system with positively associated component states is just the system reliability in the extreme case of independent components. Similarly, each component in a parallel system is a minimal path set, and the upper bound in Eq. (3.3) for a parallel system is just the reliability of the parallel system with the same marginals and independent component states. This dependence of the reliability bounds on the independence assumption can be seen from the formulation of the bounds: they are completely determined by the marginals. Though the reliability bounds hold for any coherent system with arbitrary dependence structure, the bounds in Eq. (3.3) are exactly the same for two systems with the same marginals, but different dependence structures, which is obviously not a desirable result.

Copula theory provides a direct tool to characterize the influence of dependence on system reliability. Recall that the system reliability, $R = \Pr(\phi(\boldsymbol{X}) = 1)$, is determined by the joint distribution of $\boldsymbol{X}$. The existence result in Sklar (1959) guarantees that any F can be expressed equivalently as F_j's and a copula function C. In a binary system, $F_j(x_j) = (1-\pi_j)\mathbf{1}(0 \leq x_j < 1) + \mathbf{1}(x_j \geq 1)$, where $\mathbf{1}(\cdot)$ denotes the indicator function. Write $R = R_{\boldsymbol{\pi},C}$ to explicitly express the dependence of system reliability on the marginals and the copula function. The component states are independent if $C(u_1,\ldots,u_p) = \prod_{j=1}^{p} u_p$, which is the independence copula function. Our goal is to incorporate the copula function C into the reliability bounds. To achieve this, we first need some ordering definitions to compare two copula functions. Since a copula function is a

CDF for multivariate uniform random variables, it is natural to use the definitions of stochastic ordering of positive dependence for multivariate CDFs.

Definition 3.1 (Joe (1997), Section 2.1)**.** *Let* $C(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p)$ *be the CDF of* $U_1, \ldots, U_p$ *and* $\bar{C}(u_1, \ldots, u_p) = \Pr(U_1 > u_1, \ldots, U_p > u_p)$ *be the corresponding survival function. For two CDFs,* C_1 *and* C_2*,* C_1 *is more positive upper orthant dependent (PUOD) than* C_2 *if*

$$\bar{C}_1(u_1, \ldots, u_p) \geq \bar{C}_2(u_1, \ldots, u_p), \ \forall (u_1, \ldots, u_p) \in \mathbb{R}^p.$$

C_1 *is more positive lower orthant dependent (PLOD) than* C_2 *if*

$$C_1(u_1, \ldots, u_p) \geq C_2(u_1, \ldots, u_p), \ \forall (u_1, \ldots, u_p) \in \mathbb{R}^p.$$

Some remarks about the definitions. First, the concepts of PUOD and PLOD only describe the ordering of positive dependence. Specifically, more PUOD (PLOD) indicates that the random variables are more likely to be large (small) together. Though these definitions cannot capture the ordering of negative dependence, they are sufficient for reliability modeling because the component states are usually positively associated.

Theorem 2.4 of Joe (1997) shows that for any positively associated random variables, $U_1, \ldots, U_p$, and arbitrary $u_1, \ldots, u_p$, $\Pr(U_1 \leq u_1, \ldots, U_p \leq u_p) \geq \prod_{j=1}^{p} \Pr(U_j \leq u_j)$ and $\Pr(U_1 > u_1, \ldots, U_p > u_p) \geq \prod_{j=1}^{p} \Pr(U_j > u_j)$, where the equalities hold when $U_1, \ldots, U_p$ are independent. In other words, the CDF of any positively associated random variables is both more PUOD and PLOD than the corresponding CDF under independence. Note that PUOD is equivalent to PLOD for $p = 2$, but this equivalence fails to hold in general for $p > 2$.

Based on the previous definitions, the following theorem summarizes the effects of dependence on series and parallel system reliability, which are immediate results of Theorem 3.2 and Theorem 4.2 in Navarro and Spizzichino (2010).

Theorem 3.1. *Consider a coherent system with* p *positively associated component states* $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$*,* $p \geq 2$*. Let* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ *denote the marginals and* C *the copula function for*

the joint distribution of component states. The system reliability, denoted as $R_{\boldsymbol{\pi},C}$, satisfies

1. *For a series system, if C_2 is more PUOD than C_1, then $R_{\boldsymbol{\pi},C_1} \leq R_{\boldsymbol{\pi},C_2}$.*

2. *For a parallel system, if C_2 is more PLOD than C_1, then $R_{\boldsymbol{\pi},C_1} \geq R_{\boldsymbol{\pi},C_2}$.*

Theorem 3.1 can be directly shown by applying the reliability representation for series and parallel systems in Navarro and Spizzichino (2010) to discrete distributions. Theorem 3.1 implies any positively associated series (parallel) system is at least (most) as reliable as the corresponding series (parallel) system under independence as shown in Esary and Proschan (1970). Theorem 3.1 also implies that the reliability of a series (parallel) system increases (decreases) with the order of positive dependence measured by PUOD (PLOD), which agrees with Theorem 2.40 of Aven and Jensen (1999) for continuous marginal distributions. Notice that the converse of Theorem 3.1 is true for continuous marginals but not true for discrete marginals. That is, we cannot conclude that for two series systems with the same component reliabilities, the copula function for the system with higher reliability is definitely more PUOD than the other copula function. This difference is due to the lack of uniqueness of the copula function under discrete marginals (see examples in Genest and Nešlehová (2007)). Notice also that the order of reliability for discrete random variables corresponds to the usual stochastic order for continuous random variables. Additional results on other stochastic orders for system lifetime are given in Navarro et al. (2015).

The results for series and parallel systems are building blocks for results applying to more general systems. For general system structures, the effect of dependence becomes more complex because it also depends on the specifics of the system structure. Recall that the system structure can be summarized by the minimal path sets or minimal cut sets. Denote $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_s$ as the minimal path sets and $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m$ as the minimal cut sets. For a set $J \in \{1, \ldots, p\}$, we introduce the notation $(\mathbf{1}_{-J}, \boldsymbol{u}_J)$ as the p-dimensional vector whose jth element equals to u_j if $j \in J$ and 1 otherwise. We define $(\mathbf{0}_{-J}, \boldsymbol{u}_J)$ similarly. The following theorem characterizes the influence of dependence on the reliability for a general coherent system.

Theorem 3.2. *Consider a coherent system with p positively associated component states $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$, $p \geq 2$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ denote the marginals and C the copula function for the joint distribution of component states. Denote s the number of minimal path sets and m the number of minimal cut sets.*

1. *Define*
$$h_C(\boldsymbol{u}) = \sum_{j=1}^{s} (-1)^{j+1} \sum_{1 \leq k_1 < \cdots < k_j \leq s} \bar{C}\Big((\mathbf{0}_{-\cup_{i=1}^{j} \boldsymbol{P}_{k_i}}, \boldsymbol{u}_{\cup_{i=1}^{j} \boldsymbol{P}_{k_i}})\Big).$$
If $h_{C_2}(\boldsymbol{u}) \geq h_{C_1}(\boldsymbol{u})$ for any $\boldsymbol{u} \in [0,1]^p$, then $R_{\boldsymbol{\pi},C_1} \leq R_{\boldsymbol{\pi},C_2}$.

2. *Define*
$$g_C(\boldsymbol{u}) = \sum_{j=1}^{m} (-1)^{j+1} \sum_{1 \leq k_1 < \cdots < k_j \leq m} C\Big((\mathbf{1}_{-\cup_{i=1}^{j} \boldsymbol{C}_{k_i}}, \boldsymbol{u}_{\cup_{i=1}^{j} \boldsymbol{C}_{k_i}})\Big).$$
If $g_{C_2}(\boldsymbol{u}) \geq g_{C_1}(\boldsymbol{u})$ for any $\boldsymbol{u} \in [0,1]^p$, then $R_{\boldsymbol{\pi},C_1} \geq R_{\boldsymbol{\pi},C_2}$.

Theorem 3.2 shows that the effect of dependence on reliability for a general system structure can be characterized through the functions $h_C(\boldsymbol{u})$ or $g_C(\boldsymbol{u})$. The function $h_C(\boldsymbol{u})$ was called the *domination function* in Navarro et al. (2014) and was used to represent the distribution of system lifetime as a *generalized distorted distribution* in Navarro et al. (2015). The proof of Theorem 3.2 is straightforward by converting the discrete component states to latent continuous variables via the data augmentation technique in Smith and Khaled (2012) and applying the results in Navarro et al. (2015) to the augmented data. Note that for discrete system states, we are interested in the order of reliability, which corresponds to the usual stochastic order for continuous random variables. Additional stochastic orders for continuous random variables, such as the failure rate order, are also studied in Navarro et al. (2014) and Navarro et al. (2015). Note also that the functions $h_C(\boldsymbol{u})$ and $g_C(\boldsymbol{u})$ are similar to the representations in the hyperminimal and hypermaximal distributions in Navarro et al. (2007).

The results for series and parallel systems are simply special cases of these more general results. To see this, notice that for a series system, the set $\{1, \ldots, p\}$ is the only minimal path set and $s = 1$. It can be shown that $h_C(\boldsymbol{u}) = \bar{C}(\boldsymbol{u})$ in this case. Similarly, for a parallel system, the

set $\{1, \ldots, p\}$ is the only minimal cut set and thus $m = 1$. It can be checked that $g_C(\boldsymbol{u}) = C(\boldsymbol{u})$ for a parallel system. Therefore, Theorem 3.2 extends the results for series and parallel systems to a general coherent system.

The previous results on system reliability hold for a general copula function C. In practice, a specific copula function has to be chosen from some parametric family to model the dependence. One popular class of copula functions for joint distributions with $p > 2$ is the *elliptical copulas*, where the copulas are constructed from elliptical distributions such as multivariate Gaussian distributions and multivariate t-distributions (Fang et al., 2002). The Gaussian copula is the most popular elliptical copula function, and it has been used in various problems for modeling joint distributions (Xue-Kun Song, 2000; Pitt et al., 2006; Li et al., 2011). The Gaussian copula is constructed from a multivariate Gaussian distribution with zero mean and correlation matrix $\boldsymbol{\Sigma}$. The copula function can be written as

$$C(u_1, \ldots, u_p) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)),$$

where Φ_p is the CDF of a p-dimensional $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, and Φ is the CDF for a standard univariate normal distribution. Note that the copula function C is completely determined by the off-diagonal elements of the correlation matrix $\boldsymbol{\Sigma}$. Throughout this project, we use a Gaussian copula to capture dependence in all of the numerical examples. The reason for this choice is that the concepts of positively associated random variables, PUOD, and PLOD have direct and simple interpretations in Gaussian copulas, which we summarize in the following propositions.

Proposition 3.1. *Assume the joint distribution of component states $\boldsymbol{X}$ can be expressed using a Gaussian copula with correlation matrix $\boldsymbol{\Sigma}$. The component states $\boldsymbol{X}$ are positively associated $\Leftrightarrow$ $\boldsymbol{\Sigma} \geq \mathbf{0}$ element-wise.*

Proposition 3.2. *Assume two Gaussian copulas, C_1 and C_2, with correlation matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. C_1 is more PUOD and PLOD than C_2 if $\boldsymbol{\Sigma}_1 \geq \boldsymbol{\Sigma}_2$ element-wise.*

Proposition 3.1 states that the condition of positively associated component states is equiv-

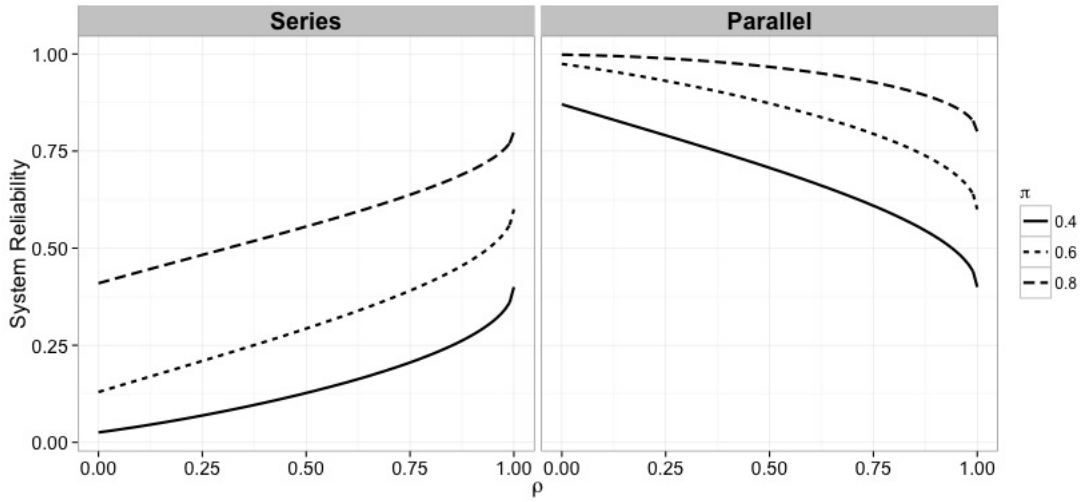

Figure 3.1: For fixed π, the series system reliability increases with ρ, while parallel system reliability decreases with ρ.

alent to the condition that all the off-diagonal elements of the correlation matrix $\mathbf{\Sigma}$ are nonnegative. Proposition 3.2 states that the orderings of copula functions in terms of PUOD and PLOD are are preserved by the element-wise ordering of $\mathbf{\Sigma}$.

For an illustrative example of the propositions, consider a system of four components with $\pi_j = \pi$, $j = 1, \ldots, 4$. Assume the dependence structure of $\boldsymbol{X}$ can be captured by a Gaussian copula with correlation matrix $\mathbf{\Sigma}$, where all the off-diagonal elements of $\mathbf{\Sigma}$ equal ρ for some $0 \leq \rho \leq 1$. The nonnegativity of ρ comes from Proposition 1. It can be easily seen that when $\rho = 0$, the component states $\boldsymbol{X}$ are independent. Figure 3.1 shows the system reliabilities of series and parallel systems as a function of ρ for different choices of π. It clearly shows that the series system becomes more reliable as ρ increases, while the parallel system becomes more reliable as ρ decreases. These results agree with the conclusions from Proposition 3.2.

3.2.2 Component Importance with Dependence

Component importance measures are used to evaluate the relative importance of each component to the system and have been applied in wide range of problems (Kuo and Zhu, 2012).

In general, the ranking of component importance is of more interest than the numerical values. Recall that the reliability importance of the jth component is defined as the probability that the jth component is critical to the system. That is, $\boldsymbol{I}_j = \mathbb{E}[\phi(1_j, \boldsymbol{X}_{-j})] - \mathbb{E}[\phi(0_j, \boldsymbol{X}_{-j})]$. It is clear that $\boldsymbol{I}_j$ depends on the joint distribution of $\boldsymbol{X}$; therefore, both the marginals and the dependence structure can affect the ranking of component importance. When the component states are independent, it is known that the component with the lowest reliability is the most important to a series system ("a chain is as strong as its weakest link") (Barlow and Proschan, 1975). Similarly, it can be shown that the component with the highest reliability is the most important to a parallel system. Though the definition of reliability importance applies for dependent systems, no properties have been described for series and parallel systems with arbitrary dependence structures.

The copula framework established in the previous section can be also applied to study the influence of dependence on component importance. Note that in both series and parallel systems, two components are equivalent in the system structure and have the same reliability importance if they have the same component reliabilities. We are interested in how the dependence structure of component states can change the ranking of reliability importance. Consider the following theorem.

Theorem 3.3. *Assume a coherent system with p positively associated component states $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$, $p \geq 2$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ denote the component reliabilities. Let $\pi_j = \pi_k$. Let $C_{j,-k}$ denote the copula for the CDF of $(X_j, \boldsymbol{X}_{-(j,k)})$, and $C_{k,-j}$ the copula for the CDF of $(X_k, \boldsymbol{X}_{-(j,k)})$, where $\boldsymbol{X}_{-(j,k)}$ denotes the states of all the components except the jth and kth components.*

1. *For a series system, if $C_{k,-j}$ is more PUOD than $C_{j,-k}$, then $\boldsymbol{I}_j \geq \boldsymbol{I}_k$.*

2. *For a parallel system, if $C_{k,-j}$ is more PLOD than $C_{j,-k}$, then $\boldsymbol{I}_j \geq \boldsymbol{I}_k$.*

Theorem 3.3 characterizes the effect of dependence on the ranking of reliability importance. It can be easily seen that in dependent systems, two components may have different reliability

importance even if they have the same marginals. More specifically, the ordering of $\boldsymbol{I}_j$ and $\boldsymbol{I}_k$ depends on $C_{j,-k}$ and $C_{k,-j}$, which capture the dependence structure of X_j and X_k with the rest of component states. In other words, the component importance of the jth component relies on both the marginal reliability and its relationship to the other components.

We now provide an intuitive explanation of the conclusions in Theorem 3.3. Imagine for the jth component that we can replace it by a new component that is independent of the rest of the components such that the system reliability remains the same after the replacement. Denote the reliabilities of the original jth component and the new component as π_j and π_j^*, respectively. $\pi_j = \pi_j^*$ if X_j is independent of the rest of the components, and π_j can be different from π_j^* if the independence assumption fails. One interpretation for the new component is that the original jth component "behaves" like the new component to the system in terms of system reliability. Depending on the dependence structure of $\boldsymbol{X}$, π_i^* may differ from π_j^* even if π_i is exactly the same as π_j for some $i \neq j$.

Under this imaginary replacement, it can be shown that for a series system with $\pi_j = \pi_k$, if $C_{k,-j}$ is more PUOD than $C_{j,-k}$, then $\pi_j^* \leq \pi_k^*$. Loosely speaking, this means that the jth component "behaves" like a less reliable component to the system than the kth component under the independence assumption. Note that $\pi_j^* \leq \pi_k^*$ can hold regardless of the ordering of the true marginals π_j and π_k. Theorem 3.3 implies that even if the two components have the exactly same π_j's, the one with the smaller π_j^* is more important to the series system. This conclusion agrees with the "a chain is as strong as its weakest link" principle for independent series systems. Intuitively, this is reasonable, because π_j^* summarizes the information from both the marginal and the dependence structure, and now the "weakest" link depends on both the marginals and the dependence. Similarly, it can be shown that for a parallel system, the condition that $C_{k,-j}$ is more PLOD than $C_{j,-k}$ given $\pi_j = \pi_k$ implies $\pi_j^* \geq \pi_k^*$.

The previous explanation assumes that π_j^* always exists and lies in the range of 0 and 1. The following proposition justifies the existence of such replacement.

Proposition 3.3. *For a coherent system with arbitrary dependence structure, there always*

exists $0 \leq \pi_j^* \leq 1$*, such that replacing the* j*th component with a component that is independent of the rest of the components and has reliability* π_j^* *keeps the system reliability unchanged.*

We now consider the case of a Gaussian copula. Recall that a Gaussian copula is completely determined by the correlation matrix $\boldsymbol{\Sigma}$. It turns out that the conditions in Theorem 3.3 can be easily translated into element-wise orderings of the corresponding sub-matrices of $\boldsymbol{\Sigma}$ based on the following proposition.

Proposition 3.4. *If the joint distribution of component states* $\boldsymbol{X}$ *has marginal* $\boldsymbol{\pi}$ *and a Gaussian copula with correlation matrix* $\boldsymbol{\Sigma}$*, then the joint distribution* $\boldsymbol{X}_{-j}$ *has marginal* $\boldsymbol{\pi}_{-j}$ *and a Gaussian copula with correlation matrix* $\boldsymbol{\Sigma}_{(-j)\times(-j)}$*, where* $\boldsymbol{\Sigma}_{(-j)\times(-j)}$ *is the sub-matrix of* $\boldsymbol{\Sigma}$ *without the* j*th row and column.*

For an illustrative example, consider a four-component system where $(X_1, \ldots, X_4)$ has marginals $\pi_j = \pi$ for $j = 1, \ldots, 4$ and a Gaussian copula with correlation matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.1 & 0.2 & 0.4 \\ 0.1 & 1 & 0.5 & 0.7 \\ 0.2 & 0.5 & 1 & 0.9 \\ 0.4 & 0.7 & 0.9 & 1 \end{bmatrix}.$$

Though the components have the same marginals, their component importance can be different due to the dependence of the component states. By Proposition 3.4, $C_{1,-2}$ and $C_{2,-1}$ are still Gaussian copulas with correlation matrices

$$\boldsymbol{\Sigma}_{(-2)\times(-2)} = \begin{bmatrix} 1 & 0.2 & 0.4 \\ 0.2 & 1 & 0.9 \\ 0.4 & 0.9 & 1 \end{bmatrix}, \boldsymbol{\Sigma}_{(-1)\times(-1)} = \begin{bmatrix} 1 & 0.5 & 0.7 \\ 0.5 & 1 & 0.9 \\ 0.7 & 0.9 & 1 \end{bmatrix}.$$

From the results of Proposition 3.2, $C_{2,-1}$ is more PUOD and PLOD than $C_{1,-2}$. From Theorem 3.3, we obtain $\boldsymbol{I}_1 > \boldsymbol{I}_2$ for both series and parallel systems. Similarly, it can be easily

checked that $\boldsymbol{I}_1 > \boldsymbol{I}_2 > \boldsymbol{I}_3 > \boldsymbol{I}_4$ for both series and parallel systems in this example.

3.2.3 Extensions to Multi-State Systems

The copula framework established in the previous sections does not require a specific parametric assumption about the joint distribution of component states. This generality enables us to extend the previous results from binary systems to multi-state systems. In a multi-state system, some components can have a range of states of performance, varying from perfect functioning to complete failure with degraded states in between. The system thus can also have multiple states (Barlow and Wu, 1978; El-Neweihi et al., 1978). We consider dependent multi-state systems by assuming the component states are positively associated. Assume the state of the component, X_j, is a discrete random variable taking values in $\boldsymbol{S}$, $\boldsymbol{S} = \{0, 1, \ldots, K\}$, where K indicates perfect functioning and 0 indicates failure. That is, each component has $K + 1$ states. Denote $\pi_{jk} = \Pr(X_j = k)$ for $1 \leq j \leq p$ and $0 \leq k \leq K$. For those components with fewer than $K + 1$ states, we simply have some π_{jk}'s equal to zero. Denote by $\boldsymbol{\pi}$ the $p \times (K+1)$ matrix that stores all the values of π_{jk}'s. Note that a multi-state system reduces to a binary system when $K = 1$.

The structure function, $\phi(\boldsymbol{X})$, is defined as a mapping from $\boldsymbol{S}^n$ to $\boldsymbol{S}$. El-Neweihi et al. (1978) generalized the idea of binary coherent systems to multi-state systems, where $\phi(\boldsymbol{X})$ satisfies the following three conditions: (i) $\phi(\boldsymbol{X})$ is increasing in each component; (ii) For level k of component j, there exists a vector $\boldsymbol{x}_{-j}$ such that $\phi(k_j, \boldsymbol{x}_{-j}) = k$, with $\phi(l_j, \boldsymbol{x}_{-j}) \neq k$ for $l \neq k$, $j = 1, \ldots, p$ and $k = 0, \ldots, K$; (iii) $\phi(k, \ldots, k) = k$ for $k = 0, 1, \ldots, K$. Systems with structure functions satisfying these three conditions are referred to as *multi-state coherent systems*. Note that a binary coherent system requires (i) and (iii), and condition (ii) is automatically satisfied for $K = 1$. Examples of $\phi(\boldsymbol{X})$ include $\phi(\boldsymbol{X}) = \min_{j=1}^{p} X_j$ for multi-state series systems, and $\phi(\boldsymbol{X}) = \max_{j=1}^{p} X_j$ for multi-state parallel systems.

Given the marginal parameters, we are interested in the influence of dependence on the expected system state, $\mathbb{E}[\phi(\boldsymbol{X})]$. The expected system state can be seen as one of the generalizations of system reliability in the binary system. Under the definitions in Barlow and

Wu (1978) and El-Neweihi et al. (1978), a multi-state coherent system has minimal path sets $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_s$ and and minimal cut sets $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m$ if

$$\phi(\boldsymbol{X}) = \max_{1 \le j \le s} \min_{i \in \boldsymbol{P}_j} X_i = \min_{1 \le j \le m} \max_{i \in \boldsymbol{C}_j} X_i.$$

That is, the state of the multi-state system is the "worst" state in the "best" path set and the "best" state in the "worst" cut set. Based on these generalized minimal path sets and minimal cut sets, the following theorem extends the results of Theorem 3.2 to multi-state systems.

Theorem 3.4. *Consider a multi-state coherent system with p positively associated component states $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$, $p > 2$. Let $\boldsymbol{\pi}_{p \times (K+1)}$ denote the marginal parameters and C the copula function for the joint distribution of component states. Denote s the number of minimal path sets and m the number of minimal cut sets.*

1. *Define*

$$h_C(\boldsymbol{u}) = \sum_{j=1}^{s} (-1)^{j+1} \sum_{1 \le k_1 < \cdots < k_j \le s} \bar{C}\Big((\boldsymbol{0}_{-\cup_{i=1}^{j} \boldsymbol{P}_{k_i}}, \boldsymbol{u}_{\cup_{i=1}^{j} \boldsymbol{P}_{k_i}}) \Big).$$

   If $h_{C_2}(\boldsymbol{u}) \ge h_{C_1}(\boldsymbol{u})$ for any $\boldsymbol{u} \in [0,1]^p$, then $\mathbb{E}_{\boldsymbol{\pi}, C_1}[\phi(\boldsymbol{X})] \le \mathbb{E}_{\boldsymbol{\pi}, C_2}[\phi(\boldsymbol{X})]$.

   When the system structure is series, $h_C(\boldsymbol{u}) = \bar{C}(\boldsymbol{u})$.

