STATISTICAL LEARNING METHODS FOR HIGH-DIMENSIONAL CLASSIFICATION AND REGRESSION

Hannan Yang

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Quefeng Li

Danyu Lin

Yufeng Liu

Guorong Wu

Donglin Zeng

ABSTRACT

Hannan Yang: Statistical Learning Methods for High-dimensional
Classification and Regression
(Under the direction of Quefeng Li and Danyu Lin)

With the recent advancement of technology, large and heterogeneous data containing enormous variables of mixed types have become increasingly popular, great challenges in computation and theory have arisen for classical methods in classification and regression. It is of great interest to develop new statistical methods that are computationally efficient and theoretically sound for classification and regression using high-dimensinoal and heterogeneous data. In this dissertation, we specifically address the problems in the computation of high-dimensional linear discriminant analysis, and in high-dimensional linear regression and ordinal classification with mixed covariates.

First, we propose an efficient greedy search algorithm that depends solely on closed-form formulae to learn a high-dimensional linear discriminant analysis (LDA) rule. We establish theoretical guarantee of its statistical properties in terms of variable selection and error rate consistency; in addition, we provide an explicit interpretation of the extra information brought by an additional feature in a LDA problem under some mild distributional assumptions. We demonstrate that this new algorithm drastically improves computational speed compared with other high-dimensional LDA methods, while maintaining comparable or even better classification performance through extensive simulation studies and real data analysis.

Second, we propose a semiparametric Latent Mixed Gaussian Copula Regression (LMGCR) model to perform linear regression for high-dimensional mixed data. The model assumes that the observed mixed covariates are generated from latent variables that follow the Gaussian copula. We develop an estimator of the regression coefficients in LMGCR and prove its estimation and variable selection consistency. In addition, we devise a prediction rule given by LMGCR and quantify its prediction error under mild conditions. We demonstrate that the proposed model has superior performance in both coefficient estimation and prediction through extensive simulation studies and real data analysis.

Finally, we propose a semiparametric Latent Mixed Gaussian Copula Classification (LMGCC) rule to perform classification of ordinal response using unnormalized high-dimensional data. Our classification rule learns the Bayes rule derived from joint modeling of ordinal response and continuous features through a latent Gaussian copula model. We develop an estimator of the regression coefficients in predicting the latent response and prove its estimation and variable selection consistency. In addition, we establish that our devised LMGCC has error rate consistency. We demonstrate that the proposed method has superior performance in ordinal classification through extensive simulation studies and real data analysis.

To my parents and Youjia.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude and appreciation to my advisors, Drs. Quefeng Li and Danyu Lin. Dr. Danyu Lin recruited me into the program and set a high standard in research for me at the early stage. He served as a counselor for both my graduate studies and dissertation research, showed me the depth and the breadth of biostatistics research by hosting a series of group seminars, and encouraged me to think critically about the relevant statistical methodologies. I am grateful that Dr. Danyu Lin's experience navigated me through the exploration of thesis topics. Dr. Quefeng Li served as my primary advisor on dissertation research and collaborative consultation with the Division of Cardiology at UNC Health Care. He whole-heartedly directed me to complete all three papers for dissertation research by patiently discussing most of the technical details with me, supervising the progress of software development and manuscript writing, and constantly being responsive to my questions on both academics and career planning. I am grateful to have Dr. Quefeng Li as my advisor, his guidance contributed tremendously to the completion of my dissertation.

Next, I would like to thank many other mentors. I thank the rest of my committee members, Drs. Yufeng Liu, Guorong Wu, and Donglin Zeng. They shared their unique perspectives on my dissertation research and generously provided constructive feedbacks for me to further polish the papers. I thank my collaborators Drs. Joseph Rossi and Anil Gehi for supporting my graduate research asistantship and introducing me to various interesting projects in cardiology research, this collaborative experience contributed significantly to my job hunting in the pharmaceutical industry.

Lastly, I would like to dedicate this work to my beloved parents, Mr. Shaoqun Yang and Mrs. Xin Tang, and my significant other, Ms. Youjia "Jacqueline" Yang. My parents supported me unconditionally to chase my dream throughout this journey and encouraged me to persevere during tough times. Ms. Youjia Yang light up my daily life by her love and companionship, she firmly stood by my side while I experienced all the ups and downs. Dad, Mom, and Youjia, I am fortuante to have you as my family, and I am thrilled to start a new chapter of life with you.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE REVIEW 4

2.1 High-dimensional Linear Discriminant Analysis 4

2.2 Copula-based Statistical Learning for Regression 7

2.2.1 Unsupervised Learning 7

2.2.2 Supervised Learning for Regression 11

2.3 Copula-based Statistical Learning for Classification 13

CHAPTER 3: AN EFFICIENT GREEDY SEARCH ALGORITHM FOR HIGH-DIMENSIONAL LINEAR DISCRIMINANT ANALYSIS 15

3.1 Introduction 15

3.2 An Efficient Greedy Search Algorithm 17

3.3 Theoretical Properties 22

3.4 Relaxation of the Normality Assumption 27

3.5 Simulation Studies 28

3.6 An Application to Cancer Subtype Classification 33

3.7 Discussion 34

3.8 Technical Details 35

3.8.1 Proofs 35

3.8.2 Supporting Lemmas and their Proofs 42

3.8.3 Additional Results in Cancer Subtype Analysis 48

CHAPTER 4: HIGH-DIMENSIONAL SEMIPARAMETRIC LATENT GAUSSIAN COPULA REGRESSION FOR MIXED DATA 50

4.1 Introduction 50

4.2 Latent Gaussian Copula Regression Model . . . 52

4.3 Estimation . . . 54

4.4 Statistical Properties . . . 57

4.5 Prediction . . . 58

4.5.1 Imputation of Latent Variables . . . 58

4.5.2 Prediction Error . . . 60

4.5.3 Comparison with a Naive Method . . . 62

4.6 Simulations . . . 63

4.7 Real Data Analysis . . . 68

4.8 Technical Details . . . 69

4.8.1 Bridge Functions and Estimation of $\mathbf{\Sigma}$ . . . 69

4.8.2 Mean Squared Imputation Error of $\widetilde{\mathbf{u}}$ . . . 70

4.8.3 Variable Selection Results for Real Data Analysis . . . 72

4.8.4 Proofs of Main Theorems . . . 73

4.8.5 Supporting Lemmas and Their Proofs . . . 86

CHAPTER 5: MULTI-CLASS CLASSIFICATION VIA LATENT MIXED GAUSSIAN COPULA MODEL . . . 102

5.1 Introduction . . . 102

5.2 Methodology . . . 104

5.2.1 Latent Mixed Gaussian Copula Classification . . . 104

5.2.2 Estimation . . . 106

5.3 Theoretical Properties . . . 109

5.4 Simulations . . . 112

5.5 Real Data Analysis . . . 117

5.6 Technical Details . . . 120

5.6.1 Proofs of Main Theorems . . . 120

5.6.2 Supporting Lemmas and their Proofs . . . 128

BIBLIOGRAPHY . . . 132

LIST OF TABLES

3.1 Numerical performance of the five classifiers in classifying cancer subtypes. 34

LIST OF FIGURES

3.1 Numerical performance of the seven classifiers in Scenario 1. Panel (c) is a zoomed plot of panel (b) for the GS-LDA , SLDA, SVM, and Logistic-L1. 31

3.2 Numerical performance of the seven classifiers in Scenario 2. Panel (d) is a zoomed plot of panel (c) for the GS-LDA , SLDA, SVM, and Logistic-L1. 31

3.3 Numerical performance of the GS-LDA in Scenario 3. 32

3.4 Numerical performance of the GS-LDA in Scenario 4. 32

3.5 Variable selection performance of the three classifiers in classifying cancer subtypes: panel (a) for the GS-LDA; panel (b) for the ROAD; and panel (c) for the Logistic-L1. 49

4.1 Estimation error, variable selection performance, and MSPEs for the five competitors in Scenario 1. 65

4.2 Estimation error, variable selection performance, and MSPEs for the five competitors in Scenario 2. 65

4.3 MSPEs for the five competitors in Scenario 3. 66

4.4 MSPEs for the four competitors in Scenario 4. 67

4.5 Prediction of Theft-Related Crime Responses by our and Naive-LASSO methods. 68

4.6 Most Frequently Selected Variables by Our and Naive Methods for Predicting Theft-Related Crimes. 72

5.1 Comparison of overall error rate for the six competitors in Scenario 1. Red lines indicate the Bayes errors. 114

5.2 Comparison of overall error rate for the six competitors in Scenario 2. Red lines indicate the Bayes errors. 114

5.3 Comparison of overall and average within-group error rate for the six competitors in Scenario 3. Red lines indicate overall and average within-group error rates for the Bayes rule. 115

5.4 Comparison of overall and average within-group error rate for the six competitors in Scenario 4. Red lines indicate the overall and average within-group error rates for the Bayes rule. 116

5.5 Performance comparison of six competitors on classifying breast cancer progression status. 118

CHAPTER 1
INTRODUCTION

Classification and regression are central statistical problems: one aims at assigning a subject to one of several classes based on certain features, while the other attempts to find the relationship between a response and the covariates. There are numerous real-life applications that fall in the category of these two tasks. In genomics studies, the microarray technology has generated massive gene expression measurements for classifying subtypes of cancers and many other diseases (Golub et al. 1999; Gordon et al. 2002); in clinical practices of cardiology, assessments on electronic health records may help determine the predictors for pacemaker implantation (Mazzella et al. 2021) or understand whether political stress has any impact on arrhythmias (Rosman et al. 2020); in social science, community level demographics help unravel the important factors for predicting crime patterns (Buczak and Gifford 2010), general social survey (GSS) could shed light on the association between religious practice and health (Idler et al. 2003); in the study of Alzheimer's disease, baseline multi-modal neuroimage provides early prediction on the disease progression (Doyle et al. 2014).

Classical methods in classification, such as linear discriminant analysis, and in regression, such as linear or generalized linear regression model, have become widely used before the big data era. With the emergence of large and heterogeneous data containing enormous variables of mixed types, great challenges in computation and theory have arisen for classical methods in classification and regression. For example, The Cancer Genome Atlas (TCGA) database integrates clinical information with gene expressions, methylations and copy number variations; the UK Biobank database integrates clinical information with genotyes from whole genome sequencing and even imaging data; the Alzheimer's Disease Neuroimaging Initiative (ADNI) integrates patient demographics with neuroimaging measurements from multiple modalities. These databases collect immeasurable and detailed information, which could significantly improve the prediction of certain clinical phenotypes and help understand the mechanism behind the chronic diseases. However, such large databases are practically imperfect for statistical tasks, as they are generally in the form with high-dimensionality, missingness, non-normality, and heterogeneity. High-dimensionality illustrates the case when the number of variables exceeds the sample size, which has led to tremndous ill-posed problems for statistical research. Modeling the missing mechanism in the data has been one central topic in statistical research for decades, maximum likelihood approach, multiple imputation, and Bayesian approach are well developed but generally computationally intensive. The presence of mixed types (including continuous, binary, ordinal, and truncated) of variables is common when integrating multiple modalities of data during classification or regression. Handling mixed data types generally requires applying appropriate transformations before classification and regression (Carroll and Ruppert 1988), but such transformations are usually subjective and it remains unclear whether choosing certain transformations can correctly specify the functional form of the variables. Currently, with the advancement of computation power, scholars have proposed viable methods in different applications of high-dimensional data analysis. However, these methods are mostly relying on assumptions that oversimplify the complexity of real datasets, e.g. the violation of normality assumption, mixed types of variables, and are practically inefficient in terms of computation. Hence, there is an urgent need to tackle more complex high-dimensional data for the study of statistical learning methods.

It is of great interest to develop new statistical methods that are computationally efficient and theoretically sound for classification and regression using large and heterogeneous data. In this dissertation, we specifically address the three problems: how to efficiently solve the high-dimensional linear discriminant analysis problem, how to unify mixed and non-normal types of covariates in high-dimensional regression, how to unify non-normal features for high-dimensional ordinal classification.

First, we propose an efficient greedy search algorithm that depends solely on closed-form formulae to learn a high-dimensional linear discriminant analysis (LDA) rule. We establish theoretical guarantee of its statistical properties in terms of variable selection and error rate consistency; in addition, we provide an explicit interpretation of the extra information brought by an additional feature in a LDA problem under some mild distributional assumptions. We demonstrate that this new algorithm drastically improves computational speed compared with other high-dimensional LDA methods, while maintaining comparable or even better classification performance through simulation studies and a real data application to cancer genomics.

Second, we propose a semiparametric Latent Mixed Gaussian Copula Regression (LMGCR) model to perform linear regression for high-dimensional mixed data. The model assumes that the observed mixed covariates are generated from latent variables that follow the Gaussian copula, and the observed response is generated by a linear model of the latent covariates. We develop an estimator of the regression coefficients in LMGCR and prove its estimation and variable selection consistency. We devise an imputation procedure with closed-form formulae to recover the latent covariates. In addition, we devise a prediction rule given by LMGCR and quantify its prediction error under mild conditions. We demonstrate that the proposed model has superior performance in both coefficient estimation and prediction through extensive simulation studies. We also apply the proposed method to analyze community crimes using a crime data from the UCI Machine Learning Repository.

Finally, we propose a semiparametric Latent Mixed Gaussian Copula Classification (LMGCC) rule to perform multi-class classification of ordinal response using high-dimensional non-normal data. With the latent mixed Gaussian copula model, we can jointly model the ordinal response and the continuous features by assuming there exists some latent continuous variable generating the response so that this latent variable and the features jointly follow a Gaussian copula. We devise the LMGCC rule that learns the Bayes rule under our model. We prove that the regression coefficients for predicting the latent variable has estimation and variable selection consistency, and establish the misclassification error rate consistency of LMGCC. We demonstrate that LMGCC has superior and robust classification performance than other multi-class classifiers through extensive simulation studies. We apply LMGCC to classify the progression of breast cancer using a baseline FNA image data from the UCI Machine Learning Repository.

CHAPTER 2
LITERATURE REVIEW

In this chapter, we review some existing literature that motivated the subsequent development of our dissertation.

2.1 High-dimensional Linear Discriminant Analysis

We consider a binary classification problem where we intend to assign a class label $Y \in \{0, 1\}$ to a subject given its features $\boldsymbol{x}$. The class label has a prior distribution of $P(Y = k) = \pi_k$, for $k = 0, 1$. Suppose $\boldsymbol{x}_k \in \mathbb{R}^p$ denotes a p-dimensional vector of features from the kth class that follows a normal distribution of $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are assumed to be independent. The Bayes rule of this classification problem under zero-one loss is given by $D_{Bayes}(\boldsymbol{x}) = I(\boldsymbol{\delta}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \log(\pi_1/\pi_0))$, where $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, $\boldsymbol{\mu} = (1/2)(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$ and $\boldsymbol{x}$ is a new observation. The corresponding Bayes error is given by $R_{Bayes} = \Phi(-\sqrt{\Delta_p}/2)$, where $\Delta_p = \boldsymbol{\delta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$ is the Mahalanobis distance between the centroids of the two classes and Φ is the cumulative distribution function of the standard normal distribution. In practice, the Bayes rule is unknown. A classification rule is learned based on the training data $\boldsymbol{X} = \{\boldsymbol{x}_{ki}; k = 0, 1; i = 1, \ldots, n_k\}$, where $\boldsymbol{x}_{ki}$'s are independent and identically distributed (i.i.d) samples from the kth class and n_k is the sample size of the kth class with $n = n_0 + n_1$. Then the rule is applied to classify a new observation $\boldsymbol{x}$, which is assumed to be independent of the training data.

The linear discriminant analysis (LDA), was widely used before the big data era (Anderson 1958). It directly learns the Bayes rule by estimating the unknown parameters involved. For the classical LDA method, the unknown parameters in the Bayes rule are replaced with their maximum likelihood estimators; this LDA rule has the form of

$$D_{LDA}(\boldsymbol{x}) = I(\widehat{\boldsymbol{\delta}}^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) \leq \log(\widehat{\pi}_1/\widehat{\pi}_0))$$

where

$$\widehat{\pi}_k = n_k/n,\ \widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_{ki},\ \widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1, \tag{2.1.1}$$

$$\widehat{\boldsymbol{\mu}} = (1/2)(\widehat{\boldsymbol{\mu}}_0 + \widehat{\boldsymbol{\mu}}_1),\ \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=0}^{1}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ki} - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{ki} - \widehat{\boldsymbol{\mu}}_k)^T. \tag{2.1.2}$$

But in the high-dimensional setting where $p > n$, the classical LDA method is no longer feasible, as $\widehat{\boldsymbol{\Sigma}}$ is not invertible. Even if we replace $\widehat{\boldsymbol{\Sigma}}^{-1}$ with a generalized matrix inverse, Bickel and Levina (2004) showed that the resulting rule has an asymptotic misclassification error of 1/2, which is as bad as random guessing. This is essentially due to the error accumulation in estimating those high-dimensional parameters in the classifier. To avoid this issue in the high-dimensional setting, many regularized methods have been proposed (Clemmensen et al. 2011; Witten and Tibshirani 2011; Shao et al. 2011; Cai and Liu 2011; Fan et al. 2012; Mai et al. 2012; Han et al. 2013). In particular, Shao et al. (2011) proposed a sparse linear discriminant analysis (SLDA) rule

$$D_{SLDA}(\boldsymbol{x}) = I(\widetilde{\boldsymbol{\delta}}^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) \le \log(\widehat{\pi}_1/\widehat{\pi}_0))$$

where $\widetilde{\boldsymbol{\delta}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are the thresholding estimators that $\widetilde{\boldsymbol{\delta}} = (\widetilde{\delta}_j)$ with $\widetilde{\delta}_j = \widehat{\delta}_j I(|\widehat{\delta}_j| > t_\delta)$ and $\widehat{\delta}_j$ is the jth element of $\widehat{\boldsymbol{\delta}}$, and $\widetilde{\boldsymbol{\Sigma}} = (\widetilde{\sigma}_{ij})$ with $\widetilde{\sigma}_{ii} = \widehat{\sigma}_{ii}$, $\widetilde{\sigma}_{ij} = \widehat{\sigma}_{ij} I(|\widehat{\sigma}_{ij}| > t_\sigma)$ for $i \neq j$ and $\widehat{\sigma}_{ij}$ is the (i, j)th element of $\widehat{\boldsymbol{\Sigma}}$. They showed that the SLDA's misclassification error still converges to the Bayes error given that the thresholds t_δ and t_σ are chosen properly and given some sparsity conditions on both $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$. Instead of separately estimating $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, as the SLDA does, two other methods directly estimate the slope of the Bayes rule $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ by solving convex optimization problems. For example, the linear programming discriminant (LPD) method (Cai and Liu 2011) estimates $\boldsymbol{\beta}$ by solving

$$\widehat{\boldsymbol{\beta}}_{LPD} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\beta}\|_1 \text{ subject to } \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\delta}}\|_\infty \le \lambda,$$

where λ is a tuning parameter and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ is as defined in (2.1.1) and (2.1.2). The regularized optimal affine discriminant (ROAD) method (Fan et al. 2012) estimates $\boldsymbol{\beta}$ by solving

$$\widehat{\boldsymbol{\beta}}_{ROAD} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\ (1/2)\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 + (\gamma/2)(\boldsymbol{\beta}^T\widehat{\boldsymbol{\delta}} - 1)^2,$$

where λ and γ are tuning parameters and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ is as defined in (2.1.1) and (2.1.2). Then, replacing $\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\delta}}$ in the SLDA rule with $\widehat{\boldsymbol{\beta}}_{LPD}$ or $\widehat{\boldsymbol{\beta}}_{ROAD}$ gives the corresponding LPD or ROAD rule. Both papers showed that the resulting rules' misclassification error asymptotically converges to the Bayes error, given some sparsity condition on $\boldsymbol{\beta}$.

In general, these methods showed that as long as the unknown population parameters satisfy some sparsity assumptions, building a regularized LDA classifier accordingly can yield a consistent classification rule, in the sense that its misclassification error converges to the Bayes error. For example, Shao et al. (2011) showed that if the difference of population means and the covariance matrices are sparse, utilizing thresholding estimators (Bickel and Levina 2008a) can still yield a consistent rule. On the other hand, Cai and Liu (2011), Fan et al. (2012) and Mai et al. (2012) separately developed distinct consistent rules while assuming the slope of the Bayes rule is sparse. Han et al. (2013) further relaxed the normality assumption on these rules and extended them to more general distributions by using a Gaussian copula method.

Although these rules are guaranteed to be consistent, learning the rules is computationally difficult. For example, both SLDA and LPD need to first compute $\widehat{\boldsymbol{\Sigma}}$, which requires $O(np^2)$ operations. For SLDA, it requires additional operations to obtain the regularized estimators $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\delta}}$. Besides that, inverting $\widetilde{\boldsymbol{\Sigma}}$ costs another $O(p^{2+\epsilon})$ operations for some $\epsilon \in (0, 1]$, depending on the actual algorithm used to invert a matrix. Finally, computing the product $\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\delta}}$ costs $O(p^2)$ operations. Thus the total computational cost of SLDA is at least $\max\{O(np^2), O(p^{2+\epsilon})\}$. For LPD, the optimization problem can be solved by the primal-dual interior-point method (Candes et al. 2007). As shown by Candes et al. (2007), in the scenario that $p \gg n$, each iteration requires solving an $n \times n$ linear system $(O(n^2))$ and updating the matrix for the system $(O(np^2))$, which also requires evaluating $\widehat{\boldsymbol{\Sigma}}$. Such an evaluation already takes $O(np^2)$ operations. Therefore, let T be the number of iterations for the interior-point method to converge. The total computational cost for LPD is $\max\{O(Tnp^2), O(Tn^2)\}$. For ROAD, if one chooses to evaluate $\widehat{\boldsymbol{\Sigma}}$ first and then solve the optimization problem, the computational cost is at least $O(np^2)$. A more computationally efficient solution is to use the fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle (2009). In each iteration of FISTA, the computational cost to compute the gradient is $O(np)$ and it is shown in Beck and Teboulle (2009) that FISTA needs at least $O(n^{1/4})$ iterations to converge. Thus, the total computational cost for FISTA to solve the ROAD problem is at least $O(n^{5/4}p)$. One may also choose to use the covariance-based method (Friedman et al. 2010) to calculate the gradient. However, its efficiency depends on the choice of tuning parameters and the initial value so that the total computational cost is hard to be quantified in general. Therefore, for large data sets with considerably large sample size and ultra-high dimension, it takes a long time to learn these rules, which motivates us to develop computationally more efficient algorithm with theoretical gaurantee that can solve the high-dimensional linear discriminant analysis problem.

2.2 Copula-based Statistical Learning for Regression

Due to the semiparametric nature that the marginal transformations are unspecified, copula model is well-known for modelling the joint distribution of continuous variables with skewed marginal distributions. The robustness of copula model on a broad class of distributions makes it one of the most favorable choices in statistical learning, especially regression problems with random design. Some corresponding examples are discussed below.

2.2.1 Unsupervised Learning

For an unsupervised problem of estimating correlations among mixed variables, some recent works have provided solutions using copula-based method (Liu et al. 2009; 2012; Fan et al. 2017; Feng and Ning 2019; Yoon et al. 2020).

Specifically, Liu et al. (2009; 2012) proposed a Gaussian copula model to estimate correlation among continuous variables. For a random vector $\mathbf{z} \in \mathbf{R}^p$, if $\mathbf{f}(\mathbf{z}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{f} = (\mathrm{f}_1, ..., \mathrm{f}_p)^T$, f_j is an unspecified monotonically increasing function and $\boldsymbol{\Sigma}$ is a correlation matrix, then $\mathbf{z}$ is said to follow a nonparanormal distribution, denoted by $\mathbf{z} \sim NPN(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{f})$. Such model extends the multivariate Gaussian distribution to Gaussian copula, but still restricts the elements of $\mathbf{z}$ to be all continuous. Liu et al. (2009) proposed to estimate $\boldsymbol{\Sigma}$ by using the normalized score method based on normalizing $\mathbf{z}$, which required estimating the marginal transformation function $\mathbf{f}$. Later Liu et al. (2012) proposed to estimate $\boldsymbol{\Sigma}$ by bridging the elements of Kendall's tau correlation matrix based on $\mathbf{z}$ to elements of $\boldsymbol{\Sigma}$. If we observe n i.i.d samples of variables Z_j and Z_k, then the Kendall's tau correlation estimate between Z_j and Z_k is given by

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \mathrm{sgn}(Z_{ij} - Z_{i'j})\mathrm{sgn}(Z_{ik} - Z_{i'k}), 1 \leq j < k \leq p.$$

Since $\tau_{jk} = E(\widehat{\tau}_{jk}) = 2\mathrm{sin}^{-1}(\Sigma_{jk})/\pi$ for $j \neq k$, the bridged correlation estimator was given by

$\widehat{\Sigma}_{jk} = \sin(\pi \widehat{\tau}_{jk}/2)$ for $j \neq k$. Such method circumvent the estimation of $\mathbf{f}$, resulting in optimal convergence rate. This idea motivated several following methods for estimating $\mathbf{\Sigma}$ given mixed types of variables.

Fan et al. (2017) proposed a latent Gaussian copula model to simultaneously model the joint distribution of continuous and binary variables. They assumed that the observed variable X_j related to the latent variable Z_j based on the following transformations

$$X_j = \begin{cases} Z_j, & \text{for } j \in \mathcal{C}; \\ I(Z_j > C_j), & \text{for } j \in \mathcal{B}, \end{cases}$$

where $\mathcal{C}$ and $\mathcal{B}$ are the index sets of continuous and binary variables, $(C_j)_{j \in \mathcal{B}}$ is a vector of unknown thresholds for binary variables. The observed variable $\mathbf{x}$ is said to follow a latent nonparanormal distribution, denoted by $LNPN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{f}, \mathbf{C})$. The estimation for $\mathbf{\Sigma}$ used the idea of bridging Kendall's tau correlation τ_{jk} based on $\mathbf{x}$ to Σ_{jk}, and the bridge functions are given as the following,

$$\tau_{jk} = F_{jk}(\Sigma_{jk}) = \begin{cases} 2\sin^{-1}(\Sigma_{jk})/\pi, & \text{for } j \in \mathcal{C},\ k \in \mathcal{C}; \\ 2(\Phi_2(\Delta_j, \Delta_k, \Sigma_{jk}) - \Phi(\Delta_j)\Phi(\Delta_k)), & \text{for } j \in \mathcal{B},\ k \in \mathcal{B}; \\ 4\Phi_2(\Delta_k, 0, \Sigma_{jk}/\sqrt{2}) - 2\Phi(\Delta_k), & \text{for } j \in \mathcal{C},\ k \in \mathcal{B}; \end{cases}$$

where $\Delta_j = \mathrm{f}_j(C_j)$ for $j \in \mathcal{B}$, Φ_2 is the two-dimensional standard multivariate Gaussian distribution function. Since the parameters Δ_j for $j \in \mathcal{B}$ can be estimated by some moment estimators,

$$\widehat{\Delta}_j = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} X_{ij}), \text{ for } j \in \mathcal{B},$$

and these bridge functions are all invertible, then the estimate $\widehat{\Sigma}_{jk}$ can be obtained by solving $\widehat{\tau}_{jk} = \widehat{F}_{jk}(\widehat{\Sigma}_{jk})$, which still circumvent the estimation of transformation $\mathbf{f}$.

Feng and Ning (2019) generalized the latent Gaussian copula model to to handle ordinal and categorical variables with arbitrarily many levels. They modeled an ordinal variable X_j by

$$X_j = \sum_{k=1}^{N_j} I(Z_j > C_{jk}), \text{ for } j \in \mathcal{O};$$

where $\mathcal{O}$ is the index set of ordinal variables, and $C_{j1} < ... < C_{jN_j}$ $(j \in \mathcal{O})$ are the N_j unknown thresholds for an ordinal variable with $N_j + 1$ levels, and incorporated the ordinal variables to the latent nonparanormal distribution. To estimate the latent correlation involving ordinal variables, Feng and Ning (2019) proposed the following ensemble approach. Suppose X_j and X_k are ordinal variables with $N_j + 1$ and $N_k + 1$ levels, then we let $X_{ij}^{(p)} = I(X_{ij} \geq p)$, $p = 1, ..., N_j$, and $X_{ik}^{(q)} = I(X_{ik} \geq q)$, $q = 1, ..., N_k$. With the dichotomized variables $X_{ij}^{(p)}$ and $X_{ik}^{(q)}$, we can estimate $\Delta_j^{(p)} = \mathrm{f}_j(C_{jp})$ and $\Delta_k^{(q)} = \mathrm{f}_k(C_{kq})$ by

$$\widehat{\Delta}_j^{(p)} = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} X_{ij}^{(p)}),\ \widehat{\Delta}_k^{(q)} = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} X_{ik}^{(q)}),\ \text{for } j, k \in \mathcal{O}.$$

Using the bridge functions for binary variables, we can estimate the latent correlations between each pair of these binary variables by solving $\widehat{F}_{jk}(\widehat{\Sigma}_{jk}^{(p,q)}) = \widehat{\tau}_{jk}^{(p,q)}$, where

$$\widehat{\tau}_{jk}^{(p,q)} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \mathrm{sgn}(X_{ij}^{(p)} - X_{i'j}^{(p)}) \mathrm{sgn}(X_{ik}^{(q)} - X_{i'k}^{(q)}),\ p = 1...N_j,\ q = 1...N_k.$$

Finally, they proposed to use the weighted average of these latent correlations to obtain the point estimator of the correlation between ordinal variables, which has the form of

$$\widehat{\Sigma}_{jk} = \sum_{q=1}^{N_k} \sum_{p=1}^{N_j} \widehat{\Sigma}_{jk}^{(p,q)} w_{jk}^{(p,q)}.$$

If X_j is ordinal and X_k is of other types, a similar estimator can be constructed as $\widehat{\Sigma}_{jk} = \sum_{p=1}^{N_j} \widehat{\Sigma}_{jk}^{(p)} w_{jk}^{(p)}$, for $p = 1, ..., N_j$, where

$$\widehat{\tau}_{jk}^{(p)} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \mathrm{sgn}(X_{ij}^{(p)} - X_{i'j}^{(p)}) \mathrm{sgn}(X_{ik} - X_{i'k}),\ \text{and } \widehat{F}_{jk}(\widehat{\Sigma}_{jk}^{(p)}) = \widehat{\tau}_{jk}^{(p)}.$$

In the above estimators, the weights must satisfy $0 \leq w_{jk}^{(p,q)} \leq 1, \sum_{q=1}^{N_k} \sum_{p=1}^{N_j} w_{jk}^{(p,q)} = 1$, and $0 \leq w_{jk}^{(p)} \leq 1, \sum_{p=1}^{N_j} w_{jk}^{(p)} = 1$. For simplicity, it suffices to use $w_{jk}^{(p,q)} = 1/(N_j N_k)$ and $w_{jk}^{(p)} = 1/N_j$. Such ensemble estimator also circumvent the estimation of $\mathbf{f}$ and its statistical propoerties are similar to the estimators given in Fan et al. (2017).

Yoon et al. (2020) proposed a truncated latent Gaussian copula model to deal with truncated variables. They modeled a truncated variable X_j by

$$X_j = I(Z_j > C_j)Z_j, \text{ for } j \in \mathcal{T};$$

where $\mathcal{T}$ is the index set for truncated variables, $\mathbf{C}_{\mathcal{T}} = (C_j)_{j \in \mathcal{T}}$ is a vector of unknown thresholds truncated variables, and incorporated the truncated variables to the latent nonparanormal distribution that contains continuous, binary and truncated variables. The estimation for $\mathbf{\Sigma}$ still used the idea of bridging Kendall's tau correlation τ_{jk} based on $\mathbf{x}$ to Σ_{jk}, and the bridge functions are given as the following,

$$\tau_{jk} = F_{jk}(\Sigma_{jk}) = \begin{cases} 2(1-\Phi(\Delta_j))\Phi(\Delta_k) - 2\Phi_3(-\Delta_j, \Delta_k, 0; \mathbf{\Sigma}_{3a}) \\ \quad - 2\Phi_3(-\Delta_j, \Delta_k, 0; \mathbf{\Sigma}_{3b}), & \text{for } j \in \mathcal{T},\ k \in \mathcal{B}; \\ -2\Phi_2(-\Delta_j, 0; 1/\sqrt{2}) + 4\Phi_3(-\Delta_j, 0, 0; \mathbf{\Sigma}_3), & \text{for } j \in \mathcal{T},\ k \in \mathcal{C}; \\ -2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \mathbf{\Sigma}_{4a}) + 2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \mathbf{\Sigma}_{4b}), & \text{for } j \in \mathcal{T},\ k \in \mathcal{T}. \end{cases}$$

$$\mathbf{\Sigma}_{3a} = \begin{bmatrix} 1 & -\Sigma_{jk} & 1/\sqrt{2} \\ -\Sigma_{jk} & 1 & -\Sigma_{jk}/\sqrt{2} \\ 1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} & 1 \end{bmatrix}, \mathbf{\Sigma}_{3b} = \begin{bmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & -\Sigma_{jk}/\sqrt{2} \\ -1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} & 1 \end{bmatrix},$$

$$\mathbf{\Sigma}_3 = \begin{bmatrix} 1 & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ 1/\sqrt{2} & 1 & \Sigma_{jk} \\ \Sigma_{jk}/\sqrt{2} & \Sigma_{jk} & 1 \end{bmatrix},$$

$$\mathbf{\Sigma}_{4a} = \begin{bmatrix} 1 & 0 & 1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} \\ 0 & 1 & -\Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -\Sigma_{jk}/\sqrt{2} & 1 & -\Sigma_{jk} \\ -\Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & -\Sigma_{jk} & 1 \end{bmatrix},$$

$$\mathbf{\Sigma}_{4b} = \begin{bmatrix} 1 & \Sigma_{jk} & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ \Sigma_{jk} & 1 & \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} & 1 & \Sigma_{jk} \\ \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & \Sigma_{jk} & 1 \end{bmatrix},$$

where $\Delta_j = \mathrm{f}_j(C_j)$ for $j \in \mathcal{B} \cup \mathcal{T}$,, Φ_d is the cumulative distribution function of the d-dimensional standard normal distribution. Since the parameters Δ_j for $j \in \mathcal{B} \cup \mathcal{T}$ can be estimated by some moment estimators,

$$\widehat{\Delta}_j = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} X_{ij}), \text{ for } j \in \mathcal{B};$$

$$\widehat{\Delta}_j = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} I(X_{ij} > 0)), \text{ for } j \in \mathcal{T},$$

and these bridge functions are all invertible, then the estimate $\widehat{\Sigma}_{jk}$ can be obtained by solving $\widehat{\tau}_{jk} = \widehat{F}_{jk}(\widehat{\Sigma}_{jk})$, which again circumvent the estimation of transformation $\mathbf{f}$.

In summary, these methods assume that there exist some latent continuous variables that generate the observed mixed variables, and the latent continuous variables follow a joint standard normal distribution, after applying some marginal transformations. The rank-based correlation is invariant of the transformations and could be bridged elementwisely to the latent correlation, which circumvent the estimation of transformations. Hence copula-based methods can be applied to a series of unsupervised learning problems with mixed data, such as graph estimation (Liu et al. 2009; 2012; Fan et al. 2017; Feng and Ning 2019), principal component analysis (Fan et al. 2017) and canonical correlation analysis (Yoon et al. 2020).

2.2.2 Supervised Learning for Regression

In terms of the supervised learning for regression problem, a few copula-based methods have been developed to handle the problem in the low-dimensional setting (Sungur 2005; Pitt et al. 2006; Crane and Hoek 2008; Masarotto et al. 2012; Noh et al. 2013). For example, Masarotto et al. (2012) proposed a general framework for the inference and model diagnosis using Gaussian copula when the responses are dependent. Noh et al. (2013) proposed a plug-in estimator of the regression function for a general copula regression. However, these methods only handle a low-dimensional copula regression model. More recently, Cai and Zhang (2018) proposed a high-dimensional Gaussian copula regression model. They assume that the response and the covariates jointly follows a Gaussian copula. $(\mathbf{x}^T, Y)^T$ jointly follows the distribution of $NPN(\mathbf{0}, \check{\mathbf{\Sigma}}, \check{\mathbf{f}})$ where $\check{\mathbf{f}} = (\mathbf{f}, \mathrm{f}_0)$ and $\check{\mathbf{\Sigma}}$ is the correlation matrix of $(\mathbf{f}(\mathbf{x})^T, \mathrm{f}_0(Y))^T$. This assumption implies that their regression model is

$$\mathrm{f}_0(Y) = \mathbf{f}(\mathbf{x})^T\boldsymbol{\theta} + \epsilon, \tag{2.2.1}$$

where $\boldsymbol{\theta} = \check{\boldsymbol{\Sigma}}_{xx}^{-1}\check{\boldsymbol{\Sigma}}_{xy}$, $\epsilon \sim N(0, 1 - \check{\boldsymbol{\Sigma}}_{xy}^{T}\check{\boldsymbol{\Sigma}}_{xx}^{-1}\check{\boldsymbol{\Sigma}}_{xy})$ and is independent of $\mathbf{f}(\mathbf{x})$, $\check{\boldsymbol{\Sigma}}_{xy} = \mathrm{E}(\mathrm{f}_0(Y)\mathbf{f}(\mathbf{x}))$ and $\check{\boldsymbol{\Sigma}}_{xx} = \mathrm{E}(\mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^T)$. Since both the response Y and the covariates $\mathbf{x}$ were restricted to continuous variables, they developed a rank-based method using results from Liu et al. (2012) to estimate the coefficients in their model and established the oracle properties of their proposed estimator. Specifically, they obtained $\widehat{\boldsymbol{\Sigma}}$ for $\check{\boldsymbol{\Sigma}}$ directly using the method from Liu et al. (2012), then extract its submatrices $\widehat{\boldsymbol{\Sigma}}_{xx}$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ to formulate the L_1 penalized estimator $\widehat{\boldsymbol{\theta}}$

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbf{R}^p} \frac{1}{2}\boldsymbol{\theta}^T\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\theta} - \boldsymbol{\theta}^T\widehat{\boldsymbol{\Sigma}}_{xy} + \lambda\|\boldsymbol{\theta}\|_1,$$

where $\widehat{\boldsymbol{\Sigma}}_{xx}$ is a positive definite estimator for $\check{\boldsymbol{\Sigma}}_{xx}$, λ is the tuning parameter for the convex optimization. Furthermore, they estimated the marginal transformation $\mathbf{f}$ and obtained the following predictor for the response

$$\widehat{Y} = \widehat{\mathrm{f}}_0^{-1}\Big(\sum_{1 \le j \le p} \widehat{\mathrm{f}}_j(X_j)\widehat{\theta}_j\Big), \tag{2.2.2}$$

where the transformation function estimates are given by

$$\widehat{\mathrm{f}}_j(t) = \Phi^{-1}(\widehat{F}_j(t)), \text{ for } j = 1, ..., p; \widehat{\mathrm{f}}_0(t) = \Phi^{-1}(\widetilde{F}_0(t)),$$

$\widehat{F}_j(t)$ is just the empirical distribution function for X_j using the training set, while $\widetilde{F}_0(t)$ is the winsorized empirical distribution function for Y using the training set with winsorization level $1/n^2$, and $\widehat{\mathrm{f}}_0^{-1}$ is the generalized inverse for $\widehat{\mathrm{f}}_0$, defined as

$$\widehat{\mathrm{f}}_0^{-1}(t) = \inf\{x \in \mathbf{R} : \widehat{\mathrm{f}}_0(x) > t\}.$$

This model has several limitations. First, $\mathbf{x}$ in (2.2.1) only contains continuous variables, the corresponding theory of copula regression has never been studied in the latent Gaussian copula model context. Secondly, in terms of prediction, the predicted value given by (2.2.2) has to be one of the responses in the training set, since an estimator of f_0 must be obtained from the training set to predict Y. With the works shown in section 2.2.1, it is our motivation to build a regression model based on latent Gaussian copula model for mixed types of covariates and carefully study its theoretical properties in estimating the regression coefficients and prediction. Our prediction should avoid estimating the unknown transformation f_0 for the response Y so that the predicted value will not be restricted to training set.

2.3 Copula-based Statistical Learning for Classification

In terms of the supervised learning for classification problem, Han et al. (2013) proposed a high-dimensional copula discriminant analysis rule. For a binary classification problem with class label $Y \in \{0, 1\}$ with equal prior probabilities, they assumed that the p-dimensional features for each class has $\mathbf{x}_0 \sim NPN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}, \mathbf{f})$ and $\mathbf{x}_1 \sim NPN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}, \mathbf{f})$, where μ_0 and μ_1 are the means for the transformed features $\mathbf{f}(\mathbf{x}_0)$ and $\mathbf{f}(\mathbf{x}_1)$, $\boldsymbol{\Sigma}$ is the common covariance matrix for the transformaed features $\mathbf{f}(\mathbf{x})$ in each class. In this setting, the corresponding Bayes rule is $D_{Bayes}(\mathbf{x}) = I((\mathbf{f}(\mathbf{x}) - \boldsymbol{\mu})^T\boldsymbol{\beta} \leq 0)$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$, $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ and $\mathbf{x}$ is a new observation. Similar to ROAD introduced in section 2.1, they proposed to estimate the parameter $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}}_{coda} = \underset{\boldsymbol{\beta}\in\mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \frac{\nu}{2}(\boldsymbol{\beta}^T\widehat{\boldsymbol{\delta}} - 1)^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.3.1}$$

where ν is set to be n_0n_1/n^2, $\widehat{\boldsymbol{\Sigma}} = n_0/n \cdot \widehat{\boldsymbol{\Sigma}}_0 + n_1/n \cdot \widehat{\boldsymbol{\Sigma}}_1$, $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{\boldsymbol{\Sigma}}_1$ are both obtained by $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{V}}^{1/2}\widehat{\mathbf{R}}\widehat{\mathbf{V}}^{1/2}$, $\widehat{\mathbf{V}}$ is the diagonal matrix with variance estimates for $\mathrm{f}_j(X_j)$, $\widehat{\mathbf{R}}$ is the estimated correlation matrix by bridging Spearman's rho/Kendall's tau and it has

$$\widehat{R}^{\rho}_{jk} = \begin{cases} 2\sin(\pi\widehat{\rho}_{jk}/6), & \text{for } j \neq k \\ 1, & \text{for } j = k \end{cases} \quad \textit{and} \quad \widehat{R}^{\tau}_{jk} = \begin{cases} \sin(\pi\widehat{\tau}_{jk}/2), & \text{for } j \neq k \\ 1, & \text{for } j = k \end{cases},$$

$\widehat{\boldsymbol{\delta}}$ is obtained by (2.1.1) and $\widehat{\boldsymbol{\mu}}$ is obtained by (2.1.2). Each of the marginal transformations $\mathbf{f}$ is estimated by

$$\begin{aligned} &\widehat{\mathrm{f}}_j(t) = (n_0/n)\widehat{\mathrm{f}}_{0j}(t) + (n_1/n)\widehat{\mathrm{f}}_{1j}(t), j = 1, ..., p \\ &\widehat{\mathrm{f}}_{0j}(t) = \widehat{\boldsymbol{\mu}}_0 + \widehat{V}_{jj}^{-1/2}\Phi^{-1}(\widetilde{F}_{0j}(t)), \\ &\widehat{\mathrm{f}}_{1j}(t) = \widehat{\boldsymbol{\mu}}_1 + \widehat{V}_{jj}^{-1/2}\Phi^{-1}(\widetilde{F}_{1j}(t)) \end{aligned}$$

where $\widetilde{F}_{0j}$ and $\widetilde{F}_{1j}$ are winsorized empirical distribution functions for X_j using samples from class 0 and from class 1 respectively, the winsorization level is set to be $1/(2n)$.

So the copula discriminant analysis rule is given by

$$D_{coda}(\mathbf{x}) = I((\widehat{\mathbf{f}}(\mathbf{x}) - \widehat{\boldsymbol{\mu}})^T \widehat{\boldsymbol{\beta}} \leq 0).$$

As an application, He et al. (2020) proposed an integrative copula discriminant analysis rule by substituting the penalty in (2.3.1) with the penalty in (5) of Li and Li (2018) to make binary classification with features from multiple genomic modalities.

This classification method has several limitations. First, discriminant analysis primarily handles binary classification and requires further extension to multi-class classification, which commonly appears with an ordinal response. Second, this copula discriminant analysis has strong restrictions about the marginal transformations $\mathbf{f}$. If we want the parameters $\widehat{\boldsymbol{\Sigma}}$, $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ to be identifiable while we do not estimate the transformations $\mathbf{f}$, then $\mathbf{f}$ has to preserve the population means and standard deviations

$$\mathrm{E}(X_j) = \mathrm{E}(\mathrm{f}_j(X_j)), \mathrm{Var}(X_j) = \mathrm{Var}(\mathrm{f}_j(X_j)), j = 1, ..., p.$$

Furthermore, the marginal transformations $\mathbf{f}$ must be in the Subgaussian Transformation Function Class (Han et al. 2013) to obtain fast convergence rate for $\widehat{\boldsymbol{\mu}}$ and $\widehat{\mathbf{Y}}$. Although these assumptions on $\mathbf{f}$ resolved the identifiability issue for $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$ and $\mathbf{V}$, they cannot be verified in practice and restrict the model to a smaller class of distributions.

Motivated by the works from sections 2.2.1 and 2.2.2, we can jointly model mixed variables by latent Gaussian copula model, which includes the ordinal response and the continuous features. With this motivation, we can assume that the ordinal response is generated by a latent continuous variable so that this latent variable and the continuous features jointly follow a Gaussian copula, or equivalently, the latent variable satisfy a regression model with the covariates being the continuous features. The regression coefficient estimates solely relies on bridging the rank-based correlation estimator, which has no restriction on $\mathbf{f}$ except monotonicity. We can derive the Bayes rule under the joint model and devise a Fisher consistent classification rule for ordinal classification.

CHAPTER 3
AN EFFICIENT GREEDY SEARCH ALGORITHM FOR HIGH-DIMENSIONAL LINEAR DISCRIMINANT ANALYSIS

3.1 Introduction

Classification—assigning a subject to one of several classes based on certain features—is an important statistical problem. However, the recent emergence of big data poses great challenges, for it requires the efficient use of many features for classification. A simple classifier, namely, linear discriminant analysis (LDA) was widely used before the big data era (Anderson 1958). However, as Bickel and Levina (2004) have shown, when the number of features exceeds the sample size, a traditional LDA is no longer applicable, owing to the accumulation of errors when estimating the unknown parameters. To deal with the high-dimensional LDA problem, a number of regularization methods have been proposed (Clemmensen et al. 2011; Witten and Tibshirani 2011; Shao et al. 2011; Cai and Liu 2011; Fan et al. 2012; Mai et al. 2012; Han et al. 2013). Early works by Clemmensen et al. (2011) and Witten and Tibshirani (2011) proposed solving the regularized Fisher's discriminant problem using sparsity-induced penalties. However, at that stage, there was little theory on the statistical properties of such classifiers. These properties were subsequently studied in more detail after additional regularized LDA classifiers (Shao et al. 2011; Cai and Liu 2011; Fan et al. 2012; Mai et al. 2012; Han et al. 2013) had been proposed to deal with the high-dimensional LDA problem.

In general, these methods showed that as long as the unknown population parameters satisfy some sparsity assumptions, building a regularized LDA classifier can yield a consistent classification rule, in the sense that its misclassification error converges to the Bayes error. For example, Shao et al. (2011) showed that if the difference between the population means and the covariance matrices are sparse, using thresholding estimators (Bickel and Levina 2008a) can still yield a consistent rule. On the other hand, Cai and Liu (2011), Fan et al. (2012), and Mai et al. (2012) separately developed distinct consistent rules while assuming that the slope of the Bayes rule is sparse. Han et al. (2013) relaxed the normality assumption on these rules, extending them to more general distributions using a Gaussian copula method.

Although these rules are guaranteed to be consistent, learning the rules is computationally difficult: the rule proposed by Shao et al. (2011) must invert a high-dimensional covariance matrix, and the other aforementioned rules must solve large-scale optimization problems. In particular, it takes a long time to learn these rules when the dimension is ultrahigh. Therefore, we propose a computationally efficient classifier that can be learned without needing to invert large matrices or solve large-scale optimization problems. Our proposed classifier is based solely on closed-form formulae.

Our method is motivated by a recent study (Li and Li 2018) on the Bayes error of the LDA problem. Li and Li (2018) showed that the Bayes error always decreases when new features are added to the Bayes rule, and that this decrease is fully characterized by the increment of the Mahalanobis distance between the two classes. We therefore develop an efficient greedy search algorithm to learn the increment of the Mahalanobis distance. Unlike many other methods, this algorithm does not estimate all population parameters; instead, it selects discriminative features in a sequential way, and computes the classification rule as it does so. Our method is therefore scalable for ultrahigh-dimensional LDA problems. We show that the proposed method admits both variable selection and error rate consistency when the classes follow some general distributions.

To prove these theoretical properties, we first establish a concentration result for the estimated increment of the Mahalanobis distance and the true increment. This result characterizes the trade-off between the gains of using more features for classification and the additional estimation error it produces. We also offer an explicit interpretation of how much information a new feature adds to the LDA problem; this interpretation holds for a general class of distributions, and is new to the LDA literature. We then show that if the slope of the Bayes rule is exactly sparse, our method can asymptotically recover its nonzero elements, and that our method's misclassification error converges to the Bayes error. These results also hold under a general class of elliptical distributions. We demonstrate numerically that our method achieves comparable or even better classification performance with a much shorter training time than other LDA-based methods.

The rest of the chapter is organized as follows. Section 3.2 presents our efficient greedy search algorithm. Section 3.3 describes the statistical properties of the proposed method in terms of its variable selection and error rate consistency. Section 3.4 relaxes the normality assumption, and shows that the statistical properties hold for a general class of distributions. Section 3.5 presents extensive numerical studies that compare the proposed method with other existing methods, demonstrating the proposed method's superiority in terms of both computational efficiency and classification performance under various scenarios. In Section 3.6, we apply our method to microarray data to classify cancer subtypes, and show that our method renders a more meaningful classification rule. All technical proofs are given in Section 3.8.

