ABSTRACT

LIU, ZHONGKAI. Classification and Variable Selection Methods for Ultrahigh Dimensional and Imbalanced Data. (Under the direction of Howard D. Bondell.)

Classification and variable selection play important roles in machine learning and statistical applications. Classification methods are used in a broad range of application areas, from medical diagnosis to anomaly detection in signal analysis, from credit risk screening to quality control, from image segmentation to information retrieval. Variable selection or feature selection, whose purpose is to eliminate irrelevant variables, is undoubtedly important, especially in high dimensional applications.

Although there are a large number of classification and variable selection methods so far, classification and variable selection on imbalanced data, i.e. a large skew in the class distribution, is a challenging problem. Evaluation of classifiers via the receiver operating characteristic (ROC) curve is common in binary classification. Techniques to develop classifiers that optimize the area under the ROC curve have been proposed. However, for imbalanced data, the ROC curve tends to give an overly optimistic view. Realizing its disadvantages of dealing with imbalanced data, Precision-Recall (PR) curves have recently become a basis for assessing classification methods on class-imbalanced data.

In this thesis work, we focus on classification and variable selection methods for ultrahigh dimensional and imbalanced data. Specifically, we investigate two types of problems. The first problem is related to the imbalanced data sets, i.e. there are many more examples from some classes than from others. The imbalanced problem is prevalent in many applications, including fraud and churn detection, text categorization, medical diagnosis, detection of software defects, etc. The second problem is on ultrahigh dimensions, where the dimension of variables is much larger than the sample size. Due to the huge improvement in data gathering and data processing mechanisms, ultrahigh dimensional data arises in various areas of modern scientific research using quantitative measurements.

The thesis is organized as follows. In Chapter 1, we provide introductions and backgrounds through an overview of current popular classification and variable selection methods, including their advantages and disadvantages. Chapter 2 is devoted to proposing an efficient classification approach with imbalanced data based on estimating the risk score function via maximization of the area under the Precision-Recall curve. The binormal Precision-Recall idea is extended to variable selection in Chapter 3. We propose

a regularized binormal Precision-Recall algorithm, which applies the threshold gradient descent regularization (TGDR) method to select important variables that can maximize the area under the Precision-Recall curve (AUCPR) in a binormal framework. Chapter 4 shows our solution to the variable selection problem on ultrahigh dimensions. We propose a principal components adjusted variable screening method, which uses top principal components as surrogate covariates to account for the variability of the omitted predictors in generalized linear models. The proposed approaches are demonstrated with both nice theoretical properties and superior numerical performances compared with the competing methods.

Classification and Variable Selection Methods for Ultrahigh Dimensional and Imbalanced Data

by
Zhongkai Liu

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:

Tao Pang	Peter Bloomfield
Rui Song	Howard D. Bondell Chair of Advisory Committee

DEDICATION

To my loving family.

BIOGRAPHY

The author was born in Zibo, Shandong, China. After graduating from Nankai High School in 2007, he started his Statistics career in Shanghai University of Finance and Economics, majoring in Statistics, and minoring in Finance. In 2011, he was awarded Shanghai Outstanding Graduate and granted a full scholarship to North Carolina State University to pursue a Doctoral degree in Statistics. His research areas include machine learning, classification, variable selection, high dimensional data analysis, asset pricing, and risk management, under the valuable guidance of his advisor Dr. Howard Bondell, Dr. Rui Song, and Dr. Tao Pang. He will graduate with his Doctoral degree in Statistics in May, 2016.

ACKNOWLEDGEMENTS

I would like to thank my advisor and Director of Statistics Graduate Programs, Dr. Howard Bondell, for his insightful guidance and constant help for my study and research. As an expert in variable selection and classification methods, he is connecting the academic research with the industry world closely, which grows my passion for statistics and research. The weekly meetings with him over 3 years bring me the inspiration and open my horizon not only in the PhD research but in my career development as well. I feel lucky to have such an excellent professor as my advisor.

I would also like to extend my appreciation to Dr. Rui Song, Dr. Tao Pang, and Dr. Peter Bloomfield, for their selfless encouragement and valuable mentorship. Admittedly, Dr. Rui Song is my first research mentor who initiates me into the statistics research. I sincerely express my gratitude to her for the generous guidance and considerate patience during our project on high dimensional data analysis, which is a great start in my PhD career. I regard Dr. Tao Pang as a lifelong mentor who shows me the magic of quantitative analysis in the financial market. Talking with him is thought-provoking, leading to fruitful remarkable work including our published paper *An Efficient Grid Lattice Algorithm for Pricing American-style Options* on *International Journal of Financial Markets and Derivatives* and more smart ideas about risk management in progress. I really appreciate the precious quantitative research work with him. Dr. Peter Bloomfield is my professional mentor on a business analytical project in Allegro MicroSystems, where I completed my first internship. I have learned a lot form his intelligence and diligence throughout the project, not only on predictive modeling and time series analysis skills, but also the passion and strictness for the truth. It is my great honor to have you all in my PhD committee.

I am also grateful to all the faculty and staff in the Department of Statistics at North Carolina State University for offering a wonderful study environment, a comprehensive collection of courses, and a great number of industry opportunities. I really appreciate the precious working experiences in SAS Institute during my fourth year and in Allegro MicroSystems during my third year, which are provided as Co-op Programs by the department. The big data course by Dr. Lexin Li and Dr. Brian Reich, the computing course by Dr. Hua Zhou, the time series course by Dr. Soumendra Lahiri, the advanced statistical inference course by Dr. Dennis Boos, the categorical data analysis course by Dr.

Daowen Zhang, the experimental statistics course by Dr. Jason Osborne, and so on, play significant parts in my statistician career. All the training helps me get the challenging data science internship in AT&T Big Data Team in Silicon Valley during the summer of 2015, a great opportunity to work with the smartest engineers and inspire my deeper enthusiasm about data science.

Last but not least, I would like to express my deepest gratitude to my family and friends for their unconditional love and support. To my loving parents, your constant caring and understanding is the source of my power. I miss you all the time. To my dear friends, thank you for walking with me whenever I need support, walking ahead of me whenever I need guidance, walking behind me whenever I need someone to watch my back. To myself, PhD is a lifelong learning path. A Doctoral degree is not the end, not even the beginning of the end, but perhaps the end of the beginning. Live and learn!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Classification and variable selection are hot topics in machine learning and statistical applications. Classification methods solve the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations whose category membership is already known, in a broad range of application areas, from medical diagnosis to anomaly detection in signal analysis, from credit risk screening to quality control, from image segmentation to information retrieval. Feature or variable selection is to eliminate irrelevant variables to enhance the generalization performance of a given learning algorithm, especially in high dimensional applications. In regression and classification problems with a large number of predictors, if only some of the variables contribute to the response, overfitting appears to be a critical concern for statistical analysis. Consequently, finding the optimal subset of variables among the pool is necessary and momentous.

The empirical accuracy, or correct classification rate, plays an important role in measuring the performance of each classification algorithm. The empirical accuracy criterion is straightforward, but depends on a particular choice of threshold and when facing an imbalanced dataset may not be an appropriate measure (Japkowicz and Stephen, 2002; Brodersen et al., 2010a). The accuracy is influenced by the decision threshold c. To overcome the above limitations, the receiver operating characteristic (ROC) curve (Hanley

and McNeil, 1982) appeared as a graphical plot that illustrates the performance of a binary classifier system as its decision threshold is varied. However, Davis and Goadrich (2006) pointed out the disadvantage of ROC curves whenever there is a large skew in the class distribution. In terms of classification of highly imbalanced data, ROC curves tend to present an overly optimistic view of an algorithm's performance. In fact, there are many situations where a large skew exists, such as anomaly detection for rare events. As an alternative to ROC curves for tasks with imbalanced data, Precision-Recall (PR) curves (Raghavan et al., 1989) have recently become a basis for assessing classification methods. Davis and Goadrich (2006) discussed the relationship between ROC and Precision-Recall curves, and Craven (2005); Davis et al. (2005); Kok and Domingos (2005); Singla and Domingos (2005) cited Precision-Recall curves as a better tool considering its performance in front of imbalanced data.

In the classification context, there are many feature selection methods. In general, they can be categorized into two groups. The first group of methods are those that incorporate variable selection algorithms as part of the classification procedure such as Classification And Regression Tree (CART) (Breiman et al., 1984), random forests (Breiman, 2001), and GUIDE (Loh, 2002). CART introduces the variable importance as the decrease of node impurity measures and thus provides a subset of optimal splitting variables to be the most significant variables after pruning (Lewis, 2000), which has been widely applied (Bittencourt and Clarke, 2004; Questier et al., 2005; Gey and Nedelec, 2005; Brezigar-Masten and Masten, 2012). Random forests propose the variable importance as a permutation importance index, ordering and selecting variables by importance scores and some filter or wrapper methods. Many works appreciate random forests for variable selection since they can handle interactions between variables (Archer and Kimes, 2008; Rodenburg et al., 2008; Yang and Gu, 2009). Genuer et al. (2010) proposed a strategy involving a ranking of explanatory variables using the random forests score of importance and a stepwise ascending variable introduction strategy. Using random forests, Hapfelmeier and Ulm (2013) developed a new variable selection approach based on the theoretical framework of permutation tests. Sauve and Tuleau-Malot (2014) designed an automatic and exhaustive procedure which relies on the use of the CART algorithm. GUIDE (short for Generalized, Unbiased Interaction Detection and Estimation) is a tree-based method by building piecewise constant and piecewise linear regression models with univariate splits. Based on GUIDE, Loh (2009) introduced techniques for variable selection by

controlling the search for local interactions and employing more effective variable and split selection strategies. Loh (2012) presented an alternative method, derived from the GUIDE classification and regression tree algorithm, that employs recursive partitioning to determine the degree of importance of the variables. Recently Tang et al. (2014) proposed a nonparametric method for variable selection and classification called Categorical Adaptive Tube Covariate Hunting (CATCH) that constructs a nonparametric measure of the relational strength between each predictor and the categorical response, and retains those predictors whose relationship to the response is above a certain threshold. The second group includes methods that incorporate variable selection by applying shrinkage methods with norm constraints (Frank and Friedman, 1993) on the parameters that generate sparse vectors of parameter estimates. Bradley and Mangasarian (1998) investigated the use of L_1 penalty (Tibshirani, 1996) in the support vector machine (SVM) (Vapnik and Vapnik, 1998) to do the variable selection. Zhu et al. (2004) proposed an algorithm for calculating the solution path of the L_1 SVM as a function of its tuning parameter. Wang and Shen (2007); Zhang et al. (2008) implemented variable selection with classification based on SVM, and Qiao et al. (2008) incorporated variable selection into linear discriminant analysis (Fisher, 1936; McLachlan, 2004). Liu and Wu (2012) designed a SVM-based variable selection method in regression and classification via a new penalty that combines the L_0 and L_1 penalties.

Although so far there are a large number of classification and variable selection methods, problem occurs when facing imbalanced data sets, i.e. there are many more examples from some classes than from others (Chawla et al., 2004). In the binary classification setting, imbalanced data means one of the classes (the negative group) is significantly larger than the other (the positive group) in terms of number of instances. The imbalanced problem is prevalent in many applications, including fraud and churn detection (Anil Kumar and Ravi, 2008), text categorization (Zheng et al., 2004), medical diagnosis (Van Hulse et al., 2009), detection of software defects (Park et al., 2013), etc. Given that the Precision-Recall curve is recommended to evaluate classifiers, it seems natural to use this criterion to estimate the classifier directly. This is what is proposed in this thesis work. In a binormal framework (Brodersen et al., 2010b), we propose an efficient classification approach with imbalanced data based on estimating the risk score function via maximization of the area under the Precision-Recall curve, as well as a regularized binormal Precision-Recall algorithm for variable selection in the classification context, which applies the threshold

gradient descent regularization method (Friedman and Popescu, 2003) to maximize the area under the Precision-Recall curve.

Another problem we consider is ultrahigh dimensions, where the dimension of variables is much larger than the sample size. Due to the huge improvement in data gathering and data processing mechanisms, ultrahigh dimensional data arises in various areas of modern scientific research using quantitative measurements. However, it is often the case that only a relatively small subset of the predictors contribute to the response. Considering the computation cost, statistical accuracy and robustness in the ultrahigh dimensional problem, lots of traditional statistical methods fail to work, including bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Dantzig selection (Candes and Tao, 2007), and other folded concave regularization methods (Fan and Lv, 2011; Zhang and Zhang, 2012). In this thesis work, we propose a principal component-adjusted screening (PCAS) method for generalized linear models, which uses principal components as surrogate covariates to account for omitted covariates in marginal screening.

We demonstrate via both theoretical arguments, as well as extensive simulations and real data analysis, that the proposed ideas work well in classification and variable selection for ultrahigh dimensional and imbalanced data.

CHAPTER 2

Optimal Classification of Imbalanced Data

2.1 Introduction

Consider a classification problem with n independent observations. Denote the observed data by $\{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$ with $\mathbf{X}_i \in \mathcal{R}^p$ and $Y_i \in \{0, 1\}$. We consider the linear risk score as our basis for classification so that we classify to $Y = 1$ if $\boldsymbol{\beta}^T\mathbf{X} > c$ based on some decision threshold c. Note that any monotone transformation of $\boldsymbol{\beta}^T\mathbf{X}$ can be applied, so this is essentially a single index model.

Assuming classification based on the linear risk score, the task is to estimate $\boldsymbol{\beta}$. Methods based on misclassification rate or surrogates, such as support vector machines (SVMs) (Cortes and Vapnik, 1995), rely on a fixed threshold that is estimated as an intercept. Since the overall performance of a classifier can be measured by the area under the curve (AUC), either ROC or PR curve, $\boldsymbol{\beta}$ can be estimated by maximizing the AUC value. Pepe et al. (2006) proposed an ROC estimator defined as the maximizer of the empirical AUC of ROC curve (AUCROC). Ma and Huang (2005) used a sigmoid approximation to the empirical area under the ROC curve as the objective function for classification and designed a threshold gradient descent regularization method for estimation and variable selection in the linear risk score function. Another group of methods were developed using the binormal AUCROC (Dorfman and Alf, 1968; Metz and Kronman, 1980; Metz and

Pan, 1999; Pepe, 2003), which assumes a pair of normal distributions underlying the positive and negative groups. Pepe (2003) showed that the binormal AUCROC provides more valuable information and exhibits more stability than the empirical AUC in the low dimensional situation, while in the high dimensional case the binormal AUCROC is computationally more affordable than the empirical one. Ma et al. (2006) implemented the threshold gradient descent regularization method for estimation and selection with the binormal AUCROC as the objective function.

On the other hand, an empirical Precision-Recall curve is difficult to use because of its sensitivity to idiosyncracies in the data, especially at high precisions (Davis and Goadrich, 2006). Clémençon and Vayatis (2009) addressed this problem by applying a nonparametric approach to derive a smooth Precision-Recall curve. Boyd et al. (2013) gave a detailed overview of current methods to compute the area under the empirical Precision-Recall curve, including lower trapezoid, average precision, interpolated median, and confidence interval estimation. In order to overcome the disadvantages of the empirical version, Brodersen et al. (2010b) discussed the Precision-Recall curve on the basis of the binormal model.

Although discussion of the Precision-Recall curve appears in the literature, its usage has been solely for evaluation of computing classifiers. However, given that it is recommended to evaluate classifiers, it seems natural to use this criterion to estimate the classifier directly. This is what is proposed in this thesis work. We propose an approach for binary classification with imbalanced data in a binormal framework. The approach is to estimate the risk score function via maximization of the area under the Precision-Recall curve based on the binormal assumption. We demonstrate via both theoretical arguments, as well as simulation and real data, that the proposed approach outperforms approaches based on the area under the ROC curve.

The chapter is organized as follows. In section 2.2, we introduce the relevant background on the binormal assumption, ROC, and Precision-Recall curves. Section 3.2 is devoted to deriving the estimators and their asymptotic distributions under the ROC and Precision-Recall framework respectively. Three simulation settings are presented in section 3.3 and two real data analysis results are illustrated in section 3.4. Section 3.5 gives discussions. All technical details are provided in the Appendix.

2.2 ROC and Precision-Recall curves

2.2.1 Preliminaries

A confusion matrix is a structure that represents the binary classification result for a particular choice of threshold, c, as shown in Table 2.1. We follow the notations in Davis and Goadrich (2006). We call $Y = 1$ a positive example, and $Y = 0$ a negative example. True Positives (TP) represent examples correctly labeled as positive, while False Positives (FP) represent negative examples but labeled incorrectly as positive. True Negatives (TN) refer to negative examples that are correctly labeled as negative. Similarly, False Negatives (FN) correspond to those indeed positive while labeled as negative examples. Based on the confusion matrix, as a function of the threshold, we can construct an ROC curve or a Precision-Recall curve. Furthermore, based on the binormal assumption, it becomes a parametric confusion matrix, and a functional form of the ROC curve and Precision-Recall curve can be derived. As a result, the area under the curve, including AUCROC and AUCPR, can be computed.

Table 2.1: Confusion Matrix.

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

2.2.2 Binormal assumption

For a given threshold c, we classify to $Y = 1$ if the decision score $\boldsymbol{\beta}^T\mathbf{X} > c$, and $Y = 0$ otherwise. Dorfman and Alf (1968); Brodersen et al. (2010b) described the binormal assumption, which is to assume the decision scores to follow two independent Gaussian distributions, one for the positive group and one for the negative group expressed as

$$\boldsymbol{\beta}^T\mathbf{X}|Y = 1 \sim \text{Normal}(\mu_p, \sigma_p^2),$$
$$\boldsymbol{\beta}^T\mathbf{X}|Y = 0 \sim \text{Normal}(\mu_n, \sigma_n^2),$$

where without loss of generality, assume $\mu_p \geq \mu_n$. If not, we can change the sign of $\boldsymbol{\beta}$.

Zou and Hall (2000); Krzanowski and Hand (2009) displayed broad applications of the binormal assumption. Note that the normality assumption is on a linear combination of $\mathbf{X}$ rather than $\mathbf{X}$ itself. This assumption on a linear combination would be more likely to hold, particularly for moderate to large p, due to the Central Limit Theorem effect.

2.2.3 ROC curve

The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), where FPR and TPR are computed based on the information from the confusion matrix.

$$\text{TPR} = \frac{TP}{TP + FN},$$
$$\text{FPR} = \frac{FP}{FP + TN}.$$

Based on the binormal assumption it follows that

$$\text{TPR} = P(\boldsymbol{\beta}^T\mathbf{X} > c|Y = 1) = 1 - \Phi\left(\frac{c - \mu_p}{\sigma_p}\right), \tag{2.1}$$

$$\text{FPR} = P(\boldsymbol{\beta}^T\mathbf{X} > c|Y = 0) = 1 - \Phi\left(\frac{c - \mu_n}{\sigma_n}\right). \tag{2.2}$$

After some algebra, the relationship betweeen TPR and FPR, i.e. the ROC curve by (FPR, TPR) pairs is of the form:

$$\text{TPR} = \Phi\left(\frac{\mu_p - \mu_n}{\sigma_p} + \frac{\sigma_n}{\sigma_p}\Phi^{-1}(\text{FPR})\right), \tag{2.3}$$

where $\Phi(\cdot)$ represents the cumulative density function of the standard normal distribution.

Integrating the right hand side of (2.3) on the interval $(0, 1)$ gives the AUC value. Bamber (1975) showed that the area under a ROC curve is equal to the probability that the learner will assign a higher score to a randomly drawn positive sample than to a randomly drawn negative sample. Consequently, under the binormal assumption and

linear classifier, the area under ROC curve can be equivalently represented as

$$\begin{aligned} \text{AUCROC} &= P(\boldsymbol{\beta}^T\mathbf{X}_p > \boldsymbol{\beta}^T\mathbf{X}_n) = P(\boldsymbol{\beta}^T\mathbf{X}_p - \boldsymbol{\beta}^T\mathbf{X}_n > 0) \\ &= 1 - \Phi\left(-\frac{\mu_p - \mu_n}{\sqrt{\sigma_p^2 + \sigma_n^2}}\right) = \Phi\left(\frac{\mu_p - \mu_n}{\sqrt{\sigma_p^2 + \sigma_n^2}}\right), \end{aligned} \tag{2.4}$$

where $\mathbf{X}_p$ denotes a random draw from the distribution $\mathbf{X}|Y = 1$ and $\mathbf{X}_n$, a random draw from the distribution $\mathbf{X}|Y = 0$.

