Journal of Machine Learning Research 14 (2013) 629-671 Submitted 3/12; Revised 10/12; Published 2/13

CODA: High Dimensional Copula Discriminant Analysis

Fang Han FHAN@JHSPH.EDU
Department of Biostatistics
Johns Hopkins University
Baltimore, MD 21205, USA

Tuo Zhao TOURZHAO@JHU.EDU
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA

Han Liu HANLIU@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Editor: Tong Zhang

Abstract

We propose a high dimensional classification method, named the *Copula Discriminant Analysis* (CODA). The CODA generalizes the normal-based linear discriminant analysis to the larger Gaussian Copula models (or the nonparanormal) as proposed by Liu et al. (2009). To simultaneously achieve estimation efficiency and robustness, the nonparametric rank-based methods including the Spearman's rho and Kendall's tau are exploited in estimating the covariance matrix. In high dimensional settings, we prove that the sparsity pattern of the discriminant features can be consistently recovered with the parametric rate, and the expected misclassification error is consistent to the Bayes risk. Our theory is backed up by careful numerical experiments, which show that the extra flexibility gained by the CODA method incurs little efficiency loss even when the data are truly Gaussian. These results suggest that the CODA method can be an alternative choice besides the normal-based high dimensional linear discriminant analysis.

Keywords: high dimensional statistics, sparse nonlinear discriminant analysis, Gaussian copula, nonparanormal distribution, rank-based statistics

1. Introduction

High dimensional classification is of great interest to both computer scientists and statisticians. Bickel and Levina (2004) show that the classical low dimensional normal-based linear discriminant analysis (LDA) is asymptotically equivalent to random guess when the dimension d increases fast compared to the sample size n, even if the Gaussian assumption is correct. To handle this problem, a sparsity condition is commonly added, resulting in many follow-up works in recent years. A variety of methods in sparse linear discriminant analysis, including the nearest shrunken centroids (Tibshirani et al., 2002; Wang and Zhu, 2007) and feature annealed independence rules (Fan and Fan, 2008), are based on a working independence assumption. Recently, numerous alternative approaches have been proposed by taking more complex covariance matrix structures into consideration (Fan et al., 2010; Shao et al., 2011; Cai and Liu, 2012; Mai et al., 2012).

A binary classification problem can be formulated as follows: suppose that we have a training set $\{(\boldsymbol{x}_i, y_i), i = 1, ..., n\}$ independently drawn from a joint distribution of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in \{0,1\}$. The target of the classification is to determine the value of Y given a new data point $\boldsymbol{x}$. Let $\psi_0(\boldsymbol{x})$ and $\psi_1(\boldsymbol{x})$ be the density functions of $(\boldsymbol{X}|Y=0)$ and $(\boldsymbol{X}|Y=1)$, and the prior probabilities $\pi_0 = \mathbb{P}(Y=0)$, $\pi_1 = \mathbb{P}(Y=1)$. It is well known that the Bayes rule classifies a new data point $\boldsymbol{x}$ to the second class if and only if

$$\log \psi_1(\boldsymbol{x}) - \log \psi_0(\boldsymbol{x}) + \log(\pi_1/\pi_0) > 0. \tag{1}$$

Specifically, when $(\boldsymbol{X}|Y=0) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, $(\boldsymbol{X}|Y=1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\pi_0 = \pi_1$, Equation (1) is equivalent to the following classifier:

$$g^*(\boldsymbol{x}) := I((\boldsymbol{x} - \boldsymbol{\mu}_a)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d > 0),$$

where $\boldsymbol{\mu}_a := \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0}{2}$, $\boldsymbol{\mu}_d := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, and $I(\cdot)$ is the indicator function. It is well known then that for any linear discriminant rule with respect to $\boldsymbol{w} \in \mathbb{R}^d$:

$$g_{\boldsymbol{w}}(\boldsymbol{X}) := I((\boldsymbol{X} - \boldsymbol{\mu}_a)^T \boldsymbol{w} > 0), \tag{2}$$

the corresponding misclassification error is

$$\mathcal{C}(g_{\boldsymbol{w}}) = 1 - \Phi\left(\frac{\boldsymbol{w}^T \boldsymbol{\mu}_d}{\sqrt{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}}}\right), \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian. By simple calculation, we have

$$\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d \in \underset{\boldsymbol{w} \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{C}(g_{\boldsymbol{w}}),$$

and we denote by $\boldsymbol{\beta}^* := \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$. In exploring discriminant rules with a similar form as Equation (2), both Tibshirani et al. (2002) and Fan and Fan (2008) assume a working independence structure for Σ. However this assumption is often violated in real applications.

Alternatively, Fan et al. (2010) propose the Regularized Optimal Affine Discriminant (ROAD) approach. Let $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\mu}}_d$ be consistent estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_d$. To minimize $\mathcal{C}(g_{\boldsymbol{w}})$ in Equation (3), the ROAD minimizes $\boldsymbol{w}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{w}$ with $\boldsymbol{w}^T \widehat{\boldsymbol{\mu}}_d$ restricted to be a constant value, that is,

$$\min_{\boldsymbol{w}^T \widehat{\boldsymbol{\mu}}_d = 1} \{\boldsymbol{w}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{w}, \text{subject to} \, ||\boldsymbol{w}||_1 \le c\}.$$

Later, Cai and Liu (2012) propose another version of the sparse LDA, which tries to make $\boldsymbol{w}$ close to the Bayes rule's linear term $\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$ in the ℓ_∞ norm (detailed definitions are provided in the next section), that is,

$$\min_{\boldsymbol{w}} \{\, ||\boldsymbol{w}||_1, \text{subject to} \, ||\widehat{\boldsymbol{\Sigma}} \boldsymbol{w} - \widehat{\boldsymbol{\mu}}_d||_\infty \le \lambda_n\}. \tag{4}$$

Equation (4) turns out to be a linear programming problem highly related to the Dantzig selector (Candes and Tao, 2007; Yuan, 2010; Cai et al., 2011).

Very recently, Mai et al. (2012) propose another version of the sparse linear discriminant analysis based on an equivalent least square formulation of the LDA. We will explain it in more details in Section 3. In brief, to avoid the "the curse of dimensionality", an ℓ_1 penalty is added in all three methods to encourage a sparsity pattern of $\boldsymbol{w}$, and hence nice theoretical properties can be obtained under certain regularity conditions. However, though significant process has been made, all these methods require the normality assumptions which can be restrictive in applications.

There are three issues with regard to high dimensional linear discriminant analysis: (1) How to estimate $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ accurately and efficiently (Rothman et al., 2008; Friedman et al., 2007; Ravikumar et al., 2009; Scheinberg et al., 2010); (2) How to incorporate the covariance estimator to classification (Fan et al., 2010; Shao et al., 2011; Cai and Liu, 2012; Witten and Tibshirani, 2011; Mai et al., 2012); (3) How to deal with non-Gaussian data (Lin and Jeon, 2003; Hastie and Tibshirani, 1996). In this paper, we propose a high dimensional classification method, named the Copula Discriminant Analysis (CODA), which addresses all the above three questions.

To handle non-Gaussian data, we extend the underlying conditional distributions of $(\boldsymbol{X}|Y=0)$ and $(\boldsymbol{X}|Y=1)$ from Gaussian to the larger nonparanormal family (Liu et al., 2009). A random variable $\boldsymbol{X}=(X_1,...,X_d)^T$ belongs to a nonparanormal family if and only if there exists a set of univariate strictly increasing functions $\{f_j\}_{j=1}^d$ such that $(f_1(X_1),...,f_d(X_d))^T$ is multivariate Gaussian.

To estimate $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ robustly and efficiently, instead of estimating the transformation functions $\{f_j\}_{j=1}^d$ as Liu et al. (2009) did, we exploit the nonparametric rank-based correlation coefficient estimators including the Spearman's rho and Kendall's tau, which are invariant to the strictly increasing functions f_j. They have been shown to enjoy the optimal parametric rate in estimating the correlation matrix (Liu et al., 2012; Xue and Zou, 2012). Unlike previous analysis, a new contribution of this paper is that we provide an extra condition on the transformation functions which guarantees the fast rates of convergence of the marginal mean and standard deviation estimators, such that the covariance matrix can also be estimated with the parametric rate.

To incorporate the estimated covariance matrix into high dimensional classification, we show that the ROAD (Fan et al., 2010) is connected to the lasso in the sense that if we fix the second tuning parameter, these two problems are equivalent. Using this connection, we prove that the CODA is variable selection consistent.

Unlike the parametric cases, one new challenge for the CODA is that the rank-based covariance matrix estimator may not be positive semidefinite which makes the objective function nonconvex. To solve this problem, we first project the estimated covariance matrix into the cone of positive semidefinite matrices (using elementwise sup-norm). It can be proven that the theoretical properties are preserved in this way.

Finally, to show that the expected misclassification error is consistent to the Bayes risk, we quantify the difference between the CODA classifier $\widehat{g}^{npn}$ and the Bayes rule g^*. To this end, we measure the convergence rate of the estimated transformation function $\{\widetilde{f}_j\}_{j=1}^d$ to the true transformation function $\{f_j\}_{j=1}^d$. Under certain regularity conditions, we show that

$$\sup_{I_{n,\gamma}} |\widetilde{f}_j - f_j| = O_P\left(n^{-\frac{\gamma}{2}}\right), \quad \forall\, j \in \{1,2,\ldots,d\},$$

over an expanding site $I_{n,\gamma}$ determined by the sample size n and a parameter γ (detailed definitions will be provided later). $I_{n,\gamma}$ is set to go to $(-\infty,\infty)$. Using this result, we can show that:

$$\mathbb{E}(\mathcal{C}(\widehat{g}^{npn})) = \mathcal{C}(g^*) + o(1).$$

A related approach to our method has been proposed by Lin and Jeon (2003). They also consider the Gaussian copula family. However, their focus is on fixed dimensions. In contrast, this paper focuses on increasing dimensions and provides a thorough theoretical analysis.

The rest of this paper is organized as follows. In the next section, we briefly review the nonparanormal estimators (Liu et al., 2009, 2012). In Section 3, we present the CODA method. We give a theoretical analysis of the CODA estimator in Section 4, with more detailed proofs collected in the appendix. In Section 5, we present numerical results on both simulated and real data. More discussions are presented in the last section.