2. *Define*

$$g_C(\boldsymbol{u}) = \sum_{j=1}^{m} (-1)^{j+1} \sum_{1 \le k_1 < \cdots < k_j \le m} C\Big((\boldsymbol{1}_{-\cup_{i=1}^{j} \boldsymbol{C}_{k_i}}, \boldsymbol{u}_{\cup_{i=1}^{j} \boldsymbol{C}_{k_i}}) \Big).$$

   If $g_{C_2}(\boldsymbol{u}) \ge g_{C_1}(\boldsymbol{u})$ for any $\boldsymbol{u} \in [0,1]^p$, then $\mathbb{E}_{\boldsymbol{\pi}, C_1}[\phi(\boldsymbol{X})] \ge \mathbb{E}_{\boldsymbol{\pi}, C_2}[\phi(\boldsymbol{X})]$.

   When the system structure is parallel, $g_C(\boldsymbol{u}) = C(\boldsymbol{u})$.

The reliability importance of a component in a binary coherent system is defined as the probability of that component being critical to the system. This definition can be naturally generalized to multi-state systems. For a multi-state system, Barlow and Wu (1978) define the jth component as critical to the system at the kth state if when the jth component in the kth

state, the system is in the kth state and when the jth component is *not* in the kth state, the system is *not* in the kth state.

They further define the importance measure of the jth component at the kth system state, denoted as I_{jk}, as the probability that the jth component is critical to the system at the kth state. Proposition 2.5 of Barlow and Wu (1978) shows that $\boldsymbol{I}_{jk} = \Pr(E_1 \cap E_2 \cap E_3)$, where $E_1 = \{\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j} : x_j = k \Rightarrow \phi(\boldsymbol{x}) = k\}$, $E_2 = \{\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j} : x_j < k \Rightarrow \phi(\boldsymbol{x}) < k\}$ and $E_3 = \{\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j} : x_j > k \Rightarrow \phi(\boldsymbol{x}) > k\}$. As expected, $\boldsymbol{I}_{j1} = \Pr(\phi(1, \boldsymbol{x}_{-j}) = 1, \phi(0, \boldsymbol{x}_{-j}) = 0)$ when $K = 1$, which agrees with the importance measure in a binary system. Based on this definition, the following theorem extends the results of Theorem 3.3 to a multi-state system.

Theorem 3.5. *Consider a multi-state coherent system with p positively associated component states $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$, $p > 2$. Let $\boldsymbol{\pi}_{p\times(K+1)}$ denote the marginal parameters. Let $\boldsymbol{\pi}_j = \boldsymbol{\pi}_{j'}$, where $\boldsymbol{\pi}_j$ is the jth row of $\boldsymbol{\pi}$ for $j = 1, \ldots, p$. Let $C_{j,-j'}$ be the copula for the CDF of $(X_j, \boldsymbol{X}_{-(j,j')})$ and $C_{j',-j}$ be the copula for the CDF of $(X_{j'}, \boldsymbol{X}_{-(j,j')})$.*

1. *For a series system, if $C_{j',-j}$ is more PUOD than $C_{j,-j'}$, then $\boldsymbol{I}_{jk} \geq \boldsymbol{I}_{j'k}$ for $k = 1, \ldots, K$.*

2. *For a parallel system, if $C_{j',-j}$ is more PLOD than $C_{j,-j'}$, then $\boldsymbol{I}_{jk} \geq \boldsymbol{I}_{j'k}$ for $k = 1, \ldots, K$.*

We add a final remark on the generalization to multi-state systems. The conclusion of Theorem 3.4 is with respect the expected system state $\mathbb{E}_{\boldsymbol{\pi},C}[\phi(\boldsymbol{X})]$. Another natural generalization of system reliability from binary systems to multi-state systems is the probability of the system being perfectly functioning, that is, $\Pr(\phi(\boldsymbol{X}) = K)$. It can be shown that the results still hold if one changes all of the expected system states in Theorem 3.4 into the probabilities of the system being perfectly functioning.

3.3 Implementation

In this section, we briefly discuss the implementation for modeling dependent systems using Gaussian copulas and Markov chain Monte Carlo (MCMC). The reasons for inference under

the Bayesian framework are twofold. First, the direct estimation approach is computationally prohibitive for discrete data (see examples in Smith and Khaled (2012)). Second, confidence intervals for copula parameters are readily available under the Bayesian paradigm, and these are important for assessing whether the associations between components are significant or not.

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ be the component states in the ith instance, and $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Consider augmented data $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^\intercal$. Write $F_j(x_{ij}^-)$ for the left-hand limit of F_j at x_{ij}, and $c(\boldsymbol{u}) = \partial C(\boldsymbol{u})/\partial \boldsymbol{u}$ for the density of the copula function C. Smith and Khaled (2012) showed for the joint density $f(\boldsymbol{x}, \boldsymbol{u}) = \prod_{i=1}^n \prod_{j=1}^p \mathbf{1}\big(F_j(x_{ij}^-) \leq u_{ij} \leq F_j(x_{ij})\big) c(\boldsymbol{u}_i)$, the marginal of $\boldsymbol{x}$ is exactly the density of the joint distribution of $\boldsymbol{x}$ with marginal F_j's and copula function C. Using the Gaussian copula density derived in Xue-Kun Song (2000), the augmented likelihood can be written as

$$
\begin{aligned}
f(\boldsymbol{x}, \boldsymbol{u} | \boldsymbol{\pi}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2} \prod_{i=1}^n \exp\big[\frac{1}{2} \Phi^{-1}(\boldsymbol{u}_i)^\intercal (\boldsymbol{I} - \boldsymbol{\Sigma}) \Phi^{-1}(\boldsymbol{u}_i)\big] \\
\times \prod_{j=1}^p \pi_j^{x_{ij}} (1 - \pi_j)^{1 - x_{ij}} \mathbf{1}\big(F_j(x_{ij}^-) \leq u_{ij} \leq F_j(x_{ij})\big),
\end{aligned}
$$

where $\Phi^{-1}(\boldsymbol{u}_i) = (\Phi^{-1}(u_{i1}), \ldots, \Phi^{-1}(u_{ip}))^\intercal$. Given a prior distribution $p(\boldsymbol{\pi}, \boldsymbol{\Sigma})$, the posterior distribution given the augmented data is simply $p(\boldsymbol{\pi}, \boldsymbol{\Sigma} | \boldsymbol{x}, \boldsymbol{u}) \propto p(\boldsymbol{\pi}, \boldsymbol{\Sigma}) \times f(\boldsymbol{x}, \boldsymbol{u} | \boldsymbol{\pi}, \boldsymbol{\Sigma})$.

The following sampling scheme follows the data augmentation scheme described in Smith and Khaled (2012). The main idea of data augmentation is to sample the augmented data and the parameters iteratively. The sampling repeats the following two steps:

1. Let $\boldsymbol{u}_{(j)} = (u_{1j}, \ldots, u_{nj})^\intercal$. For $j = 1, \ldots, p$,
   (a) Sample from $p(\pi_j | \pi_{k \neq j}, \boldsymbol{u}_{(k \neq j)}, \boldsymbol{\Sigma}, \boldsymbol{x})$,
   (b) Sample from $p(\boldsymbol{u}_{(j)} | \pi_j, \boldsymbol{u}_{(k \neq j)}, \boldsymbol{\Sigma}, \boldsymbol{x})$,
2. Sample from $p(\boldsymbol{\Sigma} | \boldsymbol{u})$.

In all of the numerical examples of this project, we choose a Unif(0,1) as the prior for the π_j's and $p(\boldsymbol{\Sigma}) \propto 1$ as the prior for $\boldsymbol{\Sigma}$, as suggested in Barnard et al. (2000). Under these prior

distributions, it can be shown that step (1b) is equivalent to sampling from a truncated normal distribution. For steps (1a) and (2), we use the Metropolis-Hasting algorithm with a Gaussian distribution centered at the current values as the proposal distribution.

3.4 Simulations

3.4.1 System Reliability with Dependence

In this subsection, we examine the performance of the Gaussian copula for modeling dependent system reliability. We consider two data generation processes. In the first, the joint component state $\boldsymbol{X}$ is generated via a Gaussian copula. In the second, we examine the flexibility and robustness of the Gaussian copula by inducing the dependence through a common shock model. These two models correspond to the two scenarios in which the copula model is correctly specified and mis-specified, respectively.

- Model 1: $p = 4$. The joint distribution of $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^\intercal$ follows Bernoulli marginals with $\boldsymbol{\pi} = (0.2, 0.4, 0.6, 0.8)^\intercal$ and a Gaussian copula with correlation matrix $\boldsymbol{\Sigma} = \rho \mathbf{1}_p^\intercal \mathbf{1}_p + (1-\rho)\mathbf{I}_p$ for some $0 \leq \rho \leq 1$, where $\mathbf{I}_p$ is p-by-p identity matrix.

- Model 2: $p = 4$. Assume there is a common shock with probability Pr(Shock). Given that the shock happens, all of the components fail. $\Pr(\boldsymbol{X} = \mathbf{0} | \text{Shock happens}) = 1$. Given the shock does not happen, the component states $\boldsymbol{X}$ are independent Bernoulli trials with $\boldsymbol{\pi} = (0.2, 0.4, 0.6, 0.8)^\intercal$.

We generate $n = 300$ independent observations of $\left\{\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i4})^\intercal\right\}_{1 \leq i \leq n}$. The Gaussian copula model (denoted as "Copula") is fitted using the non-informative priors and the data-augmented sampling procedure discussed in Section 3.3. We draw 510,000 samples, with the first 10,000 samples dropped for burn-in. The parameters $\boldsymbol{\pi}$ and $\boldsymbol{\Sigma}$ were saved every 10th iteration thereafter, resulting in 50,000 posterior samples. Four different system structures are considered in the experiments, including a series system, a parallel system, and a 3-out-of-4 system. We also

consider a mixed system with the structure function $\phi(\boldsymbol{X}) = \min\big\{\max(X_1, X_2), \max(X_3, X_4)\big\}$. The mixed system consists of two subsystems in series, where the first subsystem is the first two components in parallel and the second subsystem is the last two components in parallel.

For each of these system structures, the induced posterior distribution on system reliability is estimated, and the posterior median is recorded as the estimator for the corresponding system reliability. For comparison, we also fit an independence model (denoted as "Indep"), where the component states are assumed to be independent. Under the same Unif(0,1) priors for the π_j's, the posterior of π_j under the independence assumption is $Beta(1 + n_j, 1 + n - n_j)$, where n_j is the number of instances with $x_{ij} = 1$ among the n samples, $j = 1, \ldots, 4$. The whole procedure is repeated 100 times to examine the variability of the estimators for both the copula model and the independence model.

Table 3.1 summarizes the results based on the data generated by Model 1 with $\rho = 0, 0.2, 0.5$ and 0.8. We calculated the bias, variance, and the mean-square-error (MSE) of the posterior medians from the copula and independence models based on the 100 replications. Note that for $\rho = 0$, the component states are essentially independent. When $\rho = 0$, the estimators for the independence model have smaller MSE than those from copula model in all the four system structures. It can be seen that both the copula and independence models have small bias, and the difference in the MSE is mainly due to the larger variance of the copula estimator. The superiority of the independence model in this case is not surprising, because it takes advantage of the additional assumption of independence, which is not available for the copula model.

For the scenarios with $\rho > 0$, the estimators from the copula model dominate those from the independence model in MSE for all structures. The advantage of the copula model becomes more pronounced as ρ moves away from 0. A closer look at the results reveals that though the independence model tends to provide estimators with less variability, those estimators are biased. The estimators under the independence assumption systematically underestimate the reliability in series systems and overestimate the reliability in parallel systems, which agrees with the conclusions from Theorem 3.1. In contrast, the estimators from the copula model

Table 3.1: Results for system reliability based on Gaussian copula data generation for 100 replications

Dependence	Criterion	Method	Series	Parallel	3-out-of-4	Mixed
$\rho = 0$	Bias(%)	Copula	**0.02**	-0.16	**0.17**	0.44
		Indep	-0.06	**-0.09**	-0.34	**-0.38**
	Var($\times 10^{-4}$)	Copula	0.71	0.86	6.01	7.62
		Indep	**0.27**	**0.36**	**4.53**	**5.71**
	MSE($\times 10^{-4}$)	Copula	0.71	0.89	6.04	7.82
		Indep	**0.27**	**0.37**	**4.64**	**5.86**
$\rho = 0.2$	Bias(%)	Copula	**-0.07**	**0.33**	**0.14**	**0.19**
		Indep	-3.13	3.29	-3.24	1.43
	Var($\times 10^{-4}$)	Copula	2.00	1.53	**6.15**	**7.94**
		Indep	**0.45**	**0.37**	6.22	8.53
	MSE($\times 10^{-4}$)	Copula	**2.00**	**1.64**	**6.17**	**7.98**
		Indep	10.28	11.25	16.73	10.59
$\rho = 0.5$	Bias(%)	Copula	**-0.41**	**0.55**	**-0.12**	**0.17**
		Indep	-8.35	8.42	-7.29	3.38
	Var($\times 10^{-4}$)	Copula	2.61	2.64	**4.59**	**5.94**
		Indep	**0.39**	**0.34**	5.11	6.62
	MSE($\times 10^{-4}$)	Copula	**2.79**	**2.94**	**4.61**	**5.97**
		Indep	70.11	71.28	58.31	18.07
$\rho = 0.8$	Bias(%)	Copula	**-0.89**	**0.96**	**-0.23**	**0.26**
		Indep	-13.75	13.82	-10.06	6.14
	Var($\times 10^{-4}$)	Copula	4.92	3.80	**7.77**	**8.57**
		Indep	**0.65**	**0.57**	10.07	11.73
	MSE($\times 10^{-4}$)	Copula	**5.71**	**4.73**	**7.82**	**8.64**
		Indep	189.78	191.82	111.32	49.55

Table 3.2: Results of system reliability based on shock model data generation for 100 replications

Pr(Shock)	Criterion	Method	Series	Parallel	3-out-of-4	Mixed
0.3	Bias(%)	Copula	**1.38**	**3.91**	**0.41**	**-2.76**
		Indep	-1.70	17.35	-8.53	-4.36
	Var($\times 10^{-4}$)	Copula	0.73	7.08	2.69	5.10
		Indep	**0.32**	**2.54**	**1.68**	**4.29**
	MSE($\times 10^{-4}$)	Copula	**2.64**	**22.37**	**2.86**	**12.74**
		Indep	2.93	303.76	74.51	23.32
0.5	Bias(%)	Copula	**1.25**	**3.07**	**-0.14**	**-2.98**
		Indep	-1.67	21.82	-9.87	-7.39
	Var($\times 10^{-4}$)	Copula	0.56	**5.97**	2.21	3.70
		Indep	**<0.01**	6.28	**0.60**	**3.36**
	MSE($\times 10^{-4}$)	Copula	**2.13**	**15.42**	**2.23**	**12.58**
		Indep	2.80	482.56	98.19	58.01

correctly capture the effect of dependence on system reliability and control the bias even when component states are strongly correlated ($\rho = 0.8$). It is also interesting to note that in more complex systems, such as the 3-out-of-4 and the mixed systems, the copula model actually gives estimators with slightly smaller variance than the independence model when the dependence is high. Overall, the results suggest that the copula model is more desirable than simply ignoring dependence when there is no strong prior knowledge supporting the independence assumption.

Table 3.2 summarizes the results of Model 2 based on the common shock model with the probability of shock equal to 0.3 and 0.5. This nonzero probability induces dependence among the component states by encouraging the components to fail together. This kind of dependence model is common (Marshall and Olkin, 1967; Hokstad, 1988). Under this data generation method, the dependence structure modeled by the Gaussian copula is mis-specified. In general, the findings are similar to those from the previous example with correct model specification. More specifically, the estimators from the copula model still control the bias reasonably well and dominate the estimator assuming independence in terms of MSE for all the system structures. The results suggest that the Gaussian copula has flexibility in modeling the dependence structure and robustness against possible model mis-specification.

3.4.2 Component Importance with Dependence

In this subsection, we investigate the performance of modeling component importance under dependence using a Gaussian copula. We generate $n = 300$ data points. The component states are generated using Bernoulli marginals with constant component reliability $\pi_j = \pi = 0.6$ and a Gaussian copula with correlation matrix given in the example of Section 3.2.2. With the constant marginals, the ordering of the component importance is entirely determined by the dependence structure. The same MCMC sampling described in Section 3.4.1 is implemented to obtain posterior samples of $\boldsymbol{\pi}$ and $\boldsymbol{\Sigma}$. When the goal of the analysis is the rank of parameters, it has been shown that simply ranking the posterior means or medians can perform poorly (Laird and Louis, 1989). Let R_j denote the true rank of the jth component importance, which is defined as $R_j = \sum_{k=1}^{p} \mathbf{1}(\boldsymbol{I}_j \geq \boldsymbol{I}_k)$. The larger the component importance, the higher the rank. The posterior expected rank, $\hat{R}_j$, is defined by the summation of the posterior probabilities $\hat{R}_j = \sum_{k=1}^{p} \Pr(\boldsymbol{I}_j \geq \boldsymbol{I}_k | \boldsymbol{x})$. Note that the posterior expected rank may not be an integer. Following Laird and Louis (1989), we use the ordering of the posterior expected rank as the posterior estimate of the ordering of the component importance. The same procedure is repeated 100 times.

Table 3.3 summarizes the estimated ranking of component importance over 100 replications. The true values of component importance are also provided for each system structure. The estimated ranking of component importance from the independence model is almost uniform across the components. For the four system structures considered in the experiments, the components importance are the same if we assume constant component reliabilities and independence. The estimated ranking from the copula model, however, agrees with the true ranking of components importance, with high percentages over the replications. When the true components importance are separated by a relatively large gap, as is the case in the 3-out-4 system, the copula model demonstrates a high probability of recovering the true ordering. In the mixed system, where the true components importance are close to each other, the estimated ranking still agrees with the true ranking reasonably well, although the estimated order is less stable. In general,

Table 3.3: Counts of component importance rankings over 100 replications

Series Truth: $\boldsymbol{I}_1 = 0.42 > \boldsymbol{I}_2 = 0.37 > \boldsymbol{I}_3 = 0.33 > \boldsymbol{I}_4 = 0.29$								
	Copula				Indep			
Rank	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp3	Comp4
1st	98	2	0	0	31	19	19	31
2nd	2	96	2	0	23	27	24	26
3rd	0	2	98	0	25	24	30	21
4th	0	0	0	100	21	30	27	22
Parallel Truth: $\boldsymbol{I}_1 = 0.23 > \boldsymbol{I}_2 = 0.18 > \boldsymbol{I}_3 = 0.14 > \boldsymbol{I}_4 = 0.11$								
	Copula				Indep			
Rank	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp3	Comp4
1st	99	1	0	0	31	19	20	30
2nd	1	95	4	0	25	24	29	22
3rd	0	4	96	0	23	29	23	25
4th	0	0	0	100	21	28	28	23
3-out-of-4 Truth: $\boldsymbol{I}_1 = 0.18 < \boldsymbol{I}_2 = 0.24 < \boldsymbol{I}_3 = 0.29 < \boldsymbol{I}_4 = 0.34$								
	Copula				Indep			
Rank	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp3	Comp4
1st	0	0	1	99	32	18	20	30
2nd	0	0	99	1	21	29	23	27
3rd	0	100	0	0	26	23	29	22
4th	100	0	0	0	21	30	28	21
Mixed Truth: $\boldsymbol{I}_1 = 0.17 < \boldsymbol{I}_2 = 0.22 < \boldsymbol{I}_3 = 0.25 < \boldsymbol{I}_4 = 0.28$								
	Copula				Indep			
Rank	Comp1	Comp2	Comp3	Comp4	Comp1	Comp2	Comp3	Comp4
1st	0	0	4	96	25	27	25	23
2nd	5	23	68	4	27	21	22	30
3rd	9	64	27	0	23	29	25	23
4th	86	13	1	0	25	23	28	24

these findings suggest the superiority of copula model in estimating the ranking of component importance under dependence.

3.5 Real Data

In this section, we analyze the two stockpile test data sets described in Section 1.2 using copula models. We are especially interested in the performance of the copula model as compared to the independence model when we have data sets with small sample sizes.

The first data set consists of 169 tests on a four component system. The observed data can be denoted as $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{169})^\intercal \in \mathbb{R}^{169\times 4}$ and $\boldsymbol{x}_S = (x_{S,1}, \ldots, x_{S,169})^\intercal \in \mathbb{R}^{169}$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i4})^\intercal$, $x_{ij} = 1$ if the jth component passes the ith test and 0 otherwise, $x_{S,i} = 1$ if the system passes the ith test, $i = 1, \ldots, 169$, $j = 1, \ldots, 4$.

There are many missing values in the data. If we drop all of the tests with missing values, only 57 tests are left with complete data. It is possible that many of these missing values are actually "pass," because the data recorders were not consistent about recording successful trials. We consider two analyses: one using the complete data only, and one that fills data by treating the missing values as passes (denoted as dropped-missing data and fill-missing data, respectively). The sample sizes for the two data sets are 57 and 169, respectively.

We first calculate summary statistics for the data. The maximum likelihood estimator (MLE) of component reliability using component-level data is simply the percentage of passes for that component. Using dropped-missing data, we have $\hat{\pi}_1 = 42/57$, $\hat{\pi}_2 = 38/57$, $\hat{\pi}_3 = 55/57$, and $\hat{\pi}_4 = 48/57$. The naive estimate of system reliability under the independence assumption is $\prod_{j=1}^{4} \hat{\pi}_j = 39.9\%$. The MLE of system reliability using the system-level data is the percentage of system passes: for this data, we have $\hat{R} = 31/57 = 54.4\%$. Notice that the observed proportion of system passes is much higher than the estimate from the component-level data assuming independence.

We also calculate the summary statistics for the fill-missing data and find $\hat{\pi}_1 = 154/169$, $\hat{\pi}_2 = 121/169$, $\hat{\pi}_3 = 157/169$, and $\hat{\pi}_4 = 160/169$. The MLE of system reliability using the

system-level data is $\hat{R} = 104/169 = 61.5\%$, which is still quite different from $\prod_{j=1}^{4} \hat{\pi}_j = 57.4\%$. Such deviations in the estimates of system reliability are frequently observed in practice and can be results from ignorance or improper modeling of dependence and/or wrongly specified system structure (Graves et al., 2010). We assume the system structure is correctly specified and that the discrepancies in system reliability come from dependence of the component states.

According to the results in Theorem 3.1, the summary statistics suggest positive dependence. Our goal is to capture this dependence via a copula model and provide a more accurate estimate of system reliability and component importance. Note that Anderson-Cook (2008) proposed an estimator using both component-level and system-level data. In our analyses, we use the component-level data only and treat the system-level data as ground truth on system reliability to evaluate the performance.

The non-informative priors from Section 3.3 are used in the analysis. We implement the MCMC procedure described in Section 3.4 for both the complete and imputed data to get 50,000 posterior samples of $\{\boldsymbol{\pi}, \boldsymbol{\Sigma}\}$. We also sample 50,000 $\boldsymbol{\pi}$'s from the posterior under an independence assumption. Figure 3.2 shows the posterior of $\boldsymbol{\pi}$ under both the copula and independence models using the dropped-missing data. It is clear that the two models give almost identical results on the marginals. This is not surprising, as the copula model decomposes the joint distribution into marginals and dependence structure separately. Similar results are seen for the fill-missing data, so the plot is omitted. Figure 3.3 summarizes the posterior distribution of $\boldsymbol{\Sigma}$ using the dropped-missing data. As expected, some off-diagonal elements have large posterior probabilities above zero, suggesting the corresponding components tend to pass and fail together. The posterior distribution of $\boldsymbol{\Sigma}$ using the fill-missing data is similar, and is shown in Figure 3.4.

To examine the performance of modeling system reliability, the posterior distribution of R, computed using only component-level data, is summarized in Figure 3.5. For both the dropped-missing data and the fill-missing data, the observed system reliability tends to be much higher than the posterior median of the independence model, suggesting that the independence model

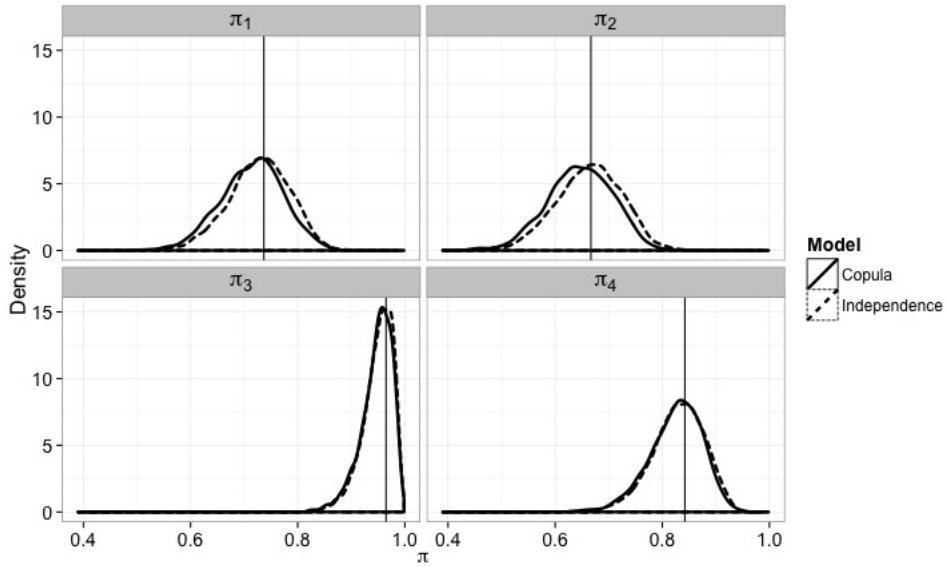

Figure 3.2: Posterior densities of $\boldsymbol{\pi}$ from the first data set with dropped-missing data. The copula model and independence model give almost identical marginals estimations, which agree with the MLE's (vertical lines).