3.2 An Efficient Greedy Search Algorithm

Consider a binary classification problem where the class label $Y \in \{0, 1\}$ has a prior distribution of $P(Y = k) = \pi_k$, for $k = 0, 1$. Suppose $\boldsymbol{x}_k \in \mathbb{R}^p$ denotes a p-dimensional vector of features from the kth class that follows the normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are assumed to be independent. The Bayes rule of this classification problem is given by $D_{Bayes}(\boldsymbol{x}) = I(\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \leq \log(\pi_1/\pi_0))$, where $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, $\boldsymbol{\mu} = (1/2)(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$, and $\boldsymbol{x}$ is a new observation. The corresponding Bayes error is given by $R_{Bayes} = \Phi(-\sqrt{\Delta_p}/2)$, where $\Delta_p = \boldsymbol{\delta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ is the Mahalanobis distance between the centroids of the two classes and Φ is the cumulative distribution function of the standard normal distribution.

In practice, the Bayes rule is unknown. A classification rule is learned from the training data $\boldsymbol{X} = \{\boldsymbol{x}_{ki}; k = 0, 1; i = 1, \ldots, n_k\}$, where $\boldsymbol{x}_{ki}$ are independent and identically distributed (i.i.d) samples from the kth class and n_k is the sample size of the kth class, with $n = n_0 + n_1$. Then, the rule is applied to classify a new observation $\boldsymbol{x}$, which is assumed to be independent of the training data. For the classical LDA method, the unknown parameters in the Bayes rule are replaced with their maximum likelihood estimators; this LDA rule has the form

$$D_{LDA}(\boldsymbol{x}) = I(\widehat{\boldsymbol{\delta}}^T\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) \leq \log(\widehat{\pi}_1/\widehat{\pi}_0)),$$

where

$$\widehat{\pi}_k = n_k/n,\ \widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{x}_{ki},\ \widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1, \tag{3.2.1}$$

$$\widehat{\boldsymbol{\mu}} = (1/2)(\widehat{\boldsymbol{\mu}}_0 + \widehat{\boldsymbol{\mu}}_1),\ \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=0}^{1}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ki} - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{ki} - \widehat{\boldsymbol{\mu}}_k)^T. \tag{3.2.2}$$

However, in the high-dimensional setting, where $p > n$, the classical LDA method is no longer feasible, because $\widehat{\boldsymbol{\Sigma}}$ is not invertible. Even if we replace $\widehat{\boldsymbol{\Sigma}}^{-1}$ with a generalized matrix inverse, Bickel and Levina (2004) showed that the resulting rule has an asymptotic misclassification error of 1/2, which is as bad as random guessing. This is essentially due to the error accumulation when estimating the high-dimensional parameters in the classifier. To avoid this issue in the high-dimensional setting, many regularized methods have been proposed (Clemmensen et al. 2011; Witten and Tibshirani 2011; Shao et al. 2011; Cai and Liu 2011; Fan et al. 2012; Mai et al. 2012; Han et al. 2013). In particular, Shao et al. (2011) proposed the sparse linear discriminant analysis (SLDA) rule

$$D_{SLDA}(\boldsymbol{x}) = I(\widetilde{\boldsymbol{\delta}}^T \widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}) \leq \log(\widehat{\pi}_1/\widehat{\pi}_0)),$$

where $\widetilde{\boldsymbol{\delta}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are the thresholding estimators. Here, $\widetilde{\boldsymbol{\delta}} = (\widetilde{\delta}_j)$, with $\widetilde{\delta}_j = \widehat{\delta}_j I(|\widehat{\delta}_j| > t_\delta)$ and $\widehat{\delta}_j$ is the jth element of $\widehat{\boldsymbol{\delta}}$, and $\widetilde{\boldsymbol{\Sigma}} = (\widetilde{\sigma}_{ij})$, with $\widetilde{\sigma}_{ii} = \widehat{\sigma}_{ii}$, $\widetilde{\sigma}_{ij} = \widehat{\sigma}_{ij} I(|\widehat{\sigma}_{ij}| > t_\sigma)$, for $i \neq j$, and $\widehat{\sigma}_{ij}$ is the (i,j)th element of $\widehat{\boldsymbol{\Sigma}}$. They showed that the SLDA's misclassification error still converges to the Bayes error, given that the thresholds t_δ and t_σ are chosen properly and given some sparsity conditions on both $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$. Instead of separately estimating $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$, as the SLDA does, two other methods directly estimate the slope of the Bayes rule $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ by solving convex optimization problems. For example, the linear programming discriminant (LPD) method (Cai and Liu 2011) estimates $\boldsymbol{\beta}$ by solving

$$\widehat{\boldsymbol{\beta}}_{LPD} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_1 \text{ subject to } \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\delta}}\|_\infty \leq \lambda,$$

where λ is a tuning parameter and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ is defined in (3.2.1) and (3.2.2). The regularized optimal affine discriminant (ROAD) method (Fan et al. 2012) estimates $\boldsymbol{\beta}$ by solving

$$\widehat{\boldsymbol{\beta}}_{ROAD} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \ (1/2)\boldsymbol{\beta}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_1 + (\gamma/2)(\boldsymbol{\beta}^T \widehat{\boldsymbol{\delta}} - 1)^2,$$

where λ and γ are tuning parameters and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Sigma}})$ is defined in (3.2.1) and (3.2.2). Then, replacing $\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\delta}}$ in the SLDA rule with $\widehat{\boldsymbol{\beta}}_{LPD}$ or $\widehat{\boldsymbol{\beta}}_{ROAD}$ gives the corresponding LPD or ROAD rule. Both papers showed that the resulting misclassification errors converge asymptotically to the Bayes error, given some sparsity condition on $\boldsymbol{\beta}$.

However, these methods all rely on evaluating and inverting a large matrix or solving a large-scale optimization problem. When p is huge, it can become computationally expensive to run these methods. Instead, we propose an efficient greedy search algorithm that does not require inverting a matrix or solving an optimization problem. Moreover, our method does not need to evaluate the whole sample covariance matrix in advance, but rather computes its elements as it goes, depending on which features enter the classification rule. We compare the computational complexity of these methods with that of ours later in this section. As shown in the numerical studies, our method has a much shorter learning time than these methods do, and even better classification performance.

The Bayes error of the LDA problem is fully characterized by the Mahalanobis distance Δ_p. Recently, Li and Li (2018) proved that Δ_p is a monotonically increasing function of p, which implies that the Bayes error always decreases when more features are involved. Therefore, we propose a greedy search algorithm that operates by learning the increment of the Mahalanobis distance. At each step of our algorithm, we seek the variable that results in the largest increment of the Mahalanobis distance. Such a variable can be regarded as the most informative, given those selected in the previous steps. We terminate the iterations when the increment is smaller than a predefined threshold. We show that the iterations are based on closed-form formulae, and the algorithm does not need to compute the whole covariance matrix; therefore, it is computationally efficient.

Let S be an arbitrary subset of $\{1,\dots,p\}$, and s be the size of S. Let $\Delta_s = \boldsymbol{\delta}_S^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S$ be the Mahalanobis distance involving only variables in S, where $\boldsymbol{\delta}_S$ is a subvector of $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_{SS}$ is a submatrix of $\boldsymbol{\Sigma}$ with indices in S. For an arbitrary $c \notin S$, let

$$\Delta_{s+1} = \begin{pmatrix} \boldsymbol{\delta}_S^T & \delta_c \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{SS} & \boldsymbol{\Sigma}_{Sc} \\ \boldsymbol{\Sigma}_{Sc}^T & \sigma_{cc} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\delta}_S \\ \delta_c \end{pmatrix}$$

be the Mahalanobis distance by adding a new variable indexed by c, and let $\theta_{Sc} = \Delta_{s+1} - \Delta_s$ be the increment of the Mahalanobis distance. Using an argument analogous to Proposition 1 of Li and Li (2018), we can show that

$$\theta_{Sc} = \frac{(\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Omega}_{SS} \boldsymbol{\delta}_S)^2}{\sigma_{cc} - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Omega}_{SS} \boldsymbol{\Sigma}_{Sc}} \geq 0, \tag{3.2.3}$$

where $\boldsymbol{\Omega}_{SS} = \boldsymbol{\Sigma}_{SS}^{-1}$. A proof of (3.2.3) is given in the Appendix. Moreover, under the normality assumption, we find that θ_{Sc} has a clear interpretation. Let $\boldsymbol{z} = \boldsymbol{x}_0 - \boldsymbol{x}_1$. Using the conditional distribution of the multivariate normal distribution, we can easily show that

$$\theta_{Sc} = \frac{2\{\mathrm{E}(z_c|\boldsymbol{z}_S = \mathbf{0})\}^2}{\mathrm{Var}(z_c|\boldsymbol{z}_S = \mathbf{0})},$$

where z_c is the cth element of $\boldsymbol{z}$ and $\boldsymbol{z}_S$ is the subvector of $\boldsymbol{z}$ with indices in S. This result shows that the contribution of a new variable does not depend on its marginal difference between the two classes (i.e., $\mathrm{E}(z_c)$), but rather on its effect size, conditional on other variables (i.e., the standardized $\mathrm{E}(z_c|\boldsymbol{z}_S = \mathbf{0})$). In the extreme case in which there is no difference in the cth variable between the two classes (i.e., $\mathrm{E}(z_c) = 0$), adding such a variable to those in S can still reduce the Bayes error if $\mathrm{E}(z_c|\boldsymbol{z}_S = \mathbf{0}) \neq 0$. This interpretation seems to be new in the LDA literature. In Section 3.4, we show that this interpretation not only holds for the normal distribution, but also holds for all elliptical distributions.

In practice, θ_{Sc} is unknown. We propose a greedy search algorithm based on learning θ_{Sc} from the training data. From (3.2.3), if we replace $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ with the corresponding estimators $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\Sigma}}$ in (3.2.1) and (3.2.2), we can easily obtain an estimator of θ_{Sc}. However, this naive method requires the computation of all elements of $\widehat{\boldsymbol{\Sigma}}$ and the inversion of its submatrices. We show that there is a more efficient method of computing $\widehat{\theta}_{Sc}$ that computes elements of $\widehat{\boldsymbol{\Sigma}}$ as it goes, with no need to invert a matrix.

At the initial step, we set the selected set $\widehat{S}_0 = \emptyset$. For all $1 \leq c \leq p$, we calculate $\widehat{\delta}_c^2/\widehat{\sigma}_{cc}$, where $\widehat{\delta}_c$ is the cth element of $\widehat{\boldsymbol{\delta}}$ and $\widehat{\sigma}_{cc}$ is the (c, c)th element of $\widehat{\boldsymbol{\Sigma}}$. We choose $\widehat{s}_1$ to be the index such that $\widehat{\delta}_c^2/\widehat{\sigma}_{cc}$ is maximized, and set the selected set $\widehat{S}_1 = \{\widehat{s}_1\}$ and the candidate set $\widehat{C}_1 = \{1, \ldots, p\}\backslash\{\widehat{s}_1\}$. To simplify the calculations in the subsequent steps, we compute and store $\widehat{\boldsymbol{\Omega}}_1 = \widehat{\sigma}^{-1}_{\widehat{s}_1\widehat{s}_1}$ and the submatrix $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_1\widehat{C}_1}$, that is, the sample covariance of the selected and candidate variables. At this step, $\widehat{\boldsymbol{\Omega}}_1$ and $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_1\widehat{C}_1}$ are a scalar and a vector of $p-1$ elements, respectively. As shown below, storing these two matrices is the key to enabling a fast computation. At the kth step, we compute

$$\widehat{\theta}_{\widehat{S}_{k-1}c} = (\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{S}_{k-1}c}\widehat{\boldsymbol{\Omega}}_{k-1}\widehat{\boldsymbol{\delta}}_{\widehat{S}_{k-1}})^2(\widehat{\sigma}_{cc} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{S}_{k-1}c}\widehat{\boldsymbol{\Omega}}_{k-1}\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_{k-1}c})^{-1},$$

for all $c \in \widehat{C}_{k-1}$. Because $\widehat{\delta}_c$, $\widehat{\sigma}_{cc}$, $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_{k-1}\widehat{C}_{k-1}}$, and $\widehat{\boldsymbol{\Omega}}_{k-1}$ have all been stored in previous steps, computing $\widehat{\theta}_{\widehat{S}_{k-1}c}$ is fast. Then, we select $\widehat{s}_k$ to be the index that maximizes $\widehat{\theta}_{\widehat{S}_{k-1}c}$ for all $c \in \widehat{C}_{k-1}$, and let $\widehat{S}_k = \widehat{S}_{k-1} \cup \{\widehat{s}_k\}$ and $\widehat{C}_k = \widehat{C}_{k-1} \backslash \{\widehat{s}_k\}$. Next, we update $\widehat{\boldsymbol{\Omega}}_k$, the estimated precision matrix of all selected variables in $\widehat{S}_k$. Using the Woodbury matrix identity, we have

$$\widehat{\boldsymbol{\Omega}}_k = \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{S}_k\widehat{S}_k} = \begin{pmatrix} \widehat{\boldsymbol{\Omega}}_{k-1} + \widehat{\alpha}_k \widehat{\boldsymbol{\rho}}_k \widehat{\boldsymbol{\rho}}_k^T & -\widehat{\alpha}_k \widehat{\boldsymbol{\rho}}_k \\ -\widehat{\alpha}_k \widehat{\boldsymbol{\rho}}_k^T & \widehat{\alpha}_k \end{pmatrix},$$

where $\widehat{\boldsymbol{\rho}}_k = \widehat{\boldsymbol{\Omega}}_{k-1} \widehat{\boldsymbol{\Sigma}}_{\widehat{S}_{k-1}\widehat{s}_k}$ and $\widehat{\alpha}_k = (\widehat{\sigma}_{\widehat{s}_k\widehat{s}_k} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{S}_{k-1}\widehat{s}_k} \widehat{\boldsymbol{\Omega}}_{k-1} \widehat{\boldsymbol{\Sigma}}_{\widehat{S}_{k-1}\widehat{s}_k})^{-1}$. Note that $\widehat{\sigma}_{\widehat{s}_k\widehat{s}_k}$, $\widehat{\boldsymbol{\Omega}}_{k-1}$, and $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_{k-1},\widehat{s}_k}$ can all be read directly from previously stored objects, allowing $\widehat{\boldsymbol{\Omega}}_k$ to be computed efficiently. Next, we update $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_k\widehat{C}_k}$ by letting $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_k\widehat{C}_k} = (\widehat{\boldsymbol{\Sigma}}_{\widehat{C}_k\widehat{S}_{k-1}} \; \widehat{\boldsymbol{\Sigma}}_{\widehat{C}_k\widehat{s}_k})^T$. This quantity is needed to calculate $\widehat{\theta}_{\widehat{S}_k c}$ in the next iteration. Note that $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}_k\widehat{S}_{k-1}}$ is the submatrix of $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}_{k-1}\widehat{S}_{k-1}}$ without its $\widehat{s}_k$th row, which has been stored in the $(k-1)$th iteration. Therefore, we need only compute $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}_k\widehat{s}_k} \in \mathbb{R}^{p-k}$, which is the sample covariance between the newly selected variable $\widehat{s}_k$ and the candidate variables in $\widehat{C}_k$. We calculate the number of operations computed at the kth iteration. First, it takes $O(k^2(p-k+1)) = O(k^2 p)$ operations to calculate $\widehat{\theta}_{\widehat{S}_{k-1}c}$. Then, we obtain $\widehat{\boldsymbol{\Sigma}}_{\widehat{C}_k\widehat{s}_k}$ at a cost of $O(np)$ operations. Finally, we update $\widehat{\boldsymbol{\Omega}}_k$ at a cost of $O(k^2)$ operations. Thus, at the kth iteration, our algorithm costs $O(np)$ operations. Therefore, up to the kth iteration, the total computational cost is $O(knp)$. As discussed in Section 3.3, the total number of iterations is close to K, which is the number of nonzero elements in $\boldsymbol{\beta}$ and is much smaller than n. Thus, the total computational cost of our algorithm is $O(Knp)$.

On the other hand, both the SLDA and the LPD need to first compute $\widehat{\boldsymbol{\Sigma}}$, which requires $O(np^2)$ operations. For the SLDA, it requires additional operations to obtain the regularized estimators $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\delta}}$. Futhermore, inverting $\widetilde{\boldsymbol{\Sigma}}$ costs another $O(p^{2+\epsilon})$ operations for some $\epsilon \in (0, 1]$, depending on the algorithm used to invert the matrix. Finally, computing the product $\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{\delta}}$ costs $O(p^2)$ operations. Thus, the total computational cost of the SLDA is at least $\max\{O(np^2), O(p^{2+\epsilon})\}$, which is much slower than ours when p is big. For the LPD, the optimization problem can be solved using the primal-dual interior-point method (Candes et al. 2007). As shown by Candes et al. (2007), when $p \gg n$, each iteration requires solving an $n \times n$ linear system ($O(n^2)$) and updating the matrix for the system ($O(np^2)$), which also requires evaluating $\widehat{\boldsymbol{\Sigma}}$. Such an evaluation already takes $O(np^2)$ operations. Therefore, let T be the number of iterations for the interior-point method to converge. The total computational cost for the LPD is $\max\{O(Tnp^2), O(Tn^2)\}$, which is clearly slower than ours. For the ROAD, if one chooses to evaluate $\widehat{\boldsymbol{\Sigma}}$ first and then solve the optimization problem, the computational cost is at least $O(np^2)$. A computationally more efficient solution is to use the fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle (2009). In each iteration of the FISTA, the cost of computing the gradient is $O(np)$ if we use a store-and-compute method that is more efficient than evaluating $\widehat{\boldsymbol{\Sigma}}$ before the iterations start. Furthermore, Theorem 4.4 in Beck and Teboulle (2009) shows that the FISTA needs at least $O(n^{1/4})$ iterations to converge. Thus, the total computational cost for the FISTA to solve the ROAD problem is at least $O(n^{5/4}p)$. Therefore, our method is still faster than the FISTA, especially when K is small. One may also choose to use the covariance-based method (Friedman et al. 2010) to calculate the gradient. However, its efficiency depends on the choice of the tuning parameters and the initial value, so that the total computational cost is difficult to quantify, in general.

In conclusion, the greedy search algorithm keeps track of the index sets of selected variables $\widehat{S}_k$ and candidate variables $\widehat{C}_k$, and iteratively updates $\widehat{\boldsymbol{\Omega}}_k$ and $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}_k\widehat{C}_k}$. It does not need to compute the whole $\widehat{\boldsymbol{\Sigma}}$ in advance; instead, it computes its elements as it goes based on selected and candidate variables. It relies solely on closed-form formulae to learn the increments of the Mahalanobis distance, without requiring a matrix inversion or solving an optimization problem. The algorithm terminates when no candidate variable produces an increment of the Mahalanobis distance of at least τ, which is a predefined stopping threshold that can be regarded as a tuning parameter that must be tuned using cross-validation. The greedy search algorithm is summarized in Algorithm 1.

Denote $\widehat{\mathcal{M}}$ and $\widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{M}}}$ as the last $\widehat{S}_k$ and $\widehat{\boldsymbol{\Omega}}_k$, respectively, when the greedy search algorithm terminates. We let $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}} = \widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{M}}}\widehat{\boldsymbol{\delta}}_{\widehat{\mathcal{M}}}$ and propose the following greedy search linear discriminant analysis (GS-LDA) rule:

$$D_{GS\text{-}LDA}(\boldsymbol{x}) = I(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}}^T(\boldsymbol{x}_{\widehat{\mathcal{M}}} - \widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{M}}}) \leq \log(\widehat{\pi}_1/\widehat{\pi}_0)),$$

where $\boldsymbol{x}_{\widehat{\mathcal{M}}}$ and $\widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{M}}}$ are subvectors of $\boldsymbol{x}$ and $\widehat{\boldsymbol{\mu}}$, respectively, with indices in $\widehat{\mathcal{M}}$.

3.3 Theoretical Properties

Here, we give two theoretical results for the statistical properties of the GS-LDA rule. First, we show that if $\boldsymbol{\beta}$ is exactly sparse, the greedy search algorithm can correctly select its nonzero elements with high probability. Second, we show that the misclassification rate of the GS-LDA rule

Algorithm 3.1: The greedy search algorithm.

```
Initialization: Compute and store $\widehat{\boldsymbol{\delta}}$ and the diagonal elements of $\widehat{\mathbf{\Sigma}}$
                using (3.2.1) and (3.2.2).
                Set $\widehat{s}_1 = \operatorname{argmax}_{j \le p} \widehat{\delta}_j^2 / \widehat{\sigma}_{jj}$, $\widehat{S}_1 = \{\widehat{s}_1\}$,
                $\widehat{C}_1 = \{1, \ldots, p\} \backslash \{\widehat{s}_1\}$, $\widehat{\mathbf{\Omega}}_1 = \widehat{\sigma}^{-1}_{\widehat{s}_1 \widehat{s}_1}$.
                Compute and store $\widehat{\mathbf{\Sigma}}_{\widehat{S}_1 \widehat{C}_1}$.
At the $k$th iteration:
    Let $\widehat{\theta}_{\widehat{S}_{k-1} c} = \frac{(\widehat{\delta}_c - \widehat{\mathbf{\Sigma}}^T_{\widehat{S}_{k-1} c} \widehat{\mathbf{\Omega}}_{k-1} \widehat{\boldsymbol{\delta}}_{\widehat{S}_{k-1}})^2}{\widehat{\sigma}_{cc} - \widehat{\mathbf{\Sigma}}^T_{\widehat{S}_{k-1} c} \widehat{\mathbf{\Omega}}_{k-1} \widehat{\mathbf{\Sigma}}_{\widehat{S}_{k-1} c}}$ for all $c \in \widehat{C}_{k-1}$.
    Set $\widehat{s}_k = \operatorname{argmax}_{c \in \widehat{C}_{k-1}} \widehat{\theta}_{\widehat{S}_{k-1} c}$, $\widehat{S}_k = \widehat{S}_{k-1} \cup \{\widehat{s}_k\}$, $\widehat{C}_k = \widehat{C}_{k-1} \backslash \{\widehat{s}_k\}$,
        $\widehat{\boldsymbol{\rho}}_k = \widehat{\mathbf{\Omega}}_{k-1} \widehat{\mathbf{\Sigma}}_{\widehat{S}_{k-1} \widehat{s}_k}$, $\widehat{\alpha}_k = (\widehat{\sigma}_{\widehat{s}_k \widehat{s}_k} - \widehat{\mathbf{\Sigma}}^T_{\widehat{S}_{k-1} \widehat{s}_k} \widehat{\mathbf{\Omega}}_{k-1} \widehat{\mathbf{\Sigma}}_{\widehat{S}_{k-1} \widehat{s}_k})^{-1}$,
        $\widehat{\mathbf{\Sigma}}_{\widehat{S}_k \widehat{C}_k} = (\widehat{\mathbf{\Sigma}}_{\widehat{C}_k \widehat{S}_{k-1}} \;\; \widehat{\mathbf{\Sigma}}_{\widehat{C}_k \widehat{s}_k})^T$
        and $\widehat{\mathbf{\Omega}}_k = \begin{pmatrix} \widehat{\mathbf{\Omega}}_{k-1} + \widehat{\alpha}_k \widehat{\boldsymbol{\rho}}_k \widehat{\boldsymbol{\rho}}_k^T & -\widehat{\alpha}_k \widehat{\boldsymbol{\rho}}_k \\ -\widehat{\alpha}_k \widehat{\boldsymbol{\rho}}_k^T & \widehat{\alpha}_k \end{pmatrix}$.
Iterate until $\widehat{\theta}_{\widehat{S}_k c} < \tau$ for all $c \in \widehat{C}_k$, where $\tau$ is a predefined stopping
threshold.
```

converges asymptotically to the Bayes error.

We begin by introducing some notation. For a matrix $\boldsymbol{A}$, a set S, and an index c, we denote $\boldsymbol{A}_{Sc}$ as the cth column of $\boldsymbol{A}$ with row indices in S. Denote $\boldsymbol{A}_{SS}$ as the submatrix of $\boldsymbol{A}$ with row and column indices in S. Denote $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ as the minimum and maximum eigenvalues, respectively, of $\boldsymbol{A}$. For two sequences a_n and b_n, we write $a_n \lesssim b_n$ if $a_n \le C b_n$ for a generic positive constant C, write $a_n = o(b_n)$ if $a_n / b_n \to 0$, and $a_n \gg b_n$ if $b_n = o(a_n)$. To simplify the nonasymptotic statements, we assume throughout this paper that C_0 is an arbitrarily large positive constant and C_1 is some generic positive constant, which may vary from line to line. In addition, we assume $n_k / n \to \ell \in (0, 1)$ for $k = 0, 1$, and we assume normality in this section.

We first give a concentration result of the estimated increment $\widehat{\theta}_{Sc}$ for an arbitrary set S with s elements and an arbitrary $c \notin S$. We introduce the following regularity conditions.

Condition 1. $\max_{j \le p} |\delta_j| \le M < \infty$.

Condition 2. $0 < m \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le M < \infty$.

Condition 1 requires that the elements of $\boldsymbol{\delta}$ be bounded, and condition 2 requires that the eigenvalues of $\mathbf{\Sigma}$ be bounded away from zero and ∞; these two conditions are also used in other works (Shao et al. 2011; Cai and Liu 2011). These are mild boundedness conditions that simplify the nonasymptotic statement of the concentration results. Next, we give the concentration result of $\widehat{\theta}_{Sc} - \theta_{Sc}$.

Theorem 3.1. *Under Conditions 1–2 and if $s^2\sqrt{(\log p)/n} = o(1)$, it holds that*

$$P\left(|\widehat{\theta}_{Sc} - \theta_{Sc}| \lesssim s^2\sqrt{(\log p)/n}\max\{s^2\sqrt{(\log p)/n}, \sqrt{\theta_{Sc}}, \theta_{Sc}\}\right) \geq 1 - C_A p^{-C_B},$$

where C_A is a generic positive constant and C_B is an arbitrary large positive constant.

Theorem 3.1 shows that the concentration of $\widehat{\theta}_{Sc} - \theta_{Sc}$ depends on s and θ_{Sc}. It implies that when $\theta_{Sc} > 1$, $\widehat{\theta}_{Sc} - \theta_{Sc} = O_P(s^2\theta_{Sc}\sqrt{(\log p)/n})$, and when $s^4(\log p)/n < \theta_{Sc} < 1$, $\widehat{\theta}_{Sc} - \theta_{Sc} = O_P(s^2\sqrt{\theta_{Sc}(\log p)/n})$. When $\theta_{Sc} = 0$, it implies that $\widehat{\theta}_{Sc} - \theta_{Sc} = O_P(s^4(\log p)/n)$. These results show that it is more difficult to estimate a larger θ_{Sc}. In addition, when s gets larger, so does the estimation error, owing to the accumulation of estimation errors when estimating the unknown parameters, because when s gets larger, more parameters need to be estimated.

Theorem 3.1 is critical in studying the statistical properties of the GS-LDA rule. First, it indicates that as long as there is a large enough gap between θ_{Sc} and $\theta_{Sc'}$ for any $c' \neq c$, the indicator functions $I(\widehat{\theta}_{Sc} > \widehat{\theta}_{Sc'}) = I(\theta_{Sc} > \theta_{Sc'})$ hold with high probability. In other words, the order of $\widehat{\theta}_{Sc}$ and $\widehat{\theta}_{Sc'}$ reflects the true order of θ_{Sc} and $\theta_{Sc'}$. As shown below, this is the key to guaranteeing that the greedy search algorithm reaches variable selection consistency. Theorem 3.1 also gives guidance on how to choose the stopping threshold τ. When $\widehat{\theta}_{Sc}$ is small, it indicates that θ_{Sc} is small or equal to zero. At that stage, adding additional variables does not improve the classification, and we should thus terminate the greedy search. More details on how to choose τ are given in Theorem 3.2. Finally, we give a corollary for the special case of $S = \emptyset$. This result is useful for proving the property of the initial iteration of our algorithm. Its proof follows directly from that of Theorem 1.

Corollary 3.1. *Under the conditions of Theorem 3.1, when $S = \emptyset$, it holds that*

$$P\left(|\widehat{\theta}_{Sc} - \theta_{Sc}| \lesssim \sqrt{(\log p)/n}\right) \geq 1 - C_A p^{-C_B},$$

where C_A is a generic positive constant and C_B is an arbitrary large positive constant.

Next, we prove that, if $\boldsymbol{\beta}$ is exactly sparse, in the sense that many of its elements are zero, the greedy search method can recover the support of $\boldsymbol{\beta}$ with high probability. Let $\mathcal{M} = \{j : \beta_j \neq 0\}$ be the support of $\boldsymbol{\beta}$, K be the number of elements in $\mathcal{M}$, and $\widehat{\mathcal{M}}$ be as defined in Section 3.2. We have the following variable selection consistency result.

Theorem 3.2. *Under Conditions 1–2 and*

Condition 3. $0 < m \leq \min_{S \subset \mathcal{M}} \max_{c \in \mathcal{M} \backslash S} \theta_{Sc} \leq \max_{S \subset \mathcal{M}} \max_{c \in \mathcal{M} \backslash S} \theta_{Sc} \leq M < \infty$*;*

Condition 4. $\min_{S \subset \mathcal{M}} (\max_{c \in \mathcal{M} \backslash S} \theta_{Sc} - \max_{c \notin \mathcal{M}} \theta_{Sc}) \gg K^2 \sqrt{(\log p)/n}$*;*

if $K^2 \sqrt{(\log p)/n} = o(1)$*, and we choose* $\tau \asymp K^4 (\log p)/n$*, it holds that*

$$P\left(\widehat{\mathcal{M}} = \mathcal{M}\right) \geq 1 - C_A K p^{-C_B},$$

where C_A *is a generic positive constant and* C_B *is an arbitrary large positive constant.*

Condition 3 requires that for any $S \subset \mathcal{M}$, if we add another variable in $\mathcal{M}$ to those in S, the true increment θ_{Sc} should be bounded away from zero and infinity. The lower bound is mild, because, as shown in (3.2.3), θ_{Sc} is always nonnegative; because $\mathcal{M}$ contains all discriminative features, the lower bound requires only that at least one additional feature in $\mathcal{M}$ should produce a large enough increment of θ_{Sc} to pass the threshold. The upper bound is mainly introduced to simplify the expression. As shown in Theorem 3.1, the concentration of $\widehat{\theta}_{Sc}$ also depends on the magnitude of θ_{Sc}, requiring all θ_{Sc} to be bounded away from infinity. This enables us to have a more succinct nonasymptotic statement in Theorem 3.2. Condition 4 requires that for any $S \subset \mathcal{M}$, the maximum increment produced by adding another feature in $\mathcal{M}$ should surpass the increment by adding a feature outside of $\mathcal{M}$ when the true θ_{Sc} is known. Such a condition naturally requires that adding a discriminative feature in $\mathcal{M}$ should bring more information than adding a non-discriminative one outside $\mathcal{M}$. As shown in Theorem 3.1, the component $K^2 \sqrt{(\log p)/n}$ is the estimation error of $\widehat{\theta}_{Sc}$ to θ_{Sc}. Thus, once Condition 4 is assumed, with high probability, we have $\max_{c \in \mathcal{M} \backslash S} \widehat{\theta}_{Sc} > \max_{c \notin \mathcal{M}} \widehat{\theta}_{Sc}$. This is the key to ensuring that the greedy search algorithm chooses the informative features in $\mathcal{M}$. As reflected by the concentration result in Theorem 1, the choice of τ is essentially the order of $\widehat{\theta}_{Sc}$ when $\theta_{Sc} = 0$. This guarantees the exclusion of non-informative features. Finally, the assumption of $K^2 \sqrt{(\log p)/n} = o(1)$ is a sparsity assumption on $\boldsymbol{\beta}$, which is similar to the condition needed for the LPD and ROAD methods; see Cai and Liu (2011) and Fan et al. (2012).

Given the variable selection consistency, we next establish the error rate consistency of the GS-LDA rule. Without loss of generality, we assume that $\pi_0 = \pi_1 = 1/2$. By definition, the misclassification error of the GS-LDA rule is

$$
\begin{aligned}
R_{GS\text{-}LDA}(\boldsymbol{X}) &= (1/2)P\left(D_{GS\text{-}LDA}(\boldsymbol{x}) = 0 | \boldsymbol{x} \text{ comes from Class 1}\right) \\
&\quad + (1/2)P\left(D_{GS\text{-}LDA}(\boldsymbol{x}) = 1 | \boldsymbol{x} \text{ comes from Class 0}\right) \\
&= \frac{1}{2}\sum_{k=0}^{1}\Phi\left(\frac{(-1)^k\widehat{\boldsymbol{\beta}}^T_{\widehat{\mathcal{M}}}(\boldsymbol{\mu}_{k\widehat{\mathcal{M}}} - \bar{\boldsymbol{x}}_{k\widehat{\mathcal{M}}}) - \widehat{\boldsymbol{\delta}}^T_{\widehat{\mathcal{M}}}\widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{M}}}\widehat{\boldsymbol{\delta}}_{\widehat{\mathcal{M}}}/2}{\sqrt{\widehat{\boldsymbol{\delta}}^T_{\widehat{\mathcal{M}}}\widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{M}}}\boldsymbol{\Sigma}_{\widehat{\mathcal{M}}\widehat{\mathcal{M}}}\widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{M}}}\widehat{\boldsymbol{\delta}}_{\widehat{\mathcal{M}}}}}\right).
\end{aligned}
\tag{3.3.1}
$$

The following theorem establishes the error rate consistency.

Theorem 3.3. *Under Conditions 1–4, if $K^2\sqrt{(\log p)/n} = o(1)$ and we choose $\tau \asymp K^4(\log p)/n$, it holds that*

(a) $R_{GS\text{-}LDA}(\boldsymbol{X}) = \Phi(-\sqrt{\Delta_p}/2\{1 + O_P(K\sqrt{(\log p)/n})\})$

(b) $R_{GS\text{-}LDA}(\boldsymbol{X})/R_{Bayes} - 1 = O_P(\sqrt{(\log p)/n})$, when $\Delta_p < \infty$

(c) $R_{GS\text{-}LDA}(\boldsymbol{X})/R_{Bayes} - 1 = O_P(\max\{\Delta_p^{-1}, K^2\sqrt{(\log p)/n}\})$
when $\Delta_p \to \infty$.

Theorem 3 proves that the ratio of $R_{GS\text{-}LDA}(\boldsymbol{X})/R_{Bayes}$ converges to one in probability. Statement (a) shows that the convergence rate of $R_{GS\text{-}LDA}(\boldsymbol{X})$ to R_{Bayes} depends on K; for a larger K, the convergence is slower because more parameters need to be estimated. In statements (b) and (c), we show that the ratio of $R_{GS\text{-}LDA}(\boldsymbol{X})/R_{Bayes}$ converges to one in probability. Because Δ_p itself can diverge, R_{Bayes} can converge to zero. Thus, statements (b) and (c) are stronger than showing $R_{GS\text{-}LDA}(\boldsymbol{X}) - R_{Bayes} \to 0$ in probability. Our result indicates that $R_{GS\text{-}LDA}(\boldsymbol{X})$ can converge to zero as fast as R_{Bayes} does, even when $R_{Bayes} \to 0$. Furthermore, we show that the convergence rates differ depending on whether Δ_p is bounded. Finally, similarly to the LPD and ROAD, our method relies on the sparsity assumption on $\boldsymbol{\beta}$ to reach the error rate consistency, as shown in Theorem 3.3. This assumption is needed to avoid the accumulation of errors when estimating $\boldsymbol{\beta}$ that could ruin the error rate consistency (Bickel and Levina 2004). In addition, Theorem 3.3 relies on the variable selection consistency established in Theorem 3.2.

3.4 Relaxation of the Normality Assumption

Although we have proved the theoretical results under the normality assumption, these results can still hold when the two classes follow more general distributions. As discussed in Shao et al. (2011), the Bayes rule remains the same, as long as there exists a unit vector $\boldsymbol{\gamma}$ such that, for any real number t and $k = 0, 1$, it holds that

$$P(\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_k - \boldsymbol{\mu}_k) \leq t) = \Psi(t),$$

where $\Psi(t)$ is a cumulative distribution function with a density that is symmetric around zero and does not depend on $\boldsymbol{\gamma}$. In this case, the Bayes error is $\Psi(-\sqrt{\Delta_p}/2)$, which is still a decreasing function of the Mahalanobis distance Δ_p. Two key conditions for such a result are that the density function is symmetric around zero and the two classes have an equal covariance. Distributions satisfying this condition include the class of elliptical distributions with a density function of $c_p|\boldsymbol{\Sigma}|^{-1/2}f((\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))$, where f is a monotone function in $[0, \infty)$ and c_p is a normalization constant. Examples of elliptical distributions include the multivariate normal, t, and double exponential distributions. Given such a general distributional assumption, the increment of Mahalanobis distance still quantifies how much a new variable can reduce the Bayes error. Interestingly, for all elliptical distributions, θ_{Sc} still has the form

$$\theta_{Sc} = \frac{C\{\mathrm{E}(z_c|\boldsymbol{z}_S = \mathbf{0})\}^2}{\mathrm{Var}(z_c|\boldsymbol{z}_S = \mathbf{0})},$$

where the positive constant C depends on the type of the distribution, and is equal to two if it is normal. The proof uses the conditional distribution of the elliptical distributions, which is similar to what the multivariate normal distribution admits; see Theorem 2.18 of Fang et al. (2018). Thus, for all elliptical distributions, the contribution of a new variable to the classification depends on its effect size, conditional on other variables (i.e., the standardized $\mathrm{E}(z_c|\boldsymbol{z}_S = \mathbf{0})$).

However, under the more general distributional assumption, the convergence rates established in Theorems 3.1–3.3 may change. A closer look at the proofs reveals that the convergence rates depend on the tail probability, which is characterized by $\Psi(-x)$. The tail probability is the key to establishing the critical concentration results of $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\Sigma}}$, upon which the proofs are built; see

Lemma 1 in the Supplementary Material. In general, we assume that

$$0 < \lim_{x \to \infty} \frac{x^{-\omega} \exp(-cx^{\varphi})}{\Psi(-x)} < \infty, \tag{3.4.1}$$

where $\varphi \in [0, 2]$, $c \in (0, \infty)$, and $\omega \in (0, \infty)$ are some constants. In particular, when Ψ is standard normal, (3.4.1) holds with $\omega = 1$, $c = 1/2$, and $\varphi = 2$. Then, if Ψ satisfies (3.4.1) with $\varphi = 2$, the same exponential-type concentration can be established so that all results in Theorem 3.1–3.3 remain the same. When Ψ satisfies (3.4.1) with $\varphi < 2$, the tail of the distribution is heavier. In that case, if we assume the moment condition that

$$\max_{k,j \leq p} \mathrm{E}|x_{k,j}|^{2\nu} < \infty, \text{ for some } \nu > 0, \text{ and } k = 0, 1, \tag{3.4.2}$$

where $x_{k,j}$ is the jth element of $\boldsymbol{x}_k$, a polynomial-type concentration can be established so that Theorems 3.1–3.3 hold with all $(\log p)/n$ terms being replaced by $p^{4/\nu}/n$. In this case, p is only allowed to grow polynomially with n. These results are analogous to the discussions in Section 4 of Shao et al. (2011). To improve the convergence rates, one can replace $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\Sigma}}$ with robust estimators, such as the Huber estimator or the median-of-means estimator; see Avella-Medina et al. (2018). Correspondingly, the greedy search algorithm can be built upon these robust estimators. Once such robust estimators are used in the algorithm, even under the moment assumption in (3.4.2), Theorems 1–3 can still hold with the same exponential rate of convergence, using the concentration results established in Avella-Medina et al. (2018).

3.5 Simulation Studies

We investigate the numerical performance of our proposed method under four different scenarios. In the first two scenarios, we compare the classification performance and the execution time of the proposed GS-LDA method with those of other LDA-based methods. These include the sparse discriminant analysis by Clemmensen et al. (2011) (SDA), the SLDA, LPD, and ROAD, and some other well-known classifiers in machine learning, such as the support vector machine (SVM) with a linear kernel and the logistic regression with an L_1-penalty (Logistic-L1). In the other two scenarios, we investigate the performance of the GS-LDA method for some ultrahigh-dimensional settings that involve tens of thousands of features. In these two scenarios, most existing methods cannot handle such high dimensions, and we only aim at testing the viability of GS-LDA. We implement the SDA using the **sparseLDA** package. We implement the SLDA using our own coding of the algorithm given in Shao et al. (2011). We implement the LPD using the **linprogPD** function from the **clime** package (https://github.com/rluo/clime). The implementation for the ROAD comes from the publicly available package developed by the authors. We implement the SVM using the **e1071** package, and implement the Logistic-L1 using the **glmnet** package. The optimal tuning parameters for each method are chosen using a grid search with five-fold cross-validation. The execution time is recorded as the time taken by each algorithm on a computing cluster with an Intel Xeon 3.4GHz CPU, with the tuning parameters fixed at their optimal values.

In the first two scenarios, we consider the following two choices of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$:

Scenario 1: $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\mu}_{\mathbf{1}} = (1, ..., 1, 0, ..., 0)^T$, where the first 10 elements are ones, and the rest are zeros. $\boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$, where $\sigma_{ij} = 0.8^{|i-j|}$, for $1 \leq i, j \leq p$.

Scenario 2: $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\mu}_{\mathbf{1}} = \boldsymbol{\Sigma}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (0.25, ..., 0.25, 0, ..., 0)^T$, with the first 10 elements of $\boldsymbol{\beta}$ equal to 0.25 and the rest zeros; $\boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$, where $\sigma_{ij} = 0.8^{|i-j|}$, for $1 \leq i, j \leq p$.

For each scenario, we generate 200 training samples for each of the two classes from $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$; we let the dimension p vary from 500 to 2000, with an increment of 500. We independently generate another 800 samples from each of these distributions as the test set. In Scenario 1, we set $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ to be exactly sparse. This scenario is the same as Model 3 considered in Cai and Liu (2011). In Scenario 2, we set $\boldsymbol{\beta}$ to be exactly sparse because such a condition is imposed for the proposed GS-LDA method. We use five-fold cross-validation to choose the optimal tuning parameters in all seven methods. For each scenario, we run 100 replicates. We report the average misclassification rates and the execution time for each classifier in Figures 3.1 and 3.2. For Scenario 2, we also report the variable selection performance on $\boldsymbol{\beta}$ by the GS-LDA, ROAD, and Logistic-L1. We measure the variable selection performance by sensitivity and specificity. Sensitivity is defined as the proportion of nonzero elements of $\boldsymbol{\beta}$ that are estimated as nonzero, and specificity is defined as the proportion of zero elements of $\boldsymbol{\beta}$ that are estimated as zero.

Figures 3.1 and 3.2 show that the GS-LDA has the best classification performance for all choices of p and under both scenarios. Its computational speed is also much faster than that of the other LDA-based classifiers, especially when the dimension p is high. It is also faster than the SVM and Logistic-L1, implemented by the **e1071** and **glmnet** packages, respectively, which are known to be computationally efficient. In both scenarios, when $p = 2000$, the average execution time is around 190 seconds for the SDA, 100 seconds for the ROAD, 20 seconds for the LPD, 3 seconds for the SVM, and 1 second for the SLDA, but only 0.05 second for the GS-LDA. The execution time for the Logistic-L1 depends on the Hessian matrix of the likelihood and the sparsity of $\boldsymbol{\beta}$. In Scenario 1, when $\boldsymbol{\beta}$ is weakly sparse, the GS-LDA is still faster than the Logistic-L1. In Scenario 2, when $\boldsymbol{\beta}$ is exactly sparse, their computational time is comparable. This is mainly because the **glmnet** package only updates nonzero components of $\boldsymbol{\beta}$ along its iterations. The proposed GS-LDA method thus offers a substantial boost in computational speed over most its competitors, while rendering an excellent classification rule. In terms of variable selection performance, in Scenario 2, the GS-LDA has similar specificity to that of the ROAD and Logistic-L1, and better sensitivity than the ROAD. However, owing to its lower sensitivity than the Logistic-L1, the GS-LDA has slightly higher misclassification error than that of the Logstic-L1 in this secnario. Finally, note that because the GS-LDA adds one variable at a time, the variation of its errors can be smaller than that of other optimization based methods, where the numbers of variables are determined by some tuning parameters, and small changes can result in multiple new variables being included in the classification rule. This can be seen from Figures 3.1 and 3.2.

To further investigate how many dimensions the GS-LDA method can efficiently handle, we simulate two additional scenarios, where we choose $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ as follows:
Scenario 3: $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\mu}_{\mathbf{1}} = (1, ..., 1, 0, ..., 0)^T$, where the first 10 elements are ones and the rest are zeros; $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, where $\boldsymbol{\Omega} = (\omega_{ij})_{p\times p}$ and $\omega_{ij} = \sqrt{ij}\{2I(i = j \neq p) + I(i = j = p) - I(|i - j| = 1)\}$.
Scenario 4: $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\mu}_{\mathbf{1}} = \boldsymbol{\Sigma\beta}$, where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$, with the first 10 elements of $\boldsymbol{\beta}$ being ones and the rest being zeros; $\boldsymbol{\Sigma}$ is the same as in Scenario 3.
For each of these two scenarios, we generate 100 training samples for each of the two classes from $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ and set the dimension $p = 1 \times 10^4, 3 \times 10^4, 5 \times 10^4$, and 1×10^5. We independently generate another 400 samples for each of the two classes as the test set. Scenarios 3 and 4 are analogues to scenarios 1 and 2: $\boldsymbol{\delta}$ is exactly sparse in Scenario 3, and $\boldsymbol{\beta}$ is exactly sparse in Scenario 4. Again, we use five-fold cross-validation to choose the optimal stopping threshold for the GS-LDA. For each scenario, we run 100 replicates. We report the average misclassification rates and the execution time for the GS-LDA in Figures 3 and 4.

Figures 3.3 and 3.4 show that in both scenarios the GS-LDA method still performs well when the

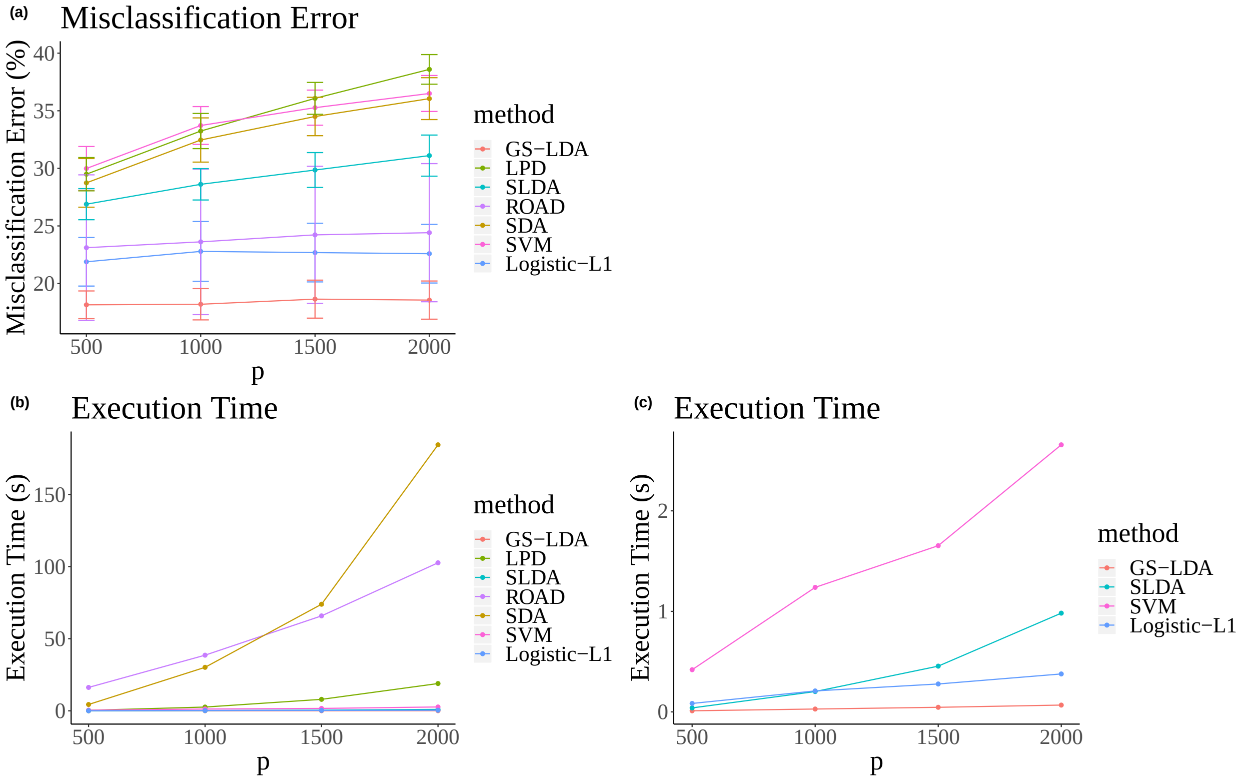


Figure 3.1. Numerical performance of the seven classifiers in Scenario 1. Panel (c) is a zoomed plot of panel (b) for the GS-LDA , SLDA, SVM, and Logistic-L1.

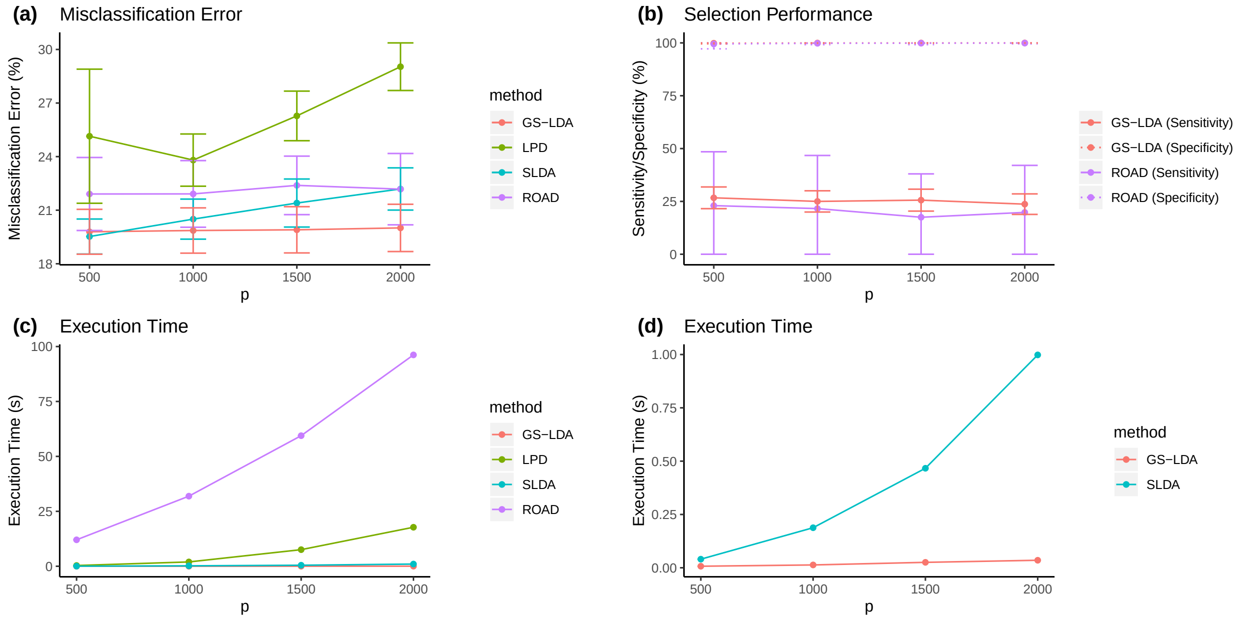


Figure 3.2. Numerical performance of the seven classifiers in Scenario 2. Panel (d) is a zoomed plot of panel (c) for the GS-LDA , SLDA, SVM, and Logistic-L1.