2. Background

We start with notations: for any two real values $a,b \in \mathbb{R}$, $a \wedge b := \min(a,b)$. Let $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1,...,v_d)^T \in \mathbb{R}^d$. Let $\boldsymbol{v}_{-j} := (v_1,...,v_{d-1},v_{j+1},...,v_d)^T$ and $\mathbf{M}_{-j,-k}$ be the matrix with $\mathbf{M}$'s j-th row and k-th column removed, $\mathbf{M}_{j,-k}$ be $\mathbf{M}$'s j-th row with the k-th column removed, $\mathbf{M}_{-j,k}$ be $\mathbf{M}$'s k-th column with the j-th row removed. Moreover, $\boldsymbol{v}$'s subvector with entries indexed by I is denoted by $\mathbf{v}_I$, $\mathbf{M}$'s submatrix with rows indexed by I and columns indexed by J is denoted by $\mathbf{M}_{IJ}$, $\mathbf{M}$'s submatrix with all rows and columns indexed by J is denoted by $\mathbf{M}_{\cdot J}$, $\mathbf{M}$'s submatrix with all rows and columns indexed by J is denoted by $\mathbf{M}_J$. For $0 < q < \infty$, we define

$$||\boldsymbol{v}_0|| := \text{card}(\text{support}(\boldsymbol{v})), \quad ||\boldsymbol{v}||_q := \left(\sum_{i=1}^{d} |v_i|^q\right)^{1/q}, \text{ and } ||\boldsymbol{v}||_\infty := \max_{1\le i\le d} |v_i|.$$

We define the matrix $\ell_{\max}$ norm as the elementwise maximum value: $||\mathbf{M}||_{\max} := \max\{|M_{ij}|\}$ and the ℓ_∞ norm as $||\mathbf{M}||_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |M_{ij}|$. $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the smallest and largest eigenvalues of $\mathbf{M}$. We define the matrix operator norm as $||\mathbf{M}||_{\text{op}} := \lambda_{\max}(\mathbf{M})$.

2.1 The *nonparanormal*

A random variable $\boldsymbol{X} = (X_1,...,X_d)^T$ is said to follow a *nonparanormal* distribution if and only if there exists a set of univariate strictly increasing transformations $f = \{f_j\}_{j=1}^d$ such that:

$$f(\boldsymbol{X}) = (f_1(X_1),...,f_d(X_d))^T := \boldsymbol{Z} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = (\mu_1,...,\mu_d)^T$, $\boldsymbol{\Sigma} = [\Sigma_{jk}]$, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^0 = [\Sigma^0_{jk}]$ are the mean, covariance, concentration and correlation matrices of the Gaussian distribution $\boldsymbol{Z}$. $\{\sigma_j^2 := \Sigma_{jj}\}_{j=1}^d$ are the corresponding marginal variances. To make the model identifiable, we add two constraints on f such that f preserves the population means and standard deviations. In other words, for $1 \le j \le d$,

$$\mathbb{E}(X_j) = \mathbb{E}(f_j(X_j)) = \mu_j; \;\; \text{Var}(X_j) = \text{Var}(f_j(X_j)) = \sigma_j^2.$$

In summary, we denote by such $\boldsymbol{X} \sim NPN(\boldsymbol{\mu},\boldsymbol{\Sigma},f)$. Liu et al. (2009) prove that the nonparanormal is highly related to the Gaussian Copula (Clemen and Reilly, 1999; Klaassen and Wellner, 1997).

2.2 Correlation Matrix and Transformation Functions Estimations

Liu et al. (2009) suggest a normal-score based correlation coefficient matrix to estimate $\mathbf{\Sigma}^0$. More specifically, let $\boldsymbol{x}_1,...,\boldsymbol{x}_n \in \mathbb{R}^d$ be n data point where $\boldsymbol{x}_i = (x_{i1},\ldots,x_{id})^T$. We define

$$\widetilde{F}_j(t;\delta_n,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) := T_{\delta_n}\left(\frac{1}{n}\sum_{i=1}^{n} I(x_{ij} \le t)\right), \tag{5}$$

to be the winsorized empirical cumulative distribution function of X_j. Here

$$T_{\delta_n}(x) := \begin{cases} \delta_n, & \text{if} \quad x < \delta_n, \\ x, & \text{if} \quad \delta_n \le x \le 1-\delta_n, \\ 1-\delta_n, & \text{if} \quad x > 1-\delta_n. \end{cases}$$

In particular, the empirical cumulative distribution function $\widehat{F}_j(t;\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) := \widetilde{F}_j(t;0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ by letting $\delta_n = 0$. Let $\Phi^{-1}(\cdot)$ be the quantile function of standard Gaussian, we define

$$\widetilde{f}_j(t) = \Phi^{-1}\big(\widetilde{F}_j(t)\big),$$

and the corresponding sample correlation estimator $\widehat{\mathbf{R}}^{ns} = [\widehat{R}^{ns}_{jk}]$ to be:

$$\widehat{R}^{\text{ns}}_{jk} := \frac{\frac{1}{n}\sum_{i=1}^{n}\widetilde{f}_j(x_{ij})\widetilde{f}_k(x_{ik})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widetilde{f}_j^2(x_{ij})}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widetilde{f}_k^2(x_{ik})}}.$$

Liu et al. (2009) suggest to use the truncation level $\delta_n = \frac{1}{4n^{1/4}\sqrt{\pi\log n}}$ and prove that

$$||\widehat{\mathbf{R}}^{ns} - \mathbf{\Sigma}^0||_{\max} = O_p\left(\sqrt{\frac{\log d \log^2 n}{n^{1/2}}}\right).$$

In contrast, Liu et al. (2012) propose a different approach for estimating the correlations, called the *Nonparanormal* SKEPTIC. The Nonparanormal SKEPTIC exploits the Spearman's rho and Kendall's tau to directly estimate the unknown correlation matrix.

In specific, let r_{ij} be the rank of x_{ij} among $x_{1j},\ldots,x_{nj}$ and $\bar{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$. We consider the following two statistics:

$$\text{(Spearman's rho)} \;\; \widehat{\rho}_{jk} = \frac{\sum_{i=1}^{n}(r_{ij}-\bar{r}_j)(r_{ik}-\bar{r}_k)}{\sqrt{\sum_{i=1}^{n}(r_{ij}-\bar{r}_j)^2\cdot\sum_{i=1}^{n}(r_{ik}-\bar{r}_k)^2}},$$

$$\text{(Kendall's tau)} \;\; \widehat{\tau}_{jk} = \frac{2}{n(n-1)}\sum_{1\le i<i'\le n}\text{sign}\left(x_{ij}-x_{i'j}\right)\left(x_{ik}-x_{i'k}\right).$$

and the correlation matrix estimators:

$$\widehat{R}^{\rho}_{jk} = \begin{cases} 2\sin\left(\frac{\pi}{6}\widehat{\rho}_{jk}\right) & j \ne k \\ 1 & j = k \end{cases} \quad \text{and} \quad \widehat{R}^{\tau}_{jk} = \begin{cases} \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right) & j \ne k \\ 1 & j = k \end{cases}.$$

Let $\widehat{\mathbf{R}}^{\rho} = [\widehat{R}^{\rho}_{jk}]$ and $\widehat{\mathbf{R}}^{\tau} = [\widehat{R}^{\tau}_{jk}]$. Liu et al. (2012) prove the following key result:

Lemma 1 *For any $n \geq \frac{21}{\log d} + 2$, with probability at least $1 - 1/d^2$, we have*

$$||\widehat{\mathbf{R}}^{\rho} - \mathbf{\Sigma}^0||_{\max} \leq 8\pi\sqrt{\frac{\log d}{n}}.$$

For any $n > 1$, with probability at least $1 - 1/d$, we have

$$||\widehat{\mathbf{R}}^{\tau} - \mathbf{\Sigma}^0||_{\max} \leq 2.45\pi\sqrt{\frac{\log d}{n}}.$$

In the following we denote by $\widehat{\mathbf{S}}^{\rho} = [\widehat{S}^{\rho}_{jk}] = [\widehat{\sigma}_j\widehat{\sigma}_k\widehat{R}^{\rho}_{jk}]$ and $\widehat{\mathbf{S}}^{\tau} = [\widehat{S}^{\tau}_{jk}] = [\widehat{\sigma}_j\widehat{\sigma}_k\widehat{R}^{\tau}_{jk}]$, with $\{\widehat{\sigma}^2_j, j = 1,\ldots.d\}$ the sample variances, to be the Spearman's rho and Kendall's tau covariance matrix estimators. As the correlation matrix based on the Spearman's rho and Kendall's tau statistics have similar theoretical performance, in the following sections we omit the superscript ρ and τ and simply denote the estimated correlation and covariance matrices by $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{S}}$.

The following theorem shows that $\widetilde{f}_j$ converges to f_j uniformly over an expanding interval with high probability. This theorem will play a key role in analyzing the classification performance of the CODA method. Here we note that a similar version of this theorem has been shown in Liu et al. (2012), but our result is stronger in terms of extending the region of I_n to be optimal (check the appendix for detailed discussions on it).

Theorem 2 *Let $g_j := f_j^{-1}$ be the inverse function of f_j. In Equation (5), let $\delta_n = \frac{1}{2n}$. For any $0 < \gamma < 1$, we define*

$$I_n := \left[g_j\left(-\sqrt{2(1-\gamma)\log n}\right), g_j\left(\sqrt{2(1-\gamma)\log n}\right)\right],$$

then $\sup_{t \in I_n} |\widetilde{f}_j(t) - f_j(t)| = O_P\left(\sqrt{\frac{\log\log n}{n^{\gamma}}}\right)$.

3. Methods

Let $\boldsymbol{X}_0 \in \mathbb{R}^d$ and $\boldsymbol{X}_1 \in \mathbb{R}^d$ be two random variables with different means $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$ and the same covariance matrix $\mathbf{\Sigma}$. Here we do not pose extra assumptions on the distributions of $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$. Let $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n_0}$ be n_0 data points i.i.d drawn from $\boldsymbol{X}_0$, $\boldsymbol{x}_{n_0+1},\ldots,\boldsymbol{x}_n$ be n_1 data points i.i.d drawn from $\boldsymbol{X}_1$, $n = n_0 + n_1$. Denote by $\mathbf{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^T$, $\boldsymbol{y} = (y_1,...,y_n)^T = (-n_1/n,\ldots,-n_1/n, n_0/n,\ldots, n_0/n)^T$ with the first n_0 entries equal to $-n_1/n$ and the next n_1 entries equal to n_0/n. We have $n_0 \sim \text{Binomial}(n,\pi_0)$ and $n_1 \sim \text{Binomial}(n,\pi_1)$. In the sequel, without loss of generality, we assume that $\pi_0 = \pi_1 = 1/2$. The extension to the case where $\pi_0 \neq \pi_1$ is straightforward (Hastie et al., 2001).