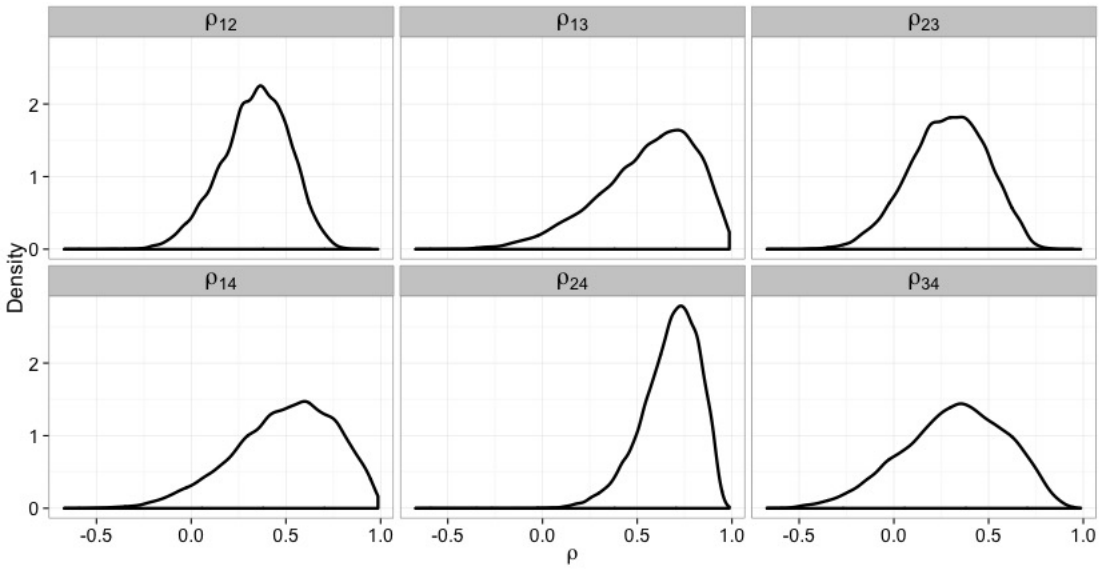

Figure 3.3: Posterior densities of $\boldsymbol{\Sigma} = (\rho_{ij})$ from the first data set with dropped-missing data. The positive associations are captured by $\rho > 0$.

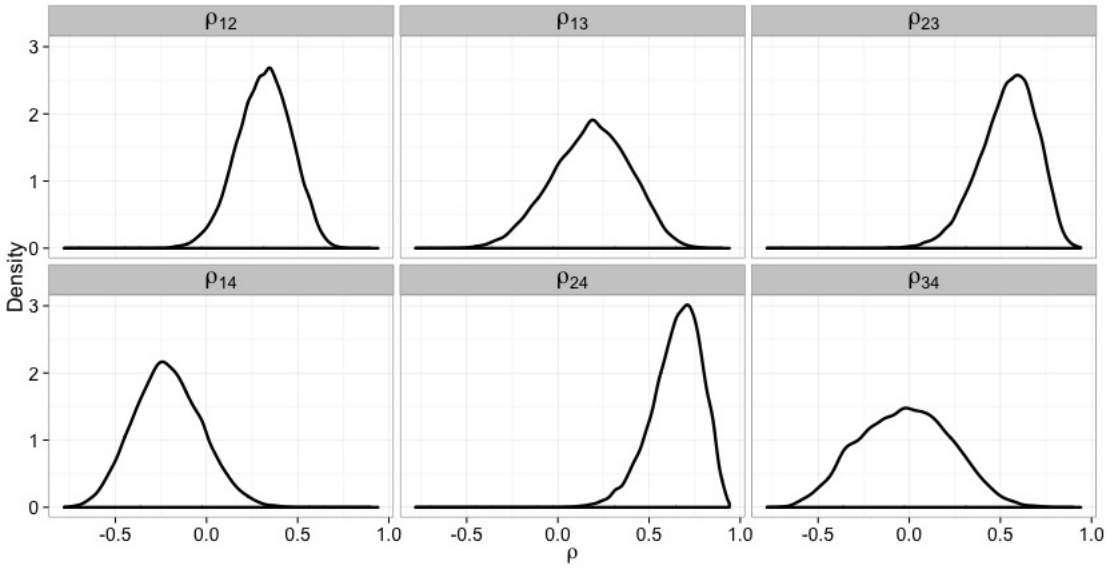

Figure 3.4: Posterior distributions of $\mathbf{\Sigma} = (\rho_{ij})$ from the first data set with the fill-missing data.

underestimates the system reliability. To further adjust the variability of the observed data, we check the posterior predictive distributions. More specifically, for each posterior sample of system reliability, we calculate the number of predicted system passes by sampling from a binomial distribution with number of trials equals to the number of tests in the data set. The histograms in Figure 3.6 clearly show that the independence model fits the data poorly, while the copula model agrees with the observed system-level data reasonably well. These results support that even for small sample sizes (here, $n = 57$), capturing dependence can improve estimation.

Based on the posterior distributions of $\boldsymbol{\pi}$ and $\mathbf{\Sigma}$, we can calculate the posterior distributions of component importance, $\boldsymbol{I}_j$'s. Figure 3.7 summarizes the posterior distributions of $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_4$. Recall that under the independence assumption, the component with the least reliability is the most important to the system. This is reflected in the posterior distributions from the independence model, where the posterior distribution of $\boldsymbol{I}_2$ is stochastically larger than the others. The posteriors from the copula model, however, suggest different results. Figure 3.7 shows that after

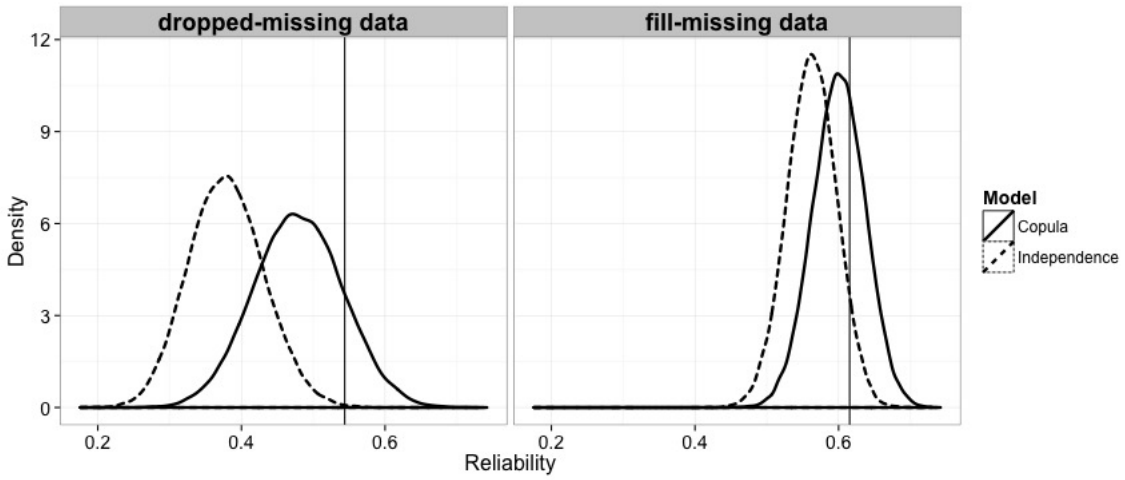

Figure 3.5: Posterior densities of system reliability from the first data set. The vertical line shows the true system reliability from system-level data. For either the dropped-missing data (n=57) or fill-missing data (n=169), the copula model fits the system-level data better.

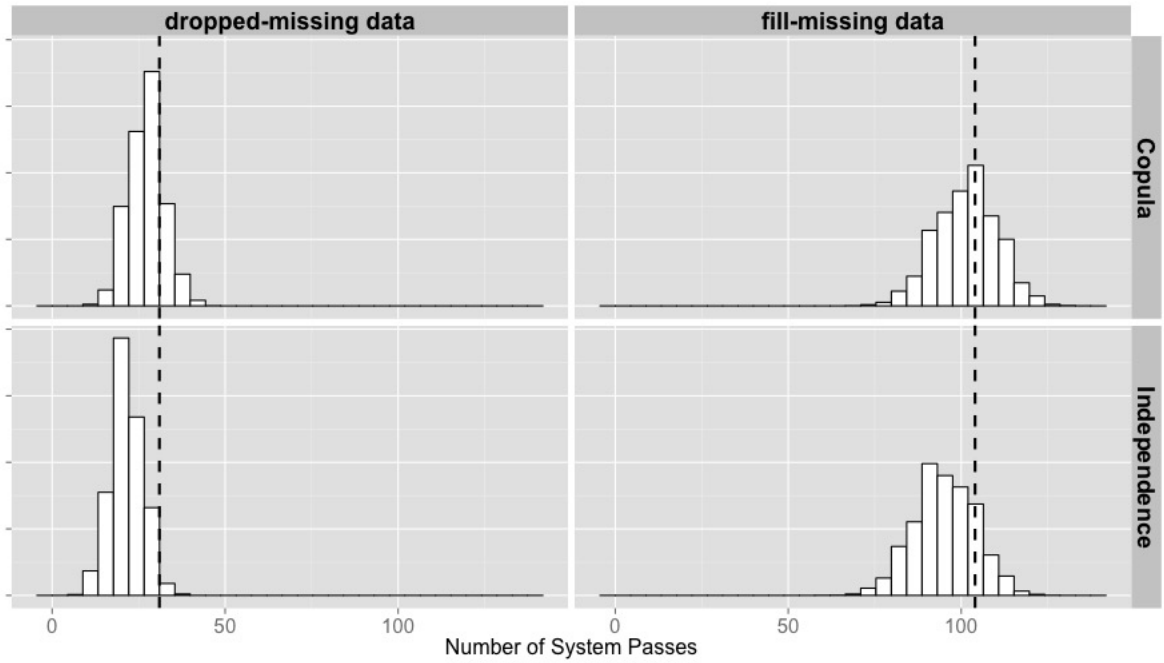

Figure 3.6: Posterior predictive distributions of system passes from the first data set. For either the dropped-missing data (n=57) or fill-missing data (n=169), the predictive distribution from copula model is closer to the observed number of system passes (dashed lines).

adjusting for the effects of dependence, component 1 and component 2 are equally important to the system, though their component reliabilities are different. We also consider component importance using the fill-missing data, and the results are similar for the copula and independence models. This example shows that in practice, the ranking of component importance relies on the underlying assumptions about the dependence structure.

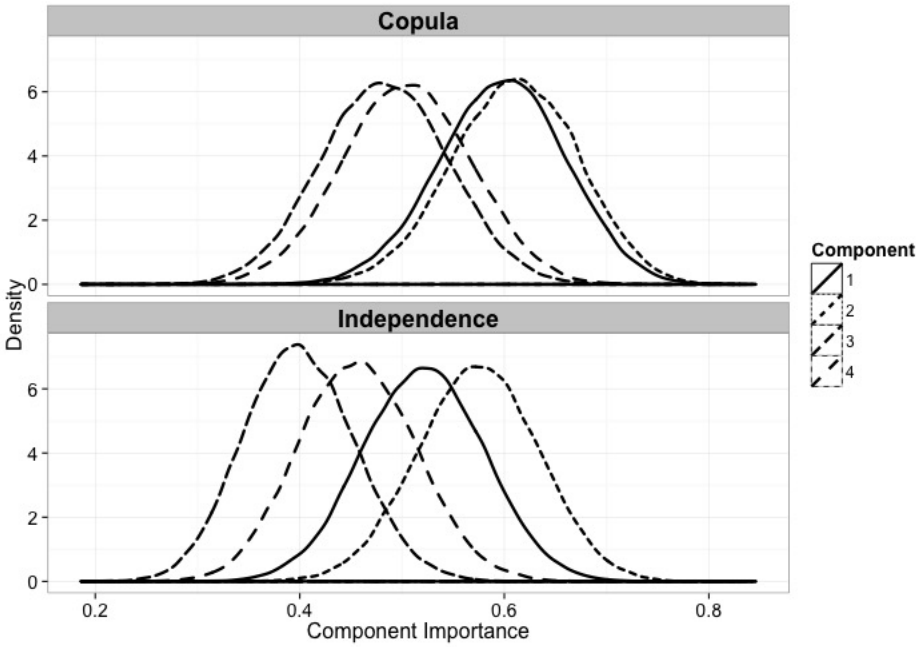

Figure 3.7: Posterior distributions of component importance from the first data set with dropped-missing data. The ranking of component reliability importance takes into account the dependence structure under the copula model.

We implement the same analyses on a second data set. The second data set consists of 181 tests on a four components system. If we drop all the 131 records with missing data, the naive estimate of system reliability under independence is 52.3%, while the observed percentage of system passes is 54%, suggesting that independence is a reasonable assumption in this example. The results from the copula and independence models are similar and are thus omitted. If we fill all the missing values as passes, the resulting naive estimate of system reliability under

independence is 34%, but the observed percentage of system passes is only 24%. According the results in Theorem 3.1, any coherent series system with positively associated components is at least as reliable as the series system under independence. This contradiction to our results suggests that the assumption of positively associated random variables is likely untrue. In fact, Figure 3.8 shows that some off-diagonal elements of $\mathbf{\Sigma}$ have a large posterior probability of being negative, which by Proposition 3.1 verifies that the positive association assumption is unlikely to hold. Although the component states are not positively associated, the copula model still correctly captures the correct system reliability, and the independence model suffers from considerable bias, as suggested by the histograms in Figure 3.9. Follow-up analyses are required to decide whether the negative dependence is acceptable or the discrepancy in system reliability is due to other reasons such as the mis-specified system structure. For one possible model to capture such a discrepancy, see Graves et al. (2010).

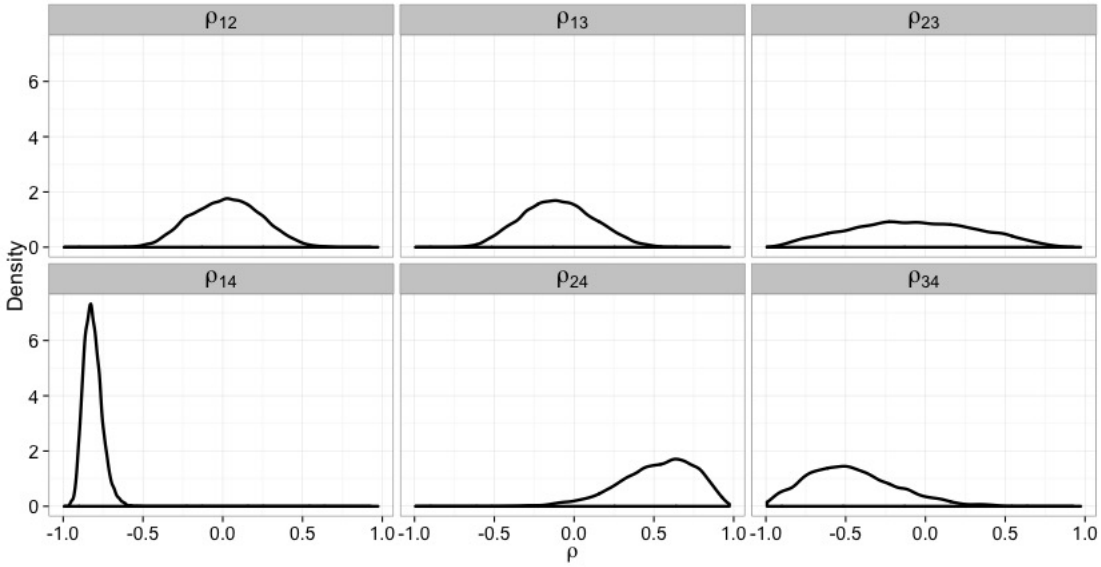

Figure 3.8: Posterior distributions of $\mathbf{\Sigma} = (\rho_{ij})$ from the second data set with the fill-missing data. It can be seen that ρ_{14} and ρ_{34} have large posterior probabilities of being negative, suggesting the positive association assumption among components fails.

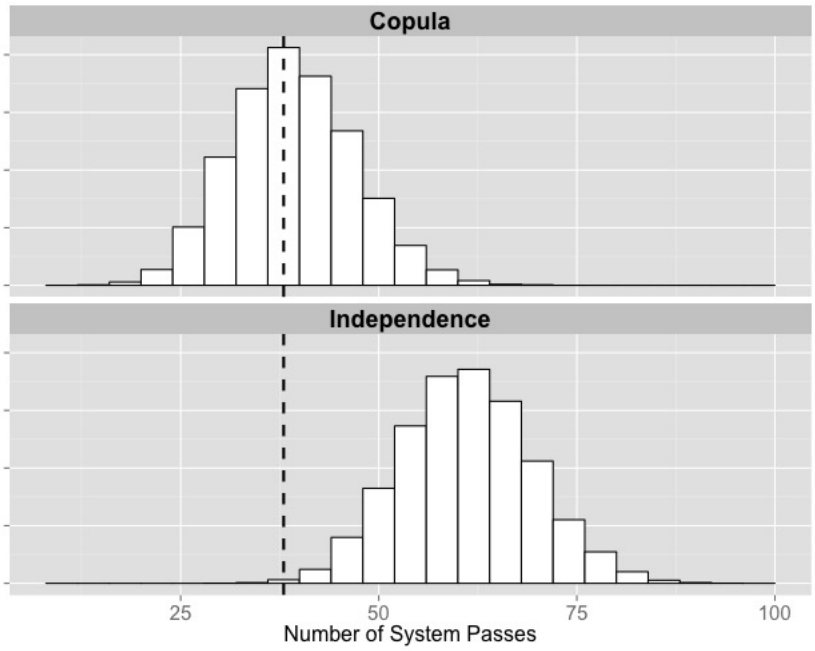

Figure 3.9: Posterior predictive distributions of system passes from the second data set using the fill-missing data. The dashed lines show the observed number of system passes. The copula model fits the observed system-level data better than the independence model.

3.6 Proofs

Proof of Theorem 3.1. We first introduce latent variables $\boldsymbol{U} = (U_1, \ldots, U_p)^\intercal$ such that the joint density of the component states $\boldsymbol{X}$ and the latent variables $\boldsymbol{U}$ is

$$f(\boldsymbol{X}, \boldsymbol{U}) = \prod_{j=1}^{p} \mathbf{1}\big(F_j(X_j^-) \leq U_j \leq F_j(X_j)\big) c(\boldsymbol{U}),$$

where $F_j(X_j^-)$ is the left-hand limit of F_j at X_j and $c(\boldsymbol{U}) = \partial C(\boldsymbol{U})/\partial \boldsymbol{U}$ is the density of the copula function C. Recall that for series systems, $R_{\boldsymbol{\pi},C} = \mathbb{E}_{\boldsymbol{\pi},C}[\phi(\boldsymbol{X})] = \Pr(\boldsymbol{X} = \mathbf{1}; \boldsymbol{\pi}, C)$. By Proposition 1 in Smith and Khaled (2012),

$$\begin{aligned}
\Pr(\boldsymbol{X} = \mathbf{1}; \boldsymbol{\pi}, C_1) &= \int \prod_{j=1}^{p} \mathbf{1}\big(F_j(1^-) \leq U_j \leq F_j(1)\big) c_1(\boldsymbol{U}) d\boldsymbol{U} \\
&= \int \prod_{j=1}^{p} \mathbf{1}\big(1 - \pi_j \leq U_j \leq 1\big) c_1(\boldsymbol{U}) d\boldsymbol{U} \\
&= \bar{C}_1(1 - \pi_1, \ldots, 1 - \pi_p),
\end{aligned}$$

where $\bar{C}_1(u_1, \ldots, u_p) = \Pr(U_1 > u_1, \ldots, U_p > u_p)$ is the corresponding survival function for the copula function C_1. By the definition of positive upper orthant dependent (PUOD), we have $R_{\boldsymbol{\pi},C_2} > R_{\boldsymbol{\pi},C_1}$ if C_2 is more PUOD than C_1.

The proof is similar for parallel systems and thus is omitted. □

Proof of Theorem 3.2. We prove the case for $h_C(\boldsymbol{u})$. The result for $g_C(\boldsymbol{u})$ can be shown using similar techniques. Notice that $R = \Pr(\phi(\boldsymbol{X}) = 1) = \Pr(\cup_{1 \leq j \leq s}(X_i = 1, i \in \boldsymbol{P}_j))$. By the

inclusion-exclusion formula,

$$
\begin{aligned}
\Pr(\cup_{1\le j\le s}(X_i=1, i\in \boldsymbol{P}_j)) =& \sum_{j=1}^{s}(-1)^{j+1}\sum_{1\le k_1<\cdots<k_j\le s}\Pr\left(X_i=1, i\in\cup\boldsymbol{P}_{k_j}\right)\\
=& \sum_{j=1}^{s}(-1)^{j+1}\sum_{1\le k_1<\cdots<k_j\le s}\int\prod_{i\in\cup\boldsymbol{P}_{k_j}}\mathbf{1}\big(F_i(1^-)\le U_i\le F_i(1)\big)c(\boldsymbol{U})d\boldsymbol{U}\\
=& \sum_{j=1}^{s}(-1)^{j+1}\sum_{1\le k_1<\cdots<k_j\le s}\bar{C}\Big((\mathbf{0}_{-\cup\boldsymbol{P}_{k_j}},(\mathbf{1}-\boldsymbol{\pi})_{\cup\boldsymbol{P}_{k_j}})\Big)\\
=& h_C(1-\pi_1,\ldots,1-\pi_p).
\end{aligned}
$$

If $h_{C_2}(\mathbf{1}-\boldsymbol{\pi})\ge h_{C_1}(\mathbf{1}-\boldsymbol{\pi})$, then $R_{\boldsymbol{\pi},C_2}\ge R_{\boldsymbol{\pi},C_1}$. □

Proof of Proposition 3.1. Write $\boldsymbol{\Sigma}=(\rho_{ij})$. We first prove that association of $\boldsymbol{X}$ implies $\rho_{ij}\ge 0$ for $1\le i,j,\le p$. We prove this by contradiction. Assume there exists some $\rho_{ij}<0$. By the properties of positively associated random variables, $\Pr(X_i=1,X_j=1)\ge\Pr(X_i=1)\Pr(X_j=1)$. By Proposition 1 of Smith and Khaled (2012),

$$
\Pr(X_i=1,X_j=1)=\Pr(\Phi^{-1}(U_i)>\Phi^{-1}(1-\pi_i),\Phi^{-1}(U_j)>\Phi^{-1}(1-\pi_j)),
$$

where Φ is the CDF of standard normal distribution, and the uniform random variables U_i and U_j satisfy

$$
(\Phi^{-1}(U_i),\Phi^{-1}(U_j))^{\intercal}\sim N(\mathbf{0},\begin{pmatrix}1 & \rho_{ij}\\ \rho_{ij} & 1\end{pmatrix}).
$$

Because $\rho_{ij}<0$ and Theorem A2 of Joe (1990),

$$
\begin{aligned}
&\Pr\left(\Phi^{-1}(U_i)>\Phi^{-1}(1-\pi_i),\Phi^{-1}(U_j)>\Phi^{-1}(1-\pi_j)\right)\\
&\qquad<\Pr\left(\Phi^{-1}(U_i)>\Phi^{-1}(1-\pi_i))\Pr(\Phi^{-1}(U_i)>\Phi^{-1}(1-\pi_i)\right).
\end{aligned}
$$

Note that $\Pr(\Phi^{-1}(U_i)>\Phi^{-1}(1-\pi_i))=\pi_i=\Pr(X_i=1)$. This contradicts with the property

of positively associated random variables and thus the proof is complete.

We next show $\boldsymbol{\Sigma} \geq 0$ element-wise implies $\boldsymbol{X}$ are positively associated. Assume $\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. By Corollary 4 of Joag-Dev et al. (1983), $\boldsymbol{Z}$ are positively associated random variables. To connect $\boldsymbol{Z}$ to $\boldsymbol{X}$, define the probability mass function $f(X_j) = \delta\Big(F_j^-\big(\Phi^{-1}(Z_j)\big)\Big)$, where δ is the Dirac mass function and F_j^- is the quantile function of the Bernoulli distribution with π_j, $j = 1, \ldots, p$. Smith and Khaled (2012) shows that under this definition, the distribution of $\boldsymbol{X}$ is exactly one with marginals $\boldsymbol{\pi}$ and copula function C. Note that F_j^- and Φ^{-1} are increasing functions. By property (P4) of Esary and Proschan (1970), the increasing functions of positively associated random variables are positively associated, which completes the proof. □

Proof of Proposition 3.2. The proof is a direct result of Theorem A2 of Joe (1990) and the fact that Φ^{-1} is an increasing function and thus is omitted here. □

Proof of Theorem 3.3. Using the same techniques in the proof of Theorem 3.1, it can be shown that in a series system,

$$\begin{aligned}
\boldsymbol{I}_j =& \Pr(X_k = 1, \boldsymbol{X}_{-(j,k)} = \mathbf{1}) - 0, \\
=& \int \prod_{k \neq j} \mathbf{1}\big(1 - \pi_k \leq U_k \leq 1\big) c_{k,-j}(\boldsymbol{U}_{-j}) d\boldsymbol{U}_{-j} \\
=& \bar{C}_{k,-j}(1 - \pi_k, \mathbf{1} - \boldsymbol{\pi}_{-(j,k)}),
\end{aligned}$$

where $c_{k,-j}$ is the density of the copula function $C_{k,-j}$. Similarly it can be shown $\boldsymbol{I}_k = \bar{C}_{j,-k}(1 - \pi_j, \mathbf{1} - \boldsymbol{\pi}_{-(j,k)})$. By the fact that $\pi_j = \pi_k$ and $C_{k,-j}$ is more PUOD than $C_{j,-k}$, we have $\boldsymbol{I}_j \geq \boldsymbol{I}_k$.