2.2.4 Precision-Recall curve

Precision-Recall curves instead plot Precision, measuring the fraction of examples classified as positive that are truly positive, versus Recall (same as TPR). The key difference is in the use of Precision instead of FPR. This is akin to the use of False Discovery Rate instead of Type I error in multiple testing problems. Unlike the ROC curve, the Precision-Recall curve will also depend on the fraction of positive examples, denoted as π. Based on the confusion matrix, binormal assumption and linear classifier, Recall and Precision can be expressed as

$$\begin{aligned} &\text{Recall} = \text{TPR} = \frac{TP}{TP + FN} = P(\boldsymbol{\beta}^T\mathbf{X} > c|Y = 1) = 1 - \Phi(\frac{c - \mu_p}{\sigma_p}), \\ &\text{Precision} = \frac{TP}{TP + FP} = P(Y = 1|\boldsymbol{\beta}^T\mathbf{X} > c) \\ &= \frac{P(\boldsymbol{\beta}^T\mathbf{X} > c|Y = 1)P(Y = 1)}{P(\boldsymbol{\beta}^T\mathbf{X} > c|Y = 1)P(Y = 1) + P(\boldsymbol{\beta}^T\mathbf{X} > c|Y = 0)P(Y = 0)} \text{ (By Bayes' Rule)} \\ &= \frac{\pi \times \text{Recall}}{\pi \times \text{Recall} + (1 - \pi)[1 - \Phi(\frac{c - \mu_n}{\sigma_n})]}. \end{aligned}$$

By linking Recall and Precision through the threshold c, the relationship between Precision and Recall, i.e. the Precision-Recall curve, can be expressed as

$$\text{Precision} = \frac{\pi \times \text{Recall}}{\pi \times \text{Recall} + (1 - \pi)\Phi(\frac{\mu_n - \mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(\text{Recall}))}. \tag{2.5}$$

The area under the Precision-Recall curve is then

$$\text{AUCPR} = \int_0^1 \frac{\pi \times t}{\pi \times t + (1-\pi)\Phi(\frac{\mu_n - \mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))} dt, \tag{2.6}$$

where $t = \text{Recall}$.

The PR space has a clear advantage over the ROC space when facing imbalanced data, and this difference exists because in this domain the number of negative examples greatly exceeds the number of positive examples. Consequently, a large change in the number of False Positives can lead to a small change in the false positive rate used in ROC analysis. Precision, on the other hand, by comparing False Positives to True Positives rather than the total number of negative examples, captures the effect of the large number of negative examples on the algorithm's performance.

Consider two competing algorithms for classification on an imbalanced data set with 100 positive examples and 2000 negative examples. Figure 2.1 shows the performance of each algorithm in ROC and PR space.

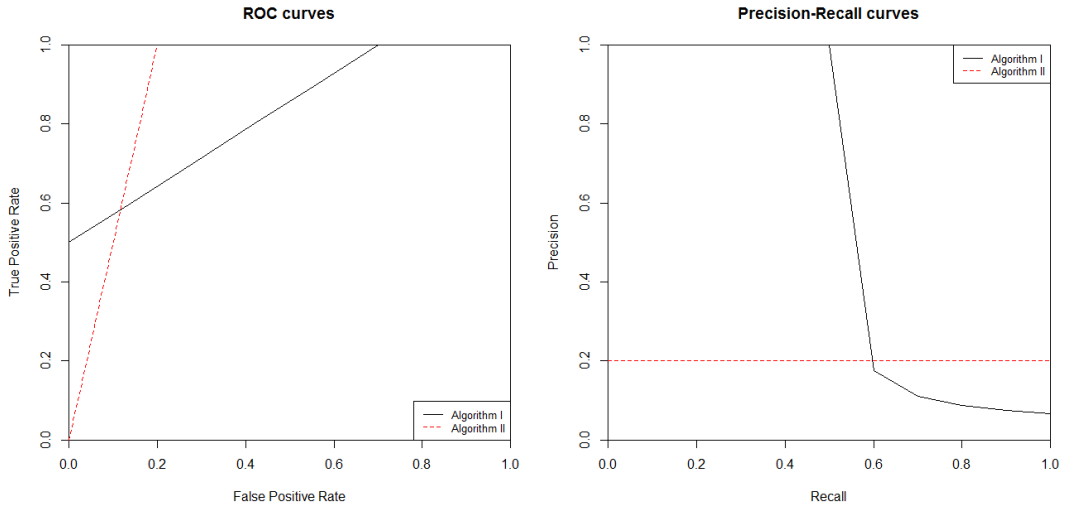

Figure 2.1: The difference between comparing algorithms in ROC vs PR space.

Based on the the area under the ROC curve, we would choose algorithm II because of its larger AUCROC (0.9) compared to algorithm I (0.825). However, algorithm II does

not perform well in terms of false discovery rate. The false discovery rate (FDR) is the proportion of false discovered positives among the total number of predicted positives, which is defined in Equation (2.7) based on Table 2.1.

$$\text{FDR} = \frac{FP}{TP + FP}. \tag{2.7}$$

Consider, for example, the first time each algorithm is able to select 50 out of the 100 positive examples, i.e. the first time each curve reaches a height of TPR=0.5 on the left panel of Figure 2.1. For algorithm II, this occurs at the point $(0.1, 0.5)$, i.e. 200 false positives and 50 true positives, algorithm II gives a high false discovery rate of $200/(200 + 50) = 0.8$. On the other hand, algorithm I can achieve the same level of true positive rate (0.5) without making any false positives, leading to zero false discovery rate. In practice, in most typical cases, we prefer algorithm I although it has a comparatively smaller AUCROC. The reason for this is the large amount of area that includes regions of high false positive rate. Unlike the ROC space, the PR space suggests to choose algorithm I based on its larger AUCPR (0.598) compared to algorithm II (0.2). Consequently, in terms of classifying the imbalanced data, ROC curves tend to present an overly optimistic view of an algorithm's performance, while Precison-Recall curves give a more appropriate picture.

2.3 Estimators and asymptotic properties

For $j = 0, 1$, we denote $\boldsymbol{\mu}_j = \text{E}(\mathbf{X}|Y = j)$ and $\Sigma_j = \text{Var}(\mathbf{X}|Y = j)$. We assume that $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, Σ_0 and Σ_1 exist, and at least one of $\{\Sigma_0, \Sigma_1\}$ is non-singular. The goal is to find the linear score function which maximizes the area under the curve, either AUCROC or AUCPR. Note that since both approaches consider the full range of thresholds, c, the vector $\boldsymbol{\beta}$ is only defined up to a scalar multiple, as the multiple can be absorbed into c. For identifiability, we fix the first component such that $|\beta_1| = 1$.

2.3.1 ROC curve

According to the binormal assumption and Equation (2.4),

$$\text{AUCROC} = \Phi\left(\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{\sqrt{\boldsymbol{\beta}^T(\Sigma_1 + \Sigma_0)\boldsymbol{\beta}}}\right). \tag{2.8}$$

Due to the monotonicity of the cumulative distribution function, $\Phi(\cdot)$, the maximizer of $\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{\sqrt{\boldsymbol{\beta}^T(\Sigma_1+\Sigma_0)\boldsymbol{\beta}}}$ will maximize the AUCROC. Finally, we may solve the equivalent problem $\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\beta}}{\boldsymbol{\beta}^T(\Sigma_1+\Sigma_0)\boldsymbol{\beta}}$ subject to $\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \geq 0$.

Proposition 1. *For symmetric matrix A, the solution to the problem $\max_{x\in\mathbb{R}^n} x^TAx$ subject to $x^Tx = 1$ will be the eigenvector for the largest eigenvalue of A, and the maximum value of x^TAx is the largest eigenvalue. In particular, any vector which is proportional to the eigenvector for the largest eigenvalue of A will maximize $\frac{x^TAx}{x^Tx}$.*

Let $\boldsymbol{\gamma} = (\Sigma_1 + \Sigma_0)^{1/2}\boldsymbol{\beta}$, then

$$\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\beta}}{\boldsymbol{\beta}^T(\Sigma_1 + \Sigma_0)\boldsymbol{\beta}} = \frac{\boldsymbol{\gamma}^T(\Sigma_1 + \Sigma_0)^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T(\Sigma_1 + \Sigma_0)^{-1/2}\boldsymbol{\gamma}}{\boldsymbol{\gamma}^T\boldsymbol{\gamma}}. \tag{2.9}$$

Note that, since both Σ_0 and Σ_1, are non-negative definite and at least one is non-singular, it follows that $\Sigma_1+\Sigma_0$ is positive definite. By Proposition 1, a solution to (2.9) will be proportional to the first eigenvector of $(\Sigma_1 + \Sigma_0)^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T(\Sigma_1 + \Sigma_0)^{-1/2}$ and thus, since this is a rank one matrix, it follows that $\boldsymbol{\gamma} \propto (\Sigma_1 + \Sigma_0)^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Consequently,

$$\boldsymbol{\beta} \propto (\Sigma_1 + \Sigma_0)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{2.10}$$

Let $\mathbb{S}_j = \{i : \mathbf{Y}_i = j\},\ j = \{0, 1\}$ be the index set to distinguish the positive and negative group, and let $n_j = |\mathbb{S}_j|$. The sample version of Equation (2.10) will be

$$\hat{\boldsymbol{\beta}} = (S_1 + S_0)^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0), \tag{2.11}$$

where S_1 and $\bar{\mathbf{X}}_1$ are the sample variance and mean for predictors in the positive group, S_0 and $\bar{\mathbf{X}}_0$ for the negative group.

2.3.2 Precision-Recall curve

Area under the Precision-Recall curve

According to the binormal assumption and Equation (2.6),

$$\text{AUCPR} = \int_0^1 \frac{\pi t}{\pi t + (1-\pi)\Phi\left(\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)}{\sqrt{\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}}} + \frac{\sqrt{\boldsymbol{\beta}^T\Sigma_1\boldsymbol{\beta}}\Phi^{-1}(t)}{\sqrt{\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}}}\right)} dt. \tag{2.12}$$

Since there is no closed form solution to Equation (2.12), we can solve it numerically via algorithms such as gradient descent, which is the approach that we implement in the examples.

In practice, $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1)$ is again replaced by its sample counterpart $(\bar{\mathbf{X}}_0, \bar{\mathbf{X}}_1, S_0, S_1)$.

Asymptotic distribution of the estimator

For simplicity of presentation, we consider the bivariate case here. For $j = 0, 1$, let $\mathrm{E}(\mathbf{X}|Y=j) = \boldsymbol{\mu}_j = (\mu_{j1}, \mu_{j2})^T$, and

$$\mathrm{Var}(\mathbf{X}|Y=j) = \Sigma_j = \begin{pmatrix} \sigma_{j1}^2 & c_j \\ c_j & \sigma_{j2}^2 \end{pmatrix},$$

where c_j denotes the covariance between X_1 and X_2 given $Y = j$. To create identifiability in $\boldsymbol{\beta}$, we fix $|\beta_1| = 1$. The goal is to estimate the linear score function in the form of $bX_1 + \beta X_2$ ($b \in \{-1, 1\}$) which maximizes the area under the Precision-Recall curve. The population version of the problem can be described as

$$\begin{aligned} (b_0, \beta_0) = \underset{b\in\{-1,1\},\beta\in\mathbb{R}}{\text{argmax}} \int_0^1 \frac{\pi t}{\pi t + (1-\pi)\Phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))} dt, \quad \text{s.t.} \\ \mu_p = b\mu_{11} + \beta\mu_{12}, \quad \mu_n = b\mu_{01} + \beta\mu_{02}, \\ \sigma_p^2 = \sigma_{11}^2 + \beta^2\sigma_{12}^2 + 2b\beta c_1, \quad \sigma_n^2 = \sigma_{01}^2 + \beta^2\sigma_{02}^2 + 2b\beta c_0. \end{aligned} \tag{2.13}$$

Proposition 2. *Under the binormal assumption, the solution, or population parameter, (b_0, β_0), that maximizes the area under the Precision-Recall curve satisfies*

$$\lambda_{10}\mu_{12} + \lambda_{20}\mu_{02} + 2\lambda_{30}\beta_0\sigma_{12}^2 + 2\lambda_{30}b_0c_1 + 2\lambda_{40}\beta_0\sigma_{02}^2 + 2\lambda_{40}b_0c_0 = 0, \tag{2.14}$$

where the functions, $\lambda_{j0} = \lambda_j(b_0, \beta_0, \mu_{p0}, \mu_{n0}, \sigma_{p0}, \sigma_{n0})$ *for* $j = 1, \cdots, 4$ *are given in the Appendix A, with* $\mu_{p0} = b_0\mu_{11} + \beta_0\mu_{12}$*, and* $\mu_{n0}, \sigma_{p0}, \sigma_{n0}$ *are defined similarly.*

The proof of proposition 2 is also given in the Appendix A.

Assume that $\mathbf{X}|Y = j$ follows a bivariate normal distribution with mean $\boldsymbol{\mu}_j$ and variance Σ_j for $j = 0, 1$. By applying Neyman-Pearson (NP) theory to binary hypothesis testing, we can reach an optimal risk score function by finding the most powerful test. Specifically, the null hypothesis is that the example belongs to the positive group, while the alternative hypothesis sets the example to the negative group. By NP theorem as well as the bivariate normal distribution setting, we will assign an example with observed $\mathbf{X}$ to the positive group if

$$\begin{aligned} &\frac{(2\pi)^{-1}|\Sigma_1|^{-1/2}\exp(-(\mathbf{X}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{X}-\boldsymbol{\mu}_1))}{(2\pi)^{-1}|\Sigma_0|^{-1/2}\exp(-(\mathbf{X}-\boldsymbol{\mu}_0)'\Sigma_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0))} > \eta_0 \\ \Longrightarrow\ &(\mathbf{X}-\boldsymbol{\mu}_0)'\Sigma_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0) - (\mathbf{X}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{X}-\boldsymbol{\mu}_1) > \eta_1 \\ \Longrightarrow\ &\mathbf{X}'(\Sigma_0^{-1} - \Sigma_1^{-1})\mathbf{X} + 2(\Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_0^{-1}\boldsymbol{\mu}_0)'\mathbf{X} > \eta, \end{aligned} \tag{2.15}$$

where η_0, η_1, and η are all constants.

If assuming equal covariance matrices, i.e. $\Sigma_0 = \Sigma_1 = \Sigma$, then the optimal score function, which is the left hand side of (2.15), will be linear with $\boldsymbol{\beta}^\star \propto \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. It can be shown that any $(b_0, \beta_0) \propto \boldsymbol{\beta}^\star$ is also the unique set of solutions to (A.4). Hence the sample version of (A.1) will be a consistent estimator for the optimal classifier. The estimator using the AUCROC, will also converge to the same quantity. Otherwise, for nonequal covariance matrix, the optimal score will be quadratic. In this case, or the case of non-normality, the true β_0 will satisfy Equation (A.4). In that situation, the maximizer for AUCROC will differ from that of AUCPR.

If the predictors are conditionally independent within each group, i.e. $c_1 = c_0 = 0$, then Equation (A.4) becomes

$$\lambda_{10}\mu_{12} + \lambda_{20}\mu_{02} + 2\lambda_{30}\beta_0\sigma_{12}^2 + 2\lambda_{40}\beta_0\sigma_{02}^2 = 0. \tag{2.16}$$

Consider a sample study with n examples, including n_1 positive ones with $\{X_{11i}, X_{12i}, Y = 1\}$ and n_0 negative ones with $\{X_{01i}, X_{02i}, Y = 0\}$. Each predictor is assumed conditionally independent within the group. In practice this may not hold. If the predictors are not conditionally independent within each group, with the assumption that the eigenvectors

of the covariance matrix in the positive group and that in the negative group are the same, we can simply transform to the principal components as our predictors to build the classifier.

Theorem 1. *Assume conditional independence among predictors, if* $\frac{n_1}{n} \to \pi$ *for* $0 < \pi < 1$, *then* $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$, *where* V *is defined in the Appendix B, and* β_0 *satisfies* (2.16).

The proof is given in the Appendix B.

2.4 Simulation

In this section, we present several simulation settings to evaluate the performance of the proposed procedure as well as the widely used ROC method. Recall that in the bivariate normal distribution case, Formula (2.15) implies a linear optimal score function with $\boldsymbol{\beta} \propto \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ if assuming equal covariance matrix ($\Sigma_0 = \Sigma_1$), otherwise a quadratic score function for non-equal covariance matrix. Simulation setting I and II will describe these two cases.

For each simulation setting, we apply two classification procedures, including ROC method and Precision-Recall (PR) method as introduced above. The false discovery rate (FDR), defined in Equation (2.7), is used as a measure of the classification effectiveness of each method. All the simulation models consider the bivariate classifier, whatever it is linear or quadratic.

2.4.1 Simulation Model I

The first configuration with the sample size $n = 400$ and proportion of positive examples $\pi = 20\%$ is to generate predictors according to $\mathbf{X}|Y = 1 \sim \text{BN}(1, 1, 2, 2, 0)$ and $\mathbf{X}|Y = 0 \sim \text{BN}(-1, -1, 2, 2, 0)$. By Formula (2.15), as a result of equal variance, the optimal score function is in the linear form of $X_1 + X_2$.

On the basis of 200 simulations, by maximizing the area under the Precision-Recall curve (AUCPR) in Equation (2.12), the linear risk score function is $X_1 + 1.0149X_2$. And maximizing area under the ROC curve (AUCROC) in Equation (2.8) gives $X_1 + 1.0193X_2$. The results are shown in Table 2.2, where the standard deviation for each coefficient estimate and its corresponding mean square error (MSE) is in the parenthesis.

Table 2.2: Estimates of the risk score function by PR and ROC methods under simulation Model I.

Method	β_0	MSE(β_0)	β_1	MSE(β_1)
PR	1 (0)	0 (0)	1.0149 (0.2142)	0.04588 (0.06388)
ROC	1 (0)	0 (0)	1.0193 (0.2475)	0.06132(0.1028)

We can see that the risk score functions obtained from either maximizing AUCPR or AUCROC are quite close to the optimal one, leading to the similarity in the false discovery rate by each approach. However, the smaller MSE as well as a smaller standard deviation for Precision-Recall one indicates that maximizing AUCPR is more efficient. Furthermore, we vary the proportion of positive examples to obtain the sample variance as well as the asymptotic variance of the Precision-Recall estimate through referring to Theorem 1. Table 2.3 indicates a close result between the two.

Table 2.3: Comparison between asymptotic and sample variance of PR estimate under simulation Model I.

Positive proportion	Asymptotic variance	Sample variance
0.1	35.8	36.3
0.2	23.5	22.4
0.3	13.7	15.0
0.4	13.0	14.0

2.4.2 Simulation Model II

The second configuration with the sample size $n = 400$ and proportion of positive examples $\pi = 20\%$ is to generate the predictors according to $\mathbf{X}|Y = 1 \sim \text{BN}(0, 0, 2, 0.5, 0)$ and $\mathbf{X}|Y = 0 \sim \text{BN}(-1, -1, 0.3, 1, 0)$. By the Neyman-Pearson lemma, because of non-equal variances, the UMP test suggests to use the quadratic score function $391X_1^2 - 108X_2^2 + 800X_1 + 72X_2$. However, in practice, linear classifiers are often used. Considering using the linear function to do the classification, based on 200 simulations, by maximizing AUCPR the linear risk score function is $X_1 + 0.44X_2$, while maximizing AUCROC gives

$X_1 + 3.59X_2$. Note that these solutions differ significantly. The angle between the two lines is 43°. Figure 2.2 shows the scatter plot of the example dataset with PR and ROC linear classifiers, which are just the direction instead of the particular lines. Here we display the classifiers with thresholds that pick out half of the true positives by each method. Estimation results are summarized in Table 2.4, where the standard deviation for each coefficient estimate is in the parenthesis. To compare the two classifiers, we consider the false discovery rate for each classifier as a function of the number of true positive discoveries. Figure 2.3 shows the false discovery rate by both PR and ROC methods.