Define

$$\widehat{\boldsymbol{\mu}}_0 = \frac{1}{n_0}\sum_{i:y_i=-n_1/n}\boldsymbol{x}_i, \quad \widehat{\boldsymbol{\mu}}_1 = \frac{1}{n_1}\sum_{i:y_i=n_0/n}\boldsymbol{x}_i, \quad \widehat{\boldsymbol{\mu}}_d = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0, \quad \widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_i \boldsymbol{x}_i,$$

$$\mathbf{S}_0 = \frac{1}{n_0}\sum_{i:y_i=-n_1/n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_0)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_0)^T, \quad \mathbf{S}_1 = \frac{1}{n_1}\sum_{i:y_i=n_0/n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_1)^T,$$

$$\mathbf{S}_b = \frac{1}{n}\sum_{i=0}^{1} n_i(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}})^T = \frac{n_0 n_1}{n^2}\widehat{\boldsymbol{\mu}}_d\widehat{\boldsymbol{\mu}}_d^T, \quad \mathbf{S}_w = \frac{n_0\mathbf{S}_0 + n_1\mathbf{S}_1}{n}.$$

3.1 The Connection between ROAD and Lasso

In this subsection, we first show that in low dimensions, LDA can be formulated as a least square problem. Motivated by such a relationship, we further show that in high dimensions, the lasso can be viewed as a special case of the ROAD. Such a connection between the ROAD and lasso will be further exploited to develop the CODA method.

When $d < n$, we define the population and sample versions of the LDA classifiers as:

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d, \quad \widehat{\boldsymbol{\beta}}^* = \mathbf{S}_w^{-1}\widehat{\boldsymbol{\mu}}_d.$$

Similarly, a least square estimator has the formulation:

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \mathop{\mathrm{argmin}}_{\beta_0, \boldsymbol{\beta}} ||\boldsymbol{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta}||_2^2, \tag{6}$$

where $\mathbf{1} := (1, 1, \ldots, 1)^T$. The following lemma connects the LDA to simple linear regression. The proof is elementary, for self-containess, we include the proof here.

Lemma 3 *Under the above notations, $\widehat{\boldsymbol{\beta}} \propto \widehat{\boldsymbol{\beta}}^*$. Specifically, when $n_1 = n_0$, the linear discriminant classifier $g^*(x)$ is equivalent to the following classifier*

$$\widehat{l}(\boldsymbol{x}) = \begin{cases} 1, & \textit{if} \quad \widehat{\beta}_0 + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} > 0 \\ 0, & \textit{otherwise}. \end{cases}.$$

Proof Taking the first derivatives of the right hand side of (6), we have $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}$ satisfying:

$$n\beta_0 + (n_0\widehat{\boldsymbol{\mu}}_0 + n_1\widehat{\boldsymbol{\mu}}_1)^T\boldsymbol{\beta} = 0, \tag{7}$$

$$(n_0\widehat{\boldsymbol{\mu}}_0 + n_1\widehat{\boldsymbol{\mu}}_1)\beta_0 + (n\mathbf{S}_w + n_1\widehat{\boldsymbol{\mu}}_1\widehat{\boldsymbol{\mu}}_1^T + n_0\widehat{\boldsymbol{\mu}}_0\widehat{\boldsymbol{\mu}}_0^T)\boldsymbol{\beta} = \frac{n_0n_1}{n}\widehat{\boldsymbol{\mu}}_d. \tag{8}$$

Combining Equations (7) and (8), we have

$$(\mathbf{S}_w + \mathbf{S}_b)\widehat{\boldsymbol{\beta}} = \frac{n_0n_1}{n^2}\widehat{\boldsymbol{\mu}}_d.$$

Noticing that $\mathbf{S}_b\widehat{\boldsymbol{\beta}} \propto \widehat{\boldsymbol{\mu}}_d$, it must be true that

$$\mathbf{S}_w\widehat{\boldsymbol{\beta}} = \left(\frac{n_1n_0}{n^2}\widehat{\boldsymbol{\mu}}_d - \mathbf{S}_b\widehat{\boldsymbol{\beta}}\right) \propto \widehat{\boldsymbol{\mu}}_d.$$

Therefore, $\widehat{\boldsymbol{\beta}} \propto \mathbf{S}_w^{-1}\widehat{\boldsymbol{\mu}}_d = \widehat{\boldsymbol{\beta}}^*$. This completes the proof of the first assertion.
Moreover, noticing that by (7), $\widehat{l}(\boldsymbol{x}) = 1$ is equivalent to

$$\widehat{\beta}_0 + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} = -\left(\frac{n_0\widehat{\boldsymbol{\mu}}_0 + n_1\widehat{\boldsymbol{\mu}}_1}{n}\right)^T\widehat{\boldsymbol{\beta}} + \boldsymbol{x}^T\widehat{\boldsymbol{\beta}} = (\boldsymbol{x} - \frac{n_1\widehat{\boldsymbol{\mu}}_1 + n_0\widehat{\boldsymbol{\mu}}_0}{n})^T\widehat{\boldsymbol{\beta}} > 0.$$

because $\widehat{\boldsymbol{\beta}} \propto \widehat{\boldsymbol{\beta}}^*$, $n_1 = n_0$ and $\mathrm{sign}(\widehat{\boldsymbol{\beta}}) = \mathrm{sign}(\widehat{\boldsymbol{\beta}}^*)$ (see Lemma 6 for details), we have $g^*(\boldsymbol{x}) = \widehat{l}(\boldsymbol{x})$. This proves the second assertion. ■

In the high dimensional setting, the following lemma shows that the ROAD is connected to the lasso.

Lemma 4 *We define:*

$$\widehat{\boldsymbol{\beta}}^{\lambda_n,\nu}_{ROAD} = \operatorname*{argmin}_{\beta} \frac{1}{2}\boldsymbol{\beta}^T\mathbf{S}_w\boldsymbol{\beta}+\lambda_n||\boldsymbol{\beta}||_1 + \frac{\nu}{2}(\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d - 1)^2, \tag{9}$$

$$\widehat{\boldsymbol{\beta}}^{\lambda_n}_{*} = \operatorname*{argmin}_{\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d=1} \frac{1}{2n}||\boldsymbol{y}-\widetilde{\mathbf{X}}\boldsymbol{\beta}||_2^2+\lambda_n||\boldsymbol{\beta}||_1, \tag{10}$$

$$\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}||\boldsymbol{y}-\widetilde{\mathbf{X}}\boldsymbol{\beta}||_2^2+\lambda_n||\boldsymbol{\beta}||_1, \tag{11}$$

where $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_{n\times 1}\widehat{\boldsymbol{\mu}}^T$ *is the globally centered version of* $\mathbf{X}$. *We then have* $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{*} = \widehat{\boldsymbol{\beta}}^{\lambda_n,\infty}_{ROAD}$ *and* $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO} = \widehat{\boldsymbol{\beta}}^{\lambda_n,\nu^*}_{ROAD}$ *where* $\nu^* = \frac{n_1 n_0}{n^2}$.

Proof Noticing that the right hand side of Equation (11) has the form:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO} &= \operatorname*{argmin}_{\beta}\left(\frac{1}{2n}||\boldsymbol{y}-\widetilde{\mathbf{X}}\boldsymbol{\beta}||_2^2+\lambda_n||\boldsymbol{\beta}||_1\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\beta}}\left(\frac{1}{2n}\boldsymbol{y}^T\boldsymbol{y} - \frac{n_1n_0}{n^2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d + \frac{1}{2}\boldsymbol{\beta}^T(\mathbf{S}_w+\mathbf{S}_b)\boldsymbol{\beta}+\lambda_n||\boldsymbol{\beta}||_1\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\beta}}\left(\frac{1}{2}\boldsymbol{\beta}^T\mathbf{S}_w\boldsymbol{\beta} + \frac{n_1n_0}{2n^2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d\widehat{\boldsymbol{\mu}}_d^T\boldsymbol{\beta} - \frac{n_1n_0}{n^2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d+\lambda_n||\boldsymbol{\beta}||_1\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\beta}}\left(\frac{1}{2}\boldsymbol{\beta}^T\mathbf{S}_w\boldsymbol{\beta} + \frac{1}{2}\frac{n_1n_0}{n^2}(\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d-1)^2+\lambda_n||\boldsymbol{\beta}||_1\right).
\end{aligned}$$

And similarly

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\lambda_n}_{*} &= \operatorname*{argmin}_{\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d=1}\left(\frac{1}{2}\boldsymbol{\beta}^T\mathbf{S}_w\boldsymbol{\beta} + \frac{1}{2}\frac{n_1n_0}{n^2}(\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d-1)^2+\lambda_n||\boldsymbol{\beta}||_1\right) \\
&= \operatorname*{argmin}_{\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d=1}\left(\frac{1}{2}\boldsymbol{\beta}^T\mathbf{S}_w\boldsymbol{\beta}+\lambda_n||\boldsymbol{\beta}||_1\right).
\end{aligned}$$

This finishes the proof. ■

Motivated by the above lemma, later we will show that $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO}$ is already variable selection consistent.

3.2 Copula Discriminant Analysis

In this subsection we introduce the Copula Discriminant Analysis (CODA). We assume that

$$\boldsymbol{X}_0 \sim NPN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}, f), \quad \boldsymbol{X}_1 \sim NPN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}, f).$$

Here the transformation functions are $f = \{f_j\}_{j=1}^d$. In this setting, the corresponding Bayes rule can be easily calculated as:

$$g^{npn}(\boldsymbol{x}) = \begin{cases} 1, & \text{if} \quad (f(\boldsymbol{x})-\boldsymbol{\mu}_a)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d > 0, \\ 0, & \text{otherwise.}\end{cases} \tag{12}$$

By Equation (12) the Bayes rule is the sign of the log odds: $(f(\boldsymbol{x})-\boldsymbol{\mu}_a)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$. Therefore, similar to the linear discriminant analysis, if there is a sparsity pattern on $\boldsymbol{\beta}^* := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$, a fast rate is expected.

