ABSTRACT

XU, YINGZI. Binormal Precision-Recall and ROC Classification and Variable Selection. (Under the direction of Dr. Howard D. Bondell.)

Methods for classification and for variable selection within classification problems have become increasingly important in the area of statistics and machine learning. Classification is widely used in medical diagnosis, anomaly detection, information retrieval and various other areas. Variable selection, also known as feature selection, identifies and removes redundant or irrelevant variables, in order to avoid the overfitting problems and thus construct a model with better performance. To evaluate the classification performance, both Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve are highly informative, and the areas under these curves are used as popular measurement for comparing different classifiers.

In Chapter 2, we propose a novel approach for binary classification based on maximizing the area under the ROC curve and the PR curve under the assumption that the decision values for the two groups follow a binormal distribution. The key idea is to estimate the optimal classifier by either maximizing the area under ROC curve with a closed form derived, or maximizing the area under PR curve with a threshold gradient descent algorithm. Both methods utilize nonparametric functions, e.g. radial basis functions and b-splines to approximate the true decision value function which overcome the fully parametric assumption used in linear classifiers. Simulation studies and real data application show that the proposed methods outperform random forest and linear classifier regardless of whether the assumption is correct or mis-specified.

In Chapter 3, we consider the problem of variable selection in the classification process and propose three approaches based on maximizing the area under the ROC curve under

the binormal assumption. The first one is estimated through minimizing the euclidean distance of between the sparse coefficients and the dense binormal AUCROC maximizer with l_1 constraint on the coefficients. The second approach is a modification of the first one which is based on least square approximation. The third one utilizes minorization-maximization (MM) algorithm with coordinate descent to solve a reformulated problem which is related to the eigenvalue problem. Numerical studies and real data application show that two of the proposed methods are very useful and achieve superior performance in terms of both classification and variable selection when compared with other methods.

Binormal Precision-Recall and ROC Classification and Variable Selection

by
Yingzi Xu

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

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

Dr. Yichao Wu	Dr. Ana-Maria Staicu
Dr. Donald E. K. Martin	Dr. Howard D. Bondell Chair of Advisory Committee

DEDICATION

To my family.

BIOGRAPHY

The author was born in Hangzhou, China in January, 1990. She attended Zhejiang University (ZJU) for her undergraduate study, majoring in Applied Biology and minoring in Finance from 2008 to 2012. At the fourth year of her undergraduate, she discovered her interest in statistics and joined the Department of Statistics at North Carolina State University (NCSU) through the $3 + X$ program between ZJU and NCSU. 2011-2012 school year is an overlap year where she worked on both her undergraduate and her master degree. In 2013, she was granted M.S. degree in Statistics en route and she continued to pursue her Ph.D. degree under the guidance of Dr. Howard Bondell.

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisor, Dr. Howard Bondell for his enormous help during my graduate study at North Carolina State University. I really appreciate his insightful guidance, generous support and unending encouragement throughout these years. He has always been patient and encouraging whenever I encountered difficulties during my research. I feel so lucky to have him as my advisor. Without his guidance and constant feedback this PhD would not have been achievable.

I would also like to thank all my committee members, Dr. Yichao Wu, Dr. Donald Martin and Dr. Ana-Maria Staicu for their valuable suggestions and comments on the dissertation work. They have always been generous for their precious time during the whole process. I thank Dr. Praveen Kolar for serving on my committee as the graduate school representative. I also thank Dr. John Monahan for being my academic advisor during my master's degree, for caring me about my study and life, for guiding me into the world of statistics.

I appreciate the excellent academic support and industry opportunities provided by the Department of Statistics at North Carolina State University. I am grateful to all the faculty members for offering a comprehensive collection of lectures which gets my graduate career started on the right foot and provides me with the indispensable foundation for becoming a statistician. I would also like to thank all the staff members for their great service to the department. I also owe my gratitude to my mentors, Jeffery Painter from GlaxoSmithKline, John Schneider and Yolanda Ingram from SAS Institute where I worked as a Graduate Industrial Trainee; and Jack Vance from Maxpoint where I worked as a summer intern.

Finally, and most importantly, I would like to thank my husband, Zhou Li, who has

been by my side all these years, and without whom, I would not have had the courage to start this journey in the first place. I thank my Mom and Dad for always believing in me and encouraging me along the way and for all those things of life beyond doing a PhD.

Thank you, to all of you !

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Methods for classification and for variable selection within classification problems have become increasingly important in the area of statistics and machine learning. The goal of classification is to predict the category for each unseen object. It usually include two main phrases. In the training phrase, we have a training set of data, in which we observe the class and the features for a bunch of objects. Based on the training set with known classes, classification algorithms seek a model for the class attribute as a function of the independent variables, which is also called a classifier. In the application phase, they apply

the previously estimated classifier on the new and unseen datasets to determine the class of each object (Friedman et al., 2001). It is widely used in medical diagnosis, anomaly detection, information retrieval and various other areas. Variable selection, also known as feature selection, is the process of identifying and removing redundant or irrelevant variables, in order to avoid the overfitting problems and thus construct a model with better performance.

There are many classification algorithms, mainly in five categories summarized by Kotsiantis et al. (2007). The first group of methods are logic based algorithms, like decision trees that use a tree-structure graph of decisions to classify the instances (Quinlan, 1986). The second groups are perceptron based algorithms (Rosenblatt, 1962), like Artificial Neural Networks (ANN) which is composed of many highly interconnected processing neurones working together to solve the problem (Rumelhart et al., 1985; Zhang, 2000). The third is statistical learning algorithm which provides a probability that an instance belongs to a class. Some examples are Linear Discriminant Analysis (LDA) (Friedman, 1989), logistic regression (Cox, 1958), Naive Bayes (Nilsson, 1965), Bayesian Networks (Jensen, 1996), etc. The fourth is instance-based learning, like k nearest neighbor (kNN) where an instance is classified by a majority vote of its neighbors (Cover and Hart, 1967). The last group is the Support Vector Machine (SVM) that tries to find a hyperplane that separate the two classes with the maximum margin (Burges, 1998). There are also ensemble methods that combine individual classifiers together to achieve a more accurate and precise system with the cost of increased computation and comprehensibility. Random Forest is one of the examples that combines many decision trees using an idea of bagging (Liaw and Wiener, 2002).

With all these classifiers described above, one important thing is to measure their performance after the modeling process. A common practice is to use the receiver operating characteristic (ROC) curve which is highly informative about the classifier performance. The ROC curve plots the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1- specificity) as the cutoff value is varied. The area under the ROC curve (AUCROC) is a popular measure of classification accuracy (Hanley and McNeil, 1982). A classifier with a larger area under the ROC curve is often considered to be a better classifier under the ROC criteria. To calculate the area under the ROC curve (AUCROC), there are two main approaches. One is to calculate the area under a parametric assumption, most common being the binormal assumption. A binormal assumption is to assume that the decision values of the positive group and the negative group follow two independent Gaussian distributions (Dorfman and Alf Jr, 1968). The other approach is to calculate the AUCROC empirically without any distribution assumption.

Instead of using the area under the ROC curve (AUCROC) as a measurement tool to compare the performance of various classifiers, some have proposed to use the area as an objective function to obtain a classifier in the first place (Su and Liu, 1993; Liu et al., 2005; Pepe and Thompson, 2000; Pepe et al., 2006). The problem with these methods lies in that they either assume a normality for the linear combination of the covariates for the two groups, which may generate unsatisfactory result for non-normal data, or they utilize empirical AUCROC which might be unstable and computational intensive.

Although the ROC curve is very useful to measure the classification performance, when encountering imbalanced data, it might be overly optimistic (Davis and Goadrich, 2006). The Precision-Recall (PR) curve on the other hand, is used as an alternative to

the ROC curve when facing imbalanced data (Davis et al., 2005; He and Garcia, 2009). It plots the precision (true discovery rate, TDR) against the recall (true positive rate, TPR) for various thresholds. Davis and Goadrich (2006) discussed the relationship between the ROC curve and the Precision-Recall curve. Liu and Bondell (2016) proposed a linear classifier that maximizes the binormal area under the Precision-Recall curve (AUCPR) for highly imbalanced data. However, it may perform unsatisfactorily for non-normal data.

In Chapter 2, we relax the fully parametric assumption, while avoiding the pitfalls of using the empirical curves. We assume that there exists a transformation function of the original covariates that makes it follow a binormal distribution. With this relaxed assumption, the classifier no longer needs to be linear, it could take any form of the original covariates. We then construct the classifier that either maximizes the binormal AUCROC or maximizes the binormal AUCPR without specifying the form of the transformation function. Non-parametric functions are employed to approximate the true transformation function, e.g. B-spline functions and radial basis functions are explored. Empirical performance of the proposed methods is evaluated through several simulation studies and a real data application.

The classifiers based on maximizing the area under the ROC curve (AUCROC) or the area under the Precision-Recall curve (AUCPR) discussed above use all of the predictors to construct the classifiers. However, when redundant or irrelevant variables are used for constructing the classifier, several problems arise. Firstly, they would cause overfitting problems and may lead to a worse classifier. Secondly, in some areas like medical diagnosis, which variables are import for the disease diagnosis is of big interest. Thirdly, knowing

the exact relevant variables helps to get insight into the nature of the problem and makes the classification model much easier to interpret. Therefore, variable selection within the classification domain becomes a very important and inevitable task. There exists a lot of techniques for variable selection, mainly classified in three categories. First is the best subset selection which compares models with all possible set of predictors and uses criteria such as cross-validation (Stone, 1974), AIC (Akaike, 1974), and BIC (Schwarz et al., 1978) to determine the best model. Exhaustive search of all the subsets is very computationally expensive since the number of subsets to be considered grows rapidly with the number of predictors. Another group of methods based on forward or backward stepwise variable selection comes to the rescue. It adds or removes individual predictors in each step, based on their statistical significance (Hocking, 1976). Tibshirani (1996) pointed out that since the predictors are either retained or dropped in the model, the model selected can be very unstable. In other words, small changes in the data may lead to a very different model being selected. To overcome the limitation inherited in stepwise variable selection, the third group of methods are developed using regularization techniques to shrink the coefficients estimates towards zero. Examples include lasso (Tibshirani, 1996) which uses an l_1 penalty to achieve a sparse solution, grouped lasso (Yuan and Lin, 2006; Meier et al., 2008) where variables are selected in groups, elastic net (Zou and Hastie, 2005) which uses a combination of l_1 and l_2 penalty, Dantzig selector (Candes and Tao, 2007) which is a modification of the lasso, SCAD (Fan and Li, 2001) which is a non-concave penalty, etc.

In Chapter 3, we incorporate variable selection based on the binormal AUCROC. We consider a fundamental problem where we assume the linear combination rather than any

linear or nonlinear transformation (discussed in Chapter 2) of the original covariates in both groups follow a binormal distribution. We propose three classification and variable selection methods based on the binormal AUCROC. The first one is estimated through minimizing the euclidean distance between the sparse coefficients and the dense binormal AUCROC maximizer with an l_1 constraint on the coefficients. The second approach is a modification of the first one which is based on least square approximation. The third one utilizes a minorization-maximization (MM) algorithm with coordinate descent to solve a reformulated problem which is related to the eigenvalue problem. Simulation studies are conducted in order to evaluate the finite sample performance of the proposed methods. Real data applications are also conducted.

CHAPTER

2

BINORMAL ROC AND PRECISION-RECALL CLASSIFICATION

2.1 Introduction

Let $Y \in \{0, 1\}$ denote the binary outcome and $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ denote the p-dimensional vector of covariates. We classify $Y = 1$ if $f(\boldsymbol{X}) \geq c$ for some threshold c with $f(\boldsymbol{X})$ some unknown function of the covariates, otherwise $Y = 0$. The goal is to

estimate the decision function $f(\boldsymbol{X})$.

Since the area under the ROC curve (AUCROC) can be used as a measurement tool to compare the performance of various classifiers (Hanley and McNeil, 1982), a very intuitive way is to use the area as an objective function to obtain a classifier. Many approaches were proposed for constructing a classifier to improve diagnostic accuracy by maximizing binormal AUCROC. Su and Liu (1993) proposed to maximize the area under the ROC curve (AUCROC) so that the best classifier could be achieved under the ROC criteria. Under the multivariate normality assumption on the original covariates, they derived the optimal linear combination of these covariates that maximizes the AUCROC. Liu et al. (2005) derived a linear combination that achieves higher sensitivity over a certain range of specificity. Hsu and Hsueh (2013) derived a linear combination that maximizes the binormal partial area under the ROC curve given a pre-specified range. Ma et al. (2006) proposed a threshold gradient descent algorithm to perform parameter estimation and variable selection of the linear classifier simultaneously that maximizes the binormal AUCROC. These methods assume the decision value to be a linear combination of the original covariates and that they follow the binormal assumption. The binormal assumption on the linear combination of the covariates are sometimes too strong and may perform unsatisfactorily when the data follow non-normal distributions.

To relax the parametric distribution assumption, Pepe and Thompson (2000), Pepe et al. (2006), Ma and Huang (2005) proposed to find the best linear combination for classification by using the empirical AUCROC as the objective function. While this method is robust due to its distribution-free assumption, it has two limitations. Firstly, for small sample sizes, the empirical AUCROC might be unstable and differ greatly from the ex-

pected AUCROC because of small perturbations. Secondly, for high-dimensional or large datasets, this empirical optimization process would be quite computationally intensive (Ma et al., 2006).

Although the ROC curve is a popular tool for measuring classification performance, the precision recall (PR) curve can expose more differences between classifiers that are not obvious in ROC space when dealing with highly imbalanced data (Davis and Goadrich, 2006). Davis and Goadrich (2006) studied the relationship between these two curves by showing that a curve dominates in ROC space if and only if it dominates in PR space. However, algorithms that optimize the area under the ROC curve do not necessarily optimize the area under the PR curve (AUCPR). Binormal assumptions on PR curves were discussed by Brodersen et al. (2010). They showed that the binormal PR curve outperforms the empirical PR estimates since the latter would be highly imprecise for the true curve when the sample size is small and data is imbalanced, which ironically is the situation in which the curve is most useful. Liu and Bondell (2016) proposed a linear classifier that maximizes the binormal AUCPR for highly imbalanced data. However, it may perform unsatisfactorily for non-normal data.