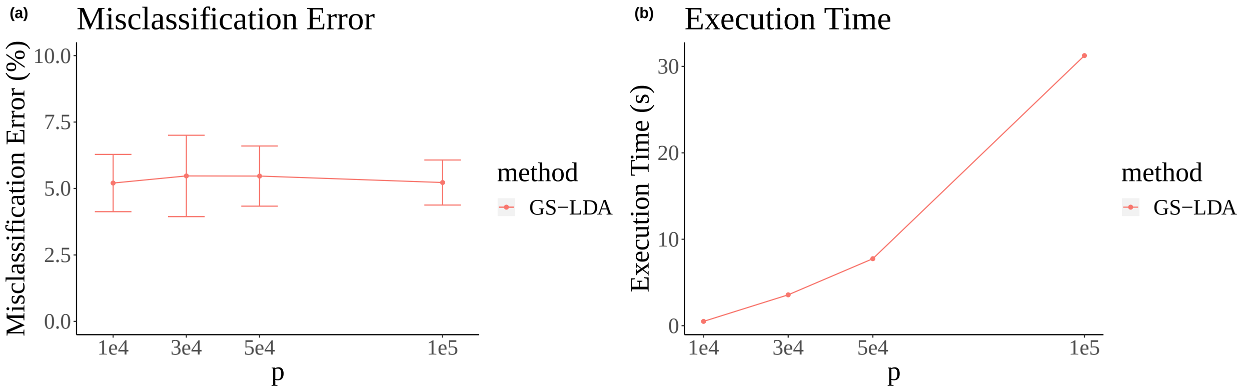


Figure 3.3. Numerical performance of the GS-LDA in Scenario 3.

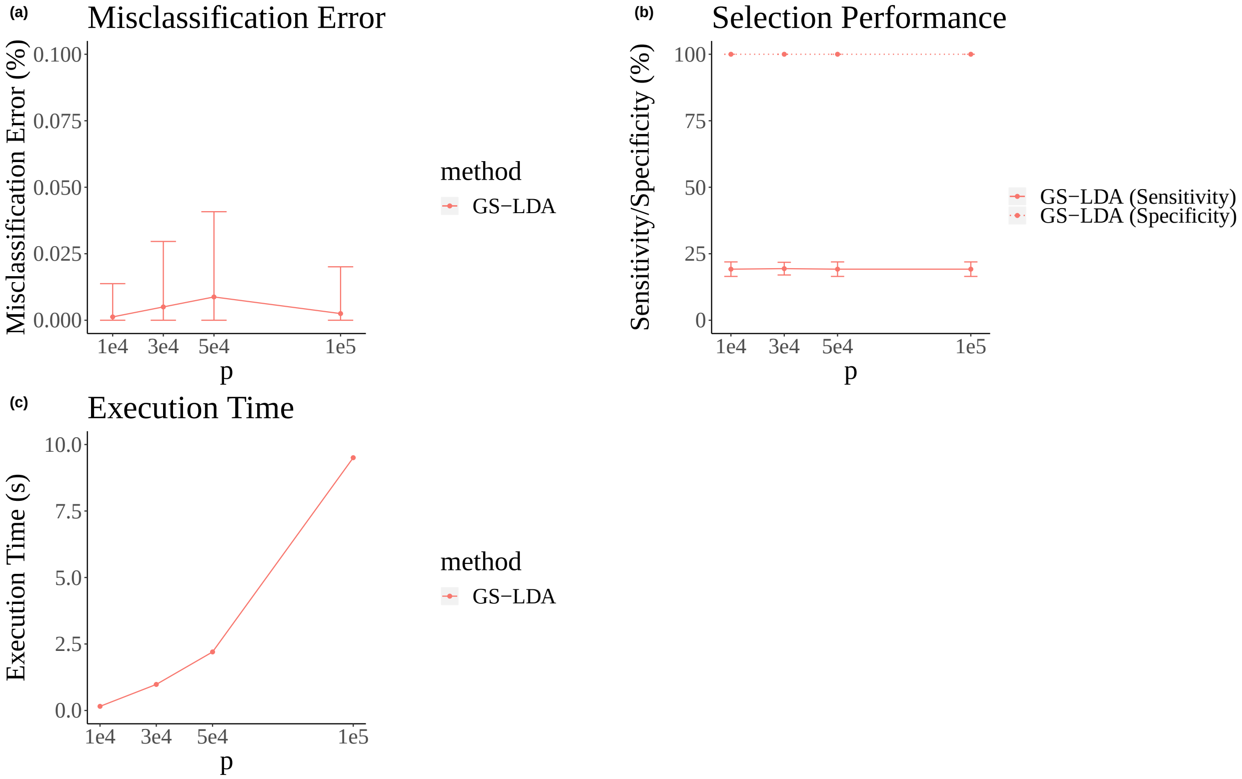


Figure 3.4. Numerical performance of the GS-LDA in Scenario 4.

dimension is ultrahigh. When the dimension p grows, the misclassification error remains stable, and the execution time grows only moderately with p. Even when p is as big as 1×10^5, the execution time of the GS-LDA is only tens of seconds. In contrast, the SLDA, LPD, and ROAD cannot solve such a problem within tens of hours. As a result, they are excluded from the comparison in these two scenarios.

3.6 An Application to Cancer Subtype Classification

To further illustrate the advantage of the GS-LDA method, we apply it to microarray data for classifying cancer subtypes. This data set contains 82 breast cancer subjects, with 41 ER-positive and 41 ER-negative. These subjects are sequenced using the Affymetrix Human Genome U133A Array, which measures the gene expression using 22283 probes. The raw data are available in the Gene Expression Omnibus database with the accession name GSE22093.

We randomly split the data set into a training set of 60 samples and a test set of 22 samples, repeating the random split 100 times. Each time, we learn the GS-LDA, SLDA, ROAD, SVM, and Logistic-L1 from the training set, and obtain their misclassification errors by applying them to the test set. The LPD and SDA methods are excluded from this study because they cannot finish the training within 24 hours. The tuning parameters in these methods are chosen using five-fold cross-validation. The misclassification errors and the execution times of these methods are summarized in Table 3.1.

The GS-LDA method performs well for this data set, with a mean misclassification error of only 2.4%. On average, this is less than one error among the 22 samples in a testing set. This misclassification error is 47% better than that of the SVM, 56% better than that of the ROAD, 72% better than that of the SLDA, and equals to that of the Logistic-L1. In term of computational speed, the GS-LDA runs for only 0.3 seconds on average, which is over 5000 times faster than the ROAD, over 1000 times faster than the SLDA, over 10 times faster than the SVM, and close to that of the Logistic-L1 on this data set.

Interestingly, we found that in the 100 splits of the data set, the GS-LDA method frequently selected one particular variable: the expression of the ESR1 gene measured by probe “205225_at.” This variable was selected 95 times by the GS-LDA. It was selected first 84 times, and was the only variable selected 56 times. **?** defined the ER status by whether the subject’s measured ESR1 expression using probe “205225_at” was higher than 10.18. In other words, the true decision rule is

Methods	Misclassification Error (%)			Execution Time (s)
	Lower Quartile	Median	Upper Quartile	
GS-LDA	0	0	4.55	0.30
SLDA	4.55	9.09	13.64	355
ROAD	4.55	4.55	9.09	1576
SVM	0	4.55	5.68	3.08
Logistic-L1	0	0	4.55	0.10

Table 3.1. Numerical performance of the five classifiers in classifying cancer subtypes.

$D_{true}(\boldsymbol{x}) = I(x_{205225\_at} > 10.18)$. We compare the selection frequency of the GS-LDA, ROAD and Logistic-L1 over the 100 random splits; see Figure S1 in the Supplementary Material. Here, we find that the GS-LDA selects the true "205225_at" probe much more often with fewer false positives than the other two methods.

To further illustrate the merit of the GS-LDA method, we use all subjects, including both the positive and the negative groups, for training; the resulting GS-LDA rule is $D_{GS\text{-}LDA}(\boldsymbol{x}) = I(x_{205225\_at} > 10.50)$, which is very close to how the ER status is defined in the original study. When we train the ROAD, SLDA, and Logistic-L1 rules using the full data, we find that they also include probe "205225_at," but the ROAD includes another 15 probes, the Logistic-L1 includes another 28 probes, and the SLDA includes all probes. It is obvious that the GS-LDA gives a rule that is much closer to the truth.

3.7 Discussion

We have developed an efficient greedy search algorithm for performing an LDA with high-dimensional data. Motivated by the monotonicity property of the Mahalanobis distance, which characterizes the Bayes error of the LDA problem, our algorithm sequentially selects the features that produce the largest increments of Mahalanobis distance. In other words, it sequentially selects the most informative features until no additional feature can bring enough extra information to improve the classification. Our algorithm is computationally much more efficient than existing optimization-based or thresholding methods, because it does not need to solve an optimization problem or invert a large matrix. Indeed, it does not even need to compute the whole covariance matrix in advance; rather, it computes matrix elements as it goes in order to update the classification rule. All calculations are based on some closed-form formulae. We proved that such an algorithm results in a GS-LDA rule that is both variable selection and error rate consistent, under a mild distributional assumption.

In practice, our method can also be modified to yield a nonlinear classification boundary by using nonlinear kernels or Gaussian copulas (Han et al. 2013). Our method may also be extended to a multicategory discriminant analysis. Using similar ideas to those in Pan et al. (2016) and Mai et al. (2019), we can translate a multicategory problem into multiple binary classification problems, to which our method is applicable. Finally, note that our method requires the key assumption that the two classes have the same covariance. If this is not the case, it becomes a quadratic discriminant analysis (QDA) problem. The Bayes error of such a problem has a much more completed form (Li and Shao 2015). We leave developing an efficient algorithm to solve the high-dimensional QDA problem as a topic for future research.

3.8 Technical Details

3.8.1 Proofs

Proof of Theorem 3.1. It follows from Lemmas 3.1 and 3.4 that

$$P\left(|(\widehat{\sigma}_{cc} - \widehat{\mathbf{\Sigma}}_{Sc}^T \widehat{\mathbf{\Sigma}}_{SS}^{-1} \widehat{\mathbf{\Sigma}}_{Sc}) - (\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc})| \lesssim s^2 \sqrt{(\log p)/n}\right) \geq 1 - C_A p^{-C_B},$$

where C_A only depends on C_1 and C_4 in Lemmas 3.1 and 3.4, and C_B is an arbitrarily large constant. Since $\mathbf{\Sigma}_{S\cup\{c\},S\cup\{c\}}$ is a submatrix of $\mathbf{\Sigma}$ with row and column indices in $S \cup \{c\}$ and is positive definite, it follows from Condition 2 and Theorem 4.3.17 of Horn and Johnson (2012) that for any $c \notin S$,

$$0 < m \leq \lambda_{\min}(\mathbf{\Sigma}_{S\cup\{c\},S\cup\{c\}}) \leq \lambda_{\max}(\mathbf{\Sigma}_{S\cup\{c\},S\cup\{c\}}) \leq M < \infty.$$

Since $\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc}$ is the Schur complement of $\mathbf{\Sigma}_S$ in $\mathbf{\Sigma}_{S\cup\{c\},S\cup\{c\}}$, it follows that $\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc} \geq m > 0$ for all $c \notin S$. Then we have

$$P\left(|(\widehat{\sigma}_{cc} - \widehat{\mathbf{\Sigma}}_{Sc}^T \widehat{\mathbf{\Sigma}}_{SS}^{-1} \widehat{\mathbf{\Sigma}}_{Sc})^{-1} - (\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc})^{-1}| \lesssim s^2 \sqrt{(\log p)/n}\right) \geq 1 - C_A p^{-C_B}, \quad (3.8.1)$$

where C_A only depends on C_1 and C_4, and C_B is an arbitrarily large constant. On the other hand, with probability at least $1 - C_A p^{-C_B}$, we have

$$\begin{aligned}
&|(\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S)^2 - (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)^2| \\
&\leq |(\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S) - (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_{SS})|^2 \\
&\quad + 2|(\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S) - (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)| \cdot |\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S| \qquad (3.8.2) \\
&\lesssim (s^2 \sqrt{(\log p)/n})^2 + (s^2 \sqrt{(\log p)/n}) \cdot |\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S| \\
&\lesssim (s^2 \sqrt{(\log p)/n}) \cdot \max(s^2 \sqrt{(\log p)/n}, \sqrt{\theta_{Sc}}),
\end{aligned}$$

where the last inequality follows from Condition 2.

Therefore, (3.8.1) and (3.8.2) together imply that, with probability at least $1 - C_A p^{-C_B}$, we have

$$\begin{aligned}
&|\widehat{\theta}_{Sc} - \theta_{Sc}| \\
&= |(\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S)^2 (\widehat{\sigma}_{cc} - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\Sigma}}_{Sc})^{-1} - (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)^2 (\sigma_{cc} - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{Sc})^{-1}| \\
&\leq |(\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S)^2 - (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)^2| \cdot |(\widehat{\sigma}_{cc} - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\Sigma}}_{Sc})^{-1} - (\sigma_{cc} - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{Sc})^{-1}| \\
&\quad + |(\widehat{\delta}_c - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S)^2 - (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)^2| (\sigma_{cc} - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{Sc})^{-1} \\
&\quad + |(\widehat{\sigma}_{cc} - \widehat{\boldsymbol{\Sigma}}_{Sc}^T \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \widehat{\boldsymbol{\Sigma}}_{Sc})^{-1} - (\sigma_{cc} - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{Sc})^{-1}| (\delta_c - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)^2 \\
&\lesssim s^4 (\log p)/n \max(s^2 \sqrt{(\log p)/n}, \sqrt{\theta_{Sc}}) + s^2 \sqrt{(\log p)/n} \max(s^2 \sqrt{(\log p/n)}, \sqrt{\theta_{Sc}}) \\
&\quad + s^2 \sqrt{(\log p)/n} \theta_{Sc} \\
&\lesssim s^2 \sqrt{(\log p)/n} \max(s^2 \sqrt{(\log p)/n}, \sqrt{\theta_{Sc}}, \theta_{Sc}).
\end{aligned}$$

□

Proof of Theorem 3.2. Let $\emptyset = \widehat{S}_0 \subset \widehat{S}_1 \subset \cdots$ be the sequence of selected indices given by the greedy search algorithm. The key of the proof is to show that, with high probability, $\widehat{S}_k \subset \mathcal{M}$ for all $k \leq K - 1$, and $\widehat{\mathcal{M}} = \widehat{S}_K = \mathcal{M}$.

When $k = 0$, it follows from Corollary 3.1 and the union bound that

$$P\left(\max_{c \leq p} |\widehat{\theta}_{Sc} - \theta_{Sc}| \lesssim \sqrt{(\log p)/n}\right) \geq 1 - C_A p^{-C_B}, \text{ for } S = \emptyset.$$

Condition 4 implies that $\max_{c \in \mathcal{M}} \theta_{Sc} - \max_{c \notin \mathcal{M}} \theta_{Sc} \gg K^2\sqrt{(\log p)/n} \geq \sqrt{(\log p)/n}$. These two results together imply that

$$P\left(\max_{c \in \mathcal{M}} \widehat{\theta}_{Sc} > \max_{c \notin \mathcal{M}} \widehat{\theta}_{Sc}\right) \geq 1 - C_A p^{-C_B}, \text{ for } S = \emptyset.$$

It further implies that $P\left(\widehat{S}_1 \subset \mathcal{M}\right) \geq 1 - C_A p^{-C_B}$.

When $k = 1$, we prove that

$$P\left(\max_{c \in \mathcal{M} \setminus \widehat{S}_1} \widehat{\theta}_{\widehat{S}_1 c} > \max_{c \notin \mathcal{M}} \widehat{\theta}_{\widehat{S}_1 c}\right) \geq 1 - C_A p^{-C_B}. \tag{3.8.3}$$

This further gives $P\left(\widehat{S}_2 \subset \mathcal{M}\right) \geq 1 - C_A p^{-C_B}$, where C_A is treated as a generic postic constant. Let events

$$\begin{aligned}
E_1 &= \left\{\widehat{S}_1 \subset \mathcal{M}\right\}, \\
A_1 &= \left\{\max_{c \in \mathcal{M} \setminus \widehat{S}_1} \theta_{\widehat{S}_1 c} - \max_{c \notin \mathcal{M}} \theta_{\widehat{S}_1 c} \gg K^2\sqrt{(\log p)/n}\right\}, \\
A_2 &= \left\{\max_{c \in \mathcal{M} \setminus \widehat{S}_1} |\widehat{\theta}_{\widehat{S}_{1c}} - \theta_{\widehat{S}_1 c}| \lesssim K^2\sqrt{(\log p)/n}\right\}, \\
A_3 &= \left\{\max_{c \notin \mathcal{M}} |\widehat{\theta}_{\widehat{S}_{1c}} - \theta_{\widehat{S}_1 c}| \lesssim K^2\sqrt{(\log p)/n}\right\}.
\end{aligned}$$

Note that $A_1 \cap A_2 \cap A_3 \subset \left\{\max_{c \in \mathcal{M} \setminus \widehat{S}_1} \widehat{\theta}_{\widehat{S}_1 c} > \max_{c \notin \mathcal{M}} \widehat{\theta}_{\widehat{S}_1 c}\right\}$. Therefore,

$$P\left(\max_{c \in \mathcal{M} \setminus \widehat{S}_1} \widehat{\theta}_{\widehat{S}_1 c} > \max_{c \notin \mathcal{M}} \widehat{\theta}_{\widehat{S}_1 c}\right) \geq 1 - P\left(\overline{A_1}\right) - P\left(\overline{A_2}\right) - P\left(\overline{A_3}\right). \tag{3.8.4}$$

Under Condition 4, $E_1 \subset A_1$, therefore, $P\left(\overline{A_1}\right) \leq P\left(\overline{E_1}\right) \leq C_A p^{-C_B}$. It follows from Theorem 3.1, Condition 3, and the union bound that $P(\overline{A_2}) \leq C_A p^{-C_B}$, and $P(\overline{A_3}) \leq C_1 p^{-C_B}$. These three results, together with (3.8.4), proves (3.8.3). By the same argument, it holds that $\widehat{S}_k \subset \mathcal{M}$ for all $k \leq K$ with probability at least $1 - (2k-1)C_A p^{-C_B}$. Since $\mathcal{M}$ contains K elements, we further have $\widehat{S}_K = \mathcal{M}$.

Next, we show that at the $(K+1)$th iteration, the greedy search algorithm terminates with high probability if we choose $\tau \asymp K^4(\log p)/n$. First, we show that $\theta_{\mathcal{M}c} = 0$ for all $c \notin M$. By definition,

$\theta_{\mathcal{M}c} = \Delta_{\mathcal{M}\cup\{c\}} - \Delta_{\mathcal{M}} = \boldsymbol{\beta}_{\mathcal{M}\cup\{c\}}^T \boldsymbol{\Sigma}_{\mathcal{M}\cup\{c\},\mathcal{M}\cup\{c\}} \boldsymbol{\beta}_{\mathcal{M}\cup\{c\}} - \boldsymbol{\beta}_{\mathcal{M}}^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}} \boldsymbol{\beta}_{\mathcal{M}} = 0$. Then, Theorem 3.1 implies that

$$P\left(\max_{c\notin\mathcal{M}} |\widehat{\theta}_{\mathcal{M}c}| \leq K^4(\log p)/n\right) \geq 1 - C_A p^{-C_B}.$$

Hence, by choosing $\tau \asymp K^4(\log p)/n$, the greedy search program terminates with high probability, i.e., $P(\widehat{\mathcal{M}} = \widehat{S}_K | \widehat{S}_K = \mathcal{M}) \geq 1 - C_A p^{-C_B}$. Then,

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}\right) &= P\left(\widehat{\mathcal{M}} = \widehat{S}_K, \widehat{S}_K = M\right) = P\left(\widehat{\mathcal{M}} = \widehat{S}_K | \widehat{S}_K = \mathcal{M}\right) P\left(\widehat{S}_K = \mathcal{M}\right) \\ &\geq (1 - C_A p^{-C_B})(1 - (2K-1)C_A p^{-C_B}) \geq 1 - C_A K p^{-C_B}. \end{aligned}$$

□

Proof of Theorem 3.3. We prove the result conditioning on the event that $\{\widehat{\mathcal{M}} = \mathcal{M}\}$, which holds with probability tending to 1. We first bound $\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T \widehat{\boldsymbol{\Omega}}_{\mathcal{M}} \widehat{\boldsymbol{\delta}}_{\mathcal{M}}$. By Lemma 3.3, we have

$$\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \widehat{\boldsymbol{\delta}}_{\mathcal{M}} - \boldsymbol{\delta}_{\mathcal{M}}^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \boldsymbol{\delta}_{\mathcal{M}} = O_P\left(K\sqrt{(\log p)/n}\right).$$

By Condition 3, $K \lesssim \Delta_p = \boldsymbol{\delta}_{\mathcal{M}}^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \boldsymbol{\delta}_{\mathcal{M}}$. Therefore,

$$\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \widehat{\boldsymbol{\delta}}_{\mathcal{M}} - \boldsymbol{\delta}_{\mathcal{M}}^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \boldsymbol{\delta}_{\mathcal{M}} = O_P\left(\Delta_p\sqrt{(\log p)/n}\right). \tag{3.8.5}$$

Then, by Lemma 3.4 we have

$$|\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T (\widehat{\boldsymbol{\Omega}}_{\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}) \widehat{\boldsymbol{\delta}}_{\mathcal{M}}| = O_P\left(\Delta_p K\sqrt{(\log p)/n}\right). \tag{3.8.6}$$

It follows from the triangular inequality and (3.8.5) and (3.8.6) that

$$\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T \widehat{\boldsymbol{\Omega}}_{\mathcal{M}} \widehat{\boldsymbol{\delta}}_{\mathcal{M}} = \Delta_p \left\{1 + O_P(K\sqrt{(\log p)/n})\right\}. \tag{3.8.7}$$

Next, we bound $\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T \widehat{\boldsymbol{\Omega}}_{\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}} \widehat{\boldsymbol{\Omega}}_{\mathcal{M}} \widehat{\boldsymbol{\delta}}_{\mathcal{M}}$. It follows from Lemma 3.2 that $\|\widehat{\boldsymbol{\Omega}}_{\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\| = O_P\left(K\sqrt{(\log p)/n}\right)$. This result, together with Condition 2, imply that $\|\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\| =$

$O_P(1)$. Then, using the same argument as in the proof of Lemma 3.4, we have

$$\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T(\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}} - \widehat{\boldsymbol{\Omega}}_{\mathcal{M}})\widehat{\boldsymbol{\delta}}_{\mathcal{M}} = O_P\left(\Delta_p K\sqrt{(\log p)/n}\right).$$

This result, together with (3.8.7), gives

$$\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}} = \Delta_p\left\{1 + O_P(K\sqrt{(\log p)/n})\right\}. \tag{3.8.8}$$

Then, we have

$$\begin{aligned}\frac{\widehat{\boldsymbol{\beta}}_{\mathcal{M}}^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}})}{\sqrt{\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}}} &= \frac{(\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}})^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}})}{\sqrt{\Delta_p\{1 + O_P(K\sqrt{(\log p)/n})\}}} \\ &\quad + \frac{\boldsymbol{\beta}_{\mathcal{M}}^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}})}{\sqrt{\Delta_p\{1 + O_P(K\sqrt{(\log p)/n})\}}}.\end{aligned}$$

Since the leading term $\Delta_p^{-1/2}\boldsymbol{\beta}_{\mathcal{M}}^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}}) \sim N(0, 1/n_1)$, we have

$$\frac{\boldsymbol{\beta}_{\mathcal{M}}^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}})}{\sqrt{\Delta_p\{1 + O_P(K\sqrt{(\log p)/n})\}}} = \frac{O_P(1/\sqrt{n})}{\sqrt{1 + O_P(K\sqrt{(\log p)/n})}}.$$

Since $K\sqrt{(\log p)/n} \le K^2\sqrt{(\log p)/n} = o(1)$, the leading term can be simplified as

$$\begin{aligned}\frac{\boldsymbol{\beta}_{\mathcal{M}}^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}})}{\sqrt{\Delta_p\{1 + O_P(K\sqrt{(\log p)/n})\}}} &= O_P(1/\sqrt{n})(1 + O_P(K\sqrt{(\log p)/n})) \\ &= O_P(1/\sqrt{n}) + O_P(K\sqrt{\log p}/n).\end{aligned}$$

Since $1/\sqrt{n} = o(\sqrt{K/n})$ and $K\sqrt{\log p}/n = o(\sqrt{K/n})$, we have

$$\frac{\widehat{\boldsymbol{\beta}}_{\mathcal{M}}^T(\bar{\boldsymbol{x}}_{1\mathcal{M}} - \boldsymbol{\mu}_{1\mathcal{M}})}{\sqrt{\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^T\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}}} = O_P\left(\sqrt{K/n}\right). \tag{3.8.9}$$

Then, it follows from (3.8.7), (3.8.8), and (3.8.9) that

$$
\begin{aligned}
&\frac{-\widehat{\boldsymbol{\beta}}_{\mathcal{M}}^{T}(\boldsymbol{\mu}_{1\mathcal{M}}-\bar{\boldsymbol{x}}_{1\mathcal{M}})-\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}/2}{\sqrt{\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}\mathcal{M}}}}\\
&=\frac{-\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}/2}{\sqrt{\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}}}-\frac{\widehat{\boldsymbol{\beta}}_{\mathcal{M}}^{T}(\boldsymbol{\mu}_{1\mathcal{M}}-\bar{\boldsymbol{x}}_{1\mathcal{M}})}{\sqrt{\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}}}\\
&=\frac{-\Delta_p\left(1+O_P(K\sqrt{(\log p)/n})\right)}{2\sqrt{\Delta_p\left(1+O_P(K\sqrt{(\log p)/n})\right)}}+O_P(\sqrt{K/n}) \qquad (3.8.10)\\
&=-\frac{\sqrt{\Delta_p}\left(1+O_P(K\sqrt{(\log p)/n})\right)}{2}+O_P(\sqrt{K/n})\\
&=-\frac{\sqrt{\Delta_p}\left(1+O_P(K\sqrt{(\log p)/n})\right)}{2},
\end{aligned}
$$

where in the second-to-last equation, we use the fact that $\{1+O_P(K\sqrt{(\log p)/n})\}^{-1/2}=1+O_P(K\sqrt{(\log p)/n})$, and in the last equation, we use $\sqrt{K/n}=o(K\{\Delta_p(\log p)/n\}^{1/2})$. Using the same argument, we also have

$$
\frac{\widehat{\boldsymbol{\beta}}_{\mathcal{M}}^{T}(\boldsymbol{\mu}_{0\mathcal{M}}-\bar{\boldsymbol{x}}_{0\mathcal{M}})-\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}/2}{\sqrt{\widehat{\boldsymbol{\delta}}_{\mathcal{M}}^{T}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}\widehat{\boldsymbol{\Omega}}_{\mathcal{M}}\widehat{\boldsymbol{\delta}}_{\mathcal{M}}}}=-\frac{\sqrt{\Delta_p}\left(1+O_P(K\sqrt{(\log p)/n})\right)}{2}. \qquad (3.8.11)
$$

Equations (3.8.10) and (3.8.11) together prove statement (a).

To prove (b), we use the fact that $R_{Bayes}=\Phi(-\sqrt{\Delta_p}/2)$ and a well-known result of the normal cumulative distribution function (Shao et al. 2011): that

$$
\frac{x}{1+x^2}e^{-x^2/2}\leq\Phi(-x)\leq\frac{1}{x}e^{-x^2/2},\ \ \text{for all } x>0. \qquad (3.8.12)
$$

First, when $\Delta_p<\infty$, by the Mean Value Theorem, we have

$$
R_{GS\text{-}LDA}(\boldsymbol{X})=R_{Bayes}+\phi(\widetilde{x})O_P\left(K\sqrt{\Delta_p(\log p)/n}\right)=R_{Bayes}+\phi(\widetilde{x})O_p(\sqrt{(\log p)/n}),
$$

where $\widetilde{x}$ is a number between $-\sqrt{\Delta_p}/2$ and $-\sqrt{\Delta_p}(1+O_p(K\sqrt{(\log p)/n}))/2$. In the last equation, we use the fact that $K\asymp\Delta_p<\infty$, which is implied by Conditions 1, 2 and 4, since $\Delta_p<\infty$, R_{Bayes}

is bounded away from 0. Then, we have

$$\frac{R_{GS\text{-}LDA}(\boldsymbol{X})}{R_{Bayes}} = 1 + \frac{\phi(\widetilde{x})}{R_{Bayes}} O_p(\sqrt{(\log p)/n}).$$

Then, the boundedness of the normal density function and R_{Bayes} implies that

$$\frac{R_{GS\text{-}LDA}(\boldsymbol{X})}{R_{Bayes}} - 1 = O_p(\sqrt{(\log p)/n}).$$

This proves statement (b).

When $\Delta_p \to \infty$, let $a_n = K\sqrt{(\log p)/n}$. Noting that $a_n = o(K^2\sqrt{(\log p)/n}) = o(1)$, it follows from statement (a) and (3.8.12) that

$$\begin{aligned}
\frac{R_{GS\text{-}LDA}(\boldsymbol{X})}{R_{Bayes}} &\leq \frac{\frac{1}{\sqrt{\Delta_p}/2(1+O_p(a_n))} e^{-(\frac{\sqrt{\Delta_p}}{2}(1+O_p(a_n)))^2/2}}{\frac{\sqrt{\Delta_p}/2}{1+(\sqrt{\Delta_p}/2)^2} e^{-(\frac{\sqrt{\Delta_p}}{2})^2/2}} \\
&\leq \frac{4+\Delta_p}{\Delta_p\{1+O_p(a_n)\}} e^{-\frac{\Delta_p}{8}(1-(1+O_p(a_n))^2)} \\
&\leq \frac{4+\Delta_p}{\Delta_p\{1+O_p(a_n)\}} e^{O_p(\Delta_p a_n)}.
\end{aligned}$$

Since $\Delta_p a_n \lesssim K^2\sqrt{(\log p)/n} = o(1)$, by the Taylor expansion, we have

$$\begin{aligned}
\frac{R_{GS\text{-}LDA}(\boldsymbol{X})}{R_{Bayes}} &\leq \frac{4+\Delta_p}{\Delta_p}(1+O_P(a_n))(1+O_P(\Delta_p a_n)) \leq \frac{4+\Delta_p}{\Delta_p}(1+O_P(\Delta_p a_n)) \\
&= (1+\frac{4}{\Delta_p})(1+O_P(\Delta_p a_n)) \leq 1 + O_P\left(\Delta_p^{-1}\right) + O_P\left(\Delta_p a_n\right).
\end{aligned}$$

Using a similar argument, we can show that

$$\begin{aligned}
\frac{R_{GS\text{-}LDA}(\boldsymbol{X})}{R_{Bayes}} &\geq \frac{\Delta_p}{4+\Delta_p}(1+O_p(\Delta_p a_n)) = (1-\frac{4}{4+\Delta_p})(1+O_p(\Delta_p a_n)) \\
&\geq 1 - O_P\left(\Delta_p^{-1}\right) - O_P\left(\Delta_p a_n\right).
\end{aligned}$$

Combining the lower and upper bounds for $R_{GS\text{-}LDA}(\boldsymbol{X})/R_{Bayes}$, we obtain

$$\frac{R_{GS\text{-}LDA}(\boldsymbol{X})}{R_{Bayes}} - 1 = O_P\left(\max\{\Delta_p^{-1}, \Delta_p a_n\}\right) = O_P\left(\max\{\Delta_p^{-1}, K^2\sqrt{(\log p)/n}\}\right).$$

This proves statement (c). □

Proof of (3.2.3). We use a similar argument to the proof of Proposition 1 given by Li and Li (2018). Letting $\alpha = (\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc})^{-1}$, we have

$$\begin{aligned}\Delta_{s+1} &= \begin{pmatrix} \boldsymbol{\delta}_S^T & \delta_c \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{SS} & \mathbf{\Sigma}_{Sc} \\ \mathbf{\Sigma}_{Sc}^T & \sigma_{cc} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\delta}_S \\ \delta_c \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\delta}_S^T & \delta_c \end{pmatrix} \begin{pmatrix} (\mathbf{\Sigma}_{SS} - \sigma_{cc}^{-1} \mathbf{\Sigma}_{Sc} \mathbf{\Sigma}_{Sc}^T)^{-1} & -\alpha \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc} \\ -\alpha \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} & \alpha \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_S \\ \delta_c \end{pmatrix}.\end{aligned}$$

By the Sherman–Morrison–Woodbury formula,

$$(\mathbf{\Sigma}_{SS} - \sigma_{cc}^{-1} \mathbf{\Sigma}_{Sc} \mathbf{\Sigma}_{Sc}^T)^{-1} = \mathbf{\Sigma}_{SS}^{-1} + \alpha \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc} \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1}.$$

Then we have

$$\begin{aligned}\Delta_{s+1} &= \begin{pmatrix} \boldsymbol{\delta}_S^T & \delta_c \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{SS}^{-1} + \alpha \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc} \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} & -\alpha \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc} \\ -\alpha \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} & \alpha \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_S \\ \delta_c \end{pmatrix} \\ &= \boldsymbol{\delta}_S^T \mathbf{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S + \alpha \boldsymbol{\delta}_S^T \mathbf{\Sigma}_{SS}^{-1} \mathbf{\Sigma}_{Sc} \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \delta_S - 2\alpha \delta_c \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S + \alpha \delta_c^2 \\ &= \Delta_s + \alpha (\delta_c - \mathbf{\Sigma}_{Sc}^T \mathbf{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S)^2.\end{aligned}$$

Hence, we have

$$\theta_{Sc} = \Delta_{s+1} - \Delta_s = \frac{(\delta_c - \mathbf{\Sigma}_{Sc}^T \mathbf{\Omega}_{SS} \boldsymbol{\delta}_S)^2}{\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Omega}_{SS} \mathbf{\Sigma}_{Sc}},$$

where $\mathbf{\Omega}_{SS} = \mathbf{\Sigma}_{SS}^{-1}$. With same argument as in the proof of Theorem 3.1,
$\sigma_{cc} - \mathbf{\Sigma}_{Sc}^T \mathbf{\Omega}_{SS} \mathbf{\Sigma}_{Sc} > 0$ for any $c \notin S$. Thus, $\theta_{Sc} \geq 0$.

□

3.8.2 Supporting Lemmas and their Proofs

Lemma 3.1. *Under Conditions 1 and 2, there exists a constant t_0 such that for all $0 < t < t_0$, the following results hold.*

(a) $P\left(\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq t\right)\leq p^2C_1e^{-C_2nt^2}$*, where* C_1 *and* C_2 *are some generic positive constants.*

(b) $P\left(\max_{j\leq p}|\widehat{\delta}_j-\delta_j|\geq t\right)\leq pC_1e^{-C_2nt^2}$*, where* C_1 *and* C_2 *are some generic positive constants.*

Proof of Lemma 3.1. These are standard concentration inequalities that follow from the normality assumption. The proof of (a) can be found in the proof of Lemma 3 of Bickel and Levina (2008b), and (b) is a result obtained by applying the Chernoff method. □

Lemma 3.2. *Under Condition 2 and if* $s\sqrt{\log(p)/n}=o(1)$*, it holds that*

$$P\left(\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|\lesssim s\sqrt{(\log p)/n}\right)\geq 1-C_1p^{-C_0};$$
$$P\left(\|\widehat{\mathbf{\Sigma}}_{SS}^{-1}-\mathbf{\Sigma}_{SS}^{-1}\|\lesssim s\sqrt{(\log p)/n}\right)\geq 1-C_1p^{-C_0},$$

where C_1 *is some generic positive constant and* C_0 *is a sufficiently large constant.*

Proof of Lemma 3.2. We have

$$\|\widehat{\mathbf{\Sigma}}_{SS}^{-1}-\mathbf{\Sigma}_{SS}^{-1}\|=\|\widehat{\mathbf{\Sigma}}_{SS}^{-1}(\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS})\mathbf{\Sigma}_{SS}^{-1}\|\leq\|\widehat{\mathbf{\Sigma}}_{SS}^{-1}\|\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|\|\mathbf{\Sigma}_{SS}^{-1}\|. \tag{3.8.13}$$

First, we bound $\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|$. By definition,

$$\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|\leq\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|_1=\max_{i\in S}\sum_{j\in S}|\widehat{\sigma}_{ij}-\sigma_{ij}|.$$

Then, it follows from Lemma 3.1 that

$$\begin{aligned}P\left(\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|\geq t\right)\leq P\left(\max_{i\in S}\sum_{j\in S}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq t\right)\leq P\left(\max_{i,j}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq t/s\right)\\ \leq p^2C_1e^{-C_2nt^2/s^2}.\end{aligned} \tag{3.8.14}$$

Letting $t=C_Ds\sqrt{(\log p)/n}$ for some large generic positive constant C_D and $C_0=C_2C_D$, we have

$$P\left(\|\widehat{\mathbf{\Sigma}}_{SS}-\mathbf{\Sigma}_{SS}\|\geq C_Ds\sqrt{(\log p)/n}\right)\leq C_1p^{2-C_2C_D}\leq C_1p^{-C_0}.$$

Next, we bound $\|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\|_2$. Note that $\|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\|_2 = 1/\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{SS})$. By Weyl's inequality,

$$\lambda_{\min}(\boldsymbol{\Sigma}_{SS}) \leq \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{SS}) + \lambda_{\max}(\boldsymbol{\Sigma}_{SS} - \widehat{\boldsymbol{\Sigma}}_{SS}) \leq \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{SS}) + \|\widehat{\boldsymbol{\Sigma}}_{SS} - \boldsymbol{\Sigma}_{SS}\|$$

Then, it follows from Condition 2 and (3.8.14) that

$$P\left(\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}) \leq \frac{1}{m - C_0 s\sqrt{(\log p)/n}}\right) \geq 1 - C_1 p^{2-C_2C_D} \geq 1 - C_1 p^{-C_0}.$$

By Condition 2 and (3.8.13), we have

$$P\left(\|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1} - \boldsymbol{\Sigma}_{SS}^{-1}\| \leq \frac{C_0 s\sqrt{(\log p)/n}}{m(m - C_0 s\sqrt{(\log p)/n})}\right) = P\left(\|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1} - \boldsymbol{\Sigma}_{SS}^{-1}\| \lesssim s\sqrt{(\log p)/n}\right)$$
$$\geq 1 - C_1 p^{2-C_2C_D} \geq 1 - C_1 p^{-C_0},$$

where in the first equality, we use the fact that as $s\sqrt{(\log p)/n} = o(1)$, $m - C_0 s\sqrt{(\log p)/n} \geq m/2$. □

Lemma 3.3. *Under Condition 1–2, and if $s\sqrt{(\log p)/n} = o(1)$, the following results hold.*

$$P\left(|\widehat{\boldsymbol{\delta}}_S^T \boldsymbol{\Sigma}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S| \lesssim s\sqrt{(\log p)/n}\right) \geq 1 - C_3 p^{-C_0},$$
$$P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \widehat{\boldsymbol{\Sigma}}_{Sc} - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{Sc}| \lesssim s\sqrt{(\log p)/n}\right) \geq 1 - C_3 p^{-C_0},$$
$$P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S - \boldsymbol{\Sigma}_{Sc}^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S| \lesssim s\sqrt{(\log p)/n}\right) \geq 1 - C_3 p^{-C_0}.$$

where C_3 is a positive constant depending on the C_1, and C_0 is a sufficiently large constant.

Proof of Lemma 3.3. To prove the first result, we have

$$\widehat{\boldsymbol{\delta}}_S^T \boldsymbol{\Sigma}_{SS}^{-1} \widehat{\boldsymbol{\delta}}_S = \boldsymbol{\delta}_S^T \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\delta}_S + 2\boldsymbol{\delta}_S^T \boldsymbol{\Sigma}_{SS}^{-1} (\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S) + (\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S)^T \boldsymbol{\Sigma}_{SS}^{-1} (\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S).$$

Then, we have

$$\begin{aligned}
&P\left(|\widehat{\boldsymbol{\delta}}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S|\geq t\right)\\
&=P\left(|2\boldsymbol{\delta}_S{}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)+(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|\geq t\right)\\
&\leq P\left(|2\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|+(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)\geq t\right)\\
&\leq P\left(|2\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|\geq t/2\right)+P\left((\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)\geq t/2\right).
\end{aligned}$$

By Cauchy-Schwarz inequality and Conditions 1 and 2, we have

$$\begin{aligned}
|\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|&\leq(\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S)^{1/2}\{(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)\}^{1/2}\\
&\leq(1/m)(\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S)^{1/2}\{(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)^T(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)\}^{1/2}\\
&\leq(sM/m)\max_{i,j\leq p}|\widehat{\delta}_{ij}-\delta_{ij}|.
\end{aligned}$$

We also have

$$(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)\leq(s/m)(\max_{j\leq p}|\widehat{\delta}_j-\delta_j|)^2.$$

Then, we have

$$\begin{aligned}
&P\left(|\widehat{\boldsymbol{\delta}}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S|\geq t\right)\\
&\leq P\left((sM/m)\max_{j\leq p}|\widehat{\delta}_j-\delta_j|\geq t/4\right)+P\left((s/m)(\max_{j\leq p}|\widehat{\delta}_j-\delta_j|)^2\geq t/2\right).
\end{aligned}$$

Letting $t=C_0s\sqrt{(\log p)/n}$ for some large enough constant C_0, then it follows from Lemma 3.1 that

$$P\left(|\widehat{\boldsymbol{\delta}}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S|\lesssim s\sqrt{(\log p)/n}\right)\geq 1-C_3p^{-C_0},$$

where C_3 is some positive constant depending on the C_1.

To prove the second result, note that

$$\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\Sigma}}_{Sc}=\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{Sc}+2\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})+(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc}).$$

Then, we have

$$\begin{aligned}
&P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{Sc}|\geq t\right)\\
&\leq P\left(|2\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|+|(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|\geq t\right)\\
&\leq P\left(|\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|\geq t/4\right)+P\left(|(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|\geq t/2\right).
\end{aligned}$$

By Cauchy-Schwarz inequality and Condition 2,

$$\begin{aligned}
|\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|&\leq(\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{Sc})^{1/2}\{(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})\}^{1/2}\\
&\leq(1/m)(\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{Sc})^{1/2}\{(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})\}^{1/2}\\
&\leq(sM/m)\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|,
\end{aligned}$$

where in the last inequality, we use the fact that $|\sigma_{ij}|\leq\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}\leq\lambda_{\max}(\boldsymbol{\Sigma})\leq M$, for all $i,j\leq p$. Also under Condition 2, we have

$$(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})\leq(s/m)(\max_{j\leq p}|\widehat{\sigma}_j-\sigma_j|)^2.$$

Then we have

$$\begin{aligned}
&P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{Sc}|\geq t\right)\\
&\leq P\left((sM/m)\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq t/4\right)+P\left((s/m)(\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|)^2\geq t/2\right).
\end{aligned}$$

Letting $t=C_0s\sqrt{(\log p)/n}$, for some large constant C_0. Then, it follows from Lemma 3.1 that

$$P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{Sc}|\lesssim s\sqrt{(\log p)/n}\right)\geq 1-C_3p^{-C_0},$$

where C_3 is some positive constant depending on the C_1 .

To prove the third result, note that

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S\\
&=\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S+\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})+\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)+(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S).
\end{aligned}$$

Then, we have

$$
\begin{aligned}
&P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S|\geq t\right)\\
&\leq P\left(|\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|\geq t/3\right)+P\left(|\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|\geq t/3\right)\\
&\quad+P\left(|(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|\geq t/3\right).
\end{aligned}
$$

By Cauchy-Schwarz inequality and Conditions 1 and 2, we have

$$
\begin{aligned}
|\boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})|&\leq(sM/m)\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|;\\
|\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)|&\leq(sM/m)\max_{j\leq p}|\widehat{\delta}_j-\delta_j|;\\
(\widehat{\boldsymbol{\Sigma}}_{Sc}-\boldsymbol{\Sigma}_{Sc})^T\boldsymbol{\Sigma}_{SS}^{-1}(\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\delta}_S)&\leq(s/m)(\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|)(\max_{j\leq p}|\widehat{\delta}_j-\delta_j|).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S|\geq t\right)\\
&\leq P\left((sM/m)\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq t/3\right)+P\left((sM/m)\max_{j\leq p}|\widehat{\delta}_j-\delta_j|\geq t/3\right)\\
&\quad+P\left((s/m)(\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|)(\max_{j\leq p}|\widehat{\delta}_j-\delta_j|)\geq t/3\right)\\
&\leq P\left((sM/m)\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq t/3\right)+P\left((sM/m)\max_{j\leq p}|\widehat{\delta}_j-\delta_j|\geq t/3\right)\\
&\quad+P\left(\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq\sqrt{mt/(3s)}\right)+P\left(\max_{j\leq p}|\widehat{\delta}_j-\delta_j|\geq\sqrt{mt/(3s)}\right).
\end{aligned}
$$

Letting $t=C_0s\sqrt{(\log p)/n}$ for some large constant C_0, it follows from Lemma 3.1 that

$$
P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S-\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S|\lesssim s\sqrt{(\log p)/n}\right)\geq 1-C_3p^{-C_0},
$$

where C_3 is some positive constant depending on the C_1. □

Lemma 3.4. *Under Condition 1–2 and if* $s\sqrt{(\log p)/n} = o(1)$*, the following results hold.*

$$
\begin{aligned}
P\left(|\widehat{\boldsymbol{\delta}}_S^T\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S| \lesssim s^2\sqrt{(\log p)/n}\right) &\geq 1 - C_4p^{-C_0};\\
P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\widehat{\boldsymbol{\Sigma}}_{Sc} - \boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{Sc}| \lesssim s^2\sqrt{(\log p)/n}\right) &\geq 1 - C_4p^{-C_0};\\
P\left(|\widehat{\boldsymbol{\Sigma}}_{Sc}^T\widehat{\boldsymbol{\Sigma}}_{SS}^{-1}\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\delta}_S| \lesssim s^2\sqrt{(\log p)/n}\right) &\geq 1 - C_4p^{-C_0};
\end{aligned}
$$

where C_4 *is some positive constant depending on the* C_1 *and* C_3 *and* C_0 *is a sufficiently large constant.*

Proof of Lemma 3.4. By definition,

$$
\begin{aligned}
&|\widehat{\boldsymbol{\delta}}_S^T(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1} - \boldsymbol{\Sigma}_{SS}^{-1})\widehat{\boldsymbol{\delta}}_S| \leq \|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1} - \boldsymbol{\Sigma}_{SS}^{-1}\|\widehat{\boldsymbol{\delta}}_S^T\widehat{\boldsymbol{\delta}}_S\\
&\leq \|\widehat{\boldsymbol{\Sigma}}_{SS}^{-1} - \boldsymbol{\Sigma}_{SS}^{-1}\|\{\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S + 2(\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S)^T\boldsymbol{\delta}_S + (\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S)^T(\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S)\}.
\end{aligned}
$$

By Condition 1, $\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S = O(s)$. It follows from Lemma 3.1 that $(\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S)^T\boldsymbol{\delta}_S = o_P(s)$ and $(\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S)^T(\widehat{\boldsymbol{\delta}}_S - \boldsymbol{\delta}_S) = o_P(s)$. Then, it follows from Lemma 3.2 that

$$
P\left(|\widehat{\boldsymbol{\delta}}_S^T(\widehat{\boldsymbol{\Sigma}}_{SS}^{-1} - \boldsymbol{\Sigma}_{SS}^{-1})\widehat{\boldsymbol{\delta}}_S| \lesssim s^2\sqrt{(\log p)/n}\right) \geq 1 - C_4p^{-C_0}.
$$

This result, together with Lemma 3.3 and the triangular inequality, prove the first result. The other two results can be proved by a similar argument, noting that $\boldsymbol{\Sigma}_{Sc}^T\boldsymbol{\Sigma}_{Sc} = O(s)$.

□

3.8.3 Additional Results in Cancer Subtype Analysis

Figure 3.5 shows the variable selection performance of the GS-LDA, ROAD and Logistic-L1 in cancer subtype analysis.

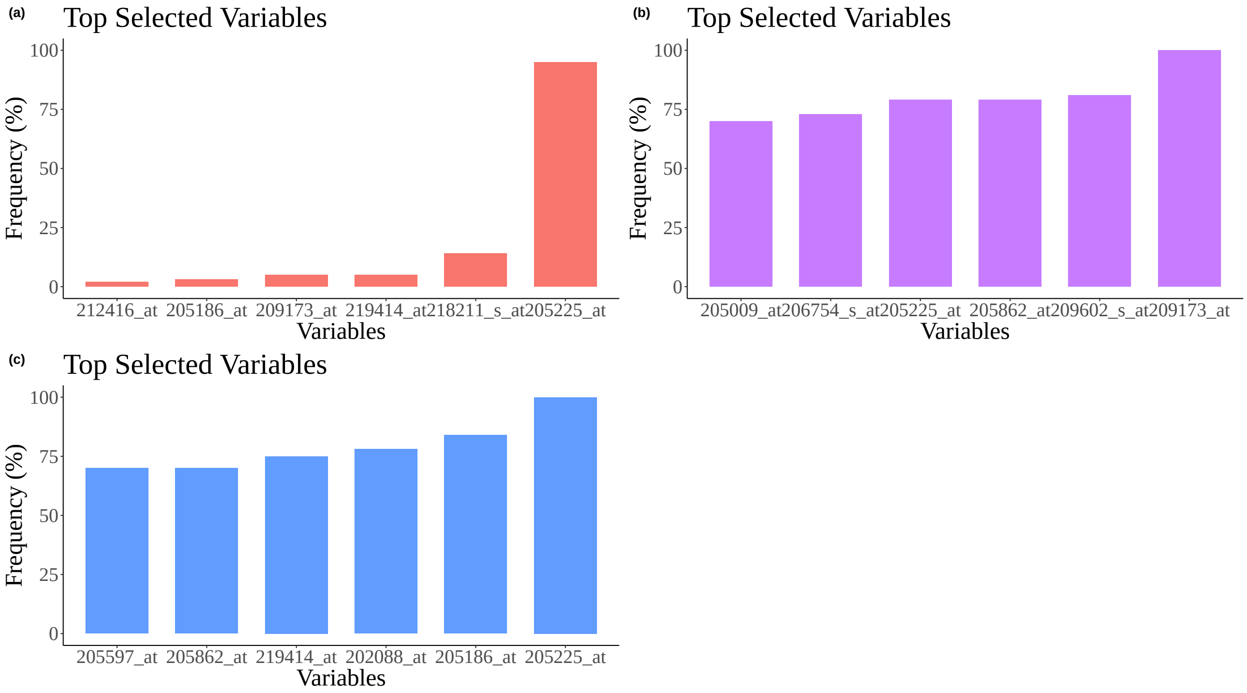


Figure 3.5. Variable selection performance of the three classifiers in classifying cancer subtypes: panel (a) for the GS-LDA; panel (b) for the ROAD; and panel (c) for the Logistic-L1.