Inspired by Lemma 3 and Lemma 4, to recover the sparsity pattern, we propose the ℓ_1 regularized minimization equation:

$$\widehat{\boldsymbol{\beta}}_{npn} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left(\frac{1}{2}\boldsymbol{\beta}^T\widehat{\mathbf{S}}\boldsymbol{\beta} + \frac{\nu}{2}(\boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_d - 1)^2 + \lambda_n||\boldsymbol{\beta}||_1 \right). \tag{13}$$

Here $\widehat{\mathbf{S}} = n_0/n \cdot \widehat{\mathbf{S}}_0 + n_1/n \cdot \widehat{\mathbf{S}}_1$, $\widehat{\mathbf{S}}_0$ and $\widehat{\mathbf{S}}_1$ are the Spearman's rho/Kendall's tau covariance matrix estimators of $[\boldsymbol{x}_1,...,\boldsymbol{x}_{n_0}]^T$ and $[\boldsymbol{x}_{n_0+1},...,\boldsymbol{x}_n]^T$, respectively. When ν is set to be $\frac{n_0n_1}{n^2}$, $\widehat{\boldsymbol{\beta}}_{npn}$ parallels the ℓ_1 regularization formulation shown in Equation (11); when ν goes to infinity, $\widehat{\boldsymbol{\beta}}_{npn}$ reduces to $\widehat{\boldsymbol{\beta}}_*^{\lambda_n}$ shown in Equation (10).

For any new data point $\boldsymbol{x} = (x_1,\ldots,x_d)^T$, reminding that the transforms f_j preserves the mean of x_j, we assign it to the second class if and only if

$$(\widehat{f}(\boldsymbol{x}) - \widehat{\boldsymbol{\mu}})^T\widehat{\boldsymbol{\beta}}_{npn} > 0,$$

where $\widehat{f}(\boldsymbol{x}) = (\widehat{f}_1(x_1),\ldots,\widehat{f}_d(x_d))^T$ with

$$\widehat{f}_j(x_j) = \left(n_0\widehat{f}_{0j}(x_j) + n_1\widehat{f}_{1j}(x_j)\right)/n, \quad \forall\, j \in \{1,\ldots,d\}.$$

Here $\widehat{f}_{0j}$ and $\widehat{f}_{1j}$ are defined to be:

$$\widehat{f}_{0j}(t) := \widehat{\mu}_0 + \widehat{S}_{jj}^{-1/2}\Phi^{-1}\left(\widetilde{F}_j(t;\delta_{n_0},\boldsymbol{x}_1,...,\boldsymbol{x}_{n_0})\right),$$

and

$$\widehat{f}_{1j}(t) := \widehat{\mu}_1 + \widehat{S}_{jj}^{-1/2}\Phi^{-1}\left(\widetilde{F}_j(t;\delta_{n_1},\boldsymbol{x}_{n_0+1},...,\boldsymbol{x}_n)\right).$$

Here we use the truncation level $\delta_n = \dfrac{1}{2n}$. The corresponding classifier is named $\widehat{g}^{npn}$.

3.3 Algorithms

To solve the Equation (13), when ν is set to be $\frac{n_0n_1}{n^2}$, Lemma 4 has shown that it can be formulated as a ℓ_1 regularized least square problem and hence popular softwares such as *glmnet* (Friedman et al., 2009, 2010) or *lars* (Efron et al., 2004) can be applied.

When ν goes to infinity, the Equation (13) reduces to the ROAD, which can be efficiently solved by the augmented Lagrangian method (Nocedal and Wright, 2006). More specifically, we define the augmented Lagrangian function:

$$\mathcal{L}(\boldsymbol{\beta},u) = \frac{1}{2}\boldsymbol{\beta}^T\widehat{\mathbf{S}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 + \nu u(\widehat{\boldsymbol{\mu}}^T\boldsymbol{\beta} - 1) + \frac{\nu}{2}(\widehat{\boldsymbol{\mu}}^T\boldsymbol{\beta} - 1)^2,$$

where $u \in \mathbb{R}$ is the rescaled Lagrangian multiplier and $\nu > 0$ is the augmented Lagrangian multiplier. We can obtain the optimum to Equation (9) using the following iterative procedure. Suppose at the k-th iteration, we already have the solution $\boldsymbol{\beta}^{(k)}$, $u^{(k)}$, then at the $(k+1)$-th iteration,

• Step.1 Minimize $\mathcal{L}(\boldsymbol{\beta}, u)$ with respect to $\boldsymbol{\beta}$. It can be efficiently solved by coordinate descent. We rearrange

$$\widehat{\mathbf{S}} = \begin{pmatrix} \widehat{S}_{j,j} & \widehat{\mathbf{S}}_{j,-j} \\ \widehat{\mathbf{S}}_{-j,j} & \widehat{\mathbf{S}}_{-j,-j} \end{pmatrix}, \ \ \widehat{\boldsymbol{\mu}} = (\widehat{\mu}_j, \widehat{\boldsymbol{\mu}}_{-j}^T)^T$$

and $\boldsymbol{\beta} = (\beta_j, \boldsymbol{\beta}_{-j}^T)^T$. Then we can iteratively update β_j by the formula

$$\beta_j^{(k+1)} = \frac{\mathrm{Soft}\left(\nu\widehat{\mu}_j\left(1 - u^{(k)} - \widehat{\boldsymbol{\mu}}_{-j}^T\boldsymbol{\beta}_{-j}^{(k)}\right) - \widehat{\mathbf{S}}_{j,-j}\boldsymbol{\beta}_{-j}^{(k)}, \ \lambda\right)}{\widehat{S}_{j,j} + \nu\widehat{\mu}_j^2},$$

where $\mathrm{Soft}(x,\lambda) := \mathrm{sign}(x)(|x| - \lambda)^+$. It is observed that a better empirical performance can be achieved by updating each β_j only once.

• Step.2 Update u using the formula

$$u^{(k+1)} = u^{(k)} + \widehat{\boldsymbol{\mu}}^T\boldsymbol{\beta}^{(k+1)} - 1.$$

This augmented Lagrangian method has provable global convergence. See Chapter 17 of Nocedal and Wright (2006) for discussions in details. Our empirical simulations show that this algorithm is more accurate than Fan et al. (2010)'s method.

To solve Equation (13), we also need to make sure that $\widehat{\mathbf{S}}$, or equivalently $\widehat{\mathbf{R}}$, is positive semidefinite. Otherwise, Equation (13) is not a convex optimization problem and the above algorithm may not even converge. Heuristically, we can truncate all of the negative eigenvalues of $\widehat{\mathbf{R}}$ to zero. In practice, we project $\widehat{\mathbf{R}}$ into the cone of the positive semidefinite matrices and find solution $\widetilde{\mathbf{R}}$ to the following convex optimization problem:

$$\widetilde{\mathbf{R}} = \operatorname*{arg\,min}_{\mathbf{R} \succeq 0} ||\widehat{\mathbf{R}} - \mathbf{R}||_{\max}, \tag{14}$$

where $\ell_{\max}$ norm is chosen such that the theoretical properties in Lemma 1 can be preserved. In specific, we have the following corollary:

Lemma 5 *For all $t \geq 32\pi\sqrt{\frac{\log d}{n \log 2}}$, the minimizer $\widetilde{\mathbf{R}}$ to Equation* (14) *satisfies the following exponential inequality:*

$$\mathbb{P}(|\widetilde{R}_{jk} - \Sigma_{jk}^0| \geq t) \leq 2\exp\left(-\frac{nt^2}{512\pi^2}\right), \quad \forall\, 1 \leq j,k \leq d.$$

Proof Combining Equation (A.23) and Equation (A.28) of Liu et al. (2012), we have

$$\mathbb{P}(|\widehat{R}_{jk} - \Sigma_{jk}^0| > t) \leq 2\exp\left(-\frac{nt^2}{64\pi^2}\right).$$

Because $\boldsymbol{\Sigma}^0$ is feasible to Equation (14), $\widetilde{\mathbf{R}}$ must satisfy that:

$$||\widehat{\mathbf{R}} - \widetilde{\mathbf{R}}||_{\max} \leq ||\widehat{\mathbf{R}} - \boldsymbol{\Sigma}^0||_{\max}.$$

Using Pythagorean Theorem, we then have

$$\begin{aligned}\mathbb{P}(|\widetilde{R}_{jk}-\Sigma^0_{jk}|\geq t) &\leq \mathbb{P}(|\widetilde{R}_{jk}-\widehat{R}_{jk}|+|\widehat{R}_{jk}-\Sigma^0_{jk}|\geq t)\\ &\leq \mathbb{P}(||\widetilde{\mathbf{R}}-\widehat{\mathbf{R}}||_{\max}+||\widehat{\mathbf{R}}-\mathbf{\Sigma}^0||_{\max}\geq t)\\ &\leq \mathbb{P}(||\widehat{\mathbf{R}}-\mathbf{\Sigma}^0||_{\max}\geq t/2)\\ &\leq d^2\exp\left(-\frac{nt^2}{256\pi^2}\right)\\ &\leq 2\exp\left(\frac{2\log d}{\log 2}-\frac{nt^2}{256\pi^2}\right).\end{aligned}$$

Using the fact that $t\geq 32\pi\sqrt{\frac{\log d}{n\log 2}}$, we have the result. ■

Therefore, the theoretical properties in Lemma 1 also hold for $\widetilde{\mathbf{R}}$, only with a slight loose on the constant. In practice, it has been found that the optimization problem in Equation (14) can be formulated as the dual of a graphical lasso problem with the smallest possible tuning parameter that still guarantees a feasible solution (Liu et al., 2012). Empirically, we can use a surrogate projection procedure that computes a singular value decomposition of $\widehat{\mathbf{R}}$ and truncates all of the negative singular values to be zero. And then we define $\widetilde{\mathbf{S}}:=[\widetilde{S}_{jk}]=[\widehat{\sigma}_j\widehat{\sigma}_k\widetilde{R}_{jk}]$ to be the projected Spearman's rho/Kendall's tau covariance matrices, which can be plugged into Equation (13) to obtain an optimum.