In this chapter, we relax the fully parametric assumption, while avoiding the pitfalls of using the empirical curves. We assume that there exists a transformation $T : \mathbb{R}^p \to \mathbb{R}$ of the original covariates such that $T(\boldsymbol{X})$ now obeys the binormal assumption. With this relaxed assumption, the classifier no longer needs to be linear, it could take any form of the original covariates. We then construct the classifier that either maximizes the binormal AUCROC or maximizes the binormal AUCPR without specifying the form of the transformation function. Non-parametric functions are employed to approximate

the true transformation function, e.g. B-spline functions and radial basis functions are explored in this chapter.

The rest of the chapter is organized as follows. Section 2.2 first reviews the ROC and the precision-recall curves. Then we describe the details of the binormal model assumption and provide a review of B-splines and radial basis functions to be used in fitting. Section 2.3 provides the estimation details of the proposed methods. Section 2.4 shows the simulation results for different settings, it also shows the performance of the classifier when the model assumption is mis-specified. Section 2.5 demonstrates the approach on vertebral column data and Section 2.6 concludes.

2.2 Methods

2.2.1 Review of ROC and Precision-Recall Curves

Based on the decision function, for every fixed threshold, the confusion matrix can be constructed as shown in Table 1. True positives (TP) denote instances correctly classified as positives. False positives (FP) refer to negative instances incorrectly classified as positive. True negatives (TN) associate with negatives correctly classified as negative. False negatives (FN) refer to positive instances incorrectly classified as negative.

Table 2.1: Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

In particular, the following threshold-dependent metrics are commonly used.

$$\begin{aligned} \text{TPR} &= \frac{TP}{TP+FN}, \\ \text{FPR} &= \frac{FP}{FP+TN}, \\ \text{Precision} &= \frac{TP}{TP+FP}, \\ \text{Recall} &= \frac{TP}{TP+FN}. \end{aligned} \tag{2.1}$$

The population versions of the above metrics (2.1) can be written as

$$\begin{aligned} \text{TPR}(c) &= P(f(\boldsymbol{X}) \geq c | Y = 1), \\ \text{FPR}(c) &= P(f(\boldsymbol{X}) \geq c | Y = 0), \\ \text{Precision}(c) &= P(Y = 1 | f(\boldsymbol{X}) \geq c), \\ \text{Recall}(c) &= P(f(\boldsymbol{X}) \geq c | Y = 1), \end{aligned} \tag{2.2}$$

for any threshold c.

Given the above quantities, the ROC curve plots the TPR against the FPR as the threshold varies. A popular summary of the ROC curve is the area under the ROC curve

(AUCROC) and it is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Bamber, 1975). Define $\boldsymbol{X}_0$ as a random draw from $\boldsymbol{X}|Y=0$, $\boldsymbol{X}_1$ as a random draw from $\boldsymbol{X}|Y=1$, then the area under the ROC curve can be calculated as

$$\text{AUCROC} = P\bigg(f(\boldsymbol{X}_1) > f(\boldsymbol{X}_0)\bigg). \tag{2.3}$$

Another way to derive the area under the ROC curve is through the integration as

$$\text{AUCROC} = \int \text{TPR}(c)\, d\{\text{FPR}(c)\}. \tag{2.4}$$

However, the derivation is simplified via the representation in (2.3). The Precision-Recall (PR) curve is a plot of the precision against the recall across different thresholds. A useful summary of the PR curve is the area under the PR curve (AUCPR) which could be calculated as

$$\text{AUCPR} = \int \text{Precision}(c)\, d\{\text{Recall}(c)\}. \tag{2.5}$$

2.2.2 Binormal Assumption

We assume that the true decision values are a function of the covariates and follow two independent Gaussian distributions for the positive and the negative group, which is also known as the binormal model (Dorfman and Alf Jr, 1968) as follows,

$$f(\boldsymbol{X})|Y=0 \sim N(\nu_0, \sigma_0^2), \tag{2.6}$$

$$f(\boldsymbol{X})|Y=1 \sim N(\nu_1, \sigma_1^2), \tag{2.7}$$

where $f(\cdot)$ is a function $\mathbb{R}^p \to \mathbb{R}^1$. The same function $f(\cdot)$ is applied on both the positive and the negative groups. Here, the specific form of the function $f(\cdot)$ does not need to be known and it could take any form of the original covariates. The only assumption is that there exists such a function that can transform the data into binormal decision values for the two groups. As long as such function exists, the best classifier can be constructed by maximizing either the binormal AUCROC or the binormal AUCPR. This assumption overcomes the strict case of the linear classifier, since the actual decision value may be quadratic, cubic or even more complex on the covariates instead of simply linear.

2.2.3 Area under the ROC curve and the PR curve

Under the binormal assumption, at any threshold c, we further derive the threshold dependent metrics (2.2) based on the model given in (2.6) and (2.7) as follows,

$$
\begin{aligned}
\mathrm{TPR}(c) &= P(f(\boldsymbol{X}) \geq c | Y = 1) \\
&= P(\frac{f(\boldsymbol{X}) - \nu_1}{\sigma_1} \geq \frac{c - \nu_1}{\sigma_1} | Y = 1) \\
&= \Phi(\frac{\nu_1 - c}{\sigma_1}), \\
\mathrm{FPR}(c) &= P(f(\boldsymbol{X}) \geq c | Y = 0) \\
&= P(\frac{f(\boldsymbol{X}) - \nu_0}{\sigma_0} \geq \frac{c - \nu_0}{\sigma_0} | Y = 0) \\
&= \Phi(\frac{\nu_0 - c}{\sigma_0}), \\
\mathrm{Precision}(c) &= P(Y = 1 | f(\boldsymbol{X}) \geq c) \\
&= \frac{P(f(\boldsymbol{X}) \geq c | Y = 1) P(Y = 1)}{P(f(\boldsymbol{X}) \geq c | Y = 1) P(Y = 1) + P(f(\boldsymbol{X}) \geq c | Y = 0) P(Y = 0)} \\
&= \frac{q\Phi(\frac{\nu_1 - c}{\sigma_1})}{q\Phi(\frac{\nu_1 - c}{\sigma_1}) + (1 - q)\Phi(\frac{\nu_0 - c}{\sigma_0})}, \\
\mathrm{Recall}(c) &= \mathrm{TPR}(c) = \Phi(\frac{\nu_1 - c}{\sigma_1}),
\end{aligned}
\tag{2.8}
$$

where $q = P(Y = 1)$ is the probability of an instance belong to the positive class and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Then the area under the ROC curve (2.3) and the area under the PR curve (2.5) can be

further derived as

$$\text{AUCROC} = \Phi\left(\frac{\nu_1 - \nu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right), \tag{2.9}$$

$$\text{AUCPR} = \int_0^1 \frac{qt}{qt + (1-q)\Phi\left[\frac{\nu_0 - \nu_1}{\sigma_0} + \frac{\sigma_1}{\sigma_0}\Phi^{-1}(t)\right]}\, dt. \tag{2.10}$$

In order to compute the AUCROC or the AUCPR under the binormal assumption, we need to find the values $\nu_0, \nu_1, \sigma_0, \sigma_1$ which are constants that depend on the function, $f(\cdot)$.

2.2.4 Non-parametric Approximation

To find the unknown function $f(\boldsymbol{X})$, nonparametric functions are utilized to approximate the true function. This chapter will focus on two types of basis functions, b-spline functions and radial basis functions, but other choices are possible.

2.2.4.1 B-Spline Functions

In this section, we assume that the true function can be approximated by a combination of arbitrary functions of the covariates and the contribution of each covariate is additive. This is the so-called additive model first suggest by Buja et al. (1989). The function of each covariate is further specified to be a combination of b-splines where a b-spline is a piecewise polynomial function of degree d with continuous derivative of order $d-1$. To formulate this approximation,

$$f(\boldsymbol{X}) \approx \sum_{i=1}^{p} g_i(X_i), \tag{2.11}$$

where $g_i(X_i) = \sum_{j=1}^{K} \beta_{ij} B_{j,m}(X_i)$. Here, $B_{j,m}$ is the j^{th} b-spline function of order m. For simplicity, we set the order for the basis expansion to $m = 4$ which leads to the cubic spline functions. For simplicity, the number of basis functions is assumed to be the same for each variable. The number of basis functions K for each variable is then the only tuning parameter in this model.

2.2.4.2 Radial Basis Functions

An alternative approach to approximate a given function is to use a set of radial basis functions (RBF). Since the RBF approach is based on the distance from a covariate vector to a given knot location, it implicitly includes the interactions between the full set of covariates, rather than an additive model. The approximation function is represented as a sum of K radial basis functions with form

$$f(\boldsymbol{X}) \approx \sum_{i=1}^{K} \beta_i h_i(\boldsymbol{X}), \tag{2.12}$$

where $h_i(\boldsymbol{X}) = \exp \frac{||\boldsymbol{X} - \boldsymbol{X}_{c(i)}||^2}{\rho}$ is the i^{th} radial basis function associated with the i^{th} center $\boldsymbol{X}_{c(i)}$. Each of the radial basis functions is weighted by an appropriate coefficient β_i. Each covariate is normalized to have mean 0 and variance 1 before performing the radial basis function transformation. A k-means clustering method is used to find the clusters and the center points with a pre-specified number of centers based on the normalized data. The number of centers K and the scale parameter ρ are the tuning parameters.

2.3 Estimation

Under the binormal assumption, the best classifier in terms of ROC criteria is obtained by maximizing the area under the ROC curve, the best classifier regards to PR criteria is obtained by maximizing the area under the PR curve.

After the nonparametric approximation, we get a set of transformed variables $\boldsymbol{Z} = (Z_1(\boldsymbol{X}), \ldots, Z_L(\boldsymbol{X}))^T$ where $\boldsymbol{Z}$ are $(B_{j,m}(X_i))_{i=1,\ldots,p,j=1,\ldots,K}$ for the b-spline transformation and $(h_j(\boldsymbol{X}))_{j=1,\ldots,K}$ for the radial basis function transformation. The binormal assumption could then be rewritten as

$$\boldsymbol{Z}^T\boldsymbol{\beta}|Y = 0 \sim \boldsymbol{N}(\nu_0, \sigma_0^2), \tag{2.13}$$

$$\boldsymbol{Z}^T\boldsymbol{\beta}|Y = 1 \sim \boldsymbol{N}(\nu_1, \sigma_1^2), \tag{2.14}$$

where $\boldsymbol{\beta} \in \mathbb{R}^L$ denotes the corresponding coefficients which need to be estimated. Note that $L = K$ for the radial basis functions, and $L = pK$ for the B-splines.

We further assume that $\boldsymbol{Z}$ has mean $\boldsymbol{\mu}_0$ and variance $\boldsymbol{\Sigma}_0$ for group $Y = 0$, and has mean $\boldsymbol{\mu}_1$ and variance $\boldsymbol{\Sigma}_1$ for group $Y = 1$. Define $\boldsymbol{\mu}_0$ with entries $\mu_{0j} = E[Z_j(\boldsymbol{X})|Y = 0]$ and $\boldsymbol{\mu}_1$ with entries $\mu_{1j} = E[Z_j(\boldsymbol{X})|Y = 1]$. Define $\boldsymbol{\Sigma}_0$ such that $\Sigma_{0jk} = \mathrm{Cov}[Z_j(\boldsymbol{X}), Z_k(\boldsymbol{X})|Y = 0]$, define $\boldsymbol{\Sigma}_1$ such that $\Sigma_{0jk} = \mathrm{Cov}[Z_j(\boldsymbol{X}), Z_k(\boldsymbol{X})|Y = 1]$.

The binormal AUCROC and AUCPR would then be calculated based on $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \boldsymbol{\beta}$ so that $\nu_0 = \boldsymbol{\mu_0}\boldsymbol{\beta}$, $\nu_1 = \boldsymbol{\mu_1}\boldsymbol{\beta}$, $\sigma_0^2 = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}$, and $\sigma_1^2 = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}$.

2.3.1 Estimation using AUCROC

After the nonparametric approximation, the binormal AUCROC in (2.9) is then given by

$$\text{AUCROC}(\boldsymbol{\beta}) \;=\; \Phi\left(\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{\beta}}}\right), \tag{2.15}$$

where Φ is the standard normal cumulative distribution function and $\boldsymbol{\beta}$, $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ are specified previously.

To maximize (2.15) , it is equivalent to maximizing $\dfrac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{\beta}}}$, since Φ is a monotone increasing function. Furthermore, up to a change in sign, it is the same as maximizing the squared form $\dfrac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\beta}}{\boldsymbol{\beta}^T(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{\beta}}$, which is a generalized eigenvalue problem and could be solved directly with the closed-form solution as

$$\boldsymbol{\beta}_{\text{ROC}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\{\text{AUCROC}(\boldsymbol{\beta})\} \propto (\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0). \tag{2.16}$$

We then estimate $\boldsymbol{\beta}_{\text{ROC}}$ by plugging in the sample version, to obtain

$$\hat{\boldsymbol{\beta}}_{\text{ROC}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\{\widehat{\text{AUCROC}}(\boldsymbol{\beta})\} \propto (\hat{\boldsymbol{\Sigma}}_0+\hat{\boldsymbol{\Sigma}}_1)^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0), \tag{2.17}$$

where $\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_0, \hat{\boldsymbol{\Sigma}}_1$ are the sample mean and the sample covariance of the transformed variables $\boldsymbol{Z}$ for the two groups.