CHAPTER 4
HIGH-DIMENSIONAL SEMIPARAMETRIC LATENT GAUSSIAN COPULA REGRESSION FOR MIXED DATA

4.1 Introduction

Regression analysis, finding the relationship between a response and the covariates, is a central statistical problem. Among all regression models, the linear regression model is the most popular one to deal with continuous response. With the emergence of big data containing enormous variables of mixed types, it poses great challenges on how to handle high dimensionality, non-normality, and heterogeneity of the data. To deal with high-dimensional linear regression model, a number of regularization methods (Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Negahban et al. 2012) have been proposed to yield sparse models and their theoretical properties have been well studied. To deal with non-normality of covariates, transformations, such as the Box-Cox transformation, Fisher's z-transformation and variance stablizing transformation, have been frequently applied to overcome potential violations of model assumptions (Carroll and Ruppert 1988). To handle variables of mixed types in the linear regression, the common practice would apply certain transformations on continuous and truncated covariates, and create dummy variables for categorical or ordinal variables. However, the choices of transformations could be subjective. It also remains unclear if applying these transformations can gaurantee to resolve non-normality and heterogeneity issues. Moreover, in some applications, the non-continuous, e.g. binary, ordinal, and truncated, variables can be generated from some latent continuous variables subject to unknown thresholds or detection limit. In these applications, it can be of interest to assess the association between the continuous response and the latent continuous variables behind the observed mixed variables. Above all, it is desirable to have a unified framework to solve this problem.

For an unsupervised problem of estimating correlations among mixed variables, some recent works proposed copula-based methods (Liu et al. 2009; 2012; Fan et al. 2017; Feng and Ning 2019; Yoon et al. 2020). Specifically, Liu et al. (2009; 2012) proposed a Gaussian copula model to estimate correlations among continuous variables. Fan et al. (2017) proposed a latent Gaussian copula model to simultaneously handle continuous and binary variables. Feng and Ning (2019) generalized the latent Gaussian copula model to handle ordinal and categorical variables. Yoon et al. (2020) proposed a truncated latent Gaussian copula model to handle truncated variables. These methods assume that there exists some latent continuous variables that generate the observed mixed variables, and the latent continuous variables follow a standard multivariate normal distribution, after applying some transformations. These methods propose to use rank-based quantities to estimate the correlations. They can be applied to a series of unsupervised learning problems, such as graph estimation, principal component analysis and canonical correlation analysis.

A few copula-based methods have also been developed to handle the supervised learning problem in the low-dimensional setting (Sungur 2005; Pitt et al. 2006; Crane and Hoek 2008; Masarotto et al. 2012; Noh et al. 2013). For example, Masarotto et al. (2012) proposed a general framework for the inference and model diagnosis using Gaussian copula when the responses are dependent. Noh et al. (2013) proposed a plug-in estimator of the regression function for a general copula regression. However, these methods only handle low-dimensional copula regression models. Recently, Cai and Zhang (2018) proposed a high-dimensional Gaussian copula regression model. They assume that the response and the covariates follow a Gaussian copula and developed a rank-based method to estimate the corresponding coefficients and established the oracle properties of their proposed estimator. However, all these works only allow continuous variables. There still lack unified approaches to handle high-dimensional mixed variables in a regression problem.

To this end, we propose a semiparametric latent Gaussian copula regression model to study the association between a continuous response variable with high-dimensional mixed covariates. The main contributions of the paper are as follows. First, our model gives a unified framework to handle mixed variables in a linear regression model. Second, we develop an imputation procedure to recover the latent variables in the test set to perform prediction with our model. The imputation procedure only depends on some closed-form formula. Finally, we quantify the prediction error of our method and compare it with the naive method that directly regresses the response on the observed covariates.

The rest of the chapter is organized as the following. Section 4.2 provides background and details on the proposed latent mixed Gaussian copula regression model. Section 4.3-4.4 describes the estimation of regression coefficients and investigates its statistical properties in terms of estimation and variable selection consistency. Section 4.5 provides details on the prediction of our method, and studies the corresponding prediction error with a comparison to that of the naive method. Section 4.6 presents extensive simulation studies to compare the proposed method with the naive method, demonstrating the superiority of the proposed method in terms of both estimation and prediction accuracy. Section 4.7 compares the prediction performances of the proposed method with the naive method using a communities crime data set from the UCI machine learning repository. All technical proofs are given in Section 4.8.

4.2 Latent Gaussian Copula Regression Model

We first introduce some notations. For a vector $\mathbf{a} \in \mathbb{R}^p$, let $\|\mathbf{a}\|_\infty = \max_{1 \le j \le p} |a_j|$, $\|\mathbf{a}\|_1 = \sum_{j=1}^p |a_j|$, $\|\mathbf{a}\|_2 = (\sum_{j=1}^p a_j^2)^{1/2}$ denote its max, L_1-, and Euclidean norms. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p \times p}$, let $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, $\|\mathbf{A}\|_\infty = \max_i \sum_{1 \le j \le p} |a_{ij}|$, $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ be the minimum and maximum eigenvalues of $\mathbf{A}$. For a symmetric matrix $\boldsymbol{\Sigma}$, we write $\boldsymbol{\Sigma} > \mathbf{0}$ if $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$. For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \le cb_n$. $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

First we review the Gaussian copula models proposed by Liu et al. (2009; 2012); Fan et al. (2017); Yoon et al. (2020); Feng and Ning (2019). For a random vector $\mathbf{z} \in \mathbb{R}^p$, if $\mathbf{f}(\mathbf{z}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{f} = (\mathrm{f}_1, ..., \mathrm{f}_p)^T$, f_j is a monotonically increasing function and $\boldsymbol{\Sigma}$ is a correlation matrix, then $\mathbf{z}$ is said to follow a nonparanormal distribution, denoted by $\mathbf{z} \sim NPN(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{f})$. We assume that the observed variable X_j relates to the latent variable Z_j based on the following transformations:

$$X_j = \begin{cases} Z_j, & \text{for } j \in \mathcal{C}; \\ I(Z_j > C_j), & \text{for } j \in \mathcal{B}; \\ \sum_{k=1}^{N_j} I(Z_j > C_{jk}), & \text{for } j \in \mathcal{O}; \\ I(Z_j > C_j) Z_j, & \text{for } j \in \mathcal{T}; \end{cases}$$

where $\mathcal{C}$, $\mathcal{B}$, $\mathcal{O}$, $\mathcal{T}$ are the index sets of continuous, binary, ordinal and truncated variables, $\mathbf{C}_{\mathcal{B}\cup\mathcal{T}} = (C_j)_{j \in \mathcal{B}\cup\mathcal{T}}$ is a vector of unknown thresholds for binary and truncated variables, and $C_{j1} < ... < C_{jN_j}$ are the N_j unknown thresholds for an ordinal variable. We call that $\mathbf{x}$ follows a latent mixed nonparanormal distribution, denoted by $\mathbf{x} \sim LMNPN(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{f}, \mathbf{C}, \mathbf{N})$, where $\mathbf{C} = (\mathbf{C}_{\mathcal{B}\cup\mathcal{T}}, \mathbf{C}_{\mathcal{O}})$ is a list of thresholds for binary, truncated, and ordinal variables, $\mathbf{N} = (N_j)_{j \in \mathcal{O}}$ is the number of thresholds for ordinal variables. We remark that the binary variable can be treated as a special case of ordinal variable with $N_j = 1$. The nonparanormal distribution was first studied by Liu et al. (2009; 2012) for continuous variables. Fan et al. (2017) extended it to include binary variables. With the same spirit, Yoon et al. (2020) and Feng and Ning (2019) further incorporated truncated and ordinal variables into the framework.

Different from these works that study the correlations in $\mathbf{x}$, our goal is to quantify the association between $\mathbf{x}$ and a continuous response Y. We assume $\mathbf{x}$ follows $LMNPN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{f}, \mathbf{C}, \mathbf{N})$, Y is observable and follows the Latent Mixed Gaussian Copula Regression model that

$$Y = \beta_0^* + \mathbf{f}(\mathbf{z})^T \boldsymbol{\beta}^* + \epsilon, \tag{4.2.1}$$

where ϵ has zero mean, finite variance, and $\mathrm{E}\{\epsilon|\mathbf{f}(\mathbf{z})\} = 0$.

A similar Gaussian copula regression problem was studied by Cai and Zhang (2018). Different from (4.2.1), they assumed that $(\mathbf{x}^T, Y)^T \sim NPN(\mathbf{0}, \check{\mathbf{\Sigma}}, \check{\mathbf{f}})$ where $\check{\mathbf{f}} = (\mathbf{f}, \mathrm{f}_0)$ and $\check{\mathbf{\Sigma}}$ is the correlation matrix of $(\mathbf{f}(\mathbf{x})^T, \mathrm{f}_0(Y))^T$. This assumption implies that $\mathrm{f}_0(Y) = \mathbf{f}(\mathbf{x})^T \boldsymbol{\theta} + \epsilon$, where $\boldsymbol{\theta} = \check{\mathbf{\Sigma}}_{xx}^{-1} \check{\mathbf{\Sigma}}_{xy}$, $\epsilon \sim N(0, 1 - \check{\mathbf{\Sigma}}_{xy}^T \check{\mathbf{\Sigma}}_{xx}^{-1} \check{\mathbf{\Sigma}}_{xy})$ and independent of $\mathbf{f}(\mathbf{x})$, $\check{\mathbf{\Sigma}}_{xy} = \mathrm{E}(\mathrm{f}_0(Y)\mathbf{f}(\mathbf{x}))$ and $\check{\mathbf{\Sigma}}_{xx} = \mathrm{E}(\mathbf{f}(\mathbf{x})\mathbf{f}(\mathbf{x})^T)$.

There are some critical differences between our model and the model in Cai and Zhang (2018). First, the response in our model is observable but not in their model because f_0 is unknown. Second, our model allows mixed covariates while theirs only allow continuous covariates. Third, our model relaxes normality assumption on $(\mathbf{f}(\mathbf{z})^T, Y)^T$, while they required that $(\mathbf{f}(\mathbf{x})^T, \mathrm{f}_0(Y))^T$ are jointly normal. Consequently, the noise in our model can depend on the latent covariates and follow non-normal distributions, while the noise in their model should be normally distributed and independent of the covariates. The most important difference is that their predicted value has to be one of the responses in the training set. In fact, the predicted value of Y given by Cai and Zhang (2018) has the form of $\widehat{\mathrm{f}}_0^{-1}(\widehat{\mathbf{f}}(\mathbf{x})^T \widehat{\boldsymbol{\theta}})$, where $\widehat{\mathbf{f}}$ and $\widehat{\boldsymbol{\theta}}$ are estimators of $\mathbf{f}$ and $\boldsymbol{\theta}$, and $\widehat{\mathrm{f}}_0^{-1}(t) = \inf\{x \in \mathbb{R} : \widetilde{F}_0(x) \geq \Phi(t)\}$ with $\widetilde{F}_0$ being the winsorized empirical distribution function of training responses and Φ being the distribution function of the standard normal distribution. Since $\widetilde{F}_0$ is a step function that increments only at the training responses, it restricts the predicted value to be one of them. On the contrary, our model does not have such a big restriction on prediction. In summary, compared with Cai and Zhang (2018), our model is more flexible to handle mixed variables, allows heteroscedastic errors, and does not restrict predicted values to be subsets of the training responses.

4.3 Estimation

A natural estimator of β_0^* is $\widehat{\beta}_0 = (1/n)\sum_{i=1}^n Y_i$. Let $\boldsymbol{\delta} = \mathrm{E}(\mathbf{f}(\mathbf{z})Y)$. Since $\boldsymbol{\delta} = \boldsymbol{\Sigma}\boldsymbol{\beta}^*$, it can be seen that $\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\, \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}/2 - \boldsymbol{\delta}^T\boldsymbol{\beta}$. Then, if we have estimators $(\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\delta}})$ for $(\boldsymbol{\Sigma}, \boldsymbol{\delta})$, we can estimate $\boldsymbol{\beta}^*$ by solving

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\mathrm{argmin}}\, \frac{1}{2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\delta}}^T\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1, \tag{4.3.1}$$

where $\|\boldsymbol{\beta}\|_1$ is an L_1-penalty function, and λ is a tuning parameter, which can be chosen by cross-validation. The problem in (4.3.1) is a convex optimization problem, which can be solved by the proximal gradient descent algorithm (Boyd and Vandenberghe 2004). We summarize its details in Algorithm 4.1. Next, we discuss how to obtain $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}$ based on observed data.

Algorithm 4.1: The proximal gradient algorithm for solving (4.3.1).

```
Initialization: Set $\widehat{\boldsymbol{\beta}}^{(0)} \in \mathbf{R}^p$ and $t = 0.8/\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}) \in \mathbf{R}$.
At the $k$th iteration, let
    $\widehat{\boldsymbol{\beta}}^{(k)} = \mathrm{prox}_{\mathrm{tg}}\left[\widehat{\boldsymbol{\beta}}^{(k-1)} - t\nabla L\{\widehat{\boldsymbol{\beta}}^{(k-1)}\}\right] = \mathrm{s}\left(\widehat{\boldsymbol{\beta}}^{(k-1)} - t\{\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{(k-1)} - \widehat{\boldsymbol{\delta}}\}, \lambda_1 t\right)$,
where $\mathrm{s}(x, \lambda) = \mathrm{sgn}(x)(|x| - \lambda)_+$ is the soft-thresholding function.
Iterate until $\|\widehat{\boldsymbol{\beta}}^{(k)} - \widehat{\boldsymbol{\beta}}^{(k-1)}\|_2 \le \rho$, where $\rho$ is a user-specified stopping threshold.
```

The estimation of $\boldsymbol{\Sigma}$ using observed mixed variables has been studied by Fan et al. (2017), Yoon et al. (2020), and Feng and Ning (2019). They propose to first obtain the Kendall's tau correlation between X_j and X_k, and rely on bridge functions to map it to the correlation between latent $\mathrm{f}_j(Z_j)$ and $\mathrm{f}_k(Z_k)$, which is defined as $\widetilde{\Sigma}_{jk}$. The estimated Kendall's tau correlation between X_j and X_k is given by

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1\le i<i'\le n} \mathrm{sgn}(X_{ij} - X_{i'j})\mathrm{sgn}(X_{ik} - X_{i'k}),\ 1 \le j, k \le p.$$

Fan et al. (2017) derived the bridge functions for pairwise correlations among binary and continuous variables. Yoon et al. (2020) further derived bridge functions between truncated and continuous/binary variables. A complete list of these bridge functions is shown in Section 4.8. Let $\widehat{F}_{jk}$ be one of the bridge functions. Then, $\widetilde{\Sigma}_{jk}$ can be obtained by solving $\widehat{F}_{jk}(\widetilde{\Sigma}_{jk}) = \widehat{\tau}_{jk}$. Fan et al. (2017)

and Yoon et al. (2020) have proved that all these bridge functions are invertible.

For ordinal variables, Feng and Ning (2019) proposed to dichotomize them into multiple binary variables. Suppose X_j and X_k are ordinal variables with $N_j + 1$ and $N_k + 1$ levels. Let $X_{ij}^{(p)} = I(X_{ij} \geq p)$, for $p = 1, ..., N_j$, and $X_{ik}^{(q)} = I(X_{ik} \geq q)$, for $q = 1, ..., N_k$. Then, the thresholds $\Delta_j^{(p)} = \mathrm{f}_j(C_{jp})$ and $\Delta_k^{(q)} = \mathrm{f}_k(C_{kq})$ can be estimated by

$$\widehat{\Delta}_j^{(p)} = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} X_{ij}^{(p)}),\ \widehat{\Delta}_k^{(q)} = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} X_{ik}^{(q)}),\ \text{for } j, k \in \mathcal{O}. \tag{4.3.2}$$

Using the bridge function for binary variables, the latent correlation between each pair of these binary variables can be obtained by solving $\widehat{F}_{jk}(\widetilde{\Sigma}_{jk}^{(p,q)}) = \widehat{\tau}_{jk}^{(p,q)}$, where

$$\widehat{\tau}_{jk}^{(p,q)} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \mathrm{sgn}(X_{ij}^{(p)} - X_{i'j}^{(p)})\mathrm{sgn}(X_{ik}^{(q)} - X_{i'k}^{(q)}),\ p = 1...N_j,\ q = 1...N_k.$$

Finally, $\widetilde{\Sigma}_{jk}$ can be calculated by $\widetilde{\Sigma}_{jk} = \sum_{q=1}^{N_k} \sum_{p=1}^{N_j} \widetilde{\Sigma}_{jk}^{(p,q)} w_{jk}^{(p,q)}$. If X_j is ordinal and X_k is of other types, a similar estimator can be constructed as $\widetilde{\Sigma}_{jk} = \sum_{p=1}^{N_j} \widetilde{\Sigma}_{jk}^{(p)} w_{jk}^{(p)}$, for $p = 1, ..., N_j$, where

$$\widehat{\tau}_{jk}^{(p)} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \mathrm{sgn}(X_{ij}^{(p)} - X_{i'j}^{(p)})\mathrm{sgn}(X_{ik} - X_{i'k}),\ \text{and } \widehat{F}_{jk}(\widetilde{\Sigma}_{jk}^{(p)}) = \widehat{\tau}_{jk}^{(p)}.$$

In these estimators, the weights must satisfy $0 \leq w_{jk}^{(p,q)} \leq 1, \sum_{q=1}^{N_k} \sum_{p=1}^{N_j} w_{jk}^{(p,q)} = 1$, and $0 \leq w_{jk}^{(p)} \leq 1, \sum_{p=1}^{N_j} w_{jk}^{(p)} = 1$. For simplicity, we use $w_{jk}^{(p,q)} = 1/(N_j N_k)$ and $w_{jk}^{(p)} = 1/N_j$.

Fan et al. (2017), Yoon et al. (2020), and Feng and Ning (2019) proved that $\widetilde{\mathbf{\Sigma}}$ is consistent to $\mathbf{\Sigma}$; see Lemma 4.1 in Section 4.8. To be used in (4.3.1), we require the estimator to be positive definite. Then, we project $\widetilde{\mathbf{\Sigma}}$ into the cone of positive definite matrices by solving $\widehat{\widetilde{\mathbf{\Sigma}}} = \mathrm{argmin}_{\mathbf{\Sigma} > \mathbf{0}} \|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$. Such a problem can be solved by Zhao et al. (2014).

To estimate $\boldsymbol{\delta}$, we need to develop new bridge functions. We propose to bridge δ_j with $\mathrm{E}(\mathrm{f}_j(X_j)Y)$ for $j \in \mathcal{C}$; with $\mathrm{E}(X_jY)$ for $j \in \mathcal{B}$; and with $\mathrm{E}(I(X_j > 0)\mathrm{f}_j(X_j)Y)$ for $j \in \mathcal{T}$. We summarize these bridge functions in Theorem 4.1.

Theorem 4.1. *Suppose $\boldsymbol{x} \sim LMNPN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{f}, \mathbf{C}, \mathbf{N})$ and Y follows (4.2.1). Then the bridge* *functions of $\boldsymbol{\delta}$ are given by*

$$F_j(\delta_j) = \begin{cases} \mathrm{E}(\mathrm{f}_j(X_j)Y) = \delta_j, & \textit{for } j \in \mathcal{C}; \\ \mathrm{E}(X_jY) = \beta_0^*(1-\Phi(\Delta_j)) + \phi(\Delta_j)\delta_j, & \textit{for } j \in \mathcal{B}; \\ \mathrm{E}(I(X_j > 0)\mathrm{f}_j(X_j)Y) = \beta_0^*\phi(\Delta_j) + C(\Delta_j)\delta_j, & \textit{for } j \in \mathcal{T}, \end{cases}$$

where $\Delta_j = \mathrm{f}_j(C_j)$, $C(\Delta_j) = E(I(\mathrm{f}_j(Z_j) > \Delta_j)\mathrm{f}_j(Z_j)^2) = \Delta_j\phi(\Delta_j) + 1 - \Phi(\Delta_j)$.

These bridge functions are all linear in δ_j, thus they are invertible. Using Theorem 4.1, we propose the following plug-in estimator of $\boldsymbol{\delta}$.

$$\widehat{\delta}_j = \begin{cases} n^{-1}\sum_{i=1}^n \widehat{\mathrm{f}}_j(X_{ij})Y_i, & \text{for } j \in \mathcal{C}; \\ \phi(\widehat{\Delta}_j)^{-1}(n^{-1}\sum_{i=1}^n X_{ij}Y_i - \widehat{\beta}_0(1-\Phi(\widehat{\Delta}_j))), & \text{for } j \in \mathcal{B}; \\ \sum_{p=1}^{N_j} w_j^{(p)}\phi(\widehat{\Delta}_j^{(p)})^{-1}[n^{-1}\sum_{i=1}^n X_{ij}^{(p)}Y_i - \widehat{\beta}_0\{1-\Phi(\widehat{\Delta}_j^{(p)})\}], & \text{for } j \in \mathcal{O}; \\ C(\widehat{\Delta}_j)^{-1}(n^{-1}\sum_{i=1}^n I(X_{ij} > 0)\widehat{\mathrm{f}}_j(X_{ij})Y_i - \widehat{\beta}_0\phi(\widehat{\Delta}_j)), & \text{for } j \in \mathcal{T}. \end{cases}$$

In the above formulae, for $j \in \mathcal{C} \cup \mathcal{T}$, we estimate $\mathrm{f}_j(t)$ by

$$\widehat{\mathrm{f}}_j(t) = \Phi^{-1}(\widetilde{F}_j(t)), \tag{4.3.3}$$

where $\widetilde{F}_j(t)$ is the winsorized empirical cumulative distribution function defined on $t \in \mathbb{R}$ for $j \in \mathcal{C}$ and $t > C_j$ for $j \in \mathcal{T}$. It has the form of

$$\widetilde{F}_j(t) = \varphi_n I(\widehat{F}_j(t) < \varphi_n) + \widehat{F}_j(t)I(\varphi_n \le \widehat{F}_j(t) \le 1-\varphi_n) + (1-\varphi_n)I(\widehat{F}_j(t) > 1-\varphi_n),$$

where $\widehat{F}_j(t) = (1/n)\sum_{i=1}^n I(X_{ij} \le t)$ and φ_n is often chosen to be $1/(2n)$. For $j \in \mathcal{B} \cup \mathcal{T}$, we estimate Δ_j by

$$\widehat{\Delta}_j = \begin{cases} \Phi^{-1}(1 - n^{-1}\sum_{i=1}^n X_{ij}), & \text{for } j \in \mathcal{B}; \\ \Phi^{-1}(1 - n^{-1}\sum_{i=1}^n I(X_{ij} > 0)), & \text{for } j \in \mathcal{T}. \end{cases} \tag{4.3.4}$$

For $j \in \mathcal{O}$, $\widehat{\Delta}_j^{(p)}$ is given by (4.3.2). Next, we show that $\widehat{\boldsymbol{\delta}}$ is consistent to $\boldsymbol{\delta}$.

Theorem 4.2. *Suppose the following conditions hold.*

Condition 1. $\|\epsilon\|_{\psi_2} < M$ *for some* $M > 0$*, where* $\|\epsilon\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathrm{E}\,|\epsilon|^p)^{1/p}$*.*

Condition 2. $\|\boldsymbol{\beta}^*\|_1 < M$ *for some* $M > 0$*.*

Condition 3. $\max_{j\in\mathcal{B}\cup\mathcal{T}} |\Delta_j| \leq M$ *and* $\max_{j\in\mathcal{O},p=1,\ldots,N_j} |\Delta_j^{(p)}| \leq M$ *for some* $M > 0$*.*

If $p = O(n^{\xi})$ *for an arbitrary* $\xi > 0$*, it holds that*

$$\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty = O_p\left(\frac{(\log p)^{1/2}(\log n)^{1/4}}{n^{1/4}}\right).$$

Condition 1 requires ϵ to follow the sub-Gaussian distribution. Condition 2 requires $\|\boldsymbol{\beta}^*\|_1$ to be bounded. Condition 3 requires the thresholds to be bounded. In Cai and Zhang (2018), they needed to estimate $\check{\boldsymbol{\Sigma}}_{xy} = \mathrm{E}(\mathrm{f}_0(Y)\mathbf{f}(\mathbf{x}))$. They proposed to estimate the jth element of $\check{\boldsymbol{\Sigma}}_{xy}$ by $\widehat{\Sigma}_{jy} = \sin(\pi\widehat{\tau}_{jy}/2)$, where $\widehat{\tau}_{jy}$ is the Kendall's tau estimator for correlation between Y and X_j. They proved that the resulting estimator $\widehat{\boldsymbol{\Sigma}}_{xy}$ has $\|\widehat{\boldsymbol{\Sigma}}_{xy} - \check{\boldsymbol{\Sigma}}_{xy}\|_\infty = O_p(\sqrt{\log p/n})$. However, this method requires normality assumption on $(\mathbf{f}(\mathbf{z})^T, \mathrm{f}_0(Y))$, and it only applies to continuous variables.

In summary, compared to Cai and Zhang (2018), our estimator does not require the normality assumption on $(\mathbf{f}(\mathbf{z})^T, Y)$ or $(\mathbf{f}(\mathbf{z})^T, \mathrm{f}_0(Y))$. Without these assumptions, we have to pay the price of estimating f_j for $j \in \mathcal{C}\cup\mathcal{T}$, which makes our convergence rate of $\|\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|_\infty$ to be slower than theirs, even though our estimator remains consistent when $p = O(n^{\xi})$ for any arbitrary $\xi > 0$. On the other hand, even with normality assumption, their method needs to estimate f_0 to predict Y. Thus, their predicted values must be subsets of training responses, which is another strong restriction.

4.4 Statistical Properties

We rely on the general M-estimation theory (Negahban et al. 2012) to study the statistical properties of $\widehat{\boldsymbol{\beta}}$. We define $\mathcal{M} = \{j : \beta_j^* \neq 0\}$ and $s = \|\mathcal{M}\|_0 = \sum_{j=1}^p I(\beta_j^* \neq 0)$. Theorem 4.3 gives the upper bounds of the estimation errors and Theorem 4.4 proves the variable selection consistency.

Theorem 4.3. *Suppose Conditions 1–3 and the following conditions hold.*

Condition 4. $\max_{1\leq j<k\leq p} |\Sigma_{jk}| \leq 1 - \delta$ *for some* $\delta > 0$*.*

Condition 5. $m \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq M$ *for some* m *and* $M > 0$*.*

If $s\sqrt{(\log p)/n} = o(1)$*,* $p = O(n^{\xi})$ *for an arbitrary* $\xi > 0$*,* *and* $\lambda = C(\log p)^{1/2}(\log n)^{1/4}n^{-1/4}$ *for some sufficiently large constant* C*, then*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_p\left(\frac{s^{1/2}(\log p)^{1/2}(\log n)^{1/4}}{n^{1/4}}\right),$$
$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(\frac{s(\log p)^{1/2}(\log n)^{1/4}}{n^{1/4}}\right).$$

Theorem 4.4. *Suppose Conditions 1–5 and the following conditions hold.*

Condition 6. $\|\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \leq M$ *for some* $M > 0$.

Condition 7. $\|\boldsymbol{\Sigma}_{\mathcal{M}^c\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \leq (1-\alpha)(1-\varepsilon)$ *for some* $\alpha > 0$ *and* $\varepsilon > 0$.

Condition 8. $\min_{j\in\mathcal{M}} |\beta_j^*| \gg ((\log p)(\log n)^{1/2}n^{-1/2})^{\gamma/2}$ *for some* $0 < \gamma < 1$.

If $s^2\sqrt{(\log p)/n} = o(1)$*,* $p = O(n^\xi)$ *for an arbitrary* $\xi > 0$*,*
and $\lambda = C((\log p)^{1/2}(\log n)^{1/4}n^{-1/4})^\gamma$*, where* $0 < \gamma < 1$ *and* C *is some sufficiently large constant, then with probability tending to 1, we have* $\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}\|_\infty \lesssim \lambda$ *and* $\widehat{\mathcal{M}} = \mathcal{M}$.

Condition 4 aims to avoid multicolinearity among latent variables. Condition 5 is a technical condition needed in the proof. Condition 6 requires that $\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}$ is invertible and assumes that the sup-norm of its inverse is bounded by a constant. Condition 7 is a standard irrepresentable condition that requires the important and unimportant variables cannot be highly correlated. It is well known that such a condition is needed for the variable selection consistency of the L_1-penalized methods. Condition 8 is a beta-min condition requiring that the minimal signal to be bounded away from zero. Given these conditions, Theorem 4.4 shows that $\widehat{\boldsymbol{\beta}}$ is variable selection consistent and gives uniformly consistent estimators of the nonzero components of $\boldsymbol{\beta}^*$.

4.5 Prediction

4.5.1 Imputation of Latent Variables

Let $\mathbf{x}_{test}$ be a new sample. If $\mathbf{x}_{test}$ contains binary, ordinal or truncated variables, the corresponding latent variables are not observable. Thus, we propose a method to impute them based on the observed variables. Let $\hat{\mathbf{u}}$ be the imputed value for the latent variable $\mathbf{f}(\mathbf{z}_{test})$. Then, the prediction is given by $\widehat{Y} = \widehat{\beta}_0 + \hat{\mathbf{u}}^T\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \hat{\mathbf{u}}^T_{\widehat{\mathcal{M}}}\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}}$. We will compare the Mean Squared Prediction Error (MSPE) of our method with that of the oracle method, which assumes that β_0^*, $\boldsymbol{\beta}^*$, $\mathbf{f}$, and $\boldsymbol{z}_{test}$ are all observed. In the oracle setting, we define the prediction of the response as $Y^* = \beta_0^* + \mathbf{f}(\mathbf{z}_{test})^T\boldsymbol{\beta}^*$. We will show that the MSPE of $\widehat{Y}$ can converge to the MSPE of Y^* up to a non-vanishing term due to the imputation error of latent variables behind binary, ordinary and truncated variables. For these variables, the truncation always leads to the loss of information, which cannot be fully recovered. Besides, we will also compare the MSPE of our method with that of a naive method, which regresses the response directly on the observed mixed variables.

First, we discuss how to impute latent variables if $\mathbf{f}$, $\mathbf{C}$ and $\mathbf{\Sigma}$ are known. We denote such an imputed value of $\mathbf{f}(\mathbf{z}_{test})$ as $\tilde{\mathbf{u}}$. We describe the imputation for each type of variables. For $j \in \mathcal{C}$, since $X_{test,j} = Z_{test,j}$, $\widetilde{u}_j = \mathrm{f}_j(X_{test,j})$ is the imputed value for $\mathrm{f}_j(Z_{test,j})$. For $j \in \mathcal{B} \cup \mathcal{O}$, we propose to impute $\mathrm{f}_j(Z_{test,j})$ by its expectation conditional on its observed value and other continuous variables. That is, $\widetilde{u}_j = \mathrm{E}\{\mathrm{f}_j(Z_{test,j})|\mathbf{v}_{test,\mathcal{I}}\}$, where $\mathbf{v}_{test,\mathcal{I}} = (V_{test,k})$, $V_{test,k} = \mathrm{f}_k(X_{test,k})$ for $k \in \mathcal{C}$, and $V_{test,k} = X_{test,k}$ for $k \notin \mathcal{C}$. We discuss in Proportion **??** how to choose a proper set $\mathcal{I}$ to condition on. For $j \in \mathcal{T}$, we impute $\mathrm{f}_j(Z_{test,j})$ depending on whether $X_{test,j} = 0$. When $X_{test,j} > 0$, $X_{test,j} = Z_{test,j}$. Therefore, the imputed value is $\widetilde{u}_j = \mathrm{f}_j(X_{test,j})$. When $X_{test,j} = 0$, we propose to impute it by $\widetilde{u}_j = \mathrm{E}\{\mathrm{f}_j(Z_{test,j})|\mathbf{v}_{test,\mathcal{I}}\}$. We define the Mean Squared Imputation Error (MSIE) of $\widetilde{u}_j$ by $MSIE(\widetilde{u}_j;\mathcal{I}) = \mathrm{E}[\widetilde{u}_j - \mathrm{f}_j(Z_{test,j})]^2$.

Proposition 4.1. *For $j \in \mathcal{B}$ and any $\widetilde{\mathcal{C}} \subset \mathcal{C}$, it holds that*

$$(1-\widetilde{\xi}_j)\left[1-\mathrm{E}_{L_j}\left\{\frac{\phi(L_j)^2}{\Phi(L_j)(1-\Phi(L_j))}\right\}\right] = MSIE(\widetilde{u}_j;\widetilde{\mathcal{C}}\cup\{j\}) < MSIE(\widetilde{u}_j;\widetilde{\mathcal{C}}) = 1-\widetilde{\xi}_j,$$

where $\widetilde{\xi}_j = \mathbf{\Sigma}^T_{\widetilde{\mathcal{C}}j}\mathbf{\Sigma}^{-1}_{\widetilde{\mathcal{C}}\widetilde{\mathcal{C}}}\mathbf{\Sigma}_{\widetilde{\mathcal{C}}j}$, $L_j \sim N(\Delta_j/(1-\widetilde{\xi}_j)^{1/2}, \widetilde{\xi}_j/(1-\widetilde{\xi}_j))$, and $\Delta_j = \mathrm{f}_j(C_j)$.

Proposition 4.1 indicates that, to impute the latent variables behind binary variables, conditioning on its observed value and other observed continuous variables guarantees to have smaller MSIE than solely conditioning on continuous variables. Besides, through numerical experiments we find that the MSIE decreases as $\widetilde{\xi}_j$ increases. It was proved in Proposition 1 of Li and Li (2018) that if there is a sequence $\widetilde{\mathcal{C}}_1 \subset \ldots \subset \widetilde{\mathcal{C}}_k \subset \widetilde{\mathcal{C}}_{k+1} \subset \cdots \subset \mathcal{C}$, $\widetilde{\xi}_j$ increases as k increases. In other words, the more continuous variables we condition on, the larger $\widetilde{\xi}_j$ is. These two results together indicate that we should condition on all continuous variables. On the other hand, if we condition on other observed binary, ordinal, or truncated variables, it needs to evaluate multiple integrals which can be computationally prohibitive. Besides, it is unclear if conditioning on these variables can further reduce the MSIE. Considering all these aspects, we choose $\mathcal{I} = \{j\} \cup \mathcal{C}$. Then,

$$\widetilde{u}_j = \begin{cases} \mathrm{f}_j(X_{test,j}), & \text{for } j \in \mathcal{C}; \\ \sqrt{1-\xi_j}\{I(X_{test,j}=0)\dfrac{-\phi(l_j)}{\Phi(l_j)} + I(X_{test,j}=1)\dfrac{\phi(l_j)}{1-\Phi(l_j)}\} & \\ + \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,\mathcal{C}}), & \text{for } j \in \mathcal{B}; \\ \sqrt{1-\xi_j}\{\sum_{k=0}^{N_j} I(X_{test,j}=k)\dfrac{\phi(l_{jk})-\phi(l_{j(k+1)})}{\Phi(l_{j(k+1)})-\Phi(l_{jk})}\} & \\ + \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,\mathcal{C}}), & \text{for } j \in \mathcal{O}; \\ f_j(X_{test,j})I(X_{test,j}>0) & \\ + \{-\dfrac{\sqrt{1-\xi_j}\phi(l_j)}{\Phi(l_j)} + \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,\mathcal{C}})\}I(X_{test,j}=0), & \text{for } j \in \mathcal{T}, \end{cases} \tag{4.5.1}$$

where $\xi_j = \mathbf{\Sigma}_{\mathcal{C}j}^T \mathbf{\Sigma}_{\mathcal{C}\mathcal{C}}^{-1} \mathbf{\Sigma}_{\mathcal{C}j}$, $\boldsymbol{\eta}_j = \mathbf{\Sigma}_{\mathcal{C}\mathcal{C}}^{-1}\mathbf{\Sigma}_{\mathcal{C}j}$; for $j \in \mathcal{B} \cup \mathcal{T}$,
$l_j = (\Delta_j - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,\mathcal{C}}))/\sqrt{1-\xi_j}$ and $\Delta_j = \mathrm{f}_j(C_j)$; for $j \in \mathcal{O}$,
$l_{jk} = (\Delta_j^{(k)} - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,\mathcal{C}}))/\sqrt{1-\xi_j}$ and $\Delta_j^{(k)} = \mathrm{f}_j(C_{jk})$. The expression of $MSIE(\widetilde{u}_j, \mathcal{I})$ is given in Theorem 4.6 in Section 4.8.

In practice, we need to estimate parameters in (4.5.1) using training data. For f_j, we estimate it by $\widehat{\mathrm{f}}_j$ as defined in (4.3.3). For Δ_j of $j \in \mathcal{B} \cup \mathcal{T}$, we estimate it by (4.3.4); for $\Delta_j^{(k)}$ of $j \in \mathcal{O}$, we estimate it by (4.3.2). For $\boldsymbol{\eta}_j$, we propose to estimate it by solving

$$\widehat{\boldsymbol{\eta}}_j = \underset{\boldsymbol{\eta}_j}{\operatorname{argmin}} \ (1/2)\boldsymbol{\eta}_j^T \widehat{\mathbf{\Sigma}}_{\mathcal{C}\mathcal{C}} \boldsymbol{\eta}_j - \boldsymbol{\eta}_j^T \widehat{\mathbf{\Sigma}}_{\mathcal{C}j} + \lambda_2 \|\boldsymbol{\eta}_j\|_1, \tag{4.5.2}$$

where $\widehat{\mathbf{\Sigma}}_{\mathcal{C}\mathcal{C}}$ and $\widehat{\mathbf{\Sigma}}_{\mathcal{C}j}$ are submatrices of $\widehat{\mathbf{\Sigma}}$, and we impose an L_1-penalty on $\boldsymbol{\eta}_j$ to regulate the problem. We assume $\boldsymbol{\eta}_j$ is weakly sparse in the sense that $\boldsymbol{\eta}_j \in B_q(R_q) = \{\boldsymbol{\theta} \in \mathbb{R}^{p_1} : \sum_{j=1}^{p_1} |\theta_j|^q < R_q\}$ for $j \in \mathcal{B} \cup \mathcal{O} \cup \mathcal{T}$, where $q \in (0,1]$, and $p_1 = \|\mathcal{C}\|_0$. The problem (4.5.2) can also be solved by the proximal gradient descent algorithm. Moreover, ξ_j can be estimated by $\widehat{\xi}_j = \widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{\Sigma}}_{\mathcal{C}j}$. Plugging all these estimators into (4.5.1) gives the imputed value $\hat{\mathbf{u}}$. Finally, we remark that we only need to perform such imputations for $j \in \widehat{\mathcal{M}}$.

4.5.2 Prediction Error

We define the Mean Squared Prediction Error (MSPE) of our method as $MSPE = (1/n_{test})\sum_{i=1}^{n_{test}}(\widehat{Y}_i - Y_i^*)^2$. In the rest of this section, our arguments are conditioning on the event that $\{\widehat{\mathcal{M}} = \mathcal{M}\}$. The

MSPE can be decomposed as

$$
\begin{aligned}
MSPE = (\widehat{\beta}_0 - \beta_0^*)^2 &+ {\boldsymbol{\beta}^*}^T_{\mathcal{M}} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{\mathbf{u}}_{i\mathcal{M}} - \widetilde{\mathbf{u}}_{i\mathcal{M}})^{\otimes 2} \right\} \boldsymbol{\beta}^*_{\mathcal{M}} \\
&+ {\boldsymbol{\beta}^*}^T_{\mathcal{M}} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widetilde{\mathbf{u}}_{i\mathcal{M}} - \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{M}})^{\otimes 2} \right\} \boldsymbol{\beta}^*_{\mathcal{M}} \\
&+ (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}})^T \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}(\mathbf{z}_{test,i})^{\otimes 2}_{\mathcal{M}} \right\} (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + R
\end{aligned}
\tag{4.5.3}
$$

where the first term is due to the estimation error of $\widehat{\beta}_0$, the second term is due to the estimation error of $\widehat{\mathbf{u}}$ to $\widetilde{\mathbf{u}}$, the third term is due to imputation error by $\widetilde{\mathbf{u}}_{\mathcal{M}}$, the fourth term is due to estimation error of $\widehat{\boldsymbol{\beta}}$, and R represents the cross-product and higher-order remainder terms. Among them, only the third term does not vanish since it does not depend on training samples. All the rest terms vanish when the training sample size divergences. Letting $\mathcal{A} = (\mathcal{B} \cup \mathcal{O} \cup \mathcal{T}) \cap \mathcal{M}$ and $s_{\mathcal{A}} = \|\mathcal{A}\|_0 = \sum_{j \in \mathcal{B} \cup \mathcal{O} \cup \mathcal{T}} I(\beta_j^* \neq 0)$, Theorem **??** quantifies the MSPE.

Theorem 4.5. *Suppose Conditions 1–8 hold and $R_q < \infty$. If $p = O(n^{\xi})$ for an arbitrary $\xi > 0$ and $s^2\sqrt{\log p/n} = o(1)$, $\lambda_1 \asymp a_n^{\gamma/2}$ for some $0 < \gamma < 1$, $\lambda_2 \asymp \sqrt{\log p/n}$, it follows that*

$$
MSPE = \begin{cases} {\boldsymbol{\beta}^*}^T_{\mathcal{A}} (\mathrm{E}(\mathbf{u}^*_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}})^{\otimes 2}) \boldsymbol{\beta}^*_{\mathcal{A}} + O_p(\sqrt{sa_n} \vee \sqrt{s_{\mathcal{A}} b_n}), & \textit{when } \mathcal{A} \neq \emptyset; \\ O_p(sa_n), & \textit{when } \mathcal{A} = \emptyset, \end{cases}
$$

where $a_n = (\log p)(\log n)^{1/2} n^{-1/2}$ and $b_n = (\log p/n)^{1-qr}$ for some $r \in (0,1)$.

Theorem 4.5 shows that when $\mathcal{A} = \emptyset$, the MSPE converges to 0 in a rate of $O_P(sa_n)$. This is because the response only associates with some continuous variables whose MSIEs are zero; see Theorem 4.6. As such, the convergence rate is dominated by the estimation error of $\widehat{\boldsymbol{\beta}}_{\mathcal{M}}$. When $\mathcal{A} \neq \emptyset$, since the MSIE of binary, ordinary and truncated variables cannot vanish, there is a non-vanishing term of ${\boldsymbol{\beta}^*}^T_{\mathcal{A}} (E(\mathbf{u}^*_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}})^{\otimes 2}) \boldsymbol{\beta}^*_{\mathcal{A}}$. The term of $O_P(\sqrt{sa_n})$ is due to the estimation error of $\widehat{\boldsymbol{\beta}}_{\mathcal{M}}$. The term of $O_P(\sqrt{s_{\mathcal{A}} b_n})$ is due to the estimation error of $\widehat{\mathbf{u}}$ to $\widetilde{\mathbf{u}}$. They both vanish as $n \to \infty$. This shows that up to a non-vanishing error that can never be recovered from the training data, our method can accurately predict the response when the training size is large enough.

4.5.3 Comparison with a Naive Method

We compare our method's prediction error with a naive method. We assume the true model follows (4.2.1). Let $\mathbf{x}_{aug} = (1, \mathbf{x}^T)^T$, $\mathbf{f}(\mathbf{z})_{aug} = (1, \mathbf{f}(\mathbf{z})^T)^T$, and $\boldsymbol{\beta}^*_{aug} = (\boldsymbol{\beta}^*_0, \boldsymbol{\beta}^{*^T})^T$. The naive method regresses Y directly on the observed $\mathbf{x}_{aug}$ and estimates $\boldsymbol{\gamma}^*$, where

$$\boldsymbol{\gamma}^* = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \mathrm{E}(Y - \mathbf{x}_{aug}^T \boldsymbol{\gamma}^*)^2 = \mathbf{H}\boldsymbol{\beta}^*_{aug}, \tag{4.5.4}$$

and $\mathbf{H} = \mathrm{E}(\mathbf{x}_{aug}\mathbf{x}_{aug}^T)^{-1}\mathrm{E}(\mathbf{x}_{aug}\mathbf{f}(\mathbf{z})_{aug}^T)$. However, in the true model $\mathrm{E}\{Y|\mathbf{x}_{aug}\}$ is a nonlinear function of $\mathbf{x}_{aug}$ but the naive method mistakenly treats it as a linear function. By (4.2.1), $\mathrm{E}\{Y|\mathbf{z}\} = \mathbf{f}(\mathbf{z})_{aug}^T\boldsymbol{\beta}^*_{aug}$. Thus, we can view the naive method as imputing $\mathbf{f}(\mathbf{z})_{aug}$ by $\mathbf{H}^T\mathbf{x}_{aug}$. We define the oracle predictions of our and naive methods by $\widehat{Y}_{ora} = \widetilde{\mathbf{u}}^T\boldsymbol{\beta}^* + \beta^*_0$, and $\widehat{Y}^{naive}_{ora} = \mathbf{x}_{aug}^T\boldsymbol{\gamma}^* = \widetilde{\mathbf{u}}_{naive}^T\boldsymbol{\beta}^* + \beta^*_0$, where $\widetilde{\mathbf{u}}$ is given by (4.5.1) and

$$\widetilde{\mathbf{u}}_{naive} = \mathrm{E}(\mathbf{f}(\mathbf{z})\mathbf{x}^T)\mathrm{E}(\mathbf{x}\mathbf{x}^T)^{-1}\left\{\mathbf{x} - \frac{1 - \mathrm{E}(\mathbf{x})^T\mathrm{E}(\mathbf{x}\mathbf{x})^{-1}\mathbf{x}}{1 - \mathrm{E}(\mathbf{x})^T\mathrm{E}(\mathbf{x}\mathbf{x})^{-1}\mathrm{E}(\mathbf{x})}\mathrm{E}(\mathbf{x})\right\}. \tag{4.5.5}$$

In such definitions, we assume that all parameters are known and highlight that $\mathbf{x}$ and $\mathbf{f}(\mathbf{z})$ in (4.5.5) do not contain the intercept. $\widetilde{\mathbf{u}}_{naive}$ can be treated as an imputation of $\mathbf{f}(\mathbf{z})$ by regressing it on $\mathbf{x}$ using a linear model. Since $\mathbf{f}(\mathbf{z})$ is not a linear function of $\mathbf{x}$, such a naive method is subject to serious model mis-specification error and $\widetilde{\mathbf{u}}_{naive}$ generally has larger MSIE than $\widetilde{\mathbf{u}}$.

Next, we compare the predictions of these two methods by taking parameter estimation into account. Let $\boldsymbol{Y} \in \mathbb{R}^n$, $\mathbf{X}_{aug} = (\mathbf{1}, \mathbf{X}) \in \mathbb{R}^{n\times(p+1)}$ and $\mathbf{f}(\mathbf{Z})_{aug} = (\mathbf{1}, \mathbf{f}(\mathbf{Z})) \in \mathbb{R}^{n\times(p+1)}$ be the response, observed and latent variables in the training set, and $\mathbf{x}_{test,aug} = (1, \mathbf{x}_{test}^T)^T \in \mathbb{R}^{p+1}$. For simplicity, we assume $p < n$. For the naive method, we consider estimating $\boldsymbol{\gamma}^*$ by the Ordinary Least Squares estimator $\widehat{\boldsymbol{\gamma}}^{OLS} = (\mathbf{X}_{aug}^T\mathbf{X}_{aug})^{-1}\mathbf{X}_{aug}^T\boldsymbol{Y}$. Then, its prediction error is given by

$$\widehat{Y}^{naive} = \widehat{\mathbf{u}}_{naive}^T\boldsymbol{\beta}^* + \beta^*_0 + \mathbf{x}_{test,aug}^T(\mathbf{X}_{aug}^T\mathbf{X}_{aug})^{-1}\mathbf{X}_{aug}^T\boldsymbol{\epsilon}, \tag{4.5.6}$$

where

$$\widehat{\mathbf{u}}_{naive} = \frac{1-\nu_1}{n-\nu_2}\mathbf{f}(\mathbf{Z})^T\mathbf{1} + \mathbf{f}(\mathbf{Z})^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\left\{\mathbf{x}_{test} - \frac{1-\nu_1}{n-\nu_2}\mathbf{X}^T\mathbf{1}\right\},$$

$\nu_1 = \mathbf{1}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{test}$, and $\nu_2 = \mathbf{1}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{1}$. Comparing $\widehat{Y}^{naive}$ with $\widehat{Y}^{naive}_{ora}$, we can find that estimating $\boldsymbol{\gamma}^*$ leads to an estimation error in $\widehat{\mathbf{u}}_{naive}$ and an extra term of $\mathbf{x}_{test,aug}^T(\mathbf{X}_{aug}^T\mathbf{X}_{aug})^{-1}\mathbf{X}_{aug}^T\boldsymbol{\epsilon}$, which converges to zero. Even if $\widehat{\mathbf{u}}_{naive}$ converges to $\widetilde{\mathbf{u}}_{naive}$ when sample size diverges, it still suffers from model misspecification. On the other hand, our method's prediction is $\widehat{\mathbf{u}}^T\widehat{\boldsymbol{\beta}}+\widehat{\beta}_0$. Comparing it with $\widehat{Y}^{naive}$, we find that even though $\widehat{\mathbf{u}}$ is better than $\widehat{\mathbf{u}}_{naive}$ in terms of imputing latent variables, our method's prediction needs to account for estimation errors in $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$, while $\widehat{Y}^{naive}$ in (4.5.6) only contain the true parameters β_0^* and $\boldsymbol{\beta}^*$. This suggests that when the improvement in imputation error surpasses the estimation errors of $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$, our method has advantage over the naive method. In empirical studies, we find that when the latent continuous variables are highly skewed, our method performs better than the naive method. In Section 4.6, we use simulation studies to further illustrate this point.