3.4 Computational Cost

Compared to the corresponding parametric methods like the ROAD and the least square formulation proposed by Mai et al. (2012), one extra cost of the CODA is the computation of $\widetilde{\mathbf{R}}$, which can be solved in two steps: (1) computing $\widehat{\mathbf{R}}$; (2) projecting $\widehat{\mathbf{R}}$ to the cone of the positive semidefinite matrices. In the first step, computing $\widehat{\mathbf{R}}$ requires the calculation of $d(d-1)/2$ pairwise Spearman's rho or Kendall's tau statistics. As shown in Christensen (2005) and Kruskal (1958), $\widehat{\mathbf{R}}$ can be computed with the cost $O(d^2 n\log n)$. In the second step, to obtain $\widetilde{\mathbf{R}}$ requires estimating a full path of estimates by implementing the graphical lasso algorithm. This approach shows good scalability to very high dimensional data sets (Friedman et al., 2007; Zhao et al., 2012). Moreover, in practice we can use a surrogate projection procedure, which can be solved by implementing the SVD decomposition of $\widehat{\mathbf{R}}$ once.

4. Theoretical Properties

In this section we provide the theoretical properties of the CODA method. We set $\nu=(n_0n_1)/n^2$ in Equation (13). With such a choice of ν, we prove that the CODA method is variable selection consistent and has an oracle property. We define

$$\mathbf{C}:=\mathbf{\Sigma}+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T/4. \tag{15}$$

To calculate $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO}$ in Equation (11), we define $\widetilde{\mathbf{\Sigma}}$:

$$\widetilde{\mathbf{\Sigma}}:=\widetilde{\mathbf{S}}+\frac{n_0n_1}{n^2}\cdot(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)^T.$$

We then replace $\frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$ with $\widetilde{\mathbf{\Sigma}}$ in Equation (11).

It is easy to see that $\widetilde{\mathbf{\Sigma}}$ is a consistent estimator of C. We define $\lfloor c \rfloor$ the greatest integer strictly less than the real number c. For any subset $T \subseteq \{1,2,\ldots,d\}$, let $\widetilde{\mathbf{X}}_T$ be the $n \times |T|$ matrix with the vectors $\{\widetilde{\mathbf{X}}_{\cdot i}, i \in T\}$ as columns. We assume that $\boldsymbol{\beta}^*$ is sparse and define $S := \{i \in 1,...,d | \beta_i^* \neq 0\}$ with $|S| = s$, $s \ll n$. we denote by

$$\boldsymbol{\beta}^{**} := \mathbf{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \quad \text{where } \beta^{**}_{\max} := \max_{j \in S}(|\beta^{**}_j|), \text{ and } \beta^{**}_{\min} := \min_{j \in S}(|\beta^{**}_j|).$$

Recalling that $\boldsymbol{\beta}^* = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$, the next lemma claims that $\boldsymbol{\beta}^{**} \propto \boldsymbol{\beta}^*$ and therefore $\boldsymbol{\beta}^{**}$ is also sparse, and hence $\boldsymbol{\beta}^{**}_S = (\mathbf{C}_{SS})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S$.

Lemma 6 *Let $\boldsymbol{\beta}^* = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d$. $\boldsymbol{\beta}^{**}$ is proportional to $\boldsymbol{\beta}^*$. Especially, we have*

$$\boldsymbol{\beta}^{**} = \frac{4\boldsymbol{\beta}^*}{4 + \boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d}.$$

Proof Using the Binomial inverse theorem (Strang, 2003), we have

$$\begin{aligned}\boldsymbol{\beta}^{**} &= (\mathbf{\Sigma} + \frac{1}{4}\boldsymbol{\mu}_d\boldsymbol{\mu}_d^T)^{-1}\boldsymbol{\mu}_d = \left(\mathbf{\Sigma}^{-1} - \frac{\frac{1}{4}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d\boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}}{1 + \frac{1}{4}\boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d}\right)\boldsymbol{\mu}_d = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d - \frac{\frac{1}{4}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d(\boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d)}{1 + \frac{1}{4}\boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d} \\ &= \left(1 - \frac{\boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d}{4 + \boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d}\right)\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d = \frac{4\boldsymbol{\beta}^*}{4 + \boldsymbol{\mu}_d^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_d}.\end{aligned}$$

This completes the proof. ∎

We want to show that $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO}$ recovers the sparsity pattern of the unknown $\boldsymbol{\beta}^*$ with high probability. In the sequel, we use $\widehat{\boldsymbol{\beta}}$ to denote $\widehat{\boldsymbol{\beta}}^{\lambda_n}_{LASSO}$ for notational simplicity. We define the variable selection consistency property as:

Definition 7 (Variable Selection Consistency Property) *We say that a procedure has the variable selection consistency property $\mathcal{R}(\mathbf{X}, \boldsymbol{\beta}^{**}, \lambda_n)$ if and only if there exists a λ_n and an optimal solution $\widehat{\boldsymbol{\beta}}$ such that $\widehat{\boldsymbol{\beta}}_S \neq 0$ and $\widehat{\boldsymbol{\beta}}_{S^c} = 0$.*

Furthermore, to ensure variable selection consistency, the following condition on the covariance matrix is imposed:

Definition 8 *A positive definite matrix C has the Irrepresentable Conditions (IC) property if*

$$||\mathbf{C}_{S^cS}(\mathbf{C}_{SS})^{-1}||_\infty := \psi < 1.$$

This assumption is well known to secure the variable selection consistency of the lasso procedure and we refer to Zou (2006), Meinshausen and Bühlmann (2006), Zhao and Yu (2007) and Wainwright (2009) for more thorough discussions.

The key to prove the variable selection consistency is to show that the marginal sample means and standard deviations converge to the population means and standard deviations in a fast rate for the nonparanormal. To get this result, we need extra conditions on the transformation functions $\{f_j\}_{j=1}^d$. For this, we define the Subgaussian Transformation Function Class.

Definition 9 (Subgaussian Transformation Function Class) *Let $Z \in \mathbb{R}$ be a random variable following the standard Gaussian distribution. The Subgaussian Transformation Function Class* $\mathrm{TF}(K)$ *is defined as the set of functions $g : \mathbb{R} \to \mathbb{R}$ which satisfies:*

$$\mathbb{E}|g(Z)|^m \leq \frac{m!}{2} K^m, \quad \forall\, m \in \mathbb{Z}^+.$$

Remark 10 *Here we note that for any function $g : \mathbb{R} \to \mathbb{R}$, if there exists a constant $L < \infty$ such that*

$$g(z) \leq L \quad \text{or} \quad g'(z) \leq L \quad \text{or} \quad g''(z) \leq L, \;\; \forall\, z \in \mathbb{R}, \tag{16}$$

then $g \in \mathrm{TF}(K)$ for some constant K. To show that, we have the central absolute moments of the standard Gaussian distribution satisfying, $\forall\, m \in \mathbb{Z}^+$:

$$\begin{aligned} &\mathbb{E}|Z|^m \leq (m-1)!! < m!!, \\ &\mathbb{E}|Z^2|^m = (2m-1)!! < m! \cdot 2^m. \end{aligned} \tag{17}$$

Because g satisfies the condition in Equation (16)*, using Taylor expansion, we have for any $z \in \mathbb{R}$,*

$$g(z) \leq |g(0)| + L \text{ or } |g(z)| \leq |g(0)| + L|z|, \text{ or } |g(z)| \leq |g(0)| + |g'(0)z| + Lz^2. \tag{18}$$

Combining Equations (17) *and* (18)*, we have $\mathbb{E}|g(Z)|^m \leq \frac{m!}{2} K^m$ for some constant K. This proves the assertion.*

The next theorem provides the variable selection consistency result of the proposed procedure. It shows that under certain conditions on the covariance matrix Σ and the transformation functions, the sparsity pattern of β^{**} can be recovered with a parametric rate.

Theorem 11 (Sparsity Recovery) *Let $\boldsymbol{X}_0 \sim NPN(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}, f)$, $\boldsymbol{X}_1 \sim NPN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}, f)$. We assume that* $\mathbf{C}$ *in Equation* (15) *satisfies the IC condition and $||(\mathbf{C}_{SS})^{-1}||_\infty = D_{\max}$ for some $0 < D_{\max} < \infty$, $||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0||_\infty = \Delta_{\max}$ for some $0 < \Delta_{\max} < \infty$ and $\lambda_{\min}(\mathbf{C}_{SS}) > \delta$ for some constant $\delta > 0$. Then, if we have the additional conditions:*

Condition 1: λ_n is chosen such that

$$\lambda_n < \min\left\{ \frac{3\beta^{**}_{\min}}{64 D_{\max}}, \; \frac{3\Delta_{\max}}{32} \right\};$$

Condition 2: Let $\sigma_{\max}$ be a constant such that

$$0 < 1/\sigma_{\max} < \min_j\{\sigma_j\} < \max_j\{\sigma_j\} < \sigma_{\max} < \infty, \max_j |\mu_j| \leq \sigma_{\max},$$

and $g = \{g_j := f_j^{-1}\}_{j=1}^d$ satisfies

$$g_j^2 \in \mathrm{TF}(K), \quad \forall\, j \in \{1, \ldots d\},$$

where $K < \infty$ is a constant,

then there exist positive constants c_0 and c_1 only depending on $\{g_j\}_{j=1}^d$, such that for large enough n

$$\begin{aligned}&\mathbb{P}(\mathcal{R}(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n))\\ &\geq 1-\underbrace{\left[2ds\cdot\exp\left(-\frac{c_0n\varepsilon^2}{s^2}\right)+2d\cdot\exp\left(-\frac{4c_1n\lambda_n^2(1-\psi-2\varepsilon D_{\max})^2}{(1+\psi)^2}\right)\right]}_{A}\\ &-\underbrace{\left[2s^2\exp\left(-\frac{c_0n\varepsilon^2}{s^2}\right)+2s\exp(-c_1n\varepsilon^2)\right]}_{B}-\underbrace{2s^2\exp(-\frac{c_0n\delta^2}{4s^2})}_{C}-\underbrace{2\exp\left(-\frac{n}{8}\right)}_{D},\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\left(||\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0n_1}-\boldsymbol{\beta}^{**}||_\infty\leq 228D_{\max}\lambda_n\right)\\ &\geq 1-\underbrace{\left[2s^2\exp\left(-\frac{c_0n\varepsilon^2}{s^2}\right)+2s\exp(-c_1n\varepsilon^2)\right]}_{B}-\underbrace{2\exp\left(-\frac{n}{8}\right)}_{D},\end{aligned}\qquad(19)$$

whenever ε satisfies that, for large enough n,

$$\begin{aligned}64\pi\sqrt{\frac{\log d}{n\log 2}}\leq\varepsilon<\min\Bigg\{&1,\ \frac{1-\psi}{2D_{\max}},\ \frac{2\lambda_n(1-\psi)}{D_{\max}(4\lambda_n+(1+\psi)\Delta_{\max})},\\ &\frac{\omega}{(3+\omega)D_{\max}},\ \frac{\Delta_{\max}\omega}{6+2\omega},\ \frac{4\lambda_n}{D_{\max}\Delta_{\max}},8\lambda_n^2\Bigg\}.\end{aligned}$$

Here $\omega:=\frac{\beta^{**}_{\min}}{\Delta_{\max}D_{\max}}$ and $\delta\geq 128\pi s\sqrt{\frac{\log d}{n\log 2}}$.