2.3.2 Estimation using AUCPR

After the nonparametric approximation, the binormal AUCPR in (2.10) can be rewritten as

$$\text{AUCPR}(\boldsymbol{\beta}) = \int_0^1 \frac{qt}{qt + (1-q)\Phi\left[\frac{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}} + \frac{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}}\Phi^{-1}(t)\right]} dt, \quad (2.18)$$

where Φ is the standard normal cumulative distribution function, Φ^{-1} is the inverse of the standard normal cumulative distribution function, $q = P(Y = 1)$ is the probability of an instance belonging to the positive class. We plug in the sample versions, $\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_0, \hat{\boldsymbol{\Sigma}}_1$, and would like to construct a classifier that maximizes this area, which is

$$\hat{\boldsymbol{\beta}}_{\text{PR}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\{\widehat{\text{AUCPR}}(\boldsymbol{\beta})\}. \quad (2.19)$$

Note that, if we had assumed $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ to be equal or proportional, then maximizing AUCPR would be equivalent to minimizing $\frac{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}}$, which leads to the optimal classifier with the form $\hat{\boldsymbol{\Sigma}}_0^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$. So under this assumption, the solution would be proportional to the maximizer for AUCROC. However, this assumption is too strong and it is difficult to even envision to be true in reality since $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ are the variances for the transformed variables $\boldsymbol{Z}$. Therefore, we seek the solution without the assumption of equal variance via a numerical method.

A gradient descent algorithm with backtracking line search is applied to this problem. When implementing this algorithm, we choose the initial step size $s = 1$. It has much

less computational cost than choosing an infinitesimal step size , e.g. $s = 1 \times 10^{-4}$. Instead of only moving an infinitesimal amount along the search direction, performing the line search allows us to start moving a relatively large step size along the direction and if needed it will iteratively shrink the step size s until the criterion is met. It is more efficient since one can move a larger amount without worry about moving too far to miss the local optimum by iteratively shrinking the step size. The full algorithm is given by Algorithm 1.

We compute the derivative of AUCPR with respect to $\boldsymbol{\beta}$ as follows,

$$\begin{aligned}\frac{dAUCPR}{d\boldsymbol{\beta}} = \int_0^1 & \frac{-qt}{\left(qt + (1-q)\Phi\left[\frac{(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}} + \frac{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}}\Phi^{-1}(t)\right]\right)^2} \times (1-q) \\ & \times \phi\left[\frac{(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}} + \frac{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}}\Phi^{-1}(t)\right] \\ & \times \left(\frac{\boldsymbol{\mu}_0-\boldsymbol{\mu}_1}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}} + \frac{\Phi^{-1}(t)\boldsymbol{\Sigma}_1\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta}}\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}}} - \frac{(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\beta}\boldsymbol{\Sigma}_0\boldsymbol{\beta}}{\sqrt{(\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta})^3}} - \frac{\Phi^{-1}(t)\sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}_1\boldsymbol{\beta}}\boldsymbol{\Sigma}_0\boldsymbol{\beta}}{\sqrt{(\boldsymbol{\beta}^T\boldsymbol{\Sigma}_0\boldsymbol{\beta})^3}}\right) dt\end{aligned} \tag{2.20}$$

Algorithm 1 describes the details about the implementation of the gradient descent algorithm. For classification purpose, both the AUCROC and the AUCPR estimators are only identifiable up to a scale constant. That is, instead of the absolute magnitude of $\boldsymbol{\beta}$, only the direction of $\boldsymbol{\beta}$, which represent the relative effects, can be estimated from the AUCROC and AUCPR objective function. For identifiability, we introduce an anchor variable whose estimated coefficient is set to a constant. Without loss of generosity, for the b-spline transformation, we assume the first transformed feature to be the anchor;

for the radial basis function transformation, we assume the feature associated with the center surrounded by the largest number of observations as the anchor; for the linear classifier, we assume the first variable to be the anchor. Since the objective AUCPR is not a convex function, only a local maximum would be achieved with a specific starting value. Thus, different initial values need to be tried and the one with the largest AUCPR would be the final solution. In our method, we use three initial values as $\boldsymbol{\beta}_{\text{initial}}$, (1) all zeros but the anchor are set to 1, (2) all zeros but the anchor are set to -1, (3) the AUCROC maximizer, $\hat{\boldsymbol{\beta}}_{\text{ROC}}$.

Algorithm 1 Gradient Descent for Maximizing AUCPR

$\boldsymbol{\beta}^{(0)} \leftarrow \boldsymbol{\beta}_{\text{initial}}$
while iter $<$ maxiter **do**
    Compute the gradient $gd = \frac{dAUCPR}{d\boldsymbol{\beta}}$. Denote the j^{th} component of gd as gd_j.
    Make $gd_{\text{anchor}} = 0$, then if $\max_j\{|gd_j|\} = 0$, stop the iterations.
    begin backtracking line search with initial step size s
    while siter $<$ smaxiter **do**
        $\boldsymbol{\beta}^{\text{temp}} \leftarrow \boldsymbol{\beta}^{(\text{iter})} + s \times gd$
        Compute $\text{AUCPR}^{\text{temp}}$ based on $\boldsymbol{\beta}^{\text{temp}}$
        if $\text{AUCPR}^{\text{temp}} \leq \text{AUCPR}^{(\text{iter})}$ **then**
            $s \leftarrow s/2$
            siter $\leftarrow$ siter+1
        else
            break
        end if
    end while
    end line search

    $\boldsymbol{\beta}^{(\text{iter})} \leftarrow \boldsymbol{\beta}^{\text{temp}}$
    $\text{AUCPR}^{(\text{iter})} \leftarrow \text{AUCPR}^{\text{temp}}$
    if $|\text{AUCPR}^{(\text{iter})} - \text{AUCPR}^{(\text{iter-1})}| < \text{tol} \times \text{AUCPR}^{(\text{iter-1})}$ **then**
        break
    end if
    iter $\leftarrow$ iter+1
end while
return $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\text{iter})}$

Use a different $\boldsymbol{\beta}_{\text{initial}}$, repeat steps 1-24.
Compare AUCPR got from different $\boldsymbol{\beta}_{\text{initial}}$, choose $\hat{\boldsymbol{\beta}}$ with the largest AUCPR.

2.3.3 Tuning Parameter Selection

For b-spline approximation, the number of basis functions K for each variable is the tuning parameter. For radial basis function approximation, the number of centers K and the scale parameter ρ are the tuning parameters. In our numerical studies, we propose using the V-fold cross validation (CV) to choose the tuning parameters. For a pre-defined integer V, we randomly partition the data into V complementary subsets of equal sizes. For estimation using AUCROC, we select the tuning parameters that maximize the ROC CV score as

$$\text{ROC CV Score} = \frac{1}{V}\sum_{i=1}^{V} \widehat{\text{AUCROC}}^{(i)}\left(\hat{\boldsymbol{\beta}}_{\text{ROC}}^{(-i)}\right), \tag{2.21}$$

where $\hat{\boldsymbol{\beta}}_{\text{ROC}}^{(-i)}$ is the estimated coefficient without the samples in the i-th fold and $\widehat{\text{AUCROC}}^{(i)}$ is the binormal AUCROC estimator with the data in the i-th fold. For estimation using AUCPR, we choose the tuning parameters to maximize the PR CV score given by

$$\text{PR CV Score} = \frac{1}{V}\sum_{i=1}^{V} \widehat{\text{AUCPR}}^{(i)}\left(\hat{\boldsymbol{\beta}}_{\text{PR}}^{(-i)}\right), \tag{2.22}$$

where $\hat{\boldsymbol{\beta}}_{\text{PR}}^{(-i)}$ is the estimated coefficient without the data in the i-th fold and $\widehat{\text{AUCPR}}^{(i)}$ is the binormal AUCPR estimator with the data in the i-th fold.

2.4 Simulation Study

In this section, we conduct several simulations with 200 replicates each to assess the finite sample performance of the proposed classification method using the AUCROC and

the AUCPR. In each simulation, we generate a training sample to fit the model and a separate test set of size 10000 to compare the performance for different classifiers. For methods utilizing radial basis functions and b-spline functions, we conduct a 5-fold cross validation as described in Section 2.3.3 on the training sample to determine the best tuning parameters. The coefficients would then be estimated on the whole training data using the best tuning parameters. For other methods, we estimate the coefficients directly on the whole training data. Three benchmark methods are used for comparison purposes. Two are linear classifiers maximizing the binormal AUCROC, and the binormal AUCPR, respectively. The third is Random Forests, a very popular ensemble method for classification.

We experiment different combinations of the training sample size and the proportion of positive examples. Training sample size of 400 and 1000 are considered. Both imbalanced case with $q = 0.1$ and balanced case with $q = 0.5$ are considered. For each combination, we consider three different settings such that the true function $f(\boldsymbol{x})$ includes some of the four following pieces, $10x_1^2$, $5\log(x_2)$, $\exp(x_3)$, $10\log(x_2)\exp(x_3)$ in each setting. The first three pieces represent the main effect and the last piece represents the interaction between variables.

2.4.1 Setting 1: Binormal Additive Data

Given $y = 0$, we generate $f(\boldsymbol{x}) \sim N(46, 2^2)$. Given $y = 1$, we generate $f(\boldsymbol{x}) \sim N(51, 2^2)$. Let $f(\boldsymbol{x}) = 10x_1^2 + 5\log(x_2) + \exp(x_3)$. First we generate x_3 so that $\exp(x_3) \sim \text{Uniform}(10, 15)$, then we generate x_2 so that $\log(x_2) \sim \text{Uniform}\left(0, \dfrac{f(\boldsymbol{x}) - \exp(x_3)}{5 + \exp(x_3)}\right)$, lastly we generate x_1 such that $x_1^2 = \Big(f(\boldsymbol{x}) - 5\log(x_2) - \exp(x_3)\Big)/10$. In this setting, our estimated classi-

fiers will not depend on variable x_3 much since x_3 is generated from the same distribution for both the positive and the negative group.

The average false discovery rate (FDR) against the true positive rate (TPR) out of 200 replicates are shown in Figure 2.1(b), Figure 2.2(b), Figure 2.3(b), Figure 2.4(b). Each represents training sample size of 400 and positive example proportion of 10%, training sample size of 400 and positive example proportion of 50%, training sample size of 1000 and positive example proportion of 10%, training sample size of 1000 and positive example proportion of 50%, respectively.

For each TPR, a lower false discovery rate is desired. The results of the AUCPR classifier is similar to the AUCROC classifier and are not shown here. In all the different combinations of the sample size and the positive example proportions, our proposed classifiers, both the AUCROC classifier and the AUCPR classifier are able to find the optimal classifier close to the truth and they outperform both the linear classifiers and the random forest method in terms of false discovery rate (FDR). The linear classifiers have the worst performance, since they reach high FDR very quickly and have higher FDR than all the other methods. This is expected, since the simulation data are nonlinear, so that the fully parametric linear classifiers fail to capture the true decision value function. The random forest classifier performs better than the linear classifier since it does not make any distribution assumption for the data. Nevertheless, it does not capture the nonlinear decision boundary well and is still worse than our proposed methods.

When changing from an imbalanced case ($q = 0.1$) to a balanced case ($q = 0.5$), or the sample size becomes larger, the classification problem becomes easier. Hence all the methods perform better, with our proposed method still performs the best but the

superiority of our proposed method upon others becomes less severe.

The b-spline approximation is slightly better than the radial basis function here in terms of FDR. In this setting, the true function is an additive model, and the b-spline approximation is good enough to capture this while the radial basis function approximation is more complex and not necessary.

For either of the nonparametric approximation, the performance of the AUCROC maximizer and the AUCPR maximizer are very similar. Although in theory, when data are highly imbalanced, the PR curve is better than the ROC curve and can show much more differences when comparing various classifiers, the AUCPR maximizer has very similar performance with the AUCROC maximizer since the AUCPR maximizer is achieved by numerical algorithm after nonparametric approximation and would have more uncertainty introduced.

2.4.2 Setting 2: Binormal Interactive Data

We consider the same binormal distribution as in the setting 1. Given $y = 0$, we generate $f(\boldsymbol{x}) \sim N(46, 2^2)$. Given $y = 1$, we generate $f(\boldsymbol{x}) \sim N(51, 2^2)$. Here, a different transformation function $f(\cdot)$ is considered. Let $f(\boldsymbol{x}) = 10x_1^2 + 5\log(x_2) + \exp(x_3) + 10\log(x_2)\exp(x_3)$ which includes interaction between x_2 and x_3. We first generate x_3 such that $\exp(x_3) \sim \text{Uniform}(10, 15)$, then generate x_2 such that $\log(x_2) \sim \text{Uniform}\left(0, \dfrac{f(x) - \exp(x_3)}{5 + 10\exp(x_3)}\right)$, lastly we generate x_1 such that $x_1^2 = \Big(f(\boldsymbol{x}) - 5\log(x_2) - \exp(x_3) - 10\log(x_2)\exp(x_3)\Big)/10$. Compared to setting 1, setting 2 considers the case where variables have interaction with each other rather than a simple additive model. In addition, x_3 becomes important in our estimated classifiers since the interaction piece

$10\log(x_2)\exp(x_3)$ is significantly different for the two groups.

The performance of the proposed classifiers and the other underlying classifiers are shown in Figure 2.1(b), Figure 2.2(b), Figure 2.3(b), Figure 2.4(b). Each represents training sample size of 400 and positive example proportion of 10%, training sample size of 400 and positive example proportion of 50%, training sample size of 1000 and positive example proportion of 10%, training sample size of 1000 and positive example proportion of 50%, respectively.

The results are mostly similar as setting 1 with our proposed classifiers being the best. The difference lies in that the radial basis function approximation is better than the b-spline approximation. This is due to the fact that the true model has interaction between variables, and radial basis functions are designed to capture this while the b-spline approximation is not.