For parallel systems, the proof is similar and thus is omitted. □

Proof of Proposition 3.3. Denote the new component by X_j^*. It can be seen that

$$R = \pi_j^* \Pr\big(\phi(\boldsymbol{X}) = 1 \big| X_j^* = 1\big) + (1 - \pi_j^*) \Pr\big(\phi(\boldsymbol{X}) = 1 | X_j^* = 1\big).$$

By the independence between X_j^* and $\boldsymbol{X}_{-j}$,

$$\begin{aligned} R =& \pi_j^* \Pr\big(\phi(1_j, \boldsymbol{X}_{-j}) = 1\big) + (1 - \pi_j^*) \Pr\big(\phi(0_j, \boldsymbol{X}_{-j}) = 1\big) \\ =& \Pr\big(\phi(0_j, \boldsymbol{X}_{-j}) = 1\big) + [\Pr\big(\phi(1_j, \boldsymbol{X}_{-j}) = 1\big) - \Pr\big(\phi(0_j, \boldsymbol{X}_{-j}) = 1\big)]\pi_j^*. \end{aligned}$$

Because ϕ is increasing in each element, and it can only take values in $\{0, 1\}$, it can be easily seen that R is increasing in the reliability of the new component π_j^*. Therefore it suffices to show that $\Pr\big(\phi(0_j, \boldsymbol{X}_{-j}) = 1\big) \leq R \leq \Pr\big(\phi(1_j, \boldsymbol{X}_{-j}) = 1\big)$. This can be seen by writing R as $\Pr\big(\phi(X_j, \boldsymbol{X}_{-j}) = 1\big)$ and the increasing property and the range of ϕ. □

Proof of Proposition 3.4. Assume the joint distribution of $\boldsymbol{X}$ has Bernoulli marginals $\boldsymbol{\pi}$ and a Gaussian copula function C. Denote the probability mass function by $f(\boldsymbol{x}) = \Pr(\boldsymbol{X} = \boldsymbol{x})$. Note that

$$\Pr(\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}) = f(1_j, \boldsymbol{x}_{-j}) + f(0_j, \boldsymbol{x}_{-j}).$$

By Proposition 1 of Smith and Khaled (2012),

$$\begin{aligned} f(1_j, \boldsymbol{x}_{-j}) &= \int \prod_{k \neq j} \mathbf{1}\big(F_k(x_k^-) \leq U_k \leq F_k(x_k^-)\big)\big(1 - \pi_j \leq U_j \leq 1\big) c(\boldsymbol{U}) d\boldsymbol{U}, \\ f(0_j, \boldsymbol{x}_{-j}) &= \int \prod_{k \neq j} \mathbf{1}\big(F_k(x_k^-) \leq U_k \leq F_k(x_k^-)\big)\big(0 \leq U_j \leq 1 - \pi_j\big) c(\boldsymbol{U}) d\boldsymbol{U}. \end{aligned}$$

Therefore $\Pr(\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}) = \Pr(\cap_{k \neq j}\{F_k(x_k^-) \leq U_k \leq F_k(x_k^-)\})$, where $(\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_p))^\intercal$ follows a multivariate normal distribution with mean zero and correlation matrix $\boldsymbol{\Sigma}$. By the property of Gaussian distributions, $(\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_{j-1}), \Phi^{-1}(U_{j+1}), \ldots, \Phi^{-1}(U_p))^\intercal$ follows multivariate normal with mean zero and correlation matrix $\boldsymbol{\Sigma}_{(-j)\times(-j)}$. Denote $c_{(-j)\times(-j)}$ the density of a Gaussian copula $C_{(-j)\times(-j)}$ with correlation matrix $\boldsymbol{\Sigma}_{(-j)\times(-j)}$. Therefore,

$$\Pr(\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}) = \int \prod_{k \neq j} \mathbf{1}\big(F_k(x_k^-) \leq U_k \leq F_k(x_k)\big) c_{(-j)\times(-j)}(\boldsymbol{U}_{-j}) d\boldsymbol{U}_{-j}.$$

which completes the proof. $\square$

Proof of Theorem 3.4. We prove the case for $h_C(\boldsymbol{u})$. The results for $g_C(\boldsymbol{u})$ can be obtained similarly. Note that

$$\mathbb{E}[\phi(\boldsymbol{X})] = \sum_{k=0}^{K} k \Pr(\phi(\boldsymbol{X}) = k) = \sum_{k=1}^{K} \Pr(\phi(\boldsymbol{X}) \geq k).$$

Therefore, it suffices to show that if $h_{C_2}(\boldsymbol{u}) > h_{C_1}(\boldsymbol{u})$, $\Pr\big(\phi(\boldsymbol{X}) \geq k; \boldsymbol{\pi}, C_2\big) \geq \Pr\big(\phi(\boldsymbol{X}) \geq k; \boldsymbol{\pi}, C_1\big)$ for all $k = 1, \ldots, K$. By the definition of minimal path sets in multi-state systems,

$$\begin{aligned}
\Pr\big(\phi(\boldsymbol{X}) \geq k\big) =& \Pr(\max_{1\leq j\leq s} \min_{i\in \boldsymbol{P}_j} X_i \geq k) \\
=& \Pr(\cup_{1\leq j\leq s}(X_i \geq k, i \in \boldsymbol{P}_{k_j})) \\
=& \sum_{j=1}^{s} (-1)^{j+1} \sum_{1\leq k_1<\cdots<k_j\leq s} \Pr(X_i \geq k, i \in \cup \boldsymbol{P}_{k_j}) \\
=& \sum_{j=1}^{s} (-1)^{j+1} \sum_{1\leq k_1<\cdots<k_j\leq s} \int \prod_{i\in\cup \boldsymbol{P}_{k_j}} \mathbf{1}\big(F_i(k^-) \leq U_i \leq F_i(1)\big) c(\boldsymbol{U}) d\boldsymbol{U} \\
=& \sum_{j=1}^{s} (-1)^{j+1} \sum_{1\leq k_1<\cdots<k_j\leq s} \bar{C}\Big(\big(\mathbf{0}_{-\cup \boldsymbol{P}_{k_j}}, (\sum_{l<k} \boldsymbol{\pi}_l)_{\cup \boldsymbol{P}_{k_j}}\big)\Big), \\
=& h_C(\sum_{l<k} \pi_{1l}, \ldots, \sum_{l<k} \pi_{pl}),
\end{aligned}$$

where $\boldsymbol{\pi}_l$ is the lth column of the matrix $\boldsymbol{\pi}$. This implies that if

$$h_{C_2}(\sum_{l<k} \pi_{1l}, \ldots, \sum_{l<k} \pi_{pl}) \geq h_{C_1}(\sum_{l<k} \pi_{1l}, \ldots, \sum_{l<k} \pi_{pl}),$$

we have $\mathbb{E}_{\boldsymbol{\pi}, C_2}[\phi(\boldsymbol{X})] \geq \mathbb{E}_{\boldsymbol{\pi}, C_1}[\phi(\boldsymbol{X})]$. $\square$

Proof of Theorem 3.5. We prove the results for a series system. The result for parallel systems can be proved similarly. By Proposition 2.5 of Barlow and Wu (1978), $I_{jk} = \Pr(E_1 \cap E_2 \cap E_3)$, where $E_1 = \{\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j} : x_j = k \Rightarrow \phi(\boldsymbol{x}) = k\}$, $E_2 = \{\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j} : x_j < k \Rightarrow \phi(\boldsymbol{x}) < k\}$ and

$E_3 = \{\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j} : x_j > k \Rightarrow \phi(\boldsymbol{x}) > k\}$. Under $\phi(\boldsymbol{X}) = \min_j(X_j)$, it can be easily seen that $\Pr(E_1 \cap E_2 \cap E_3) = \Pr(\boldsymbol{X}_{-j} > \mathbf{k})$ for $k < K$ and $\Pr(E_1 \cap E_2 \cap E_3) = \Pr(\boldsymbol{X}_{-j} = \mathbf{k})$ for $k = K$. Using the same techniques in the proof of Theorem 3.4, it can be shown that for $k < K$

$$\begin{aligned}
&\Pr(\boldsymbol{X}_{-j} > \mathbf{k}) \\
=&\bar{C}_{j',-j}(\sum_{l\le k}\pi_{j'l}, \sum_{l\le k}\pi_{1l}, \ldots, \sum_{l\le k}\pi_{(j'-1)l}, \sum_{l\le k}\pi_{(j'+1)l}, \ldots, \sum_{l\le k}\pi_{(j-1)l}, \sum_{l\le k}\pi_{(j+1)l}, \ldots, \sum_{l\le k}\pi_{pl}).
\end{aligned}$$

Then the theorem holds for $k < K$ by the definition of PUOD. For $k = K$ the proof is similar and thus is omitted here. $\square$

Chapter 4

Longitudinal Tensor Regression

This chapter is organized as follows. We introduce the problem and notation in Section 4.1. The proposed method is discussed in Section 4.2, followed by implementation details in Section 4.3. Section 4.4 presents the main theoretical results. Simulations and a real data example are given in Section 4.5 and Section 4.6, respectively. Proofs can be found in Section 4.7.

4.1 Introduction

In recent years, an increasing number of *longitudinal* neuroimaging studies have rapidly emerged (Zhang et al., 2012), where brain images are collected for multiple subjects at multiple time points. Our motivating example is a study from the Alzheimer's Disease Neuroimaging Initiative (ADNI), described in Section 1.2. This data is both high dimensional and dependent. Each image is in the form of a multidimensional array, i.e., *tensor*. A $32 \times 32 \times 32$ MRI image involves $32^3 = 32,768$ parameters, whereas the number of subjects is only 88. In addition, the repeated images are also temporally correlated. Therefore, this longitudinal imaging analysis is particularly challenging.

Most existing neuroimaging methods use only the baseline imaging data, ignoring all the information at the follow-up time points. A small group of researchers recently started to use longitudinal images for individual-based classification (Misra et al., 2009; Davatzikos et al., 2009;

McEvoy et al., 2011; Hinrichs et al., 2011), and for cognitive score prediction (Zhang et al., 2012). There also emerged studies regressing longitudinal images on covariates (Skup et al., 2012; Li et al., 2013) and using functional principal components to quantify longitudinal images (Shinohara et al., 2011). However, despite those excellent efforts and the increasing availability of longitudinal imaging data, there is a paucity of effective solutions for longitudinal imaging analysis. There is thus a substantial demand for a systematic development of new longitudinal imaging analytical methods.

Suppose there are n training subjects, and for the i-th subject, there are observations over m_i time points. For simplicity, we assume $m_i = m$ and the same time points for all subjects. The observed data consist of $\{(Y_{ij}, \boldsymbol{X}_{ij}), i = 1, \ldots, n, j = 1, \ldots, m\}$, where Y_{ij} denotes the target response and $\boldsymbol{X}_{ij} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is a D-dimensional array representing the image. Write $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^\intercal$. A key attribute of longitudinal data is that the observations from different subjects are commonly viewed as independent, but the observations from the same subject are *correlated*. That is, the intra-subject covariance matrix, $\mathrm{Var}(\boldsymbol{Y}_i) \in \mathbb{R}^{m \times m}$, is not a diagonal matrix, but has some structure.

Since the seminal work of Liang and Zeger (1986), there has been a substantive literature on statistical analysis of longitudinal data. See Prentice and Zhao (1991); Li (1997); Qu et al. (2000); Xie and Yang (2003); Balan and Schiopu-Kratina (2005); Song et al. (2009); Wang (2011), among many others. However, all those studies take the covariates as a vector, whereas in imaging regression, covariates take the form of a multi-dimensional array. Naively turning an array into a vector would result in extremely high dimensionality and also destroy all inherent spatial information in images. There has been some recent development of statistical regression models for a scalar response on an image/tensor predictor; for instance, Caffo et al. (2010); Reiss and Ogden (2010); Zhou et al. (2013); Goldsmith et al. (2014); Wang et al. (2014), among others. Although those methods directly work with a tensor covariate, none has taken longitudinal tensors into account, and thus none is immediately applicable to our longitudinal imaging study.

We use the following notation throughout this chapter. The *inner product* between two tensors is $\langle \boldsymbol{B}, \boldsymbol{X} \rangle = \langle \mathrm{vec}\boldsymbol{B}, \mathrm{vec}\boldsymbol{X} \rangle = \sum_{i_1,\ldots,i_D} \beta_{i_1\ldots i_D} x_{i_1\ldots i_D}$, where the $\mathrm{vec}(\boldsymbol{B})$ *operator* stacks the entries of a tensor $\boldsymbol{B}$ into a column vector. The *outer product*, $\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D$, of D vectors $\boldsymbol{b}_d \in \mathbb{R}^{p_d}$ is a $p_1 \times \cdots \times p_D$ array with entries $(\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D)_{i_1\cdots i_D} = \prod_{d=1}^{D} b_{di_d}$. The *mode-d matricization*, $\boldsymbol{B}_{(d)}$, flattens a tensor $\boldsymbol{B}$ into a $p_d \times \prod_{d' \neq d} p_{d'}$ matrix such that the $(i_1, \ldots, i_D)$ element of the array $\boldsymbol{B}$ maps to the (i_d, j) element of the matrix $\boldsymbol{B}_{(d)}$, where $j = 1 + \sum_{d' \neq d}(i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$. A tensor $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is said to admit a *rank-R CANDECOMP/PARAFAC (CP) decomposition* (Kolda and Bader, 2009), if

$$\boldsymbol{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \tag{4.1}$$

where $\boldsymbol{\beta}_d^{(r)} \in \mathbb{R}^{p_d}, d = 1, \ldots, D, r = 1, \ldots, R$, are all column vectors, and $\boldsymbol{B}$ cannot be written as a sum of less than R outer products. The decomposition (4.1) is often represented by a shorthand, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$, where $\boldsymbol{B}_d = [\boldsymbol{\beta}_d^{(1)}, \ldots, \boldsymbol{\beta}_d^{(R)}] \in \mathbb{R}^{p_d \times R}$. If a tensor $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admits a rank-R decomposition (4.1), then

$$\boldsymbol{B}_{(d)} = \boldsymbol{B}_d(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1)^{\mathsf{T}} \text{ and } \mathrm{vec}\,\boldsymbol{B} = (\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1)\mathbf{1}_R,$$

where $\odot$ denotes the *Khatri-Rao product* (Rao and Mitra, 1971) of two matrices $\boldsymbol{B}_d \in \mathbb{R}^{p_d \times r}$ and $\boldsymbol{B}_{d'} \in \mathbb{R}^{p_{d'} \times r}$ such that $\boldsymbol{B}_d \odot \boldsymbol{B}_{d'} = \left[\boldsymbol{\beta}_d^{(1)} \otimes \boldsymbol{\beta}_{d'}^{(1)}, \boldsymbol{\beta}_d^{(2)} \otimes \boldsymbol{\beta}_{d'}^{(2)}, \ldots, \boldsymbol{\beta}_d^{(R)} \otimes \boldsymbol{\beta}_{d'}^{(R)}\right] \in \mathbb{R}^{p_d p_{d'} \times R}$, and $\otimes$ denotes the *Kronecker product*.

4.2 Methodology

The GEE method has been widely employed for analyzing correlated longitudinal data since the pioneering work of Liang and Zeger (1986). It requires specification of the first two moments of the conditional distribution of the response given the covariates, $\mu_{ij} = E(Y_{ij}|\boldsymbol{X}_{ij})$ and $\sigma_{ij}^2 = \mathrm{Var}(Y_{ij}|\boldsymbol{X}_{ij})$. Following Liang and Zeger (1986), we assume Y_{ij} is from an exponential family

with a canonical link. Then $\mu_{ij}(\boldsymbol{B}) = \mu(\theta_{ij})$ and $\sigma_{ij}^2(\boldsymbol{B}) = \phi\mu^{(1)}(\theta_{ij})$, $i = 1, \ldots, n,\ j = 1, \ldots, m$, where $\mu(\cdot)$ is a differentiable canonical link function, $\mu^{(1)}(\cdot)$ is its first derivative, θ_{ij} is the linear systematic part, and ϕ is an over-dispersion parameter. In this chapter we simply set $\phi = 1$, while the extension to a general ϕ is straightforward. θ_{ij} is associated with the covariates via the relation

$$\theta_{ij} = \langle \boldsymbol{B}, \boldsymbol{X}_{ij} \rangle, \tag{4.2}$$

where $\boldsymbol{B}$ is the coefficient tensor of the same size as $\boldsymbol{X}$ that captures effects of every array element of $\boldsymbol{X}$ on $\boldsymbol{Y}$. The GEE estimator of $\boldsymbol{B}$ is then defined as the solution of

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{B})}{\partial \mathrm{vec}(\boldsymbol{B})} \boldsymbol{V}_i^{-1} \big\{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B}) \big\} = \boldsymbol{0}, \tag{4.3}$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^\intercal$, $\boldsymbol{\mu}_i(\boldsymbol{B}) = [\mu_{i1}(\boldsymbol{B}), \ldots, \mu_{im}(\boldsymbol{B})]^\intercal$, and $\boldsymbol{V}_i = \mathrm{cov}(\boldsymbol{Y}_i)$ is the response covariance matrix of the i-th subject. The first component in (4.3) is the derivative of $\boldsymbol{\mu}_i(\boldsymbol{B})$ with respect to the vector $\mathrm{vec}(\boldsymbol{B}) \in \mathbb{R}^{\prod_d p_d}$. As such, there are a total of $\prod_d p_d$ estimating equations to solve in (4.3). For regression with image covariates, this dimension is prohibitively high and usually far exceeds the sample size.

In this work, we impose a low rank structure on the coefficient array $\boldsymbol{B}$. More specifically, we assume $\boldsymbol{B}$ in model (4.2) follows a CP structure in (4.1), $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$, where $\boldsymbol{B}_d = [\boldsymbol{\beta}_d^{(1)}, \ldots, \boldsymbol{\beta}_d^{(R)}] \in \mathbb{R}^{p_d \times R}$. Then the systematic part in (4.2) becomes

$$\begin{aligned} \theta_{ij} &= \langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \boldsymbol{X}_{ij} \rangle \\ &= \langle (\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1) \boldsymbol{1}_R, \mathrm{vec} \boldsymbol{X}_{ij} \rangle. \end{aligned} \tag{4.4}$$

Adopting (4.4), we propose the *tensor generalized estimating equations* estimator of $\boldsymbol{B}$, defined

as the solution of

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{B})}{\partial \boldsymbol{\beta}_{\boldsymbol{B}}} \boldsymbol{V}_i^{-1} \big\{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B}) \big\} = \boldsymbol{0}, \tag{4.5}$$

where $\boldsymbol{\beta}_{\boldsymbol{B}} = \mathrm{vec}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$, and the subscript $\boldsymbol{B}$ is to remind that $\boldsymbol{\beta}$ is constructed based on the CP decomposition of a given coefficient tensor $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$. Introducing the CP structure into GEE has two important implications. First, comparing to the classical GEE (4.3), the derivative in (4.5) is now with respect to $\boldsymbol{\beta}_{\boldsymbol{B}} \in \mathbb{R}^{R\sum_d p_d}$. Consequently, the number of estimating equations is reduced from the exponential order $\prod_d p_d$ to the linear order $R\sum_d p_d$. This substantial reduction in dimensionality is the key to enable effective estimation and inference with a limited sample size. Second, under this structure, any two elements $\beta_{i_1 \ldots i_d}$ and $\beta_{j_1 \ldots j_d}$ in $\boldsymbol{B}$ share common parameters if $i_d = j_d$ for any $d = 1, \ldots, D$. In consequence, the coefficients are correlated if they share the same spatial locations along any one of the tensor modes. This implicitly incorporates the spatial structure of the tensor coefficient.

Examining (4.5), the true intra-subject covariance structure $\boldsymbol{V}_i$ is usually unknown in practice. The classical GEE adopts a working covariance matrix, specified through a working correlation matrix $\boldsymbol{R}$. That is, $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}(\boldsymbol{B})\boldsymbol{R}\boldsymbol{A}_i^{1/2}(\boldsymbol{B})$, where $\boldsymbol{A}_i(\boldsymbol{B})$ is an $m \times m$ diagonal matrix with $\sigma_{ij}^2(\boldsymbol{B})$ on the diagonal and $\boldsymbol{R}$ is the m-by-m working intra-subject correlation matrix. Some commonly used correlation structures include independence, autocorrelation (AR), compound symmetry, and unstructured correlation, among others. The correlation matrix $\boldsymbol{R}$ may involve additional parameters, which can be estimated using residual-based moment method.

By both adopting this working covariance/correlation idea, and explicitly evaluating the derivative in (4.5), we finally arrive at the formal definition of the tensor GEE estimator, which is the solution $(\widehat{\boldsymbol{B}})$ of the following estimating equations

$$\sum_{i=1}^{n} [\boldsymbol{J}_1,\ \boldsymbol{J}_2, \ldots, \boldsymbol{J}_D]^{\intercal} \mathrm{vec}(\boldsymbol{X}_i) \boldsymbol{A}_i^{1/2}(\boldsymbol{B}) \widehat{\mathbf{R}}^{-1} \boldsymbol{A}_i^{-1/2}(\boldsymbol{B}) \big\{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B}) \big\} = \boldsymbol{0}, \tag{4.6}$$

where $\widehat{\boldsymbol{R}}$ is an estimated correlation matrix, $\mathrm{vec}(\boldsymbol{X}_i) = (\mathrm{vec}(\boldsymbol{X}_{i1}), \ldots, \mathrm{vec}(\boldsymbol{X}_{im}))$ is a $\prod_{d=1}^{D} p_d \times$

m matrix, $\boldsymbol{J}_d$ is the $\prod_{d=1}^{D} p_d \times Rp_d$ Jacobian matrix of the form $\boldsymbol{\Pi}_d \times [(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \otimes \boldsymbol{I}_{p_d}]$, where $\boldsymbol{\Pi}_d$ is the $(\prod_{d=1}^{D} p_d)$-by-$(\prod_{d=1}^{D} p_d)$ permutation matrix that reorders $\text{vec}\boldsymbol{B}_{(d)}$ to obtain $\text{vec}\boldsymbol{B}$, i.e., $\text{vec}\boldsymbol{B} = \boldsymbol{\Pi}_d \times \text{vec}\boldsymbol{B}_{(d)}$. Note that $\mu^{(1)}(\theta_{ij})$ has been canceled by the diagonals on the matrix $\boldsymbol{A}_i^{-1}$ due to the property of canonical link. For ease of presentation, we denote the left hand side of equation (4.6) as $\boldsymbol{s}(\boldsymbol{B})$, and write the tensor GEE (4.6) as $\boldsymbol{s}(\boldsymbol{B}) = \boldsymbol{0}$.

4.3 Implementation

Directly solving the tensor generalized estimating equations (4.6) with respect to $\boldsymbol{B}$ can be computationally intensive, as the mean of the response given the covariates is nonlinear in the parameters and the Jacobian matrices $\boldsymbol{J}_1, \ldots, \boldsymbol{J}_D$ also depend on the unknown parameters. We propose a block relaxation algorithm to iteratively solve the sub-GEE for $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D$ one-at-a-time, while keeping all other components fixed. Specifically, when updating $\boldsymbol{B}_d \in \mathbb{R}^{p_d \times R}$, the systematic part $\theta_{ij}(\boldsymbol{B})$ can be rewritten as

$$\theta_{ij}(\boldsymbol{B}) = \langle \boldsymbol{B}, \boldsymbol{X}_{ij} \rangle = \langle \boldsymbol{B}_d, \boldsymbol{X}_{ij(d)}(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \rangle,$$

where $\boldsymbol{X}_{ij(d)}$ is the mode-d matricization of the tensor $\boldsymbol{X}_{ij}$. As such, the systematic part $\theta_{ij}(\boldsymbol{B})$ becomes linear in $\boldsymbol{B}_d$. The Jacobian matrix $\boldsymbol{J}_d$ is free of $\boldsymbol{B}_d$ and depends on the covariates and fixed parameters only. Consequently, each step reduces to a standard GEE problem with Rp_d parameters, which can be solved using standard statistical softwares.

Another problem of importance is to choose the rank R for the coefficient array $\boldsymbol{B}$ in its CP decomposition. This can be viewed as a *model selection* problem. Pan (2001) proposed a quasi-likelihood independence model criterion for the classical GEE model selection, where the core idea is to evaluate the likelihood under the independence working correlation assumption.

In our tensor GEE setup, we adopt a similar idea,

$$\mathrm{BIC}(R) = -2\ell(\widehat{\boldsymbol{B}}(R); \boldsymbol{I}_m) + \log(n)p_e, \tag{4.7}$$

where $\ell(\widehat{\boldsymbol{B}}(R); \boldsymbol{I}_m)$ is the log-likelihood evaluated at the tensor GEE estimator $\widehat{\boldsymbol{B}}(R)$, with a working rank R and the independence working correlation structure $\boldsymbol{I}_m$. For simplicity, we call this criterion BIC, as the term $\log(n)$ is used. We choose R that minimizes this criterion among a series of working ranks.

We also consider adding regularization into our proposed method to achieve region selection. Selecting brain subregions that are highly relevant to the disease outcome is of vital scientific interest. It allows researchers to concentrate on subregions of brains for improved understanding of the disease pathology, and for hypothesis generation and validation. In our setup, this region selection translates to sparse estimation of the elements of the coefficient tensor $\boldsymbol{B}$, and is analogous to the intensively studied variable selection in classical vector-valued regression. We adopt the L_1 type regularization to achieve the goal of region selection. Specifically, we consider the following regularized tensor GEE

$$n^{-1}\boldsymbol{s}(\boldsymbol{B}) - \begin{pmatrix} \partial_{\beta_{11}^{(1)}} P_\lambda(|\beta_{11}^{(1)}|, \rho_n) \\ \vdots \\ \partial_{\beta_{di}^{(r)}} P_\lambda(|\beta_{di}^{(r)}|, \rho_n) \\ \vdots \\ \partial_{\beta_{Dp_D}^{(R)}} P_\lambda(|\beta_{Dp_D}^{(R)}|, \rho_n) \end{pmatrix} = \mathbf{0}, \tag{4.8}$$

where $P_\lambda(|\beta|, \rho_n)$ is a scalar penalty function, ρ_n is the penalty tuning parameter, λ is an index for the penalty family, and $\partial_\beta P_\lambda(|\beta|, \rho_n)$ is the subgradient with respect to the argument β. We consider two specific penalty functions: the Lasso (Tibshirani, 1996) in which $P_\lambda(|\beta|, \rho_n) = \rho_n|\beta|$ with $\lambda = 1$, and the SCAD (Fan and Li, 2001), in which $\partial/\partial|\beta| P_\lambda(|\beta|, \rho_n) = \rho_n \left\{1_{\{|\beta| \le \rho_n\}} + (\lambda\rho_n - |\beta|)_+/(\lambda - 1)1_{\{|\beta| > \rho_n\}}\right\}$, $\lambda > 2$.