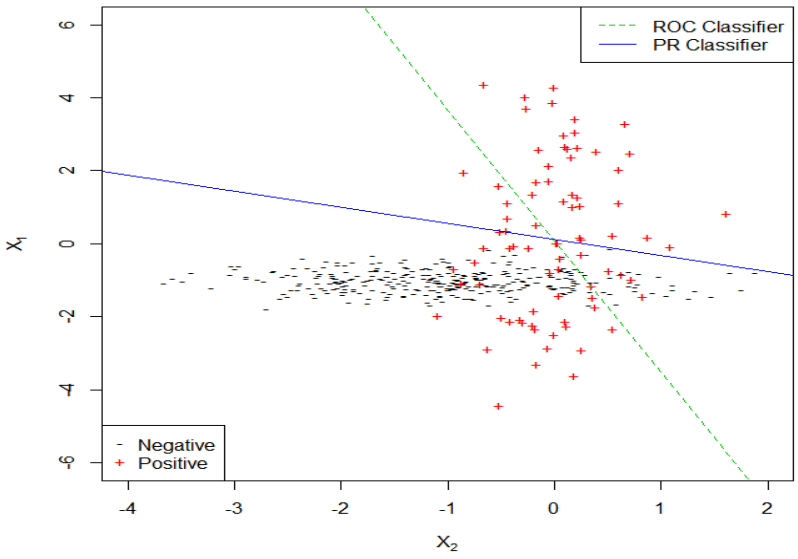

Figure 2.2: Scatter plot of the example dataset with PR and ROC linear classifiers under simulation Model II.

If the goal is to control the false discovery rate at a reasonable level, which is often the case, then the Precision-Recall approach stands out with a much smaller false discovery rate compared to the ROC, crossing when the false discovery rate reaches about 50% and approximately 75% of the true positives are found.

In fact, we can derive the population linear risk score function for each method by numerically solving Equation (2.12) and (2.8) using the true $\boldsymbol{\mu}$ and Σ. By optimization

Table 2.4: Estimates of the risk score function by PR and ROC methods under simulation Model II.

Curve	β_0	β_1
PR	1 (0)	0.4418 (0.1075)
ROC	1 (0)	3.5936 (1.5326)

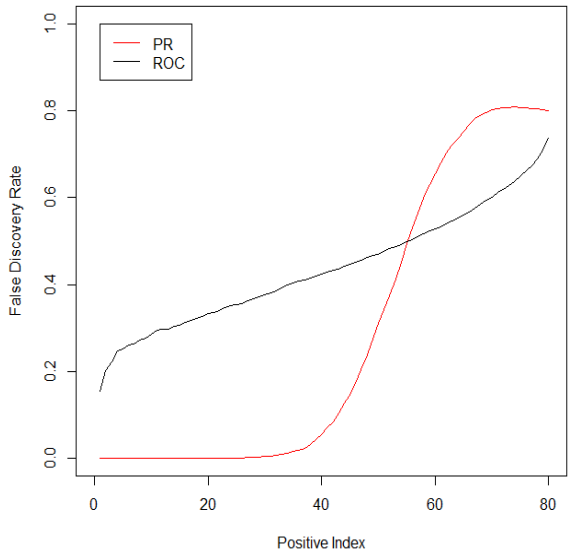

Figure 2.3: False discovery rate by PR and ROC methods under simulation Model II.

methods, the Precision-Recall approach suggests to use $X_1 + 0.43X_2$ while the ROC method proposes $X_1 + 3.27X_2$, similar to the results from the sample. Once we project down to the linear combination, since the optimal classifier is quadratic, we can consider constructing a quadratic classifier in the form of $(X_1 + 0.43X_2)^2 + a(X_1 + 0.43X_2)$ from the linear classifier suggested by the PR curve, and compute the maximum area under the PR curve and ROC curve respectively by optimizing over a. Similarly, we can also construct a quadratic classifier $(X_1 + 3.27X_2)^2 + b(X_1 + 3.27X_2)$ based on the ROC result, compute the maximum area under each curve by optimizing over b. After the tuning procedure, the optimal a and b are found to maximize the area under the curve with results summarized in Table 2.5.

Table 2.5: Comparison in the area under the PR and ROC curve among different classifiers.

Type	Classifier Function	AUCPR	AUCROC
NP Lemma	$Y = 391X_1^2 - 108X_2^2 + 800X_1 + 72X_2$	0.9193	0.9640
Precision Recall	$Y = (X_1 + 0.43X_2)^2 + a(X_1 + 0.43X_2)$	0.8101 ($a_{opt} = 3.07$)	0.8747 ($a_{opt} = 3.07$)
ROC	$Y = (X_1 + 3.27X_2)^2 + b(X_1 + 3.27X_2)$	0.5487 ($b_{opt} = 19.57$)	0.8466 ($b_{opt} = 20.28$)

It can be seen that the binormal Precision-Recall method outperforms binormal ROC method in terms of larger area under the curve, including both AUCPR and AUCROC. Although the optimal classifier is quadratic, if we want to simplify the classification problem by using linear classifier, it is suggested to use the risk score function generated by maximizing AUCPR, instead of AUCROC. By Precision-Recall method, not only we can better control the false discovery rate, but we can gain a larger area under the curve through constructing a quadratic classifier from a linear one as well.

2.4.3 Simulation Model III

Unlike the first two simulation models that come with bivariate normal distributions, the third configuration with the sample size $n = 400$ and proportion of positive examples $\pi = 20\%$ introduces mixture normal distributions. Hence, the binormal assumption is violated, as well as the linearity of the optimal rule. Specifically, we generate predictors according to $\mathbf{X}|Y = 0 \sim \text{BN}(0, 0, 1.2, 0.35, 0)$, and $X_1|Y = 1 \sim 0.95N(1, 1) + 0.05N(-1, 0.2)$ and

$X_2|Y = 1 \sim 0.95N(-0.8, 1.2) + 0.05N(1.5, 0.2)$. Figure 2.4 displays the false discovery rate by both PR and ROC methods, indicating a substantial improvement using the Precision-Recall approach over the ROC method in terms of the classification performance while controlling the false discovery rate at a low level.

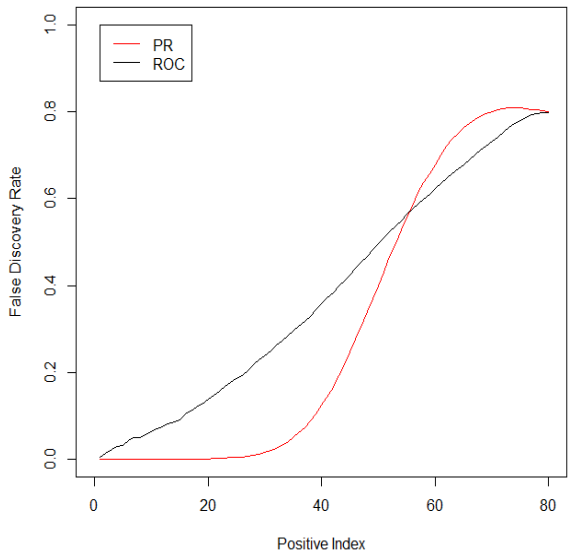

Figure 2.4: False discovery rate by PR and ROC methods under simulation Model III.

2.5 Real data analysis

To illustrate the classification performance of the binormal Precision-Recall approach, we analyze two real data sets, which are downloaded from the UC Irvine machine learning repository (Bache and Lichman, 2013) and transformed to become highly imbalanced according to Fan et al. (2014). The first one is the Abalone data from Nash (1994), aiming to predict the age of abalone from 7 continuous physical inputs. Following Fan et al. (2014), we create a highly imbalanced binary classification data set called Abalone9_18 by focusing only on those observations with class label "9" and "18", and regard class 18 as the minority class with proportion 5.7%. The second data set is Vehicle, first used in

Siebert (1987), involving classification of a given silhouette as one of four types of vehicles based on a set of 18 continuous features. In order to make the data imbalanced, we follow Fan et al. (2014) to choose class label "van" as the minority class with ratio 23.5%. The data is summarized in Table 3.1.

Table 2.6: Characteristics of datasets.

Dataset	Samples	Majority	Minority	Minority Proportion	Predictors
Abalone9_18	731	689	42	5.7%	7
Vehicle	846	647	199	23.5%	18

For each data set, we randomly split both the majority group and minority group into two parts evenly, forming a training dataset consisting of half of majority samples and half of minority samples, as well as a testing dataset with the other halves. Besides the binormal Precision-Recall method and binormal ROC method discussed above, we also implement the non-parametric approach to maximize AUCROC, which uses a sigmoid approximation without the binormal assumption (Ma and Huang, 2005). The three methods are conducted on the training data set to estimate their corresponding linear classifiers with performance measured on the testing data set based on 200 repetitions, where the relationship between average false discovery rate and true positive (minority) examples are shown in Figure 2.5.

For the Abalone9_18 data, the binormal Precision-Recall approach performs best among the three in terms of controlling false discovery rate at a lower level, up to around 20% in classification of correctly identifying 50% of the minority class. For the Vehicle data, each of the three methods does a good job because of the low false discovery rate, but comparatively the binormal Precision-Recall approach outperforms the binormal ROC over the full range, while outperforming the non-parametric approach when successfully finding 80% of the minority examples in the testing data set.

2.6 Discussions

In this chapter, we present an approach in a binormal framework, which is to estimate a linear classifier that maximizes the area under the Precision-Recall curve. The asymptotic

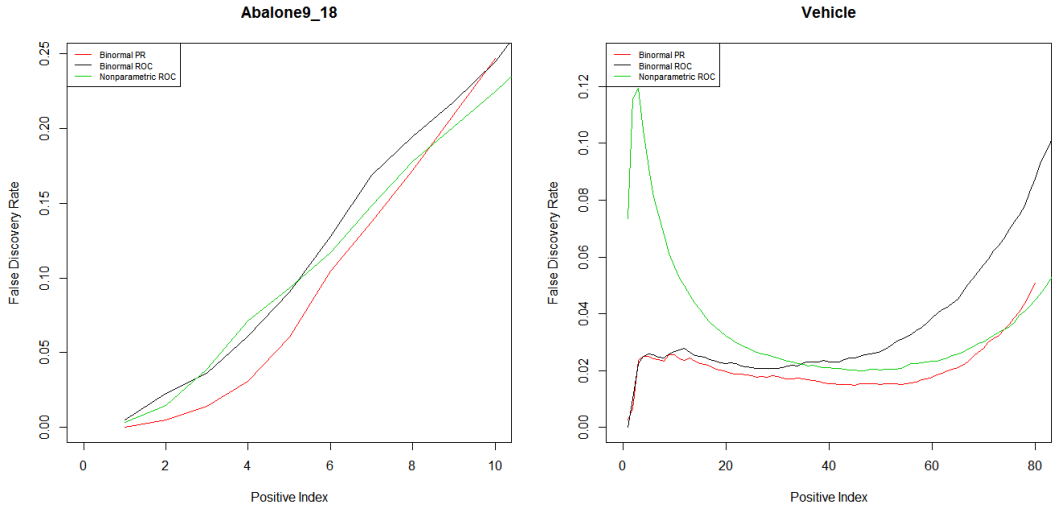

Figure 2.5: Relationship between average false discovery rate and true positive examples.

distribution of the estimate has been derived and compared with that in the binormal ROC approach. Simulation as well as real data results indicate that the binormal Precision-Recall approach has an improvement over the binormal ROC method and the TGDR algorithm in terms of controlling the false discovery rate at a lower level and a smaller asymptotic variance for the estimates in the classifier.

Like the regularized binormal ROC method (Ma et al., 2006), it may be reasonable to implement the threshold gradient descent regularization (TGDR) algorithm (Friedman and Popescu, 2003) in the binormal Precision-Recall framework if variable selection was the goal. Any other common distributional assumptions, such as gamma and exponential distribution, can also be an alternative. Furthermore, in front of irregular data, we can still build the proposed binormal Precision-Recall framework by applying some power transformations to the original data, like Box-Cox transformation (Box and Cox, 1964).

CHAPTER 3

Feature Selection for Imbalanced Data

3.1 Introduction

Feature or variable selection, whose purpose is to eliminate irrelevant variables to enhance the generalization performance of a given learning algorithm, is an important topic in machine learning, especially in high dimensional applications. In regression and classification problems with a large number of predictors, if only some of the variables contribute to the response, overfitting appears to be a critical concern for statistical analysis. Consequently, finding the optimal subset of variables among the pool is necessary and momentous.

Although there exists to be a lot of variable selection methods within the classification context, problem occurs when facing imbalanced data sets, i.e. there are many more examples from some classes than from others (Chawla et al., 2004). In this thesis we focus on variable selection for binary classification and class-imbalanced data means one of the classes (the negative) is significantly larger in terms of instances than the other (the positive). Furthermore, the cost of misclassifying an example from the positive class (minority group) is larger than misclassifying one from the negative (majority group) (He et al., 2009). The imbalanced problem is prevalent in many applications, including fraud and churn detection (Anil Kumar and Ravi, 2008), text categorization (Zheng et al., 2004), medical diagnosis (Van Hulse et al., 2009), detection of software defects (Park et al., 2013),

etc. Maldonado et al. (2014) designed a feature selection method for high-dimensional class-imbalanced data sets using SVM. Since the receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982) plays an important role in measuring the performance of a binary classifier system as its decision threshold is varied, Ma and Huang (2005) applied a sigmoid approximation to the empirical area under the ROC curve (AUCROC) as the objective function for classification and implemented the threshold gradient descent regularization (TGDR) algorithm (Friedman and Popescu, 2003) for estimation and variable selection in the linear risk score function. However, ROC curves tend to present an overly optimistic view of an algorithm's performance in terms of classification of highly imbalanced data (Davis and Goadrich, 2006). Precision-Recall (PR) curves (Raghavan et al., 1989), an alternative to ROC curves for tasks with imbalanced data, have become a basis for assessing classification methods. Consequently, Liu and Bondell (2016) proposed a classification algorithm with imbalanced data based on estimating the risk score function via maximization of the area under the Precision-Recall curve (AUCPR) in a binormal framework (Brodersen et al., 2010b).

In this chapter, we propose a regularized binormal Precision-Recall algorithm for variable selection in the classification context. It consists of two stages. The first stage is to compute the area under the Precision-Recall curve (AUCPR) in a binormal framework. With the binormal AUCPR criterion, we apply the threshold gradient descent regularization (TGDR) method for variable selection, which is the second stage. The proposed variable selection approach works well, especially when facing class-imbalanced data sets. We demonstrate via both simulations and real data analysis, that our method outperforms that based on the area under the ROC curve.

The chapter is organized as follows. Section is devoted to illustrating the proposed regularized binormal Precision-Recall algorithm. Two simulation settings are presented in section and real data analysis is discussed in section . Section gives discussions.

3.2 Regularized Binormal Precision-Recall Algorithm

For $j = 0, 1$, we denote $\boldsymbol{\mu}_j = \mathrm{E}(\mathbf{X}|Y = j)$ and $\Sigma_j = \mathrm{Var}(\mathbf{X}|Y = j)$. We assume that $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, Σ_0 and Σ_1 exist, and at least one of $\{\Sigma_0, \Sigma_1\}$ is non-singular. (Liu and Bondell, 2016) proved that in the binormal framework, the maximizer of AUCROC, defined in

Equation (2.4), is given by

$$\boldsymbol{\beta}_{ROC} \propto (\Sigma_1 + \Sigma_0)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{3.1}$$

However, in order to maximize AUCPR, defined in Equation (2.6), we solve it numerically via algorithms such as gradient descent because there is no closed-form solution. In practice, $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1)$ is replaced by its sample counterpart $(\bar{\mathbf{X}}_0, \bar{\mathbf{X}}_1, S_0, S_1)$.

3.2.1 Anchor Variable

Since both ROC and PR estimators consider the full range of thresholds, c, the vector $\boldsymbol{\beta}$ is only identifiable up to a scalar multiple. For identifiability, prior to analysis, we need to define the anchor variable, the one whose estimated coefficient will be set as a constant. This is the same as the anchor biomarker in the biology field. Ma et al. (2006) used an adjusted t-statistic, similar to a simple shrinkage method (Cui et al., 2005), to define the anchor biomarker. In this thesis work, we fix the anchor variable such that $|\beta_{anchor}| = 1$.

In the PR approach, we compute the AUCPR according to (2.6) by using each variable separately, and the anchor variable is defined as the one with maximum AUCPR. Similarly, in the ROC method, the anchor variable is defined as the one with maximum AUCROC in (2.4).

3.2.2 Threshold Gradient Descent Regularization (TGDR)

Friedman and Popescu (2003) proposed the threshold gradient descent regularization (TGDR) algorithm, establishing a parameter path in the high-dimensional coefficient space using the gradient descent method and afterwards identifying the best model along this parameter path. Let $R(\boldsymbol{\beta})$ denote the objective function that we want to maximize, $\boldsymbol{\beta}(v)$ denote the parameter path indexed by $v \in [0, \infty)$, Δv denote the infinitesimal positive increment as in ordinary gradient descent methods. For any threshold $\tau \in [0, 1]$, the TGDR algorithm is described as follows.

In the solution path, the anchor variable will always be the first one to be selected because we set its gradient at 0 in every iteration, and its magnitude is fixed at 1. The number of iterations K and threshold τ jointly determine the property of other estimated $\boldsymbol{\beta}$. When $\tau \approx 0$, $\hat{\boldsymbol{\beta}}$ is dense for all values of K. On the contrary, when $\tau \approx 1$, then $\hat{\boldsymbol{\beta}}$ will be sparse for small K and remain so for a relatively large number of iterations, but

Algorithm 1 TGDR

1: **procedure** TGDR
2: Initialize $\boldsymbol{\beta}(0) = \mathbf{0}$ and $v = 0$, except that $|\beta_{anchor}| = 1$.
3: **repeat**
4: Compute the negative gradient $g(v) = -\partial R(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}(v)$. Denote the i-th component of $g(v)$ as $g_i(v)$, and set the component for anchor variable at 0. If $\max_i\{|g_i(v)|\} = 0$, stop the iteration.
5: Compute the vector $f(v)$ with the i-th component $f_i(v) = I\{|g_i(v)| \geq \tau \cdot \max_i |g_i(v)|\}$, an indicator function.
6: Update $\boldsymbol{\beta}(v + \Delta v) = \boldsymbol{\beta}(v) - \Delta v \times g(v) \times f(v)$ and replace v by $v + \Delta v$, where the product of f and g is componentwise.
7: **until** K times, a large number to guarantee a full parameter path.
8: Output nonzero β's in $\boldsymbol{\beta}$.
9: **end procedure**

will become dense eventually. At the extreme case, i.e. $\tau = 1$, the TGDR algorithm usually increases in the direction of a single covariate in each iteration, which mimics the incremental forward stagewise strategy (Hastie et al., 2005). When τ is in the middle range, the characteristics of $\boldsymbol{\beta}$ are between those for $\tau = 0$ and $\tau = 1$. Friedman and Popescu (2003) used cross validation techniques to tune the parameters K and τ.

3.2.3 Regularized Binormal Precision-Recall Algorithm

The regularized binormal Precision-Recall algorithm is a combination of the area under the binormal Precision-Recall curve, defined in (2.6), and the threshold gradient descent regularization method described above. In other words, the regularized binormal Precision-Recall algorithm is the TGDR approach with binormal AUCPR as the objective function, i.e.

$$R_{PR}(\boldsymbol{\beta}) = \int_0^1 \frac{\pi t}{\pi t + (1-\pi)\Phi\left(\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}{\sqrt{\boldsymbol{\beta}^T \Sigma_0 \boldsymbol{\beta}}} + \frac{\sqrt{\boldsymbol{\beta}^T \Sigma_1 \boldsymbol{\beta}}\Phi^{-1}(t)}{\sqrt{\boldsymbol{\beta}^T \Sigma_0 \boldsymbol{\beta}}}\right)} dt. \tag{3.2}$$

And the negative gradient at $\boldsymbol{\beta}$ is given by

$$-\frac{\partial R_{PR}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \int_0^1 \frac{\pi(1-\pi)t\phi(T_1)(T_2+T_3)}{(\pi t+(1-\pi)\Phi(T_1))^2}dt, \tag{3.3}$$

$$\begin{aligned} where \quad T_1 &= \frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)}{\sqrt{\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}}} + \frac{\sqrt{\boldsymbol{\beta}^T\Sigma_1\boldsymbol{\beta}}\Phi^{-1}(t)}{\sqrt{\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}}}, \\ T_2 &= \frac{(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}-\boldsymbol{\beta}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\Sigma_0\boldsymbol{\beta}}{\left(\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}\right)^{3/2}}, \\ T_3 &= \Phi^{-1}(t)\frac{\Sigma_1\boldsymbol{\beta}\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}-\boldsymbol{\beta}^T\Sigma_1\boldsymbol{\beta}\Sigma_0\boldsymbol{\beta}}{\left(\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\beta}\right)^{3/2}\left(\boldsymbol{\beta}^T\Sigma_1\boldsymbol{\beta}\right)^{1/2}}. \end{aligned}$$

In conclusion, we apply the TGDR algorithm to maximize the AUCPR in (3.2) with the negative gradient function in (3.3), record the parameter solution path and select important variables.