2.4.3 Setting 3: Non-normal Interactive Data

Setting 1 and setting 2 both consider the situations where the binormal assumption is correct. However, the performance of the proposed method under the cases where the model is mis-specified is also of interest. In setting 3, we let the true function follow two different shifted Gamma distribution for the two groups. Given $y = 0$, we generate $f(\boldsymbol{x}) \sim \text{Gamma}(2, 1) + 30$. Given $y = 1$, we generate $f(\boldsymbol{x}) \sim \text{Gamma}(8, 1) + 30$. We consider the same transformation function $f(\cdot)$ which includes interaction between the variables as in setting 2 and uses the same generation scheme for x_1, x_2, x_3 as in setting 2.

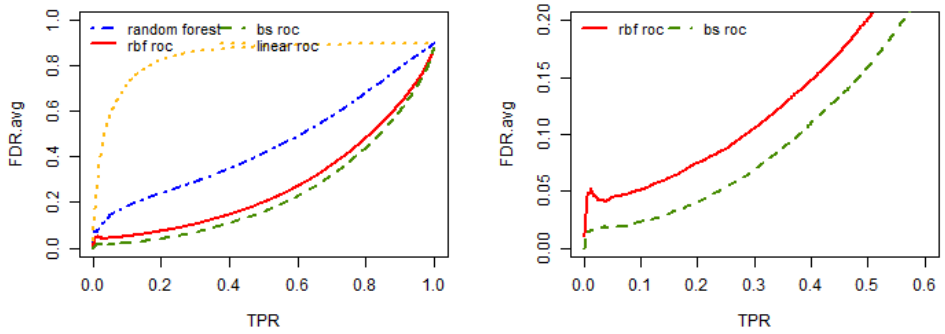
Figure 2.1(c), Figure 2.2(c), Figure 2.3(c), Figure 2.4(c) show the result for setting

3. Each represents training sample size of 400 and positive example proportion of 10%, training sample size of 400 and positive example proportion of 50%, training sample size of 1000 and positive example proportion of 10%, training sample size of 1000 and positive example proportion of 50%, respectively. We can see that when the binormal assumption is seriously violated as in setting 3, the proposed methods still give very satisfactory performance which is better than the random forest method and the linear classifiers.

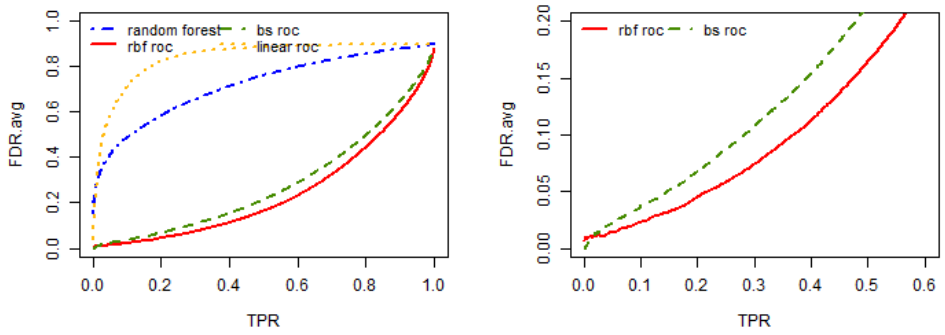
2.5 Application to Vertebral Column Data

We applied our methods on the vertebral column data from the UCI data repository (Lichman, 2013). The data consists of 310 records with 210 abnormal patients and 100 normal patients. Each patient has six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine which are pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis.

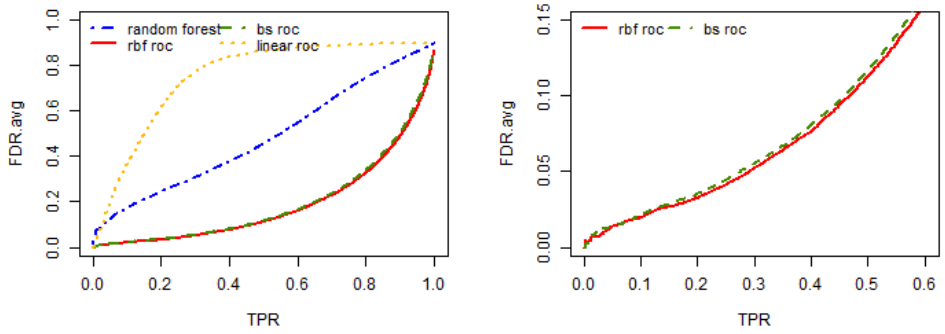
The data are randomly split into two parts, 75% training and 25% testing. A 5-fold cross validation is conducted on the training set for the methods using radial basis functions and b-spline approximation. The partition process is repeated 50 times. Average false discovery rate out of the 50 testing data sets is plotted in Figure 2.5. The AUCPR classifiers have very similar performance with the AUCROC classifiers and are not shown in the plot. It can be seen that the AUCROC maximizer with the radial basis function approximation is the best among the methods in terms of FDR. Although random forest is a bit better when TPR is smaller than 0.3, the improvement is very little and we could always choose a cutoff value with a lot higher TPR without much sacrifice in the FDR, e.g. 0.6 TPR with only 0.1 FDR. Here, the radial basis function approximation performs



(a) Setting 1: Binormal non-interactive data

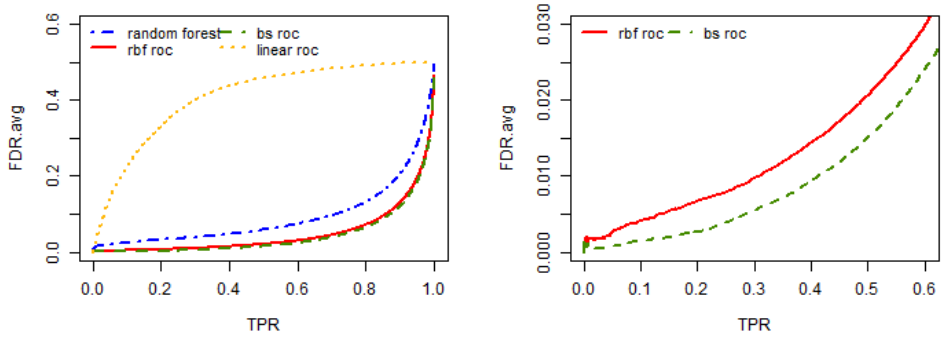


(b) Setting 2: Binormal interactive data

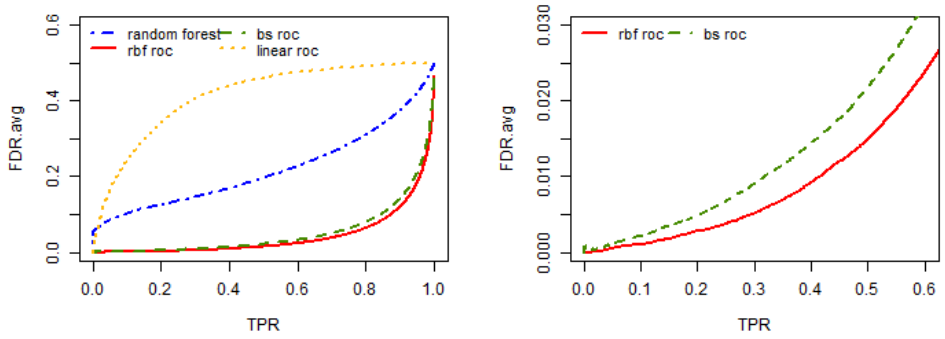


(c) Setting 3: Non-normal interactive data

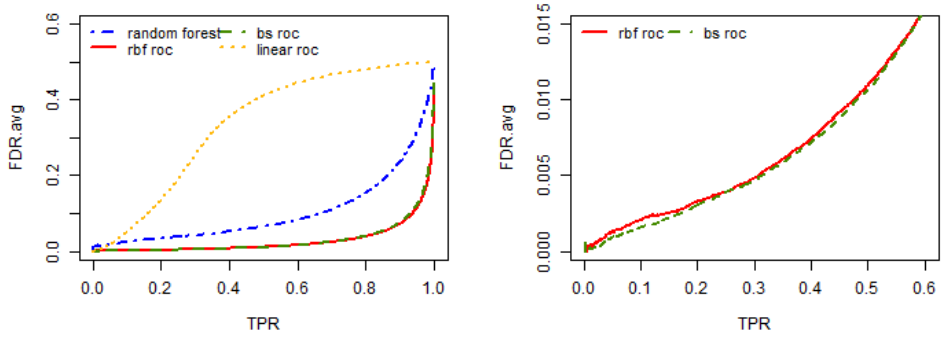
Figure 2.1: Average false discovery rate against the true positive rate out of 200 replicates for all the three simulation settings when the sample size is 400 and the positive example proportion is $q = 0.1$. The left panels are true positive rate (TPR) on the full range $(0, 1)$ while the right panels zoom in to the TPR range of (0, 0.6).



(a) Setting 1: Binormal non-interactive data

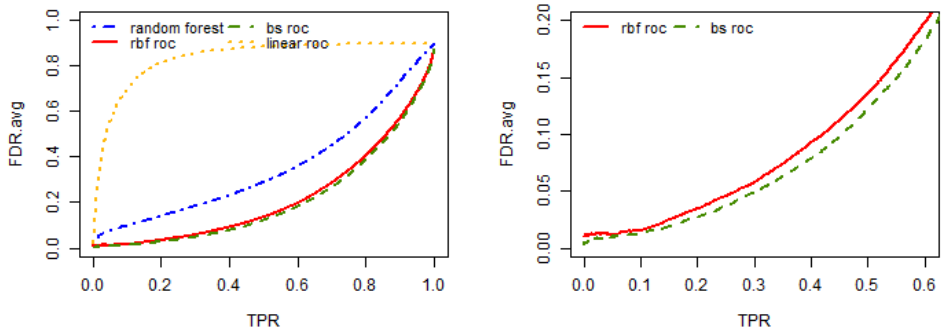


(b) Setting 2: Binormal interactive data

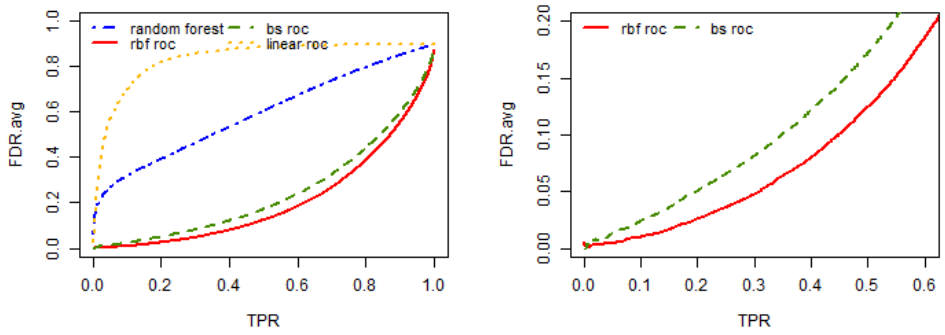


(c) Setting 3: Non-normal interactive data

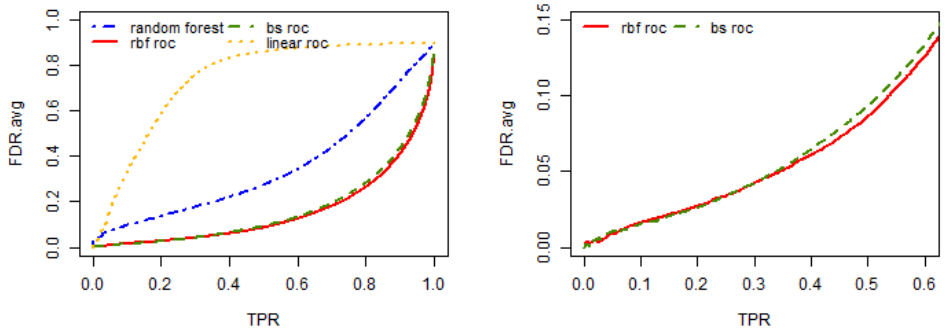
Figure 2.2: Average false discovery rate against the true positive rate out of 200 replicates for all the three simulation settings when the sample size is 400 and the positive example proportion is $q = 0.5$. The left panels are true positive rate (TPR) on the full range $(0, 1)$ while the right panels zoom in to the TPR range of (0, 0.6).