4.4 Theory

4.4.1 Regularity Conditions

We begin with a list of regularity conditions for the asymptotics of tensor GEE with a fixed number of parameters. Denote by $||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ the Euclidean norm of a vector $\boldsymbol{x}$ and $||\boldsymbol{X}||_F = \sqrt{\text{trace}\boldsymbol{X}^\top \boldsymbol{X}}$ the Frobenius norm of a matrix $\boldsymbol{X}$. Denote $\boldsymbol{N}_n$ the neighborhood of the true tensor coefficient $\{\boldsymbol{B} : ||\boldsymbol{\beta}_{\boldsymbol{B}} - \boldsymbol{\beta}_{\boldsymbol{B}_0}|| \leq \triangle n^{-1/2}\}$ for some constant $\triangle > 0$.

(A1) For some positive constant $c_1 > 0$,$||\boldsymbol{X}_{ij}||_F \leq c_1$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

(A2) The true value $\boldsymbol{B}_0$ of the unknown parameter lies in the interior of a compact parameter space $\mathcal{B}$ and follows a rank-R CP structure defined in (4.1).

(A3) Let $\boldsymbol{I}(\boldsymbol{B}) = n^{-1} \sum_{i=1}^{n} [\boldsymbol{J}_1,\ \boldsymbol{J}_2, \ldots, \boldsymbol{J}_D]^\top \text{vec}\boldsymbol{X}_i \text{vec}^\top \boldsymbol{X}_i [\boldsymbol{J}_1,\ \boldsymbol{J}_2, \ldots, \boldsymbol{J}_D]$. It is assumed that there exist two positive constants $c_2 < c_3$ such that

$$c_2 \leq \lambda_{\min}(\boldsymbol{I}(\boldsymbol{B})) \leq \lambda_{\max}(\boldsymbol{I}(\boldsymbol{B})) \leq c_3,$$

over the set $\boldsymbol{N}_n$, where $\lambda_{\min}$ and $\lambda_{\max}$ are smallest and largest eigenvalue, respectively. It is also assumed that on the same set $\boldsymbol{I}(\boldsymbol{B})$ has a constant rank.

(A4) The true intra-subject correlation matrix $\boldsymbol{R}_0$ has bounded eigenvalues from zero and infinity. There exists a positive definite matrix $\tilde{\boldsymbol{R}}$ with eigenvalues bounded away from zero and infinity, such that $\|\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}\|_F = O_p(n^{-1/2})$, where $\widehat{\boldsymbol{R}}$ is an estimator of the correlation matrix.

(A5) For $\delta > 0$ and $c_4 > 0$, $E(\|\boldsymbol{A}_i^{-1/2}(\boldsymbol{B}_0)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B}_0))\|)^{2+\delta} \leq c_4$ for all $1 \leq i \leq n$.

(A6) For some positive constant $c_5 > 0$, $||\partial\theta_{ij}(\boldsymbol{\beta}_{\boldsymbol{B}})/\partial\boldsymbol{\beta}_{\boldsymbol{B}}|| \leq c_5$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

(A7) Denote by $\mu^{(k)}(\theta_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 2, 3$, the k-th derivative of $\mu(\theta_{ij})$. For some positive constants $c_6 < c_7$, $c_8 > 0$, $c_6 < |\mu^{(1)}(\theta_{ij})| < c_7$, and $|\mu^{(k)}(\theta_{ij})| < c_8$,

over the set $\boldsymbol{N}_n$.

(A8) Denote by $\boldsymbol{H}_{ij}(\boldsymbol{B}) = \frac{\partial^2 \theta_{ij}(\boldsymbol{\beta}_{\boldsymbol{B}})}{\partial \boldsymbol{\beta}_{\boldsymbol{B}} \partial \boldsymbol{\beta}_{\boldsymbol{B}}^{\mathsf{T}}}$. That is, $\boldsymbol{H}_{ij}(\boldsymbol{B})$ is the Hessian matrix of the linear systematic part θ_{ij}. There exist two positive constants $c_9 < c_{10}$ such that for $i = 1, \ldots, n$ and $j = 1, \ldots, m$,

$$c_9 \leq \lambda_{\min}(\boldsymbol{H}_{ij}(\boldsymbol{B})) \leq \lambda_{\max}(\boldsymbol{H}_{ij}(\boldsymbol{B})) \leq c_{10},$$

over the set $\boldsymbol{N}_n$.

A few remarks are in order. Conditions (A2) and (A3) are required for model identifiability of tensor GEE (Zhou et al., 2013). We observe that, the matrix $\boldsymbol{I}(\boldsymbol{B})$ in (A3) is an $R\sum_{d=1}^{D} p_d \times R\sum_{d=1}^{D} p_d$ matrix, and thus (A3) is much weaker than the nonsingularity condition on the design matrix if one were to directly vectorize the tensor covariate. Condition (A4) is commonly imposed in the GEE literature. It only requires that $\widehat{\boldsymbol{R}}$ be a consistent estimator of some $\tilde{\boldsymbol{R}}$, in the sense $\|\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}\|_F = O_p(n^{-1/2})$. $\tilde{\boldsymbol{R}}$ needs to be well behaved in that it is positive definite with bounded eigenvalues from zero and infinity, but $\tilde{\boldsymbol{R}}$ does *not* have to be the true intra-subject correlation $\boldsymbol{R}_0$. This condition essentially leads to the robust feature in Theorem 4.1 that the tensor GEE estimate is consistent even if the working correlation structure is misspecified. Condition (A5) regulates the tail behavior of the residuals so that the noise cannot accumulate too fast, and we can employ the Lindeberg-Feller central limit theorem to control the asymptotic behavior of the residuals. Condition (A6) states that the gradients of the systematic part are well-defined. Condition (A7) concerns the canonical link and generally holds for common exponential families, for example, the binomial and the Poisson distributions. Condition (A8) ensures that the Hessian matrix $\boldsymbol{H}(\boldsymbol{B})$ of the linear systematic part, which is highly sparse, is well-behaved in a neighborhood of the true value.

4.4.2 Consistency and Asymptotic Normality

We first address two components involved in the estimating equations: the initial estimator and the correlation estimator. Recall the tensor GEE estimator $\widehat{\boldsymbol{B}}$ is obtained by solving the

equations

$$\sum_{i=1}^{n}[\boldsymbol{J}_1,\ldots,\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{B})\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{B})\big\{\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{B})\big\}=\boldsymbol{0},$$

where $\widehat{\boldsymbol{R}}$ is any estimator of the intra-subject correlation matrix satisfying condition (A4). Note that $\widehat{\boldsymbol{R}}$ is often obtained via residual-based moment method, which in turn requires an initial estimator of $\boldsymbol{B}_0$. Next, we examine some frequently used estimators of $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{R}}$.

A customary initial estimator $\widehat{\boldsymbol{B}}$ in the GEE literature is the one that assumes an independent working correlation. That is, one completely ignores potential intra-subject correlation, and the corresponding tensor GEE becomes

$$\sum_{i=1}^{n}[\boldsymbol{J}_1,\ldots,\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X}_i\big\{\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{B})\big\}=\boldsymbol{0}.$$

Denoting the equations as $\boldsymbol{s}_{init}(\boldsymbol{B})=\boldsymbol{0}$, and the solution as $\widehat{\boldsymbol{B}}_{init}$, the next Lemma shows that it is a consistent estimator of the true $\boldsymbol{B}_0$.

Lemma 4.1. *Under conditions (A1)-(A3) and (A5)-(A8), there exists a root $\widehat{\boldsymbol{B}}_{init}$ of the equations $\boldsymbol{s}_{init}(\boldsymbol{B})=\boldsymbol{0}$ satisfying*

$$\|\boldsymbol{\beta}_{\widehat{\boldsymbol{B}}_{init}}-\boldsymbol{\beta}_{\boldsymbol{B}_0}\|=O_p(n^{-1/2}).$$

Here $\boldsymbol{\beta}_{\boldsymbol{B}}=\mathrm{vec}(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D)$, and is constructed based on the CP decomposition of a given tensor $\boldsymbol{B}=[\![\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D]\!]$, as defined before. Given a consistent initial estimator of $\boldsymbol{B}_0$, there exist multiple choices for the working correlation structure, e.g., autocorrelation, compound symmetry, and the nonparametric structure (Balan and Schiopu-Kratina, 2005).

Next we establish the consistency and asymptotic normality of the tensor GEE estimator from (4.6).

Theorem 4.1. *Under conditions (A1)-(A8), there exists a root $\widehat{\boldsymbol{B}}$ of the equations $\boldsymbol{s}(\boldsymbol{B})=\boldsymbol{0}$ satisfying*

$$\|\boldsymbol{\beta}_{\widehat{\boldsymbol{B}}}-\boldsymbol{\beta}_{\boldsymbol{B}_0}\|=O_p(n^{-1/2}).$$

The key message of Theorem 4.1, as implied by condition (A4), is that the consistency of the tensor coefficient estimator $\widehat{\boldsymbol{B}}$ does *not* require the estimated working correlation $\widehat{\boldsymbol{R}}$ to be a consistent estimator of the true correlation $\boldsymbol{R}_0$. This protects us from potential misspecification of the intra-subject correlation structure. Such a robustness feature is well known for GEE estimator with vector-valued covariates. Theorem 4.1 confirms and extends this result to the tensor GEE case with image covariates. We also remark that, although the asymptotics of the classical GEE can in principle be generalized to the tensor data by directly vectorizing the coefficient array, the ultrahigh dimensionality of the parameters would have made the regularity conditions such as (A3) unrealistic. By contrast, Theorem 4.1 ensures that one could still enjoy the consistency and robustness properties, by taking into account the structural information of the tensor coefficient under the GEE framework.

Under condition (A4), we define

$$\begin{aligned}
\tilde{\boldsymbol{M}}_n(\boldsymbol{B}) &= \sum_{i=1}^{n}[\boldsymbol{J}_1,\ldots,\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{B})\tilde{\boldsymbol{R}}^{-1}\boldsymbol{R}_0\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{B})\mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i[\boldsymbol{J}_1,\ldots,\boldsymbol{J}_D],\\
\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{B}) &= \sum_{i=1}^{n}[\boldsymbol{J}_1,\ldots,\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{B})\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{B})\mathrm{vec}^{\mathsf{T}}\boldsymbol{X}_i[\boldsymbol{J}_1,\ldots,\boldsymbol{J}_D].
\end{aligned}$$

As we will show in the appendix, $\tilde{\boldsymbol{M}}_n(\boldsymbol{B})$ approximates the covariance matrix of $\boldsymbol{s}(\boldsymbol{B})$ in (4.6), while $\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{B})$ approximates the leading term of the negative gradient of $\boldsymbol{s}(\boldsymbol{B})$ with respect to $\boldsymbol{\beta}_{\boldsymbol{B}}$. Then the next theorem gives the asymptotic normality of the tensor GEE estimator.

Theorem 4.2. *Under conditions (A1)-(A8), for any vector* $\boldsymbol{b}\in\mathbb{R}^{R\sum_{d=1}^{D}p_d}$ *such that* $\|\boldsymbol{b}\|=1$, *we have*

$$\boldsymbol{b}^{\mathsf{T}}\tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{B}_0)\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{B}_0)\big(\boldsymbol{\beta}_{\widehat{\boldsymbol{B}}}-\boldsymbol{\beta}_{\boldsymbol{B}_0}\big)\to\mathrm{Normal}(0,1)\ \textit{in distribution.}$$

4.4.3 Rank Selection Consistency

Next we establish that the rank selected by BIC in (4.7) under the independent working correlation is a consistent estimator of the true rank. This result is useful in two ways. First, it

justifies, to some extent, the asymptotic study in the previous section under a known rank. Second, it improves our understanding of the *interaction* between the working correlation and the rank specification. That is, the rank selected under a potentially misspecified correlation structure is consistent, and the tensor GEE estimator under the true rank and a potentially misspecified correlation structure is consistent.

We employ the same regularity conditions (A1)-(A8) in Section 4.4.2, except that we replace (A3) by the following condition:

(A3*) There exist two positive constants $c_1 < c_2$ such that $c_1 \leq \lambda_{\min}(\boldsymbol{I}(\boldsymbol{B})) \leq \lambda_{\max}(\boldsymbol{I}(\boldsymbol{B})) \leq c_2$, for all parameter points $\boldsymbol{B}$ in the interior of the parameter space. In addition, the rank is constant over the set $\{\boldsymbol{B} : ||\boldsymbol{\beta}_{\boldsymbol{B}} - \boldsymbol{\beta}_{\boldsymbol{B}_0}|| \leq \triangle n^{-1/2}\}$ for some constant $\triangle > 0$.

The reason for requiring (A3*) is that we need to characterize the behavior of some underfitted estimators with rank smaller than the true rank. These underfitted estimators may not reside in the neighborhood of the true parameters. We note that, however, (A3*) is a fairly mild condition, and the difference between (A3*) and (A3) is small. This is because, when the dimension is fixed, $\boldsymbol{I}(\boldsymbol{B})$ has fixed dimensions, then the condition on bounded eigenvalues is essentially requiring the matrix to be non-singular. The next theorem establishes the rank selection consistency of the BIC in (4.7).

Theorem 4.3. *Let* $\widehat{R} = \arg\min BIC(R)$*, and* $R_0 = rank(\boldsymbol{B}_0)$*. For normal responses, under conditions (A1)-(A8) and the modified condition (A3*), we have*

$$\Pr(\widehat{R} = R_0) \to 1, \ \ as\ n \to \infty.$$

That is, with high probability, the rank selected by BIC recovers the true rank. From the model selection perspective, this rank selection consistency implies that neither the overfitted model with a higher rank nor the underfitted model with an insufficient rank is favored by BIC.

4.4.4 Region Selection Consistency

Recall that, under the CP tensor structure, the element in the coefficient tensor can be written as $\beta_{i_1 \ldots i_D} = \sum_{r=1}^{R} \boldsymbol{\beta}_{1i_1}^{(r)} \times \cdots \times \boldsymbol{\beta}_{Di_D}^{(r)}$. In the imaging application where $D = 3$, for example, the region at (i_1, i_2, i_3) is non-active if $\beta_{i_1,i_2,i_3} = 0$. This can be induced if one of $\{\boldsymbol{\beta}_{1i_1}^{(r)}, \boldsymbol{\beta}_{2i_2}^{(r)}, \boldsymbol{\beta}_{3i_3}^{(r)}\}$ is zero for each $r = 1, \ldots, R$. Therefore, correctly recovering the sparsity pattern of $\boldsymbol{\beta}_{\boldsymbol{B}_0}$ results in selection of active regions of the coefficient tensor $\boldsymbol{B}_0$. We next establish that, for the SCAD-based region selection in (4.8), such selection is consistent.

Theorem 4.4. *Under conditions (A1)-(A8), $\rho_n = o(1)$ and $n^{-1/2} \log n = o(\rho_n)$, these exists one solution, $\boldsymbol{\beta}_{\widehat{\boldsymbol{B}}}$, to the SCAD regularized tensor GEE such that*

$$\Pr(supp(\boldsymbol{\beta}_{\widehat{\boldsymbol{B}}}) = supp(\boldsymbol{\beta}_{\boldsymbol{B}_0})) \to 1, \ \ as\ n \to \infty,$$

where $supp(\boldsymbol{\beta})$ denotes the support of the vector $\boldsymbol{\beta}$.

4.5 Simulations

We adopt the following simulation setup. We generated the responses according to the normal linear model

$$\boldsymbol{Y}_i \sim \mathrm{MVN}(\boldsymbol{\mu}_i, \sigma^2 \mathbf{R}_0), \quad i = 1, \ldots, n,$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^\intercal$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})^\intercal$, σ^2 is a scale parameter, and $\boldsymbol{R}_0$ is the true $m \times m$ intra-subject correlation matrix. We have chosen $\boldsymbol{R}_0$ to be of an exchangeable (compound symmetric) structure with the off-diagonal coefficient $\rho_n = 0.8$. The mean function is of the form $\mu_{ij} = \boldsymbol{\gamma}^\intercal \boldsymbol{Z}_{ij} + \langle \boldsymbol{B}, \boldsymbol{X}_{ij} \rangle, i = 1, \ldots, n, j = 1, \ldots, m$, where $\boldsymbol{Z}_{ij} \in \mathbb{R}^5$ denotes the covariate vector, with all elements generated from a standard normal distribution, and $\boldsymbol{\gamma} \in \mathbb{R}^5$ is the corresponding coefficient vector, with all elements equal to one; $\boldsymbol{X}_{ij} \in \mathbb{R}^{64 \times 64}$ denotes the 2D matrix covariate, again with all elements from standard normal, and $\boldsymbol{B} \in \mathbb{R}^{64 \times 64}$ is the matrix coefficient. The entries of $\boldsymbol{B}$ take the value of 0 or 1, and contains a series of shapes as shown

in Figure 4.1, including "square", "T-shape", "disk", "triangle", and "butterfly". Our goal is to recover those shapes in $\boldsymbol{B}$ by inferring the association between Y_{ij} and $\boldsymbol{X}_{ij}$ after adjusting for $\boldsymbol{Z}_{ij}$.

We set $n = 500$ and $m = 4$ and show the tensor GEE estimates and the corresponding BIC values under three working ranks $R = 1, 2$, and 3 in Figure 4.1. We first assume that the correlation structure is correctly specified, and will study potential misspecification in the next section. In this setup, "square" has a true rank equal to 1, "T-shape" has a rank 2, and the remaining shapes have the highest possible rank 64. It is clearly seen from Figure 4.1 that, first, the tensor GEE produces a sound recovery of the true signal, even for the signals with high rank or natural shape, e.g., "disk" and "butterfly", and second, the BIC criterion (4.7) successfully identifies the correct or best approximate rank for all the signals.

We also investigate potential effect of correlation misspecification when the sample size is *small* or *moderate*. We chose the "butterfly" signal and fit the tensor GEE model with three different working correlation structures: exchangeable, which is the correct specification in our setup, autoregressive of order one (AR-1), and independent. Table 4.1 reports the averages and standard errors (in parenthesis) out of 100 simulation replicates of the squared bias, the variance, and the mean squared error (MSE) of the tensor GEE estimate. We observe that the estimator based on the correct working correlation structure, i.e., the exchangeable structure, performs better than those based on misspecified correlation structures. When the sample size is moderate ($n = 100$), all the estimators have comparable bias, while the difference in MSE mostly comes from the variance part of the estimator. This agrees with the theory that the choice of the working correlation structure affects the asymptotic variance of the estimator. When the sample size becomes relatively large ($n = 150$), all the estimators perform similarly by the scaling term of $n^{-1/2}$ on the variance. When the sample size is small ($n = 50$), all the estimators have relatively large bias, while the independence working structure yields similar results as the exchangeable structure. This suggests that, when the sample size is *limited*, using a simple independence working structure is probably preferable compared to a more complex

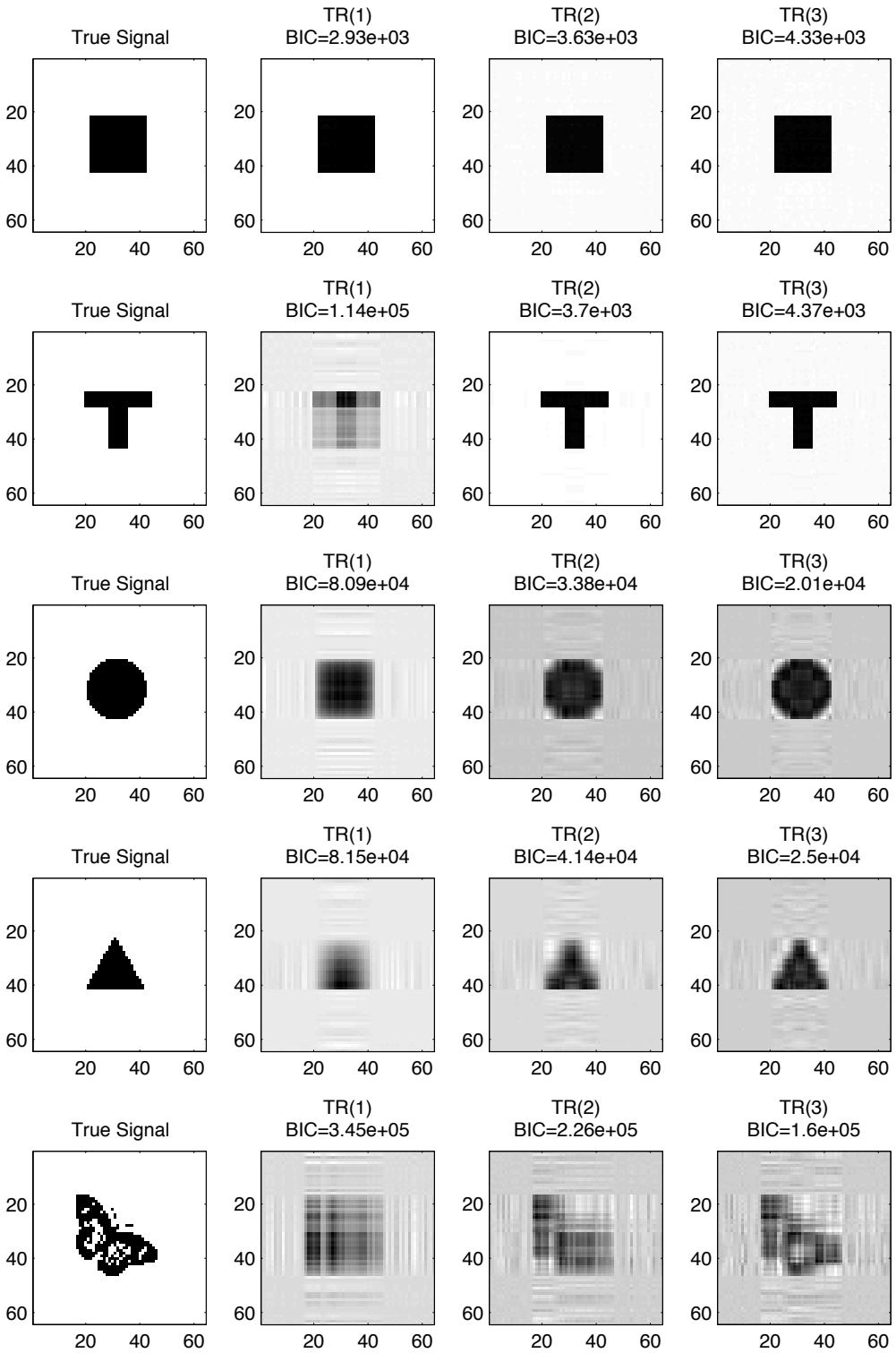

Figure 4.1: True and recovered image signals by the tensor GEE with varying ranks. $n = 500, m = 4$. The correlation structure is correctly specified. TR(R) means estimate from the rank-R tensor model.

Table 4.1: Bias, variance, and MSE of the tensor GEE estimates under various working correlation structures. Reported are the average out of 100 simulation replicates. The true intra-subject correlation is exchangeable with $\rho_n = 0.8$.

n	m	Working Correlation	Bias^2	Variance	MSE
50	10	Exchangeable	122.0	383.6	**505.6(7.9)**
		AR-1	139.1	530.0	669.1(15.8)
		Independence	119.1	393.9	513.0(11.0)
100	10	Exchangeable	85.8	128.9	**214.7(2.2)**
		AR-1	88.0	159.1	247.1(3.0)
		Independence	93.0	141.2	234.2(2.8)
150	10	Exchangeable	86.1	51.3	**137.2(0.6)**
		AR-1	85.6	56.0	141.6(0.6)
		Independence	84.9	62.3	147.2(0.9)

correlation structure.

Nevertheless, we should bear in mind that the above observations are for the average behavior of the estimate. Figure 4.2 shows two snapshots of the estimated signals under the three working correlations with $n = 100$. The top panel is one replicate where the estimates are "close" to the average in the sense that the bias, variance and MSE values for this single data realization are similar to those averages reported in Table 4.1. Consequently, the visual qualities of the three recovered signals are similar. The bottom panel, on the other hand, shows another replicate where the estimates are "far away" from the average. Then for this particular data, the quality of the estimated signal under the correct working correlation structure is superior than the ones under the incorrect specifications. Such an observation suggests that, as long as the sample size of the longitudinal imaging study is moderate to large, a longitudinal model should be favored over the one that totally ignores potential intra-subject correlation.