Similarly, based on the binormal AUCROC in (2.4), the regularized binormal ROC algorithm applies the TGDR algorithm to maximize (3.4) with the negative gradient function in (3.5).

$$R_{ROC}(\boldsymbol{\beta}) = \Phi\left(\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{\sqrt{\boldsymbol{\beta}^T(\Sigma_1+\Sigma_0)\boldsymbol{\beta}}}\right). \tag{3.4}$$

$$-\frac{\partial R_{ROC}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \phi\left(\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{\sqrt{\boldsymbol{\beta}^T(\Sigma_1+\Sigma_0)\boldsymbol{\beta}}}\right)\frac{\boldsymbol{\beta}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)(\Sigma_1+\Sigma_0)\boldsymbol{\beta}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\boldsymbol{\beta}^T(\Sigma_1+\Sigma_0)\boldsymbol{\beta}}{\left(\boldsymbol{\beta}^T(\Sigma_1+\Sigma_0)\boldsymbol{\beta}\right)^{3/2}}. \tag{3.5}$$

3.3 Simulation

We present two simulation settings to evaluate the performance of the proposed PR-based variable selection algorithm as well as the ROC-based approach. For each simulation setting, the analysis consists of two stages. In the first stage, we apply two variable selection methods, including the regularized binormal PR algorithm and the regularized binormal ROC algorithm, on the simulated training data set. In the meantime, we record the solution path, i.e. the sequence of selected variables. During the application of TGDR

algorithm, we fix $\tau = 1$ to guarantee the sparse solution and $K = 5000$, which is large enough to produce a full parameter path for our settings. In the second stage, we follow the solution path and sequentially use a certain number of selected variables to implement the classification on a simulated large testing data set, and compare the performance between the two methods. The false discovery rate (FDR) in Equation (2.7) and the false positive rate (FPR) in Equation (2.1) are used as measures of the variable selection effectiveness of each method. The simulation settings consider the normal and non-normal situations respectively.

3.3.1 Simulation Model I

In the first configuration, we use variable size $p = 10$ and proportion of positive examples $\pi = 20\%$. Both the training data set with the sample size $n = 400$ and the testing data set with the sample size $n = 10000$ are to generate the first two predictors according to binormal distributions (BN), i.e.

$$(X_1, X_2)^T | Y = 0 \sim \text{BN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 \\ 0 & 0.1 \end{pmatrix}\right),$$

$$(X_1, X_2)^T | Y = 1 \sim \text{BN}\left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 0.4 & 0.9 \\ 0.9 & 2.5 \end{pmatrix}\right),$$

and other covariates $\mathbf{X} = (X_3, \cdots, X_{10})^T$ in both negative group (given $Y = 0$) and positive group (given $Y = 1$) are generated from a common multivariate normal distribution with mean vector $\mathbf{2}$ and compound symmetric covariance matrix Σ, where $\rho = \Sigma_{ij} = 0.4$, when $i \neq j$.

Based on the Neyman-Pearson theory (Neyman and Pearson, 1992), the optimal risk score function via the most powerful test in this setting only consists of the first two variables, X_1 and X_2, because of the common probability density function form for the other covariates. The training data set is generated repeatedly for 100 times, where we apply the regularized binormal PR algorithm and the regularized binormal ROC algorithm as well as record the solution path on each simulated training data. The PR-based algorithm always regards X_2 as the anchor variable and the second selected variable in the solution path is X_1 each time. On the other hand, the ROC-based method

always chooses X_1 as the anchor variable, while the second selected variable in the solution path varies.

Following the solution path generated by each algorithm, we consider using linear functions to do the classification on the testing data set with the sample size $n = 10000$. Start from using the anchor variable only, we sequentially add one variable in the existing classifier. To compare the two classifiers, we consider the false discovery rate for each classifier as a function of the proportion of true positive discoveries, as well as the false positive rate for each classifier as a function of the proportion of true positive discoveries. Figure 3.1 and 3.2 respectively show the false discovery rate and false positive rate by using the two algorithms.

In practice, when facing imbalanced data, the goal is often to control the false discovery rate or false positive rate at a reasonable level. Figure 3.1 tells us that the Precision-Recall algorithm stands out with a much smaller false discovery rate compared to the ROC, crossing when the false discovery rate reaches about 70% and approximately 60% of the true positives are found. In addition, Figure 3.2 shows that the PR has a smaller false positive rate before it reaches 35%. The trend is already obvious when only using one variable in the classifier. The performance is slightly improved when using two variables (X_2 and X_1) in the PR-based classifier, while more variables do not help any more.

In conclusion, the regularized binormal PR algorithm selects X_2 and X_1 as the important variables, which is consistent to the Neyman-Pearson theory, while the regularized binormal ROC algorithm fails in terms of variable selection in this setting. It is also suggested to follow the solution path generated by the regularized binormal PR algorithm to create classifiers in order to better control the false discovery rate and false positive rate in dealing with imbalanced data sets.

3.3.2 Simulation Model II

Unlike the first simulation model that comes with bivariate normal distributions for important variables, the second configuration with the same variable size $p = 10$ and proportion of positive examples $\pi = 20\%$ introduces mixture normal distributions. Specifically, we generate the training data set with sample size $n = 400$ and the testing data set with

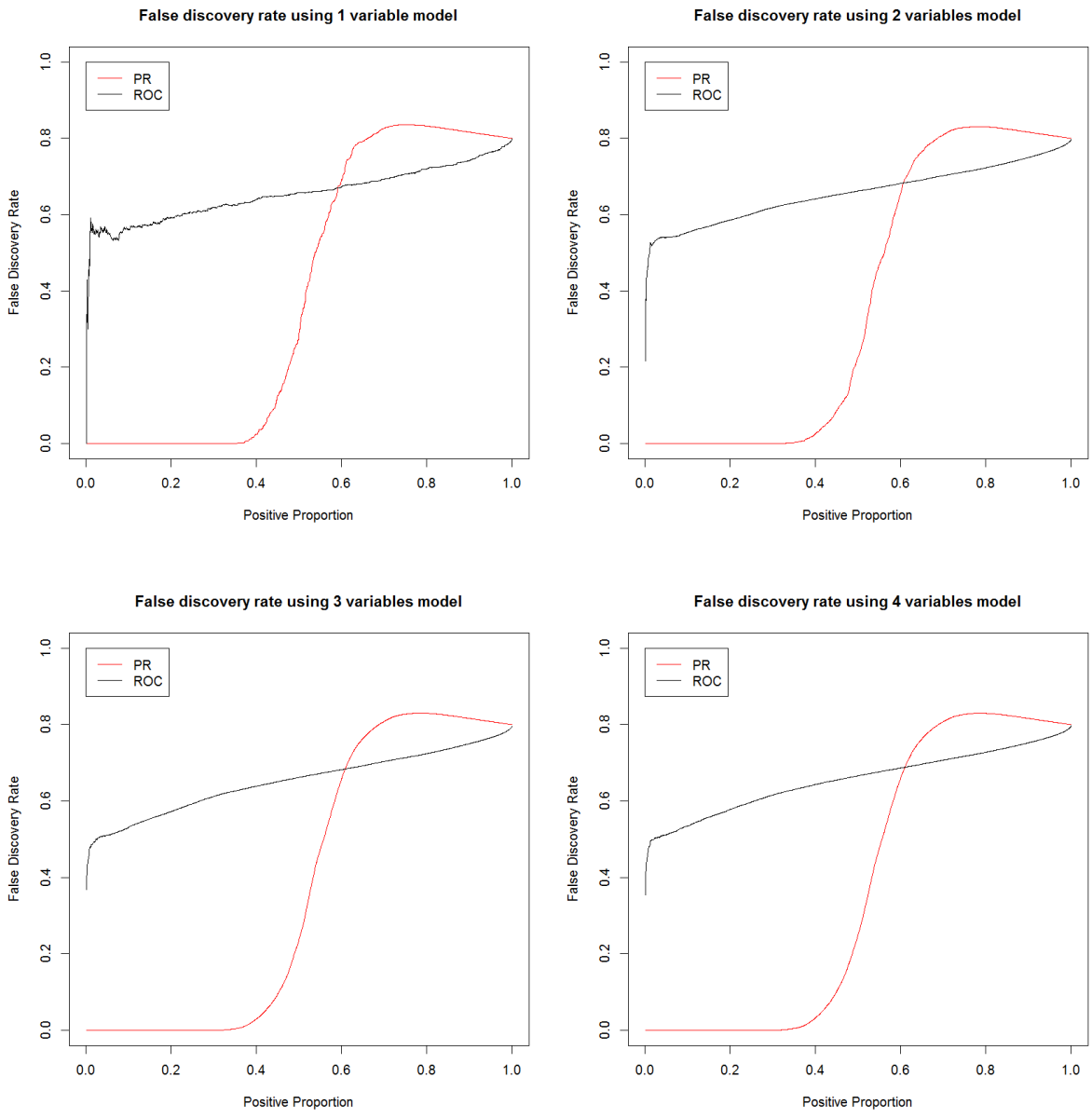

Figure 3.1: False discovery rate by PR and ROC methods under simulation Model I.

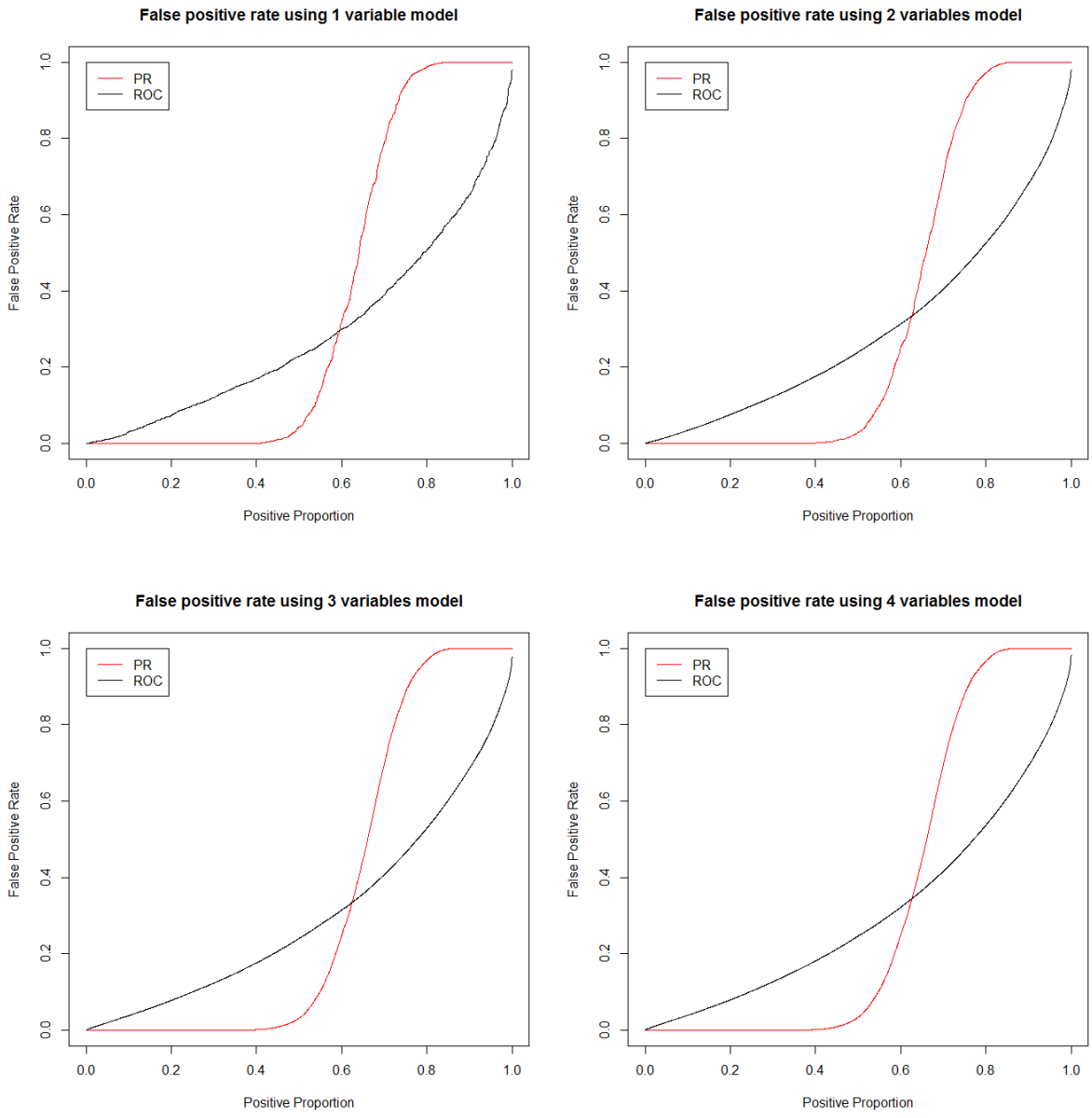

Figure 3.2: False positive rate by PR and ROC methods under simulation Model I.

sample size $n = 10000$ according to

$$
\begin{aligned}
&(X_1, X_2)^T | Y = 0 \sim \text{BN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.44 & 0 \\ 0 & 0.1225 \end{pmatrix}\right), \\
&X_1 | Y = 1 \sim 0.95N(1, 1) + 0.05N(-1, 0.04), \\
&X_2 | Y = 1 \sim 0.95N(-0.8, 1.44) + 0.05N(1.5, 0.04).
\end{aligned}
$$

and other covariates $\mathbf{X} = (X_3, \cdots, X_{10})^T$ in both negative group (given $Y = 0$) and positive group (given $Y = 1$) are generated from a common multivariate normal distribution with mean vector $\mathbf{2}$ and compound symmetric covariance matrix Σ, where $\rho = \Sigma_{ij} = 0.4$, when $i \neq j$.

Following the same two-stage procedure as in simulation model I, the regularized binormal PR algorithm picks X_2 as the anchor variable and X_1 as the second important variable in each repetition, and it also appears to be (X_2, X_1) or (X_1, X_2) in the solution path of the regularized binormal ROC algorithm. Although both algorithms catch the right important variables (X_2 and X_1) during the early stage of the solution path, consistent to the Neyman-Pearson result, the classification performances on the testing data set by classifiers of the two algorithms differ a lot in terms of false discovery rate and false positive rate, shown in Figure 3.3 and 3.4.

Figure 3.3 and 3.4 indicate a substantial improvement using the regularized binormal PR algorithm over the ROC-based method in terms of the classification performance while controlling the false discovery rate as well as the false positive rate at a low level. The Precision-Recall algorithm produces a smaller false discovery rate than the ROC algorithm when it reaches around 70%, and a smaller false positive rate when it reaches nearly 40%. The difference between performances of two algorithms has been significant since using only one variable in the classifier. Furthermore, classifiers with two variables are better than one variable model for both methods, but this difference remains constant since three variables.

To sum up, in this non-normal setting, although both the regularized binormal PR and ROC algorithm select the same important variables, X_2 and X_1, the PR algorithm proposes a better classifier, controlling the false discovery rate and false positive rate well when working with imbalanced data.

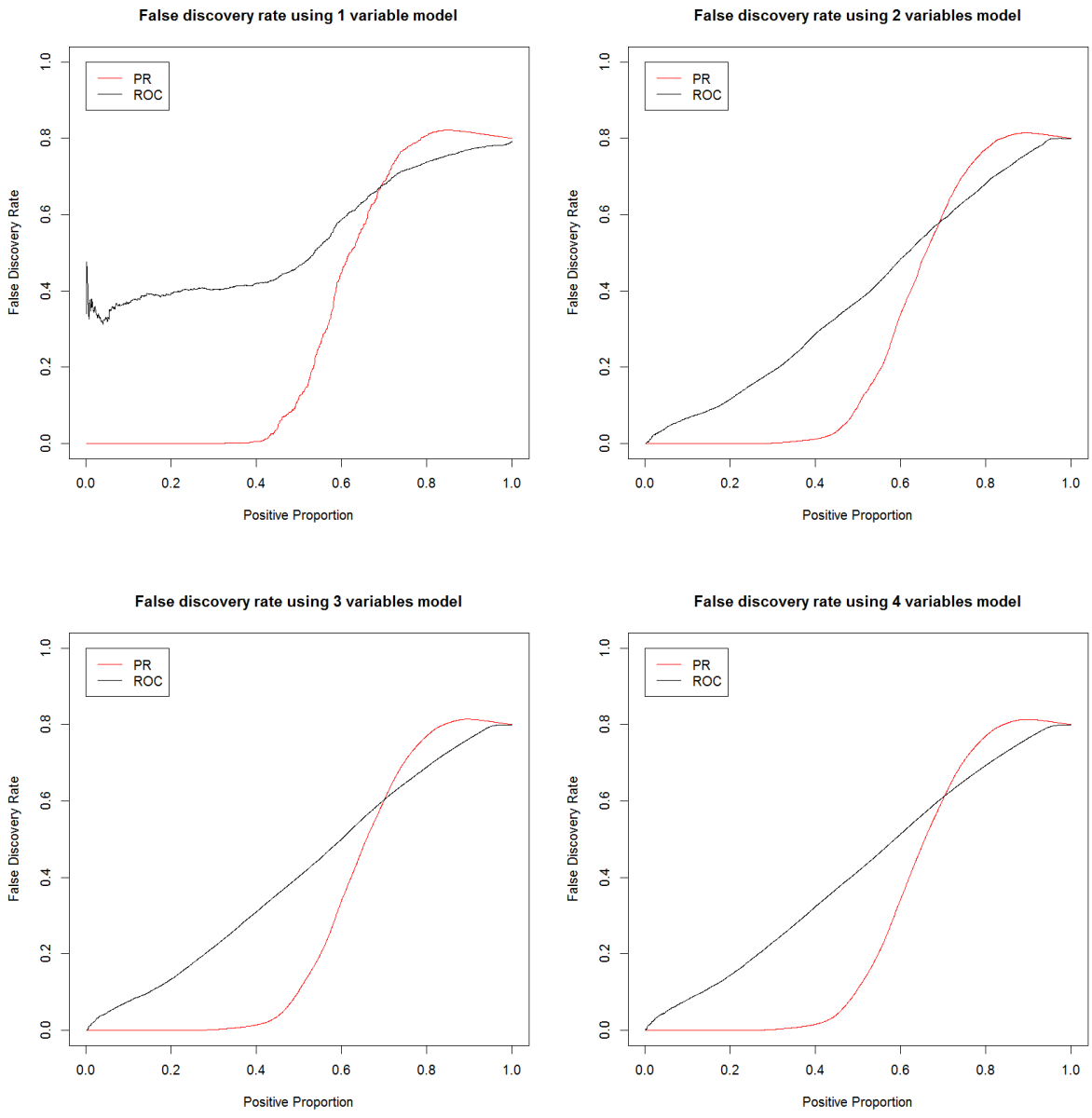

Figure 3.3: False discovery rate by PR and ROC methods under simulation Model II.