(a) Setting 1: Binormal non-interactive data

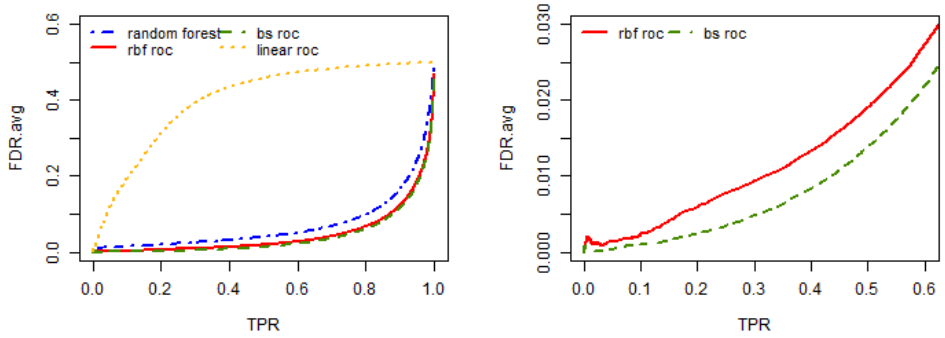


(b) Setting 2: Binormal interactive data

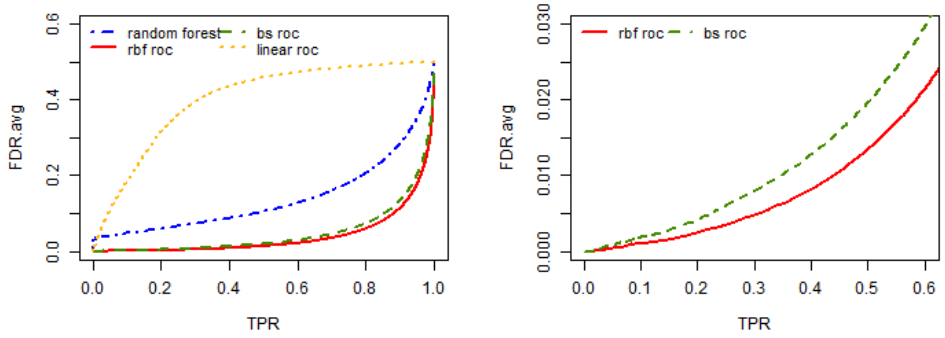


(c) Setting 3: Non-normal interactive data

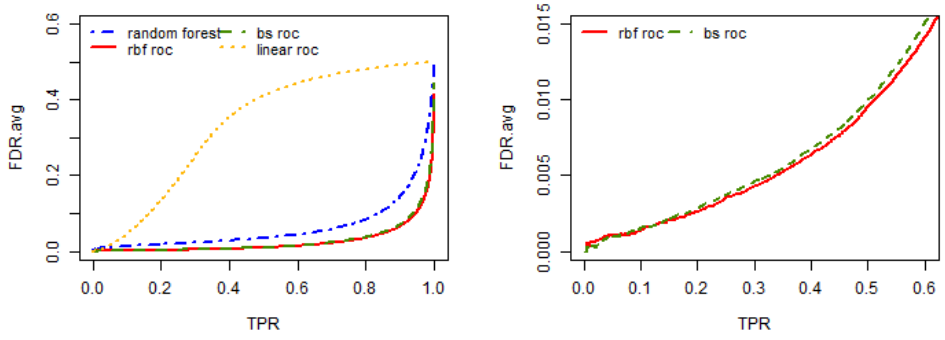
Figure 2.3: Average false discovery rate against the true positive rate out of 200 replicates for all the three simulation settings when the sample size is 1000 and the positive example proportion is $q = 0.1$. The left panels are true positive rate (TPR) on the full range $(0, 1)$ while the right panels zoom in to the TPR range of (0, 0.6).



(a) Setting 1: Binormal non-interactive data



(b) Setting 2: Binormal interactive data



(c) Setting 3: Non-normal interactive data

Figure 2.4: Average false discovery rate against the true positive rate out of 200 replicates for all the three simulation settings when the sample size is 1000 and the positive example proportion is $q = 0.5$. The left panels are true positive rate (TPR) on the full range $(0, 1)$ while the right panels zoom in to the TPR range of (0, 0.6).

better than the b-spline approximation likely due to strong interactions between these covariates.

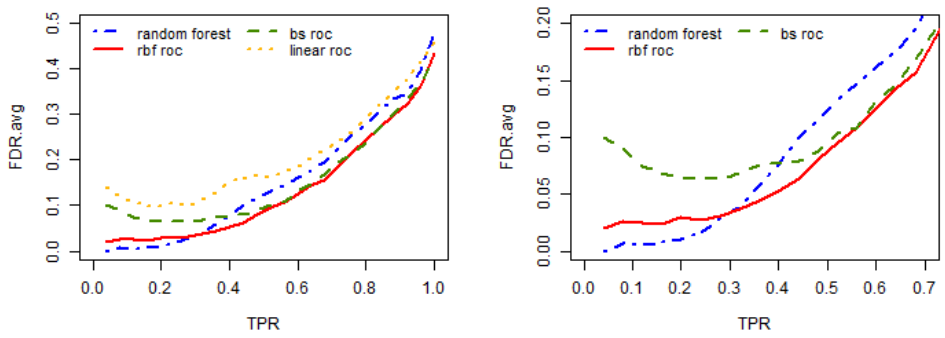


Figure 2.5: Average false discovery rate against the true positive rate for vertebral column data set from UCI data repository (Lichman, 2013). The left panel is true positive rate (TPR) on the full range (0,1) while the right panel zooms in to the TPR range of (0, 0.7).

2.6 Conclusion

In this chapter, we propose using the binormal AUCROC and the binormal AUCPR as the objective function for binary classification using a nonparametric approximation of the true decision value function. Both simulation studies and real data application show that the proposed methods outperform linear classifiers and random forest classifiers. The proposed methods have a very flexible assumption that only assumes the existence of a function which could transform the data to a binormal distribution without the need of knowing the specific form of the function. It is more suitable for the real world data

than the fully parametric assumption in the linear classifier. When this assumption is seriously violated, the proposed methods still give very satisfactory performance which is better than the random forest method and the linear classifier.

When there is a strong interaction, the radial basis function approximation would lead to better results. When the truth is an additive model, the b-spline approximation would be suggested. Other nonparametric approximations can also be explored according to real situations.

CHAPTER

3

BINORMAL ROC CLASSIFICATION AND VARIABLE SELECTION

3.1 Introduction

Variable selection, also known as feature selection, is the process of identifying and removing redundant or irrelevant variables. When redundant or irrelevant variables are used for constructing the classifier, several problems arise. Firstly, they would cause overfit-

ting problems and may lead to a worse classifier. Secondly, in some areas like medical diagnosis, which variables are import for the disease diagnosis is of big interest. Thirdly, knowing the exact relevant variables helps to get insight into the nature of the problem and makes the classification model much easier to interpret. Therefore, variable selection within the classification domain becomes a very important task.

Recall that the classification problem consists of a binary response $Y \in \{0, 1\}$ and a p-dimensional vector of covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$. In this chapter, rather than making a binormal assumption on any general transformation $f(\boldsymbol{X})$ as discussed in Chapter 2, we consider a specific case where $f(\boldsymbol{X})$ being $\boldsymbol{\beta}^T\boldsymbol{X}$ follows the binormal distribution where $\boldsymbol{\beta}$ is the corresponding coefficients that need to be estimated. In addition, we consider the decision function $f(\boldsymbol{X})$ to be "linear risk scores" $\boldsymbol{\beta}^T\boldsymbol{X}$. The goal here is to estimate $\boldsymbol{\beta}$.

To incorporate variable selection based on binormal AUCROC, Ma et al. (2006) proposed a threshold gradient descent algorithm to perform parameter estimation and variable selection of the linear classifier simultaneously that maximizes the binormal AUCROC. Yu and Park (2014) proposed a l_1 penalized regression estimator based on the binormal AUCROC.

In this chapter, we propose three classification and variable selection methods that are based on the binormal AUCROC. The first one is estimated through minimizing the euclidean distance between the sparse coefficients and the dense binormal AUCROC maximizer with an l_1 constraint on the coefficients. The second approach is a modification of the first one which is based on least squares approximation. The third one utilizes the minorization-maximization (MM) algorithm with coordinate descent to solve

a reformulated problem which is related to the eigenvalue problem.

The rest of the chapter is organized as follows. Section 3.2 provides the details of the binormal model assumption and the binormal AUCROC. Section 3.3 illustrates the proposed classification and variable selection methods. Simulation results for different settings are provided in Section 3.4 and real data application results are shown in Section 3.5. Section 3.6 concludes.

3.2 Preliminaries

3.2.1 Binormal Assumption

Use the binormal assumption on $f(\boldsymbol{X})$ given in (2.6) and (2.7), we consider a specific case with $f(\boldsymbol{X}) = \boldsymbol{\beta}^T\boldsymbol{X}$. We have

$$\boldsymbol{\beta}^T\boldsymbol{X}|Y = 0 \sim N(\nu_0, \sigma_0^2), \tag{3.1}$$

$$\boldsymbol{\beta}^T\boldsymbol{X}|Y = 1 \sim N(\nu_1, \sigma_1^2). \tag{3.2}$$

Notice that the normality of $\boldsymbol{X}$ itself is not necessary, we only require the linear combination of $\boldsymbol{X}$ to be normal, which is more likely to hold.

3.2.2 Area under the ROC curve

Define $\boldsymbol{X}|Y = j$ to have mean $\boldsymbol{\mu_j} = E(\boldsymbol{X}|Y = j)$ and variance $\boldsymbol{\Sigma}_j = \text{Var}(\boldsymbol{X}|Y = j)$ for $j = 0, 1$. Assume that at least one of $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$ is non-singular. Then the binormal area

under the ROC curve in (2.9) can be further derived as

$$\text{AUCROC}(\boldsymbol{\beta}) \;=\; \Phi\left(\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{\sqrt{\boldsymbol{\beta}^T(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{\beta}}}\right), \tag{3.3}$$

where Φ is the standard normal cumulative distribution function.

3.3 Classification and Variable Selection Methods

Xu and Bondell (2016) proved that maximizing (3.3), up to a change in sign, is the same as maximizing the squared form,

$$\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\beta}}{\boldsymbol{\beta}^T(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)\boldsymbol{\beta}}, \tag{3.4}$$

which is a generalized eigenvalue problem and that can be solved directly with closed-form solution

$$\boldsymbol{\beta}_{\text{ROC}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\{\text{AUCROC}(\boldsymbol{\beta})\} \propto (\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0), \tag{3.5}$$

up to a change in sign. We then estimate $\boldsymbol{\beta}_{\text{ROC}}$ by plugging in the sample version, to obtain

$$\hat{\boldsymbol{\beta}}_{\text{ROC}} = \underset{\boldsymbol{\beta}}{\text{argmax}}\{\widehat{\text{AUCROC}}(\boldsymbol{\beta})\} \propto (\hat{\boldsymbol{\Sigma}}_0+\hat{\boldsymbol{\Sigma}}_1)^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_0), \tag{3.6}$$

where $\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_0, \hat{\boldsymbol{\Sigma}}_1$ are the sample mean and the sample covariance of the predictors $\boldsymbol{X}$ for the two groups.

The dense solution $\hat{\boldsymbol{\beta}}_{\text{ROC}}$ utilizes all the covariates available without doing any variable selection. However, it leads to an unsatisfactory result when redundant or irrelevant

variables are used. Therefore, we study to perform the variable selection and classification simultaneously by optimizing the binormal AUCROC.

Throughout the chapter, for a vector $\boldsymbol{q} = (q_1, ..., q_p)^T \in \mathbb{R}^p$, define the l_1 norm as $\|\boldsymbol{q}\|_1 = \sum_{i=1}^p |q_i|$, l_2 norm as $\|\boldsymbol{q}\|_2 = \sqrt{\sum_{i=1}^p q_i^2}$. We propose several approaches to achieve the variable selection goal.

3.3.1 Method 1: IdenInv

The intuitive way of getting a sparse solution but as accurate as the dense solution $\hat{\boldsymbol{\beta}}_{\text{ROC}}$ is to minimize the euclidean norm between these two solutions with a lasso penalty given by

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\text{argmin}} \Big\{ \|\boldsymbol{\beta} - (\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \Big\}, \tag{3.7}$$

where $\lambda \geq 0$ is the tuning parameter that controls the sparseness of the coefficients. When $\lambda = 0$, the solution is exactly the same as the dense solution $\hat{\boldsymbol{\beta}}_{\text{ROC}}$. If λ increases, some elements of $\boldsymbol{\beta}$ are exactly zero, thus we get a more sparse solution. To solve (3.7), one can use a trick, that is to treat $(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$ as the response, and create an identity matrix $\boldsymbol{I}_p \in \mathbb{R}^{p\times p}$ as the design matrix for the independent variables. Then the LARS algorithm (Efron et al., 2004) can be utilized to get the whole solution path with various λ values. Since we utilize the identity matrix and the dense solution $\hat{\boldsymbol{\beta}}_{\text{ROC}}$ which need an inverse step, we name this method "IdenInv" throughout the chapter.

3.3.2 Method 2: LsInv

The IdenInv method does not account for the variance of $\hat{\boldsymbol{\beta}}_{\text{ROC}}$. However, using the variance of $\hat{\boldsymbol{\beta}}_{\text{ROC}}$ would lead to a more efficient estimator. Therefore, we would like to take care of the variance of $\hat{\boldsymbol{\beta}}_{\text{ROC}}$. Since the direct variance of $\hat{\boldsymbol{\beta}}_{\text{ROC}}$ is hard to compute, we consider $\hat{\boldsymbol{\beta}}^*_{\text{ROC}} = (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$, whose variance can be calculated as $\boldsymbol{V} = \text{Var}(\hat{\boldsymbol{\beta}}^*_{\text{ROC}}) = (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-T}(\frac{\boldsymbol{\Sigma}_0}{n_0} + \frac{\boldsymbol{\Sigma}_1}{n_1})(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}$. Now utilizing the least square approximation, we have

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \left\| \boldsymbol{V}^{-1/2}\hat{\boldsymbol{\beta}}^*_{\text{ROC}} - \boldsymbol{V}^{-1/2}\boldsymbol{\beta} \right\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}. \tag{3.8}$$

where $\boldsymbol{V}^{-1}$ is the inverse of $\boldsymbol{V}$ that can be decomposed into $\boldsymbol{V}^{-1} = \boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}$, and $\lambda \geq 0$ is the tuning parameter that controls the sparseness of the coefficients $\boldsymbol{\beta}$. We call this method "LsInv' since it utilizes the idea of least squares approximation and $\hat{\boldsymbol{\beta}}^*_{\text{ROC}}$, which requires an inverse step.

The LsInv method uses a more accurate objective to solve, which leads to a more accurate sparse solution $\boldsymbol{\beta}$ than the IdenInv method. When solving (3.8), we use $\hat{\boldsymbol{V}} = (\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)^{-T}(\frac{\hat{\boldsymbol{\Sigma}}_0}{n_0} + \frac{\hat{\boldsymbol{\Sigma}}_1}{n_1})(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)^{-1}$ to estimate $\boldsymbol{V}$ and use a similar trick as in IdenInv. We treat $\hat{\boldsymbol{V}}^{-1/2}\hat{\boldsymbol{\beta}}^*_{\text{ROC}}$ as the response and $\hat{\boldsymbol{V}}^{-1/2}$ as the design matrix, then employ the LARS algorithm (Efron et al., 2004) to get the whole solution path with different λ values.