Remark 12 *The above Condition 2 requires the transformation functions' inverse $\{g_j\}_{j=1}^d$ to be restricted such that the estimated marginal means and standard deviations converge to their population quantities exponentially fast. The exponential term A is set to control $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{S^c}\neq 0)$, B is set to control $\mathbb{P}(\widehat{\boldsymbol{\beta}}_S=0)$, C is set to control $\mathbb{P}(\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{SS})\leq 0)$ and D is set to control $\mathbb{P}(\frac{3}{16}\leq\frac{n_0n_1}{n^2}\leq\frac{1}{4})$. Here we note that the key of the proof is to show that: (i) there exist fast rates for sample means and standard deviations converging to the population means and standard deviations for the nonparanormal; (ii) $\widetilde{\boldsymbol{\Sigma}}_{SS}$ is invertible with high probability. $\varepsilon\geq 64\pi\sqrt{\frac{\log d}{n\log 2}}$ is used to make sure that the Lemma 5 can be applied here.*

The next corollary provides an asymptotic result of the Theorem 11.

Corollary 13 *Under the same conditions as in Theorem 11, if we further have the following Conditions 3,4 and 5 hold:*

Condition 3: $D_{\max}$, $\Delta_{\max}$, ψ and δ are constants that do not scale with (n,d,s);

Condition 4: The triplet (n,d,s) admits the scaling such that

$$s\sqrt{\frac{\log d+\log s}{n}}\to 0 \quad\text{and}\quad \frac{s}{\beta^{**}_{\min}}\sqrt{\frac{\log d+\log s}{n}}\to 0;$$

Condition 5: λ_n scales with (n,d,s) such that

$$\frac{\lambda_n}{\beta^{**}_{\min}}\to 0 \quad\text{and}\quad \frac{s}{\lambda_n}\sqrt{\frac{\log d+\log s}{n}}\to 0,$$

then

$$\mathbb{P}(\mathcal{R}(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n))\to 1.$$

Remark 14 *Condition 3 is assumed to be true, in order to give an explicit relationship among (n,d,s) and λ_n. Condition 4 allows the dimension d to grow in an exponential rate of n, which is faster than any polynomial of n. Condition 5 requires that λ_n shrinks towards zero in a slower rate than $s\sqrt{\frac{\log d+\log s}{n}}$.*

In the next theorem, we analyze the classification oracle property of the CODA method. Suppose that there is an oracle, which classifies a new data point $\boldsymbol{x}$ to the second class if and only if

$$(f(\boldsymbol{x})-\boldsymbol{\mu}_a)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d>0. \tag{20}$$

In contrast, $\widehat{g}^{npn}$ will classify $\boldsymbol{x}$ to the second class if

$$(\widehat{f}(\boldsymbol{x})-\widehat{\boldsymbol{\mu}})^T\widehat{\boldsymbol{\beta}} \quad > \quad 0,$$

or equivalently,

$$(\widehat{f}(\boldsymbol{x})-\widehat{\boldsymbol{\mu}})^T\cdot\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0n_1} \quad > \quad 0. \tag{21}$$

We try to quantify the difference between the "oracle" in Equation (20) and the empirical classifier in Equation (21). For this, we define the empirical and population classifiers as

$$\begin{aligned} G(\boldsymbol{x},\boldsymbol{\beta},f) &:= (f(\boldsymbol{x})-\boldsymbol{\mu}_a)^T\boldsymbol{\beta},\\ \widehat{G}(\boldsymbol{x},\boldsymbol{\beta},f) &:= (f(\boldsymbol{x})-\widehat{\boldsymbol{\mu}})^T\boldsymbol{\beta}. \end{aligned}$$

With the above definitions, we have the following theorem:

Theorem 15 *When the conditions in Theorem 11 hold such that $A+B+C\to 0$, furthermore b is a positive constant chosen to satisfy*

$$sn^{-c_2\cdot b}\to 0,\quad \text{where } c_2 \text{ is a constant depending only on the choice of } \{g_j\}_{j=1}^d,$$

then we have

$$\left|\widehat{G}\left(\boldsymbol{x},\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0n_1},\widehat{f}\right)-G(\boldsymbol{x},\boldsymbol{\beta}^{**},f)\right|=O_P\left(s\beta^{**}_{\max}\sqrt{\frac{\log\log n}{n^{1-b/2}}}+sD_{\max}\lambda_n\left(\sqrt{\log n}+\Delta_{\max}\right)\right).$$

Remark 16 *Using $\frac{n^2}{n_0 n_1}\widehat{\boldsymbol{\beta}}$ instead of $\widehat{\boldsymbol{\beta}}$ is for the purpose of using the Equation* (19) *in Theorem 11.*

When the conditions in Corollary 13 hold, the rate in Theorem 15 can be written more explicitly:

Corollary 17 *When $\beta^{**}_{\max}$ is a positive constant which does not scale with (n,d,s),*

$$\frac{\log s}{c_2 \log n} < b < \frac{4\log s}{\log n},$$

and the conditions in Corollary 13 hold, we have

$$\left|\widehat{G}\left(\boldsymbol{x}, \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}, \widehat{f}\right) - G(\boldsymbol{x}, \boldsymbol{\beta}^{**}, f)\right| = O_P\left(s^2 \log n \cdot \sqrt{\frac{\log d + \log s}{n}}\right),$$

by choosing $\lambda_n \asymp s\sqrt{\frac{\log n(\log d + \log s)}{n}}$.

Remark 18 *Here we note that the conditions require that $c_2 > 1/4$, in order to give an explicit rate of the classifier estimation without including b. Theorem 15 or Corollary 17 can directly lead to the result on misclassification consistency. The key proof proceeds by showing that $f(\boldsymbol{X})$ satisfies a version of the "low noise condition" as proposed by Tsybakov (2004).*

Corollary 19 (Misclassification Consistency) *Let*

$$\mathcal{C}(g^*) := \mathbb{P}(Y \cdot \mathrm{sign}(G(\boldsymbol{X}, \boldsymbol{\beta}^{**}, f)) < 0) \quad \text{and} \quad \mathcal{C}(\widehat{g}) := \mathbb{P}(Y \cdot \mathrm{sign}(\widehat{G}(\boldsymbol{X}, \widehat{\boldsymbol{\beta}}, \widehat{f})) < 0 \,|\, \widehat{\boldsymbol{\beta}}, \widehat{f}),$$

be the misclassification errors of the Bayes classifier and the CODA classifier. Then if the conditions in Theorem 15 hold, and we have the addition assumption that

$$\log n \cdot \left(s\beta^{**}_{\max}\sqrt{\frac{\log\log n}{n^{1-b/2}}} + sD_{\max}\lambda_n\left(\sqrt{\log n} + \Delta_{\max}\right)\right) \to 0;$$

or if the conditions in Corollary 17 hold, and we have the additional assumption that

$$s^2 \log^2 n \cdot \sqrt{\frac{\log d + \log s}{n}} \to 0,$$

then we have

$$\mathbb{E}(\mathcal{C}(\widehat{g})) = \mathcal{C}(g^*) + o(1).$$

5. Experiments

In this section we investigate the empirical performance of the CODA method. We compare the following five methods:

- LS-LDA: the least square formulation for classification proposed by Mai et al. (2012);
- CODA-LS: the CODA using a similar optimization formulation as the LS-LDA;
- ROAD: the Regularized Optimal Affine Discriminant method (Fan et al., 2010);

- CODA-ROAD: the CODA using a similar optimization formulation as the LS-LDA

- SLR: the sparse logistic regression (Friedman et al., 2010).

We note that the main difference among the top four methods is that the covariance matrix $\mathbf{\Sigma}$ is estimated in different ways: the ROAD and LS-LDA both assume that data are Gaussian and use the sample covariance, which introduces estimation bias for non-Gaussian data and the resulting covariance matrix can be inconsistent to $\mathbf{\Sigma}$; in contrast, the CODA method exploits the Spearman's rho and Kendall's tau covariance matrices to estimate Σ. It enjoys a $O\left(\sqrt{\frac{\log d}{n}}\right)$ convergence rate in terms of ℓ_∞ norm. In the following, the Spearman's rho estimator is applied. The Kendall's tau estimator achieves very similar performance.

The LS-LDA, CODA-LS and SLR are implemented using the R package *glmnet* (Friedman et al., 2009). We use the augmented Lagrangian multiplier algorithm to solve ROAD and CODA-ROAD. Here in computing ROAD and CODA-ROAD, ν is set to be 10.

5.1 Synthetic Data

In the simulation studies, we randomly generate $n+1000$ class labels such that $\pi_1 = \pi_2 = 0.5$. Conditioning on the class labels, we generate d dimensional predictors $\boldsymbol{x}$ from nonparanormal distribution $NPN(\boldsymbol{\mu}_0, \mathbf{\Sigma}, f)$ and $NPN(\boldsymbol{\mu}_1, \mathbf{\Sigma}, f)$. Without loss of generality, we suppose $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\beta}^{\text{Bayes}} := \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1$ with $s := ||\boldsymbol{\beta}^{\text{Bayes}}||_0$. The data are then split to two parts: the first n data points as the training set and the next 1000 data points as the testing set. We consider twelve different simulation models. The choices of $n, d, s, \mathbf{\Sigma}, \boldsymbol{\beta}^{\text{Bayes}}$ are shown in Table 1. Here the first two schemes are sparse discriminant models with difference $\mathbf{\Sigma}$ and $\boldsymbol{\mu}_1$; Model 3 is practically sparse in the sense that its Bayes rule depends on all variables in theory but can be well approximated by sparse discriminant functions.

scheme	n	d	s	$\mathbf{\Sigma}$	$\boldsymbol{\beta}^{\text{Bayes}}$
Scheme 1	100	400	20	$\Sigma_{ij} = 0.5^{\lvert i-j \rvert}$	$0.342(1,\ldots,1,0,\ldots,0)^T$
Scheme 2	400	800	20	$\Sigma_{jj} = 1, \Sigma_{ij} = 0.5, i \neq j$	$0.176(1,\ldots,1,0,\ldots,0)^T$
Scheme 3	100	200	20	$\Sigma_{ij} = 0.6^{\lvert i-j \rvert}$	$0.198(1,\ldots,1,0.001,\ldots,0.001)^T$

Table 1: Simulation Models with different $n, d, s, \mathbf{\Sigma}$ and $\boldsymbol{\beta}^{\text{Bayes}}$ listed below.