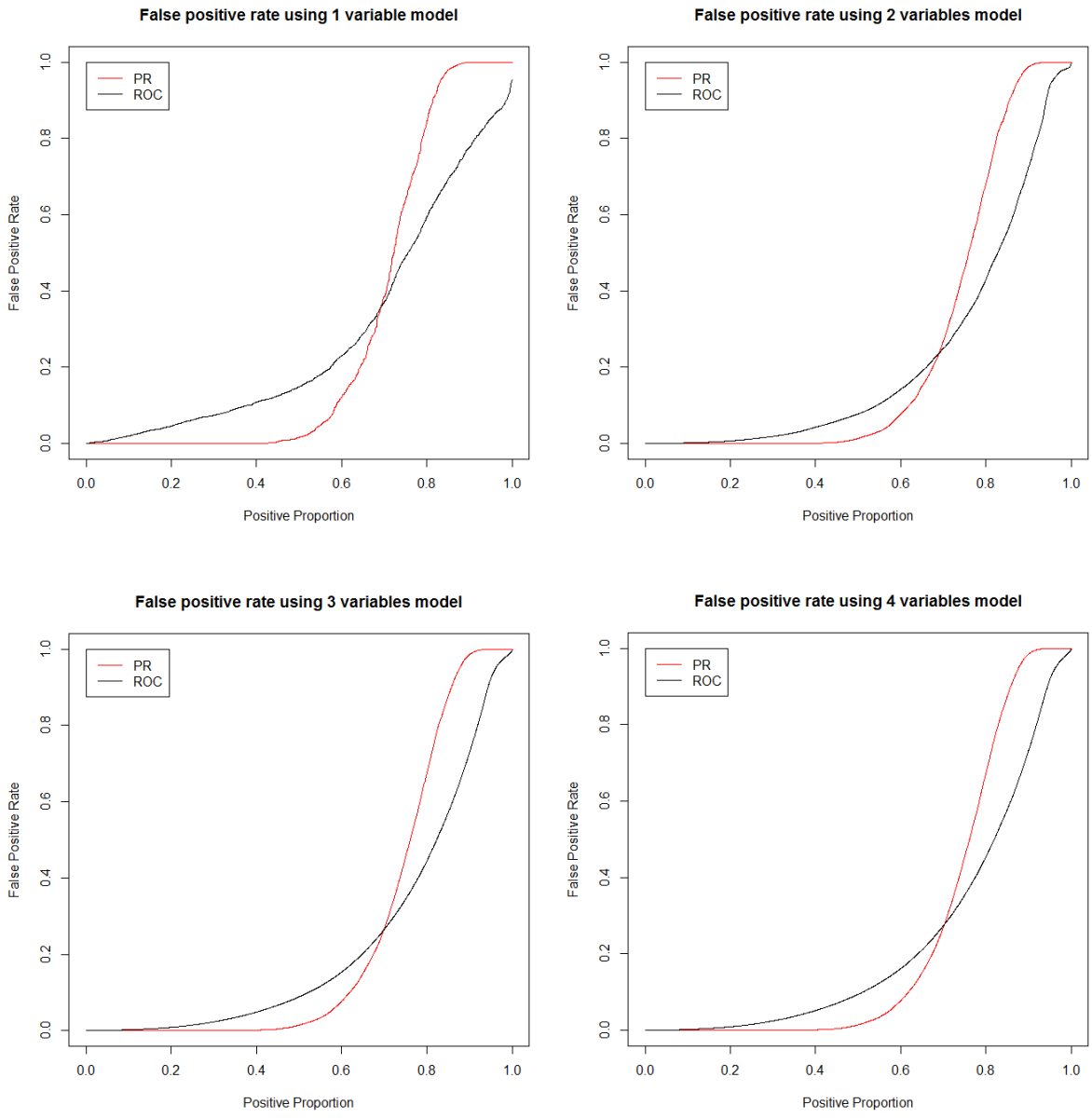

Figure 3.4: False positive rate by PR and ROC methods under simulation Model II.

3.4 Real Data Analysis

To illustrate the variable selection performance of the regularized binormal Precision-Recall algorithm, we analyze the Abalone real data set (Nash, 1994), aiming to predict the age of abalone from 7 continuous physical inputs. The data is from the UC Irvine machine learning repository (Bache and Lichman, 2013), and according to Fan et al. (2014), we transformed it to a highly imbalanced binary classification data set called Abalone9_18 by focusing only on those observations with class label "9" (young group) and "18" (old group), regarding class 18 as the minority class with proportion 5.7%. The data is summarized in Table 3.1.

Table 3.1: Characteristics of Abalone9_18 data set.

Data	Samples	Majority	Minority	Minority Proportion	Predictors
Abalone9_18	731	689	42	5.7%	7

We randomly split both the majority group and minority group in the data set into two parts evenly, forming a training dataset (365 samples) consisting of half of majority samples (344 negatives) and half of minority samples (21 positives), as well as a testing dataset (366 samples) with the other halves (345 negatives and 21 positives). Similar to the two-stage procedure in the simulation study, we apply the proposed regularized binormal PR algorithm and the regularized binormal ROC algorithm on the training data set to obtain the solution path with threshold $\tau = 1$ and iterations $K = 10,000$. Afterwards we sequentially use the first 4 variables in the solution path to implement the classification on the testing data set, and compare the performance between the two methods. Figure 3.5 and 3.6 respectively show the false discovery rate (FDR) and false positive rate (FPR) as a function of positive proportion by using the two algorithms based on 200 repetitions.

Figure 3.5 and 3.6 show an obvious preference towards the regularized binormal PR algorithm over the ROC-based method in terms of a smaller false discovery rate and false positive rate. For both methods, the classification performance is improved with the increase in the number of variables, while the improvement becomes small since the four variable model. So it suggests to choose the first three variables in the solution path for

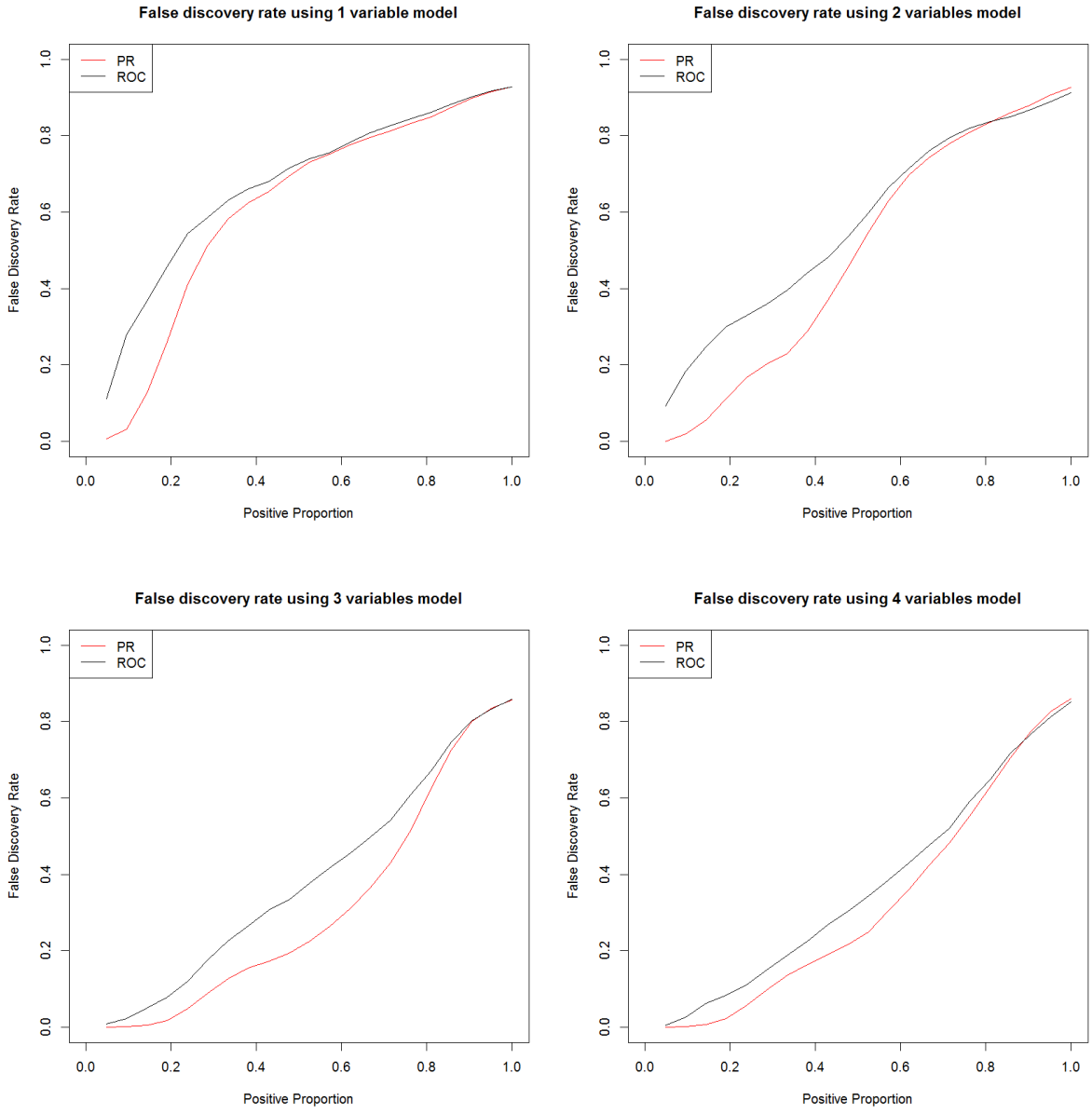

Figure 3.5: False discovery rate by regularized binormal PR and ROC methods under Abalone9_18 real data.

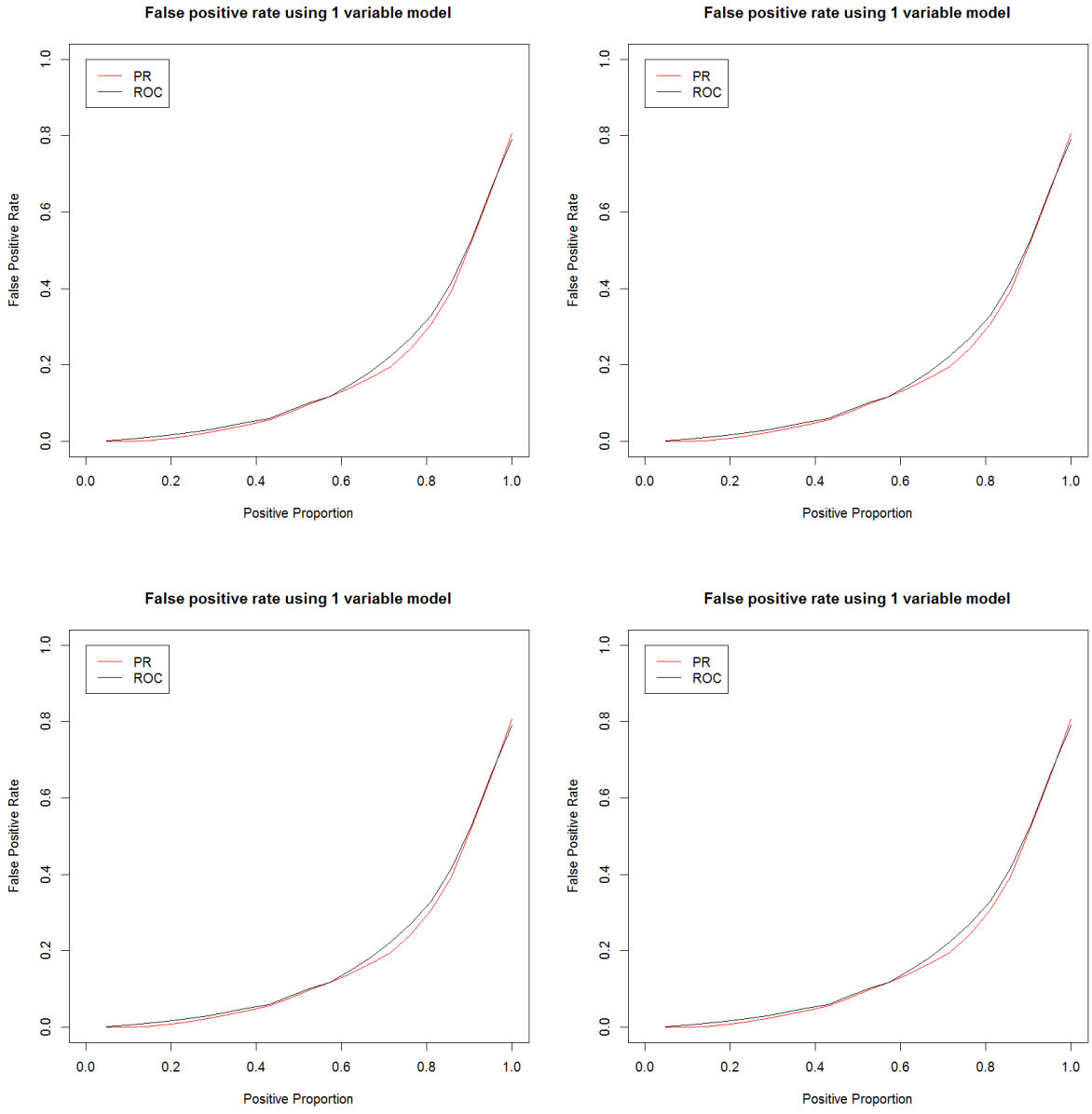

Figure 3.6: False positive rate by regularized binormal PR and ROC methods under Abalone9_18 real data.

classification. No matter how many variables are used for classification, the regularized binormal PR algorithm significantly outperforms the ROC algorithm almost over the full range. Figure 3.7 and 3.8 show 95% confidence intervals for the mean difference of FDR and FPR between the two methods separately. The mean difference, drawn in red solid lines, is negative almost over the full range. In the meantime, the 95% confidence intervals, drawn in black dashed lines, does not contain 0, further indicating the significance of better performances using the proposed regularized binormal PR algorithm.

3.5 Discussions

It is of practical importance to develop an efficient feature selection approach for class-imbalanced data. In this chapter, we propose a regularized binormal Precision-Recall algorithm for variable selection in the classification context, which applies the threshold gradient descent regularization (TGDR) method to maximize the area under the Precision-Recall curve (AUCPR) in a binormal framework. Simulation as well as real data results indicate that the regularized binormal Precision-Recall algorithm has an improvement over the regularized binormal ROC method when facing imbalanced data in terms of controlling the false discovery rate and false positive rate at a lower level.

Variable selection and classification using the regularized binormal PR algorithm for two-class settings can be extended to multi-class problems. Another interesting extension would be to compare the performance of the proposed approach to other variable selection methods, such as random forest (Breiman, 2001) and GUIDE (Loh, 2012), in dealing with curse of dimensionality problem (Donoho et al., 2000).

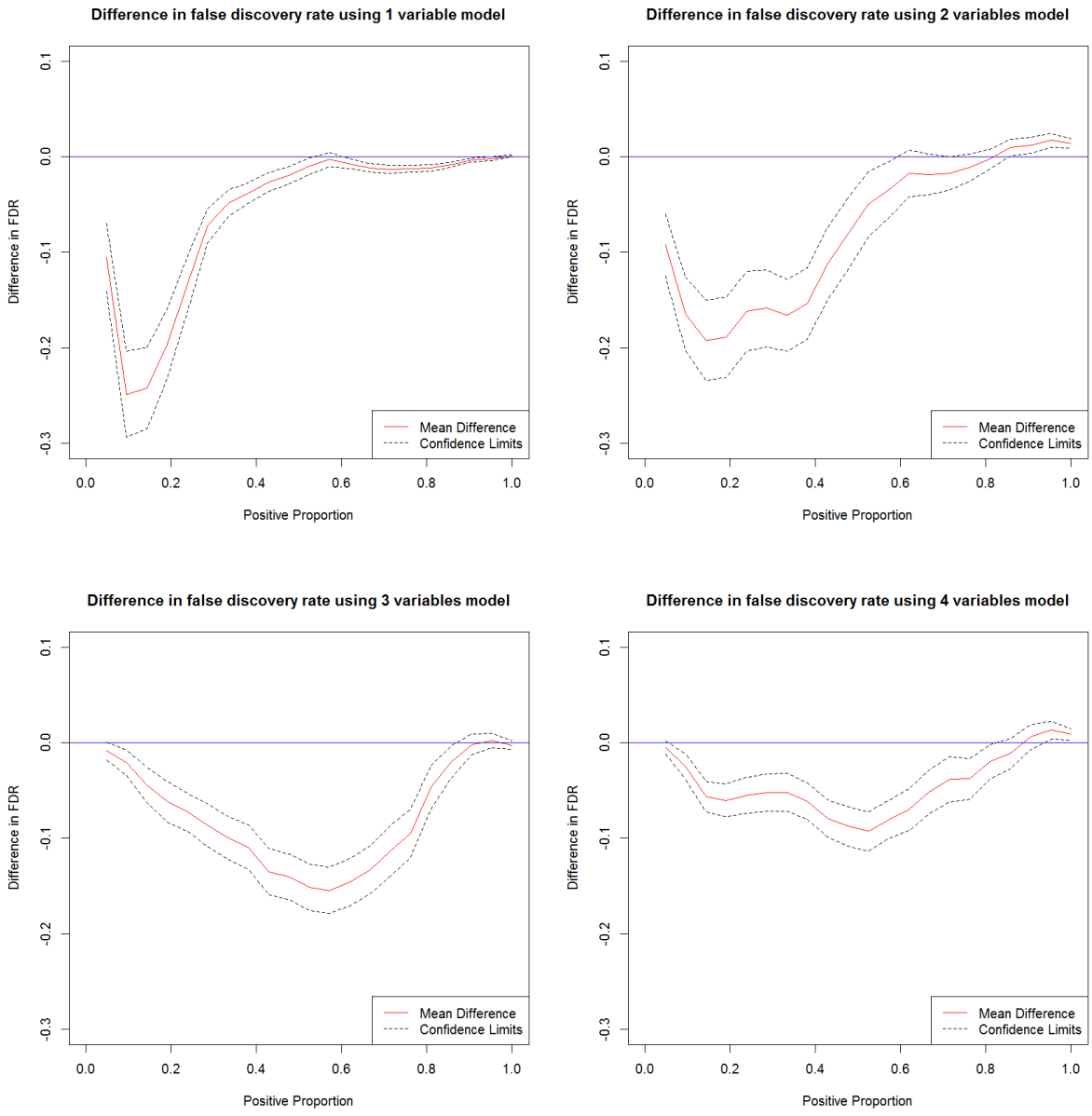

Figure 3.7: False discovery rate difference between PR and ROC methods under Abalone9_18 real data.

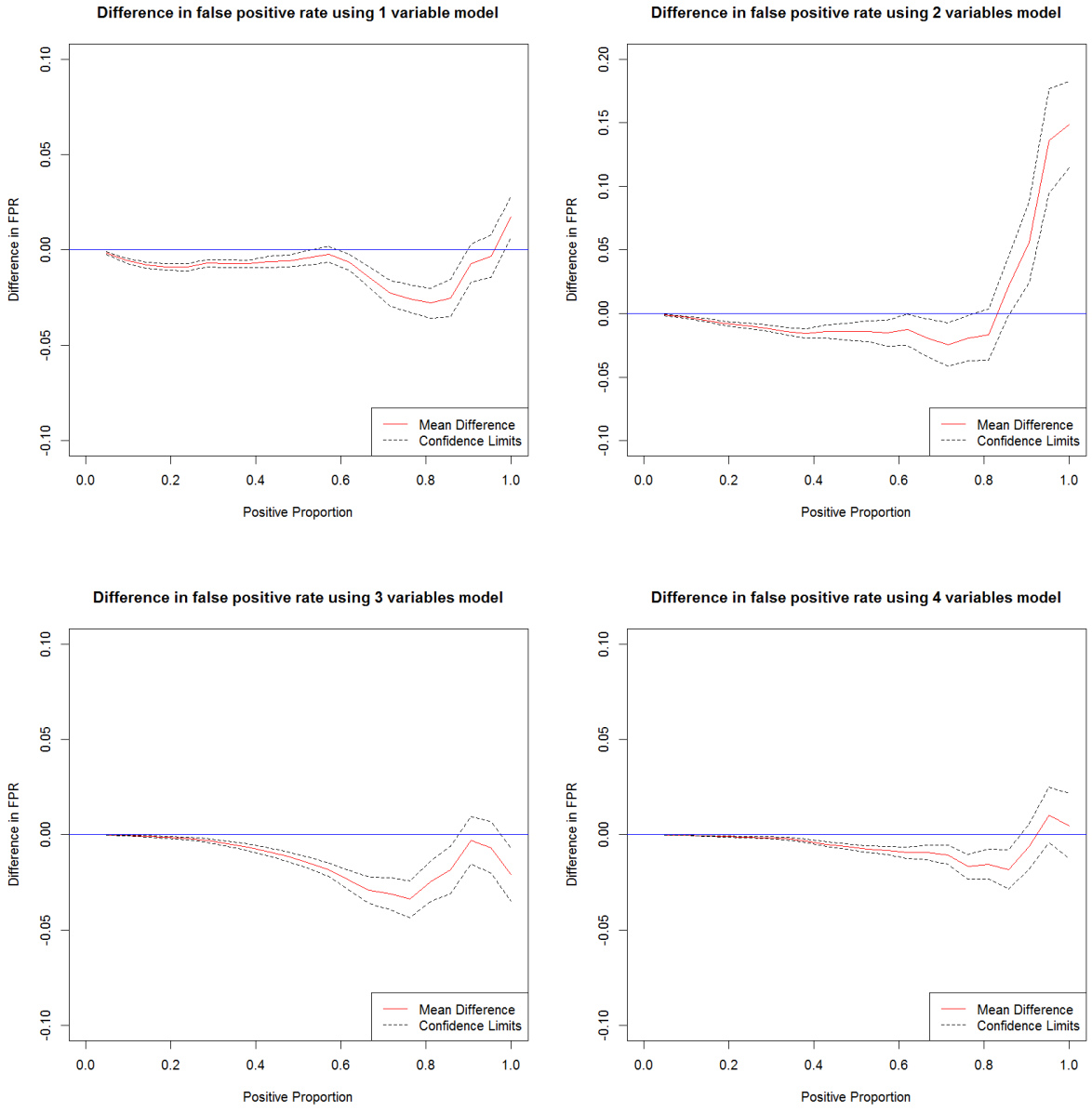

Figure 3.8: False positive rate difference between PR and ROC methods under Abalone9_-18 real data.