3.3.3 Method 3: MM

Now instead of dealing with $\hat{\boldsymbol{\beta}}_{\text{ROC}}$ in various ways, we take a step back and consider the generalized eigenvalue problem (3.4) directly, which is equivalent to

$$\text{Maximize}\Big\{\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}\Big\} \text{ subject to } \boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} \leq 1, \tag{3.9}$$

where $\boldsymbol{A} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T$ and $\boldsymbol{W} = (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)$. We define the penalized problem as

$$\boldsymbol{\beta} \;=\; \underset{\boldsymbol{\beta}}{\text{argmax}} \Big\{\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta} - \lambda\|\boldsymbol{\beta}\|_1\Big\} \text{ subject to } \boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} \leq 1 \tag{3.10}$$

where $\boldsymbol{A} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T$, $\boldsymbol{W} = (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)$ and $\lambda \geq 0$ is a tuning parameter. We can utilize the minorize-maximization (MM) algorithm to solve the problem (Lange et al., 2000; Hunter and Lange, 2004; Lange, 2004). It tries to find a surrogate function that minorizes the objective function and then optimizes the surrogate function iteratively to drive the objective function upward towards the maximum. In each iteration, we consider two scenarios. When $\boldsymbol{W}$ is a diagonal matrix, the problem is easier and each coefficient can be updated separately in each iteration (Gaynanova et al., 2016). When $\boldsymbol{W}$ is a symmetric but not diagonal matrix, a coordinate descent algorithm is employed in each iteration. A detailed proof is provided in the Appendix. The Algorithm details are shown in Algorithm 2. Throughout the Algorithm 2, we denote $ST(z, r) = \text{sgn}(z)(|z| - r)_+$.

Algorithm 2

$\boldsymbol{\beta}^{(1)} \leftarrow \boldsymbol{\beta}_{\text{initial}}$
repeat
  $m \leftarrow 1$
  if $\boldsymbol{W}$ is diagonal matrix **then**
    for $i = 1, 2, \ldots, p$ **do**
      $d_i \leftarrow ST\Big(\dfrac{(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i}{w_{ii}}, \dfrac{\lambda}{2w_{ii}}\Big)$ where w_{ii} is the i^{th} is diagonal element of $\boldsymbol{W}$ and
      $(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i$ is the i^{th} element of $\boldsymbol{A}\boldsymbol{\beta}^{(m)}$.
    end for
  else if $\boldsymbol{W}$ is symmetric but not diagonal matrix **then**
    repeat
      $\boldsymbol{d}^{(1)} \leftarrow \boldsymbol{d}_{\text{initial}}$
      $k \leftarrow 1$
      for $i = 1, 2, \ldots, p$ **do**
        $d_i^{(k+1)} \leftarrow ST\Big(\dfrac{(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i - (\boldsymbol{d}^{(k)})_{-i}^T \boldsymbol{W}_{-i}}{w_{ii}}, \dfrac{\lambda}{2w_{ii}}\Big)$
        where $\boldsymbol{d}_{-i}^{(k)} = (d_1^{(k+1)}, \cdots, d_{i-1}^{(k+1)}, d_{i+1}^{(k)}, \cdots d_p^{(k)})$,
        and $\boldsymbol{W}_{-i} = (w_{1i}, \cdots, w_{i-1,i}, w_{i+1,i}, \cdots, w_{pi})$.
      end for
      $g(\boldsymbol{d}^{(k+1)}) \leftarrow (\boldsymbol{d}^{(k+1)})^T \boldsymbol{W} \boldsymbol{d}^{(k+1)} - 2(\boldsymbol{d}^{(k+1)})^T \boldsymbol{A}\boldsymbol{\beta}^{(m)} + \lambda \|\boldsymbol{d}^{(k+1)}\|_1$
      $k \leftarrow k + 1$
    until $k = k_{\max}$ or $|g(\boldsymbol{d}^{(k+1)}) - g(\boldsymbol{d}^{(k)})| < \epsilon |g(\boldsymbol{d}^{(k)})|$
  end if
  if $\boldsymbol{d} \neq \boldsymbol{0}$ **then**
    $\boldsymbol{\beta}^{(m+1)} \leftarrow \boldsymbol{d}/(\boldsymbol{d}^T \boldsymbol{W} \boldsymbol{d})^{\frac{1}{2}}$
  else
    $\boldsymbol{\beta}^{(m+1)} \leftarrow \boldsymbol{0}$
  end if
  $h(\boldsymbol{\beta}^{(m+1)}) \leftarrow (\boldsymbol{\beta}^{(m+1)})^T \boldsymbol{A}\boldsymbol{\beta}^{(m+1)} - \lambda \|\boldsymbol{\beta}^{(m+1)}\|_1$
until $m = m_{\max}$ or $|h(\boldsymbol{\beta}^{(m+1)}) - h(\boldsymbol{\beta}^{(m)})| < \epsilon |h(\boldsymbol{\beta}^{(m)})|$

3.4 Simulation

In this section, we conduct several simulations to assess the finite sample performance of the proposed classification and variable selection methods. In each setting, we generate a training sample of size 1000 to fit the model, a separate validation set of size 1000 to find the optimal tuning parameters and finally a test set of size 1000 to compare the performance for different methods. The process is repeated 100 times for each simulation setting.

Three benchmark methods are used for comparison purposes. The first method by Yu and Park (2014) has objective $\text{argmin}_{\boldsymbol{\beta}}\{\|(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1)\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)\|^2 + \lambda \sum |\beta_j|\}$. We denote it as "IdenReg". The second is the threshold gradient descent regularization (TGDR) method discussed in Ma et al. (2006). For all the ROC related methods, e.g. our proposed methods and TGDR, we choose the tuning parameters that maximize the AUCROC in the validation set.

In these simulations, we will generate predictors $\boldsymbol{X}|Y = j$ from a multivariate normal distribution with mean $\boldsymbol{\mu_j}$ and covariance $\boldsymbol{\Sigma}_j$ in all settings. Following the proof from Liu and Bondell (2016) using Neyman-Pearson lemma, we can show that the optimal risk score function is

$$\boldsymbol{X}^T(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{X} + 2(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)^T\boldsymbol{X}. \tag{3.11}$$

If we assume $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are equal or proportional to each other, then the optimal risk score function is linear in $\boldsymbol{X}$, with coefficients $\boldsymbol{\beta} \propto (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Otherwise, if the covariance matrices are not proportional to each other, the optimal risk score would

be quadratic.

Let $q = P(Y = 1)$ be the probability of an instance belonging to the positive class. For each simulation setting, we consider both $q = 0.1$ representing the imbalanced data case and $q = 0.5$ representing the balanced case. Define $\mathbf{0}_k \in \mathbb{R}^k$ as a vector of length k with all elements equal to 0. Similarly, define $\mathbf{1}_k \in \mathbb{R}^k$ as a vector of length k with all elements equal to 1. Define $\mathbf{0}_{k\times l} \in \mathbb{R}^{k\times l}$ as a matrix with k rows and l columns with all elements equal to 0.

When assessing the performance of the variable selection process, we will use the following quantities and plots. For a model with a specific number of variables, the variable true positive rate (Variable TPR, Variable Recall) measures the proportion of important variables that are correctly identified. The variable false positive rate (Variable FPR) measures the proportion of unimportant variables that are incorrectly identified as important. The Variable Precision represents the proportion of detected important variables that are truly important. Knowing the order of the variables that come into the model, we can construct the ROC curve for the variable path which plots the Variable TPR against the Variable FPR as the number of variables in the model is varied. The Precision-Recall curve for the variable path plots the Variable Precision versus the Variable Recall when the number of variables in the model is varied is also constructed.

3.4.1 Simulation Setting 1: Equal AR(1) Variance Model

For a vector of p variables, an AR (1) covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ has the structure of

$$\mathbf{\Sigma} = \sigma^2(\rho^{|i-j|}) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & 1 \end{bmatrix}_{p\times p}, \tag{3.12}$$

where ρ is the AR(1) correlation parameter and σ^2 is the variance. Kac et al. (1953) has shown that the inverse of the AR(1) covariance is given by

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2(1-\rho^2)} \begin{bmatrix} 1 & -\rho & & & & \\ -\rho & 1+\rho^2 & -\rho & & & \\ & -\rho & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & 1+\rho^2 & -\rho \\ & & & & -\rho & 1 \end{bmatrix}_{p\times p}. \tag{3.13}$$

Let $\boldsymbol{\mu}_0 = \mathbf{0}_{500}$, $\boldsymbol{\mu}_1 = (0, \mathbf{1}_8, 0, \mathbf{0}_{490})$. Let $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 = (0.5^{|i-j|})_{500\times 500}$. $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ are equal and they follow the AR(1) variance structure with $\sigma^2 = 1$ and $\rho = 0.5$. Given $y = 0$, we generate $\boldsymbol{X} \sim MN(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$. Given $y = 1$, we generate $\boldsymbol{X} \sim MN(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1)$. According to the best risk score function (3.11), the optimal linear risk score function is linear in X_1 through X_{10} when the data are generated from multivariate normal distributions with

equal variance. Hence, we have 10 true important variables X_1 through X_{10} and the other 490 unimportant variables do not contribute to the classifier in any way. Notice that the mean for the positive group and the negative group only differ in X_2 through X_8. In other words, if one performs variable screening marginally by t-tests, one would fail to detect X_1 and X_{10} as important variables. This leads to an interesting thing to see about our proposed methods.

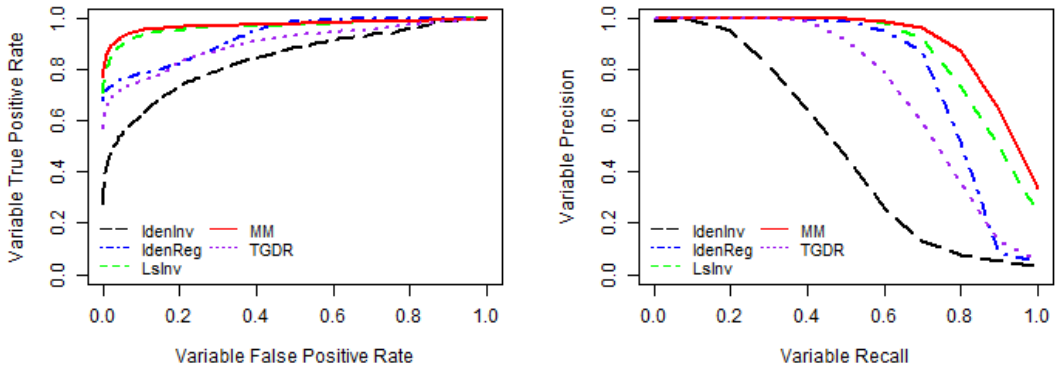
For both the imbalanced case ($q = 0.1$) and the balanced case ($q = 0.5$), Table 3.1 summarizes the AUCROC on the test data sets across 100 replicates. It also summarizes the number of important variables selected (Variable TP) and the number of unimportant variables selected (Variable FP) when using the best tuning parameter. On the other hand, Figure 3.1 plots the ROC curve and the Precision-Recall curve for the variable selection paths. When data are imbalanced and $q = 0.1$, our proposed method MM achieves the highest average binormal AUCROC on the test set across 100 replicates and it is significantly better than TGDR, IdenReg, IdenInv. The other proposed method LsInv is comparable to MM in terms of the AUCROC but is worse than MM in variable selection. It can be seen that MM selects a significantly larger number of true important variables and significantly smaller number of unimportant variables when compared with LsInv. Figure 3.1(a) shows that MM has the best variable selection path, following by LsInv, IdenReg, TGDR, IdenInv, sequentially. When $q = 0.5$, our proposed methods LsInv and MM have very similar outcomes and they both outperform other methods. LsInv and MM are able to select all of the important variables including marginally undetectable variables X_1 and X_{10} with nearly no unimportant variables selected. In Figure 3.1(b), the variable path plot also shows the superiority of LsInv and MM over

other methods like IdenInv, IdenReg and TGDR. The IdenInv estimator always performs the worst and thus we will not show its results in the later simulations and real data applications.

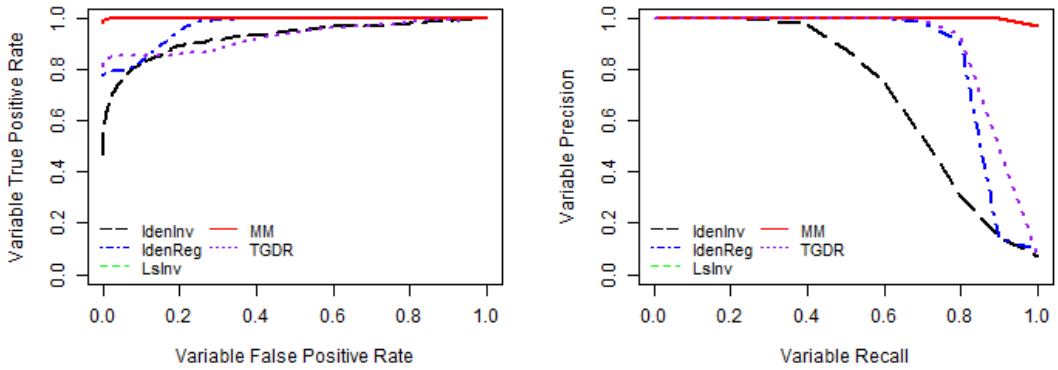
Table 3.1: Equal AR(1) Variance Simulation Performance

		IdenInv	IdenReg	LsInv	MM	TGDR
			q= 0.1			
Binormal AUCROC	Mean	0.8775	0.8884	0.8983	0.9008	0.8921
	SE	0.0024	0.0017	0.0016	0.0016	0.0020
	Median	0.8798	0.8918	0.8986	0.9029	0.8947
TP Variables Selected	Mean	5.1000	7.2500	7.9200	8.1900	6.3600
	SE	0.1168	0.0892	0.1285	0.1195	0.1345
	Median	5.0000	7.0000	8.0000	8.0000	6.0000
FP Variables Selected	Mean	9.7600	3.2700	2.5900	0.9600	1.5000
	SE	1.1339	1.0236	0.4926	0.1392	0.2541
	Median	6.0000	1.0000	1.0000	0.0000	0.0000
			q=0.5			
Binormal AUCROC	Mean	0.9085	0.8994	0.9196	0.9196	0.9162
	SE	0.0009	0.0010	0.0008	0.0008	0.0008
	Median	0.9097	0.9000	0.9203	0.9203	0.9168
TP Variables Selected	Mean	6.8800	8.2700	9.9700	9.9700	8.3900
	SE	0.0913	0.0750	0.0171	0.0171	0.0618
	Median	7.0000	8.0000	10.0000	10.0000	8.0000
FP Variables Selected	Mean	5.9200	11.6300	0.2600	0.2600	0.5100
	SE	0.6634	2.3389	0.0597	0.0597	0.1202
	Median	3.5000	0.0000	0.0000	0.0000	0.0000

For both imbalance case (q=0.1) and balance case (q=0.5), we summarize the the binormal area under the ROC curve (AUCROC) on the test data sets, the number of true important variables (TP) selected and the number of unimportant variables (FP) selected by using the best tuning parameters. The mean associated with the standard error as well as the median across 100 replicates are provided.