4.6 Simulations

We conduct simulation studies to compare our method with the naive method. We simulate for three scenarios. We set training and test set sizes to be $n = 500$ and $n_{test} = 1000$, and consider $p = 100$ and $p = 500$. In all scenarios, the errors are independent of covariates. The setup of the scenarios are as follows. **Scenario 1:** Let $\mathbf{B}_1 = \text{diag}(\mathbf{D}, \mathbf{I}_5)$ and $\boldsymbol{\Sigma} = \text{diag}(\mathbf{B}_1, \mathbf{B}_1, \mathbf{I}_{10}, ..., \mathbf{I}_{10})$, where $\mathbf{D} = (d_{ij})$, $d_{ii} = 1$ for $i = 1, ..., 5$, $d_{ij} = 0.3$ for $1 \le i \ne j \le 5$, and $\mathbf{I}_5$ is a five-dimensional identity matrix. We generate $\mathbf{f}(\mathbf{z})$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$ and choose $\mathrm{f}_j(Z_j) = Z_j^3$ for $1 \le j \le p$. The observed variables are generated by

$$X_j = \begin{cases} I(Z_j > C_j), & \text{for } j = 1 + 10t, \text{ and } 0 \le t \le p/10 - 1; \\ Z_j, & \text{otherwise,} \end{cases}$$

where $C_j = -0.3^{(1/3)}$. The response is generated by $Y = 5\sum_{j=1}^{5}\mathrm{f}_j(Z_j) + 5\sum_{j=11}^{15}\mathrm{f}_j(Z_j) + \epsilon$, where $\epsilon \sim N(0, 1)$.

Scenario 2: We choose $\boldsymbol{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = 0.3^{|i-j|}$ for $1 \le i, j \le p$, and $\mathrm{f}_j(Z_j) = \log(Z_j)$ for $1 \le j \le p$. The observed variables are generated by

$$X_j = \begin{cases} I(Z_j > C_j), & \text{for } j = 1 + 10t, \text{ and } 0 \le t \le p/10 - 1; \\ Z_j, & \text{otherwise,} \end{cases}$$

where $C_j = \exp(-0.3)$. The response is generated the same as in Scenario 1.

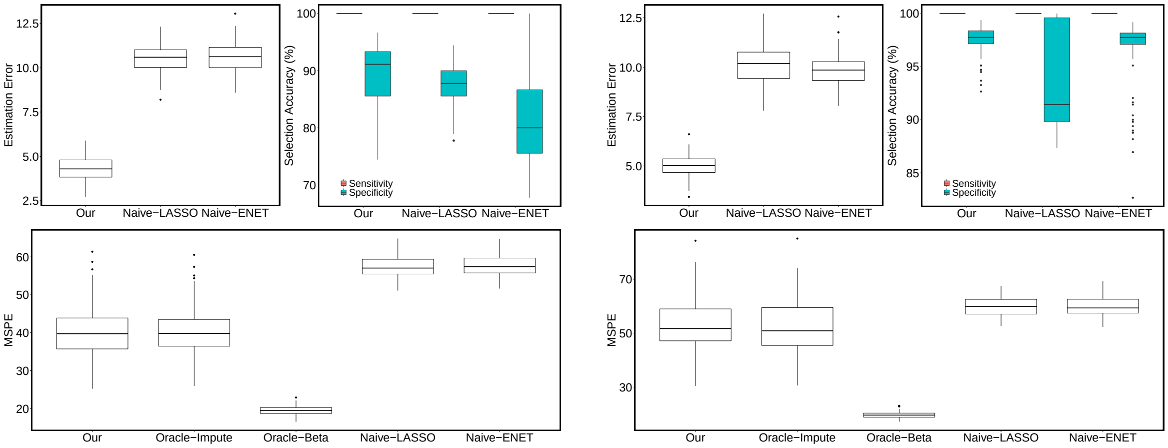
Scenario 3: We choose $\mathbf{\Sigma}$ the same as in Scenario 1 and two marginal transformation functions. We choose either $\mathrm{f}_j(Z_j) = Z_j^3$ or $\mathrm{f}_j(Z_j) = \log(Z_j)$ for all $1 \leq j \leq p$. The observed variables are generated by

$$X_j = \begin{cases} \sum_{k=1}^{2} I(\mathrm{f}_j(Z)_j > \Delta_j^{(k)}), & \text{for } j = 1 + 20t, \text{ and } 0 \leq t \leq p/20 - 1; \\ Z_j I(\mathrm{f}_j(Z)_j > \Delta_j), & \text{for } j = 11 + 20t, \text{ and } 0 \leq t \leq p/20 - 1; \\ Z_j, & \text{otherwise}, \end{cases}$$

where $\Delta_j^{(1)} = \mathrm{f}_j(C_{j1}) = -0.1$ and $\Delta_j^{(2)} = \mathrm{f}_j(C_{j2}) = 0.1$ for $j \in \mathcal{O}$, and $\Delta_j = \mathrm{f}_j(C_j) = 0.1$ for $j \in \mathcal{T}$. The response is generated by $Y = 5\sum_{j=1}^{5} \mathrm{f}_j(Z_j) + 5\sum_{j=11}^{15} \mathrm{f}_j(Z_j) + \epsilon$, where $\epsilon \sim Uniform[-1, 1]$.

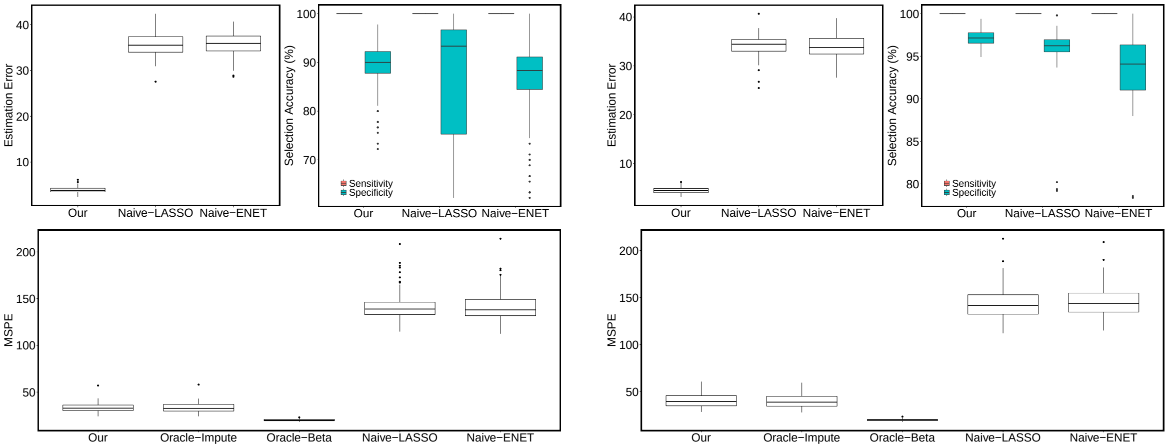
For each scenario, we independently generate n samples for the training set and n_{test} samples for the test set. For our method, we obtain the estimators of regression coefficients by solving (4.3.1) and the imputed latent variables by using methods described in Section 4.5.1. We compare it with two oracle-like methods. For the first "Oracle-impute" method, it imputes the latent variables using (4.5.1) where the parameters therein are assumed to be known and estimates regression coefficients from (4.3.1). We define its prediction by $\widehat{\beta}_0 + \widetilde{\mathbf{u}}_{\widehat{\mathcal{M}}}^T \widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}}$. For the second "Oracle-beta" method, it assumes β_0^* and $\boldsymbol{\beta}^*$ are known and imputes the latent variables by $\widehat{\mathbf{u}}$. We define its prediction by $\beta_0^* + \widehat{\mathbf{u}}_{\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}}^*$. We also involve two naive methods into the comparison. The naive methods directly regress the response on the observed variables by solving penalized Least Squares problems with either an L_1 or an elastic-net penalty. We refer to them as "Naive-LASSO" and "Naive-ENET". For each scenario, we repeat simulations for 100 times and report the estimation errors of regression coefficients, variable selection performance, and MSPEs.

It is seen from Figure 4.1 that, our method has clear advantage over Naive-LASSO and Naive-ENET in Scenario 1. It has much smaller estimation errors. In term of In variable selection, even though the sensitivity of the three methods are comparable, our method has better specificity. Finally, the MSPE of our method is smaller than that of Naive-LASSO and Naive-ENET, and comparable to the MSPE of the Oracle-impute method, suggesting that estimating unknown parameters in the imputation formulae does not worsen the MSPE. This agrees with Theorem 4.5. However, our method's MSPE is worse than that of the Oracle-beta method. One reason is due to the estimation



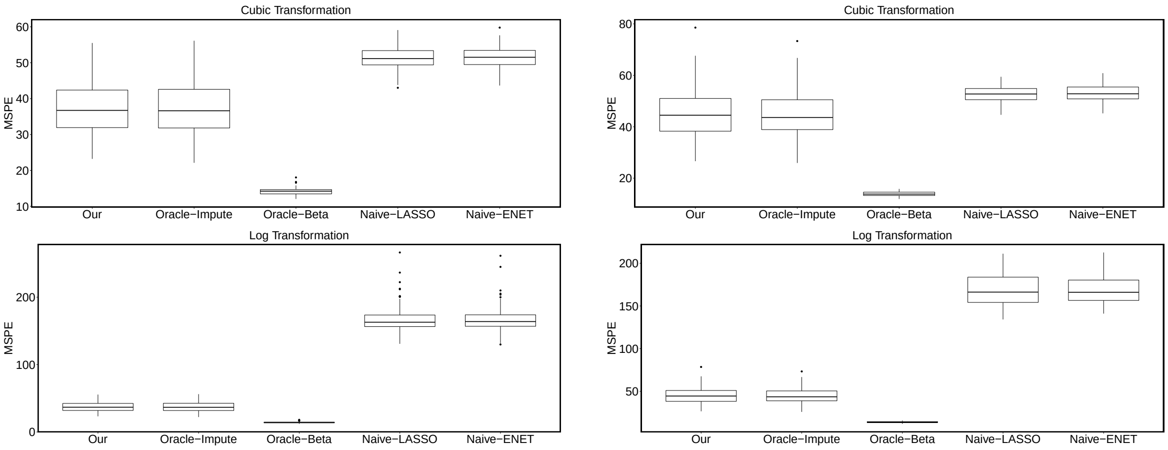
$p = 100$ $p = 500$

Figure 4.1. Estimation error, variable selection performance, and MSPEs for the five competitors in Scenario 1.



$p = 100$ $p = 500$

Figure 4.2. Estimation error, variable selection performance, and MSPEs for the five competitors in Scenario 2.



$p = 100$ $p = 500$

Figure 4.3. MSPEs for the five competitors in Scenario 3.

error in $\widehat{\boldsymbol{\beta}}$. Another reason is that our method selects some false positives, which also need to be imputed. That could enlarge the MSPE. Finally, we emphasize that these two oracle-like methods are not feasible in practice. We only use them as benchmarks to evaluate the prediction performance of the our method.

We can draw similar conclusions from Figure 4.2. Indeed, the advantage of our method becomes more apparent in Scenario 2 than in Scenario 1. The main reason is because the transformation functions in Scenario 2 are logarithmic functions, which are more skewed than the cubic transformations in Scenario 1. As discussed in Section 4.5.3, the naive method can be viewed as imputing latent variables by linear combinations of observed variables. When the transformation is non-linear, the naive method has larger mis-specification error, which gives the our method more advantage.

Scenario 3 involves more types of variables. Since it involves ordinal variables with three levels, the naive method's regression coefficients have different dimensions. Thus, we cannot compare their coefficient estimation errors and the variable selection performance with ours. However, Figure 4.3 demonstrates that our method has much better prediction. The advantage is more apparent when the transformation functions are more skewed.

We further compare our method with another two commonly-used naive methods in the following scenario, where we set $n = 3000$, $n_{test} = 6000$, and $p = 100$.

Scenario 4: Let $\mathbf{B}_1 = \text{diag}(\mathbf{D}, \mathbf{I}_{18})$ and $\boldsymbol{\Sigma} = \text{diag}(\mathbf{B}_1, \mathbf{B}_1, \mathbf{B}_1, \mathbf{B}_1, \mathbf{B}_1)$, where $\mathbf{D} = (d_{ij})$, $d_{ii} = 1$ for $i = 1, ..., 2$, $d_{ij} = 0.8$ for $1 \leq i \neq j \leq 2$, and $\mathbf{I}_{18}$ is an eighteen-dimensional identity matrix. We generate $\mathbf{f}(\mathbf{z})$ from $N(\mathbf{0}, \mathbf{\Sigma})$ and choose $\mathrm{f}_j(Z_j) = \log(Z_j)$. The observed variables are generated by

$$X_j = \begin{cases} \sum_{k=1}^{2} I(Z_j > C_{jk}), & \text{for } j = 1 + 20t, \text{ and } 0 \leq t \leq p/20 - 1; \\ Z_j, & \text{otherwise,} \end{cases}$$

where $C_{j1} = \exp(-0.5)$ and $C_{j2} = \exp(0.5)$ for $j = 1 + 20t$, and $0 \leq t \leq p/20 - 1$. The response is generated by $Y = \sum_{j=0}^{4} 10\mathrm{f}_{1+20j}(Z_{1+20j}) + \epsilon$, where $\epsilon \sim N(0, 1)$.

We consider two naive methods. The first one (Naive-1) treats the ordinal variables as if they are continuous. The second one (Naive-2) creates dichotomized dummy variables for ordinal variables. Then, they both apply inverse normal transformation to all continuous variables as shown in (4.3.3) and run the regression. These two naive methods are widely used when handling ordinal variables in a linear regression problem. We run 100 simulations for Scenario 4 and compare the MSPEs given by our, Oracle-beta, and these two naive methods; see Figure 4.4. As been discussed in Section 4.5.3, the differences between the two naive methods and the Oracle-beta method reflect the model misspecification errors of the naive methods in imputing the latent variables. The difference between our and the Oracle-beta methods reflects the estimation errors of regression coefficients. Since the model misspecification errors of the naive methods are larger than the estimation errors of our method, we have much better prediction than the two naive methods in this scenario. Figure 4.4 also reveals that dichotomizing ordinal variables or treating them as continuous gives similar prediction performance.

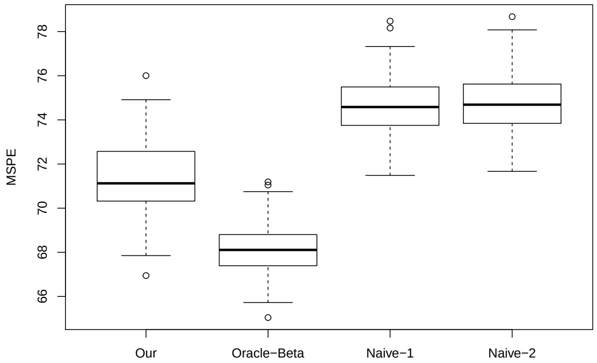


Figure 4.4. MSPEs for the four competitors in Scenario 4.

4.7 Real Data Analysis

The Communities and Crime Unnormalized Data from the UCI Machine Learning Repository contains records for 2215 communities with 147 variables about socio-economic information from the 1990 census and other surveys. We propose to predict the number of theft-related crimes, including larcenies, burglaries and auto-thefts, using the community characteristics.

We removed all variables that have missing values and screened out highly correlated variables. After screening, there are 34 variables left including one ordinal variable with three-levels (urbanization level of a community), one truncated variable (the number of people in homeless shelters), and 32 continuous variables.

We use these them to predict the number of three theft-related crimes: larcenies, burglaries and auto-thefts. A natural log transformation is applied to the three outcomes before further analysis. There are 2212 samples for all three responses. We compare our method with the Naive-LASSO in predicting the three crimes. For both methods, the optimal tuning parameters are chosen by five-fold cross-validation. We randomly split the full data into a training dataset with 1000 samples and a test set with the remaining samples. We repeat the random split 100 times. The MSPEs of these two methods are shown in Figure 4.5. It is seen that our method has much better prediction.

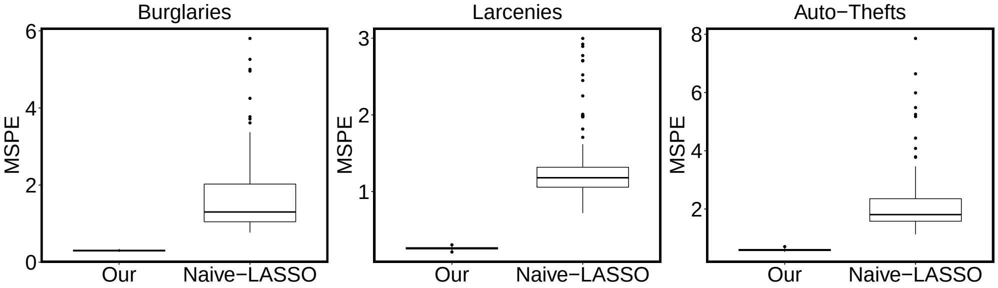


Figure 4.5. Prediction of Theft-Related Crime Responses by our and Naive-LASSO methods.

Finally, we investigate the most frequently selected covariates by both methods; see Figure 4.6 in Section 4.8. In all runs for the three kinds of crimes, our method selects land area, population density, inter-quartile range of housing rent, the number of people in homeless shelters and the percentage of kids born to never married. These selected variables are meaningful for all theft-related crimes. Larger land area and higher population density can have higher number of thefts. The inter-quartile range of housing rent can be an indicator of social disparity. More people in homeless shelters and higher percentage of kids born to never married can relate to increased number of crimes. On the contrary, the variables selected by the Naive-LASSO method vary in different runs and the frequencies for the meaningful variables are relatively low.

4.8 Technical Details

4.8.1 Bridge Functions and Estimation of $\mathbf{\Sigma}$

The pairwise bridge functions for binary, continuous and truncated variables were given in Fan et al. (2017); Yoon et al. (2020); Feng and Ning (2019), and summarized in below.

$$\widehat{F}_{jk}(r) = \begin{cases} 2\sin^{-1}(r)/\pi, & \text{for } j \in \mathcal{C},\ k \in \mathcal{C}; \\ 2(\Phi_2(\widehat{\Delta}_j, \widehat{\Delta}_k, r) - \Phi(\widehat{\Delta}_j)\Phi(\widehat{\Delta}_k)), & \text{for } j \in \mathcal{B},\ k \in \mathcal{B}; \\ 4\Phi_2(\widehat{\Delta}_k, 0, r/\sqrt{2}) - 2\Phi(\widehat{\Delta}_k), & \text{for } j \in \mathcal{C},\ k \in \mathcal{B}; \\ 2(1 - \Phi(\widehat{\Delta}_j))\Phi(\widehat{\Delta}_k) - 2\Phi_3(-\widehat{\Delta}_j, \widehat{\Delta}_k, 0; \mathbf{\Sigma}_{3a}) & \\ \quad - 2\Phi_3(-\widehat{\Delta}_j, \widehat{\Delta}_k, 0; \mathbf{\Sigma}_{3b}), & \text{for } j \in \mathcal{T},\ k \in \mathcal{B}; \\ -2\Phi_2(-\widehat{\Delta}_j, 0; 1/\sqrt{2}) + 4\Phi_3(-\widehat{\Delta}_j, 0, 0; \mathbf{\Sigma}_3), & \text{for } j \in \mathcal{T},\ k \in \mathcal{C}; \\ -2\Phi_4(-\widehat{\Delta}_j, -\widehat{\Delta}_k, 0, 0; \mathbf{\Sigma}_{4a}) + 2\Phi_4(-\widehat{\Delta}_j, -\widehat{\Delta}_k, 0, 0; \mathbf{\Sigma}_{4b}), & \text{for } j \in \mathcal{T},\ k \in \mathcal{T}. \end{cases}$$

$$\mathbf{\Sigma}_{3a} = \begin{bmatrix} 1 & -r & 1/\sqrt{2} \\ -r & 1 & -r/\sqrt{2} \\ 1/\sqrt{2} & -r/\sqrt{2} & 1 \end{bmatrix}, \mathbf{\Sigma}_{3b} = \begin{bmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} \\ -1/\sqrt{2} & -r/\sqrt{2} & 1 \end{bmatrix},$$

$$\mathbf{\Sigma}_3 = \begin{bmatrix} 1 & 1/\sqrt{2} & r/\sqrt{2} \\ 1/\sqrt{2} & 1 & r \\ r/\sqrt{2} & r & 1 \end{bmatrix},$$

$$\mathbf{\Sigma}_{4a} = \begin{bmatrix} 1 & 0 & 1/\sqrt{2} & -r/\sqrt{2} \\ 0 & 1 & -r/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -r/\sqrt{2} & 1 & -r \\ -r/\sqrt{2} & 1/\sqrt{2} & -r & 1 \end{bmatrix}, \mathbf{\Sigma}_{4b} = \begin{bmatrix} 1 & r & 1/\sqrt{2} & r/\sqrt{2} \\ r & 1 & r/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & r/\sqrt{2} & 1 & r \\ r/\sqrt{2} & 1/\sqrt{2} & r & 1 \end{bmatrix},$$

where Φ_d is the cumulative distribution function of the d-dimensional standard normal distribution, and

$$\widehat{\Delta}_j = \begin{cases} \Phi^{-1}(1 - (1/n)\sum_{i=1}^n X_{ij}), & \text{for } j \in \mathcal{B}, \\ \Phi^{-1}(1 - (1/n)\sum_{i=1}^n I(X_{ij} > 0)), & \text{for } j \in \mathcal{T}. \end{cases}$$

Lemma 4.1. *(Uniform Convergence for $\widetilde{\mathbf{\Sigma}}$ (Fan et al. 2017; Yoon et al. 2020; Feng and Ning 2019)) Suppose Conditions 3 and 4 hold. It follow that*

$$P(\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{max} \geq C\sqrt{\log p/n}) \leq C_1 p^{2-C_2C},$$

where C_1 and C_2 are generic positive constants and C is a sufficiently large constant.

4.8.2 Mean Squared Imputation Error of $\tilde{\mathbf{u}}$

Theorem 4.6. *For $\mathcal{I} = \mathcal{C} \cup \{j\}$, it holds that*

$$MSIE(\widetilde{u}_j; \mathcal{I}) = \begin{cases} 0, & \textit{for } j \in \mathcal{C}; \\ (1-\xi_j)(1 - \mathrm{E}_{L_j}(\frac{\phi(L_j)^2}{\Phi(L_j)(1-\Phi(L_j))})), & \textit{for } j \in \mathcal{B}; \\ (1-\xi_j)(1 - \sum_{k=0}^{N_j} \mathrm{E}(\frac{(\phi(L_{jk}) - \phi(L_{j(k+1)}))^2}{\Phi(L_{j(k+1)}) - \Phi(L_{jk})})), & \textit{for } j \in \mathcal{O}; \\ \Phi(\Delta_j)\xi_j(1-\xi_j) + (1-\xi_j)^2(\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)) & \\ \quad - (1-\xi_j)\mathrm{E}_{L_j}(\dfrac{\phi(L_j)^2}{\Phi(L_j)}), & \textit{for } j \in \mathcal{T}, \end{cases}$$

where $L_j \sim N(\Delta_j/(1-\widetilde{\xi}_j)^{1/2}, \widetilde{\xi}_j/(1-\widetilde{\xi}_j))$ for $j \in \mathcal{B} \cup \mathcal{T}$,
and $L_{jk} \sim N(\Delta_j^{(k)}/(1-\widetilde{\xi}_j)^{1/2}, \widetilde{\xi}_j/(1-\widetilde{\xi}_j))$ for $j \in \mathcal{O}$, and the expectation is taken with respect to L_{jk} and $L_{j(k+1)}$.

Proof of Theorem 4.6. For $j \in \mathcal{C}$, since $\widetilde{u}_j = \mathrm{f}_j(X_j)$ and $X_j = Z_j$, we have $MSIE(\widetilde{u}_j) = \mathrm{E}(\widetilde{u}_j - \mathrm{f}_j(Z_j))^2 = \mathrm{E}(\mathrm{f}_j(X_j) - \mathrm{f}_j(Z_j))^2 = 0$. We derive the the MSIEs for ordinal and truncated variables. The MSIE for binary variables will be derived in the proof of Proposition 4.1.

For $j \in \mathcal{O}$, we have

$$\widetilde{u}_j = \sqrt{1-\xi_j}\left\{\sum_{k=0}^{N_j} I(X_j = k)\frac{\phi(L_{jk}) - \phi(L_{j(k+1)})}{\Phi(L_{j(k+1)}) - \Phi(L_{jk})}\right\} + \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}) = A + \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}),$$

where A denotes the first summand. Hence,

$$MSIE(\widetilde{u}_j;\mathcal{I}) = \mathrm{E}(A^2) + \mathrm{E}((\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}) - \mathrm{f}_j(Z_j))^2) + 2\mathrm{E}(A(\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}) - \mathrm{f}_j(Z_j))),$$

where $\mathrm{E}(A^2) = (1-\xi_j)\sum_{0\le k\le N_j}\mathrm{E}((\phi(L_{jk}) - \phi(L_{j(k+1)}))^2/(\Phi(L_{j(k+1)}) - \Phi(L_{jk})))$ and

$$\mathrm{E}(A(\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}) - \mathrm{f}_j(Z_j))) = -(1-\xi_j)\sum_{k=0}^{N_j}\mathrm{E}\left(\frac{(\phi(L_{jk}) - \phi(L_{j(k+1)}))^2}{\Phi(L_{j(k+1)}) - \Phi(L_{jk})}\right).$$

Then, we have

$$MSIE(\widetilde{u}_j;\mathcal{I}) = (1-\xi_j)\left(1 - \sum_{k=0}^{N_j}\mathrm{E}\left(\frac{(\phi(L_{jk}) - \phi(L_{j(k+1)}))^2}{\Phi(L_{j(k+1)}) - \Phi(L_{jk})}\right)\right),\ for\ j \in \mathcal{O}.$$

For $j \in \mathcal{T}$, we have

$$\widetilde{u}_j = \mathrm{f}_j(Z_j)I(X_j > 0) + \left(-\frac{\sqrt{1-\xi_j}\phi(L_j)}{\Phi(L_j)} + \boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}})\right)I(X_j = 0).$$

Then, its MSIE has

$$\begin{aligned}
MSIE(\widetilde{u}_j;\mathcal{I}) &= \mathrm{E}\left(I(X_j = 0)\left(-\frac{\sqrt{1-\xi_j}\phi(L_j)}{\Phi(L_j)} + \boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}) - \mathrm{f}_j(Z_j)\right)^2\right)\\
&= \mathrm{E}\left(I(Z_j \le C_j)\left(-\frac{\sqrt{1-\xi_j}\phi(L_j)}{\Phi(L_j)} + \boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}) - \mathrm{f}_j(Z_j)\right)^2\right)\\
&= \mathrm{E}(I(Z_j \le C_j)(\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}))^2) + \mathrm{E}(I(Z_j \le C_j)\mathrm{f}_j(Z_j)^2)\\
&- 2\mathrm{E}(I(Z_j \le C_j)\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}})\mathrm{f}_j(Z_j)) - (1-\xi_j)\mathrm{E}_{L_j}(\frac{\phi(L_j)^2}{\Phi(L_j)}).
\end{aligned}$$

By some algebra, we have

$$\begin{aligned}
&\mathrm{E}(I(Z_j \le C_j)(\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}}))^2) = \Phi(\Delta_j)\xi_j(1-\xi_j) + \xi_j^2(\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)),\\
&\mathrm{E}(I(Z_j \le C_j)\mathrm{f}_j(Z_j)^2) = \Phi(\Delta_j) - \Delta_j\phi(\Delta_j),\\
&\mathrm{E}(I(Z_j \le C_j)\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{z}_{\mathcal{C}})\mathrm{f}_j(Z_j)) = \xi_j(\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)).
\end{aligned}$$

Hence,

$$\begin{aligned}
MSIE(\widetilde{u}_j;\mathcal{I}) &= \Phi(\Delta_j)\xi_j(1-\xi_j) + \xi_j^2(\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)) \\
&\quad + (\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)) - 2\xi_j(\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)) - (1-\xi_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{\Phi(L_j)}\right) \\
&= \Phi(\Delta_j)\xi_j(1-\xi_j) + (1-\xi_j)^2(\Phi(\Delta_j) - \Delta_j\phi(\Delta_j)) - (1-\xi_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{\Phi(L_j)}\right).
\end{aligned}$$

□

4.8.3 Variable Selection Results for Real Data Analysis

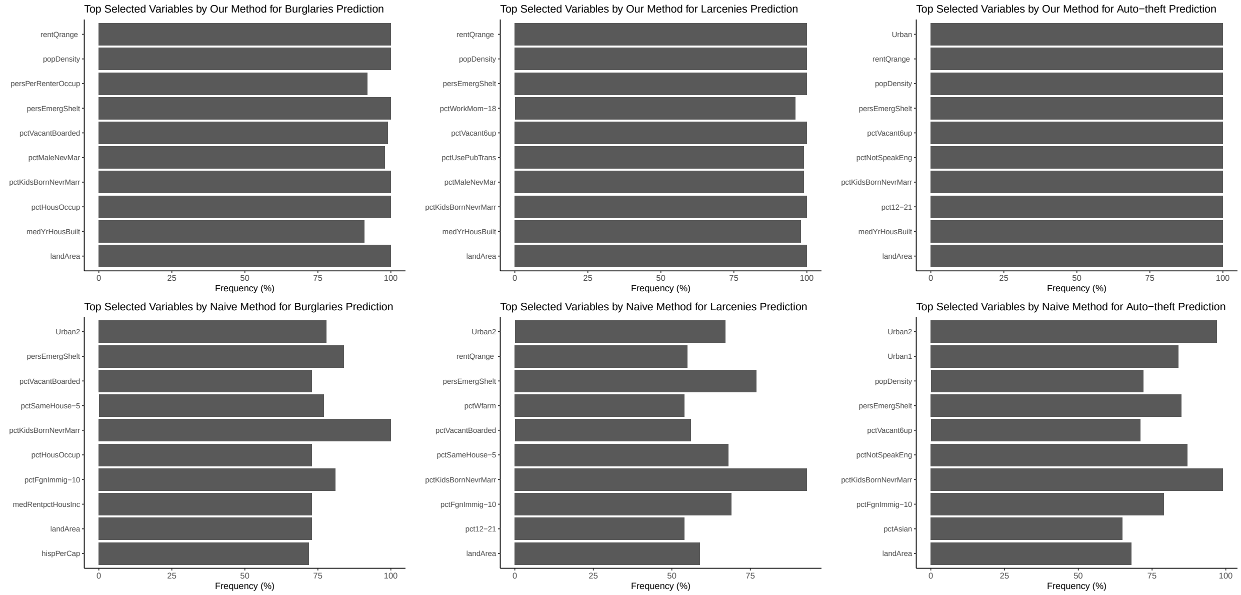


Figure 4.6. Most Frequently Selected Variables by Our and Naive Methods for Predicting Theft-Related Crimes.

A dictionary of variable names in Figure 4.6

hispPerCap: per capita income for people with hispanic heritage (continuous)

landArea: land area in square miles (continuous)

medRentpctHousInc: median gross rent as a percentage of household income (continuous)

medYrHousBuilt: median year housing units built (continuous)

pct12to21: percentage of population that is 12–21 in age (continuous)

pctAsian: percentage of population that is of asian heritage (continuous)

pctFgnImmig10: percentage of immigrants who immigated within last 10 years (continuous) pctHousOccup: percent of housing occupied (continuous)

pctKidsBornNevrMarr: percentage of kids born to never married (continuous)

pctMaleNevMar: percentage of males who have never married (continuous)

pctNotSpeakEng: percent of people who do not speak English well (continuous)

pctSameHouse5: percent of people living in the same house as in 1985 (continuous)

pctUsePubTrans: percent of people using public transit for commuting (continuous)

pctVacant6up: percent of vacant housing that has been vacant more than 6 months (continuous)

pctVacantBoarded: percent of vacant housing that is boarded up (continuous)

pctWfarm: percentage of households with farm or self employment income in 1989 (continuous)

pctWorkMom18: percentage of moms of kids under 18 in labor force (continuous)

persEmergShelt: number of people in homeless shelters (truncated)

persPerRenterOccup: mean persons per rental household (continuous)

popDensity: population density in persons per square mile (continuous)

rentQrange: rental housing - difference between upper quartile and lower quartile rent (continuous)

Urban: degree of urbanization (ordinal)

Urban1: dichotomized from Urban, indicating if Urban being 1 or 2

Urban2: dichotomized from Urban, indicating if Urban being 2

4.8.4 Proofs of Main Theorems

Proof of Theorem 4.1

Proof. For $j \in \mathcal{C}$, by definition, $X_j = Z_j$. Thus, $\mathrm{E}(\mathrm{f}_j(X_j)Y) = \mathrm{E}(\mathrm{f}_j(Z_j)Y) = \delta_j$.

For $j \in \mathcal{B}$, we have

$$\mathrm{E}(X_jY) = \mathrm{E}(X_j(\beta_0^* + \boldsymbol{\beta}^{*T}\mathbf{f}(\mathbf{z}))) = \beta_0^*(1 - \Phi(\Delta_j)) + \boldsymbol{\beta}^{*T}\mathrm{E}(X_j\mathbf{f}(\mathbf{z})).$$

By taking double expectation, we have

$$\mathrm{E}(X_j\mathrm{f}_i(Z_i)) = \mathrm{E}(X_j\mathrm{E}(\mathrm{f}_i(Z_i)|\mathrm{f}_j(Z_j))) = \mathrm{E}(X_j\mathrm{f}_j(Z_j))\Sigma_{ij} = \phi(\Delta_j)\Sigma_{ij}.$$

Hence $\mathrm{E}(X_j\mathbf{f}(\mathbf{z})) = \phi(\Delta_j)\boldsymbol{\Sigma}_j$, where $\boldsymbol{\Sigma}_j$ is the jth column of matrix $\boldsymbol{\Sigma}$. Since $\boldsymbol{\delta} = \mathrm{E}(\mathbf{f}(\mathbf{z})Y) = \boldsymbol{\Sigma}\boldsymbol{\beta}^*$, we have $\mathrm{E}(X_jY) = \beta_0^*(1 - \Phi(\Delta_j)) + \phi(\Delta_j)\boldsymbol{\Sigma}_j^T\boldsymbol{\beta}^*$. Then, $\mathrm{E}(X_jY) = \beta_0^*(1 - \Phi(\Delta_j)) + \phi(\Delta_j)\delta_j$.

For $j \in \mathcal{T}$, since $X_j = Z_j I(Z_j > C_j)$, we have

$$\mathrm{E}(I(X_j > 0)\mathrm{f}_j(X_j)Y) = \mathrm{E}(I(Z_j > C_j)\mathrm{f}_j(Z_j)Y).$$

Then,

$$\begin{aligned}
\mathrm{E}(I(Z_j > C_j)\mathrm{f}_j(Z_j)Y) &= \mathrm{E}(I(Z_j > C_j)\mathrm{f}_j(Z_j))\beta_0^* + \mathrm{E}(I(Z_j > C_j)\mathrm{f}_j(Z_j)\mathbf{f}(\mathbf{z}))^T\boldsymbol{\beta}^* \\
&= \beta_0^*\phi(\Delta_j) + \mathrm{E}(I(Z_j > C_j)\mathrm{f}_j(Z_j)^2)\boldsymbol{\Sigma}_j^T\boldsymbol{\beta}^* \\
&= \beta_0^*\phi(\Delta_j) + \mathrm{E}(I(Z_j > C_j)\mathrm{f}_j(Z_j)^2)\delta_j \\
&= \beta_0^*\phi(\Delta_j) + C(\Delta_j)\delta_j,
\end{aligned}$$

where $C(\Delta_j) = \Delta_j\phi(\Delta_j) + 1 - \Phi(\Delta_j)$. □

Proof of Theorem 4.2

Proof. For $j \in \mathcal{C}$, $\widehat{\delta}_j = n^{-1}\sum_{i=1}^n \widehat{\mathrm{f}}_j(X_{ij})Y_i$, where $\widehat{\mathrm{f}}_j(t) = \Phi^{-1}(\widetilde{F}(t))$. Then, for any $t > 0$, it holds that

$$\begin{aligned}
P\left(\left|\widehat{\delta}_j - \delta_j\right| \geq t\right) &\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n \widehat{\mathrm{f}}_j(X_{ij})Y_i - \mathrm{E}(\mathrm{f}_j(Z_j)Y)\right| \geq t\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n \mathrm{f}_j(Z_{ij})\mathbf{f}(\mathbf{z}_i)^T\boldsymbol{\beta}^* - \mathrm{E}(\mathrm{f}_j(Z_j)\mathbf{f}(\mathbf{z})^T)\boldsymbol{\beta}^*\right| \geq t/4\right) \\
&+ P\left(\left|\frac{1}{n}\sum_{i=1}^n \mathrm{f}_j(Z_{ij})\epsilon_i\right| \geq t/4\right) \\
&+ P\left(\left|\frac{1}{n}\sum_{i=1}^n (\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))\mathbf{f}(\mathbf{z}_i)^T\boldsymbol{\beta}^*\right| \geq t/4\right) \\
&+ P\left(\left|\frac{1}{n}\sum_{i=1}^n (\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))\epsilon_i\right| \geq t/4\right) \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}$$

For I, by Condition 2, we have

$$
\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(Z_{ij})\mathbf{f}(\mathbf{z}_i)^T\boldsymbol{\beta}^* - \mathrm{E}(\mathrm{f}_j(Z_j)\mathbf{f}(\mathbf{z})^T)\boldsymbol{\beta}^*\right| \geq t/4\right)\\
&\leq P\left(\max_{1\leq k\leq p_1}\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(Z_{ij})\mathrm{f}_k(Z_{ik}) - \mathrm{E}(\mathrm{f}_j(Z_j)\mathrm{f}_k(Z_k))\right|\cdot\|\boldsymbol{\beta}^*\|_1 \geq t/4\right)\\
&\leq P\left(\max_{1\leq k\leq p_1}\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(Z_{ij})\mathrm{f}_k(Z_{ik}) - \mathrm{E}(\mathrm{f}_j(Z_j)\mathrm{f}_k(Z_k))\right| \geq t/(4M)\right)\\
&\leq p_1 C_1 e^{-C_2 n t^2},
\end{aligned}
$$

where the last inequality follows from Lemma 3 in Bickel and Levina (2008b) and the union bound, and C_1 and C_2 are some generic positive constants. Then, letting $t = C(\log p_1)^{1/2}(\log n)^{1/4}n^{-1/4}$ for a sufficiently large constant C, we have $\mathrm{I} = o(p_1^{-1})$.

For II, it follows from Condition 1 that $\|\mathrm{f}_j(Z_{ij})\epsilon_i\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathrm{E}\,|\mathrm{f}_j(Z_{ij})\epsilon_i|^p)^{1/p} \leq 2M^2$. Thus, $\mathrm{f}_j(Z_j)\epsilon$ is sub-exponential. Then, it follows from the Bernstein inequality that

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(Z_{ij})\epsilon_i\right| \geq t/4\right) \leq 2\exp\left[-C^{''}\min(\frac{t^2}{64M^4}, \frac{t}{8M^2})n\right], \tag{4.8.1}
$$

where $C^{''}$ is a universal constant. Hence, letting $t = C(\log p_1)^{1/2}(\log n)^{1/4}n^{-1/4}$ for a sufficiently large constant C and $(\log p_1)(\log n)^{1/2}n^{-1/2} \to 0$, we have $\mathrm{II} = o(p_1^{-1})$.

For IV, it follows from the Cauchy-Schwarz inequality that

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))\epsilon_i\right| \geq t/4\right) \leq P\left(\left|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}\right| \geq t/4\right).
$$

By Condition 1 and Bernstein inequality, we have

$$
P\left(\left|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2} - \sigma\right| > \frac{C}{n^{1/4}}\right) \leq P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^2\right| > \frac{C\sigma}{n^{1/4}}\right) = o(n^{-1/2}),
$$

for some constant C. Then, for sufficiently large n, by the law of total probability,

$$
\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))\epsilon_i\right| \geq t/4\right) \\
&\leq P\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))^2} \geq t/(8\sigma)\right) + o(n^{-1/2}).
\end{aligned}
$$

Letting $t = C(\log p_1)^{1/2}(\log n)^{1/4}n^{-1/4}$ with a sufficiently large constant C, it follows from Lemma 4.3 that $\mathrm{IV} = o(1)$. Similarly, we can show that under the same choice of t, $\mathrm{III} = o(1)$.

Then, by the union bound, for any $t > 0$, we have

$$
\begin{aligned}
P(\|\widehat{\boldsymbol{\delta}}_{\mathcal{C}} - \boldsymbol{\delta}_{\mathcal{C}}\|_\infty \geq t) \leq\ & p_1 P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(Z_{ij})\mathbf{f}(\mathbf{z}_i)^T\boldsymbol{\beta}^* - \mathrm{E}(\mathrm{f}_j(Z_j)\mathbf{f}(\mathbf{z})^T)\boldsymbol{\beta}^*\right| \geq t/4\right) \\
&+ p_1 P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(Z_{ij})\epsilon_i\right| \geq t/4\right) \\
&+ P\left(\max_{j\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))\mathbf{f}(\mathbf{z}_i)^T\boldsymbol{\beta}^*\right| \geq t/4\right) \\
&+ P\left(\max_{j\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(Z_{ij}) - \mathrm{f}_j(Z_{ij}))\epsilon_i\right| \geq t/4\right).
\end{aligned}
$$

By choosing $t = C(\log p_1)^{1/2}(\log n)^{1/4}n^{-1/4}$ and noting that $(\log p_1)(\log n)^{1/2}n^{-1/2} \to 0$, we have $\|\widehat{\boldsymbol{\delta}}_{\mathcal{C}} - \boldsymbol{\delta}_{\mathcal{C}}\|_\infty = O_p((\log p_1)^{1/2}(\log n)^{1/4}n^{-1/4})$.

For $j \in \mathcal{B}$, $\widehat{\delta}_j = \phi(\widehat{\Delta}_j)^{-1}(1/n)\sum_{1\leq i\leq n} X_{ij}Y_i$, where $\widehat{\Delta}_j = \Phi^{-1}(1 - (1/n)\sum_{1\leq i\leq n} X_{ij})$. Then, for any $t > 0$,

$$
\begin{aligned}
P\left(\left|\widehat{\delta}_j - \delta_j\right| \geq t\right) \leq\ & P\left(\left|(\sqrt{2\pi}e^{\frac{\widehat{\Delta}_j}{2}} - \sqrt{2\pi}e^{\frac{\Delta_j}{2}})\mathrm{E}(X_jY)\right| \geq t/3\right) \\
&+ P\left(\left|\sqrt{2\pi}e^{\frac{\Delta_j}{2}}(\frac{1}{n}\sum_{i=1}^{n}X_{ij}Y_i - \mathrm{E}(X_jY))\right| \geq t/3\right) \\
&+ P\left(\left|(\sqrt{2\pi}e^{\frac{\widehat{\Delta}_j}{2}} - \sqrt{2\pi}e^{\frac{\Delta_j}{2}})(\frac{1}{n}\sum_{i=1}^{n}X_{ij}Y_i - \mathrm{E}(X_jY))\right| \geq t/3\right) \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}
$$

For I, by Conditions 2 and 4,

$$P\left(\left|(\sqrt{2\pi}e^{\frac{\widehat{\Delta}_j}{2}}-\sqrt{2\pi}e^{\frac{\Delta_j}{2}})\mathrm{E}(X_jY)\right|\geq t/3\right)\leq P\left(\left|e^{\frac{\widehat{\Delta}_j}{2}}-e^{\frac{\Delta_j}{2}}\right|\geq t/(3M)\right).$$

By Lemma A.1 in Fan et al. (2017), the function $\Phi^{-1}(y)$ is Lipschitz continuous for $y\in(\Phi(-2M),\Phi(2M))$. Given the event $A_j=\{|\widehat{\Delta}_j|\leq 2M\}$, we have

$$\left|\widehat{\Delta}_j-\Delta_j\right|\leq L_1\left|\sum_{i=1}^n X_{ij}-(1-\Phi(\Delta_j))\right|=L_1\left|(1/n)\sum_{i=1}^n X_{ij}-\mathrm{E}(X_j)\right|.$$

Then, we have

$$\begin{aligned}
&P(A_j^c)=P(|\widehat{\Delta}_j|>2M)\\
&=P\left(1-\frac{1}{n}\sum_{i=1}^n X_{ij}<\Phi(-2M)\text{ or }1-\frac{1}{n}\sum_{i=1}^n X_{ij}>\Phi(2M)\right)\\
&=P\left(\frac{1}{n}\sum_{i=1}^n X_{ij}-(1-\Phi(\Delta_j))>\Phi(\Delta_j)-\Phi(-2M)\right.\\
&\qquad\left.\text{or }\frac{1}{n}\sum_{i=1}^n X_{ij}-(1-\Phi(\Delta_j))<\Phi(\Delta_j)-\Phi(2M)\right)\\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{ij}-(1-\Phi(\Delta_j))\right|\geq\Phi(2M)-\Phi(M)\right)\\
&\leq 2\exp\left(-\frac{n}{2}(\Phi(2M)-\Phi(M))^2\right),
\end{aligned}$$

where the second to the last inequality follows from Condition 3 and the last inequality follows from the Hoeffding inequality. Furthermore, by the Hoeffding inequality and Lipschitz continuity of $\exp(x/2)$ for $x\in(-2M,2M)$, we have, for any $t>0$,

$$\begin{aligned}
&P\left(\left|e^{\widehat{\Delta}_j/2}-e^{\Delta_j/2}\right|\geq t\right)\\
&\leq P\left(\left|e^{\widehat{\Delta}_j/2}-e^{\Delta_j/2}\right|\geq t|A_j\right)+P(A_j^c)\\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{ij}-(1-\Phi(\Delta_j))\right|\geq t/(L_1L_2)\right)+2\exp\left(-\frac{n}{2}(\Phi(2M)-\Phi(M))^2\right)\\
&\leq 2\exp\left(-\frac{nt^2}{L_1^2L_2^2}\right)+2\exp\left(-\frac{n}{2}(\Phi(2M)-\Phi(M))^2\right).
\end{aligned}$$

Letting $t = C\sqrt{\log p_2/n}$ for some sufficiently large constant C, we have $\mathrm{I} = o(p_2^{-1})$.

For II, by using similar arguments as in (4.8.1), we can show that $\mathrm{II} = o(p_2^{-1})$, when $t = C\sqrt{\log p_2/n}$ for some sufficiently large constant C. For III, it is a higher-order term dominated by I and II. Then, the union bound gives $\|\widehat{\boldsymbol{\delta}}_{\mathcal{B}} - \boldsymbol{\delta}_{\mathcal{B}}\|_\infty = O_p(\sqrt{\log p_2/n})$.

For $j \in \mathcal{O}$, given the result that $\|\widehat{\boldsymbol{\delta}}_{\mathcal{B}} - \boldsymbol{\delta}_{\mathcal{B}}\|_\infty = O_p(\sqrt{(\log p_2)/n})$, the ensemble method guarantees that $\|\widehat{\boldsymbol{\delta}}_{\mathcal{O}} - \boldsymbol{\delta}_{\mathcal{O}}\|_\infty = O_p(\sqrt{(\log p_3)/n})$. Details can be found in the supplementary materials of Feng and Ning (2019).