CHAPTER 4

Variable Screening in Ultrahigh Dimensions

4.1 Introduction

We consider the problem of ultrahigh dimensional regression, i.e. the dimension of predictors used for predicting a response of interest, p, is much larger than sample size, n. It is often assumed that only a relatively small subset of the predictors contribute to the response. As a result, an efficient method of variable selection, which can be able to identify the most important predictors, plays an key role in the ultra-high dimensional regression.

One group of variable selection methods are based on penalized methods which can select variables and estimates parameters simultaneously through solving an ultrahigh dimensional regression with some pre-specified penalties leading to sparsity. These methods include bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Dantzig selection (Candes and Tao, 2007), and other folded concave regularization methods (Fan and Lv, 2011; Zhang and Zhang, 2012). When the dimension is very high, however, these methods may have heavy implementation cost and face challenges in computational feasibility.

Recently, variable screening methods have been re-discovered and advocated in the ultra-high dimensional setting, including sure independence screening (SIS) method (Fan

and Lv, 2008), marginal bridge regression based method (Huang et al., 2008) and some others. Specifically, SIS method in (Fan and Lv, 2008) selects important variables in an ultrahigh dimensional linear models based on the marginal correlations of each predictor with the response. They showed that the correlation ranking of the predictors possesses a sure independence screening property, that is, the important variables can be selected with probability close to one. Later, the marginal screening method was extended to generalized linear models (Fan and Song, 2010). Various screening methods have been developed, to name a few, tilting methods (Hall et al., 2009), generalized correlation screening (Hall and Miller, 2009), nonparametric screening (Fan et al., 2011), robust rank correlation based screening (Li et al., 2012), and quantile-adaptive model-free feature screening (He et al., 2013).

These marginal screening methods face a number of challenges. For example, if the marginal working model is too far away from the true model, it is hard to ensure the sufficient conditions for sure screening to hold. Consequently marginally unimportant but jointly important variables may not be preserved in marginal screening. Meanwhile, the marginal screening methods may include noise variables that are weakly correlated with the important predictors. It can potentially increase false positive rate.

To address these issues, in this chapter, we propose a principal component-adjusted screening (PCAS) method for generalized linear models. The key idea is to use principal components as surrogate covariates to account for omitted covariates in marginal screening. Specifically, we fit p marginal regressions by maximizing the marginal likelihood including not only the screened predictor but also some selected principal components. Then we consider an independence learning by ranking the maximum marginal likelihood estimators or maximum marginal likelihood.

PCAS method has several advantages. First, PCAS retains top principal components as surrogate covariates, thus retains the information in those predictors that are not included in the marginal screening. Second, it possesses good properties of the conditional screening to reduce the correlation among predictors and thus reduce the noise in the process of variable selection. Finally, unlike the conditional sure independence screening method (Barut et al., 2012) where certain variables are known to be responsible for the outcomes, PCAS does not need these prior information of the predictors. Extensive numerical results show that the proposed PCAS method has superior performance to the original SIS method. As an important remark, computing the principal components in the

implementation only requires eigenvalue-decomposition of an n by n matrix regardless of the dimensionality p.

The setup of generalized linear models is introduced in section 4.2. Section 4.3 discussed the computation of principal components. In section 4.4, we introduced the PCAS procedure with maximum marginal likelihood estimators (MMLE) and marginal likelihood ratio (MLR). Simulation results are presented in section 4.5 and two real data analysis results are illustrated in section 4.6. Section 4.7 gives concluding remarks.

4.2 Generalized linear models

Consider the generalized linear model where the probability density function of a response variable Y takes the form $f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}$, with known functions $b(\cdot)$, $c(\cdot)$, and the natural parameter θ. Suppose that the observed data $\{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$ are identically independent distributed copies of $(\mathbf{X}, Y)$, where $\mathbf{X}_i = (1, X_{i1}, \cdots, X_{ip})^T$ and $X_{i1}, \cdots, X_{ip}$ are p-dimensional covariates for subject i. $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ is a $(p+1)$-vector of parameter. We are interested in identifying the sparsity structure of $\boldsymbol{\beta}$ from the equation

$$E(Y|\mathbf{X} = \mathbf{x}) = b'(\theta(\mathbf{x})) = g^{-1}(\sum_{j=0}^{p} \beta_j x_j), \tag{4.1}$$

where $\mathbf{x} = \{x_0, x_1, \cdots, x_p\}^T$ is a $(p+1)$-vector with $x_0 = 1$ when considering the intercept, $b'(\theta)$ is the first order derivative of $b(\theta)$ with respect to θ and g is the link function. For demonstration purpose, in the thesis work we only take canonical link function, that is $g = (b')^{-1}$, into consideration. In this case, $\theta(x) = \sum_{j=0}^{p} \beta_j x_j$. The ordinary linear model $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$, where ε is the random error, is a special case of model (4.1) by using the identity link, i.e. $g(\mu) = \mu$. Considering binary response data, the logistic regression is another special case of model (4.1) by using the logit link $g(\mu) = \log(\mu/(1-\mu))$.

4.3 Principal component analysis

Principal component analysis is a widely used tool for high dimensional data analysis in many fields, such as signal processing and dimension reduction. Based on projecting a

dataset to another coordinate system by determining the eigenvectors and eigenvalues of the matrix, principal component analysis involves calculations of a covariance matrix of a dataset to minimize the redundancy as well as maximize the variance (Shlens, 2014). A common method to find the eigenvectors and eigenvalues is singular value decomposition (SVD), which decomposes a matrix into a set of rotation and scale matrices. Suppose $\mathbf{X}$ is a matrix with n rows and p columns and columns are normalized to be norm one. A singular value decomposition of $\mathbf{X}$ is given by $\mathbf{X}_{n\times p} = \mathbf{U}_{n\times n}(\text{diag}(\lambda_1, \ldots, \lambda_n), \mathbf{0}_{n\times(p-n)})\mathbf{V}^T_{p\times p}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices with dimensions n and p respectively and $\text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_n$. Additionally, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. Since

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\text{diag}(\lambda_1^2, \ldots, \lambda_n^2, 0, \ldots, 0)\mathbf{V}^T,$$

it is clear that the columns of $\mathbf{V}$ are the principal directions of $\mathbf{X}$. Thus, the principal components, that is, the projection of $\mathbf{X}$'s rows on these directions, should be $\mathbf{XV} = \mathbf{U}_{n\times n}\text{diag}(\lambda_1, \ldots, \lambda_n)$. In other words, each column of $\mathbf{U}$ represents each principal component up to some scale.

To calculate $\mathbf{U}$, we note $\mathbf{XX}^T = \mathbf{U}\text{diag}(\lambda_1^2, \ldots, \lambda_n^2)\mathbf{U}^T$. Therefore, if we perform an eigenvalue decomposition on $\mathbf{XX}^T$, which is a matrix with much smaller dimensions when p is much larger than n, then $\mathbf{U}'s$ columns consist of all eigenvectors.

4.4 PCAS procedure

Let $\mathcal{M}_\star = \{1 \leq j \leq p : \beta_j^\star \neq 0\}$ be the true sparse model with non-sparsity size $s = |\mathcal{M}_\star|$, where $\boldsymbol{\beta}^\star = (\beta_0^\star, \beta_1^\star, \ldots, \beta_p^\star)^T$ denotes the true value. In this thesis work, we refer to principal components adjusted models as fitting models with componentwise covariates and the first K_n principal components as offset covariates.

4.4.1 PCAS with maximum marginal likelihood estimators

PCAS maximum marginal likelihood estimators (PCAS-MMLE) $\hat{\beta}_j^M$, for $j = 1, \ldots, p$, is defined as the minimizer of the negative marginal log-likelihood

$$(\hat{\beta}_{j,0}^M, \hat{\beta}_j^M, \hat{\gamma}_{j,1}^M, \ldots, \hat{\gamma}_{j,K_n}^M)^T = \underset{\beta_0, \beta_j, \gamma_{j,1}, \cdots, \gamma_{j,K_n}}{\operatorname{argmin}} \mathbb{P}_n l(\beta_0 + \beta_j X_j + \sum_{k=1}^{K_n} \gamma_{j,k}^M U_k, Y), \quad \text{for } j = 1, \ldots, p,$$

where $l(Y; \theta) = -(\theta Y - b(\theta) + c(Y))$, $\{U_k\}$ is the kth eigenvector consisting of $\{U_{ik}\}_{i=1}^n$, and $\mathbb{P}_n f(X, Y) = n^{-1} \sum_{i=1}^n f(X_i, Y_i)$ is the empirical measure. $\hat{\beta}_j^M$ measures the strength of the conditional contribution of X_j given the first K_n principal components. These principal components represent the information of predictors except for X_j in the marginal model. The process can be rapidly computed.

Specifically, in ordinary linear models with normality assumption of random errors, the maximum likelihood estimator is identical to the ordinary least squares estimator written as

$$(\hat{\beta}_{j,0}^M, \hat{\beta}_j^M, \hat{\gamma}_{j,1}^M, \ldots, \hat{\gamma}_{j,K_n}^M)^T = \underset{\beta_0, \beta_j, \gamma_{j,1}, \cdots, \gamma_{j,K_n}}{\operatorname{argmin}} \mathbb{P}_n (Y - \beta_0 - \beta_j X_j - \sum_{k=1}^{K_n} \gamma_{j,k}^M U_k)^2, \quad \text{for } j = 1, \ldots, p.$$

We select a set of variables

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq j \leq p : |\hat{\beta}_j^M| \geq \gamma_n\}, \tag{4.2}$$

where γ_n is a given threshold value. By ranking the importance of features according to their magnitude of marginal regression coefficients adjusted for a proportion of principal components, we retain variables with large conditional contribution given these principal components. Such an independence learning helps to decrease the dimension of the parameter space from p (possibly hundreds of thousands) to a much smaller number by choosing a large γ_n, leading to a more feasible computation. Although interpretations and implications of principal components adjusted models are still biased from the joint model, the non-sparse information about the joint model can be passed along to the marginal model under a mild condition. Hence it is suitable for the purpose of variable screening.

Since the rationale to use the principal components as surrogate covariates is to account for the effect of the omitted covariates in the marginal model, we should compute the principal components based on the $p-1$ omitted covariates for each marginal regression. For simplicity of computation, we compute the principal components based on all p covariates and use these principal components as surrogate covariates. Based on our observations, the numerical performance of two methods are very close while the latter one has significantly smaller computational costs.

4.4.2 PCAS with marginal likelihood ratio

As an alternative method, we can also rank variables based on the likelihood reduction of the variable X_j given the first K_n principal components, which we call PCAS with maximum likelihood ratio (PCAS-MLR):

$$L_{j,n} = \mathbb{P}_n\{l(\hat{\beta}_j^M X_j + \sum_{k=1}^{K_n} \hat{\gamma}_{j,k}^M U_k, Y)\} - \mathbb{P}_n\{l(\hat{\beta}_0^M, Y)\}, \text{ for } j = 1, \ldots, p,$$

where $\hat{\beta}_0^M = \underset{\beta_0}{\text{argmin}} \mathbb{P}_n l(\beta_0, Y)$. Denote $\mathbf{L}_n = (L_{1,n}, L_{2,n}, \cdots, L_{p,n})^T$. Specifically, in ordinary linear models,

$$L_{j,n} = \mathbb{P}_n\{(Y - \hat{\beta}_j^M X_j - \textstyle\sum_{k=1}^{K_n} \hat{\gamma}_{j,k}^M U_k)^2\} - \mathbb{P}_n\{(Y - \hat{\beta}_0^M)^2\}, \text{ for } j = 1, \ldots, p,$$

where $\hat{\beta}_0^M = \underset{\beta_0}{\text{argmin}} \mathbb{P}_n (Y - \beta_0)^2$.

The smaller the $L_{j,n}$ is, the more the variable X_j contributes. We sort the vector $\mathbf{L}_n$ in an ascending order and choose variables according to

$$\widehat{\mathcal{N}}_{\nu_n} = \{1 \leq j \leq p : L_{j,n} \leq \nu_n\}, \tag{4.3}$$

where ν_n is a predefined thresholding parameter. PCAS-MLR ranks the importance of features according to their marginal contributions to the magnitudes of the log-likelihood function given a proportion of principal components. Unlike PCAS-MMLE method which only uses the information of magnitudes of estimators, PCAS-MLR method makes use of more information, including the magnitudes of the estimators as well as their associated variation.

4.4.3 Determining the number of selected variables

It remains open on how to choose the number of selected variables d in variable screening literature. In applications, it is common for practitioners to select a prefixed number of top-ranked variables, as the prefixed number may reflect prior knowledge of the number of susceptible predictors or budget limitations. Another commonly used procedure is to set the size of the selected set to a number less than the sample size, for example $d = [2n/\log(n)]$ (Fan and Lv, 2008), so that the follow-up analysis can be performed in a $p < n$ scenario. Data-driven procedures for determining the size of the important set are appealing but relatively limited. They include information criteria, such as AIC and BIC, and the false discovery rate (FDR) based methods (Barut et al., 2012; Zhao and Li, 2012). These methods, however, have large computational cost, especially in the ultra-high dimensional framework. Following (Fan and Lv, 2008), we used $d = [2n/\log(n)]$ in this thesis.

4.4.4 Determining the number of principal components

The choice of numbers of principal components K_n is critical for PCAS. We propose to use the following two data adaptive methods. The first method is the scree plot, a classical method in factor analysis to determine the number of principal components. As a related numerical method, we can also use the maximum eigenvalue ratio criterion (Luo et al., 2009), defined as λ_j/λ_{j+1} with $1 \leq j \leq n-1$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$. We choose the number of principal components that can maximize the eigenvalue ratio, that is,

$$\hat{k} = \text{argmax}_{1 \leq j \leq k_{\max}}(\lambda_j/\lambda_{j+1}), \tag{4.4}$$

where $k_{\max} \leq n$ is a prespecified maximum factor dimension. When the predictors' correlation structure follows a factor model, it was shown in Wang (2012) that $\hat{k}$ is consistent to the dimension of the linear subspace spanned by the column vectors of factors' matrix.

4.5 Simulations

In this section, we present several simulated linear model examples and logistic regression model examples to evaluate the performance of the proposed procedure and to demonstrate

some factors influencing the false selection rate. We implement four different scenarios to generate data. We vary the size of the nonsparse set of coefficients as well as the number of principal components from 1 to 100 for different scenarios, to gauge difficulties of simulation models on the basis of 200 simulations with sample size 500.

For each simulation setting, we apply two marginal sure independence screening (SIS) procedures based on marginal maximum likelihood estimator (MMLE) and marginal likelihood ratio (MLR), and two PCAS procedures including PCAS-MLR and PCAS-MMLE, to screen variables. The minimum model size required for each method to have a sure screening, i.e. to contain the true model $\mathcal{M}_\star$, is used as a measure of the effectiveness of a screening method. This avoids the issues of choosing the thresholding parameter. For each simulation model, we evaluate each method by summarizing the median minimum model size (MMMS) as well as its robust estimate of the standard deviation (RSD), which is the associated interquartile range (IQR) divided by 1.34.

4.5.1 Simulation Model I

The covariates $\mathbf{X} = (X_1, \cdots, X_p)^T$ is generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and compound symmetric covariance matrix Σ, where $\rho = \Sigma_{ij} = 0.4$, when $i \neq j$. The size of the non-sparse set size s is taken as $6, 8$ and 12 with the true regression coefficients recorded in Table 4.1. The MMMS of selected models with its associated RSD for linear models and logistic regression models with $p = 1000$ and $p = 10000$ are shown in Table 4.1. We recorded the results of PCAS with number of PCs taking as 1, 3, 5, 10, 30, 50 and 100 respectively. The case of zero PCs is SIS method (Fan and Song, 2010). The scree plot is provided as Figure 4.1.

Since the first principal component can explain over 40% variation, much larger than that of the rest PCs, the scree plot in Figure 4.1 suggests that the number of PCs taken should be one. In addition, the maximum eigenvalue ratio estimator gives the same choice, i.e. $\hat{k} = 1$. This is consistent with our observation in Table 4.1, where PCAS performs the best when only one PC is adjusted. The performance of PCAS method is not improved with the increase in the number of principal components. It is reasonable since the proportion of the variation that can be explained by the rest of PCs is so small that including more PCs will not be helpful to account for the additional contribution from the rest of the covariates, instead, it leads to larger estimation variation hence deteriorate the performance of PCAS. We also compute the cases for $\rho = 0.2, 0.6$ and 0.8. Since the

results demonstrate similar trend, we omit the details.

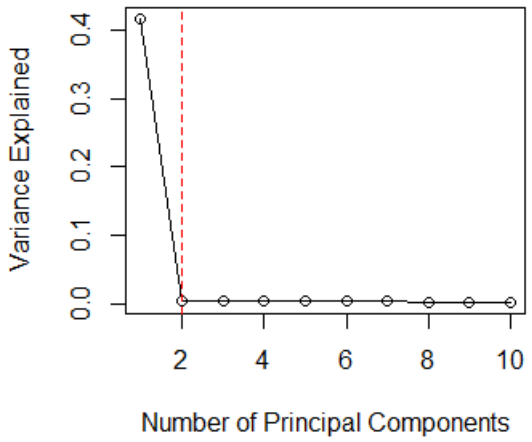

Figure 4.1: The Scree Plot for Linear Models in Simulation Model I with $p = 1000$ and $n = 500$.

4.5.2 Simulation Model II

In this model, we evenly divide all variables into five groups, and each group of variables follows a multivariate normal distribution with mean $\mathbf{0}$ and compound symmetric covariance matrix Σ_ρ, where $\rho = 0.4, 0.5, 0.6$, 0.8 and 0.9 respectively. The MMMS of the selected models with its associated RSD for linear models and logistic regression models with $p = 1000$ and $p = 10000$ are shown in Table 4.2.

PCAS-MLR seems to outperform PCAS-MMLE in terms of smaller MMMS and RSD in many cases. Unlike PCAS-MMLE which uses only the information of magnitudes of estimators, PCAS-MLR makes use of more information, including the magnitudes of the estimators as well as their associated variation.

The scree plot in Figure 4.2 suggests to choose five principal components based on the variance explained. In addition, the maximum eigenvalue ratio estimator gives the same answer, i.e. $\hat{k} = 5$. It is obvious that PCAS method with five principal components

Table 4.1: The MMMS and RSD (in parenthesis) of the simulated examples for linear and logistic regression from simulation model I with $n = 500$ when $p = 1000$ and $p = 10000$. PC=0 refers to the marginal screening in Fan and Lv (2008).

PCs	Variance	SIS-PCA-MLR	SIS-PCA-MMLE	SIS-PCA-MLR	SIS-PCA-MMLE
Setting 1, linear model with $p = 1000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.3, -0.3, 0.3, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, 3, \ldots)^T$	
0	0	13(35)	13(35)	101(96)	101(96)
1	41.5%	7(3)	7(4)	12(0)	12(0)
3	42.2%	7(3)	7(4)	12(0)	12(0)
5	42.8%	7(3)	7(3)	12(0)	12(0)
10	44.4%	7(4)	7(4)	12(0)	12(0)
30	50.2%	8(5)	8(5)	12(0)	12(0)
50	55.4%	11(8)	10(8)	12(0)	12(0)
100	66.4%	19.5(32)	19(31)	12(1)	12(1)
Setting 2, logistic regression with $p = 1000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.7, -0.7, 0.7, \ldots)^T$		$s = 8$, $\boldsymbol{\beta}^\star = (3, 4, 3, \ldots)^T$	
0	0	14(26)	14(26)	70.5(80)	64(82)
1	41.7%	7(3)	7(3)	21(31)	23(28)
3	42.4%	7(3)	7(3)	22.5(31)	24(30)
5	43.0%	7(4)	7(3)	25(29)	26(32)
10	44.6%	7(4)	8(4)	24(38)	27(38)
30	50.4%	8(7)	8(7)	38(49)	37(46)
50	55.5%	10(10)	10.5(10)	58(72)	60(80)
100	66.5%	22(34)	24.5(34)	532(460)	414(347)
Setting 3, linear model with $p = 10000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.3, -0.3, 0.3, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, 3, 4, \ldots)^T$	
0	0	90.5(501)	90.5(501)	830.5(924)	830.5(924)
1	40.3%	14.5(37)	14(35)	12(1)	12(1)
3	40.6%	15(35)	14(34)	12(1)	12(1)
5	41.0%	15(30)	14(29)	12(1)	12(1)
10	41.9%	16.5(36)	15(35)	12(1)	12(1)
30	45.2%	27(49)	25.5(45)	12(1)	12(1)
50	48.5%	36.5(100)	36.5(95)	12(2)	12(2)
100	56.1%	70.5(171)	67.5(170)	14(8)	14(7)
Setting 4, logistic regression with $p = 10000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.7, -0.7, 0.7, \ldots)^T$		$s = 8$,$\boldsymbol{\beta}^\star = (3, 4, 3, \ldots)^T$	
0	0	112(365)	112(366)	641(742)	609.5(731)
1	41.5%	15(30)	16(29)	142(339)	146(354)
3	41.8%	16(32)	17(32)	149.5(372)	160(351)
5	42.2%	15(37)	17(36)	157(392)	168.5(394)
10	43.0%	16.5(35)	17(37)	154(351)	160(367)
30	46.3%	28(51)	26(50)	259(663)	259(646)
50	49.5%	36(68)	34.5(71)	410.5(834)	455(879)
100	57.0%	78.5(206)	80.5(238)	6837(6317)	2570(3513)

adjusted outperforms SIS, and the performance of PCAS method is highly related to the number of PCs used.