Furthermore, we explore the effects of different transformation functions f by considering the following four types of the transformation functions:

- **Linear transformation:** $f_{linear} = (f^0, f^0, \ldots)$, where f^0 is the linear function;

- **Gaussian CDF transformation:** $f_{CDF} = (f^1, f^1, \ldots)$, where f^1 is the marginal Gaussian CDF transformation function as defined in Liu et al. (2009);

- **Power transformation:** $f_{power} = (f^2, f^2, \ldots)$, where f^2 is the marginal power transformation function as defined in Liu et al. (2009) with parameter 3;

- **Complex transformation:** $f_{complex} = (\underbrace{f^1, f^1, \ldots, f^1}_{s}, f^2, f^2, \ldots)$, where the first s variables are transformed through f^1, and the rest are transformed through f^2.

Then we obtain twelve models based on all the combinations of the three schemes of $n, d, s, \mathbf{\Sigma}, \boldsymbol{\beta}^{\text{Bayes}}$ and four transformation functions (linear, Gaussian CDF, power and complex). We note that the linear transformation f_{linear} is equivalent to no transformation. The Gaussian CDF transformation function is bounded and therefore preserves the theoretical properties of the CODA method. The power transformation function, on the other hand, is unbounded. We exploit f_{CDF}, f_{power} and $f_{complex}$ to separately illustrate how the CODA works when the assumptions in Section 4 hold, when these assumptions are mildly violated and when they are only violated for the variables X_j's with $(\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0))_j = 0$.

Figures 1 to 3 summarize the simulation results based on 3,000 replications for twelve models discussed above. Here the means of misclassification errors in percentage are plotted against the numbers of extracted features to illustrate the performance of different methods across the whole regularization paths.

To further show quantitative comparisons among different methods, we use two penalty parameter selection criteria. First, we use an oracle penalty parameter selection criterion. Let $S :=$ support$(\boldsymbol{\beta}^{\text{Bayes}})$ be the set that contains the s discriminant features. Let $\widehat{S}_\lambda$ be the set of nonzero values in the estimated parameters using the regularization parameter λ in different methods. In this way, the number of false positives at λ is defined as FP$(\lambda) :=$ the number of features in $\widehat{S}_\lambda$ but not in S. The number of false negatives at λ is defined as FN$(\lambda) :=$ the number of features in S but not in $\widehat{S}_\lambda$. We further define the false positive rate (FPR) and false negative rate (FNR) as

$$\text{FPR}(\lambda) := \text{FP}(\lambda)/(d-s), \text{ and } \text{FNR}(\lambda) := \text{FN}(\lambda)/s.$$

Let Λ be the set of all regularization parameters used to create the full path. The oracle regularization parameter λ^* is defined as

$$\lambda^* := \operatorname*{argmin}_{\lambda \in \Lambda} \{\text{FPR}(\lambda) + \text{FNR}(\lambda)\}.$$

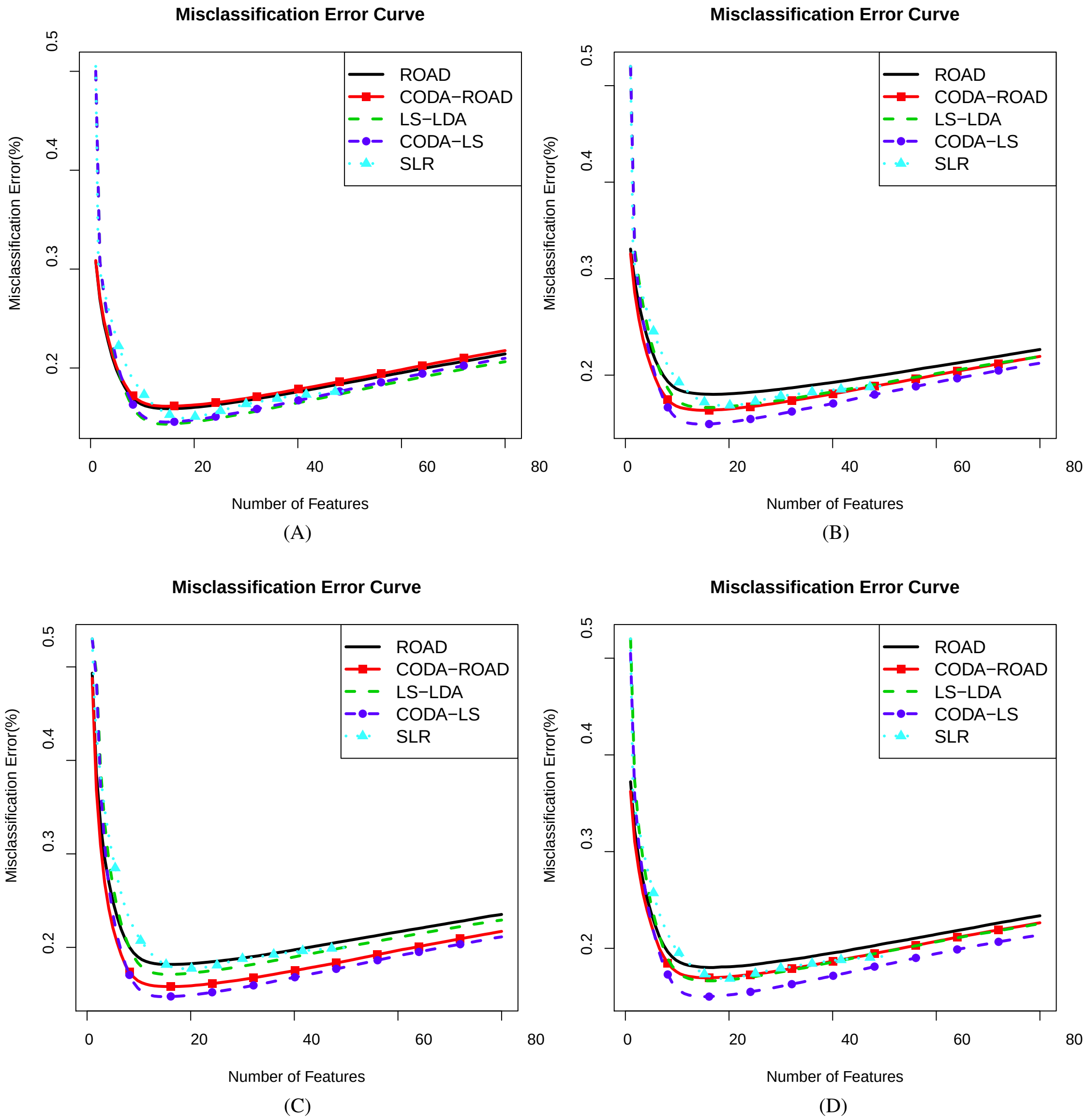
Using the oracle regularization parameter λ^*, the numerical comparisons of the five methods on the twelve models are presented in Table 2. Here Bayes is the Bayes risk and in each row the winning method is in bold. These results manifest how the five methods perform when data are either Gaussian or non-Gaussian.

Second, in practice, we propose a cross validation based approach in penalty parameter selection. In detail, for the training set, we randomly separate the data into ten folds with no overlap between each two parts. Each part has the same case and control data points. Each time we apply the above five methods to the combination of any nine folds, using a given set of regularization parameters. The parameters learned are then applied to predict the labels of the left one fold. We select the penalty parameter λ^*_{CV} to be the one that minimizes the averaged misclassification error. λ^*_{CV} is then applied to the test set. The numerical results are presented in Table 3. Here Bayes is the Bayes risk and in each row the winning method is in bold.

In the following we provide detailed analysis based on these numeric simulations.

5.1.1 NON-GAUSSIAN DATA

From Tables 2 and 3 and Figures 1 to 3, we observe that for different transformation functions f and different schemes of $\{n, d, s, \mathbf{\Sigma}, \boldsymbol{\beta}^{\text{Bayes}}\}$, CODA-ROAD and CODA-LS both significantly outperform



(A)