For the empirical performance of the regularized tensor GEE for region selection, we adopted the simulation setup described at the beginning of this sectionn, but varied the sample size n. We have implemented both the Lasso penalty and the SCAD penalty, and found their performances

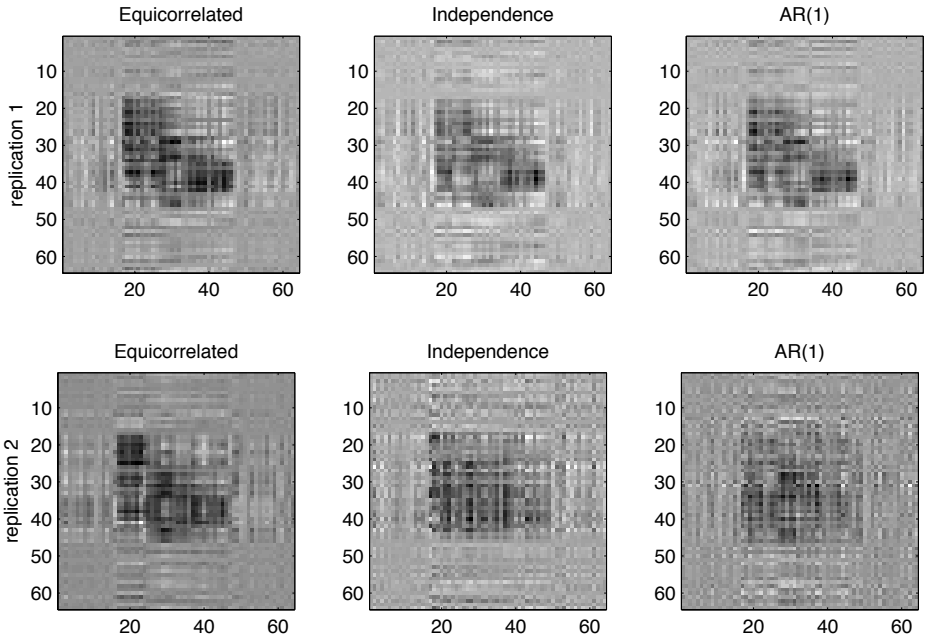

Figure 4.2: Snapshots of tensor GEE estimation with different working correlation structures. The true correlation is an equicorrelated structure. The comparison is row-wise. The first row shows a replicate where the estimates are "close" to the average behavior, and thus the visual quality of the estimates under different correlations structures are similar. The second row shows a replicate where the estimates are "far away" from the average, then the estimate under the correct correlation structure (panel 1) is superior than those under incorrect structures.

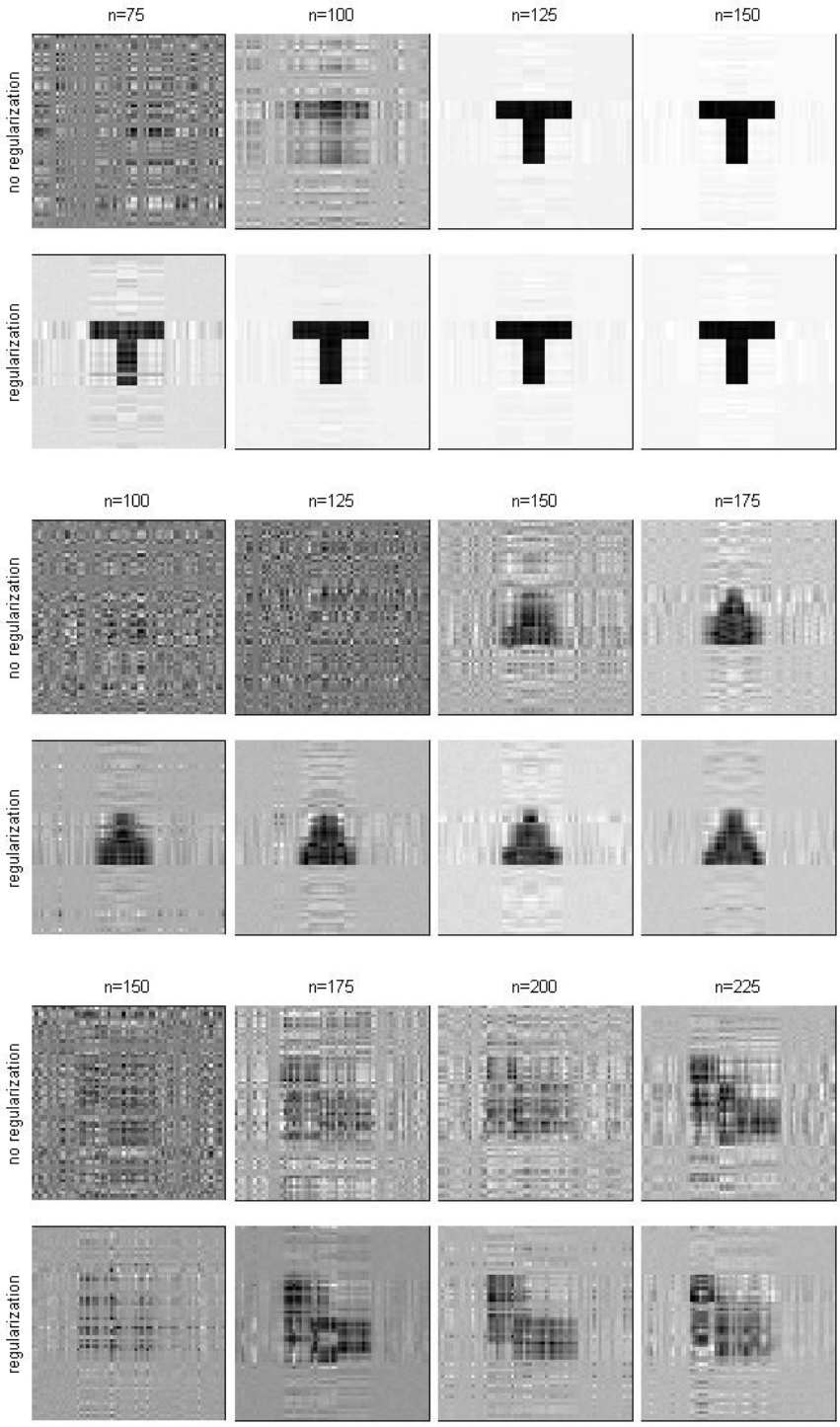

Figure 4.3: Comparison of tensor GEE estimation with and without regularization under varying sample size. $m = 4$. The matrix covariate is of size 64×64.

visually similar. As such we only report the results based on SCAD here. The estimates of three shapes, "T-shape", "triangle", and "butterfly", with and without regularizations, are shown in Figure 4.3. For the regularized tensor GEE, the penalty parameter λ was selected based on the prediction accuracy on an independent validation set. It is clearly seen from the plot that, while increasing sample size improves estimation accuracy for both tensor GEE and regularized tensor GEE, regularization leads to a more accurate recovery, especially when the sample size is limited.

4.6 Real Data

We consider the longitudinal imaging data described in Section 1.2. Two scientific goals are of interest. One is to predict the future clinical scores based on the data at previous time points, which is potentially useful for monitoring disease progression. The second is to identify brain subregions that are highly relevant to the disorder, and so to better understand the disease pathology. We fit a tensor GEE to this data. To evaluate the prediction accuracy, we first fit the tensor GEE using the data of all subjects from baseline to 12-month, and used prediction of MMSE at 18-month to select the tuning parameter. With the selected tuning parameter, we refit the model using the data from baseline to 18-month, and finally evaluate the prediction accuracy of all subjects using the "future" MMSE score at 24-month. The accuracy is evaluated by the root mean squared error (RMSE), $\{\sum_{i=1}^{n} n^{-1}(Y_{im} - \hat{Y}_{im})^2\}^{1/2}$, and the correlation, $\mathrm{Corr}(Y_{im}, \hat{Y}_{im})$. This evaluation scheme is the same as that of Zhang et al. (2012). Table 4.2 summarizes the results. It is seen that, for this real data, the best prediction was achieved under an AR(1) working correlation structure with the SCAD penalty. The corresponding RMSE and correlation were 2.147 and 0.781, which are only slightly worse than the best reported RMSE 2.035 and correlation 0.786 in Zhang et al. (2012). Note that Zhang et al. (2012) used multiple imaging modalities as well as additional biomarkers, which are supposed to improve the prediction accuracy, while our study utilizes only one imaging modality.

We applied both the Lasso and SCAD regularized tensor GEE to this data, and due to

Table 4.2: Prediction of future clinical MMSE scores using tensor GEE

	RMSE: $\{\sum_{i=1}^{n} n^{-1}(Y_{im} - \hat{Y}_{im})^2\}^{1/2}$			
Working Correlation	Independence	Equicorrelated	AR(1)	Unstructured
regularization (Lasso)	2.460	2.349	2.270	2.570
regularization (SCAD)	2.324	2.202	**2.147**	2.674
no regularization	2.526	2.427	2.429	2.628
	Correlation: $\mathrm{Corr}(Y_{im}, \hat{Y}_{im})$			
Working Correlation	Independence	Equicorrelated	AR(1)	Unstructured
regularization (Lasso)	0.705	0.733	0.747	0.700
regularization (SCAD)	0.742	0.767	**0.781**	0.658
no regularization	0.701	0.716	0.725	0.693

graphical similarity of the results, we report the SCAD estimate only. Figure 4.4 shows the estimate (marked in red) overlaid on an image of an arbitrarily chosen subject, with three views, top, side and bottom, respectively. The identified anatomical regions mainly correspond to cerebral cortex, part of temporal lobe, parietal lobe, and frontal lobe (Braak and Braak, 1991; Desikan et al., 2009; Yao et al., 2012). With AD, patients experience significant widespread damage over the brain, causing shrinkage of brain volume (Yao et al., 2012; Harasty et al., 1999) and thinning of cortical thickness (Desikan et al., 2009; Yao et al., 2012). The affected brain regions include those involved in controlling language (Broca's area) (Harasty et al., 1999), reasoning (superior and inferior frontal gyri) (Harasty et al., 1999), part of sensory area (primary auditory cortex, olfactory cortex, insula, and operculum) (Braak and Braak, 1991; Lee et al., 2013), somatosensory association area (Yao et al., 2012; Tales et al., 2005; Mapstone et al., 2003), memory loss (hippocampus) (den Heijer et al., 2010), and motor function (Buchman and Bennett, 2011). It is interesting to note that these regions are affected starting at different stages of AD, indicating the capability of the proposed method to locate brain atrophies as disease progresses. Specifically, hippocampus, which is highly correlated to memory loss, is commonly detected at the earliest stage of the disease. Regions related to language, communication, and

motor functions are normally detected at the later stages of the disease. The fact that our findings are consistent with results reported in previous studies, particularly the longitudinal studies, demonstrates the efficacy of our proposed method in identifying correct biomarkers that are closely related to AD/MCI.

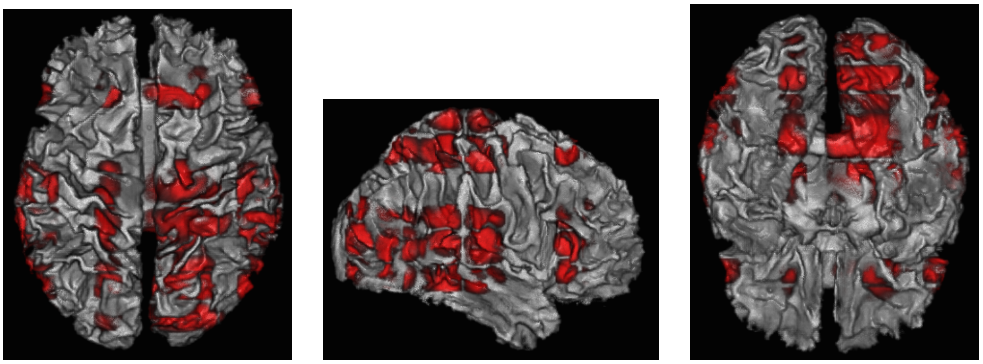

Figure 4.4: The ADNI data: regularized estimate overlaid on a randomly selected subject.

4.7 Proofs

Outline of the proofs:

To facilitate the proof, we introduce the following notation. Denote $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_{\hat{\boldsymbol{B}}}$ the estimator from tensor GEE and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{\boldsymbol{B}_0}$ the true values. Recall that the CP decomposition ensures that $\boldsymbol{B}$ is uniquely determined by $\boldsymbol{\beta}_n \in \mathbb{R}^{R\sum_{d=1}^{D} p_d}$. Denote $\boldsymbol{J}(\boldsymbol{\beta}) = [\boldsymbol{J}_1,\ \boldsymbol{J}_2, \cdots, \boldsymbol{J}_D]$, and note that under tensor structure $\partial\theta_{ij}/\partial\boldsymbol{\beta} = \boldsymbol{J}(\boldsymbol{\beta})^{\mathsf{T}}\mathrm{vec}\boldsymbol{X}_{ij}$. Recall the generalized estimating equations can be written as

$$\boldsymbol{s}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n)).$$

The proof of Lemma 4.1 is similar to that of Theorem 4.1 by dropping the terms involving the working correlation matrix and thus is omitted here.

The main technique to prove Theorem 4.1 is the sufficient condition for existence and consistency of a root of equations proposed in Ortega and Rheinboldt (2000), which also has been used in Portnoy (1984) for M-estimator and in Wang (2011) for GEE estimator with vector covariates. To check this condition, Lemmas 4.2 - 4.4 are proposed. Lemma 4.2 provides a useful approximation to the generalized estimating equations $\boldsymbol{s}_n(\boldsymbol{\beta}_0)$ based on the condition (A4) of the working correlation matrix. This facilitates the later evaluations of the moments of the generalized estimating equations by treating the intra-subject correlation as known. Lemma 4.3 further establishes the approximation of the negative gradients of the generalized estimating equations. Lemma 4.4 refines this approximation of the negative gradients at one more step, providing the foundations for the Talyor expansion of generalized estimating equations at the true value.

Based on Theorem 4.1, the proof of Theorem 4.2 is obtained by evaluating the covariance matrix of the generalized estimating equations and applying the Lindeberg-Feller central limit theorem.

The proof of Theorem 4.3 follows two steps. We show that BIC neither overestimates nor underestimates the true rank. By combing these results, the rank selection consistency is established.

Theorem 4.4 is proved by construction. We show that the oracle estimator is an approximated solution to the SCAD regularized tensor GEE.

Lemma 4.2. *Under conditions (A1)-(A8),* $||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) - \boldsymbol{s}_n(\boldsymbol{\beta}_0)|| = O_p(1)$*, where* $\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)$ *is* $\boldsymbol{s}_n(\boldsymbol{\beta}_0)$ *with* $\widehat{\boldsymbol{R}}$ *replaced by* $\tilde{\boldsymbol{R}}$*.*

Proof of Lemma 4.2. Consider

$$\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n)).$$

Denote by $\{r_{i,j}\}_{1\leq i,j\leq m}$ the (i,j)-th element of $\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}$. By condition (A4), $r_{i,j} = O_p(n^{-1/2})$.

By direct calculation,

$$
\begin{aligned}
&\boldsymbol{s}_n(\boldsymbol{\beta}_0) - \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} r_{j,m}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij} \\
=&\sum_{j=1}^{m}\sum_{k=1}^{m} r_{j,m}\Big[\sum_{i=1}^{n}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij}\Big],
\end{aligned}
$$

where $\epsilon_{ik}(\boldsymbol{\beta}_0) = \sigma_{ik}^{-1}(\boldsymbol{\beta}_0)(Y_{ik} - \mu_{ik}(\boldsymbol{\beta}_0))$. By condition (A5), (A6) and (A7),

$$
\mathbb{E}\Big[||\sum_{i=1}^{n}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij}||^2\Big] = O(n).
$$

Therefore, $||\sum_{i=1}^{n}\sigma_{ij}(\boldsymbol{\beta}_0)\epsilon_{ik}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\mathsf{T}}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_{ij}|| = O_p(\sqrt{n})$. Since $r_{i,j} = O_p(n^{-1/2})$, the proof is complete. □

Consider $\boldsymbol{D}_n(\boldsymbol{\beta}_n) = -\partial\boldsymbol{s}_n(\boldsymbol{\beta}_n)/\partial\boldsymbol{\beta}_n$, $\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) = -\partial\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_n)/\partial\boldsymbol{\beta}_n$. Lemma 4.3 establishes the approximation of the negative gradients of the estimating equations.

Lemma 4.3. *Under conditions (A1)-(A8), for some constant $\triangle > 0$,*

$$
\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}} |\lambda_{\max}[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \boldsymbol{D}_n(\boldsymbol{\beta}_n)]| = O_p(n^{1/2}),
$$

$$
\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}} |\lambda_{\min}[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \boldsymbol{D}_n(\boldsymbol{\beta}_n)]| = O_p(n^{1/2}).
$$

Proof of Lemma 4.3. Similar to Lemma C.1. of Wang (2011), it can be shown by direct calculation that

$$
\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) = \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n) + \tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n) + \tilde{\boldsymbol{D}}_{n3}(\boldsymbol{\beta}_n) + \tilde{\boldsymbol{D}}_{n4}(\boldsymbol{\beta}_n),
$$

where

$$\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n),$$

$$\tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n) = \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_i(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n),$$

$$\tilde{\boldsymbol{D}}_{n3}(\boldsymbol{\beta}_n) = -\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\boldsymbol{K}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n),$$

$$\tilde{\boldsymbol{D}}_{n4}(\boldsymbol{\beta}_n) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{e}_j^{\intercal}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n))\mathbf{H}(\boldsymbol{\beta}_n),$$

with

$$\mathbf{C}_i(\boldsymbol{\beta}_n) = \mathrm{diag}\Big(Y_{i1} - \mu_{i1}(\boldsymbol{\beta}_n), \ldots, Y_{im} - \mu_{im}(\boldsymbol{\beta}_n)\Big),$$

$$\boldsymbol{F}_i(\boldsymbol{\beta}_n) = \mathrm{diag}\Big(\mu_{i1}^{(2)}(\boldsymbol{\beta}_n), \ldots, \mu_{im}^{(2)}(\boldsymbol{\beta}_n)\Big),$$

$$\boldsymbol{K}_i(\boldsymbol{\beta}_n) = \mathrm{diag}\Big(\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n))\Big),$$

$\mathbf{e}_j^{\intercal}$ the length m vector with j-th element 1 and 0 everywhere else, and $\mathbf{H}(\boldsymbol{\beta}_n)$ is defined in condition (A8).

Let $\boldsymbol{D}_{ni}(\boldsymbol{\beta}_n)$ be defined the same as $\tilde{\boldsymbol{D}}_{ni}(\boldsymbol{\beta}_n)$, but with $\tilde{\boldsymbol{R}}$ replaced by $\widehat{\boldsymbol{R}}$, for $i = 1, \ldots, 4$. It is sufficient to prove

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \leq \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^{\intercal}[\boldsymbol{D}_{ni}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{ni}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2})$$

for any $\boldsymbol{u} \in \mathbb{R}^{R\sum_{d=1}^{D} p_d}$ such that $||\boldsymbol{u}|| = 1$, $i = 1, \ldots, 4$.

For $i = 1$, we have

$$
\begin{aligned}
&|\boldsymbol{u}^{\top}[\boldsymbol{D}_{n1}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| \\
\leq & n||\boldsymbol{u}||^2 \cdot ||\widehat{\boldsymbol{R}}^{-1} - \tilde{\boldsymbol{R}}^{-1}||_F \cdot \lambda_{\max}(\boldsymbol{A}_i(\boldsymbol{\beta}_n)) \cdot \lambda_{\max}\Big(n^{-1}\sum_{i=1}^{n} \boldsymbol{J}^{\top}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^{\top}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\Big).
\end{aligned}
$$

By condition (A3), (A4) and (A7), $|\boldsymbol{u}^{\top}[\boldsymbol{D}_{n1}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2})$ on the set $\boldsymbol{N}_n$.

For $i = 2$, we have

$$
\begin{aligned}
&|\boldsymbol{u}^{\top}[\boldsymbol{D}_{n2}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n)]\boldsymbol{u}| \\
\leq & \frac{1}{2}|\boldsymbol{u}^{\top}\sum_{i=1}^{n} \boldsymbol{J}^{\top}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_{i1}(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\top}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
& + \frac{1}{2}|\boldsymbol{u}^{\top}\sum_{i=1}^{n} \boldsymbol{J}^{\top}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_{i2}(\boldsymbol{\beta}_0)\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\top}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
\triangleq & J_{n1} + J_{n2},
\end{aligned}
$$

where we decompose $\mathbf{C}_i(\boldsymbol{\beta}_n)$ as $\mathbf{C}_{i1}(\boldsymbol{\beta}_n) + \mathbf{C}_{i2}(\boldsymbol{\beta}_0)$,

$$
\begin{aligned}
\mathbf{C}_{i1}(\boldsymbol{\beta}_n) &= \mathrm{diag}\Big(\mu_{i1}(\boldsymbol{\beta}_0) - \mu_{i1}(\boldsymbol{\beta}_n), \ldots, \mu_{im}(\boldsymbol{\beta}_0) - \mu_{im}(\boldsymbol{\beta}_n)\Big), \\
\mathbf{C}_{i2}(\boldsymbol{\beta}_0) &= \mathrm{diag}\Big(Y_{i1} - \mu_{i1}(\boldsymbol{\beta}_0), \ldots, Y_{im} - \mu_{im}(\boldsymbol{\beta}_0)\Big).
\end{aligned}
$$

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
2J_{n1} \leq & \sum_{i=1}^{n} ||\boldsymbol{u}^{\intercal}\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1}-\widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{C}_{i1}(\boldsymbol{\beta}_n)\boldsymbol{F}_i(\boldsymbol{\beta}_n)|| \\
& \times ||\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}|| \\
\leq & \sum_{i=1}^{n} ||\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}||^2 \times \lambda_{\max}\Big(\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\Big) \times \lambda_{\max}\Big(\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\Big) \times ||\widehat{\boldsymbol{R}}^{-1}-\tilde{\boldsymbol{R}}^{-1}||_F \\
& \times \max_{i,j}|\mu_{ij}^{(1)}(\widetilde{\boldsymbol{\beta}}_n)| \times \max_{i,j}|\mu_{ij}^{(2)}(\widetilde{\boldsymbol{\beta}}_n)| \times ||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0|| \\
\leq & \sum_{i=1}^{n} ||\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}||^2 \times ||\widehat{\boldsymbol{R}}^{-1}-\tilde{\boldsymbol{R}}^{-1}||_F \times \frac{\max_{i,j}\sigma_{ij}(\boldsymbol{\beta}_n)}{\min_{i,j}\sigma_{ij}^3(\boldsymbol{\beta}_n)} \\
& \times \max_{i,j}|\mu_{ij}^{(1)}(\widetilde{\boldsymbol{\beta}}_n)| \times \max_{i,j}|\mu_{ij}^{(2)}(\widetilde{\boldsymbol{\beta}}_n)| \times ||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0|| \\
\leq & ||\boldsymbol{u}||^2 \times \lambda_{\max}\Big(\sum_{i=1}^{n}\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\Big) \times ||\widehat{\boldsymbol{R}}^{-1}-\tilde{\boldsymbol{R}}^{-1}||_F \times \frac{\max_{i,j}\sigma_{ij}(\boldsymbol{\beta}_n)}{\min_{i,j}\sigma_{ij}^3(\boldsymbol{\beta}_n)} \\
& \times \max_{i,j}|\mu_{ij}^{(1)}(\widetilde{\boldsymbol{\beta}}_n)| \times \max_{i,j}|\mu_{ij}^{(2)}(\widetilde{\boldsymbol{\beta}}_n)| \times ||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||
\end{aligned}
$$

where $\widetilde{\boldsymbol{\beta}}_n$ is between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$. Under conditions (A3), (A4) and (A7), it can be easily seen now $J_{n1} \leq CnO_p(n^{-1/2})O_p(n^{-1/2}) = O_p(1)$.

For J_{n2}, recall that $\epsilon_{ij}(\boldsymbol{\beta}_0) = \sigma_{ij}^{-1}(\boldsymbol{\beta}_0)(Y_{ij}-\mu_{ij}(\boldsymbol{\beta}_0))$. By condition (A5),

$$
\begin{aligned}
& \sup_{\boldsymbol{\beta}_n\in\boldsymbol{N}_n}\mathbb{E}[J_{n2}^2] = \sup_{\boldsymbol{\beta}_n\in\boldsymbol{N}_n}\mathrm{Tr}[\mathbb{E}(J_{n2}^{\intercal}J_{n2})] \\
= & \sup_{\boldsymbol{\beta}_n\in\boldsymbol{N}_n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\mathbb{E}[\epsilon_{ij}\epsilon_{ik}]\mathrm{Tr}\Big[\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1}-\widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{e}_j\mathbf{e}_j^{\intercal}\boldsymbol{F}_i(\boldsymbol{\beta}_n) \\
& \cdot \mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\mathbf{e}_k\mathbf{e}_k^{\intercal}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1}-\widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n) \\
\leq & \sup_{\boldsymbol{\beta}_n\in\boldsymbol{N}_n} C\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}||\mathbf{e}_j^{\intercal}\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)|| \cdot ||\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{F}_i(\boldsymbol{\beta}_n)\mathbf{e}_k|| \\
& \cdot ||\mathbf{e}_k^{\intercal}\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1}-\widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\intercal}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)|| \\
& \cdot ||\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)(\tilde{\boldsymbol{R}}^{-1}-\widehat{\boldsymbol{R}}^{-1})\boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n)\mathbf{e}_j||.
\end{aligned}
$$

By conditions (A4), (A6) and (A7), $\sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} \mathbb{E}[J_{n2}^2] \leq Cn||\tilde{\boldsymbol{R}}^{-1} - \widehat{\boldsymbol{R}}^{-1}||_F^2 = O(1)$. Using similar decompositions, we can verify the results for $\boldsymbol{D}_{n3}$ and $\boldsymbol{D}_{n4}$, which completes the proof. $\square$

Based on Lemma 4.3, we can further approximate $\tilde{\mathbf{D}}_n(\boldsymbol{\beta}_n)$ by $\tilde{\mathbf{D}}_{n1}(\boldsymbol{\beta}_n)$, which are easier to evaluate. Lemma 4.4 provides this approximation.