(a) q = 0.1



(b) q=0.5

Figure 3.1: Equal AR(1) Variance Simulation ROC Curve and Precision-Recall Curve for Variable Path

3.4.2 Simulation Setting 2: Equal Block Variance Model

Let $\boldsymbol{\mu}_0 = \mathbf{0}_{500}$, $\boldsymbol{\mu}_1 = (\mathbf{1}_{10}, \mathbf{0}_{490})$. Let

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \begin{bmatrix} \boldsymbol{B}_1 & \mathbf{0}_{10\times 490} \\ \mathbf{0}_{490\times 10} & \boldsymbol{B}_2 \end{bmatrix},$$

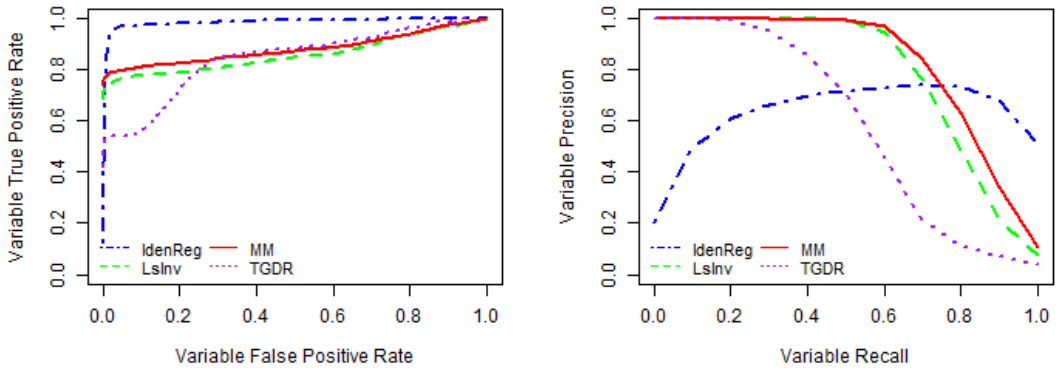
where each diagonal element of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ equals 1, and each off-diagonal element of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ equals to 0.5. Given $y = 0$, we generate $\boldsymbol{X} \sim MN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Given $y = 1$, we generate $\boldsymbol{X} \sim MN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. According to the best risk score function (3.11), the optimal linear risk score function is linear in X_1 through X_{10}. This block variance structure assumes that there is correlation among the important variables but no correlation between the important variables and the unimportant ones.

When $q = 0.1$, Table 3.2 shows that our proposed method MM achieves the highest average binormal AUCROC and it is significantly better than the benchmark methods IdenReg, TGDR. The other proposed method LsInv is comparable to MM in terms of the AUCROC but it identifies fewer important variables and more unimportant ones. Figure 3.2(a) tells us that LsInv and MM identify important variables with no error until they reach a 0.5 recall in all 100 replicates, then unimportant variables come in and IdenReg would be able to identify almost all important variables but with a lot of unimportant ones which makes IdenReg an unfavored method. When $q = 0.5$, LsInv and MM outperform other methods in terms of the binormal AUCROC and the final model it chooses. When comparing the variable selection path for different methods in Figure 3.2(b), it is similar as for the imbalanced case.

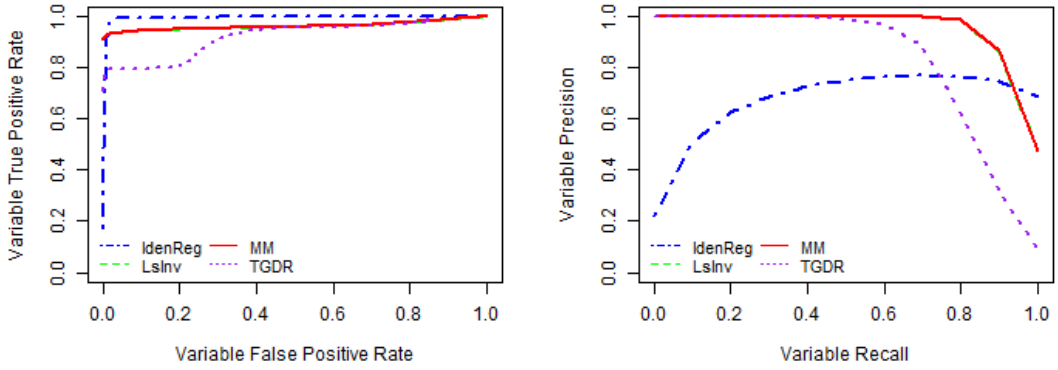
Table 3.2: Equal Block Variance Simulation Performance

		IdenReg	LsInv	MM	TGDR
		q= 0.1			
Binormal AUCROC	Mean	0.8135	0.8172	0.8195	0.8101
	SE	0.0022	0.0021	0.0021	0.0024
	Median	0.8135	0.8189	0.8198	0.8104
Variable TP	Mean	6.8700	6.4500	7.1700	4.7900
	SE	0.2177	0.1344	0.1471	0.1423
	Median	7.0000	6.0000	7.0000	5.0000
Variable FP	Mean	3.6300	0.4600	0.2600	0.8400
	SE	0.4242	0.1086	0.0836	0.1762
	Median	3.0000	0.0000	0.0000	0.0000
		q=0.5			
Binormal AUCROC	Mean	0.8220	0.8245	0.8245	0.8216
	SE	0.0013	0.0013	0.0013	0.0014
	Median	0.8216	0.8233	0.8233	0.8213
Variable TP	Mean	8.7100	9.0300	9.0400	7.2400
	SE	0.1351	0.0870	0.0864	0.1016
	Median	9.0000	9.0000	9.0000	7.0000
Variable FP	Mean	3.5700	0.1800	0.1800	0.4700
	SE	0.2772	0.0520	0.0520	0.1077
	Median	3.0000	0.0000	0.0000	0.0000

For both imbalance case (q=0.1) and balance case (q=0.5), we summarize the the binormal area under the ROC curve (AUCROC) on the test data sets, the number of true important variables (TP) selected and the number of unimportant variables (FP) selected by using the best tuning parameters. The mean associated with the standard error as well as the median across 100 replicates are provided.



(a) q = 0.1



(b) q = 0.5

Figure 3.2: Equal Block Variance Simulation ROC Curve and Precision-Recall Curve for Variable Path

3.4.3 Simulation Setting 3: Unequal Block Variance Model

In this setting, we consider the variance for different groups being unequal. Let $\boldsymbol{\mu}_0 = \mathbf{0}_{500}$, $\boldsymbol{\mu}_1 = (\mathbf{0}_{10}, \mathbf{1}_{10}, \mathbf{0}_{490})$. Let

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \boldsymbol{C}_0 & \mathbf{0}_{10\times 10} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{10\times 10} & \boldsymbol{R} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{480\times 10} & \mathbf{0}_{480\times 10} & \boldsymbol{S} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \boldsymbol{C}_1 & \mathbf{0}_{10\times 10} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{10\times 10} & \boldsymbol{R} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{480\times 10} & \mathbf{0}_{480\times 10} & \boldsymbol{S} \end{bmatrix},$$

where $\boldsymbol{C}_0 \in \mathbb{R}^{10\times 10}$ is a matrix with each diagonal element equals 1 and each off-diagonal element equals to 0.5, $\boldsymbol{C}_1 \in \mathbb{R}^{10\times 10}$ is a matrix with each diagonal element equals 1 and each off-diagonal element equals to 0.3, $\boldsymbol{R} \in \mathbb{R}^{10\times 10}$ and $\boldsymbol{S} \in \mathbb{S}^{480\times 480}$ are both matrices with each diagonal element equals 1 and each off-diagonal element equals to 0.5. Given $y = 0$, we generate $\boldsymbol{X} \sim MN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Given $y = 1$, we generate $\boldsymbol{X} \sim MN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. In order to get the best risk score function in (3.11), we need to know the inverse of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. Since both of them are diagonal blockwise matrix, it can be proved that the inverse of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are also diagonal blockwise matrices with the form of

$$\boldsymbol{\Sigma}_0^{-1} = \begin{bmatrix} \boldsymbol{C}_0^{-1} & \mathbf{0}_{10\times 10} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{10\times 10} & \boldsymbol{R}^{-1} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{480\times 10} & \mathbf{0}_{480\times 10} & \boldsymbol{S}^{-1} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_1^{-1} = \begin{bmatrix} \boldsymbol{C}_1^{-1} & \mathbf{0}_{10\times 10} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{10\times 10} & \boldsymbol{R}^{-1} & \mathbf{0}_{10\times 480} \\ \mathbf{0}_{480\times 10} & \mathbf{0}_{480\times 10} & \boldsymbol{S}^{-1} \end{bmatrix},$$

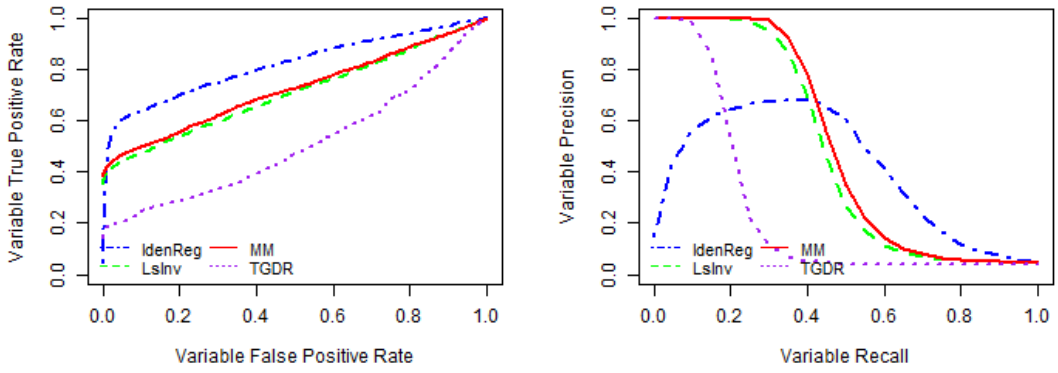
where $\boldsymbol{C}_0^{-1}$ is the inverse of $\boldsymbol{C}_0$, $\boldsymbol{C}_1^{-1}$ is the inverse of $\boldsymbol{C}_1$, $\boldsymbol{R}^{-1}$ is the inverse of $\boldsymbol{R}$, $\boldsymbol{S}^{-1}$ is the inverse of $\boldsymbol{S}$. We can further derive that the optimal linear risk score function is quadratic in X_1 through X_{10} and linear in X_{11} through X_{20}. Hence, we have 20 important variables X_1 through X_{20}.

Table 3.3 and Figure 3.3 shows the performance of different methods. The overall ranking of these different methods is similar to what we observed in Section 3.4.2 where our proposed method MM achieves the highest average binormal AUCROC and identifies more important variables with less unimportant ones. However, we are only able to identify the variables with linear effect and miss those variables with quadratic effect. This seems to be a limitation for all these methods. Nevertheless, the classification performance is still good even when we do not identify all the important variables.

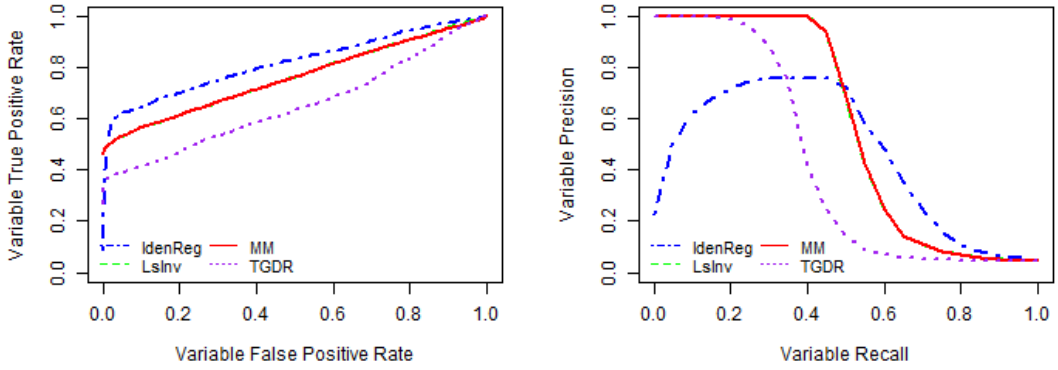
Table 3.3: Unequal Block Variance Simulation Performance

		IdenReg	LsInv	MM	TGDR
		q= 0.1			
Binormal AUCROC	Mean	0.8145	0.8199	0.8227	0.8016
	SE	0.0024	0.0024	0.0024	0.0025
	Median	0.8141	0.8218	0.8220	0.8015
Variable TP	Mean	7.5900	6.8700	7.3100	3.1600
	SE	0.2288	0.1368	0.1323	0.0861
	Median	8.0000	7.0000	7.0000	3.0000
Variable FP	Mean	4.6700	0.3200	0.1400	0.6400
	SE	0.5221	0.0839	0.0472	0.1291
	Median	3.0000	0.0000	0.0000	0.0000
		q=0.5			
Binormal AUCROC	Mean	0.8223	0.8249	0.8249	0.8188
	SE	0.0014	0.0014	0.0014	0.0014
	Median	0.8228	0.8243	0.8243	0.8194
Variable TP	Mean	8.8200	9.1300	9.1400	6.3800
	SE	0.1702	0.0812	0.0804	0.0962
	Median	9.0000	9.0000	9.0000	6.0000
Variable FP	Mean	3.5700	0.1800	0.1800	0.7100
	SE	0.2900	0.0716	0.0716	0.0868
	Median	3.0000	0.0000	0.0000	1.0000

For both imbalance case (q=0.1) and balance case (q=0.5), we summarize the the binormal area under the ROC curve (AUCROC) on the test data sets, the number of true important variables (TP) selected and the number of unimportant variables (FP) selected by using the best tuning parameters. The mean associated with the standard error as well as the median across 100 replicates are provided.