For $j \in \mathcal{T}$, we have

$$\begin{aligned}\left|\widehat{\delta}_j - \delta_j\right| \lesssim{}& \left|(C(\widehat{\Delta}_j)^{-1} - C(\Delta_j)^{-1})\mathrm{E}(\mathrm{f}_j(X_j)I(X_j > 0)Y)\right| \\ &+ \left|C(\Delta_j)^{-1}\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(X_{ij}) - \mathrm{f}_j(X_{ij}))I(X_{ij} > 0)Y_i\right| \\ &+ \left|C(\Delta_j)^{-1}(\frac{1}{n}\sum_{i=1}^{n}\mathrm{f}_j(X_{ij})I(X_{ij} > 0)Y_i - \mathrm{E}(\mathrm{f}_j(X_j)I(X_j > 0)Y))\right|.\end{aligned}$$

Using similar arguments as in Lemma 4.3, $\max_{j\in\mathcal{T}} n^{-1}\sum_{i=1}^{n}(\widehat{\mathrm{f}}_j(X_{ij}) - \mathrm{f}_j(X_{ij}))^2 I(X_{ij} > 0) = O_p((\log p_4)(\log n)^{1/2}n^{-1/2})$. The rest of the proof follows similar arguments as for $j \in \mathcal{C}$ and $j \in \mathcal{B}$. □

Proof of Theorem 4.3

Proof. We rely on the general M-estimator theory (Negahban et al. 2012) to prove the results. First, we verify that the restrictive strong convexity (RSC) condition holds with high probability. Let $L(\boldsymbol{\beta}) = (1/2)\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\delta}}^T\boldsymbol{\beta}$. We have

$$\begin{aligned}\delta L(\boldsymbol{\Delta}, \boldsymbol{\beta}^*) &= L(\boldsymbol{\beta}^* + \boldsymbol{\Delta}) - L(\boldsymbol{\beta}^*) - \nabla L(\boldsymbol{\beta}^*)^T\boldsymbol{\Delta} \\ &= \frac{1}{2}\boldsymbol{\Delta}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta},\end{aligned}$$

where $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. To verify the RSC condition, it suffices to show that $\delta L(\boldsymbol{\Delta}, \boldsymbol{\beta}^*)/\|\boldsymbol{\Delta}\|_2^2$ is bounded away from 0 for all $\boldsymbol{\Delta} \in \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}_{\mathcal{M}^c}\|_1 \le 3\|\boldsymbol{\Delta}_{\mathcal{M}}\|_1\}$. Under Conditions 3 and 4, we have

$$P(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{max} \ge C\sqrt{\frac{\log p}{n}}) \le C_1 p^{2-C_2 C},$$

for some sufficiently large positive constant C and generic positive constants C_1 and C_2. Then, with probability greater than $1 - C_1 p^{2-C_2 C}$, it holds that

$$\begin{aligned}\frac{\delta L(\mathbf{\Delta}, \boldsymbol{\beta}^*)}{\|\mathbf{\Delta}\|_2^2} = \frac{\mathbf{\Delta}^T \widehat{\mathbf{\Sigma}} \mathbf{\Delta}}{2\|\mathbf{\Delta}\|_2^2} &= \frac{\mathbf{\Delta}^T \mathbf{\Sigma} \mathbf{\Delta} + \mathbf{\Delta}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{\Delta}}{2\|\mathbf{\Delta}\|_2^2} \\ &\geq \frac{m}{2} - \frac{\|\mathbf{\Delta}\|_1^2 C\sqrt{(\log p)/n}}{2\|\mathbf{\Delta}\|_2^2} \quad (\mathit{Condition}\ 5) \\ &\geq \frac{1}{2}\left(m - C\sqrt{(\log p)/n}\left(\frac{4\|\mathbf{\Delta}_{\mathcal{M}}\|_1}{\|\mathbf{\Delta}_{\mathcal{M}}\|_2}\right)^2\right) \\ &\geq \frac{1}{2}(m - Cs\sqrt{(\log p)/n}) \geq \frac{m}{4}.\end{aligned}$$

Next, by Lemma 4.5 and Theorem 2, we have $\|\widehat{\mathbf{\Sigma}}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\delta}}\|_\infty = O_P\left((\log p)^{1/2}(\log n)^{1/4}n^{-1/4}\right)$. Then, by choosing $\lambda = C(\log p)^{1/2}(\log n)^{1/4}n^{-1/4}$ for some sufficiently large C, it follows from Lemma 4.4 that Theorem 4.3 holds. □

Proof of Theorem 4.4

Proof. By the standard convex optimization theory, any $\boldsymbol{\beta} \in \mathbf{R}^p$ satisfying the following Karush–Kuhn–Tucker conditions (Boyd and Vandenberghe 2004) is the solution to (4.3.1).

$$(\widehat{\mathbf{\Sigma}}\boldsymbol{\beta})_j - \widehat{\boldsymbol{\delta}}_j + \lambda \mathrm{sign}(\beta_j) = 0,\ for\ j \in \mathcal{M}; \tag{4.8.2}$$

$$\left|(\widehat{\mathbf{\Sigma}}\boldsymbol{\beta})_j - \widehat{\boldsymbol{\delta}}_j\right| < \lambda,\ for\ j \notin \mathcal{M}; \tag{4.8.3}$$

$$\lambda_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}) > 0. \tag{4.8.4}$$

We first show that there exists a solution $\widehat{\boldsymbol{\beta}}_{\mathcal{M}} \in \mathbf{R}^s$ to (4.8.2) in the neighbourhood $\mathcal{N} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\mathcal{M}}\|_\infty \leq C\lambda\}$ with probability tending to 1. We have

$$(\widehat{\mathbf{\Sigma}}\boldsymbol{\beta})_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} = \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}}.$$

It follows from Lemma 4.5 and Theorem 4.2 that

$$\begin{aligned}
P(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}}\|_\infty \geq Ca_n) &\leq P(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \boldsymbol{\delta}_{\mathcal{M}}\|_\infty \geq Ca_n) \\
&\quad + P(\|\widehat{\boldsymbol{\delta}}_{\mathcal{M}} - \boldsymbol{\delta}_{\mathcal{M}}\|_\infty \geq Ca_n) \\
&= o(1),
\end{aligned}$$

where $a_n = (\log p)^{1/2}(\log n)^{1/4}n^{-1/4}$. Let $\boldsymbol{\tau} = (\tau_j) \in \mathbf{R}^p$ with $\tau_j = \mathrm{sign}(\beta_j)$ for $j \in \mathcal{M}$ and $\tau_j = 0$ for $j \notin \mathcal{M}$,

$$\begin{aligned}
f(\boldsymbol{\beta}_{\mathcal{M}}) &= \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}, \\
g(\boldsymbol{\beta}_{\mathcal{M}}) &= \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}f(\boldsymbol{\beta}_{\mathcal{M}}) = \boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}} + \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}\}.
\end{aligned}$$

By Lemma 4.6, for some sufficiently large constant C, we have

$$P(\|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty \geq 2M) \leq C_1 p^{2-C_2C}.$$

Hence, by the stated choice of λ, with probability tending to 1, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}\}\|_\infty &\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}\}\|_\infty \\
&\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}}\|_\infty + \lambda) \\
&\leq 2M(Ca_n + \lambda) \lesssim \lambda.
\end{aligned}$$

Hence, when n is sufficiently large, if $(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}})_j = C\lambda$ for some sufficently large $C > 0$,

$$g(\boldsymbol{\beta}_{\mathcal{M}})_j = C\lambda - \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}})_j \geq 0,$$

and if $(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}})_j = -C\lambda$,

$$g(\boldsymbol{\beta}_{\mathcal{M}})_j = -C\lambda - \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}})_j \leq 0.$$

By the continuity of $g(\boldsymbol{\beta}_{\mathcal{M}})$ and Miranda's existence theorem, $g(\boldsymbol{\beta}_{\mathcal{M}}) = 0$ has a solution $\widehat{\boldsymbol{\beta}}_{\mathcal{M}}$ in $\mathcal{N}$

with probability tending to 1.

Second, we verify that $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathbf{0})^T$ also satisfies (4.8.3). We have

$$(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta})_{\mathcal{M}^c} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}^c} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{M^c} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + (\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\delta}})_{\mathcal{M}^c}.$$

Since $g(\boldsymbol{\beta}_{\mathcal{M}}) = 0$, we have

$$(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta})_{\mathcal{M}^c} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}^c} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}) + (\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\delta}})_{\mathcal{M}^c}.$$

By similar arguments as in Lemma 4.5, we have

$$P(\|(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\delta}})_{\mathcal{M}^c}\|_\infty \geq Ca_n) = o(1).$$

By Lemma 4.7, we have

$$P(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty \geq (1-\alpha)(1-\epsilon/2)) \leq C_1 p^{2-C_2C}.$$

Then, with probability tending to 1,

$$\begin{aligned} \|(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta})_{\mathcal{M}^c} - \widehat{\boldsymbol{\delta}}_{\mathcal{M}^c}\|_\infty &\leq (1-\alpha)(1-\epsilon/2)(Ca_n + \lambda) + Ca_n \\ &\leq (1-\alpha)(1-\epsilon/2)\lambda + (2-\epsilon/2)Ca_n \\ &< (1-\alpha)\lambda, \end{aligned}$$

where the last inequality is due to $a_n = o(\lambda)$. This verifies (4.8.3). Finally, to verify (4.8.4), Condition 5 implies that $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}}) \geq m$. Then by a similar proof, we can show that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}) \geq m/2$ with probability tending to 1. □

Proof of Proposition 4.1

Proof. For $j \in \mathcal{B}$, denote $\widetilde{\xi}_j = \mathbf{\Sigma}^T_{\widetilde{\mathcal{C}}j}\mathbf{\Sigma}^{-1}_{\widetilde{\mathcal{C}}\widetilde{\mathcal{C}}}\mathbf{\Sigma}_{\widetilde{\mathcal{C}}j}$, $\widetilde{\boldsymbol{\eta}}_j = \mathbf{\Sigma}^{-1}_{\widetilde{\mathcal{C}}\widetilde{\mathcal{C}}}\mathbf{\Sigma}_{\widetilde{\mathcal{C}}j}$, and

$$L_j = \frac{\Delta_j - \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})}{\sqrt{1-\widetilde{\xi}_j}}.$$

We first calculate $\mathrm{E}(\mathrm{f}_j(Z_j)|x_j, \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}))$. We have $\mathrm{f}_j(Z_j)|\mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) \sim N(\widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}), 1-\widetilde{\xi}_j)$. Hence, $\mathrm{f}_j(Z_j)|\mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) = \sqrt{1-\widetilde{\xi}_j}Z + \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})$, where $Z \sim N(0,1)$ is independent of Z_j and $\mathbf{z}_{\widetilde{\mathcal{C}}}$. Then, we have

$$\begin{aligned}\mathrm{E}(\mathrm{f}_j(Z_j)|X_j, \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) &= \mathrm{E}\left(\sqrt{1-\widetilde{\xi}_j}Z + \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})|I(Z > L_j)\right)\\ &= \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) + \sqrt{1-\widetilde{\xi}_j}\mathrm{E}(Z|I(Z > L_j)).\end{aligned}$$

Since $\mathrm{E}(Z|Z > L_j) = \phi(L_j)/(1-\Phi(L_j))$ and $\mathrm{E}(Z|Z \le L_j) = -\phi(L_j)/\Phi(L_j)$, $X_j|\mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) = I(Z > L_j)$, we have

$$\mathrm{E}(\mathrm{f}_j(Z_j)|X_j, \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) = \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) + \sqrt{1-\widetilde{\xi}_j}(X_j\frac{\phi(L_j)}{1-\Phi(L_j)} - (1-X_j)\frac{\phi(L_j)}{\Phi(L_j)}).$$

Now we calculate $\mathrm{E}((\mathrm{E}(\mathrm{f}_j(Z_j)|X_j, \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) - \mathrm{f}_j(Z_j))^2)$. Let $A = \sqrt{1-\widetilde{\xi}_j}(X_j\phi(L_j)/(1-\Phi(L_j)) - (1-X_j)\phi(L_j)/(\Phi(L_j)))$. We have,

$$\begin{aligned}&\mathrm{E}((\mathrm{E}(\mathrm{f}_j(Z_j)|X_j, \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) - \mathrm{f}_j(Z_j))^2)\\ &= \mathrm{E}((A + \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) - \mathrm{f}_j(Z_j))^2)\\ &= \mathrm{E}(A^2) + \mathrm{E}((\widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) - \mathrm{f}_j(Z_j))^2) + 2\mathrm{E}(A(\widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) - \mathrm{f}_j(Z_j))).\end{aligned}$$

Since

$$\mathrm{E}(A^2) = (1-\widetilde{\xi})\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{1-\Phi(L_j)} + \frac{\phi(L_j)^2}{\Phi(L_j)}\right) = (1-\widetilde{\xi})E_{L_j}\left(\frac{\phi(L_j)^2}{(1-\Phi(L_j))\Phi(L_j)}\right),$$

and

$$\begin{aligned}
\mathrm{E}(A(\widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) - \mathrm{f}_j(Z_j))) &= -\mathrm{E}(A\mathrm{f}_j(Z_j)) \\
&= -(1-\widetilde{\xi}_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{1-\Phi(L_j)} + \frac{\phi(L_j)^2}{\Phi(L_j)}\right) \\
&= -(1-\widetilde{\xi}_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{(1-\Phi(L_j))\Phi(L_j)}\right),
\end{aligned}$$

then,

$$\begin{aligned}
MSIE(\widetilde{u}_j; \widetilde{\mathcal{C}} \cup \{j\}) &= \mathrm{E}((\mathrm{E}(\mathrm{f}_j(Z_j)|X_j, \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) - \mathrm{f}_j(Z_j))^2) \\
&= \mathrm{E}((\widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}) - \mathrm{f}_j(Z_j))^2) - (1-\widetilde{\xi}_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{(1-\Phi(L_j))\Phi(L_j)}\right) \\
&= \mathrm{E}((\mathrm{E}(\mathrm{f}_j(Z_j)|\mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) - \mathrm{f}_j(Z_j))^2) - (1-\widetilde{\xi}_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{(1-\Phi(L_j))\Phi(L_j)}\right) \\
&= MSIE(\widetilde{u}_j; \widetilde{\mathcal{C}}) - (1-\widetilde{\xi}_j)\mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{(1-\Phi(L_j))\Phi(L_j)}\right).
\end{aligned}$$

Since $1-\widetilde{\xi}_j$ and $E_{L_j}(\phi(L_j)^2/((1-\Phi(L_j))\Phi(L_j)))$ are both positive, we have

$$MSIE(\widetilde{u}_j; \widetilde{\mathcal{C}} \cup \{j\}) < MSIE(\widetilde{u}_j; \widetilde{\mathcal{C}}).$$

By the normality assumption, it follows that $MSIE(\widetilde{u}_j; \widetilde{\mathcal{C}}) = \mathrm{Var}(\mathrm{f}_j(Z_j)|\mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}})) = 1-\widetilde{\xi}_j$. Hence, we have

$$MSIE(\widetilde{u}_j; \widetilde{\mathcal{C}} \cup \{j\}) = (1-\widetilde{\xi}_j)\left[1 - \mathrm{E}_{L_j}\left(\frac{\phi(L_j)^2}{1-\Phi(L_j)\Phi(L_j)}\right)\right],$$

where $L_j = (\Delta_j - \widetilde{\boldsymbol{\eta}}_j^T \mathbf{f}_{\widetilde{\mathcal{C}}}(\mathbf{z}_{\widetilde{\mathcal{C}}}))/\sqrt{1-\widetilde{\xi}_j} \sim N(\Delta_j/\sqrt{1-\widetilde{\xi}_j}, \widetilde{\xi}_j/(1-\widetilde{\xi}_j))$. □

Proof of Theorem 4.5

Proof. We have

$$
\begin{aligned}
MSPE &= (\widehat{\beta}_0 - \beta_0^*)^2 + \boldsymbol{\beta}_{\mathcal{M}}^{*T} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{\mathbf{u}}_{i\mathcal{M}} - \widetilde{\mathbf{u}}_{i\mathcal{M}})^{\otimes 2} \right\} \boldsymbol{\beta}_{\mathcal{M}}^* \\
&+ \boldsymbol{\beta}_{\mathcal{M}}^{*T} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widetilde{\mathbf{u}}_{i\mathcal{M}} - \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{M}})^{\otimes 2} \right\} \boldsymbol{\beta}_{\mathcal{M}}^* \\
&+ (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}}^*)^T \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{M}}^{\otimes 2} \right\} (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}}^*) + R \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + R.
\end{aligned}
$$

For I, we have $(\widehat{\beta}_0 - \beta_0^*)^2 = O_p(1/n)$. For II, we have

$$
\mathrm{II} \leq \| \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{\mathbf{u}}_{i\mathcal{M}} - \widetilde{\mathbf{u}}_{i\mathcal{M}})^{\otimes 2} \|_{\max} \| \boldsymbol{\beta}_{\mathcal{M}}^* \|_1^2.
$$

For $\|(1/n_{test}) \sum_{1 \leq i \leq n_{test}} (\widehat{\mathbf{u}}_{i\mathcal{M}} - \widetilde{\mathbf{u}}_{i\mathcal{M}})^{\otimes 2} \|_{\max}$, we have

$$
\| \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{\mathbf{u}}_{i\mathcal{M}} - \widetilde{\mathbf{u}}_{i\mathcal{M}})^{\otimes 2} \|_{\max} \leq \max_{j \in \mathcal{M}} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2,
$$

and

$$
\max_{j \in \mathcal{M}} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 \leq \max_{j \in \mathcal{M} \cap \mathcal{C}} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 \vee \max_{j \in \mathcal{A}} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2,
$$

where $\mathcal{A} = \mathcal{M} \cap (\mathcal{B} \cup \mathcal{O} \cup \mathcal{T})$. By Lemmas 4.3 and 4.10, we have

$$
\begin{aligned}
\max_{j \in \mathcal{M} \cap \mathcal{C}} (1/n_{test}) \sum_{1 \leq i \leq n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 &= O_p((\log p)(\log n)^{1/2} n^{-1/2}), \\
\max_{j \in \mathcal{A}} (1/n_{test}) \sum_{1 \leq i \leq n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 &\leq \sum_{j \in \mathcal{A}} (1/n_{test}) \sum_{1 \leq i \leq n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 \\
&= O_p(s_{\mathcal{A}}((\log p)(\log n)^{1/2} n^{-1/2} \vee (\log p/n)^{1-qr})).
\end{aligned}
$$

Hence, we have

$$\mathrm{II} = \begin{cases} O_p((\log p)(\log n)^{1/2} n^{-1/2}), & \text{for } \mathcal{A} = \emptyset; \\ O_p(s_{\mathcal{A}}((\log p)(\log n)^{1/2} n^{-1/2} \vee (\log p/n)^{1-qr})), & \text{for } \mathcal{A} \neq \emptyset. \end{cases}$$

For III, we have

$$\boldsymbol{\beta}^{*T}_{\mathcal{M}} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widetilde{\mathbf{u}}_{i\mathcal{M}} - \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{M}})^{\otimes 2} \right\} \boldsymbol{\beta}^{*}_{\mathcal{M}} = \boldsymbol{\beta}^{*T}_{\mathcal{A}} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widetilde{\mathbf{u}}_{i\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{A}})^{\otimes 2} \right\} \boldsymbol{\beta}^{*}_{\mathcal{A}}.$$

Since $\widetilde{\mathbf{u}}_{\mathcal{C}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{C}} = \mathbf{0}$, we have

$$\begin{aligned} \mathrm{III} = {} & \boldsymbol{\beta}^{*T}_{\mathcal{A}} \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widetilde{\mathbf{u}}_{i\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{A}})^{\otimes 2} - \mathrm{E}\left((\widetilde{\mathbf{u}}_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}})^{\otimes 2} \right) \right\} \boldsymbol{\beta}^{*}_{\mathcal{A}} \\ & + \boldsymbol{\beta}^{*T}_{\mathcal{A}} \mathrm{E}\left((\widetilde{\mathbf{u}}_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}})^{\otimes 2} \right) \boldsymbol{\beta}^{*}_{\mathcal{A}}. \end{aligned}$$

It follows from Theorem 2.26 of Wainwright (2019) that elements of $\widetilde{\mathbf{u}}_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}}$ are sub-Gaussian. Then applying Theorem 6.5 of Wainwright (2019) gives that

$$\| \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widetilde{\mathbf{u}}_{i\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test,i})_{\mathcal{A}})^{\otimes 2} - \mathrm{E}\left((\widetilde{\mathbf{u}}_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}})^{\otimes 2} \right) \|_2 = O_p(\frac{s}{n} + \sqrt{\frac{s}{n}}).$$

Hence we have

$$\mathrm{III} = \boldsymbol{\beta}^{*T}_{\mathcal{A}} \mathrm{E}\left((\widetilde{\mathbf{u}}_{\mathcal{A}} - \mathbf{f}(\mathbf{z}_{test})_{\mathcal{A}})^{\otimes 2} \right) \boldsymbol{\beta}^{*}_{\mathcal{A}} + O_p(\frac{s}{n} + \sqrt{\frac{s}{n}}).$$

For IV, we have

$$\begin{aligned} & (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}})^T \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}(\mathbf{z}_{test,i})^{\otimes 2}_{\mathcal{M}} \right\} (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}}) \\ = {} & (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}})^T \left\{ \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}(\mathbf{z}_{test,i})^{\otimes 2}_{\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}} \right\} (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}}) \\ & + (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}})^T \boldsymbol{\Sigma}_{\mathcal{M}\mathcal{M}} (\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}}) \\ = {} & O_P\left(\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^{*}_{\mathcal{M}}\|_2^2 \right) = O_p(s(\log p)(\log n)^{1/2} n^{-1/2}), \end{aligned}$$

where the last identify follows from Theorem 4.3 and Condition 5.

By comparing the upper bounds of II and IV we conclude that

$$\mathrm{I} + \mathrm{II} + \mathrm{IV} = O_p(s(\log p)(\log n)^{1/2} n^{-1/2} \vee s_{\mathcal{A}}(\log p/n)^{1-qr}) = O_p(sa_n \vee s_{\mathcal{A}} b_n).$$

If $\mathcal{A} \neq \emptyset$, the cross-products of III and each of I, II and IV are the leading terms and of the order of $O_p(\sqrt{sa_n} \vee \sqrt{s_{\mathcal{A}} b_n})$. If $\mathcal{A} = \emptyset$, then the leading term is $O_p(sa_n)$. □

4.8.5 Supporting Lemmas and Their Proofs

Lemma 4.2. *When n is sufficiently large, it follows that*

$$\max_{j \in \mathcal{C}} \sup_{t \in I_{jn}} \left| \widehat{\mathrm{f}}_j(t) - \mathrm{f}_j(t) \right| = O_p\left(\sqrt{\frac{\log\log n + (1/2)\log p_1}{n^{1/2}}} \right),$$

where $\widehat{\mathrm{f}}_j(t)$ is given by (4.3.3)*, $\mathrm{f}_j(t) = \Phi^{-1}(F_j(t))$, $I_{jn} = [\mathrm{g}_j(-\sqrt{\log n}), \mathrm{g}_j(\sqrt{\log n})]$, and $\mathrm{g}_j(u) = \mathrm{f}_j^{-1}(u) = F_j^{-1}(\Phi(u))$.*

Proof. The proof follows a similar argument as in Theorem 2 of Han et al. (2013). We first show that, for sufficiently large n,

$$P\left(\sup_{t \in I_{jn}} \left| \widehat{\mathrm{f}}_j(t) - \mathrm{f}_j(t) \right| \geq C\sqrt{\frac{\log\log n + (1/2)\log p_1}{n^{1/2}}} \right) = o(p_1^{-1}).$$

Then applying the union bound completes the proof.

By symmetry, we only focus on interval $I_{jn}^s = [\mathrm{g}_j(0), \mathrm{g}_j(\sqrt{\log n})]$. Define a series $0 = \beta_{-1} < \alpha = \beta_0 < 1 < \beta_1 < ... < \beta_\kappa$ and $I_{ijn} := [\mathrm{g}_j(\sqrt{\beta_{i-1}\log n}), \mathrm{g}_j(\sqrt{\beta_i \log n})]$. For all $i = 0, 1, ..., \kappa$, we have

$$\sup_{t \in I_{ijn}} \left| \widehat{\mathrm{f}}_j(t) - \mathrm{f}_j(t) \right| = \sup_{t \in I_{ijn}} \left| \Phi^{-1}(\widetilde{F}_j(t)) - \Phi^{-1}(F_j(t)) \right|.$$

Then by the Mean Value Theorem, there always exists some $\xi_n \in [\min\{\widetilde{F}_j(\mathrm{g}_j(\sqrt{\beta_{i-1}\log n})), F_j(\mathrm{g}_j(\sqrt{\beta_{i-1}\log n}))\}, \max\{\widetilde{F}_j(\mathrm{g}_j(\sqrt{\beta_i \log n})), F_j(\mathrm{g}_j(\sqrt{\beta_i \log n}))\}]$, such that

$$\sup_{t \in I_{ijn}} \left| \Phi^{-1}(\widetilde{F}_j(t)) - \Phi^{-1}(F_j(t)) \right| = \sup_{t \in I_{ijn}} \left| (\Phi^{-1})'(\xi_n)(\widetilde{F}_j(t) - F_j(t)) \right|.$$

Next, we bound both $(\Phi^{-1})'(\xi_n)$ and $\widetilde{F}_j(t) - F_j(t)$ in each of these small intervals.

Since $(\Phi^{-1})'(u) = 1/\phi(\Phi^{-1}(u))$, then by its monotonicity, we have

$$(\Phi^{-1})'(\xi_n) \leq (\Phi^{-1})'(\max(\widetilde{F}_j(\mathrm{g}_j(\sqrt{\beta_i \log n})), F_j(\mathrm{g}_j(\sqrt{\beta_i \log n})))).$$

By Lemma 21 of Han et al. (2013), for large enough n, it holds almost surely that

$$\begin{aligned}
\widetilde{F}_j(g_j(\sqrt{\beta_i \log n})) &\leq 4\sqrt{\frac{\log\log n}{n}}(1 - F_j(g_j(\sqrt{\beta_i \log n})))^{1/2} + F_j(g_j(\sqrt{\beta_i \log n})) \\
&= 4\sqrt{\frac{\log\log n}{n}}(1 - \Phi(\sqrt{\beta_i \log n}))^{1/2} + \Phi(\sqrt{\beta_i \log n}) \\
&\leq \Phi\left(\sqrt{\beta_i \log n} + 4\sqrt{\frac{\log\log n}{n^{1-\beta_i/2}}}\right),
\end{aligned}$$

where the last inequality follows from the supplementary material of Liu et al. (2012). Hence for large enough n, it holds almost surely that

$$\begin{aligned}
(\Phi^{-1})'(\widetilde{F}_j(g_j(\sqrt{\beta_i \log n}))) &\leq 1/\phi\left(\sqrt{\beta_i \log n} + 4\sqrt{\frac{\log\log n}{n^{1-\beta_i/2}}}\right) \\
&\asymp (\Phi^{-1})'(F_j(g_j(\sqrt{\beta_i \log n}))).
\end{aligned}$$

Then, $(\Phi^{-1})'(\xi_n) \leq C/\phi(\sqrt{\beta_i \log n}) \leq c_1 n^{\beta_i/2}$.

To bound $\widetilde{F}_j(t) - F_j(t)$, we have

$$\begin{aligned}
P\left(\sup_{t \in I_{0jn}} \left|\widetilde{F}_j(t) - F_j(t)\right| \geq \frac{1}{2n} + \sqrt{\frac{\log\log n + (1/2)\log p_1}{n}}\right) &\leq \frac{2}{p_1(\log n)^2}, \\
P\left(\sup_{t \in I_{0jn}} \left|\widetilde{F}_j(t) - F_j(t)\right| \geq 2\sqrt{\frac{\log\log n + (1/2)\log p_1}{n}}\right) &\leq \frac{2}{p_1(\log n)^2}.
\end{aligned}$$

Above all, we can bound $\widetilde{\mathrm{f}}_j(t) - \mathrm{f}_j(t)$ on I_{0jn} by

$$P\left(\sup_{t \in I_{0jn}} \left|\widetilde{\mathrm{f}}_j(t) - \mathrm{f}_j(t)\right| \geq 2c_1\sqrt{\frac{\log\log n + (1/2)\log p_1}{n^{1-\alpha}}}\right) \leq \frac{2}{p_1(\log n)^2}.$$

Now for $i = 1, ..., \kappa$, it follows from Lemma 21 of Han et al. (2013) that for large enough n,

$$P\left(\sup_{t \in I_{ijn}} \left|\widetilde{F}_j(t) - F_j(t)\right| \leq 4\sqrt{\frac{\log\log n}{n}}(1 - F_j(\mathrm{g}_j(\sqrt{\beta_{i-1}\log n})))^{1/2}\right) = 1,$$
$$P\left(\sup_{t \in I_{ijn}} \left|\widetilde{F}_j(t) - F_j(t)\right| \leq 4\sqrt{\frac{\log\log n}{n}}(\frac{n^{-\beta_{i-1}/2}}{\sqrt{\alpha\log n}})^{1/2}\right) = 1,$$
$$P\left(\sup_{t \in I_{ijn}} \left|\widetilde{F}_j(t) - F_j(t)\right| \leq 4\sqrt{\frac{\log\log n}{n^{1+\beta_{i-1}/2}}}\right) = 1.$$

Since $(\Phi^{-1})'(\xi_n) \leq C/\phi(\sqrt{\beta_i \log n}) \leq c_1 n^{\beta_i/2}$, it holds for $i = 1, ..., \kappa$ that

$$P\left(\sup_{t \in I_{ijn}} \left|\widehat{\mathrm{f}}_j(t) - \mathrm{f}_j(t)\right| \leq 4c_1\sqrt{\frac{\log\log n}{n^{1+\beta_{i-1}/2-\beta_i}}}\right) = 1.$$

By choosing $\beta_i = (2 - (1/2)^i)(1/2)$ so that $1 - \alpha = 1/2$ and $1 + \beta_{i-1}/2 - \beta_i = 1/2$, we have

$$P\left(\sup_{t \in \cup_{i=0}^{\kappa} I_{ijn}} \left|\widehat{\mathrm{f}}_j(t) - \mathrm{f}_j(t)\right| \geq 4c_1\sqrt{\frac{\log\log n + (1/2)\log p_1}{n^{1/2}}}\right) \leq \frac{2}{p_1(\log n)^2}.$$

Since $\cup_{i=0}^{\kappa} I_{ijn} = [\mathrm{g}_j(0), \mathrm{g}_j(\sqrt{(2 - 2^{-\kappa})(1/2)\log n})]$, by symmetry, the same arguments apply to $[\mathrm{g}_j(-\sqrt{(2 - 2^{-\kappa})(1/2)\log n}), \mathrm{g}_j(\sqrt{(2 - 2^{-\kappa})(1/2)\log n})]$. By letting $\kappa \to \infty$, we prove the result for $I_{jn} = [\mathrm{g}_j(-\sqrt{\log n}), \mathrm{g}_j(\sqrt{\log n})]$. □

Lemma 4.3. *For $j \in \mathcal{C}$, suppose $\{X'_{ij}\}_{i=1}^{n'}$ and $\{X_{ij}\}_{i=1}^{n}$ are two i.i.d sequences, and $n' \asymp n$. When $p_1 = O(n^{\xi})$ for an arbitrary $\xi > 0$, it holds that*

$$\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{i=1}^{n'} (\widehat{\mathrm{f}}_j(X'_{ij}) - \mathrm{f}_j(X'_{ij}))^2 = O_p((\log p_1)(\log n)^{1/2} n^{-1/2}), \tag{4.8.5}$$
$$\max_{j \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} (\widehat{\mathrm{f}}_j(X_{ij}) - \mathrm{f}_j(X_{ij}))^2 = O_p((\log p_1)(\log n)^{1/2} n^{-1/2}), \tag{4.8.6}$$

where $\widehat{\mathrm{f}}_j$ is given by (4.3.3).

Proof. We use similar arguments as in Theorem 4 of Liu et al. (2009) and only prove for (4.8.5). The proof for (4.8.6) can be done similarly.

For any arbitrary ξ, there always exists a constant $M \geq 2(1 + \xi)$. Without loss of generality, assume $M > 2$ and let $\beta = 1$. We break the interval $[\mathrm{g}_j(-\sqrt{M \log n}), \mathrm{g}_j(\sqrt{M \log n})]$ into

$$\begin{aligned}
\mathscr{M}_n &= (\mathrm{g}_j(-\sqrt{\beta \log n}), \mathrm{g}_j(\sqrt{\beta \log n})), \\
\mathscr{E}_n &= \left[\mathrm{g}_j(-\sqrt{M \log n}), \mathrm{g}_j(-\sqrt{\beta \log n})\right] \cup \left[\mathrm{g}_j(\sqrt{\beta \log n}), \mathrm{g}_j(\sqrt{M \log n})\right].
\end{aligned}$$

In $\mathscr{M}_n$, the convergence of marginal transformations has been studied in Lemma 4.2. Let $\Delta_i(j) = (\widehat{\mathrm{f}}_j(X'_{ij}) - \mathrm{f}_j(X'_{ij}))^2$ and $\Theta_t(j) = (\widehat{\mathrm{f}}_j(t) - \mathrm{f}_j(t))^2$.

For any $\epsilon > 0$, we have

$$P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{i=1}^{n'} \Delta_i(j) > \epsilon\right) = P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{i=1}^{n'} \Delta_i(j) > \epsilon, \mathscr{A}_n\right) + P(\mathscr{A}_n^c),$$

where the event $\mathscr{A}_n$ is defined as

$$\mathscr{A}_n = \{\mathrm{g}_j(-\sqrt{M \log n}) \le X'_{1j}, \ldots, X'_{n'j} \le \mathrm{g}_j(\sqrt{M \log n}), \forall j \in \mathcal{C}\}.$$

Then by Lemma 13 of Liu et al. (2009) and using the fact that $M \ge 2(\xi + 1)$, we have

$$\begin{aligned}
P(\mathscr{A}_n^c) = P\left(\max_{i=1,\ldots,n'; j \in \mathcal{C}} \left|\mathrm{f}_j(X'_{ij})\right| > \sqrt{M \log n}\right) &\le P\left(\max_{i=1,\ldots,n'; j \in \mathcal{C}} \left|\mathrm{f}_j(X'_{ij})\right| > \sqrt{2 \log(np_1)}\right) \\
&\le \frac{c}{2\sqrt{\pi \log(np_1)}}.
\end{aligned}$$

For $P\left(\max_{j \in \mathcal{C}} (1/n') \sum_{i=1}^{n'} \Delta_i(j) > \epsilon, \mathscr{A}_n\right)$, we have

$$\begin{aligned}
&P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{i=1}^{n'} \Delta_i(j) > \epsilon, \mathscr{A}_n\right) \\
&\le P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{X'_{ij} \in \mathscr{E}_n} \Delta_i(j) > \epsilon/2\right) + P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{X'_{ij} \in \mathscr{M}_n} \Delta_i(j) > \epsilon/2\right) \\
&= \mathrm{I} + \mathrm{II}.
\end{aligned}$$

For II, we have

$$\begin{aligned} P\left(\max_{j\in\mathcal{C}}\frac{1}{n'}\sum_{X'_{ij}\in\mathscr{M}_n}\Delta_i(j)>\epsilon/2\right) &\le p_1P\left(\max_{t\in\mathscr{M}_n}\Theta_t(j)>\epsilon/2\right)\\ &= p_1P\left(\sup_{t\in\mathscr{M}_n}\left|\widehat{\mathrm{f}}_j(t)-\mathrm{f}_j(t)\right|>\sqrt{\epsilon/2}\right). \end{aligned}$$

Since it follows from Lemma 4.2 that, for any $\epsilon\ge C(\log\log n+\log p_1)n^{-1/2}$,

$$P\left(\max_{t\in\mathscr{M}_n}\left|\widehat{\mathrm{f}}_j(t)-\mathrm{f}_j(t)\right|>\sqrt{\epsilon/2}\right)=o(1/p_1).$$

By choosing $\epsilon\ge C(\log p_1)n^{-1/2}$, we have II $=o(1)$.

For I, let $\theta_1=n^{\beta/2}\epsilon/(4A\sqrt{\log n})$, where A is a constant given in the Lemma 15 of Liu et al. (2009). Then,

$$\frac{n'\epsilon}{2\theta_1}-n'A\sqrt{\frac{\log n}{n^\beta}}=n'A\sqrt{\frac{\log n}{n^\beta}}>0.$$

By Lemma 15 of Liu et al. (2009), we have

$$\begin{aligned} &P\left(\frac{1}{n'}\sum_{i=1}^{n'}I(X'_{ij}\in\mathscr{E}_n)>\frac{\epsilon}{2\theta_1}\right)\\ &=P\left(\sum_{i=1}^{n'}(I(X'_{ij}\in\mathscr{E}_n)-P(X'_{ij}\in\mathscr{E}_n))>\frac{n'\epsilon}{2\theta_1}-n'P(X_{ij}\in\mathscr{E}_n)\right)\\ &\le P\left(\sum_{i=1}^{n'}(I(X'_{ij}\in\mathscr{E}_n)-P(X'_{ij}\in\mathscr{E}_n))>\frac{n'\epsilon}{2\theta_1}-n'A\sqrt{\frac{\log n}{n^\beta}}\right). \end{aligned}$$

By the Bernstein inequality, we have

$$\begin{aligned} &P\left(\frac{1}{n'}\sum_{i=1}^{n'}I(X'_{ij}\in\mathscr{E}_n)>\frac{\epsilon}{2\theta_1}\right)\\ &\le P\left(\sum_{i=1}^{n'}(I(X'_{ij}\in\mathscr{E}_n)-P(X'_{ij}\in\mathscr{E}_n))>n'A\sqrt{\frac{\log n}{n^\beta}}\right)\\ &\le\exp\left(-\frac{c_1n^{2-\beta}\log n}{c_2n^{1-\beta/2}\sqrt{\log n}+c_3n^{1-\beta/2}\sqrt{\log n}}\right)=o(1), \end{aligned}$$

since $\beta = 1$. Note that,

$$\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{X'_{ij} \in \mathscr{E}_n} \Delta_i(j) \leq \max_{j \in \mathcal{C}} \sup_{t \in \mathscr{E}_n} \Theta_t(j) \cdot \frac{1}{n'} \sum_{i=1}^{n'} I(X'_{ij} \in \mathscr{E}_n).$$

Therefore,

$$\begin{aligned} P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{X'_{ij} \in \mathscr{E}_n} \Delta_i(j) > \epsilon/2\right) &= P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{X'_{ij} \in \mathscr{E}_n} \Delta_i(j) > \epsilon/2, \max_{j \in \mathcal{C}} \sup_{t \in \mathscr{E}_n} \Theta_t(j) > \theta_1\right) \\ &\quad + P\left(\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{X'_{ij} \in \mathscr{E}_n} \Delta_i(j) > \epsilon/2, \max_{j \in \mathcal{C}} \sup_{t \in \mathscr{E}_n} \Theta_t(j) \leq \theta_1\right) \\ &\leq P\left(\max_{j \in \mathcal{C}} \sup_{t \in \mathscr{E}_n} \Theta_t(j) > \theta_1\right) + P\left(\frac{1}{n'} \sum_{i=1}^{n'} I(X'_{ij} \in \mathscr{E}_n) > \frac{\epsilon}{2\theta_1}\right) \\ &= P\left(\max_{j \in \mathcal{C}} \sup_{t \in \mathscr{E}_n} \Theta_t(j) > \theta_1\right) + o(1). \end{aligned}$$

Then we have,

$$P\left(\max_{j \in \mathcal{C}} \sup_{t \in \mathscr{E}_n} \Theta_t(j) > \theta_1\right) \leq p_1 P\left(\sup_{t \in \mathscr{E}_n} (\widehat{\mathrm{f}}(t) - \mathrm{f}(t))^2 \geq \theta_1\right) = p_1 P\left(\sup_{t \in \mathscr{E}_n} \left|\widehat{\mathrm{f}}(t) - \mathrm{f}(t)\right| \geq \sqrt{\theta_1}\right).$$

To ensure the above probability converges to 0, we choose

$$\theta_1 = \frac{n^{\beta/2}\epsilon}{4A\sqrt{\log n}} \geq 2(M+4)\log n,$$

where $\epsilon \geq C_M(\log p_1)(\log n)^{1/2} n^{-1/2}$, and $C_M = 8A(M+4)/(1+\xi)$. Note that this choice of ϵ also guarantees that $\epsilon \geq C(\log p_1)n^{-1/2}$ as required for handling the arguments in $\mathscr{M}_n$. Thus, we have

$$\max_{j \in \mathcal{C}} \frac{1}{n'} \sum_{i=1}^{n'} (\widehat{\mathrm{f}}_j(X'_{ij}) - \mathrm{f}_j(X'_{ij}))^2 = O_p((\log p_1)(\log n)^{1/2} n^{-1/2}).$$

Similarly, (4.8.6) can be shown using the above argument by subsituting $\{X'_{ij}\}_{1 \leq i \leq n'}$ with $\{X_{ij}\}_{1 \leq i \leq n}$ for any $j \in \mathcal{C}$. □

Lemma 4.4. *Let* $\widehat{\boldsymbol{\beta}}$ *be the solution of* (4.3.1), *when RSC condition holds and* $\lambda \geq 2\|\nabla L(\boldsymbol{\beta}^*)\|_\infty$, *it* *holds that*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 = O(s\lambda^2), \ \textit{and} \ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O(s\lambda).$$

Proof. Lemma 4.4 is a direct implication of Corollary 1 of Negahban et al. (2012). □

Lemma 4.5. *Under Conditions 2, 3 and 4, there exists a sufficiently large positive constant C, and some generic positive constants C_1 and C_2 such that*

$$P\left(\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \boldsymbol{\delta}\|_\infty \geq C\sqrt{\frac{\log p}{n}}\right) \leq C_1 p^{2-C_2 C}.$$

Proof. Since $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, we have

$$\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \boldsymbol{\delta}\|_\infty = \|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}\|\boldsymbol{\beta}^*\|_1.$$

By Condition 2, we have

$$\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \boldsymbol{\delta}\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}\|\boldsymbol{\beta}^*\|_1 \leq M\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}.$$

By Conditions 3 and 4, it follows from Corollary 6.1 of Fan et al. (2017), Theorem 7 of Yoon et al. (2020), Proposition 1 of Feng and Ning (2019), and (3.5) of Zhao et al. (2014) that

$$P\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \geq C\sqrt{\frac{\log p}{n}}\right) \leq C_1 p^{2-C_2 C}.$$

Then,

$$\begin{aligned} P\left(\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* - \boldsymbol{\delta}\|_\infty \geq C\sqrt{\frac{\log p}{n}}\right) &\leq P\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \cdot M \geq C\sqrt{\frac{\log p}{n}}\right) \\ &= P\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \geq C\sqrt{\frac{\log p}{n}}\right) \\ &\leq C_1 p^{2-C_2 C}. \end{aligned}$$

This completes the proof of Lemma 4.5. □

Lemma 4.6. *Under Conditions 3, 4 and 6, if $s\sqrt{(\log p)/n} = o(1)$, there exists a sufficiently large positive constant C and some generic positive constans C_1 and C_2 such that*

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \geq 2M) \leq C_1 p^{2-C_2C}.$$

Proof. By Fan et al. (2017), Yoon et al. (2020),Feng and Ning (2019), and Zhao et al. (2014), we have

$$P\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \geq C\sqrt{\frac{\log p}{n}}\right) \leq C_1 p^{2-C_2C}.$$

Then,

$$\begin{aligned} P\left(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}\|_\infty \geq Cs\sqrt{\frac{\log p}{n}}\right) &\leq P\left(s\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}\|_{\max} \geq Cs\sqrt{\frac{\log p}{n}}\right) \\ &\leq C_1 p^{2-C_2C}. \end{aligned}$$

Since

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty &= \|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} + \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}(\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}} - \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \\ &\leq \|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty + \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}} - \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}\|_\infty \|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty, \end{aligned}$$

by Conditions 3, 4 and 6, it holds with probability greater than $1 - C_1 p^{2-C_2C}$ that

$$\begin{aligned} \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty &\leq M + MCs\sqrt{\frac{\log p}{n}}\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty, \\ \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty &\leq \frac{M}{1 - MCs\sqrt{(\log p)/n}} \leq 2M. \end{aligned}$$

where the last inequality holds since $s\sqrt{\log p/n} = o(1)$. □

Lemma 4.7. *Under Conditions 3,4,6 and 7, we have*

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty > (1-\alpha)(1-\epsilon/2)) \leq C_1 p^{2-C_2C}.$$

Proof. Note that

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} &= \widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}) + (\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} + \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

For I,

$$\begin{aligned}&\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1})\|_\infty \\ &\le (\|\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty + \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty)\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}\|_\infty\|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty.\end{aligned}$$

Since by Condition 4, $\|\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty \lesssim s$, and $P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty \le Cs\sqrt{(\log p)/n}) \ge 1 - C_1p^{2-C_2C}$, we have

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1})\|_\infty \ge Cs^2\sqrt{\log p/n}) \le C_1p^{2-C_2C}.$$

Using similar arguments, for II, we have

$$P(\|(\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \ge Cs\sqrt{(\log p)/n}) \le C_1p^{2-C_2C}.$$

Hence, if $s^2\sqrt{(\log p)/n} = o(1)$, we have

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}) + (\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \le (1-\alpha)\epsilon/2) \ge 1 - C_1p^{2-C_2C}.$$

This result together with Condition 7 completes the proof. □

Lemma 4.8. *Under Conditions 3, 4 and 5, if we choose* $\lambda_2 = C\sqrt{\log p_1/n}$ *in* (3.8.3) *for some sufficiently large constant* C *and* $\log p_1/n = o(1)$*, then it holds with probability at least* $1 - C_1p^{2-C_2C}$ *that*

$$\|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_2^2 \lesssim R_q\lambda^{2-qr}, \quad \|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_1^2 \lesssim R_q^2\lambda^{2-2qr},$$

where C_1 *and* C_2 *are some positive constants and* $r \in (0,1)$*.*

Proof. Define $\mathbf{C} = \{\boldsymbol{\Delta} \in \mathbf{R}^{p_1} : \|\boldsymbol{\Delta}_{\mathbf{S}^c}\|_1 \leq 3\|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 + 4\|\boldsymbol{\eta}_{j_{\mathbf{S}^c}}\|_1\}$, where $\mathbf{S} = \{i \in \{1, 2, .., p_1\} : \left|(\boldsymbol{\eta}_j)_i\right| > \eta\}$ and $\eta > 0$ is some threshold to be chosen. For any $\boldsymbol{\Delta} \in \mathbf{C}$, we have

$$\begin{aligned}\|\boldsymbol{\Delta}\|_1 &\leq \|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 + \|\boldsymbol{\Delta}_{\mathbf{S}^c}\|_1 \leq 4\|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 + 4\|\boldsymbol{\eta}_{j_{\mathbf{S}^c}}\|_1 \\ &\leq 4\sqrt{|\mathbf{S}|}\|\boldsymbol{\Delta}\|_2 + 4R_{jq}\eta^{1-q} \\ &\leq 4\sqrt{R_q}\eta^{-q/2}\|\boldsymbol{\Delta}\|_2 + 4R_{jq}\eta^{1-q}.\end{aligned}$$

Next, we verify the RSC condition for $\boldsymbol{\Delta} \in \mathbf{C}$. We have

$$\begin{aligned}\sqrt{(1/2)\boldsymbol{\Delta}^T\widehat{\boldsymbol{\Sigma}}_{\mathcal{CC}}\boldsymbol{\Delta}} &= \sqrt{(1/2)\boldsymbol{\Delta}^T\boldsymbol{\Sigma}_{\mathcal{CC}}\boldsymbol{\Delta} + (1/2)\boldsymbol{\Delta}^T(\widehat{\boldsymbol{\Sigma}}_{\mathcal{CC}} - \boldsymbol{\Sigma}_{\mathcal{CC}})\boldsymbol{\Delta}} \\ &\geq ((m/2)\|\boldsymbol{\Delta}\|_2^2 - C\sqrt{\log p/n}\|\boldsymbol{\Delta}\|_1^2)^{1/2} \\ &\geq \sqrt{m/2}\|\boldsymbol{\Delta}\|_2 - C(\log p/n)^{1/4}\|\boldsymbol{\Delta}\|_1 \\ &\geq \sqrt{m/2}\|\boldsymbol{\Delta}\|_2 - C(\log p/n)^{1/4}(\sqrt{R_q}\eta^{-q/2}\|\boldsymbol{\Delta}\|_2 + R_q\eta^{1-q}) \\ &= \|\boldsymbol{\Delta}\|_2(\sqrt{m/2} - C(\log p/n)^{1/4}\sqrt{R_q}\eta^{-q/2}) - C(\log p/n)^{1/4}R_q\eta^{1-q}.\end{aligned}$$

When we choose $\lambda_2 = C\sqrt{\log p/n}$ and $\eta = C\lambda_2^r$ for some $0 < r < 1$, we have

$$C(\log p/n)^{1/4}\sqrt{R_q}\eta^{-q/2} = C\sqrt{R_q}(\log p/n)^{(1-qr)/4} < \sqrt{m/8},$$

when $\log p/n = o(1)$. Thus, it holds with probability at least $1 - C_1p^{2-C_2C}$ that

$$\sqrt{(1/2)\boldsymbol{\Delta}^T\widehat{\boldsymbol{\Sigma}}_{\mathcal{CC}}\boldsymbol{\Delta}} \geq \|\boldsymbol{\Delta}\|_2\sqrt{m/8} - C(\log p/n)^{1/4}R_q\eta^{1-q},$$

which verifies the RSC condition. Then, the result follows from Theorem 1 from Negahban et al. (2012). □

Lemma 4.9. *Let* $\mathrm{h}_1(x) = \phi(x)/(1-\Phi(x))$, $\mathrm{h}_2(x) = \phi(x)/\Phi(x)$, *and* $\mathrm{h}_3(x, y) = (\phi(y)-\phi(x))/(\Phi(y)-\Phi(x))$. *Then,* h_1, h_2, *and* h_3 *are uniformly Lipschitz continuous with finite Lipschitz constants.*

Proof. To show that a continuously differentiable function is uniformly Lipschitz continuous, we only need to show that its partial derivatives are all uniformly bounded. Since,

$$\mathrm{h}_1(x) = \frac{\phi(x)}{1-\Phi(x)} = \frac{\phi(x)}{\Phi(-x)} = \frac{\phi(-x)}{\Phi(-x)} = \mathrm{h}_2(-x).$$

We only need to show h_2 is uniformly Lipschitz continuous. We have,

$$\mathrm{h}_2^{'}(x) = \frac{\phi(x)(-x)\Phi(x) - \phi(x)^2}{\Phi(x)^2}.$$

which is bounded as $x \to \infty$. We prove that it is also bounded when $x \to -\infty$. By L'Hospital's rule, we have

$$\lim_{x \to -\infty} \frac{\Phi(x)}{\phi(x)/(-x)} = 1,$$

which implies that

$$\lim_{x \to -\infty} \frac{-x}{\mathrm{h}_2(x)} = \lim_{x \to -\infty} \frac{-1}{\mathrm{h}_2^{'}(x)} = 1.$$

Therefore, $\lim_{x \to -\infty} \mathrm{h}_2^{'}(x) = -1$.