Lemma 4.4. *Under conditions (A1)-(A8), for some constant $\triangle > 0$ and $\boldsymbol{u} \in \mathbb{R}^{R\sum_{d=1}^{D} p_d}$ such that $||\boldsymbol{u}|| = 1$,*

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| = \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^\top [\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2}), \tag{4.9}$$

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| = \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^\top [\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| = O_p(n^{1/2}). \tag{4.10}$$

Proof of Lemma 4.4. To prove (4.9), it is sufficient to show, for $i = 2, 3, 4$,

$$\sup_{||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| = \triangle n^{-1/2}} \sup_{\boldsymbol{u}} |\boldsymbol{u}^\top \tilde{\boldsymbol{D}}_{ni}(\boldsymbol{\beta}_n)\boldsymbol{u}| = O_p(n^{1/2}).$$

For $\tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n)$, if suffices to show

$$\sup_{\boldsymbol{\beta}_n \in \boldsymbol{N}_n} |\boldsymbol{u}^\top \sum_{i=1}^{n} \boldsymbol{J}^\top(\boldsymbol{\beta}_n) \mathrm{vec} \boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n) \tilde{\boldsymbol{R}}^{-1} \boldsymbol{A}_i^{-3/2}(\boldsymbol{\beta}_n) \mathbf{C}_i(\boldsymbol{\beta}_n) \boldsymbol{F}_i(\boldsymbol{\beta}_n) \mathrm{vec}^\top \boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n) \boldsymbol{u}| = O_p(n^{1/2}).$$

By using the decomposition $\mathbf{C}_i(\boldsymbol{\beta}_n)$ as $\mathbf{C}_{i1}(\boldsymbol{\beta}_n) + \mathbf{C}_{i2}(\boldsymbol{\beta}_0)$, the proof is similar to the proof for $|\boldsymbol{u}^\top [\boldsymbol{D}_{n2}(\boldsymbol{\beta}_n) - \tilde{\boldsymbol{D}}_{n2}(\boldsymbol{\beta}_n)]\boldsymbol{u}|$ in Lemma 4.3. We can prove the results for $\tilde{\boldsymbol{D}}_{n3}(\boldsymbol{\beta}_0)$ and $\tilde{\boldsymbol{D}}_{n4}(\boldsymbol{\beta}_0)$ in the same way, which completes the proof of (4.9).

To prove (4.10), note that

$$
\begin{aligned}
&|\boldsymbol{u}^{\top}[\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)-\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)]\boldsymbol{u}| \\
\leq&|\boldsymbol{u}^{\top}[\boldsymbol{J}^{\top}(\boldsymbol{\beta}_0)-\boldsymbol{J}^{\top}(\boldsymbol{\beta}_n)]\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\top}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
&+|\boldsymbol{u}^{\top}\boldsymbol{J}^{\top}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i[\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)-\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)]\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathrm{vec}^{\top}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
&+|\boldsymbol{u}^{\top}\boldsymbol{J}^{\top}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}[\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)-\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)]\mathrm{vec}^{\top}\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{u}| \\
&+|\boldsymbol{u}^{\top}\boldsymbol{J}^{\top}(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\mathrm{vec}^{\top}\boldsymbol{X}_i[\boldsymbol{J}(\boldsymbol{\beta}_0)-\boldsymbol{J}(\boldsymbol{\beta}_n)]\boldsymbol{u}|.
\end{aligned}
$$

The rest of the proof is similar to the proof of Lemma 4.3 and thus is omitted here. □

Proof of Theorem 4.1. Wang (2011) gave a sufficient condition for the existence and consistency of a sequence of roots $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{s}_n(\boldsymbol{\beta}_n)=0$, namely,

$$
P(\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||=\triangle n^{-1/2}}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\top}\boldsymbol{s}_n(\boldsymbol{\beta}_n)<0)\geq 1-\epsilon \tag{4.11}
$$

with $\forall\epsilon>0$ and a constant $\triangle>0$. To verify (4.11), the main idea is to approximate $\boldsymbol{s}_n(\boldsymbol{\beta}_n)$ by $\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_n)$, whose moments are easier to evaluate.

By direct calculation,

$$
\begin{aligned}
&(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\top}\boldsymbol{s}_n(\boldsymbol{\beta}_n) \\
=&(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\top}\boldsymbol{s}_n(\boldsymbol{\beta}_0)-(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\top}\boldsymbol{D}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0) \\
\triangleq&I_{n1}+I_{n2},
\end{aligned}
$$

where $\boldsymbol{\beta}_n^*$ is between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$. Further decompose I_{n1} into

$$
\begin{aligned}
I_{n1}&=(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\top}\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)+(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0)^{\top}[\boldsymbol{s}_n(\boldsymbol{\beta}_0)-\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)] \\
&\triangleq I_{n11}+I_{n12}.
\end{aligned}
$$

Note that $I_{n11} \leq \triangle n^{-1/2} \cdot ||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)||$. By condition (A6),

$$
\begin{aligned}
&\mathbb{E}[||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)||^2] \\
=&\mathbb{E}\Big\{\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^{\intercal} \tilde{\boldsymbol{R}}^{-1} \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\mathrm{vec}^{\intercal}\boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{\epsilon}_i\Big\} \\
\leq& C \cdot \sum_{i=1}^{n} \mathrm{Tr}\Big(\mathrm{vec}^{\intercal}\boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_i\Big) \\
=& C \sum_{i=1}^{n}\sum_{j=1}^{m} \cdot \mathrm{Tr}\Big(\mathrm{vec}^{\intercal}\boldsymbol{X}_{ij} \boldsymbol{J}(\boldsymbol{\beta}_n)\boldsymbol{J}^{\intercal}(\boldsymbol{\beta}_n)\mathrm{vec}\boldsymbol{X}_{ij}\Big) = O(n)
\end{aligned}
$$

for some constant $C > 0$. This implies that $I_{n11} = \triangle n^{-1/2} O_p(n^{1/2}) = \triangle O_p(1)$. For I_{n12}, by Lemma 4.2,

$$
I_{n12} \leq ||\boldsymbol{\beta}_n - \boldsymbol{\beta}_0|| \cdot ||\boldsymbol{s}_n(\boldsymbol{\beta}_0) - \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)|| = o_p(1).
$$

Therefore, I_{n1} is dominated in probability by I_{n11}.

For I_{n2} ,we decompose it into

$$
\begin{aligned}
I_{n2} =& -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\intercal}\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\
&- (\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\intercal}[\boldsymbol{D}_n(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n^*)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\
\triangleq& I_{n21} + I_{n22}.
\end{aligned}
$$

By Lemma 4.3, it can be easily checked that $I_{n22} = o_p(1)$. Next, for I_{n21},

$$
\begin{aligned}
I_{n21} =& -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\intercal}\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\
&- (\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\intercal}[\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\
&- (\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^{\intercal}[\tilde{\boldsymbol{D}}_{n}(\boldsymbol{\beta}_n^*) - \tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n^*)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\
\triangleq& I_{n21}^1 + I_{n21}^2 + I_{n21}^3.
\end{aligned}
$$

We next show that I_{n21} is dominated in probability by I_{n21}^1. Note that by conditions (A3),

(A4) and (A7),

$$
\begin{aligned}
I_{n21}^1 =& -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\top \Big[\sum_{i=1}^n \boldsymbol{J}^\top(\boldsymbol{\beta}_n) \mathrm{vec} \boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n) \tilde{\boldsymbol{R}}^{-1} \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n) \mathrm{vec}^\top \boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n) \Big] (\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \\
\leq& - n^{-1} \triangle^2 \min_i \lambda_{\min}(\boldsymbol{A}_i(\boldsymbol{\beta}_n)) \lambda_{\min}\Big(\sum_{i=1}^n \boldsymbol{J}^\top(\boldsymbol{\beta}_n) \mathrm{vec} \boldsymbol{X}_i \mathrm{vec}^\top \boldsymbol{X}_i \boldsymbol{J}(\boldsymbol{\beta}_n) \Big) \lambda_{\min}(\tilde{\boldsymbol{R}}^{-1}) \\
\leq& - C \triangle^2,
\end{aligned}
$$

for some constant $C > 0$. By Lemma 4.4, it can be checked directly that both I_{n21}^2 and I_{n21}^3 are $o_p(1)$.

Therefore, with high probability, the sign of $(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\top \boldsymbol{s}_n(\boldsymbol{\beta}_n)$ is determined by $I_{n11} + I_{n21}^1$ and is negative for sufficiently large $\triangle$, which completes the proof. □

Proof of Theorem 4.2. We first show that the normalized $\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)$ has an asymptotic normal distribution. That is, for any $\boldsymbol{b} \in \mathbb{R}^{R \sum_{d=1}^D p_d}$ such that $||\boldsymbol{b}|| = 1$,

$$
\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0) \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) \to N(0, 1), \tag{4.12}
$$

where $\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0) = \mathrm{Var}(\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0))$.

Denote $\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0) \tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0) = \sum_{i=1}^n Z_{ni}$, where

$$
Z_{ni} = \boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0) \boldsymbol{J}^\top(\boldsymbol{\beta}_0) \mathrm{vec} \boldsymbol{X}_i \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0) \tilde{\boldsymbol{R}}^{-1} \boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0),
$$

and $\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0) = \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_0)(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_0))$. Note that $\mathbb{E}(Z_{ni}) = 0$, $\mathrm{Var}(\sum_{i=1}^n Z_{ni}) = 1$. To prove (4.12), it suffices to check the Lyapunov condition. That is, for some $\delta > 0$,

$$
\sum_{i=1}^n \mathbb{E}\Big(|Z_{ni}|^{2+\delta} \Big) \to 0,
$$

as $n \to \infty$. By Cauchy-Schwarz inequality,

$$Z_{ni}^2 \leq \lambda_{\max}(\tilde{\boldsymbol{R}}^{-2})\lambda_{\max}(\boldsymbol{A}_i(\boldsymbol{\beta}_0))||\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0)||^2\gamma_{ni},$$

where $\gamma_{ni} \triangleq \boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{J}^\top(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\top\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{b}$. To evaluate $\max_{1\leq i\leq n}\gamma_{ni}$, we need to evaluate $\lambda_{\min}^{-1}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0))$. Note that

$$\begin{aligned}\boldsymbol{b}^\top\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)\boldsymbol{b} \geq & C\boldsymbol{b}^\top\Big(\sum_{i=1}^n \boldsymbol{J}^\top(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\top\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big)\boldsymbol{b}\\ \geq & C\lambda_{\min}\Big(\sum_{i=1}^n \boldsymbol{J}^\top(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\top\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big),\end{aligned}$$

which implies $\lambda_{\min}\Big(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)\Big) \geq \lambda_{\min}\Big(\sum_{i=1}^n \boldsymbol{J}^\top(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\top\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\Big)$. By condition (A3), $\lambda_{\min}^{-1}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)) = O(n^{-1})$ and hence $\max_{1\leq i\leq n}\gamma_{ni} = o(1)$.

It follows that, for any $\delta > 0$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\Big(|Z_{ni}|^{2+\delta}\Big) \leq & \sum_{i=1}^n \mathbb{E}\Big(C^{1+\delta/2}\gamma_{ni}^{1+\delta/2}||\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0)||^{2+\delta}\Big)\\ \leq & C(\max_{1\leq i\leq n}\gamma_{ni})^{\delta/2}\sum_{i=1}^n \boldsymbol{b}^\top\tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{J}^\top(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\top\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\boldsymbol{b}\\ \leq & C(\max_{1\leq i\leq n}\gamma_{ni})^{\delta/2}\lambda_{\max}\big(\sum_{i=1}^n \boldsymbol{J}^\top(\boldsymbol{\beta}_0)\mathrm{vec}\boldsymbol{X}_i\mathrm{vec}^\top\boldsymbol{X}_i\boldsymbol{J}(\boldsymbol{\beta}_0)\big)\lambda_{\min}^{-1}\big(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0)\big)\\ = & o(1)O(n)O(n^{-1}) = o(1),\end{aligned}$$

which completes the proof of (4.12).

To prove Theorem 4.2, note that because $\boldsymbol{s}_n(\widehat{\boldsymbol{\beta}}_n) = 0$, we have $\boldsymbol{s}_n(\boldsymbol{\beta}_0) = \boldsymbol{D}_n(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$,

for some $\boldsymbol{\beta}_n^*$ between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Hence,

$$
\begin{aligned}
&\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)\\
=&\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\\
&+\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\boldsymbol{D}_n(\boldsymbol{\beta}_n^*)-\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)](\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\\
&+\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)-\boldsymbol{s}_n(\boldsymbol{\beta}_0)]\\
=&J_{n1}+J_{n2}(\boldsymbol{\beta}_n^*)+J_{n3}(\boldsymbol{\beta}_0).
\end{aligned}
$$

By (4.12), it is sufficient to prove that both $\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}}|J_{n2}(\boldsymbol{\beta}_n)|$ and $|J_{n3}(\boldsymbol{\beta}_0)|$ are $o_p(1)$.

For J_{n3}, recall that $||\tilde{\boldsymbol{s}}_n(\boldsymbol{\beta}_0)-\boldsymbol{s}_n(\boldsymbol{\beta}_0)||=O_p(1)$ from Lemma 4.2. Using the previous result that $\lambda_{\min}^{-1}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0))=O(n^{-1})$, it can be easily checked that $J_{n3}^2=o_p(1)$ and hence $|J_{n3}|=o_p(1)$.

For J_{n2}, we have

$$
\begin{aligned}
&\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}}|J_{n2}(\boldsymbol{\beta}_n)|\\
\le&\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}}\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\boldsymbol{D}_n(\boldsymbol{\beta}_n)-\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n)](\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\\
&+\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}}\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n)-\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)](\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\\
&+\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle n^{-1/2}}\boldsymbol{b}^\top \tilde{\boldsymbol{M}}_n^{-1/2}(\boldsymbol{\beta}_0)[\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_n)-\tilde{\boldsymbol{D}}_{n1}(\boldsymbol{\beta}_0)](\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\\
\triangleq&I_{n1}+I_{n2}+I_{n3}.
\end{aligned}
$$

Notice that

$$
I_{n1}\le C\times|\lambda_{\max}(\boldsymbol{D}_n(\boldsymbol{\beta}_n)-\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n))|\times\lambda_{\min}^{-1/2}(\tilde{\boldsymbol{M}}_n(\boldsymbol{\beta}_0))\times||\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0||.
$$

By Lemma 4.3, we have $\sup_{||\boldsymbol{\beta}_n-\boldsymbol{\beta}_0||\le\triangle\sqrt{p/n}}|\lambda_{\max}(\boldsymbol{D}_n(\boldsymbol{\beta}_n)-\tilde{\boldsymbol{D}}_n(\boldsymbol{\beta}_n))|=O_p(\sqrt{np_n})$. Therefore, $I_{n1}=O_p(\sqrt{np_n})O(n^{-1/2})O_p(\sqrt{p/n})=O_p(pn^{-1/2})=o_p(1)$. Similarly, by Lemma 4.4, we have $I_{n2}=o_p(1)$ and $I_{n3}=o_p(1)$. Therefore J_{n1} has the same asymptotic distribution as in (4.12),

which completes the proof. $\square$

Proof of Theorem 4.3. Denote the rank-R tensor GEE estimator by $\widehat{\boldsymbol{B}}_{(R)}$. For Gaussian response, the BIC can be written as

$$BIC(R) = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \langle \boldsymbol{X}_{ij}, \widehat{\boldsymbol{B}}_{(R)}\rangle)^2 + \lambda_n R$$

where $\lambda_n = O(\log n)$.

The proof follows two steps: we need show that BIC neither overestimate nor underestimate the rank. By combining these two results, the consistency of BIC is established.

Step 1: To show BIC does not overestimate the rank, it suffices to show that for any $R > R_0$,

$$\begin{aligned}&\Pr\big(BIC(R) - BIC(R_0) > 0\big)\\ =&\Pr\big(\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) + (R - R_0)\lambda_n > 0\big) \to 1\end{aligned}$$

as $n \to \infty$, where

$$\ell(\widehat{\boldsymbol{B}}_{(R)}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \langle \boldsymbol{X}_{ij}, \widehat{\boldsymbol{B}}_{(R)}\rangle)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \theta_{ij}(\widehat{\boldsymbol{B}}_{(R)}))^2$$

by the identity link function.

Denote $\widehat{\boldsymbol{\beta}}_{(R)} = \mathrm{vec}(\widehat{\boldsymbol{B}}_{(R)1}, \ldots, \widehat{\boldsymbol{B}}_{(R)D})$, where $[\![\widehat{\boldsymbol{B}}_{(R)1}, \ldots, \widehat{\boldsymbol{B}}_{(R)D}]\!]$ is the CP-decomposition of $\widehat{\boldsymbol{B}}_{(R)}$. That is, $\widehat{\boldsymbol{\beta}}_{(R)}$ is the vector of free parameters in $\widehat{\boldsymbol{B}}_{(R)}$. By previous theorems, there exists one tensor GEE estimator $\widehat{\boldsymbol{\beta}}_{(R)}$ that is a root-n consistent estimator for $\boldsymbol{\beta}_{0(R)}$ for $R \geq R_0$, where $\boldsymbol{\beta}_{0(R)}$ is simply $\boldsymbol{\beta}_0$ with additional 0's in ranks $R_0 + 1, \ldots, R$ and $\boldsymbol{\beta}_{0(R_0)} = \boldsymbol{\beta}_0$. If we can show that $\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) = O_p(1)$, the proof is completed by the fact that $R - R_0 > 0$ and λ_n is a diverging sequence. Notice that by subtracting the same term,

$$\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) = \big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big).$$

Denote $L(\boldsymbol{\beta}) = \mathbb{E}[\ell(\boldsymbol{\beta})]$, where the expectation is taken w.r.t. Y_{ij}. We have

$$\begin{aligned}
&\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0) \\
=&\big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) + \big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big).
\end{aligned}$$

Therefore, it suffices to show that

$$\big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) = O_p(1), \tag{4.13}$$

$$\big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) = O_p(1). \tag{4.14}$$

To show (4.13), by the definition of $\boldsymbol{\beta}_0$, we have $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0$. By Taylor expansion at $\boldsymbol{\beta}_0$ and Proposition 2.3 in Zhou et al. (2013),

$$L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0) = Cn||\widehat{\boldsymbol{\beta}}_{(R_0)} - \boldsymbol{\beta}_0||^{\intercal} I(\tilde{\boldsymbol{B}}_0)||\widehat{\boldsymbol{\beta}}_{(R_0)} - \boldsymbol{\beta}_0||$$

where $I(\tilde{\boldsymbol{B}}_0)$ is determined by some $\tilde{\boldsymbol{\beta}}_0 \in \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leq \Delta n^{-1/2}\}$ via CP-decomposition. Under the condition (A3*), this term is $O_p(1)$.

Next we bound the term in (4.14). By direct algebra, it can be shown that

$$\begin{aligned}
&\big(\ell(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R_0)}) - L(\boldsymbol{\beta}_0)\big) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) - \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}[Y_{ij}]\big(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{m} \big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) \\
\leq&\sum_{i=1}^{n}\sum_{j=1}^{m} \big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)Cn^{-1/2}
\end{aligned}$$

by the condition that $\partial\theta_{ij}/\partial\boldsymbol{\beta}$ are uniformly bounded, $|\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R_0)}) - \theta_{ij}(\boldsymbol{\beta}_0)| \leq Cn^{-1/2}$ for some constant C. Denote $g_i(\boldsymbol{u}) = \sum_{j=1}^{m}\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)Cn^{-1/2}$. Notice that $\{g_i(\boldsymbol{u})\}_{i=1}^{n}$ are independent

mean zero random variables. Under the condition that $\mathrm{Var}(\boldsymbol{Y}_i)$ has bounded eigenvalues, it can be easily verified that $\mathrm{Var}(g_i(\boldsymbol{u})) = O(n^{-1})$. Therefore, $\sum_{i=1}^n g_i(\boldsymbol{u}) = O_p(1)$.

Using similar techniques, it can be shown that $\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_{0(R)}) = O_p(1)$ for $R > R_0$ as well. Therefore, for $R > R_0$, the term $BIC(R) - BIC(R_0)$ is asymptotically dominated by $(R - R_0)\log(n)$, which is always positive.

Step 2: To show BIC does not underestimate the rank, it suffices to show that for any $R < R_0$,

$$\begin{aligned}&\Pr\big(BIC(R) - BIC(R_0) > 0\big)\\ =&\Pr\big(\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)}) + (R - R_0)\lambda_n > 0\big) \to 1\end{aligned}$$

as $n \to \infty$. Notice that $n^{-1}(R - R_0)\lambda_n \to 0$ as $n \to \infty$. Therefore, if we can show that $n^{-1}\{\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)})\} \geq c$ for some constant $c > 0$, the proof is completed. Intuitively, we need to show that for any underestimated estimator, the increase of the population loss function to the one with correct rank is bounded away from zero.

Notice that

$$\begin{aligned}&\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\\ =&\big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big) + \big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big).\end{aligned}$$

Denote $\widehat{\boldsymbol{\beta}}_{(R),R_0}$ the augmented vector of $\widehat{\boldsymbol{\beta}}_{(R)}$ with 0's at the those rank $R+1, \ldots, R_0$ so that it has the same length as $\boldsymbol{\beta}_0$. By similar arguments in Step 1, for $R < R_0$

$$L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0) = Cn||\widehat{\boldsymbol{\beta}}_{(R),R_0} - \boldsymbol{\beta}_0||^{\mathsf{T}} I(\tilde{\boldsymbol{B}}_0)||\widehat{\boldsymbol{\beta}}_{(R),R_0} - \boldsymbol{\beta}_0||.$$

Notice that there exists some positive constant c_1 such that $||\widehat{\boldsymbol{\beta}}_{(R),R_0} - \boldsymbol{\beta}_0|| \geq c_1$. This is true because the elements of $\boldsymbol{\beta}_0$ at those locations for rank $R+1, \ldots, R_0$ cannot be all zeros. By the condition (A3*) that the smallest eigenvalue of $I(\boldsymbol{B})$ is bounded away from 0, it can be seen

that $n^{-1}\{L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\} \geq c_2$ for some constant $c_2 > 0$ that does not depend on R.

Similar as in Step 1,

$$\begin{aligned}&\big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big)\\ =&\sum_{i=1}^{n}\sum_{j=1}^{m}\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big).\end{aligned}$$

By the condition that the first derivative of $\theta_{ij}(\boldsymbol{\beta})$ is bounded away from infinity, $\boldsymbol{X}_{ij}$ are uniformly bounded and p is fixed, $\mathrm{Var}[\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big)] = O(1)$. Therefore $\sum_{i=1}^{n}\sum_{j=1}^{m}\big(Y_{ij} - \mathbb{E}[Y_{ij}]\big)\big(\theta_{ij}(\widehat{\boldsymbol{\beta}}_{(R)}) - \theta_{ij}(\boldsymbol{\beta}_0)\big) = O_p(\sqrt{n})$ and $n^{-1}\{\big(\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\big) - \big(L(\widehat{\boldsymbol{\beta}}_{(R)}) - L(\boldsymbol{\beta}_0)\big)\} = o_p(1)$. Combined with previous result, $n^{-1}\{\ell(\widehat{\boldsymbol{B}}_{(R)}) - \ell(\widehat{\boldsymbol{B}}_{(R_0)})\}$ dominates the term in $n^{-1}\{\ell(\widehat{\boldsymbol{\beta}}_{(R)}) - \ell(\boldsymbol{\beta}_0)\}$for sufficiently large n and is bounded away from 0, which completes the proof. □

Proof of Theorem 4.4. Write the SCAD regularized tensor GEE as

$$n^{-1}\boldsymbol{s}_n(\boldsymbol{\beta}_n) - \boldsymbol{q}_{\rho_n}(|\boldsymbol{\beta}_n|) \times \mathrm{sign}(\boldsymbol{\beta}_n),$$

where $\boldsymbol{q}_{\rho_n}(|\boldsymbol{\beta}_n|) = (q_{\rho_n}(|\beta_{n1}|), \ldots, q_{\rho_n}(|\beta_{nR\sum_{d=1}^{D} p_d}|))^{\mathsf{T}}$ is a $R\sum_{d=1}^{D} p_d$-dimensional vector of the subgradients of SCAD penalty, $q_{\rho_n}(\beta) = \rho_n\left\{1_{\{|\beta|\leq\rho_n\}} + (\lambda\rho_n - |\beta|)_+/(\lambda-1)1_{\{|\beta|>\rho_n\}}\right\}$, $\mathrm{sign}(\boldsymbol{\beta}_n) = (\mathrm{sign}(\beta_{n1}), \ldots, \mathrm{sign}(\beta_{nR\sum_{d=1}^{D} p_d}))^{\mathsf{T}}$, the symbol "$\times$" denotes component-wise product, β_{nj} is the jth element of $\boldsymbol{\beta}_n$, $j = 1, \ldots, R\sum_{d=1}^{D} p_d$. Write the support of $\boldsymbol{\beta}_0$ as $\mathcal{J} = \{j : \beta_{0j} > 0\}$.

We prove the theorem by showing the the oracle estimator, $\widehat{\boldsymbol{\beta}}_n^O$, is an approximated solution to the regularized tensor GEE. Denote the jth element of $\widehat{\boldsymbol{\beta}}_n^O$ as $\widehat{\beta}_{nj}^O$. By the definition of the oracle estimator, $\widehat{\beta}_{nj}^O = 0$ for $j \notin \mathcal{J}$. Similar as the definition in Wang et al. (2012), an approximated solution to the regularized tensor GEE, $\widehat{\boldsymbol{\beta}}_n$, is defined to satisfy

$$\Pr\Big(n^{-1}s_{nj}(\boldsymbol{\beta}_n) - q_{\rho_n}(|\beta_{nj}|)\mathrm{sign}(\beta_{nj}) = 0,\ j \in \mathcal{J}\Big) \to 1, \tag{4.15}$$

$$\Pr\Big(|n^{-1}s_{nj}(\boldsymbol{\beta}_n) - q_{\rho_n}(|\beta_{nj}|)\text{sign}(\beta_{nj})| \le \rho_n/\log n,\ j \notin \mathcal{J}\Big) \to 1. \tag{4.16}$$

The reason for this definition of the approximated solution is that the regularized tensor GEE involves non-smooth points, so the exact solution may not exist. It suffices to show that $\widehat{\boldsymbol{\beta}}_n^O$ satisfies both (4.15) and (4.16).