(a) q = 0.1



(b) q = 0.5

Figure 3.3: Unequal Block Variance Simulation ROC Curve and Precision-Recall Curve for Variable Path

3.5 Real Data Application

We applied our methods on two real data sets, the Ionosphere data (Sigillito et al., 1989) and the Surgery data (Zikeba et al., 2013) from UCI data repository (Lichman, 2013). The summary of these two data sets are described in Table 3.4. For the Ionosphere data, the goal is to predict whether the radar returns are good or bad. For the Surgery data, we are interested in predicting whether patients will die or survive in a one year period from major lung resections.

We randomly split each data set into 50% training, 25% validation and 25% testing in a stratified fashion preserving the original ratio between the two classes. The process is repeated 200 times. The performance on the test set are provided in Table 3.5. Since whether the real data satisfy our binormal assumption is unknown, we use the empirical area under the ROC curve (AUCROC) and the empirical area under the Precision-Recall curve (AUCPR) to measure the performance. In both real data sets, the proposed method MM seems to have a higher empirical AUCROC and empirical AUCPR but with fewer variables needed in the model.

Table 3.4: Summary of Datasets

Datasets	Samples	Minority Proportions	Continuous Predictors	Categorical Predictors
Ionosphere	351	35.9%	32	1
Surgery	470	14.9%	3	13

Table 3.5: Real Data Performance

		IdenReg	LsInv	MM	TGDR
		Ionosphere Data			
Empirical AUCROC	Mean	0.8813	0.8799	0.8804	0.8782
	SE	0.0031	0.0031	0.0030	0.0032
	Median	0.8876	0.8853	0.8873	0.8847
Empirical AUCPR	Mean	0.8776	0.8786	0.8796	0.8792
	SE	0.0028	0.0027	0.0026	0.0027
	Median	0.8799	0.8817	0.8836	0.8829
Variables Selected	Mean	13.8700	10.5000	10.0650	10.9550
	SE	0.5189	0.5079	0.4886	0.5077
	Median	12.0000	8.0000	8.0000	8.0000
		Surgery Data			
Empirical AUCROC	Mean	0.6262	0.6324	0.6341	0.6271
	SE	0.0046	0.0048	0.0046	0.0043
	Median	0.6311	0.6358	0.6389	0.6261
Empirical AUCPR	Mean	0.8071	0.8035	0.7988	0.8055
	SE	0.0024	0.0023	0.0021	0.0021
	Median	0.8049	0.8069	0.7943	0.8074
Variables Selected	Mean	12.6341	13.5238	11.5000	13.2439
	SE	0.4080	0.4856	0.4522	0.4491
	Median	13.0000	15.0000	12.0000	13.0000

For both real data sets, we summarize the number of variables selected by using the best tuning parameters, the empirical area under the ROC curve (AUCROC) and the empirical area under the Precision Recall curve (AUCPR) on the test sets. The mean associated with the standard error as well as the median across 200 replicates are all provided.

3.6 Conclusion

In this chapter, we propose three classification and variable selection methods: IdenInv, LsInv and MM, that are based on the binormal AUCROC. Both simulation studies and real data applications show that two of the proposed methods, LsInv and MM, are very useful and achieve superior performance in terms of both classification and variable selection when compared with other competing methods, e.g. IdenReg estimator and the threshold gradient descent regularization method (TGDR).

BIBLIOGRAPHY

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4): 387–415, 1975.

Kay H Brodersen, Cheng Soon Ong, Klaas E Stephan, and Joachim M Buhmann. The binormal assumption on precision-recall curves. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4263–4266. IEEE, 2010.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

Jesse Davis, Elizabeth S Burnside, Inês de Castro Dutra, David Page, Raghu Ramakrishnan, Vitor Santos Costa, and Jude W Shavlik. View learning for statistical relational learning: With an application to mammography. In *IJCAI*, pages 677–683. Citeseer, 2005.

Donald D Dorfman and Edward Alf Jr. Maximum likelihood estimation of parameters of signal detection theorya direct solution. *Psychometrika*, 33(1):117–124, 1968.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

Irina Gaynanova, James G Booth, and Martin T Wells. Penalized versus constrained generalized eigenvalue problems. *Journal of Computational and Graphical Statistics*, (just-accepted):1–24, 2016.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

Man-Jen Hsu and Huey-Miin Hsueh. The linear combinations of biomarkers which maximize the partial area under the roc curves. *Computational Statistics*, 28(2):647–666, 2013.

David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Finn V Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

M Kac, WL Murdock, and G Szego. On the eigen-values of certain hermitian forms. *Journal of Rational Mechanics and Analysis*, 2(6):767–802, 1953.

Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques, 2007.

Kenneth Lange. *Optimization*. New York: Springer, 2004.

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Aiyi Liu, Enrique F Schisterman, and Yan Zhu. On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 24(1):37–47, 2005.

Zhongkai Liu and Howard Bondell. Classification and variable selection methods for ultrahigh dimensional and imbalanced data. 2016.

Shuangge Ma and Jian Huang. Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24):4356–4362, 2005.

Shuangge Ma, Xiao Song, and Jian Huang. Regularized binormal roc method in disease classification using microarray data. *BMC bioinformatics*, 7(1):253, 2006.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Nils J Nilsson. Learning machines. 1965.

Margaret Sullivan Pepe and Mary Lou Thompson. Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140, 2000.

Margaret Sullivan Pepe, Tianxi Cai, and Gary Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229, 2006.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Frank Rosenblatt. Principles of neurodynamics. 1962.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

Vincent G Sigillito, Simon P Wing, Larrie V Hutton, and Kile B Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266, 1989.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

John Q Su and Jun S Liu. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355, 1993.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Daniela M Witten and Robert Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

Wenbao Yu and Taesung Park. Aucpr: An auc-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC genomics*, 15(10):1, 2014.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Guoqiang Peter Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462, 2000.

Maciej Zikeba, Jakub M Tomczak, Marek Lubicz, and Jerzy 'Swikatek. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, 2013.

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

APPENDIX

A

APPENDIX

A.1 Derivation of Algorithm 2.

Since $\boldsymbol{A} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T$ is a positive semidefinite matrix, we have $h^*(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}$ convex in $\boldsymbol{\beta}$. Then for a fix value of $\boldsymbol{\beta}^{(m)}$, we have

$$\begin{aligned} h^*(\boldsymbol{\beta}) &\geq h^*(\boldsymbol{\beta}^{(m)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(m)})^T \nabla h^*(\boldsymbol{\beta}^{(m)}) \\ &= \boldsymbol{\beta}^{(m)^T} \boldsymbol{A} \boldsymbol{\beta}^{(m)} + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(m)})^T 2\boldsymbol{A}\boldsymbol{\beta}^{(m)} \\ &= 2\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m)^T} \boldsymbol{A} \boldsymbol{\beta}^{(m)}, \end{aligned} \tag{A.1}$$

for any $\boldsymbol{\beta}$ and equality holds when $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$. We denote the objective function in (3.10) as

$$h(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta} - \lambda \|\boldsymbol{\beta}\|_1, \tag{A.2}$$

where $\lambda \geq 0$ is a tuning parameter. Hence, we can find the minorization function

$$g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) = 2\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m)^T} \boldsymbol{A} \boldsymbol{\beta}^{(m)} - \lambda \|\boldsymbol{\beta}\|_1, \tag{A.3}$$

which minorize the objective function (A.2) at $\boldsymbol{\beta}^{(m)}$ since it satisfies $h(\boldsymbol{\beta}) \geq g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)})$ for any $\boldsymbol{\beta}$ and equality holds when $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$. A minorization-maximization (MM) algorithm (Lange et al., 2000; Hunter and Lange, 2004; Lange, 2004) is used here. It initialize with $\boldsymbol{\beta}^{(0)}$ and in each iteration, we calculate

$$\boldsymbol{\beta}^{(m+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ g(\boldsymbol{\beta}|\boldsymbol{\beta}^{(m)}) \right\} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ 2\boldsymbol{\beta}^T \boldsymbol{A} \boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m)^T} \boldsymbol{A} \boldsymbol{\beta}^{(m)} - \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{A.4}$$

It can be shown that the objective function $h(\boldsymbol{\beta})$ is non-decreasing along the iterations since $h(\boldsymbol{\beta}^{(m+1)}) \geq g(\boldsymbol{\beta}^{(m+1)}|\boldsymbol{\beta}^{(m)}) \geq g(\boldsymbol{\beta}^{(m)}|\boldsymbol{\beta}^{(m)}) = h(\boldsymbol{\beta}^{(m)})$. Now let's add the constraint $\boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} \leq 1$ into the process, in each iteration, we have

$$\boldsymbol{\beta}^{(m+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ 2\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}^{(m)} - \lambda\|\boldsymbol{\beta}\|_1 \right\} \text{ subject to } \boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} \leq 1 \tag{A.5}$$

where $\lambda \geq 0$ is a tuning parameter. To solve (A.5), it is proved by Witten and Tibshirani (2011) in Proposition 2 that we can first solve

$$\boldsymbol{d} = \underset{\boldsymbol{d}}{\operatorname{argmin}} \left\{ \boldsymbol{d}^T\boldsymbol{W}\boldsymbol{d} - 2\boldsymbol{d}^T\boldsymbol{A}\boldsymbol{\beta}^{(m)} + \lambda\|\boldsymbol{d}\|_1 \right\}, \tag{A.6}$$

Then we calculate $\boldsymbol{\beta}$ as below,

$$\boldsymbol{\beta}^{(m+1)} = \begin{cases} 0 & \text{if } \boldsymbol{d} = 0, \\ \dfrac{\boldsymbol{d}}{\sqrt{\boldsymbol{d}^T\boldsymbol{W}\boldsymbol{d}}} & \text{otherwise.} \end{cases} \tag{A.7}$$

To solve (A.6), we consider two cases.

1. If $\boldsymbol{W}$ is diagonal matrix, the objective function (A.6) is separable, we can solve directly for i^{th} element of $\boldsymbol{d}$ followed by Gaynanova et al. (2016). For $i = 1, 2, \cdots, p$,

we have

$$\begin{aligned} d_i &= \underset{d_i}{\operatorname{argmin}} \Big\{ w_{ii} d_i^2 - 2(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i d_i + \lambda |d_i|_1 \Big\} \\ &= \underset{d_i}{\operatorname{argmin}} \left\{ 2w_{ii} \left[\frac{1}{2} \Big(d_i - \frac{(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i}{w_{ii}} \Big)^2 + \frac{\lambda}{2w_{ii}} |d_i| \right] \right\} \\ &= ST\Big(\frac{(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i}{w_{ii}}, \frac{\lambda}{2w_{ii}} \Big) \end{aligned} \tag{A.8}$$

where w_{ii} is the i^{th} is diagonal element of $\boldsymbol{W}$, $(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i$ is the i^{th} element of $\boldsymbol{A}\boldsymbol{\beta}^{(m)}$, and $ST(z, r) = \text{sgn}(z)(|z| - r)_+$.

2. if $\boldsymbol{W}$ is symmetric but not diagonal matrix, the objective function (A.6) is no longer separable and we can solve it via coordinate descent algorithm. In each iteration k, for $i = 1, 2, \cdots, p$,

$$\begin{aligned} d_i^{(k+1)} &= \underset{d_i}{\operatorname{argmin}} \Big\{ w_{ii} d_i^2 + 2(\boldsymbol{d}^{(k)})_{-i}^T \boldsymbol{W}_{-i} \boldsymbol{d}_i^{(k)} - 2(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i d_i + \lambda |d_i|_1 \Big\} \\ &= \underset{d_i}{\operatorname{argmin}} \left\{ 2w_{ii} \left[\frac{1}{2} \Big(d_i - \frac{(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i - (\boldsymbol{d}^{(k)})_{-i}^T \boldsymbol{W}_{-i}}{w_{ii}} \Big)^2 + \frac{\lambda}{2w_{ii}} |d_i| \right] \right\} \\ &= ST\Big(\frac{(\boldsymbol{A}\boldsymbol{\beta}^{(m)})_i - (\boldsymbol{d}^{(k)})_{-i}^T \boldsymbol{W}_{-i}}{w_{ii}}, \frac{\lambda}{2w_{ii}} \Big) \end{aligned} \tag{A.9}$$

where $\boldsymbol{d}_{-i}^{(k)} = (d_1^{(k+1)}, \cdots, d_{i-1}^{(k+1)}, d_{i+1}^{(k)}, \cdots d_p^{(k)})$, w_{ii} is the i^{th} diagonal element of $\boldsymbol{W}$, $\boldsymbol{W}_{-i} = (w_{1i}, \cdots, w_{i-1,i}, w_{i+1,i}, \cdots, w_{pi})$.