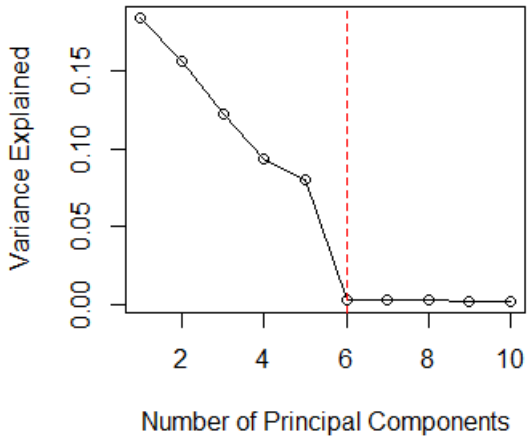

Figure 4.2: The Scree Plot for Linear Models in Simulation Model II with $p = 1000$ and $n = 500$.

4.5.3 Simulation Model III

This simulation model is adopted from (Shen and Huang, 2008), where variables are generated after creating the covariance matrix. First, we generate vectors $v_i,\ i = 1, \cdots, p$, according to a standard normal distribution and let $V = (v_1, \ldots, v_p)'$. Let C be a diagonal matrix, where among the diagonal entries, the top five values are set as 50 and the rest are randomly drawn from a standard uniform distribution. In this way we can generate covariates from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = VCV^T$. The MMMS of the selected models with its associated RSD for linear models and logistic regression models with $p = 1000$ and $p = 10000$ are shown in Table 4.3. The scree plot in Figure 4.3 suggests to choose five principal components based on the variance explained. In addition, the maximum eigenvalue ratio estimator $\hat{k} = 5$. These observations are consistent with the results in Table 4.3.

Table 4.2: The MMMS and RSD (in parenthesis) of the simulated examples for linear and logistic regression model II using different number of PCs with $n = 500$ when $p = 1000$ and $p = 10000$.

PCs	Variance	SIS-PCA-MLR	SIS-PCA-MMLE	SIS-PCA-MLR	SIS-PCA-MMLE
Setting 1, linear model with $p = 1000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.3, -0.3, 0.3, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, 3, 4, \ldots)^T$	
0	0	12(35)	12(35)	30(21)	30(21)
1	18.4%	14(37)	84(66)	30.5(21)	31.5(23)
3	45.3%	19.5(32)	177.5(88)	30(19)	34.5(28)
5	63.6%	7(2)	53(26)	12(0)	39(16)
10	64.8%	7(3)	54(28)	12(1)	38(18)
30	68.9%	9(10)	63.5(40)	12(1)	35(15)
50	72.5%	14(15)	72(51)	12(1)	30.5(15)
100	79.8%	52.5(75)	119(109)	13(2)	28(15)
Setting 2, logistic regression with $p = 1000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.7, -0.7, 0.7, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (-3, 4, -3, 4, \ldots)^T$	
0	0	13(37)	13(37)	856.5(241)	856.5(241)
1	18.7%	14.5(39)	81(73)	862(216)	879(193)
3	46.8%	19(37)	171(90)	878.5(165)	922.5(111)
5	64.7%	7(3)	50(26)	14(5)	73(35)
10	65.8%	7(3)	51(30)	15(7)	76(40)
30	69.9%	9(8)	59(39)	19(16)	84(49)
50	73.4%	12(15)	66(49)	25.5(33)	97(63)
100	80.6%	58(96)	132.5(127)	67(79)	139(98)
Setting 3, linear model with $p = 10000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.3, -0.3, 0.3, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, 3, 4, \ldots)^T$	
0	0	126.5(381)	126.5(381)	165.5(190)	165.5(190)
1	18.4%	154(357)	831(705)	168(186)	190(226)
3	46.8%	190.5(316)	1832.5(734)	154.5(186)	255(330)
5	64.6%	16(22)	489(242)	12(1)	179(113)
10	65.2%	16(29)	490(289)	12(1)	201(109)
30	67.3%	21.5(48)	549(325)	12(1)	206(118)
50	69.3%	34.5(74)	646(392)	12(2)	238(122)
100	74.0%	105(184)	860(590)	14(6)	284.5(162)
Setting 4, logistic regression with $p = 10000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.7, -0.7, 0.7, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (-3, 4, -3, 4, \ldots)^T$	
0	0	163.5(302)	163.5(302)	7978.5(2840)	7978.5(2840)
1	2.0%	160(307)	240.5(327)	8122(2625)	8147.5(2583)
3	5.3%	167.5(300)	312(336)	8039.5(2382)	8092(2325)
5	7.4%	15(26)	51(29)	39(63)	77(40)
10	8.8%	15.5(27)	52(31)	41.5(57)	78.5(39)
30	14.1%	21(45)	56.5(36)	50(91)	85.5(46)
50	19.3%	30(65)	65.5(42)	71(118)	95(61)
100	31.3%	57(125)	87(65)	139(244)	128.5(127)

Table 4.3: The MMMS and RSD (in parenthesis) of the simulated examples for linear and logistic regression model III using different number of PCs with $n = 500$ when $p = 1000$ and $p = 10000$.

PCs	Variance	SIS-PCA-MLR	SIS-PCA-MMLE	SIS-PCA-MLR	SIS-PCA-MMLE
Setting 1, linear model with $p = 1000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (3, -3, 3, -3, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, 3, 4, \ldots)^T$	
0	0	29(20)	29(20)	802.5(121)	802.5(121)
1	7.5%	157.5(173)	135.5(153)	808(229)	813.5(222)
2	14.0%	90.5(150)	50.5(120)	669.5(291)	685(284)
3	20.0%	56.5(161)	26(95)	369(372)	378.5(367)
5	30.8%	6(0)	6(0)	16(6)	15(7)
10	34.0%	6(0)	6(0)	15(6)	14.5(4)
30	44.7%	6(0)	6(0)	14.5(6)	14(6)
50	53.3%	6(0)	6(0)	14(4)	14(6)
100	69.3%	6(0)	6(0)	14(4)	15(5)
Setting 2, logistic regression with $p = 1000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (3, -3, 3, -3, \ldots)^T$		$s = 8$, $\boldsymbol{\beta}^\star = (1.3, 1, 1.3, 1, \ldots)^T$	
0	0	43.5(34)	44(35)	333.5(124)	333.5(124)
1	6.9%	24(35)	24(32)	378.5(190)	613(336)
2	13.6%	18.5(19)	18(18)	364(164)	387(151)
3	19.7%	11.5(13)	9.5(10)	291(247)	272.5(237)
5	30.8%	6(0)	6(0)	8(0)	8(0)
10	34.0%	6(0)	6(0)	8(0)	8(1)
30	44.7%	6(0)	6(0)	8(0)	8(0)
50	53.3%	6(0)	6(0)	8(0)	8(0)
100	69.2%	7(4)	8(4)	8(0)	8(1)
Setting 3, linear model with $p = 10000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.2, -0.2, 0.2, \ldots)^T$		$s = 12$, $\boldsymbol{\beta}^\star = (3, 4, 3, 4, \ldots)^T$	
0	0	98(117)	98(117)	142.5(291)	142.5(291)
1	2.1%	111(126)	106(135)	62(134)	65(142)
2	4.2%	87.5(149)	92(161)	19(72)	21(58)
3	6.2%	60(120)	54.5(123)	15(9)	15.5(11)
5	9.8%	58(127)	52(141)	12(2)	12(3)
10	11.4%	54(106)	53.5(110)	12.5(2)	13(3)
30	17.5%	73.5(152)	96.5(180)	12(3)	13(5)
50	23.3%	90(235)	101(204)	13(2)	13(3)
100	36.3%	261(383)	282.5(395)	14(13)	15(17)
Setting 4, logistic regression with $p = 10000$					
		$s = 6$, $\boldsymbol{\beta}^\star = (0.5, -0.5, 0.5, -0.5, \ldots)^T$		$s = 8$, $\boldsymbol{\beta}^\star = (1.3, 1, 1.3, 1, \ldots)^T$	
0	0	48(118)	48(118)	13(26)	13(26)
1	1.3%	42.5(110)	43(106)	13(22)	13(26)
3	3.6%	40.5(95)	44(88)	12.5(13)	13(14)
5	5.6%	38.5(73)	40(78)	11(11)	11(12)
10	7.3%	44(74)	46(84)	11(9)	12(11)
30	13.7%	61(120)	60(131)	12(17)	13(21)
50	19.7%	94.5(181)	95.5(178)	15.5(25)	16(28)
100	33.3%	186.5(351)	202(334)	34.5(66)	35(73)

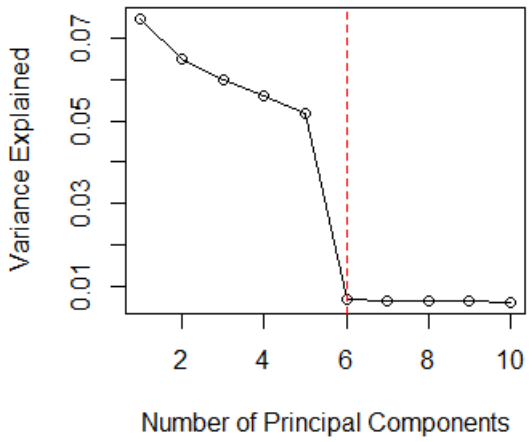

Figure 4.3: The Scree Plot for Linear Models in Simulation Model III with $p = 1000$ and $n = 500$.

4.5.4 Simulation Model IV

This simulation study imitates a genome-wide analysis where the covariates represent genotype status at each SNP across the whole genome. Furthermore, the correlation among all SNPs carries subject's ancestry information reflection latent population substructures which should be controlled when assessing each SNP effect. The covariates $\mathbf{X} = (X_1, \cdots, X_p)^T$ is generated according to the Balding-Nichols model (Balding and Nichols, 1995) as follows. First, we generate a latent variable Y^* that follows a Bernoulli distribution with parameter 0.5. Second, we generate covariates X from a multinomial distribution with parameters depending on the value of the latent variable Y^*. If $Y^* = 0$, X follows a multinomial distribution with parameters $(n, (1-p_l)^2, 2p_l(1-p_l), p_l^2)$. If $Y^* = 1$, X follows a multinomial distribution with $(n, \frac{(1-p_l)^2}{(1-p_l)^2+2Rp_l(1-p_l)+R^2p_l^2}, \frac{2Rp_l(1-p_l)}{(1-p_l)^2+2Rp_l(1-p_l)+R^2p_l^2}, \frac{R^2p_l^2}{(1-p_l)^2+2Rp_l(1-p_l)+R^2p_l^2})$ as the parameters, where p_l follows a Beta distribution with shape parameters $\frac{p(1-F_{ST})}{F_{ST}}$ and $\frac{(1-p)(1-F_{ST})}{F_{ST}}$. In addition, $F_{ST} = 0.04$ represents the genetic distance between two populations, $p = 0.5$, and the relative risk $R = 0.5$. We consider $s = 3, 6$ and 12 for different levels of sparsity. When $s = 3$, $\boldsymbol{\beta}^* = (1, 1.3, 1)^T$. When $s = 6$, $\boldsymbol{\beta}^* = (3, -3, 3, -3, 3, -3)^T$. When $s = 12$, $\boldsymbol{\beta}^\star = (1, 1.3, 1, 1.3, 1, 1.3, \ldots)^T$. The MMMS of the selected models with its associated

RSD for linear models when $s = 12$ are shown in Table 4.4.

Table 4.4: The MMMS and RSD (in parenthesis) of the simulated examples for linear model IV using different number of PCs with $s = 12$, $\boldsymbol{\beta}^\star = (1, 1.3, 1, 1.3, 1, 1.3, \ldots)^T$ when $p = 40000$ and $n = 500$.

PCs	Variance	SIS-PCA-MLR	SIS-PCA-MMLE
		(SIS-MLR)	(SIS-MMLE)
0	0	39(70)	39(70)
1	5.9%	13(4)	13(3)
3	6.3%	13(4)	13(4)
5	6.8%	13(4)	13(4)
10	8.0%	13(4)	13(4)
30	12.2%	14(6)	14(7)
50	17.0%	15(12)	15.5(11)
100	27.8%	23(35)	22(37)

PCAS and SIS can both perfectly identify important predictors when s=3 or 6, therefore the results are not shown in the table format. This may be because the independence structure among predictors leads to the equivalence between the joint population signal and the marginal population signal. We now discuss the case when $s = 6$. Based on the above simulation model, we generate iid random variables $X_1, \cdots, X_p$, $E(X_i X_j) = E(X_i)E(X_j) = 0$ for $i \neq j$ and $E(X_j^2) = 1$. When $j \leq s = 6$, $EX_jY = EX_j(\sum_{k=1}^{s} \beta_k X_k + \epsilon)$=3 or -3. When $j > s = 6$, because of the independence structure, $EX_jY = 0$. It is similar for $s = 3$ case. Although still being a model misspecification, the sparsity structure of the joint model is the same as that of the marginal model. Moreover, there is a clear gap between the marginal signal and the marginal noise. As a result, we can pick up the exact number of important variables with both PCAS and SIS.

When $s = 12$, the scree plot in Figure 4.4 recommends to take one principal component. The corresponding PCAS outperforms SIS method as principal components play a critical role in capturing the correlation structure among predictors. In addition, the performance of PCAS method is not improved by the increase in the number of principal components, indicating that a certain number of principal components can capture the information

among all the predictors reasonably well.

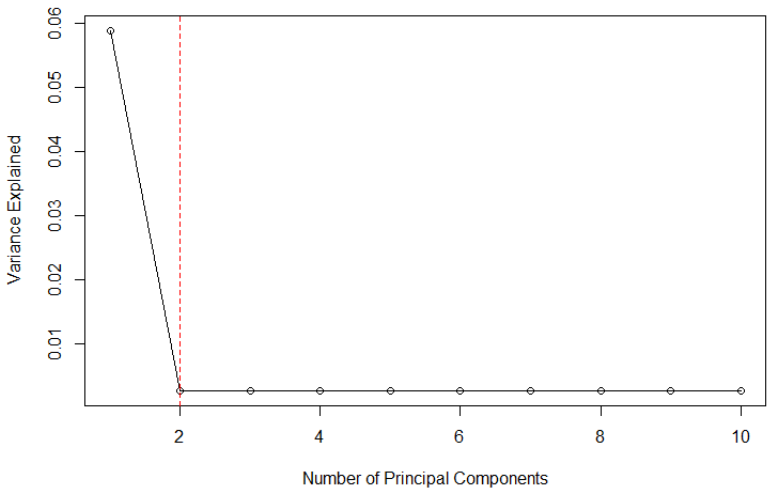

Figure 4.4: Scree Plot for Linear Models in Simulation Model IV with $p = 40000$ and $n = 500$.

4.6 Real data analysis

4.6.1 Affymetric GeneChip Rat Genome 230 2.0 Array Example

To illustrate the proposed method, we analyze the dataset reported in (Scheetz et al., 2006), where 120 12-week-old male rats were selected for harvesting of tissue from the eyes and subsequent microarray analysis. The microarrays used to analyze the RNA from the eyes of these animals contain more than 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multichip averaging method (Irizarry et al., 2003) to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale. We are interested in finding the genes that are related to the TRIM32 gene, which was found to cause Bardet-Biedl syndrome (Chiang et al., 2006) and is a genetically heterogeneous

disease of multiple organ systems, including the retina. Although more than 30,000 probe sets are represented on the Rat Genome 230 2.0 Array, many of these are not expressed in the eye tissue. We focused only on the 18,975 probes that are expressed in the eye tissue.

We applied SIS and the proposed PCAS to this dataset, where $n = 120$ and $p = 18,975$. Because the performance of PCAS-MLR is no worse than that of PCAS-MMLE, we only present the results from PCAS-MLR. With PCAS-MLR, we chose the first 2 principal components based on its scree plot shown in Figure 4.5 as well as the maximum eigenvalue ratio estimator $\hat{k} = 2$.

To evaluate the accuracy of two methods, we used cross validation and compared the prediction error (PE):

$$\text{PE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where y_i is the observed value and $\hat{y}_i$ is the predicted value. By 6-fold cross validation, we randomly partition the data into a training data set of 100 observations and a testing set of 20 observations. On the training data set, we conducted each variable screening method to select $d = 50$ variables, following the suggestion in (Fan and Lv, 2008). Based on these selected variables, we obtained the ordinary least squares (OLS) estimates of the coefficients in the linear regression model, and made a prediction on the testing data set. Then we compared the predicted response with the true response, and obtained the prediction error as well as its standard deviation. As shown in Table 4.5, PCAS-MLR gives the prediction error 0.2278, which is about 50% smaller than 0.4636 produced by SIS. Furthermore, the much smaller standard deviation of PCAS-MLR indicates that PCAS-MLR is more robust than SIS method in this data analysis.

Table 4.5: Comparison between SIS and PCA-SIS over the rats testing data.

Method	Prediction Error	Standard Deviation
SIS	0.4636	0.2563
PCA-SIS	0.2278	0.08762

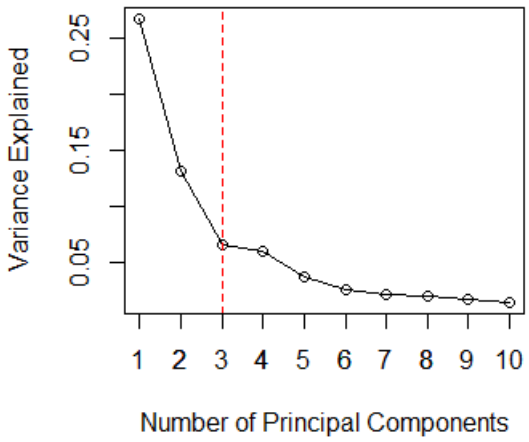

Figure 4.5: Scree Plot for Rat Genome Data with $p = 18975$ and $n = 120$.

4.6.2 European American SNP Example

Our second example is the European American SNP study in (Price et al., 2006). As part of an ongoing disease study, it consists 488 European Americans genotyped on an Affymetrix platform containing 116,204 SNPs. Similarly as in (Price et al., 2006), we use 360 observations after removing outlier individuals. We are interested in finding the SNPs that are related to the height phenotype (0/1 binary data) in European Americans, which leads to 277 variables (Price et al., 2006). We implemented the proposed method and the marginal screening method on the data set, where $n = 360$ and $p = 277$. Both the scree plot in Figure 4.6 and the maximum eigenvalue ratio estimator $\hat{k} = 6$ suggest to use 6 principal components.

Similar as before, we implemented 6-fold cross validation to partition the data into a training data set of 300 observations and a testing set of 60 observations. On the training data set, we selected $d = [2n/\log(n)]$ variables using each variable screening method, and fit the logistic regression based on these selected variables. We then made a prediction on the testing data set, and evaluated the prediction effect by the area under ROC curve (AUC). The result shows that PCAS-MLR obtains a 9.42% larger AUC value than that of SIS and a relatively smaller standard deviation, indicating that PCAS-MLR is preferred in terms of accuracy and robustness.