For (4.15), note that by consistency in Theorem 4.1, $||\widehat{\boldsymbol{\beta}}_{n\mathcal{J}}^O - \boldsymbol{\beta}_{0\mathcal{J}}|| = O_p(n^{-1/2})$, where $\widehat{\boldsymbol{\beta}}_{n\mathcal{J}}^O = \{\widehat{\beta}_{nj}^O : j \in \mathcal{J}\}$ and similarly for $\boldsymbol{\beta}_{0\mathcal{J}}$. For fixed p, there exists some constant $C > 0$ that $\min_j \beta_{0j} > C$. Therefore, $\Pr(\min_{j\in\mathcal{J}} \widehat{\beta}_{nj}^O > C) \to 1$ as $n \to \infty$. By the fact $\rho_n = o(1)$, $\Pr(\min_{j\in\mathcal{J}} \widehat{\beta}_{nj}^O > \lambda\rho_n) \to 1$. By the definition of the oracle estimator, $s_{nj}(\widehat{\boldsymbol{\beta}}_n^O) = 0$. Therefore (4.15) holds for the oracle estimator.

For (4.16), by the definition of the oracle estimator, $q_{\rho_n}(|\widehat{\beta}_{nj}^O|)\text{sign}(\widehat{\beta}_{nj}^O) = 0$ for $j \notin \mathcal{J}$. Therefore, it suffices to show $\Pr\Big(|s_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| \le n\rho_n/\log n,\ j \notin \mathcal{J}\Big) \to 1$. Note that $|s_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| \le |s_{nj}(\widehat{\boldsymbol{\beta}}_n^O) - \tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| + |\tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)|$. By Lemma 4.2 and the consistency of the oracle estimator established in Theorem 4.1,

$$\max_{j\notin\mathcal{J}} \Pr(|s_{nj}(\widehat{\boldsymbol{\beta}}_n^O) - \tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| > n\rho_n/\log n) \to 0.$$

Therefore, we only need to verify $\Pr\Big(|\tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O)| > n\rho_n/\log n,\ j \notin \mathcal{J}\Big) \to 0$.

Consider the Taylor expansion

$$\tilde{s}_{nj}(\widehat{\boldsymbol{\beta}}_n^O) = \tilde{s}_{nj}(\boldsymbol{\beta}_0) + \nabla_j(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0) + (\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)^\top \boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0),$$

where $\nabla_j(\boldsymbol{\beta}) = \partial\tilde{s}_{nj}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $\boldsymbol{\psi}_j(\boldsymbol{\beta}) = \partial^2\tilde{s}_{nj}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top$, $\boldsymbol{\beta}_n^*$ is between $\widehat{\boldsymbol{\beta}}_n^O$ and $\boldsymbol{\beta}_0$.

We first show $\Pr\Big(|\tilde{s}_{nj}(\boldsymbol{\beta}_0)| > n\rho_n/\log n,\ j \notin \mathcal{J}\Big) \to 0$. Note that

$$n^{-1}\tilde{s}_{nj}(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \boldsymbol{e}_j^\top \boldsymbol{J}^\top(\boldsymbol{\beta}_0)\text{vec}\boldsymbol{X}_i\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\tilde{\boldsymbol{R}}^{-1}\boldsymbol{\epsilon}_i(\boldsymbol{\beta}_0) \triangleq n^{-1}\sum_{i=1}^n Z_i.$$

Note that Z_i are independent random variables with mean zero. By condition (A4)-(A7), it can

be directly verified that $\mathbb{E}(|Z_i|^l) \leq l! C_1^{l-2} C_2$ for some constants $C_1 > 0$ and $C_2 > 0$. Therefore, $\Pr\left(|n^{-1}\tilde{s}_{nj}(\boldsymbol{\beta}_0)| > \rho_n/\log n\right) \leq \exp[-Cn\rho_n^2/(\log n)^2] \to 0$ is implied by the Bernstein's inequality for any $j \notin \mathcal{J}$. By the condition that $n\rho_n^2/(\log n)^2 \to \infty$, the proof of this step is completed.

We next show $\Pr\left(|\nabla_j(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| > n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 0$. Similar to the decomposition used in the proof of Lemma 4.3, we write $\nabla_j(\boldsymbol{\beta}_0) = \sum_{m=1}^4 \tilde{\boldsymbol{D}}_{njm}(\boldsymbol{\beta}_0)$, where $\tilde{\boldsymbol{D}}_{njm}(\boldsymbol{\beta}_0) = \boldsymbol{e}_j^\intercal \tilde{\boldsymbol{D}}_{nm}(\boldsymbol{\beta}_0)$ for $m = 1, \ldots, 4$. By condition (A4)-(A8), the elements of $n^{-1}\tilde{\boldsymbol{D}}_{njm}(\boldsymbol{\beta}_0)$ are uniformly bounded by a positive constant for $j \notin \mathcal{J}$ and $m = 1, \ldots, 4$. Therefore, $|\nabla_j(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| = O_p(n^{1/2}) = o_p(n\rho_n/\log n)$, which compiles the proof.

Finally, we show $\Pr\left(|(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)^\intercal \boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| > n\rho_n/\log n,\ j \notin \mathcal{J}\right) \to 0$. It can be directly verified that the elements of $n^{-1}\boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)$ are uniformly bounded by a positive constant. Therefore, $|(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)^\intercal \boldsymbol{\psi}_j(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n^O - \boldsymbol{\beta}_0)| = O_p(1)$, which completes the proof. $\square$

Chapter 5

Discussion

5.1 Contributions

This thesis consists of three projects developing some new statistical methods for high dimensional and dependent data. We summarize the contributions of the three projects as follows.

In the first project (Chapter 2), we prove that one local solution to the non-convex penalized SVMs possesses the desirable oracle property, even with a diverging number of variables. The proof technique is different from those in the existing literature, because we use a new sufficient local optimal condition based on subgradients, while existing techniques are not directly applicable to the non-smooth loss function in our setting. We also provide an algorithm with provable global convergence to the oracle estimator. To the best of our knowledge, this is the first result on the convergence of the LLA algorithm in the setup of a non-smooth loss function with a non-convex penalty.

In the second project (Chapter 3), we demonstrate the modeling of reliability under dependence via a copula approach with discrete marginal distributions. The results are established in a general framework and can be extended to multi-state system. We also characterize the influence of dependence on component reliability importance. This is the first result in the literature on the ranking of component reliability importance without the independence assumption. We

reveal the interesting connections to some well-known principles about component importance under independence. We also demonstrate the results for Gaussian copulas, which yield direct and simple interpretations.

In the third project (Chapter 4), we propose a new method for longitudinal imaging analysis. Our proposed method is based on GEE and tensor regression, which can capture the spatial structure of the image covariates and the temporal correlation within each subject simultaneously. Our proposal offers a new way for analyzing such longitudinal imaging data, extending the existing literature about GEEs where the covariate is restricted to vectors and the literature on tensor regression where the observations are assumed to be independent. We are also the first to prove the rank selection consistency under the framework of tensor regression.

5.2 Future Work

We summarize some possible directions to extend the work in this thesis.

For variable selection in support vector machine, we only study the SVMs in non-separable cases in the limit. Although the non-separable cases are important in practical applications, it would be interesting to show similar results for separable cases. Another direction for future study is the availability of an appropriate initial estimator in ultra-high dimensions. One could try to extend the work of Bickel et al. (2009) by assuming similar types of restricted eigenvalue conditions. This extension would require new techniques because both the loss function and the penalty are non-differentiable and the non-smooth locations are different in L_1-penalized SVM. The setup in Bickel et al. (2009) is a smooth loss function with a non-smooth penalty, and the setup in this thesis is a non-smooth loss function with a smooth penalty.

For reliability modeling under dependence using a copula approach, one possible extension of this work is to consider copula functions other than the Gaussian copula. For example, if prior knowledge is available that two components tend to fail together but not function together, an asymmetric copula function, such as the one based on the skewed t-distributions, may be preferred to the symmetric Gaussian copula. If there is some natural ordering of the components,

a copulas such as the vine copula (Kurowicka and Joe, 2011) can take into account the ordering information in modeling the dependence structure.

For longitudinal tensor regression, one possible future development would be a more comprehensive study on rank selection, including the selection consistency for a more general family of models, its convergence rate, and its selection under a diverging dimension. Another is the asymptotic study of our tensor GEE with a diverging dimension, which is also related to the diverging rank selection problem. One may consider more stringent regularity conditions to prove these results.

REFERENCES

An, L. T. H. and P. D. Tao (2005). The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems. *Annals of Operations Research 133*(1-4), 23–46.

Anderson-Cook, C. M. (2008). Evaluating the Series or Parallel Structure Assumption for System Reliability. *Quality Engineering 21*(1), 88–95.

Aven, T. and U. Jensen (1999). *Stochastic Models in Reliability*, Volume 41. Springer.

Balan, R. M. and I. Schiopu-Kratina (2005). Asymptotic Results with Generalized Estimating Equations for Longitudinal Data. *The Annals of Statistics 33*(2), 522–541.

Barlow, R. E. and F. Proschan (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. New York, Holt, Rinehart and Winston.

Barlow, R. E. and A. S. Wu (1978). Coherent Systems with Multi-state Components. *Mathematics of Operations Research 3*(4), 275–281.

Barnard, J., R. McCulloch, and X.-L. Meng (2000). Modeling Covariance Matrices in terms of Standard Deviations and Correlations, with Application to Shrinkage. *Statistica Sinica 10*(4), 1281–1312.

Becker, N., G. Toedt, P. Lichter, and A. Benner (2011). Elastic SCAD as a Novel Penalization Method for SVM Classification Tasks in High-dimensional Data. *BMC Bioinformatics 12*(1), 138.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *The Annals of Statistics 37*(4), 1705–1732.

Birnbaum, Z. W. (1968). On the Importance of Different Components in a Multicomponent System. Technical report, DTIC Document. Available at `http://www.dtic.mil/dtic/tr/fulltext/u2/670563.pdf`.

Bobbio, A., L. Portinale, M. Minichino, and E. Ciancamerla (2001). Improving the Analysis of Dependable Systems by Mapping Fault Trees into Bayesian Networks. *Reliability Engineering & System Safety 71*(3), 249–260.

Braak, H. and E. Braak (1991). Neuropathological Stageing of Alzheimer-related Changes. *Acta Neuropathologica 82*(4), 239–259.

Bradley, P. and O. Mangasarian (1998). Feature Selection via Concave Minimization and Support Vector Machines. In *Machine Learning Proceedings of the Fifteenth International Conference (ICML98)*, pp. 82–90.

Buchman, A. and D. Bennett (2011). Loss of Motor Function in Preclinical Alzheimer's Disease. *Expert Review Neurotherapeutics 11*(5), 665–676.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer.

Caffo, B., C. Crainiceanu, G. Verduzco, S. Joel, M. S.H., S. Bassett, and J. Pekar (2010). Two-stage Decompositions for the Analysis of Functional Connectivity for fMRI with Application to Alzheimer's Disease Risk. *NeuroImage 51*(3), 1140–1149.

Cai, T. and W. Liu (2011). A Direct Estimation Approach to Sparse Linear Discriminant Analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Chen, J. and Z. Chen (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika 95*(3), 759–771.

Claeskens, G., C. Croux, and J. Van Kerckhoven (2008). An Information Criterion for Variable Selection in Support Vector Machines. *The Journal of Machine Learning Research 9*, 541–558.

Davatzikos, C., F. Xu, Y. An, Y. Fan, and S. M. Resnick (2009). Longitudinal Progression

of Alzheimer's-like Patterns of Atrophy in Normal Older Adults: the SPARE-AD Index. *Brain 132*(8), 2026–2035.

den Heijer, T., F. van der Lijn, P. J. Koudstaal, A. Hofman, A. van der Lugt, G. P. Krestin, W. J. Niessen, and M. M. B. Breteler (2010). A 10-year Follow-up of Hippocampal Volume on Magnetic Resonance Imaging in Early Dementia and Cognitive Decline. *Brain 133*(4), 1163–1172.

Desikan, R., H. Cabral, C. Hess, W. Dillon, D. Salat, R. Buckner, B. Fischl, and A. D. N. Initiative (2009). Automated MRI Measures Identify Individuals with Mild Cognitive Impairment and Alzheimer's Disease. *Brain 132*, 2048–2057.

Durham, S. and J. Lynch (2000). A Threshold Representation for the Strength Distribution of a Complex Load Sharing System. *Journal of Statistical Planning and Inference 83*(1), 25–46.

Ebrahimi, N., N. Y. Jalali, E. S. Soofi, and R. Soyer (2014). Importance of Components for a System. *Econometric Reviews 33*(1-4), 395–420.

El-Neweihi, E., F. Proschan, and J. Sethuraman (1978). Multistate Coherent Systems. *Journal of Applied Probability 15*(4), 675–688.

Esary, J. and F. Proschan (1970). A Reliability Bound for Systems of Maintained, Interdependent Components. *Journal of the American Statistical Association 65*(329), 329–338.

Fan, J. and Y. Fan (2008). High Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics 36*(6), 2605–2637.

Fan, J. and R. Li (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society. Series B (Methodological) 70*, 849–911.

Fan, J., L. Xue, and H. Zou (2014). Strong Oracle Optimality of Folded Concave Penalized Estimation. *The Annals of Statistics*, to appear.

Fang, H.-B., K.-T. Fang, and S. Kotz (2002). The Meta-elliptical Distributions with Given Marginals. *Journal of Multivariate Analysis 82*(1), 1–16.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*, Volume 1. Springer Series in Statistics.

Genest, C. and J. Nešlehová (2007). A Primer on Copulas For Count Data. *Astin Bulletin 37*(02), 475–515.

Goldsmith, J., L. Huang, and C. Crainiceanu (2014). Smooth Scalar-on-Image Regression via Spatial bayesian variable selection. *Journal of Computational and Graphical Statistics 23*, 46–64.

Graves, T. L., C. M. Anderson-Cook, and M. S. Hamada (2010). Reliability Models for Almost-series and Almost-parallel Systems. *Technometrics 52*(2), 160–171.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine learning 46*(1), 389–422.

Harasty, J. A., G. M. Halliday, J. J. Kril, and C. Code (1999). Specific Temporoparietal Gyral Atrophy Reflects the Pattern of Language Dissolution in Alzheimer's Disease. *Brain 122*(4), 675–686.

Hinrichs, C., V. Singh, G. Xu, and S. C. Johnson (2011). Predictive Markers for AD in a Multi-modality Framework: An Analysis of MCI Progression in the ADNI Population. *NeuroImage 55*(2), 574 – 589.

Hokstad, P. (1988). A Shock Model for Common-cause Failures. *Reliability Engineering & System Safety 23*(2), 127–145.

Joag-Dev, K., M. D. Perlman, and L. D. Pitt (1983). Association of Normal Random Variables and Slepian's Inequality. *The Annals of Probability 11*(2), 451–455.

Joe, H. (1990). Multivariate Concordance. *Journal of Multivariate Analysis 35*(1), 12–30.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*, Volume 73. CRC Press.

Kim, Y., H. Choi, and H. Oh (2008). Smoothly Clipped Absolute Deviation on High Dimensions. *Journal of the American Statistical Association 103*(484), 1665–1673.

Koenker, R. (2005). *Quantile Regression*, Volume 38. Cambridge University Press.

Kolda, T. G. and B. W. Bader (2009). Tensor Decompositions and Applications. *SIAM Review 51*(3), 455–500.

Koo, J., Y. Lee, Y. Kim, and C. Park (2008). A Bahadur Representation of the Linear Support Vector Machine. *The Journal of Machine Learning Research 9*, 1343–1368.

Kuo, W. and X. Zhu (2012). *Importance Measures in Reliability, Risk, and Optimization: Principles and Applications*. John Wiley & Sons.

Kurowicka, D. and H. Joe (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific.

Kvam, P. H. and E. A. Pena (2005). Estimating Load-sharing Properties in a Dynamic Reliability System. *Journal of the American Statistical Association 100*(469), 262–272.

Laird, N. M. and T. A. Louis (1989). Empirical Bayes Ranking Methods. *Journal of Educational and Behavioral Statistics 14*(1), 29–46.

Langseth, H. and L. Portinale (2007). Bayesian Networks in Reliability. *Reliability Engineering & System Safety 92*(1), 92–108.

Lawless, J. (1983). Statistical Methods in Reliability. *Technometrics 25*(4), 305–316.

Lee, T. M., D. Sun, M.-K. Leung, L.-W. Chu, and C. Keysers (2013). Neural Activities During Affective Processing in People with Alzheimer's Disease. *Neurobiology of Aging 34*(3), 706 – 715.

Li, B. (1997). On the Consistency of Generalized Estimating Equations. In *Selected Proceedings of the Symposium on Estimating Functions (Athens, GA, 1996)*, Volume 32 of *IMS Lecture Notes Monograph Series*, pp. 115–136. Hayward, CA: Institute of Mathematical Statistics.

Li, Q., J. B. Brown, H. Huang, and P. J. Bickel (2011). Measuring Reproducibility of High-throughput Experiments. *The Annals of Applied Statistics 5*(3), 1752–1779.

Li, Y., J. H. Gilmore, D. Shen, M. Styner, W. Lin, and H. Zhu (2013). Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. *NeuroImage 72*(0), 91 – 105.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika 73*(1), 13–22.

Lin, Y., Y. Lee, and G. Wahba (2002). Support Vector Machines for Classification in Nonstandard Situations. *Machine Learning 46*(1-3), 191–202.

Mapstone, M., T. Steffenella, and C. Duffy (2003). A Visuospatial Variant of Mild Cognitive Impairment: Getting Lost Between Aging and AD. *Neurology 60*, 802–808.

Marshall, A. W. and I. Olkin (1967). A Multivariate Exponential Distribution. *Journal of the American Statistical Association 62*(317), 30–44.

Mazumder, R., J. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with non-convex penalties. *Journal of the American Statistical Association 106*(495), 1125–1138.

McEvoy, L. K., D. Holland, D. J. Hagler, C. Fennema-Notestine, J. B. Brewer, and A. M. Dale (2011). Mild Cognitive Impairment: Baseline and Longitudinal Structural MR Imaging Measures Improve Predictive Prognosis. *Radiology 259*(3), 834–843. PMID: 21471273.

Misra, C., Y. Fan, and C. Davatzikos (2009). Baseline and Longitudinal Patterns of Brain Atrophy in MCI Patients, and Their Use in Prediction of Short-term Conversion to AD: Results from ADNI. *NeuroImage 44*(4), 1415 – 1422.

Navarro, J., Y. Águila, M. A. Sordo, and A. Suárez-Llorens (2014). Preservation of Reliability Classes under the Formation of Coherent Systems. *Applied Stochastic Models in Business and Industry 30*(4), 444–454.

Navarro, J., Y. del Águila, M. A. Sordo, and A. Suárez-Llorens (2015). Preservation of Stochastic Orders under the Formation of Generalized Distorted Distributions. Applications to Coherent Systems. To appear in *Methodology and Computing in Applied Probability*. Published online first Feb. 28, 2015. DOI: 10.1007/s11009-015-9441-z.

Navarro, J., J. M. Ruiz, and C. J. Sandoval (2007). Properties of Coherent Systems with Dependent Components. *Communications in Statistics - Theory and Methods 36*(1), 175–191.

Navarro, J. and F. Spizzichino (2010). Comparisons of Series and Parallel Systems with Components Sharing the Same Copula. *Applied Stochastic Models in Business and Industry 26*(6), 775–791.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer.

Ortega, J. M. and W. C. Rheinboldt (2000). *Iterative Solution of Nonlinear Equations in Several Variables*, Volume 30. SIAM.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics 57*(1), 120–125.

Park, C., K.-R. Kim, R. Myung, and J.-Y. Koo (2012). Oracle Properties of SCAD-penalized Support Vector Machine. *Journal of Statistical Planning and Inference 142*(8), 2257–2270.

Pitt, M., D. Chan, and R. Kohn (2006). Efficient Bayesian Inference for Gaussian Copula Regression Models. *Biometrika 93*(3), 537–554.

Portnoy, S. (1984). Asymptotic Behavior of M-estimators of p Regression Parameters When p^2/n is Large. I. Consistency. *The Annals of Statistics*, 1298–1309.

Prentice, R. L. and L. P. Zhao (1991). Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. *Biometrics 47*(3), 825–839.

Qu, A., B. G. Lindsay, and B. Li (2000). Improving Generalised Estimating Equations Using Quadratic Inference Functions. *Biometrika 87*(4), 823–836.

Rao, C. R. and S. K. Mitra (1971). *Generalized Inverse of Matrices and its Applications.* John Wiley & Sons, Inc., New York-London-Sydney.

Reiss, P. and R. Ogden (2010). Functional Generalized Linear Models with Images as Predictors. *Biometrics 66*, 61–69.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Shinohara, R., C. Crainiceanu, B. Caffo, and D. Reich (2011). Longitudinal Analysis of Spatiotemporal Processes: A Case Study of Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Multiple Sclerosis. *Johns Hopkins University, Department of Biostatistics Working Papers*.

Sklar, M. (1959). *Fonctions de Répartition à n Dimensions et Leurs Marges.* Université Paris 8.

Skup, M., H. Zhu, and H. Zhang (2012). Multiscale Adaptive Marginal Analysis of Longitudinal Neuroimaging Data with Time-varying Covariates. *Biometrics 68*(4), 1083–1092.

Smith, M. S. and M. A. Khaled (2012). Estimation of Copula Models with Discrete Margins via Bayesian Data Augmentation. *Journal of the American Statistical Association 107*(497), 290–303.

Song, P. X.-K., Z. Jiang, E. Park, and A. Qu (2009). Quadratic Inference Functions in Marginal Models for Longitudinal Data. *Statistics in Medicine 28*(29), 3683–3696.

Tales, A., J. Haworth, S. Nelson, R. J. Snowden, and G. Wilcock (2005). Abnormal Visual Search in Mild Cognitive Impairment and Alzheimer's Disease. *Neurocase 11*(1), 80–84.

Tao, P. and L. An (1997). Convex Analysis Approach to DC Programming: Theory, Algorithms and Applications. *Acta Mathematica Vietnamica 22*(1), 289–355.

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

Vapnik, V. (1996). *The Nature of Statistical Learning Theory.* Springer, New York.

Vaurio, J. K. (2005). Uncertainties and Quantification of Common Cause Failure Rates and Probabilities for System Analyses. *Reliability Engineering & System Safety 90*(2), 186–195.

Wang, L. (2011). GEE Analysis of Clustered Binary Data with Diverging Number of Covariates. *The Annals of Statistics 39*(1), 389–417.

Wang, L., Y. Kim, and R. Li (2013). Calibrating Non-convex Penalized Regression in Ultra-high Dimension. *The Annals of Statistics 41*, 2505–2536.

Wang, L., Y. Wu, and R. Li (2012). Quantile Regression for Analyzing Heterogeneity in Ultra-high Dimension. *Journal of the American Statistical Association 107*(497), 214–222.

Wang, L., J. Zhou, and A. Qu (2012). Penalized Generalized Estimating Equations for High-dimensional Longitudinal Data Analysis. *Biometrics 68*(2), 353–360.

Wang, L., J. Zhu, and H. Zou (2006). The Doubly Regularized Support Vector Machine. *Statistica Sinica 16*(2), 589–615.

Wang, L., J. Zhu, and H. Zou (2007). Hybrid Huberized Support Vector Machines for Microarray Classification. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 983–990.

Wang, X., B. Nan, J. Zhu, and R. Koeppe (2014). Regularized 3D Functional Regression for Brain Image Data via Haar Wavelets. *The Annals of Applied Statistics*, in press.

Wegkamp, M. and M. Yuan (2011). Support Vector Machines with a Reject Option. *Bernoulli 17*, 1368–1385.

Welsh, A. (1989). On M-processes and M-estimation. *The Annals of Statistics 17*(1), 337–361.

Xie, M. and Y. Yang (2003). Asymptotics for Generalized Estimating Equations with Large Cluster Sizes. *The Annals of Statistics 31*(1), 310–347.

Xue-Kun Song, P. (2000). Multivariate Dispersion Models Generated from Gaussian Copula. *Scandinavian Journal of Statistics 27*(2), 305–320.

Yao, Z., B. Hu, C. Liang, L. Zhao, M. Jackson, and the Alzheimer's Disease Neuroimaging Initiative (2012). A Longitudinal Study of Atrophy in Amnestic Mild Cognitive Impairment and Normal Aging Revealed by Cortical Thickness. *PLoS One 7*(11), e48973.

Yuan, M. (2010). High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *The Journal of Machine Learning Research 99*, 2261–2286.

Zhang, C. (2010). Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics 38*(2), 894–942.

Zhang, C.-H. and J. Huang (2008). The Sparsity and Bias of the Lasso Selection in High-dimensional Linear Regression. *The Annals of Statistics 36*(4), 1567–1594.

Zhang, D., D. Shen, and Alzheimer's Disease Neuroimaging Initiative (2012). Predicting Future Clinical Changes of MCI Patients using Longitudinal and Multimodal Biomarkers. *PLoS One 7*(3), e33182.

Zhang, H., J. Ahn, X. Lin, and C. Park (2006). Gene Selection using Support Vector Machines with Non-convex Penalty. *Bioinformatics 22*(1), 88–95.

Zhou, H., L. Li, and H. Zhu (2013). Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association 108*(502), 540–552.

Zhu, J., S. Rosset, T. Hastie, and R. Tibshirani (2004). 1-norm Support Vector Machines. *Advances in Neural Information Processing Systems 16*(1), 49–56.

Zou, H. (2007). An Improved 1-norm SVM for Simultaneous Classification and Variable Selection. In *Eleventh International Conference on Artificial Intelligence and Statistics*.

Zou, H. and R. Li (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Annals of Statistics 36*(4), 1509–1533.

Zou, H. and M. Yuan (2008). The F-infinity Norm Support Vector Machine. *Statistica Sinica 18*, 379–398.