Next, we show h_3 is uniformly Lipschitz continuous. We have

$$\frac{\partial \mathrm{h}_3(x,y)}{\partial y} = \frac{\phi(y)(-y)(\Phi(y)-\Phi(x)) - (\phi(y)-\phi(x))\phi(y)}{(\Phi(y)-\Phi(x))^2},$$

which is bounded everywhere, except for $y = x$. By using L'Hospital's, we have

$$\lim_{y \to x} \frac{\partial \mathrm{h}_3(x,y)}{\partial y} < \infty.$$

Hence, $\partial \mathrm{h}_3(x,y)/\partial y$ is uniformly bounded. Similarly, we can show that $\partial \mathrm{h}_3(x,y)/\partial x$ is uniformly bounded. Thus, h_3 is uniformly Lipschitz continuous. □

Lemma 4.10. *Under Conditions 3, 4 and 5, if we choose $\lambda_2 = C\sqrt{\log p_1/n}$ in* (4.5.2) *for some sufficiently large constant C and $p = O(n^{\xi})$ for an arbitrary $\xi > 0$, we have for $j \in \mathcal{B} \cup \mathcal{O} \cup \mathcal{T}$ that $(1/n_{test}) \sum_{i=1}^{n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 = O_p(a_n \vee b_n)$, where $a_n = (\log p/n)^{1-qr}$ and $b_n = (\log p)(\log n)^{1/2} n^{-1/2}$.*

Proof. We prove for $j \in \mathcal{B}$. The proofs for $j \in \mathcal{O} \cup \mathcal{T}$ can be done analogously. Note that, $\widetilde{u}_{ij} = A_{ij} + \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})$ and $\widehat{u}_{ij} = \widehat{A}_{ij} + \widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})$, where

$$A_{ij} = X_{test,ij} \frac{\sqrt{1-\xi_j}\phi(L_{ij})}{1-\Phi(L_{ij})} - (1 - X_{test,ij}) \frac{\sqrt{1-\xi_j}\phi(L_{ij})}{\Phi(L_{ij})} \text{ and } L_{ij} = \frac{\Delta_j - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})}{\sqrt{1-\xi_j}}.$$

Hence, we only need to study $(1/n_{test}) \sum_{i=1}^{n_{test}} (\widehat{A}_{ij} - A_{ij})^2$ and $(1/n_{test}) \sum_{i=1}^{n_{test}} (\widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}) - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^2$. For the second term, it has

$$\begin{aligned} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}) - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^2 &\lesssim (\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)^T (\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})^{\otimes 2})(\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j) \\ &+ \boldsymbol{\eta}_j^T (\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}) - \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^{\otimes 2}) \boldsymbol{\eta}_j \\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

For I, we have

$$\begin{aligned} &(\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)^T (\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})^{\otimes 2})(\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j) \\ \le\ & \|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_1^2 \|\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})^{\otimes 2}\|_{\max} \\ \le\ & \|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_1^2 \|\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})^{\otimes 2} - \boldsymbol{\Sigma}_{\mathcal{CC}}\|_{\max} \\ +\ & \|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_1^2 \|\boldsymbol{\Sigma}_{\mathcal{CC}}\|_{\max} \\ \lesssim\ & \|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_1^2, \end{aligned}$$

where the last inequality follows from Bickel and Levina (2008b) and $\|\boldsymbol{\Sigma}_{\mathcal{CC}}\|_{\max} = 1$. Hence, by Lemma 4.8, we have $\mathrm{I} = O_p(a_n)$.

For II, we have

$$\begin{aligned}
&\boldsymbol{\eta}_j^T(\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})-\mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^{\otimes 2})\boldsymbol{\eta}_j\\
&\le \|\boldsymbol{\eta}_j\|_1^2\|\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})-\mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^{\otimes 2}\|_{\max}\\
&\lesssim \|\boldsymbol{\eta}_j\|_1^2\max_{j\in\mathcal{C}}\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(\widehat{\mathrm{f}}_j(X_{test,ij})-\mathrm{f}_j(X_{test,ij}))^2.
\end{aligned}$$

Hence, by Lemma 4.3, we have $\mathrm{II}=O_p(b_n)$. Combining the results for I and II, we have

$$\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(\widehat{\boldsymbol{\eta}}_j^T\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})-\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^2=O_p(a_n\vee b_n). \tag{4.8.7}$$

To bound $(1/n_{test})\sum_{i=1}^{n_{test}}(\widehat{A}_{ij}-A_{ij})^2$, we have

$$\begin{aligned}
A_{ij}&=X_{test,ij}\frac{\sqrt{1-\xi_j}\phi(L_{ij})}{1-\Phi(L_{ij})}-(1-X_{test,ij})\frac{\sqrt{1-\xi_j}\phi(L_{ij})}{\Phi(L_{ij})}\\
&=\sqrt{1-\xi_j}(X_{test,j}\mathrm{h}_1(L_{ij})-(1-X_{test,j})\mathrm{h}_2(L_{ij})),\\
\widehat{A}_{ij}&=X_{test,ij}\frac{\sqrt{1-\widehat{\xi}_j}\phi(\widehat{L}_{ij})}{1-\Phi(\widehat{L}_{ij})}-(1-X_{test,ij})\frac{\sqrt{1-\widehat{\xi}_j}\phi(\widehat{L}_{ij})}{\Phi(\widehat{L}_{ij})}\\
&=\sqrt{1-\widehat{\xi}_j}(X_{test,ij}\mathrm{h}_1(\widehat{L}_{ij})-(1-X_{test,ij})\mathrm{h}_2(\widehat{L}_{ij})),\\
L_{ij}&=\frac{\Delta_j-\boldsymbol{\eta}_j^T\mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})}{\sqrt{1-\xi_j}},\text{ and }\widehat{L}_{ij}=\frac{\widehat{\Delta}_j-\widehat{\boldsymbol{\eta}}_j^T\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})}{\sqrt{1-\widehat{\xi}_j}}.
\end{aligned}$$

Note that

$$\begin{aligned}
\widehat{L}_{ij}-L_{ij}\lesssim&(\widehat{\Delta}_j-\Delta_j+\widehat{\boldsymbol{\eta}}_j^T\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})-\boldsymbol{\eta}_j{}^T\mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))\\
&+(\frac{1}{\sqrt{1-\widehat{\xi}_j}}-\frac{1}{\sqrt{1-\xi_j}})(\Delta_j-\boldsymbol{\eta}_j{}^T\mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})).
\end{aligned} \tag{4.8.8}$$

For $\widehat{A}_{ij}-A_{ij}$, we have

$$\begin{aligned}
\widehat{A}_{ij}-A_{ij}\lesssim&(\sqrt{1-\widehat{\xi}_j}-\sqrt{1-\xi_j})(X_{test,ij}\mathrm{h}_1(L_{ij})-(1-X_{test,ij})\mathrm{h}_2(L_{ij}))\\
&+\sqrt{1-\xi_j}\{X_{test,ij}\mathrm{h}_1(\widehat{L}_{ij})-(1-X_{test,ij})\mathrm{h}_2(\widehat{L}_{ij})\\
&-(X_{test,ij}\mathrm{h}_1(L_{ij})-(1-X_{test,ij})\mathrm{h}_2(L_{ij}))\}.
\end{aligned}$$

Furthermore, we observe

$$
\begin{aligned}
X_{test,ij}\mathrm{h}_1(L_{ij}) - (1 - X_{test,ij})\mathrm{h}_2(L_{ij}) &= X_{test,ij}(\mathrm{h}_1(L_{ij}) + \mathrm{h}_2(L_{ij})) - \mathrm{h}_2(L_{ij}) \\
&< \mathrm{h}_1(L_{ij}) + \mathrm{h}_2(L_{ij}),
\end{aligned}
$$

since $\mathrm{h}_2 > 0$ and $X_{test,ij} \leq 1$. Then, we have

$$
\begin{aligned}
\widehat{A}_{ij} - A_{ij} \lesssim & (\sqrt{1 - \widehat{\xi}_j} - \sqrt{1 - \xi_j})(\mathrm{h}_1(L_{ij}) + \mathrm{h}_2(L_{ij})) \\
& + \sqrt{1 - \xi_j}(X_{test,ij}(\mathrm{h}_1(\widehat{L}_{ij}) - \mathrm{h}_1(L_{ij})) + (1 - X_{test,ij})(\mathrm{h}_2(L_{ij}) - \mathrm{h}_2(\widehat{L}_{ij}))) \\
& \lesssim (\sqrt{1 - \widehat{\xi}_j} - \sqrt{1 - \xi_j})(\mathrm{h}_1(L_j) + \mathrm{h}_2(L_j)) \\
& + \sqrt{1 - \xi_j}(\mathrm{h}_1(\widehat{L}_j) - \mathrm{h}_1(L_j) + (\mathrm{h}_2(L_j) - \mathrm{h}_2(\widehat{L}_j))).
\end{aligned}
$$

It follows from Lemma 4.8 that

$$
\begin{aligned}
\widehat{A}_{ij} - A_{ij} \lesssim & (\sqrt{1 - \widehat{\xi}_j} - \sqrt{1 - \xi_j})(\mathrm{h}_1(0) + \mathrm{h}_2(0) + (H_1 + H_2)\left|L_{ij}\right|) \\
& + \sqrt{1 - \xi_j}(H_1 + H_2)\left|\widehat{L}_{ij} - L_{ij}\right|.
\end{aligned} \tag{4.8.9}
$$

Combining (4.8.8) and (4.8.9), we have

$$
\begin{aligned}
\widehat{A}_{ij} - A_{ij} \lesssim & (\sqrt{1 - \widehat{\xi}_j} - \sqrt{1 - \xi_j})\left(\mathrm{h}_1(0) + \mathrm{h}_2(0) + (H_1 + H_2)\left|\frac{\boldsymbol{\Delta}_j - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})}{\sqrt{1 - \xi_j}}\right|\right) \\
& + \sqrt{1 - \xi_j}(L_1 + L_2)\left|(\widehat{\boldsymbol{\Delta}}_j - \Delta_j + \widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}) - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))\right| \\
& + \sqrt{1 - \xi_j}(L_1 + L_2)\left|\left(\frac{1}{\sqrt{1 - \widehat{\xi}_j}} - \frac{1}{\sqrt{1 - \xi_j}}\right)(\Delta_j - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))\right|.
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
&\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{A}_{ij} - A_{ij})^2 \\
&\lesssim (\sqrt{1-\widehat{\xi}_j} - \sqrt{1-\xi_j})^2 \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\mathrm{h}_1(0) + \mathrm{h}_2(0) + (H_1 + H_2)\,|L_{ij}|)^2 \\
&+ (1-\xi_j)(L_1+L_2)^2 \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\widehat{\Delta}_j - \Delta_j + \widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}) - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})\right)^2 \\
&+ (1-\xi_j)(L_1+L_2)^2 \left(\frac{1}{\sqrt{1-\widehat{\xi}_j}} - \frac{1}{\sqrt{1-\xi_j}}\right)^2 \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} L_{ij}^2 \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}
$$

Note that $\widehat{\xi}_j - \xi_j \lesssim \|\widehat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_1 \max_{k\in\mathcal{C}} \Sigma_{jk} + \|\boldsymbol{\eta}_j\|_1 \max_{k\in\mathcal{C}}(\widehat{\Sigma}_{jk} - \Sigma_{jk}) = O_p(\sqrt{a_n})$ by Lemmas 4.1 and 4.8. Hence Condition 5 implies that

$$
\begin{aligned}
\sqrt{1-\widehat{\xi}_j} - \sqrt{1-\xi_j} \le \frac{\left|\xi_j - \widehat{\xi}_j\right|}{\sqrt{1-\xi_j}} \lesssim \left|\xi_j - \widehat{\xi}_j\right| = O_p(\sqrt{a_n}),\\
\left(\frac{1}{\sqrt{1-\widehat{\xi}_j}} - \frac{1}{\sqrt{1-\xi_j}}\right) = O_p(\sqrt{a_n}).
\end{aligned}
$$

Then we have $\mathrm{I} + \mathrm{III} = O_p(a_n)$. For II, we have shown that $\max_{j\in\mathcal{B}} \left|\widehat{\Delta}_j - \Delta_j\right| = O_p(\sqrt{\log p_2/n})$. Then, by (4.8.7), we have $\mathrm{II} = O_p(a_n \vee b_n)$. Above all, we have

$$
\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{A}_{ij} - A_{ij})^2 = O_p(a_n \vee b_n).
$$

Combining it with the result for $(1/n_{test}) \sum_{i=1}^{n_{test}} (\widehat{\boldsymbol{\eta}}_j^T \widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}) - \boldsymbol{\eta}_j^T \mathbf{f}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}))^2$, we have the following for $j \in \mathcal{B}$,

$$
\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\widehat{u}_{ij} - \widetilde{u}_{ij})^2 = O_p(a_n \vee b_n) = O_p((\log p/n)^{1-qr} \vee (\log p)(\log n)^{1/2} n^{-1/2}).
$$

For $j \in \mathcal{O}$, $\widehat{u}_{ij}$ can be expressed as

$$\begin{aligned}\widehat{u}_{ij} &= \sqrt{1-\widehat{\xi}_j}\sum_{k=0}^{N_j} I(X_{test,ij}=k)\frac{\phi(\widehat{L}_{ijk})-\phi(\widehat{L}_{ij(k+1)})}{\Phi(\widehat{L}_{ij(k+1)})-\Phi(\widehat{L}_{ijk})}+\widehat{\boldsymbol{\eta}}_j^T\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})\\ &= \sqrt{1-\widehat{\xi}_j}\sum_{k=0}^{N_j} I(X_{test,ij}=k)(-\mathrm{h}_3(\widehat{L}_{ijk},\widehat{L}_{ij(k+1)}))+\widehat{\boldsymbol{\eta}}_j^T\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}}),\end{aligned}$$

and h_3 has been shown in Lemma 4.9 to be uniformly Lipschitz continuous. Similarly, we have

$$\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(\widehat{u}_{ij}-\widetilde{u}_{ij})^2 = O_p((\log p/n)^{1-qr}\vee(\log p)(\log n)^{1/2}n^{-1/2}).$$

For $j \in \mathcal{T}$, $\widehat{u}_{ij}$ can be expressed as

$$\widehat{u}_{ij} = I(X_{test,ij}>0)\widehat{\mathrm{f}}_j(X_{test,ij})+I(X_{test,ij}=0)(-\sqrt{1-\widehat{\xi}_j}\mathrm{h}_2(\widehat{L}_{ij})+\widehat{\boldsymbol{\eta}}_j^T\widehat{\mathbf{f}}_{\mathcal{C}}(\mathbf{x}_{test,i\mathcal{C}})).$$

By an analogous argument, we have

$$\frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(\widehat{u}_{ij}-\widetilde{u}_{ij})^2 = O_p((\log p/n)^{1-qr}\vee(\log p)(\log n)^{1/2}n^{-1/2}).$$

Hence, the proof of Lemma 4.10 is complete. □

CHAPTER 5
MULTI-CLASS CLASSIFICATION VIA LATENT MIXED GAUSSIAN COPULA MODEL

5.1 Introduction

Multi-class classification, assigning a subject to one of multple categories based on certain features, is an important task commonly appearing in statistical research. In many applications, the response for a multi-class classification problem frequently occurs with inherent ordering, which leads to an ordinal classification problem. The ordinal classification problem is traditionally handled by logistic or probit ordinal regression model under the proportional odds assumption (McCullagh 1980). With the emergence of big data containing enormous features in various scales, it poses great challenges to handle the high dimensionality and non-normality of the features. Although penalized likelihood methods have been successfully applied to high-dimensional linear, logistic regression models (Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005) and Cox proportional hazards regression model (Tibshirani 1997), it is only until recently that there have been some regularized methods for high-dimensional ordinal regression model. Archer and Williams (2012) proposed to fit an L_1 penalized continuation ratio model for ordinal response, Wurm et al. (2017) proposed to fit a family of multinomial-ordinal models with the elastic net penalty. Other than penalized likelihood methods, some works (Archer et al. 2014; Hou and Archer 2015) proposed incremental forward stagewise method as a greedy approximation to fit a family of high-dimensional ordinal regression models. However, there has not been any theoretical guarantee on the classification performance for these existing methods. To deal with non-normality of features in various scales, transformations, such as the Box-Cox transformation, Fisher z-transformation and variance stablizing transformation, have been frequently applied to overcome potential violations of model assumptions (Carroll and Ruppert 1988). However, the choice of the transformations could be subjective and error-prone when dealing with high-dimensional features. Above all, it is desirable to develop a unified framework that could handle multi-class ordinal classification without resorting to ordinal regression.

Recent progress in estimating correlations among mixed variables using copula-based methods

(Liu et al. 2009; 2012; Fan et al. 2017; Feng and Ning 2019) provides insights on high-dimensional ordinal classification problem. Specifically, Liu et al. (2009; 2012) proposed a Gaussian copula model to estimate correlations among continuous variables in various scales. Fan et al. (2017) proposed a latent Gaussian copula model to simultaneuously handle continuous and binary variables. Feng and Ning (2019) generalized the latent Gaussian copula model to handle ordinal and continuous variables. These methods assume that there exist some latent continuous variables that generate the observed mixed variables, and the latent continuous variables follow a standard multivariate normal distribution, after applying some transformations. These methods propose to use rank-based quantities to estimate the correlations. They can be applied to a series of unsupervised learning problems, such as graph estimation and principal component analysis.

For the ordinal regression problem, there has been a rich literature that model the observed ordinal response by thresholding a latent continuous variable (Hedeker and Gibbons 1994; Qu et al. 1995; Sha and Dechi 2019), then the association between the observed ordinal response and continuous features becomes equivalent to the association between the latent continuous response and continuous features. Such model has been successfully applied to studying Alzheimer's Disease progression (Doyle et al. 2014), psychometrics (Bürkner and Vuorre 2019), and many other fields. To make this latent variable model even more flexible to handle features with high dimensionality and non-normality, the latent mixed Gaussian copula model provides a unified framework for high-dimensional ordinal regression. Under this framework, it is natural to jointly model the ordinal response and the high-dimensional continuous features so that the latent response and the latent features are jointly normal and follow a linear regression model. Moreover, we can derive the Bayes rule of classifying the ordinal response under such model, which consequently leads to a Fisher consistent classification rule obtained by consistently estimating the unknown parameters.

To this end, we propose a semiparametric latent Gaussian copula classification method to classify ordinal response given high-dimensional continuous features in various scales. The main contribution of the paper are as follows. First, our method gives a unified framework to hanlde high-dimensional ordinal classification, which incorporates the traditional ordinal regression model as a special case. Second, our method has statistical properties in terms of estimation consistency and error rate consistency under mild sparsity assumption.

The rest of the paper is organized as the follows. Section 5.2 first presents the latent Gaussian copula model and the latent Gaussian copula classification rule, then explains the estimation of parameters in the latent Gaussian copula classification rule. Section 5.3 describes the statistical properties of the proposed method in terms of its estimation and error rate consistency. Section 5.4 presents extensive numerical studies that compare the proposed method with other well-known multi-class classification methods, demonstrating the proposed method's superiority in terms of classification performance and robustness under several metrics. In Section 5.5, we apply our method to an imaging dataset from UCI machine learning repository for breast cancer progression classification, and show our method results in an improved classification performance. All technical details are given in Section 5.6.

5.2 Methodology

5.2.1 Latent Mixed Gaussian Copula Classification

Consider a multi-class classification problem where $Y \in \{0, 1, ..., K\}$ denotes $K+1$ class labels, $\mathbf{x} = (X_1, ..., X_p)^T \in \mathbb{R}^p$ is a vector of continuous features. Denote $\mathcal{D} = \{(y_1, \mathbf{x}_1^T)^T, ..., (y_n, \mathbf{x}_n^T)^T\}$ as the training data. The goal of multi-class classification is to find a classification rule $D(\mathbf{x})$ from the training data so that we can assign a class label to a new subject based on his features $\mathbf{x}$. Based on the decision theory, the Bayes rule under the 0-1 loss is $D_{Bayes}(\mathbf{x}) = \text{argmax}_{k=0,1,...,K}\, P(Y = k|\mathbf{x})$, and the Bayes error $\mathcal{R}_{Bayes} = \mathrm{E}(I(D_{Bayes}(\mathbf{x}) \neq Y))$. In certain applications, the ranks of class labels have meanings. For example, the stages of Alzheimer disease or breast cancer are ordered based on the disease progression. The ratings of products indicate the customer preference. For these applications, it is important to model the distribution of the ordinal response Y. On the other hand, in the high-dimensional setting when $p > n$, it is hard to directly model $P(Y = k|\mathbf{x})$.

Instead, we propose a flexible model to model the joint distribution of $(Y, \mathbf{x})$. We first assume that the observed ordinal response Y is generated by a latent continuous variable Z such that $Y = \sum_{k=1}^{K} I(Z > C_k)$, where $\mathbf{C} = (C_1, ..., C_K)^T$ is the unknown thresholds. Then we assume that Z and $\mathbf{x}$ jointly follow a Gaussian copula. That is,

$$\begin{bmatrix} \mathrm{f}_y(Z) \\ \mathbf{f}_x(\mathbf{x}) \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} 1 & \boldsymbol{\Sigma}_{xy}^T \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right), \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \boldsymbol{\Sigma}_{xy}^T \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}, \tag{5.2.1}$$

where $\mathbf{f} = (\mathrm{f}_y, \mathbf{f}_x)$ are the unknown transformation functions, which are assumed to be monotonically increasing. The joint distribution for continuous variables $(Z, \mathbf{x})$ is called a nonparanormal (NPN) distribution, denoted by $(Z, \mathbf{x}) \sim NPN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{f})$. It was first introduced by Liu et al. (2009; 2012) to study the correlation among non-normal continuous variables. In addition, the joint distribution for mixed type variables $(Y, \mathbf{x})$ is called a latent mixed nonparanormal (LMNPN) distribution, denoted by $(Y, \mathbf{x}) \sim LMNPN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{f}, \mathbf{C}, K)$. This model was first introduced by Fan et al. (2017) to study the correlation among binary and continuous variables, and was later generalized by Feng and Ning (2019) to incorporate ordinal variables with arbitrarily many levels. Using the latent mixed nonparanormal distribution, we can simultaneously model the distribution of the ordinal response and the distribution of high-dimensional continuous features, while allowing much flexibility on the marginal transformation functions that characterize the specific distribution of the features. In the following, we derive the Bayes rule for ordinal classification under our proposed model for $P(Y, \mathbf{x})$.

Given (5.2.1), we essentially have the following regression model,

$$\mathrm{f}_y(Z) = \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* + \epsilon, \tag{5.2.2}$$

where $\boldsymbol{\beta}^* = \mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}$, $\epsilon \sim N(0, 1 - \mathbf{\Sigma}_{xy}^T \mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy})$ and is independent of $\mathbf{f}_x(\mathbf{x})$. Under (5.2.2), the Bayes rule becomes,

$$\begin{aligned} D_{Bayes}(\mathbf{x}) &= \underset{k=0,1,\ldots,K}{\operatorname{argmax}}\ P(\Delta_y^{(k)} < \mathrm{f}_y(Z) \le \Delta_y^{(k+1)} | \mathbf{f}_x(\mathbf{x})) \\ &= \underset{k=0,1,\ldots,K}{\operatorname{argmax}}\ \Phi(\frac{\Delta_y^{(k+1)} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*}{\sqrt{1-\xi_y}}) - \Phi(\frac{\Delta_y^{(k)} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*}{\sqrt{1-\xi_y}}) \\ &= \sum_{k=1}^{K} I(\mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^* > \Delta_y^{(k)}), \end{aligned} \tag{5.2.3}$$

where $\xi_y = \mathbf{\Sigma}_{xy}^T \mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}$, $\Delta_y^{(k)} = \mathrm{f}_y(C_k)$ and $-\infty = \Delta_y^{(0)} < \Delta_y^{(1)} < \ldots < \Delta_y^{(K+1)} = \infty$. In practice, we propose the latent mixed Gaussian copula classification rule (LMGCC)

$$D_{LMGCC}(\mathbf{x}) = \sum_{k=1}^{K} I(\widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} > \widehat{\Delta}_y^{(k)}), \tag{5.2.4}$$

where $\widehat{\boldsymbol{\beta}}$, $\widehat{\Delta}_y^{(k)}$ and $\widehat{\mathbf{f}}_x$ are estimators of $\boldsymbol{\beta}^*$, $\Delta_y^{(k)}$ and $\mathbf{f}_x$ to be introduced later.

Our model in (5.2.2) has a natural connection to the classical ordinal regression under proportional odds assumption, which learns the classification rule by modeling $g(P(Y \leq k|\mathbf{x})) = C_k - \boldsymbol{\omega}^T\mathbf{x}$ for $k = 1, ..., K$, where g is a pre-specified link function, $\boldsymbol{\omega}$ denotes the regression coefficients, and the distribution of the features $\mathbf{x}$ is unspecified. For the classical ordinal regression, if g is chosen as the probit function, it is equivalent to assuming $Y = \sum_{k=1}^{K} I(Z > C_k)$ where

$$Z = \boldsymbol{\omega}^T\mathbf{x} + \epsilon,$$

where $\epsilon|\mathbf{x} \sim N(0, 1)$. Hence the classical ordinal regression requires that $\mathrm{E}(Z|\mathbf{x})$ is a linear function in $\mathbf{x}$, which is prone to model misspecification for high-dimensional $\mathbf{x}$ since the functional forms of $\mathbf{x}$ are subjectively chosen from some transformations. In summary, there are two critical differences between our proposed model in (3.2.1) and the classical ordinal probit regression. First (5.2.2) specifies the distribution of the features $\mathbf{x} \sim NPN(\mathbf{0}, \boldsymbol{\Sigma}_{xx}, \mathbf{f}_x)$, while the classical ordinal regression have it unspecified. Second we allow the marginal transformation functions to be unspecified, while the classical ordinal probit regression restricts $\mathrm{E}(Z|\mathbf{x})$ to be a linear function of $\mathbf{x}$. For our proposed model (3.2.1), if we assume $\mathrm{f}_y(Z) = \sqrt{1 - \xi_y}Z$ and $\mathbf{f}_x$ is identity function, then (5.2.2) reduces to the classical ordinal regression with $\boldsymbol{\omega} = \boldsymbol{\beta}^*(1 - \xi_y)^{-1/2}$ and ϵ independent of $\mathbf{x}$. Hence, the classical ordinal probit regression can be viewed as a special case of (5.2.2). Unlike the classical ordinal regression, we utilize the properties of latent mixed Gaussian copula model to propose an M-estimation formulation to estimate $\boldsymbol{\beta}^*$ and plug-in estimators for $\Delta_y^{(k)}$ $(k = 1, ..., K)$ and $\mathbf{f}_x$.

5.2.2 Estimation

Since $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$, it can be seen that $\boldsymbol{\beta}^*$ solves the following problem

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmax}} \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} - \boldsymbol{\Sigma}_{xy}^T\boldsymbol{\beta}.$$

Inspired by such an observation, if we have an estimator $\widehat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$, then we can estimate $\boldsymbol{\beta}^*$ by solving the problem that

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\Sigma}}_{xx} - \widehat{\boldsymbol{\Sigma}}_{xy}^T\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1, \tag{5.2.5}$$

where $\widehat{\boldsymbol{\Sigma}}_{xx}$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ are submatrices of $\widehat{\boldsymbol{\Sigma}}$, $\|\boldsymbol{\beta}\|_1$ is an L_1-penalty function, and λ is a tuning parameter, which can be chosen by cross-validation. The problem in (5.2.5) is a convex optimization problem, which can be solved by a standard proximal gradient descent algorithm (Boyd and

Vandenberghe 2004). Next we discuss how to obtain $\widehat{\mathbf{\Sigma}}$.

The estimation of $\mathbf{\Sigma}$ by using the observed mixed data has been studied by Liu et al. (2012), Fan et al. (2017) and Feng and Ning (2019). For the estimation of $\mathbf{\Sigma}_{xx}$, the idea is to first estimate the correlation between X_j and X_k by the Kendall's tau correlation, then derive a bridge function to map it to the correlation between variables $\mathrm{f}_j(X_j)$ and $\mathrm{f}_k(X_k)$. The Kendall's tau for estimating the correlation between X_j and X_k is given by

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \mathrm{sgn}(X_{ij} - X_{i'j})\mathrm{sgn}(X_{ik} - X_{i'k}),\ 1 \le j, k \le p.$$

If $\mathbf{x} \sim NPN(\mathbf{0}, \mathbf{\Sigma}_{xx}, \mathbf{f}_x)$ with some unknown marginal transformation functions $\mathbf{f} = (\mathrm{f}_1, ..., \mathrm{f}_p)$, then Liu et al. (2012) and Fan et al. (2017) gave the bridge function F_{jk} for pairwise correlations among $\mathbf{x}$,

$$\widehat{F}_{jk}(r) = 2\mathrm{sin}^{-1}(r)/\pi, \text{ for } 1 \le j, k \le p,$$

where $\mathcal{C}$ denotes the set of continuous variables. Then an estimator for Σ_{jk} can be obtained by $\widetilde{\Sigma}_{jk} = \sin(\pi \cdot \widehat{\tau}_{jk}/2)$.

For the estimation of $\mathbf{\Sigma}_{xy}$, the idea is to use the ensemble method by Feng and Ning (2019). We first dichotomize Y to binary variables $Y_i^{(k)} = I(Y_i \ge k)$, $k = 1, ..., K$, estimate the correlation between X_j and $Y^{(k)}$ by the Kendall's tau correlation,

$$\widehat{\tau}_j^{(k)} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \mathrm{sgn}(Y_i^{(k)} - Y_{i'}^{(k)})\mathrm{sgn}(X_{ij} - X_{i'j}),\ k = 1, ..., K,$$

and use a bridge function to map $\widehat{\tau}_{jk}^{(k)}$ to the correlation between $\mathrm{f}_j(X_j)$ and the latent variable $\mathrm{f}_y(Z)$. The distribution of $(Y^{(k)}, X_j)$ satisfies the model proposed by Fan et al. (2017), which gave the bridge function between $Y^{(k)}$ and X_j by the following,

$$\widehat{F}_j(r) = 4\Phi_2(\widehat{\Delta}_y^{(k)}, 0, r/\sqrt{2}) - 2\Phi(\widehat{\Delta}_y^{(k)}),\ \text{for } 1 \le j \le p,$$

where Φ_d is the cumulative distribution function of the d-dimensional standard normal distribution,

$\widehat{\Delta}_y^{(k)}$ estimates $\widehat{\Delta}_y^{(k)} = \mathrm{f}_y(C_k)$ and has the following expression

$$\widehat{\Delta}_y^{(k)} = \Phi^{-1}(1 - (1/n)\sum_{i=1}^{n} Y_i^{(k)}), k = 1, ..., K. \tag{5.2.6}$$

Then, the latent correlation $\widetilde{\Sigma}_{jy}^{(k)}$ based on $Y^{(k)}$ can be obtained by solving $\widehat{F}_j(\widetilde{\Sigma}_{jy}^{(k)}) = \widehat{\tau}_j^{(k)}$. Finally, Feng and Ning (2019) proposed to use the weighted average of these latent correlations to obtain the point estimator of the correlation between Y and $\mathbf{x}$, which has the form of

$$\widetilde{\Sigma}_{jy} = \sum_{k=1}^{K} \widetilde{\Sigma}_{jy}^{(k)} w_{jy}^{(k)},$$

where the weights must satisfy $0 \leq w_{jy}^{(k)} \leq 1, \sum_{k=1}^{K} w_{jy}^{(k)} = 1$. For simplicity, it suffices to use $w_{jy}^{(k)} = 1/K$.

We remark that we do not need to estimate the marginal transformation functions when applying these estimators. Besides, Fan et al. (2017) have proved that all these bridge functions are invetible.

In order to be used in our proposed problem (5.2.5), we further need the estimator for $\mathbf{\Sigma}_{xx}$ to be positive definite. Then, we project $\widetilde{\mathbf{\Sigma}}_{xx}$ into the cone of positive definite matrices by solving $\widehat{\mathbf{\Sigma}}_{xx} = \mathrm{argmin}_{\mathbf{\Sigma}>0} \|\widetilde{\mathbf{\Sigma}}_{xx} - \mathbf{\Sigma}_{xx}\|_{\max}$. Such a problem has been studied by Zhao et al. (2014), who proposed to replace the elementwise maximum loss function with a smooth surrogate and apply accelerated proximal gradient algorithm to solve it. Finally, $\widehat{\mathbf{\Sigma}}$ is obtained once the submatrix $\widetilde{\mathbf{\Sigma}}_{xx}$ of $\widetilde{\mathbf{\Sigma}}$ is replaced by $\widehat{\mathbf{\Sigma}}_{xx}$.

To estimate the marginal transformation for the features $\mathbf{f}_x$, we propose the following estimator

$$\widehat{\mathrm{f}}_j(t) = \Phi^{-1}(\widetilde{F}_j(t)) \text{ for } j = 1, ..., p, \tag{5.2.7}$$

where $\widetilde{F}_j(t)$ is the winsorized empirical c.d.f of the jth continuous feature with the form

$$\widetilde{F}_j(t) = \delta_n I(\widehat{F}_j(t) < \delta_n) + \widehat{F}_j(t) I(\delta_n \leq \widehat{F}_j(t) \leq 1 - \delta_n) + (1 - \delta_n) I(\widehat{F}_j(t) > 1 - \delta_n),$$

where $\widehat{F}_j(t) = (1/n)\sum_{i=1}^{n} I(X_{ij} \leq t)$ is the empirical c.d.f of the jth continuous feature and δ_n is often chosen to be $1/(2n)$.

Based on our proposed estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\Delta}_y^{(k)}$ $(k = 1, ..., K)$, and $\widehat{\mathbf{f}} = (\widehat{\mathrm{f}}_1, ..., \widehat{\mathrm{f}}_p)$, we next study their statistical properties in parameter estimation and theoretical performance in classification.

5.3 Theoretical Properties

We start with introducing some notations. For vector $\mathbf{a} \in \mathbb{R}^p$, let $\|\mathbf{a}\|_\infty = \max_{1\le j\le p} |a_j|$, $\|\mathbf{a}\|_1 = \sum_{j=1}^p |a_j|$, $\|\mathbf{a}\|_2 = (\sum_{j=1}^p a_j^2)^{1/2}$ denote its max, L_1-, and Euclidean norms, respectively. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times p}$, let $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, $\|\mathbf{A}\|_\infty = \max_i \sum_{1\le j\le p} |a_{ij}|$, $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ be the minimum and maximum eigenvalues of $\mathbf{A}$ respectively. For any symmetric matrix $\mathbf{\Sigma}$, we write $\mathbf{\Sigma} > \mathbf{0}$ if $\lambda_{\min}(\mathbf{\Sigma}) > 0$. For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \le cb_n$. $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

First we rely on the general M-estimation theory (Negahban et al. 2012) to study the statistical properties of $\widehat{\boldsymbol{\beta}}$. We assume that $\boldsymbol{\beta}^* \in B_q(R_q) = \left\{\boldsymbol{\theta} \in \mathbb{R}^p : \sum_{1\le j\le p} |\theta_j|^q \le R_q\right\}$ where $q \in [0, 1)$ is a fixed constant. Without loss of generality, let $\|\boldsymbol{\beta}^*\|_1 > m$ for some $m > 0$. The following theorem quantifies the estimation error of $\widehat{\boldsymbol{\beta}}$ in the Euclidean and L_1 norms respectively.

Theorem 5.1. *Suppose the following conditions hold.*

Condition 1. $\max_{1\le j<k\le p} |\Sigma_{jk}| \le 1 - \delta$ *for some* $\delta > 0$.

Condition 2. $\max_{k=1,\dots,N_j} |\Delta_y^{(k)}| \le M$ *for some* $M > 0$.

Condition 3. $m \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le M$ *for some* m *and* $M > 0$.

If choosing $\lambda = C\|\boldsymbol{\beta}^*\|_1\sqrt{(\log p)/n}$ *for some sufficiently large constant* C*, and* $R_q\lambda^{1-q} = o(1)$*, then there exist generic constants* C_1 *and* C_2 *such that*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \lesssim R_q\lambda^{2-q} \text{ and } \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1^2 \lesssim R_q^2\lambda^{2(1-q)}$$

with probability at least $1 - C_1p^{-C_2}$.

Conditions 1-2 are standard technical conditions required for uniform convergence of $\widehat{\mathbf{\Sigma}}$, they also appears in Fan et al. (2017) and Feng and Ning (2019). Condition 3 is a technical condition needed in the proof of Theorem 5.1. Given these conditions, Theorem 5.1 shows that $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^*$ under weak sparsity assumption. The results in Theorem 5.1 are derived from Theorem 1 of Negahban et al. (2012), which also presents a similar convergence rate in Euclidean norm under weak sparsity given by Corollary 3. But the difference is that our choice of the tuning parameter λ has the term $\|\boldsymbol{\beta}^*\|_1$ while Corollary 3 in Negahban et al. (2012) does not have it. This difference is due to that we only know the convergence rate of $\widehat{\mathbf{\Sigma}}$ in $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$, so we can only bound $\|\widehat{\mathbf{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\mathbf{\Sigma}}_{xy}\|_\infty$ by $\|\widehat{\mathbf{\Sigma}}_{xx} - \mathbf{\Sigma}_{xx}\|_{\max}\|\boldsymbol{\beta}^*\|_1 + \|\widehat{\mathbf{\Sigma}}_{xy} - \mathbf{\Sigma}_{xy}\|_\infty$, while Negahban et al. (2012) assumes linear regression model directly using observed features $\mathbf{X}$ with sub-Gaussian columns and obtain fast convergence rate for $\|(\mathbf{X}^T\boldsymbol{\epsilon})/n\|_\infty$. If we impose an assumption that $\|\boldsymbol{\beta}^*\|_1 < \infty$, then our estimator $\widehat{\boldsymbol{\beta}}$ can reach the fast convergence rate in Euclidean norm as in Corollary 3 of Negahban et al. (2012). The convergence rate in L_1 norm is obtained by multipying an upper bound for the compatibility constant to the convergence rate in Euclidean norm. Next we show that $\widehat{\boldsymbol{\beta}}$ has variable selection consistency under exact sparsity.

We assume $\boldsymbol{\beta}^*$ is exactly sparse that $\boldsymbol{\beta}^* \in B_0(R_0)$, and define $\mathcal{M} = \{j : \beta_j \neq 0\}$, $R_0 = \|\mathcal{M}\|_0 = \sum_{j=1}^p I(\beta_j \neq 0)$, and $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$ and $\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}$ are submatrices of $\mathbf{\Sigma}_{xx}$. In the following theorem, we show the variable selection consistency of $\widehat{\boldsymbol{\beta}}$ under exact sparsity.

Theorem 5.2. *Suppose Conditions 1 - 3 and the following conditions hold.*

Condition 4. $\|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \leq M$ *for some* $M > 0$.

Condition 5. $\|\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \leq (1-\alpha)(1-\varepsilon)$ *for some* $\alpha > 0$ *and* $\varepsilon > 0$.

Condition 6. $\min_{j\in\mathcal{M}} |\beta_j^*| \gg (\|\boldsymbol{\beta}^*\|_1\sqrt{(\log p)/n})^\gamma$ *for some* $0 < \gamma < 1$.

If $R_0^2\sqrt{(\log p)/n} = o(1)$*, and* $\lambda = C(\|\boldsymbol{\beta}^*\|_1\sqrt{(\log p)/n})^\gamma$*, where* $0 < \gamma < 1$ *and* C *is some sufficiently large constant, then with probability at least* $1 - C_1 p^{-C_2}$*, we have* $\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}\|_\infty \lesssim \lambda$ *and* $\widehat{\mathcal{M}} = \mathcal{M}$.

Condition 4 requires that $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$ is invertible and assumes that the sup-norm of its inverse is bounded by a constant. Condition 5 is a standard irrepresentable condition that requires the important and unimportant variables cannot be highly correlated. It is well known that such a condition is needed for the variable selection consistency of the L_1-penalized methods. Condition 6 is a beta-min condition requiring that the minimal signal to be bounded away from zero. Given these conditions, Theorem 5.2 shows that $\widehat{\boldsymbol{\beta}}$ is variable selection consistent and gives uniformly consistent estimators of the nonzero components of $\boldsymbol{\beta}^*$. Theorem 5.2 will be useful when studying the misclassification error of the LMGCC rule, defined as $\mathcal{R}_{LMGCC}(\mathcal{D}) = \mathrm{E}(I(D_{LMGCC}(\mathbf{x}) \neq Y)|\mathcal{D})$ with $(\mathbf{x}, Y)$ coming from a new subject, compared to the Bayes error $\mathcal{R}_{Bayes}$, it gives explicit expressions for the probability of $\widehat{\mathcal{M}} = \mathcal{M}$ and the uniform convergence rate of $\widehat{\boldsymbol{\beta}}$. Next we analyze the classification oracle property of LMGCC.

Recall from (5.2.3) that the Bayes rule assigns a subject to the kth class if and only if

$$\Delta_y^{(k)} < \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* \leq \Delta_y^{(k+1)}. \tag{5.3.1}$$

The LMGCC rule assigns a subject to the kth class if and only if

$$\widehat{\Delta}_y^{(k)} < \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} \leq \widehat{\Delta}_y^{(k+1)}. \tag{5.3.2}$$

Theorem 5.3 evaluates the difference between the LMGCC rule and the Bayes rule, in terms of $\left|(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) - (\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)})\right|$ for $k = 1, ...K$.

Theorem 5.3. *When the conditions for Theorem 3.2 hold, if $R_0(n^b \log n)^{-1/2} = o(1)$ for some $0 < b < 1$, then we have the following for $k = 1, ..., K$.*

$$(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) - (\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)}) = O_p\left(R_0 \|\boldsymbol{\beta}^*\|_\infty \sqrt{(\log\log n)/n^{1-b/2}} + R_0 \lambda \sqrt{\log n}\right).$$

To prove Theorem 5.3, we need to know the $\widehat{\boldsymbol{\beta}}$, $\widehat{\Delta}_y^{(k)}$ $(k = 1, ..., K)$ and $\widehat{\mathbf{f}}_x$ are consistent. The consistency of $\widehat{\boldsymbol{\beta}}$ has been studied early in this section. The consistency of $\widehat{\Delta}_y^{(k)}$ has been studied in Fan et al. (2017) and Feng and Ning (2019) as they show that $\widehat{\Delta}_y^{(k)}$ is root-n consistent. Finally, the consistency of $\widehat{\mathrm{f}}_j$ has been given in Theorem 2 of Han et al. (2013), where they give a uniform convergence rate on an expanding interval $T_{jn} = [g_j(-\sqrt{b \log n}), g_j(\sqrt{b \log n})]$, where $g_j = \mathrm{f}_j^{-1}$ and $0 < b < 1$. Then given the events that $\widehat{\mathcal{M}} = \mathcal{M}$ and $X_j \in T_{jn}$ for all $j \in \mathcal{M}$, we can show that $(\widehat{\mathbf{f}_x}(\mathbf{x}) - \mathbf{f}_x(\mathbf{x}))_{\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}}^* = O_p(R_0 \|\boldsymbol{\beta}^*\|_\infty \sqrt{(\log\log n)/n^{1-b/2}})$ and $(\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}}^*)^T (\mathbf{f}_x(\mathbf{x}))_{\mathcal{M}} = O_p(R_0 \lambda \sqrt{\log n})$. The condition that $R_0(n^b \log n)^{-1/2} = o(1)$ controls $P(X_j \notin T_{jn}, \text{for some } j \in \mathcal{M}) = o(1)$. With Theorem 3, we can apply it to further show that the misclassification error rate of LMGCC rule is consistent to the Bayes error in the following corollary.

Corollary 5.1. *Suppose the conditions for Theorem 3.2 hold.*
If $R_0\left(\|\boldsymbol{\beta}^*\|_\infty \sqrt{(\log\log n)/n^{1-b/2}} + \lambda\sqrt{\log n} + (n^b \log n)^{-1/2}\right) = o(1)$ for some $0 < b < 1$, then $\mathrm{E}(\mathcal{R}_{LMGCC}(\mathcal{D})) = \mathcal{R}_{Bayes} + o(1)$.

Corollary 5.1 shows that the average misclassification error of LMGCC can converge to the Bayes error under exact sparsity. In the following section, we will demonstrate this property through simulation studies.

5.4 Simulations

We investigate the numerical performance of our proposed method under four different scenarios. In each of the four scenarios, we compare the classification performance of the proposed LMGCC method with some well-known classifiers for multi-class classification, such as the support vector machine (SVM) with radial basis function kernel, the K-nearest-neighbors algorithm (KNN), the random forest (RF), the multinomial logistic regression with an L_1-penalty (Multinomial), and the ordinal probit regression with an L_1-penalty (Ordinal) mentioned in Section 5.2. Among these candidate classifiers, the SVM, the KNN, and the RF are machine-learning-based classifiers that do not model the distribution for the response and the features; the multinomial regression is a parametric classifier that directly models $P(Y = k|\mathbf{x})$ without incorporating the order of label k, while the ordinal regression is a parametric classifier that models $P(Y \leq k|\mathbf{x})$ to handle the label order. We also compare the classification performance of the LMGCC method with the Bayes rule to examine the error rate consistency property in Corollary 3.1. We design the first two scenarios with balanced class labels, so we aim to compare the overall misclassification error $r = (1/n_{test}) \sum_{i=1}^{n_{test}} I(\widehat{y}_{test,i} \neq y_{test,i})$ of these classifiers, where n_{test} denotes the sample size in the test set. While we design the other two scenarios with unbalanced class labels, the overall misclassification error becomes inadequete to distinguish the prediction accuracy for ordinal response since we are interested in the misclassification error in both the dominant and the minority class. We use the average within-group error rate as an additional performance evaluation metric with the following form

$$\widetilde{r} = \frac{1}{K+1} \sum_{k=0}^{K} \left\{ \frac{1}{n_{test,k}} \sum_{i:y_{test,i}=k} I(\widehat{y}_{test,i} \neq y_{test,i}) \right\},$$

where $n_{test,k}$ denotes the sample size for those test samples with label k. This metric was introduced previously in Qiao and Liu (2009). In each of the four scenarios, we set the training and the test set sizes to be $n = 200$ and $n_{test} = 400$ respectively, and consider $p = 50$ and $p = 500$. The setup for the four scenarios are as follows.

Scenario 1: $\mathbf{\Sigma}_{xx} = (\sigma_{xx,ij})_{p\times p}$ where $\sigma_{xx,ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. $\boldsymbol{\beta}^* = (0.18, ..., 0.18, 0, ..., 0)^T$, where the first 10 elements are 0.18 and the rest are zeros. $\mathbf{\Sigma}_{xy} = \mathbf{\Sigma}_{xx}\boldsymbol{\beta}^*$, and $\mathbf{\Sigma}$ is obtained given $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{xy}$. We generate $(\mathrm{f}_y(Z), \mathbf{f}_x(\mathbf{x}))$ from $N(\mathbf{0}, \mathbf{\Sigma})$ and choose marginal transformations $\mathrm{f}_j(X_j) = X_j^3$ for $1 \leq j \leq p$. Set three classes to be balanced, i.e. $P(Y = k) = 1/3$ for $k = 0, 1, 2$, with $\Delta_y^{(1)} = -0.43$ and $\Delta_y^{(2)} = 0.43$ and the observed response is generated by $Y = \sum_{k=1}^{2} I(\mathrm{f}_y(Z_y) > \Delta_y^{(k)})$.

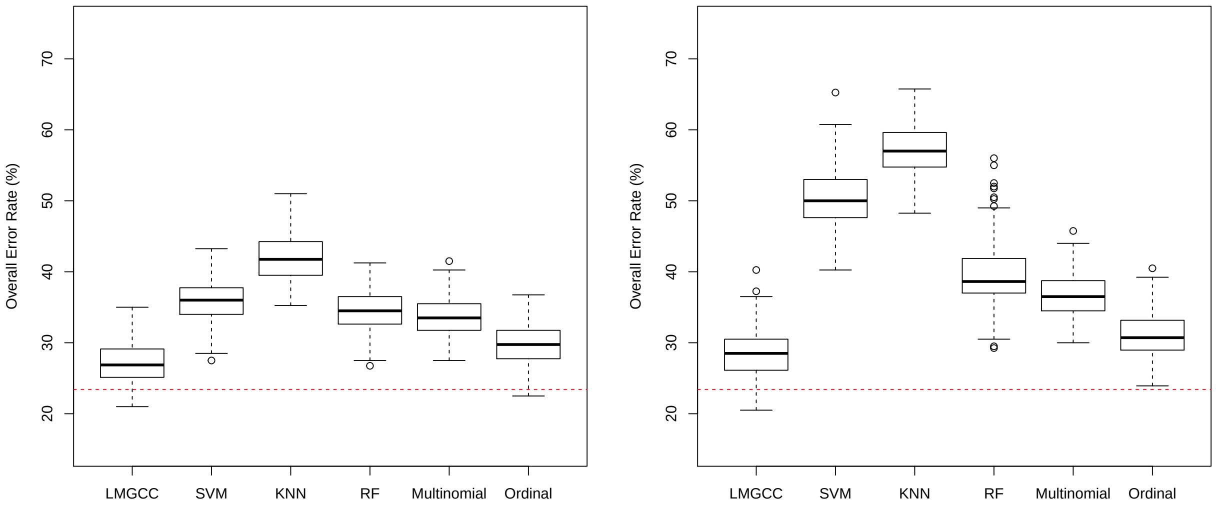
Scenario 2: All settings remain the same as Scenario 1, except the marginal transformations $\mathrm{f}_j(Z_j) = \log(Z_j)$ for $1 \leq j \leq p$.

Scenario 3: All settings remain the same as Scenario 1, except the three classes are unbalanced, i.e. $P(Y = 0) = 1/4$, $P(Y = 1) = 1/2$, $P(Y = 2) = 1/4$, with $\Delta_y^{(1)} = -0.7$ and $\Delta_y^{(2)} = 0.7$.

Scenario 4: All settings remain the same as Scenario 3, except the marginal transformations $\mathrm{f}_j(Z_j) = \log(Z_j)$ for $1 \leq j \leq p$.

For each scenario, we independently generate n samples for the training set and n_{test} samples for the test set, by generating $(\mathrm{f}_y(Z), \mathbf{f}_x(\mathbf{x}))$ from $N(\mathbf{0}, \mathbf{\Sigma})$, obtaining $\mathbf{x}$ by $\mathbf{f}_x^{-1}(\mathbf{f}_x(\mathbf{x}))$ and Y by $\sum_{k=1}^{2} I(\mathrm{f}_y(Z) > \Delta_y^{(k)})$. For the proposed LMGCC method, we obtain the estimator for the regression coefficients by solving (5.2.5), obtain the estimators for transformations by (5.2.7), and obtain the estimators for the two thresholds by (5.2.6), then the LMGCC rule is given by (5.2.4). We also use the performance of the Bayes rule given by (5.2.3) as a benchmark for each scenario; see red lines in Figures (5.1) to (5.4). We implement SVM using the **e1071** package, implement KNN using the **class** package, implement RF using the **randomForest** package, implement multinomial logistic regression with L_1-penalty using the **glmnet** package, and implement ordinal probit regression with L_1-penalty using the **ordinalNet** package. The optimal tuning parameters for each method are chosen by a grid search using five-fold cross-validation. For each scenario we repeat simulations for 100 times, report the overall misclassification error for all six methods in all scenarios and the average within-group error for all six methods in Scenario 3 and 4.