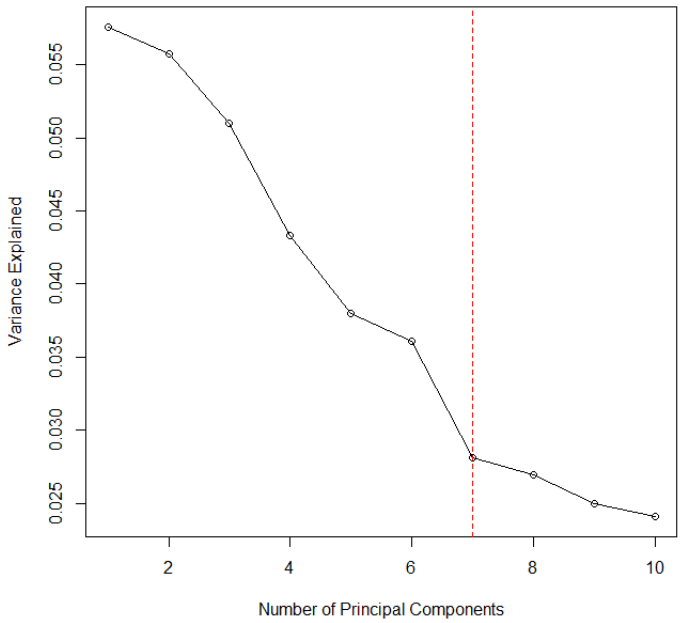

Figure 4.6: Scree Plot for SNP Data with $p = 277$ and $n = 360$.

4.7 Discussions

In this chapter, we propose a PCAS method for generalized linear models, where principal components are used as surrogate covariates to account for the variability of the omitted covariates. Compared with the marginal screening method, PCAS can represent more information of other predictors that are not included in the marginal model, and thus decrease the degree of model misspecification to a large extent. With principal components included in the marginal model, it improves the accuracy as well as the robustness of estimation when dimensionality is ultrahigh. Our proposed method shows improvement from both simulation and real data analysis results.

It is important yet challenging to decide how many principal components should be used when performing this method. In the thesis, we used maximum eigenvalue ratio estimator along with the scree plot. The theoretical property of the proposed approaches is an interesting topic for future research.

REFERENCES

Anil Kumar, D. and Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies*, 1(1):4–28.

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12.

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

Barut, E., Fan, J., and Verhasselt, A. (2012). Conditional sure independence screening. *arXiv preprint arXiv:1206.1024*.

Bittencourt, H. and Clarke, R. (2004). Feature selection by using classification and regression trees (cart). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Boyd, K., Eng, K. H., and Page, C. D. (2013). Area under the precision-recall curve: Point

estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases*, volume 8190, pages 451–466. Springer.

Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brezigar-Masten, A. and Masten, I. (2012). Cart-based selection of bankruptcy predictors for the logit model. *Expert systems with applications*, 39(11):10153–10159.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010a). The balanced accuracy and its posterior distribution. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3121–3124. IEEE.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010b). The binormal assumption on precision-recall curves. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4263–4266. IEEE.

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Chawla, N., Japkowicz, N., and Kolcz, A. (2004). Special issue on learning from imbalanced datasets, sigkdd explorations. In *ACM SIGKDD*.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006). Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl

syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292.

Clémençon, S. and Vayatis, N. (2009). Nonparametric estimation of the precision-recall curve. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 185–192. ACM.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Craven, J. B. M. (2005). Markov networks for detecting overlapping elements in sequence data. *Advances in Neural Information Processing Systems*, 17:193.

Cui, X., Hwang, J. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75.

Davis, J., Burnside, E. S., de Castro Dutra, I., Page, D., Ramakrishnan, R., Costa, V. S., and Shavlik, J. W. (2005). View learning for statistical relational learning: With an application to mammography. In *Proceeding of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 677–683.

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

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32.

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

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494).

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on*, 57(8):5467–5484.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Fan, Y., Kai, Z., and Qiang, L. (2014). A revisit to the class imbalance learning with linear support vector machine. In *Computer Science & Education (ICCSE), 2014 9th International Conference on*, pages 516–521. IEEE.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J. and Popescu, B. E. (2003). Gradient directed regularization for linear regression and classification. Technical report, Statistics Department, Stanford University.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Gey, S. and Nedelec, E. (2005). Model selection for cart regression trees. *Information Theory, IEEE Transactions on*, 51(2):658–670.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3).

Hall, P., Titterington, D., and Xue, J.-H. (2009). Tilting methods for assessing the influence of components in a classifier. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):783–803.

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

Hapfelmeier, A. and Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60:50–69.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

He, H., Garcia, E., et al. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284.

He, X., Wang, L., Hong, H. G., et al. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369.

Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Kok, S. and Domingos, P. (2005). Learning the structure of markov logic networks. In *Proceedings of the 22nd international conference on Machine learning*, pages 441–448.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC curves for continuous data*. CRC Press.

Lewis, R. J. (2000). An introduction to classification and regression tree (cart) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, pages 1–14.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877.

Liu, Y. and Wu, Y. (2012). Variable selection via a combination of the l0 and l1 penalties. *Journal of Computational and Graphical Statistics*.

Liu, Z. and Bondell, H. (2016). Binormal precision-recall curves for optimal classification of imbalanced data. *Unpublished manuscript*.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection statistica sinica (2002) vol. 12, pp. 361–386. *Statistica Sinica*, 12:361–386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, pages 1710–1737.

Loh, W.-Y. (2012). Variable selection for classification and regression in large p, small n problems. In *Probability approximations and beyond*, pages 135–159. Springer.

Luo, R., Wang, H., Tsai, C.-L., et al. (2009). Contour projected dimension reduction. *The Annals of Statistics*, 37(6B):3743–3778.

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

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

Maldonado, S., Weber, R., and Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information Sciences*, 286:228–246.

McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons.

Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal roc curves. *Journal of Mathematical Psychology*, 22(3):218–243.

Metz, C. E. and Pan, X. (1999). proper binormal roc curves: theory and maximum-likelihood estimation. *Journal of mathematical psychology*, 43(1):1–33.

Nash, W. J. (1994). *The Population Biology of Abalone (Haliotis Species) in Tasmania: Blacklip Abalone (H. Rubra) from the North Coast and the Islands of Bass Strait*. Sea Fisheries Division, Marine Research Laboratories-Taroona, Department of Primary Industry and Fisheries, Tasmania.

Neyman, J. and Pearson, E. S. (1992). *On the problem of the most efficient tests of statistical hypotheses*. Springer.

Park, B.-J., Oh, S.-K., and Pedrycz, W. (2013). The design of polynomial function-based neural network predictors for detection of software defects. *Information Sciences*, 229:40–57.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Qiao, Z., Zhou, L., and Huang, J. Z. (2008). Effective linear discriminant analysis for high dimensional, low sample size data. In *Proceeding of the World Congress on Engineering*, volume 2, pages 2–4. Citeseer.

Questier, F., Put, R., Coomans, D., Walczak, B., and Vander Heyden, Y. (2005). The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76(1):45–54.

Raghavan, V., Bollmann, P., and Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229.

Rodenburg, W., Heidema, A. G., Boer, J. M., Bovee-Oudenhoven, I. M., Feskens, E. J., Mariman, E. C., and Keijer, J. (2008). A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes. *Physiological genomics*, 33(1):78–90.

Sauve, M. and Tuleau-Malot, C. (2014). Variable selection through cart. *ESAIM. Probability and Statistics*, 18:770.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Siebert, J. P. (1987). Vehicle recognition using rule based methods. Project report, Turing Institute.

Singla, P. and Domingos, P. (2005). Discriminative training of markov logic networks. In *AAAI*, volume 5, pages 868–873.

Tang, S., Chen, L., Tsui, K.-W., and Doksum, K. (2014). Nonparametric variable selection and classification: The catch algorithm. *Computational Statistics & Data Analysis*, 72:158–175.

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

Van Hulse, J., Khoshgoftaar, T. M., Napolitano, A., and Wald, R. (2009). Feature selection with high-dimensional imbalanced data. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 507–514. IEEE.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika*, 99(1):15–28.

Wang, L. and Shen, X. (2007). On l1-norm multiclass support vector machines. *Journal of the American Statistical Association*, 102(478).

Yang, W. W. and Gu, C. C. (2009). Selection of important variables by statistical learning in genome-wide association analysis. In *BMC proceedings*, volume 3, page S70. BioMed Central Ltd.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.

Zhang, H. H., Liu, Y., Wu, Y., Zhu, J., et al. (2008). Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2:149–167.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis*, 105(1):397–411.

Zheng, Z., Wu, X., and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter*, 6(1):80–89.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004). 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56.

Zou, K. H. and Hall, W. (2000). Two transformation models for estimating an roc curve derived from continuous data. *Journal of Applied Statistics*, 27(5):621–631.

APPENDICES

APPENDIX A

Proof of Proposition 2

The population version of the problem can be described as

$$
\begin{aligned}
(b_0, \beta_0) = \underset{b\in\{-1,1\},\beta\in\mathbb{R}}{\operatorname{argmax}} \int_0^1 \frac{\pi t}{\pi t+(1-\pi)\Phi(\frac{\mu_n-\mu_p}{\sigma_n}+\frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))}dt, \quad \text{s.t.} \qquad (A.1)\\
\mu_p = b\mu_{11} + \beta\mu_{12}, \quad \mu_n = b\mu_{01} + \beta\mu_{02},\\
\sigma_p^2 = \sigma_{11}^2 + \beta^2\sigma_{12}^2 + 2b\beta c_1, \quad \sigma_n^2 = \sigma_{01}^2 + \beta^2\sigma_{02}^2 + 2b\beta c_0.
\end{aligned}
$$

We solve the optimization problem (A.1) separately for each $b \in \{-1, 1\}$ to find the overall maximum. We refer to Lagrange multipliers method to find the maxima of the function subject to four equality constraints, with the objective function f written as

$$
\begin{aligned}
f(\mu_p, \mu_n, \sigma_p, \sigma_n, \beta, \boldsymbol{\lambda}) \;=\; & \int_0^1 \frac{\pi t}{\pi t + (1-\pi)\Phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))}dt \qquad (A.2)\\
& -\lambda_1(\mu_p - b\mu_{11} - \beta\mu_{12}) - \lambda_2(\mu_n - b\mu_{01} - \beta\mu_{02})\\
& -\lambda_3(\sigma_p^2 - \sigma_{11}^2 - \beta^2\sigma_{12}^2 - 2b\beta c_1) - \lambda_4(\sigma_n^2 - \sigma_{01}^2 - \beta^2\sigma_{02}^2 - 2b\beta c_0).
\end{aligned}
$$

Setting the gradient $\nabla_{\boldsymbol{\lambda}} f(\mu_p, \mu_n, \sigma_p, \sigma_n, \beta, \boldsymbol{\lambda}) = 0$ yields the system of constraint equations listed in (A.1), and setting the gradient $\nabla_{\mu_p,\mu_n,\sigma_p,\sigma_n} f = 0$ gives Lagrange multipliers $\boldsymbol{\lambda}$,

where functions $\lambda_j = \lambda_j(b, \beta, \mu_p, \mu_n, \sigma_p, \sigma_n)$ for $j = 1, \cdots, 4$ as follows.

$$
\begin{aligned}
&\frac{\partial f}{\partial \mu_p} = 0 \Longrightarrow \lambda_1 = \int_0^1 \frac{\pi(1-\pi)t\phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))}{\sigma_n[\pi t + (1-\pi)\Phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))]^2}dt, \qquad \text{(A.3)}\\
&\frac{\partial f}{\partial \mu_n} = 0 \Longrightarrow \lambda_2 = -\lambda_1,\\
&\frac{\partial f}{\partial \sigma_p} = 0 \Longrightarrow \lambda_3 = -\frac{1}{2\sigma_p}\int_0^1 \frac{\pi(1-\pi)t\Phi^{-1}(t)\phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))}{\sigma_n[\pi t + (1-\pi)\Phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))]^2}dt,\\
&\frac{\partial f}{\partial \sigma_n} = 0 \Longrightarrow \lambda_4 = \frac{1}{2\sigma_n}\int_0^1 \frac{\pi(1-\pi)t(\mu_n - \mu_p + \sigma_p\Phi^{-1}(t))\phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))}{\sigma_n^2[\pi t + (1-\pi)\Phi(\frac{\mu_n-\mu_p}{\sigma_n} + \frac{\sigma_p}{\sigma_n}\Phi^{-1}(t))]^2}dt,
\end{aligned}
$$

where $\phi(x)$ denotes the probability density function of standard normal distribution, i.e. $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

With the gradient $\nabla_\beta f = 0$, we have

$$\lambda_{10}\mu_{12} + \lambda_{20}\mu_{02} + 2\lambda_{30}\beta_0\sigma_{12}^2 + 2\lambda_{30}b_0c_1 + 2\lambda_{40}\beta_0\sigma_{02}^2 + 2\lambda_{40}b_0c_0 = 0. \qquad \text{(A.4)}$$

APPENDIX B

Proof of Theorem 1

Let $\bar{X}_{11} = \sum_{i=1}^{n_1} X_{11i}/n_1$, $\bar{X}_{12} = \sum_{i=1}^{n_1} X_{12i}/n_1$, $\bar{X}_{01} = \sum_{i=1}^{n_0} X_{01i}/n_0$, $\bar{X}_{02} = \sum_{i=1}^{n_0} X_{02i}/n_0$, $\hat{\sigma}_{11}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{11i}-\bar{X}_{11})^2$, $\hat{\sigma}_{12}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{12i}-\bar{X}_{12})^2$, $\hat{\sigma}_{01}^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(X_{01i}-\bar{X}_{01})^2$, $\hat{\sigma}_{02}^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(X_{02i}-\bar{X}_{02})^2$, $\hat{c}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{12i}-\bar{X}_{12})(X_{11i}-\bar{X}_{11})$, and $\hat{c}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0}(X_{02i}-\bar{X}_{02})(X_{01i}-\bar{X}_{01})$. We can replace the means and variances in (A.1) by the sample versions, and let $(\hat{b}, \hat{\beta})$ be the corresponding solution. Then $\hat{\mu}_p = \hat{b}\bar{X}_{11} + \hat{\beta}\bar{X}_{12}$, $\hat{\mu}_n = \hat{b}\bar{X}_{01} + \hat{\beta}\bar{X}_{02}$, $\hat{\sigma}_p^2 = \sum_{i=1}^{n_1}(\hat{b}X_{11i} - \hat{b}\bar{X}_{11} + \hat{\beta}(X_{12i} - \bar{X}_{12}))^2/n_1$, and $\hat{\sigma}_n^2 = \sum_{i=1}^{n_0}(\hat{b}X_{01i} - \hat{b}\bar{X}_{01} + \hat{\beta}(X_{02i} - \bar{X}_{02}))^2/n_0$. Consequently, we can get the estimating equation $\hat{g}$ as follows, i.e. the sample version of Equation (A.4).

$$\begin{aligned}
\hat{g} &= \hat{\lambda}_1\bar{X}_{12} + \hat{\lambda}_2\bar{X}_{02} + 2\hat{\lambda}_3\hat{\beta}\hat{\sigma}_{12}^2 + 2\hat{\lambda}_3\hat{b}\hat{c}_1 + 2\hat{\lambda}_4\hat{\beta}\hat{\sigma}_{02}^2 + 2\hat{\lambda}_4\hat{b}\hat{c}_0 \\
&= \hat{\lambda}_1\bar{X}_{12} + \hat{\lambda}_2\bar{X}_{02} + \frac{2\hat{\lambda}_3\hat{\beta}}{n_1}\sum_{i=1}^{n_1}(X_{12i} - \bar{X}_{12})^2 + \frac{2\hat{\lambda}_3\hat{b}}{n_1}\sum_{i=1}^{n_1}(X_{12i} - \bar{X}_{12})(X_{11i} - \bar{X}_{11}) \\
&+ \frac{2\hat{\lambda}_4\hat{\beta}}{n_0}\sum_{i=1}^{n_0}(X_{02i} - \bar{X}_{02})^2 + \frac{2\hat{\lambda}_4\hat{b}}{n_0}\sum_{i=1}^{n_0}(X_{02i} - \bar{X}_{02})(X_{01i} - \bar{X}_{01}) = 0, \qquad \text{(B.1)}
\end{aligned}$$

where $\hat{\lambda}_j = \lambda_j(\hat{b}, \hat{\beta}, \hat{\mu}_p, \hat{\mu}_n, \hat{\sigma}_p, \hat{\sigma}_n)$ for $j = 1, \cdots, 4$.

Expand the estimating equation $\hat{g}$ in a 1st order Taylor series around β_0 as

$$0 = \hat{g} \approx g + g' \cdot (\hat{\beta} - \beta_0), \tag{B.2}$$

where g' denotes the first derivative of g with respect to β and $g(\beta)$ is expressed in Equation (A.4).

As a result,

$$\sqrt{n}(\hat{\beta} - \beta_0) \approx \frac{\sqrt{n}(\hat{g} - g)}{g'}. \tag{B.3}$$

By WLLN and the sample version of Equation (A.3) with plugging in sample means and variances as estimates for the population ones, we can have

$$g' \xrightarrow{p} B = \lambda_{10}^{'}\mu_{12} + \lambda_{20}^{'}\mu_{02} + 2\lambda_{30}^{'}\beta_0\sigma_{12}^2 + 2\lambda_{30}\sigma_{12}^2 + 2\lambda_{40}^{'}\beta_0\sigma_{02}^2 + 2\lambda_{40}\sigma_{02}^2 + 2\lambda_{30}^{'}b_0c_1 + 2\lambda_{40}^{'}b_0c_0, \tag{B.4}$$

where $\lambda_{j0}^{'} = \frac{\partial \lambda_{j0}}{\partial \beta_0}$ for $j = 1, \cdots, 4$.

In addition to the independence structure among predictors, by taking $\frac{n_1}{n} \to \pi$, $\frac{n_0}{n} \to 1 - \pi$ into consideration, the Central Limit Theorem (CLT) gives us

$$\sqrt{n}\left((\bar{X}_{11}, \bar{X}_{12}, \bar{X}_{01}, \bar{X}_{02}, \hat{\sigma}_{11}^2, \hat{\sigma}_{12}^2, \hat{\sigma}_{01}^2, \hat{\sigma}_{02}^2, \hat{c}_1, \hat{c}_0)^T - (\mu_{11}, \mu_{12}, \mu_{01}, \mu_{02}, \sigma_{11}^2, \sigma_{12}^2, \sigma_{01}^2, \sigma_{02}^2, c_1, c_0)^T\right) \xrightarrow{d} N(0, \Sigma), \tag{B.5}$$

where Σ is a diagonal matrix with the elements of vector $(\frac{\sigma_{11}^2}{\pi}, \frac{\sigma_{12}^2}{\pi}, \frac{\sigma_{01}^2}{1-\pi}, \frac{\sigma_{02}^2}{1-\pi}, \frac{\mu_{114}-\sigma_{11}^4}{\pi}, \frac{\mu_{124}-\sigma_{12}^4}{\pi}, \frac{\mu_{014}-\sigma_{01}^4}{1-\pi}, \frac{\mu_{024}-\sigma_{02}^4}{1-\pi}, \frac{\sigma_{11}^2\sigma_{12}^2}{\pi}, \frac{\sigma_{01}^2\sigma_{02}^2}{1-\pi})^T$ on the main diagonal, and $\mu_{ij4} = \mathrm{E}[(X_{ij} - \mu_{ij})^4]$ with $i \in \{0, 1\}$ and $j \in \{1, 2\}$.

Consequently, according to multivariate delta method, we have

$$\begin{aligned} &\sqrt{n}(\hat{g} - g) \xrightarrow{d} N(0, D = \mathbf{g'}^T \Sigma \mathbf{g'}), \\ \text{where} \quad &\mathbf{g'} = (\frac{\partial g}{\partial \mu_{11}}, \frac{\partial g}{\partial \mu_{12}}, \frac{\partial g}{\partial \mu_{01}}, \frac{\partial g}{\partial \mu_{02}}, \frac{\partial g}{\partial \sigma_{11}^2}, \frac{\partial g}{\partial \sigma_{12}^2}, \frac{\partial g}{\partial \sigma_{01}^2}, \frac{\partial g}{\partial \sigma_{02}^2}, \frac{\partial g}{\partial c_1}, \frac{\partial g}{\partial c_0})^T. \end{aligned} \tag{B.6}$$

Finally, by Equation (B.3), the asymptotic distribution of $\boldsymbol{\beta}$ is

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V), \tag{B.7}$$

where $V = D/B^2$ with B and D defined in Equation (B.4) and (B.6) separately.