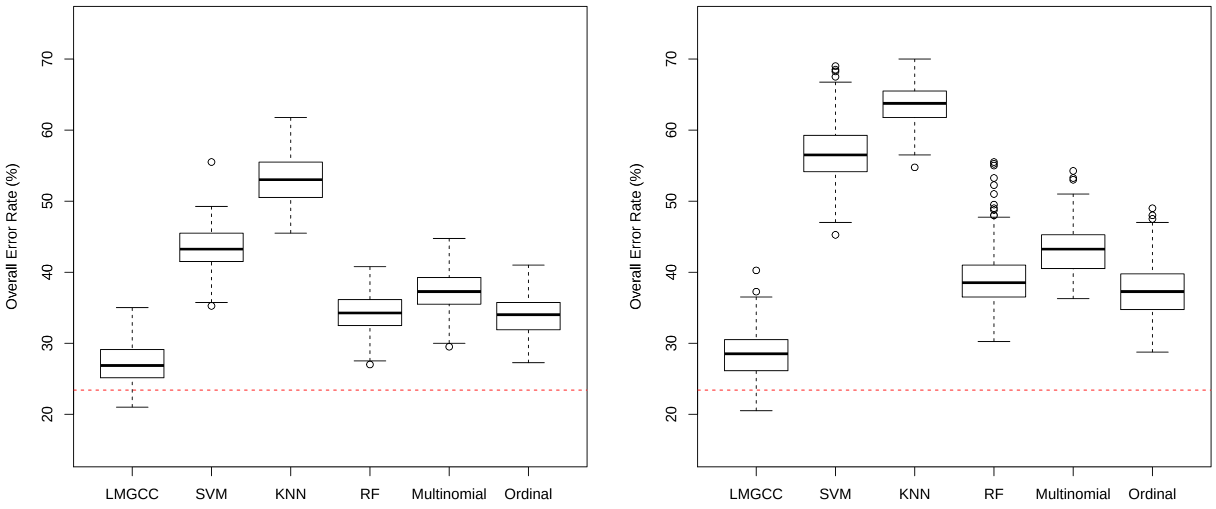
It is seen from Figure (5.1) that LMGCC has clear advantage in overall misclassification error over other methods in Scenario 1. Such advantage becomes more obvious when the dimension increases. These results are reasonable as only LMGCC identifies the joint normality structure between the latent variables for the response and the features, while allowing a nonlinear relationship between the latent response and the features. Among the competing nonparametric methods, kernel-based SVM looks for boundary that gives large margin for the mapped observed features, KNN assigns class label based on the Euclidean distance of the observed features, and random forest aggregates classification trees for observed training data, none of them builds classification rule based



(a) $p = 50$

(b) $p = 500$

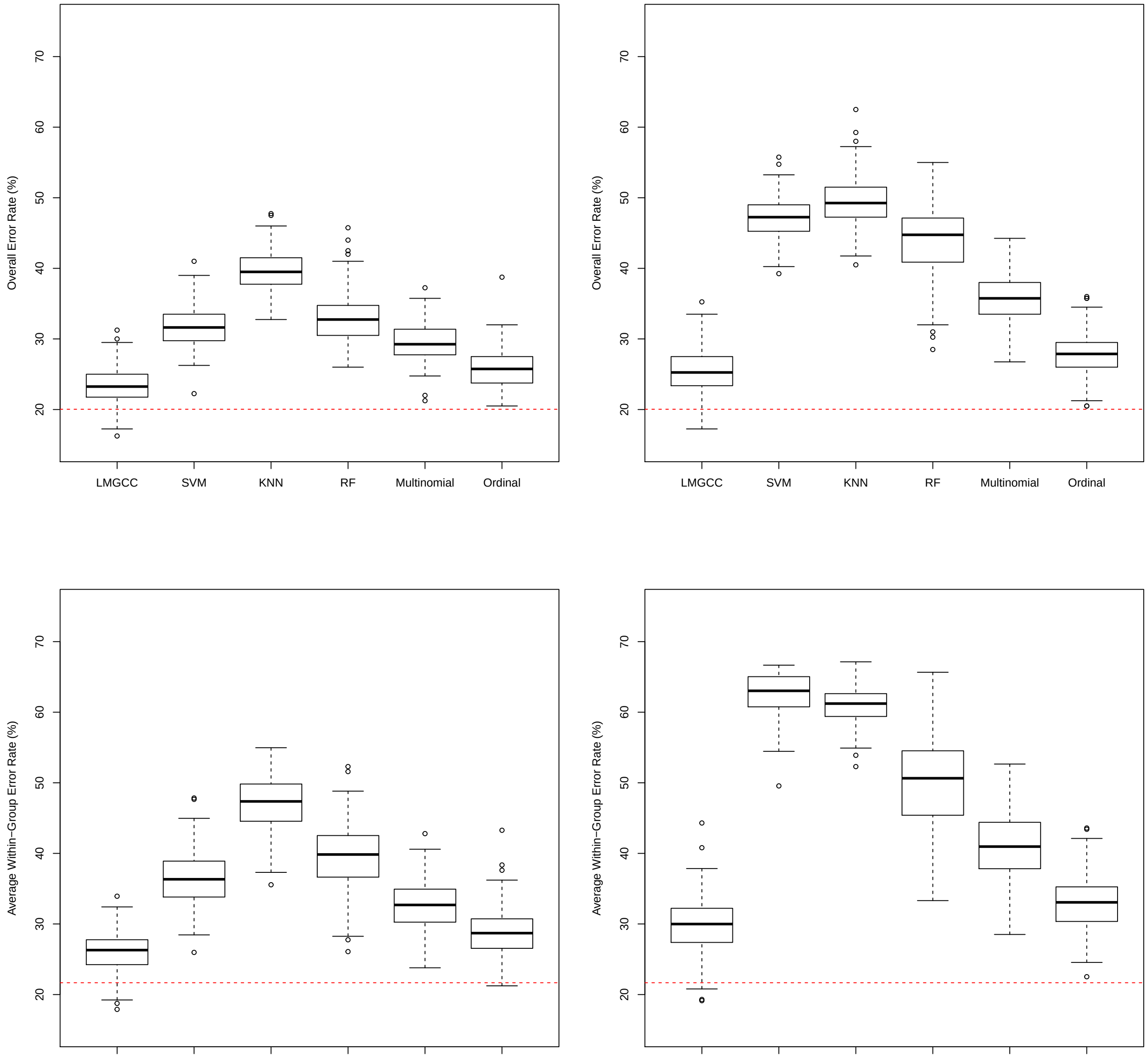
Figure 5.1. Comparison of overall error rate for the six competitors in Scenario 1. Red lines indicate the Bayes errors.



(a) $p = 50$

(b) $p = 500$

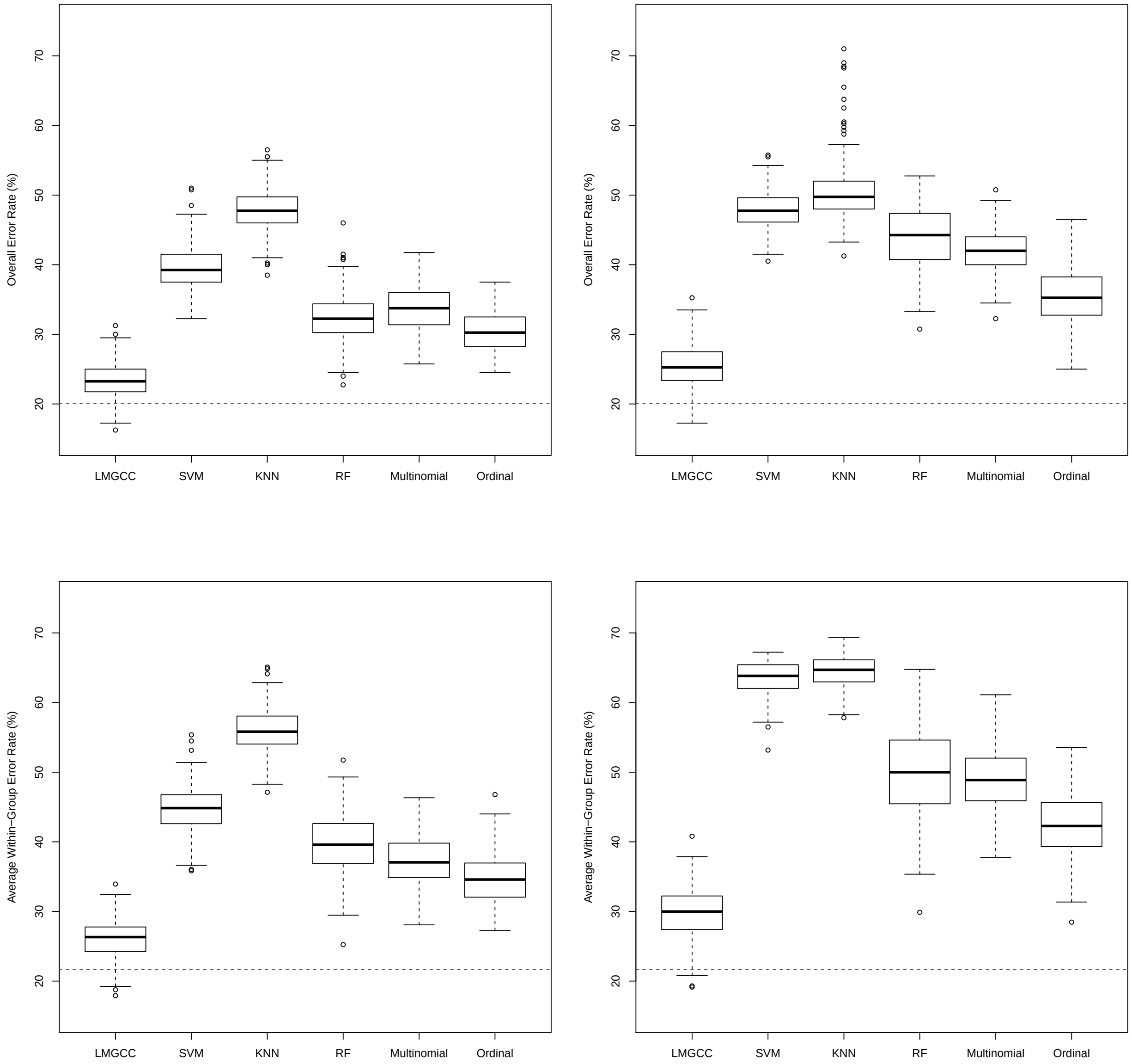
Figure 5.2. Comparison of overall error rate for the six competitors in Scenario 2. Red lines indicate the Bayes errors.



(a) $p = 50$

(b) $p = 500$

Figure 5.3. Comparison of overall and average within-group error rate for the six competitors in Scenario 3. Red lines indicate overall and average within-group error rates for the Bayes rule.



(a) $p = 50$

(b) $p = 500$

Figure 5.4. Comparison of overall and average within-group error rate for the six competitors in Scenario 4. Red lines indicate the overall and average within-group error rates for the Bayes rule.

on modeling the distribution of the response and the features. As for the competing parametric methods, both the multinomial regression and the ordinal regression classify a subject based on the posterior probability of each class given observed features, but the multinomial regression ignores the order of the labels and the model is misspecified, while the ordinal regression misspecifies the functional form of the features.

Figure (5.2) shows similar results as Figure (5.1). Furthermore, results in Figure (5.2) show our proposed LMGCC is robust to arbitrarily skewed monotone transformations, while other competing methods except random forest are susceptible in multi-class classification when the marginal transformations become more skewed.

Figure (5.3) shows both the overall and the average within-group error rate for all competitors in Scenario 3. In this scenario, the three class labels have proportion $(0.25, 0.5, 0.25)$. We are interested in the misclassification error in all three classes, so the average within-group error becomes more informative about which classifier has better multi-class classification performance. It is seen that LMGCC demonstrates even more significant advantage on the average within-group error over other competitors, and it maintains excellent classification performance for each class even when there is an unbalanced class proportion, while other competitors shows inferior average within-group error rate under both low and high dimension.

Finally, Figure (5.4) shows similar results as Figure (5.3), and again demonstrates that LMGCC has robust classification performance invariant of the change of marginal transformations even when the class proportion becomes unbalanced.

5.5 Real Data Analysis

The Wisconsin breast cancer data from the UCI Machine Learning Repository contains two datasets. The diagnostic dataset contains patient information with baseline breast cancer diagnostic results (benign or malignant). The prognostic dataset contains follow-up data for breast cancer cases, i.e. those who were diagnosed to be malignant at the baseline, including only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. During the follow-up, some cases might experience recurrence of breast cancer, indicating these are more severe cases. In this example, our goal is to use the baseline features, which are 30 variables computed from a digitized image of a fine needle aspiration (FNA) of a breast mass, to predict the status of breast cancer progression being either benign, malignant but non-recurrent, and recurrent. We combine

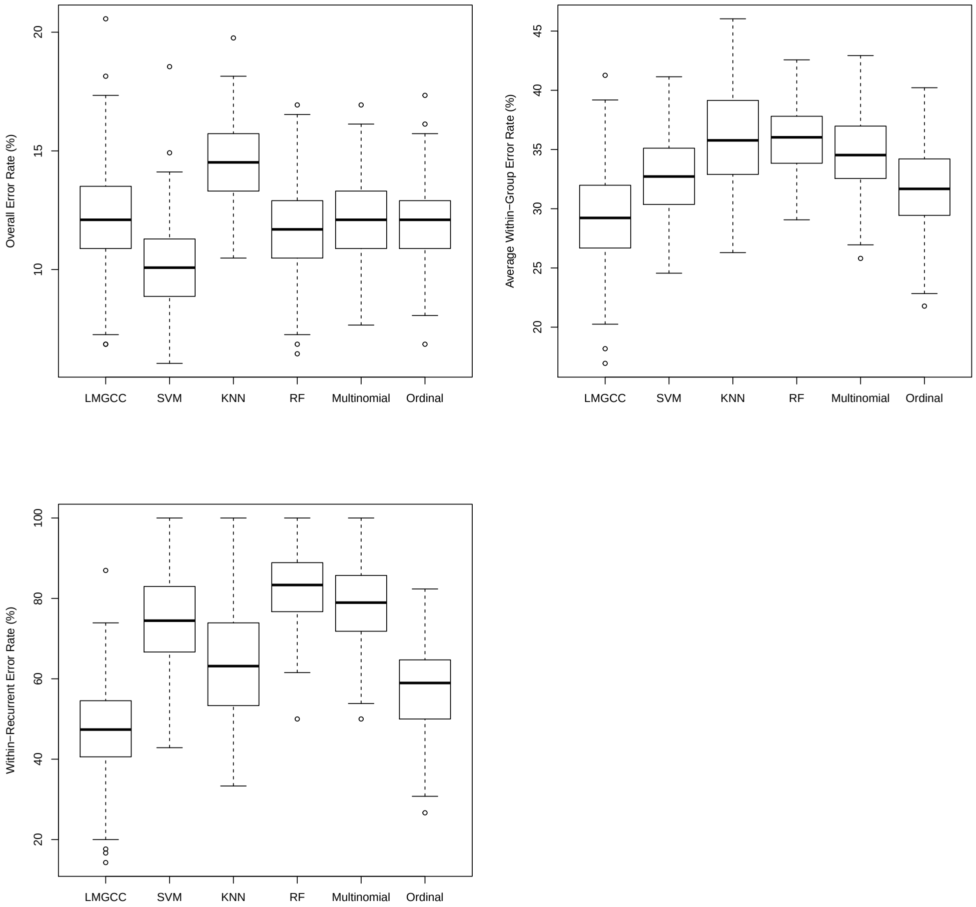


Figure 5.5. Performance comparison of six competitors on classifying breast cancer progression status.

all subjects that were diagnosed as benign at the baseline with all subjects that were diagnosed as malignant at the baseline and had follow-up status (non-recurrent or recurrent). For all the subjects, the baseline FNA image features are all available. The full dataset for our analysis contains 496 samples with three possible status and 30 features. We want to point out that the class proportion for the cancer progression status is unbalanced with 72% samples being benign, 21% samples being malignant but non-recurrent, 7% samples being recurrent, and we are particularly interested to see how well our method can predict the recurrent status. Therefore, we apply our proposed LMGCC along with SVM, KNN, random forest, multinomial logistic regression with L_1-penalty, and ordinal probit regression with L_1-penalty to this dataset and compare three performance evaluation metric: the overall misclassification error, the average within-group error, and the within-recurrent-group error. We randomly split the full data into a training set and a test set with same number of samples, then apply the six competitors separately on the training set with the optimal tuning parameters being chosen by five-fold cross-validation, and calculate their performance evaluation metrics on the test set for comparison. The above procedures are repeated 100 times and the results are shown in Figure (5.5).

The comparison of overall misclassification error shows that all classifiers yielded similar results on this highly unbalanced dataset, which is not informative because over 70% samples are benign, and classifying some of those rare malignant or even recurrent subjects to benign status can still result in good overall classification performance. In this example, we particularly care about making correct classification for those malignant and specifically recurrent subjects, so our focus will be mainly the average within-group error and the within-recurrent error. It is shown that LMGCC outperforms all other competitors by 2-5% in terms of the average within-group error. The advantage of LMGCC is much more significant when comparing the classification performance for those recurrent subjects, resulting in over 10% reduction in within-recurrent error. These results are reasonable because the baseline characteristics of the cell nuclei already indicate the severity of breast cancer, which can be viewed as the latent variable corresponding to the progression status, and LMGCC captures the relationship between the latent response and the features, which could lead to large advantage in predicting the cancer progression status.

5.6 Technical Details

5.6.1 Proofs of Main Theorems

Proof of Theorem 5.1 The proof of Theorem 5.1 modifies the proof of Corollary 3 in Negahban et al. (2012). Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Define $\mathbf{C} = \{\boldsymbol{\Delta} \in \mathbf{R}^p : \|\boldsymbol{\Delta}_{\mathbf{S}^c}\|_1 \leq 3\|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 + 4\|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1\}$, where $\mathbf{S} = \{i \in \{1, 2, .., p\} : \left|(\boldsymbol{\eta}_j)_i\right| > \eta\}$ and $\eta > 0$ is some threshold to be chosen. For any $\boldsymbol{\Delta} \in \mathbf{C}$, we have

$$\begin{aligned}
\|\boldsymbol{\Delta}\|_1 &\leq \|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 + \|\boldsymbol{\Delta}_{\mathbf{S}^c}\|_1 \leq 4\|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 + 4\|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1, \\
\|\boldsymbol{\Delta}_{\mathbf{S}}\|_1 &\leq \sqrt{|\mathbf{S}|}\|\boldsymbol{\Delta}\|_2 \leq \sqrt{R_q}\eta^{-q/2}\|\boldsymbol{\Delta}\|_2.
\end{aligned}$$

Next, we verify the RSC condition for $\boldsymbol{\Delta} \in \mathbf{C}$. Using (5.6.5) in the proof of Lemma 5.2, with probability at least $1 - C_1 p^{-C_2}$ for some generic positive constants C_1 and C_2 and a sufficiently large constant C, we have

$$\begin{aligned}
(1/2)\boldsymbol{\Delta}^T\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\Delta} &= (1/2)\boldsymbol{\Delta}^T\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\Delta} + (1/2)\boldsymbol{\Delta}^T(\widehat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma}_{xx})\boldsymbol{\Delta} \\
&\geq (m/2)\|\boldsymbol{\Delta}\|_2^2 - C\sqrt{\log p/n}\|\boldsymbol{\Delta}\|_1^2 \\
&\geq (m/2)\|\boldsymbol{\Delta}\|_2^2 - 32C\sqrt{\log p/n}(\|\boldsymbol{\Delta}_{\mathbf{S}}\|_1^2 + \|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1^2) \\
&\geq (m/2 - 32C\sqrt{\log p/n}R_q\eta^{-q})\|\boldsymbol{\Delta}\|_2^2 - 32C\sqrt{\log p/n}\|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1^2.
\end{aligned}$$

If we choose $\eta = \|\boldsymbol{\beta}^*\|_1\sqrt{\log p/n}$, and $R_q\|\boldsymbol{\beta}^*\|_1^{-q}(\log p/n)^{(1-q)/2} = o(1)$ since $R_q\lambda^{1-q} = o(1)$, then we have the following for large n

$$32C\sqrt{\log p/n}R_q\eta^{-q} = 32C \cdot R_q\|\boldsymbol{\beta}^*\|_1^{-q}(\log p/n)^{(1-q)/2} \leq m/4.$$

Thus, it holds with probability at least $1 - C_1 p^{-C_2}$ that

$$(1/2)\boldsymbol{\Delta}^T\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\Delta} \geq (m/4)\|\boldsymbol{\Delta}\|_2^2 - 32C\sqrt{\log p/n}\|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1^2. \tag{5.6.1}$$

Based on Lemma 5.2 and (5.6.1), we apply Theorem 1 of Negahban et al. (2012) with $\kappa_{\mathcal{L}} = m/4$, $\Psi^2(\bar{\mathcal{M}}) = |\mathbf{S}|$, $\tau_{\mathcal{L}}^2 = 32C\sqrt{\log p/n}\|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1^2$ and $\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) = \|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1$. Using the result that

$$\|\boldsymbol{\beta}^*_{\mathbf{S}^c}\|_1 \leq R_q \eta^{1-q},$$

we can simplify the result from Theorem 1 of Negahban et al. (2012) and obtain that with probability at least $1 - C_1 p^{-C_2}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \lesssim \|\boldsymbol{\beta}^*\|_1^{2-q} R_q (\frac{\log p}{n})^{1-q/2} + \|\boldsymbol{\beta}^*\|_1^{3-2q} R_q^2 (\frac{\log p}{n})^{2-q}.$$

Since $\|\boldsymbol{\beta}^*\|_1^{3-2q} R_q^2 (\log p/n)^{2-q} = \left(\|\boldsymbol{\beta}^*\|_1^{2-q} R_q (\log p/n)^{1-q/2}\right)^2 \|\boldsymbol{\beta}^*\|_1^{-1}$, $R_q \lambda^{2-q} = o(1)$ because $R_q \lambda^{1-q} = o(1)$, and $\|\boldsymbol{\beta}^*\|_1^{-1} < 1/m$, we may conclude that, with probability at least $1 - C_1 p^{-C_2}$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \lesssim R_q \lambda^{2-q}.$$

Furthermore we bound $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1^2$.

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1^2 &\leq \Psi^2(\bar{\mathcal{M}}) \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \\
&\leq R_q \eta^{-q} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \\
&\lesssim R_q^2 \|\boldsymbol{\beta}\|_1^{2(1-q)} (\log p/n)^{1-q} + R_q^3 \|\boldsymbol{\beta}^*\|_1^{3(1-q)} (\log p/n)^{2-3q/2}.
\end{aligned}$$

Since $R_q^3 \|\boldsymbol{\beta}^*\|_1^{3(1-q)} (\log p/n)^{2-3q/2} = \left(R_q^2 \|\boldsymbol{\beta}\|_1^{2(1-q)} (\log p/n)^{1-q}\right)^{3/2} \cdot (\sqrt{\log p/n})$, and $R_q^2 \lambda^{2(1-q)} = o(1)$, we can conclude that, with probability at least $1 - C_1 p^{-C_2}$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1^2 \lesssim R_q^2 \lambda^{2(1-q)}.$$

Proof of Theorem 5.2

Proof. By the standard convex optimization theory, any $\boldsymbol{\beta} \in \mathbf{R}^p$ satisfying the following Karush–

Kuhn–Tucker conditions (Boyd and Vandenberghe 2004) is the solution to (5.2.5).

$$(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta})_j - \widehat{\boldsymbol{\Sigma}}_{xy,j} + \lambda \text{sign}(\beta_j) = 0, \ for \ j \in \mathcal{M}; \tag{5.6.2}$$

$$\left|(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta})_j - \widehat{\boldsymbol{\Sigma}}_{xy,j}\right| < \lambda, \ for \ j \notin \mathcal{M}; \tag{5.6.3}$$

$$\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}) > 0. \tag{5.6.4}$$

We first show that there exists a solution $\widehat{\boldsymbol{\beta}}_{\mathcal{M}} \in \mathbf{R}^{R_0}$ to (5.6.2) in the neighbourhood $\mathcal{N} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\mathcal{M}}\|_\infty \leq C\lambda\}$ with probability at least $1 - C_1 p^{-C_2}$. We have

$$(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta})_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}}.$$

It follows from Lemma 5.2 that

$$P(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}}\|_\infty \geq C a_n) \leq C_1 p^{-C_2},$$

where $a_n = \|\boldsymbol{\beta}^*\|_1 \sqrt{(\log p)/n}$. Let $\boldsymbol{\tau} = (\tau_j) \in \mathbf{R}^p$ with $\tau_j = \text{sign}(\beta_j)$ for $j \in \mathcal{M}$ and $\tau_j = 0$ for $j \notin \mathcal{M}$,

$$\begin{aligned} f(\boldsymbol{\beta}_{\mathcal{M}}) &= \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}, \\ g(\boldsymbol{\beta}_{\mathcal{M}}) &= \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}} f(\boldsymbol{\beta}_{\mathcal{M}}) = \boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}} + \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}\}. \end{aligned}$$

By Lemma 5.3, for some positive constant M, we have

$$P(\|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty \geq 2M) \leq C_1 p^{-C_2}.$$

Hence, by the stated choice of λ, with probability at least $1 - C_1 p^{-C_2}$, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}\}\|_\infty &\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty \|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}\}\|_\infty \\ &\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M}\mathcal{M}}\|_\infty (\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}}\|_\infty + \lambda) \\ &\leq 2M(Ca_n + \lambda) \lesssim \lambda. \end{aligned}$$

Hence, when n is sufficiently large, if $(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}})_j = C\lambda$ for some sufficently large $C > 0$,

$$g(\boldsymbol{\beta}_{\mathcal{M}})_j = C\lambda - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}})_j \geq 0,$$

and if $(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}})_j = -C\lambda$,

$$g(\boldsymbol{\beta}_{\mathcal{M}})_j = -C\lambda - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}})_j \leq 0.$$

By the continuity of $g(\boldsymbol{\beta}_{\mathcal{M}})$ and Miranda's existence theorem, $g(\boldsymbol{\beta}_{\mathcal{M}}) = 0$ has a solution $\widehat{\boldsymbol{\beta}}_{\mathcal{M}}$ in $\mathcal{N}$ with probability tending to 1.

Second, we verify that $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathbf{0})^T$ also satisfies (5.6.3). We have

$$(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta})_{\mathcal{M}^c} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}^c} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}^c} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) + (\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\Sigma}}_{xy})_{\mathcal{M}^c}.$$

Since $g(\boldsymbol{\beta}_{\mathcal{M}}) = 0$, we have

$$(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta})_{\mathcal{M}^c} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}^c} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}\boldsymbol{\beta}^*_{\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}} + \lambda\boldsymbol{\tau}_{\mathcal{M}}) + (\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\Sigma}}_{xy})_{\mathcal{M}^c}.$$

By similar arguments as in Lemma 5.2, we have

$$P(\|(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\Sigma}}_{xy})_{\mathcal{M}^c}\|_\infty \geq Ca_n) \leq C_1 p^{-C_2}.$$

By Lemma 5.4, we have

$$P(\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \geq (1-\alpha)(1-\epsilon/2)) \leq C_1 p^{2-C_2C}.$$

Then, with probability at least $1 - C_1 p^{-C_2}$,

$$\begin{aligned}
\|(\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta})_{\mathcal{M}^c} - \widehat{\boldsymbol{\Sigma}}_{xy,\mathcal{M}^c}\|_\infty &\leq (1-\alpha)(1-\epsilon/2)(Ca_n + \lambda) + Ca_n \\
&\leq (1-\alpha)(1-\epsilon/2)\lambda + (2-\epsilon/2)Ca_n \\
&< (1-\alpha)\lambda,
\end{aligned}$$

where the last inequality is due to $a_n = o(\lambda)$. This verifies (5.6.3). Finally, to verify (5.6.4), Condition 3 implies that $\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}) \geq m$. Then by a similar proof, we can show that $\lambda_{\min}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}) \geq m/2$ with probability at least $1 - C_1 p^{-C_2}$. □

Proof of Theorem 5.3

Proof. We first bound $\widehat{\Delta}_y^{(k)} - \Delta_y^{(k)}$ for $k = 1, ..., K$. By definition, we have

$$\begin{aligned}
\Delta_y^{(k)} &= \Phi^{-1}(1 - P(Z_y > C_{yk})) = \Phi^{-1}(1 - P(Y \geq k)) \\
\widehat{\Delta}_y^{(k)} &= \Phi^{-1}(1 - (1/n)\sum_{i=1}^n Y_i^{(k)}) = \Phi^{-1}(1 - (1/n)\sum_{i=1}^n I(Y_i \geq k)).
\end{aligned}$$

By Lemma A.1 in Fan et al. (2017), the function $\Phi^{-1}(y)$ is Lipschitz continuous for $y \in (\Phi(-2M), \Phi(2M))$. Given the event $A_k = \{|\widehat{\Delta}_y^{(k)}| \leq 2M\}$, we have

$$\left|\widehat{\Delta}_y^{(k)} - \Delta_y^{(k)}\right| \leq L_1 \left|(1/n)\sum_{i=1}^n Y_i^{(k)} - (1 - \Phi(\Delta_y^{(k)}))\right| = L_1 \left|(1/n)\sum_{i=1}^n Y_i^{(k)} - \mathrm{E}(Y_i^{(k)})\right|.$$

Then, we have

$$\begin{aligned}
P(A_k^c) &= P(|\widehat{\Delta}_y^k| > 2M) \\
&= P\left(1 - \frac{1}{n}\sum_{i=1}^n Y_i^{(k)} < \Phi(-2M) \text{ or } 1 - \frac{1}{n}\sum_{i=1}^n Y_i^{(k)} > \Phi(2M)\right) \\
&= P\left(\frac{1}{n}\sum_{i=1}^n Y_i^{(k)} - (1 - \Phi(\Delta_y^{(k)})) > \Phi(\Delta_y^{(k)}) - \Phi(-2M)\right. \\
&\qquad \left.\text{or } \frac{1}{n}\sum_{i=1}^n Y_i^{(k)} - (1 - \Phi(\Delta_y^{(k)})) < \Phi(\Delta_y^{(k)}) - \Phi(2M)\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i^{(k)} - (1 - \Phi(\Delta_y^{(k)}))\right| \geq \Phi(2M) - \Phi(M)\right) \\
&\leq 2\exp\left(-\frac{n}{2}(\Phi(2M) - \Phi(M))^2\right),
\end{aligned}$$

where the second to the last inequality follows from Condition 2 and the last inequality follows from the Hoeffding inequality. Then we have the following for any $t > 0$.

$$\begin{aligned}
&P\left(\left|\widehat{\Delta}_y^{(k)} - \Delta_y^{(k)}\right| > t\right)\\
&\leq P\left(\left|\widehat{\Delta}_y^{(k)} - \Delta_y^{(k)}\right| > t|A_k\right) + P(A_k^c)\\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i^{(k)} - \mathrm{E}(Y_i^{(k)})\right| > t/L_1\right) + 2\exp\left(-\frac{n}{2}(\Phi(2M) - \Phi(M))^2\right)\\
&\leq 2\exp(-\frac{nt^2}{L_1^2}) + 2\exp\left(-\frac{n}{2}(\Phi(2M) - \Phi(M))^2\right).
\end{aligned}$$

Hence, choosing any $t \geq \sqrt{(\log p)/n}$ can let $P\left(\left|\widehat{\Delta}_y^{(k)} - \Delta_y^{(k)}\right| > t\right) = o(1)$.

Next we bound $\widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*$. We define $\{j_1, ..., j_{R_0}\} = \mathcal{M}$ to be the indices of the s important features under exact sparsity. Then further define

$$T_n = \left[g_{j_1}(-\sqrt{b\log n}), g_{j_1}(\sqrt{b\log n})\right] \times ... \times \left[g_{j_{R_0}}(-\sqrt{b\log n}), g_{j_{R_0}}(\sqrt{b\log n})\right],$$

for some $0 < b < 1$. Moreover, define events $E_1 = \{\widehat{\mathcal{M}} = \mathcal{M}\}$ and $E_2 = \{\mathbf{x} \in \mathbf{R}^p : \mathbf{x}_{\mathcal{M}} \in T_n\}$. Then we have

$$\begin{aligned}
&P\left(\left|\widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*\right| > t\right)\\
&\leq P\left(\left|\widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*\right| > t|E_1, E_2\right) + P(E_1^c) + P(E_2^c)\\
&= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

For I, given events E_1 and E_2, we can bound $\left|\widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*\right|$ by $\|\boldsymbol{\beta}^*\|_\infty \sup_{\mathbf{x}\in E_2}\|(\widehat{\mathbf{f}}_x(\mathbf{x}) - \mathbf{f}_x(\mathbf{x}))_{\mathcal{M}}\|_1 + \|\widehat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}\|_\infty \sup_{\mathbf{x}\in E_2}\|(\mathbf{f}_x(\mathbf{x}))_{\mathcal{M}}\|_1$. Using Theorem 2 from Han et al. (2013), we have

$$\sup_{\mathbf{x}\in E_2}\|(\widehat{\mathbf{f}}_x(\mathbf{x}) - \mathbf{f}_x(\mathbf{x}))_{\mathcal{M}}\|_1 = O_p(R_0\sqrt{\frac{\log\log n}{n^{1-b/2}}}).$$

By definition of E_2, we have $\sup_{\mathbf{x}\in E_2}\|(\mathbf{f}_x(\mathbf{x}))_{\mathcal{M}}\|_1 = O_p(R_0\sqrt{\log n})$. Hence we have

$$\left|\widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} - \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^*\right| = O_p\left(R_0\|\boldsymbol{\beta}^*\|_\infty\sqrt{\frac{\log\log n}{n^{1-b/2}}} + R_0\lambda\sqrt{\log n}\right).$$

For II, we have shown that II $= o(1)$ in Theorem 5.2.

For III, we use Mill's inequality on Gaussian tail to have

$$P(\mathrm{f}_j(X_j) > \sqrt{b \log n}) = O((n^b \log n)^{-1/2}),$$

and hence obtain III $= O(R_0(n^b \log n)^{-1/2}) = o(1)$.

Above all, we have shown that $P\left(\left|(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) - (\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)})\right| > t\right) = o(1)$ under the conditions of Theorem 5.2, $R_0(n^b \log n)^{-1/2} = o(1)$, and choosing $t = R_0\left(\|\boldsymbol{\beta}^*\|_\infty \sqrt{(\log\log n)/n^{1-b/2}} + \lambda\sqrt{\log n}\right)$ for some $0 < b < 1$. This concludes the proof of Theorem 5.3. □

Proof of Corollary 5.1 The proof of Corollary modifies the proof of Corollary 19 in Han et al. (2013). Denote the training data by $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, denote a future observation by $(\mathbf{x}, Y)$, denote the Bayesian rule as Y^* and the LMGCC rule by $\widehat{Y}$. We denote the LMGCC miclassification error given training data by $R_{LMGCC}(\mathcal{D})$ and denote the Bayes error by R_{Bayes}. We have the following identities.

$$\begin{aligned}
I(Y = k) &= I(\Delta_y^{(k)} < \mathrm{f}_y(Z_y) \le \Delta_y^{(k+1)}),\\
I(Y^* = k) &= I(\Delta_y^{(k)} < \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* \le \Delta_y^{(k+1)}),\\
I(\widehat{Y} = k) &= I(\widehat{\Delta}_y^{(k)} < \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} \le \widehat{\Delta}_y^{(k+1)}).
\end{aligned}$$

The indicators above take value 1 if the condition is true, and 0 if the condition is false. We can define the following variables that are equivalent with the above indicators but take value 1 if the condition is true, and -1 if the condition is false.

$$\begin{aligned}
W_k &= \mathrm{sgn}(\mathrm{f}_y(Z_y) - \Delta_y^{(k)}) + \mathrm{sgn}(\Delta_y^{(k+1)} - \mathrm{f}_y(Z_y)) - 1,\\
W_k^* &= \mathrm{sgn}(\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)}) + \mathrm{sgn}(\Delta_y^{(k+1)} - \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^*) - 1,\\
\widehat{W}_k &= \mathrm{sgn}(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) + \mathrm{sgn}(\widehat{\Delta}_y^{(k+1)} - \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}}) - 1.
\end{aligned}$$

Since we have

$$\begin{aligned} R_{LMGCC}(\mathcal{D}) &= \sum_{k=0}^{K} P(Y = k, \widehat{Y} \neq k|\mathcal{D}) = \sum_{k=0}^{K} P(Y \neq k, \widehat{Y} = k|\mathcal{D}), \\ R_{Bayes} &= \sum_{k=0}^{K} P(Y = k, Y^* \neq k) = \sum_{k=0}^{K} P(Y \neq k, Y^* = k), \end{aligned}$$

then

$$\begin{aligned} 2R_{LMGCC}(\mathcal{D}) &= \sum_{k=0}^{K} P(W_k \widehat{W}_k < 0|\mathcal{D}) \\ &= \sum_{k=0}^{K} P(W_k W_k^* + W_k(\widehat{W}_k - W_k^*) < 0|\mathcal{D}) \\ &\leq \sum_{k=0}^{K} (P(W_k W_k^* < 0|\mathcal{D}) + P(W_k(\widehat{W}_k - W_k^*) < 0|\mathcal{D})) \\ &\leq 2R_{Bayes} + \sum_{k=0}^{K} P(W_k(\mathrm{sgn}(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) - \mathrm{sgn}(\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)}))|\mathcal{D}) \\ &+ \sum_{k=0}^{K} P(W_k(\mathrm{sgn}(\widehat{\Delta}_y^{(k+1)} - \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}}) - \mathrm{sgn}(\Delta_y^{(k+1)} - \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^*))|\mathcal{D}) \\ &\leq 2R_{Bayes} + \sum_{k=0}^{K} P(\mathrm{sgn}(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) \neq \mathrm{sgn}(\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)})|\mathcal{D}) \\ &+ \sum_{k=0}^{K} P(\mathrm{sgn}(\widehat{\Delta}_y^{(k+1)} - \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\Delta_y^{(k+1)} - \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^*)|\mathcal{D}) \end{aligned}$$

Therefore,

$$\begin{aligned} 2\mathrm{E}(R_{LMGCC}(\mathcal{D})) - 2R_{Bayes} &\leq \sum_{k=0}^{K} P(\mathrm{sgn}(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) \neq \mathrm{sgn}(\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)})) \\ &+ \sum_{k=0}^{K} P(\mathrm{sgn}(\widehat{\Delta}_y^{(k+1)} - \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\Delta_y^{(k+1)} - \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^*)) \\ &= \sum_{k=1}^{K} P(\mathrm{sgn}(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) \neq \mathrm{sgn}(\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)})) \\ &+ \sum_{k=0}^{K-1} P(\mathrm{sgn}(\widehat{\Delta}_y^{(k+1)} - \widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}}) \neq \mathrm{sgn}(\Delta_y^{(k+1)} - \mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^*)) \\ &= 2\sum_{k=1}^{K} P(\mathrm{sgn}(\widehat{\mathbf{f}}_x(\mathbf{x})^T \widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}) \neq \mathrm{sgn}(\mathbf{f}_x(\mathbf{x})^T \boldsymbol{\beta}^* - \Delta_y^{(k)})). \end{aligned}$$

For the simplicity of notation, denote $G_k^* = \mathbf{f}_x(\mathbf{x})^T\boldsymbol{\beta}^* - \Delta_y^{(k)}$ and $\widehat{G}_k = \widehat{\mathbf{f}}_x(\mathbf{x})^T\widehat{\boldsymbol{\beta}} - \widehat{\Delta}_y^{(k)}$. Then the remainder of the proof only needs to show $P(\text{sgn}(\widehat{G}_k) \neq \text{sgn}(G_k^*)) = o(1)$ for $k = 1, ..., K$. Since we have

$$\begin{aligned} P(\text{sgn}(\widehat{G}_k) \neq \text{sgn}(G_k^*)) &= P(\widehat{G}_k G_k^* < 0) \\ &\leq P(|G_k^*| < t) + P(\left|\widehat{G}_k - G_k^*\right| > t) \end{aligned}$$

Using Theorem 5.3, we can choose $t = R_0\left(\|\boldsymbol{\beta}^*\|_\infty\sqrt{(\log\log n)/n^{1-b/2}} + \lambda\sqrt{\log n} + (n^b\log n)^{-1/2}\right)$ for some $0 < b < 1$ such that $P(\left|\widehat{G}_k - G_k^*\right| > t) = o(1)$. Also, because $t = o(1)$, the Gaussian probability $P(|G_k^*| < t) = o(1)$. This proves the misclassification error consistency.

5.6.2 Supporting Lemmas and their Proofs

Lemma 5.1. *(Fan et al. 2017; Feng and Ning 2019) Suppose Conditions 1 and 2 hold. It follow that*

$$P(\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{max} \geq C\sqrt{\log p/n}) \leq C_1 p^{-C_2},$$

where C_1 and C_2 are generic positive constants and C is a sufficiently large constant.

Proof. The proof of Lemma 5.1 can be found in the proofs of Theorem 6.1 in Fan et al. (2017) and Proposition 1 in Feng and Ning (2019). □

Lemma 5.2. *Under Conditions 1 and 2, there exists a sufficiently large positive constant C, and some generic positive constants C_1 and C_2 such that*

$$P(\|\widehat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\Sigma}}_{xy}\|_\infty \geq C\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log p}{n}}) \leq C_1 p^{-C_2}.$$

Proof. By Lemma 5.1 and (3.5) of Zhao et al. (2014), we have

$$P(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{max} \geq C\sqrt{\frac{\log p}{n}}) \leq C_1 p^{-C_2}, \tag{5.6.5}$$

for some sufficiently large constant C and some generic positive constants C_1 and C_2. We also have the following

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\mathbf{\Sigma}}_{xy}\|_\infty &= \|\widehat{\mathbf{\Sigma}}_{xx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy} - \widehat{\mathbf{\Sigma}}_{xy}\|_\infty \\ &= \|(\widehat{\mathbf{\Sigma}}_{xx} - \mathbf{\Sigma}_{xx})\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy} + \mathbf{\Sigma}_{xy} - \widehat{\mathbf{\Sigma}}_{xy}\|_\infty \\ &\leq \|\widehat{\mathbf{\Sigma}}_{xx} - \mathbf{\Sigma}_{xx}\|_{max}\|\boldsymbol{\beta}^*\|_1 + \|\widehat{\mathbf{\Sigma}}_{xy} - \mathbf{\Sigma}_{xy}\|_\infty.\end{aligned}$$

Since we have $\|\boldsymbol{\beta}^*\|_1 > m$ for some $m > 0$, using the result of Lemma 5.1, we have that $\|\widehat{\mathbf{\Sigma}}_{xx}\boldsymbol{\beta}^* - \widehat{\mathbf{\Sigma}}_{xy}\|_\infty \leq C\|\boldsymbol{\beta}^*\|_1\sqrt{\log p/n}$ with probability at least $1 - C_1 p^{-C_2}$ for some sufficiently large constant C and some generic positive constants C_1 and C_2.

□

Lemma 5.3. *Under Conditions 1, 2 and 4, if $R_0\sqrt{(\log p)/n} = o(1)$, there exist some generic positive constans C_1 and C_2 such that*

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \geq 2M) \leq C_1 p^{-C_2}.$$

Proof. By Fan et al. (2017),Feng and Ning (2019), and Zhao et al. (2014), we have the following for some sufficiently large constant C.

$$P\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \geq C\sqrt{\frac{\log p}{n}}\right) \leq C_1 p^{-C_2}.$$

Then,

$$\begin{aligned}P\left(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}\|_\infty \geq CR_0\sqrt{\frac{\log p}{n}}\right) &\leq P\left(R_0\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}\|_{\max} \geq CR_0\sqrt{\frac{\log p}{n}}\right) \\ &\leq C_1 p^{-C_2}.\end{aligned}$$

Since

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty &= \|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} + \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}(\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}} - \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \\ &\leq \|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty + \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty\|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}} - \widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}\|_\infty\|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty,\end{aligned}$$

by Conditions 1, 2 and 4, it holds with probability greater than $1 - C_1 p^{-C_2}$ that

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty &\leq M + MCR_0\sqrt{\frac{\log p}{n}}\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty, \\
\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty &\leq \frac{M}{1 - MCR_0\sqrt{(\log p)/n}} \leq 2M.
\end{aligned}$$

where the last inequality holds since $R_0\sqrt{\log p/n} = o(1)$. □

Lemma 5.4. *Under Conditions 1,2,4 and 5, if $R_0^2\sqrt{\log p/n} = o(1)$, then there exist some generic positive constants C_1 and C_2 such that*

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty > (1-\alpha)(1-\epsilon/2)) \leq C_1 p^{-C_2}.$$

Proof. Note that

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} &= \widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}) + (\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} + \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1} \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

For I,

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1})\|_\infty \\
&\leq (\|\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty + \|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty)\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}\|_\infty\|\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty.
\end{aligned}$$

Since by Condition 1, $\|\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty \lesssim R_0$, and $P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}\|_\infty \leq CR_0\sqrt{(\log p)/n}) \geq 1 - C_1 p^{-C_2}$, we have

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1})\|_\infty \geq CR_0^2\sqrt{\log p/n}) \leq C_1 p^{-C_2}.$$

Using similar arguments, for II, we have

$$P(\|(\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \geq CR_0\sqrt{(\log p)/n}) \leq C_1 p^{-C_2}.$$

Hence, if $R_0^2\sqrt{(\log p)/n} = o(1)$, we have

$$P(\|\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}^{-1} - \mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}) + (\widehat{\mathbf{\Sigma}}_{\mathcal{M}^c\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}^{-1}\|_\infty \leq (1-\alpha)\epsilon/2) \geq 1 - C_1 p^{-C_2}.$$

This result together with Condition 5 completes the proof. □

BIBLIOGRAPHY

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* Wiley. New York.

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., and Gentry, A. E. (2014). ordinalgmifs: An r package for ordinal regression in high-dimensional data settings. *Cancer informatics* **13,** CIN–S20806.

Archer, K. J. and Williams, A. A. (2012). l_1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine* **31,** 1464–1474.

Avella-Medina, M., Battey, H. S., Fan, J., and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **105,** 271–284.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2,** 183–202.

Bickel, P. J. and Levina, E. (2004). Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10,** 989–1010.

Bickel, P. J. and Levina, E. (2008a). covariance regularization by thresholding. *The Annals of Statistics* **36,** 2577–2604.

Bickel, P. J. and Levina, E. (2008b). regularized estimation of large covariance matrices. *The Annals of Statistics* **36,** 199–227.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization.* Cambridge university press.

Buczak, A. L. and Gifford, C. M. (2010). Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD workshop on intelligence and security informatics*, pages 1–10.

Bürkner, P.-C. and Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science* **2,** 77–101.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association* **106,** 1566–1577.

Cai, T. T. and Zhang, L. (2018). High-dimensional gaussian copula regression: Adaptive estimation and statistical inference. *Statistica Sinica* pages 963–993.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of statistics* **35,** 2313–2351.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and weighting in regression*, volume 30. CRC Press.

Clemmensen, L., Hastie, T., Witten, D., and Ersb***o***ll, B. (2011). Sparse discriminant analysis. *Technometrics* **53,** 406–413.

Crane, G. J. and Hoek, J. v. d. (2008). Conditional expectation formulae for copulas. *Australian & New Zealand Journal of Statistics* **50,** 53–67.

Doyle, O. M., Westman, E., Marquand, A. F., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Lovestone, S., Williams, S. C., et al. (2014). Predicting progression of alzheimers disease using ordinal regression. *PloS one* **9,** e105542.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74,** 745–771.

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

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Methodological)* **79,** 405–421.

Fang, K.-T., Kotz, S., and Ng, K. W. (2018). *Symmetric multivariate and related distributions*. Chapman and Hall/CRC.

Feng, H. and Ning, Y. (2019). High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 654–663. PMLR.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33,** 1.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science* **286,** 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research* **62,** 4963–4967.

Han, F., Zhao, T., and Liu, H. (2013). Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research* .

He, Y., Chen, H., Sun, H., Ji, J., Shi, Y., Zhang, X., and Liu, L. (2020). High-dimensional integrative copula discriminant analysis for multiomics data. *Statistics in Medicine* **39,** 4869–4884.

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* pages 933–944.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hou, J. and Archer, K. J. (2015). Regularization method for predicting an ordinal response using longitudinal high-dimensional genomic data. *Statistical applications in genetics and molecular biology* **14,** 93–111.

Idler, E. L., Musick, M. A., Ellison, C. G., George, L. K., Krause, N., Ory, M. G., Pargament, K. I., Powell, L. H., Underwood, L. G., and Williams, D. R. (2003). Measuring multiple dimensions of religion and spirituality for health research: Conceptual background and findings from the 1998 general social survey. *Research on Aging* **25,** 327–365.

Li, Q. and Li, L. (2018). Integrative linear discriminant analysis with guaranteed error rate improvement. *Biometrika* **105,** 917–930.

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica* pages 457–473.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40,** 2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10,**.

Mai, Q., Yang, Y., and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica* **29,** 97–111.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99,** 29–42.

Masarotto, G., Varin, C., et al. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6,** 1517–1549.

Mazzella, A. J., Sanders, M., Yang, H., Li, Q., Vavalle, J. P., and Gehi, A. (2021). Predicting need for pacemaker implantation early and late after transcatheter aortic valve implantation. *Catheterization and Cardiovascular Interventions* **97,** E588–E596.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* **42,** 109–127.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science* **27,** 538–557.

Noh, H., Ghouch, A. E., and Bouezmarni, T. (2013). Copula-based regression estimation and inference. *Journal of the American Statistical Association* **108,** 676–688.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association* **111,** 169–179.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient bayesian inference for gaussian copula regression models. *Biometrika* **93,** 537–554.

Qiao, X. and Liu, Y. (2009). Adaptive weighted learning for unbalanced multicategory classification. *Biometrics* **65,** 159–168.

Qu, Y., Piedmonte, M. R., and Medendorp, S. V. (1995). Latent variable models for clustered ordinal data. *Biometrics* pages 268–275.

Rosman, L., Salmoirago-Blotcher, E., Mahmood, R., Yang, H., Li, Q., Mazzella, A. J., Lawrence Klein, J., Bumgarner, J., and Gehi, A. (2020). Arrhythmia risk during the 2016 united states presidential election: The cost of stressful politics. *Available at SSRN 3699620* .

Sha, N. and Dechi, B. O. (2019). A bayes inference for ordinal response with latent variable approach. *Stats* **2,** 321–331.

Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics* **39,** 1241–1265.

Sungur, E. A. (2005). Some observations on copula regression functions. *Communications in Statistics?Theory and Methods* **34,** 1967–1978.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine* **16,** 385–395.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 753–772.

Wurm, M. J., Rathouz, P. J., and Hanlon, B. M. (2017). Regularized ordinal regression and the ordinalnet r package. *arXiv preprint arXiv:1706.05003* .

Yoon, G., Carroll, R. J., and Gaynanova, I. (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika* **107,** 609–625.

Zhao, T., Roeder, K., and Liu, H. (2014). Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. *Journal of Computational and Graphical Statistics* **23,** 895–922.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67,** 301–320.