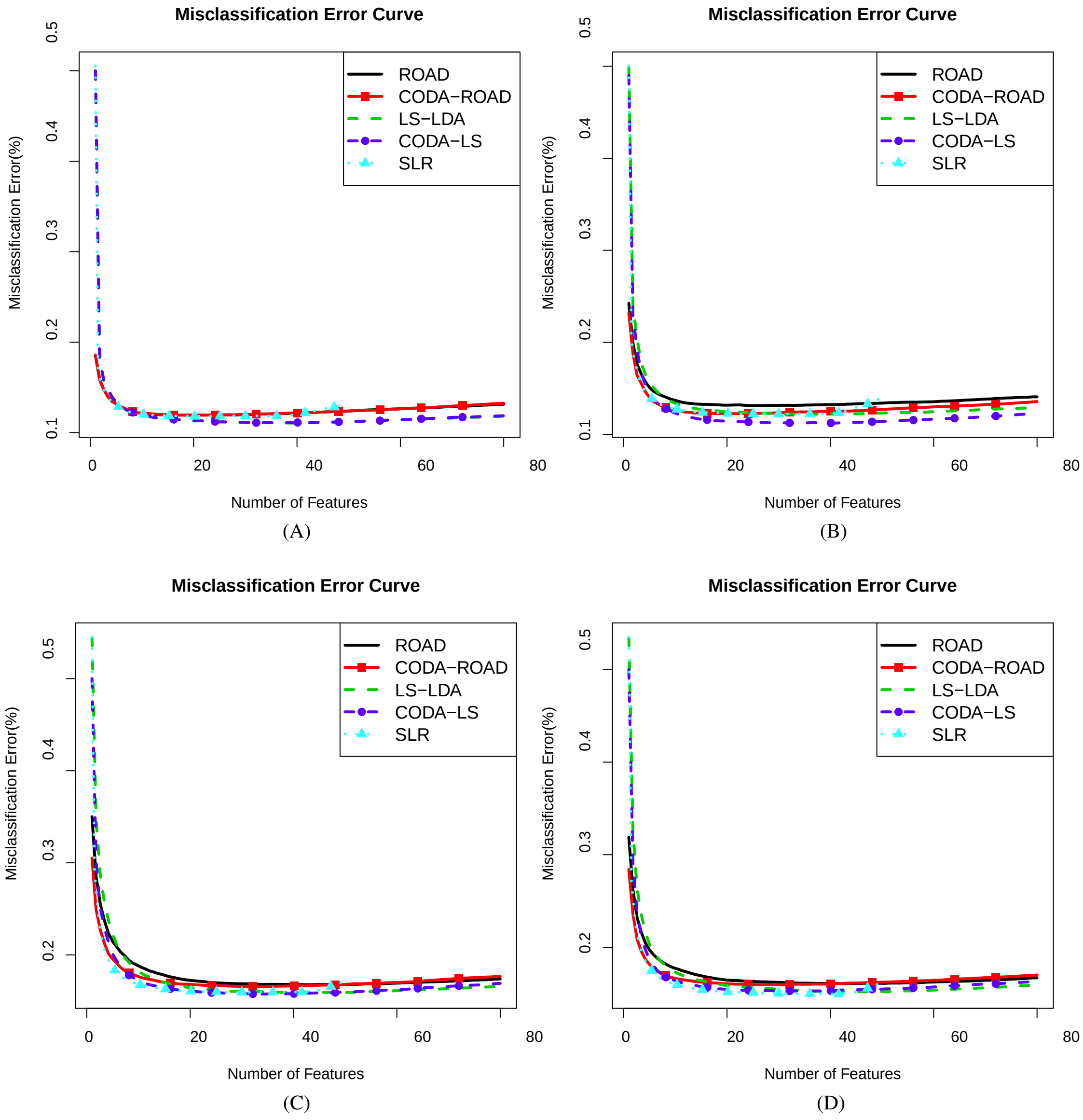
(B)

(C)

(D)

Figure 1: Misclassification error curves on Scheme 1 with four different transformation functions. (A) transformation function is f_{linear}; (B) transformation function is f_{CDF}; (C) transformation function is f_{power}; (D) transformation function is $f_{complex}$. The x-axis represents the numbers of features extracted by different methods; the y-axis represents the averaged misclassification errors in percentage of the methods on the testing data set based on 3,000 replications.

ROAD and LS-LDA, respectively. Secondly, for different transformation functions f, the differences between the two CODA methods (CODA-ROAD and CODA-LS) and their corresponding parametric methods (ROAD and LS-LDA) are comparable. This suggests that the CODA methods can beat the corresponding parametric methods when the sub-Gaussian assumptions for transformation functions



(A)

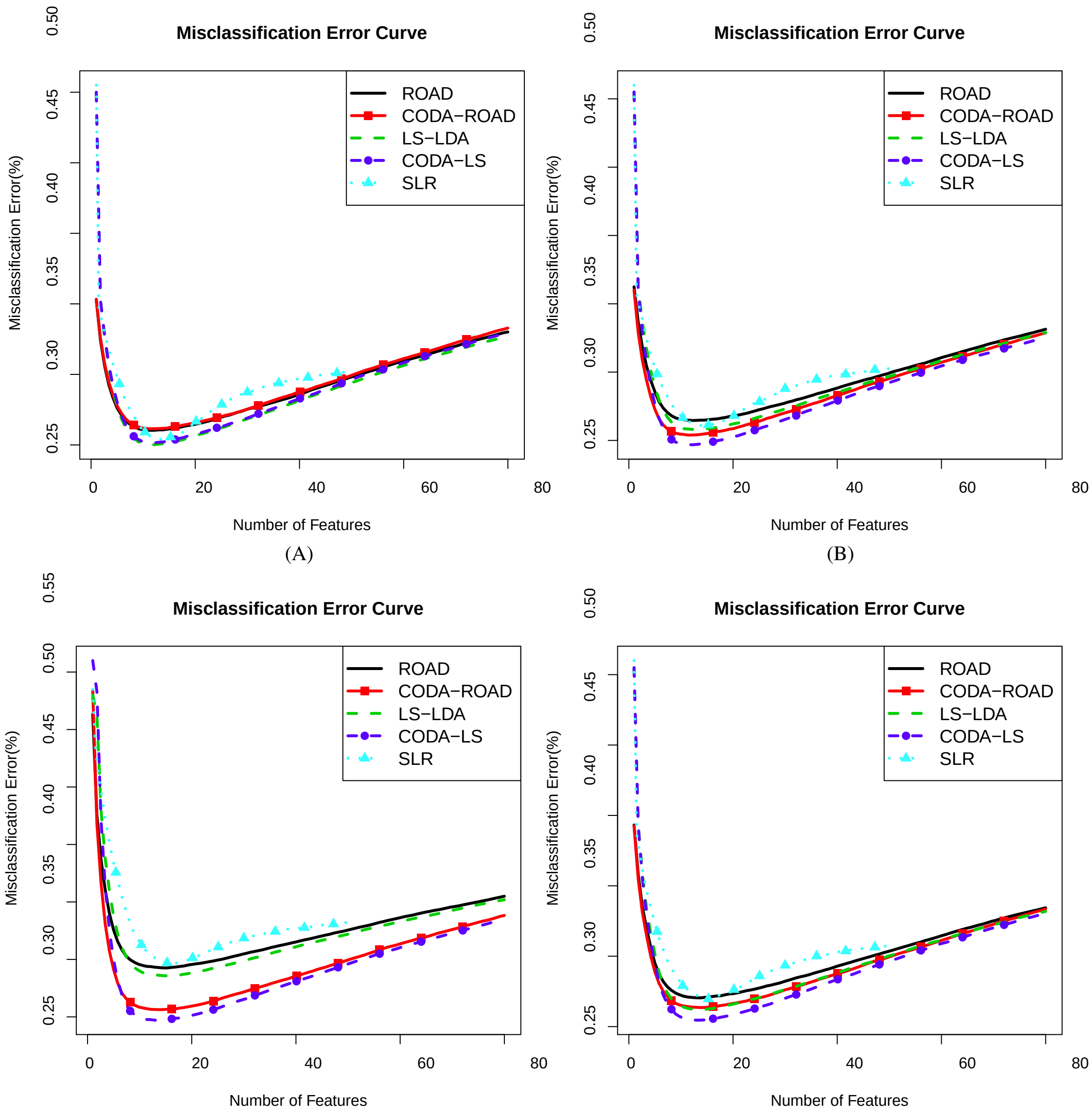
(B)

(C)

(D)

Figure 2: Misclassification error curves on Scheme 2 with four different transformation functions. (A) transformation function is f_{linear}; (B) transformation function is f_{CDF}; (C) transformation function is f_{power}; (D) transformation function is $f_{complex}$. The x-axis represents the numbers of features extracted by different methods; the y-axis represents the averaged misclassification errors in percentage of the methods on the testing data set based on 3,000 replications.

shown in Section 4 are mildly violated. Thirdly, the CODA methods CODA-LS and CODA-ROAD both outperform SLR frequently.



(A)

(B)

(C)

(D)

Figure 3: Misclassification error curves on Scheme 3 with four different transformation functions. (A) transformation function is f_{linear}; (B) transformation function is f_{CDF}; (C) transformation function is f_{power}; (D) transformation function is $f_{complex}$. The x-axis represents the numbers of features extracted by different methods; the y-axis represents the averaged misclassification errors in percentage of the methods on the testing data set based on 3,000 replications.

5.1.2 GAUSSIAN DATA

From Tables 2 and 3 and Figures 1 to 3, we observe that when the transformation function is f_{linear}, there is no significant differences between CODA-ROAD and ROAD, and between CODA-LS and LS-

Scheme	f	Bayes(%)	ROAD	CODA-ROAD	LS-LDA	CODA-LS	SLR
Scheme 1	f_{linear}	10.00	16.64(0.81)	16.90(0.84)	**15.09(0.52)**	15.29(0.53)	15.40(0.49)
	f_{CDF}	10.00	18.57(0.80)	17.17(0.84)	17.26(0.52)	**15.66(0.53)**	17.10(0.46)
	f_{power}	10.00	18.80(0.81)	16.51(0.86)	17.76(0.52)	**15.45(0.56)**	17.99(0.52)
	$f_{complex}$	10.00	18.68(0.84)	17.12(0.87)	17.40(0.54)	**15.78(0.53)**	17.26(0.57)
Scheme 2	f_{linear}	10.00	12.28(0.41)	12.34(0.38)	**11.46(0.28)**	11.48(0.28)	12.19(0.31)
	f_{cdf}	10.00	13.56(0.65)	12.83(0.72)	12.95(1.00)	**11.99(0.99)**	12.84(0.30)
	f_{power}	10.00	17.85(0.86)	17.38(0.65)	17.10(0.73)	16.65(0.50)	**16.50(0.33)**
	$f_{complex}$	10.00	16.89(1.39)	16.89(0.43)	17.20(2.43)	**16.77(0.33)**	16.91(0.47)
Scheme 3	f_{linear}	20.00	26.65(0.78)	26.69(0.77)	**25.59(0.63)**	25.70(0.64)	25.97(0.58)
	f_{cdf}	20.00	26.97(0.70)	26.03(0.78)	26.16(0.58)	**25.18(0.64)**	26.41(0.58)
	f_{power}	20.00	29.78(0.72)	26.07(0.87)	29.03(0.60)	**25.14(0.70)**	29.34(0.61)
	$f_{complex}$	20.00	27.54(0.71)	26.78(0.73)	26.70(0.62)	**25.87(0.59)**	26.87(0.57)

Table 2: Quantitative comparisons on different models with linear, Gaussian CDF, power, and complex transformations using the oracle penalty parameter selection criterion. The methods compared here are ROAD,CODA-ROAD,LS-LDA,CODA-LS and SLR. Here Bayes is the Bayes risk and the winning methods are in bold. The means of misclassification errors in percentage with their standard deviations in parentheses are presented. The results are based on 3,000 replications.

LDA. This suggests that the CODA methods can be an alternative choice besides the Gaussian-based high dimensional classification methods.

In summary, we observe that the CODA methods (CODA-LS in particular) have very good overall performance. The simulation results suggest that they can be an alternative choices besides their corresponding parametric methods. And the results also show that in our experiments the CODA methods can outperform their corresponding parametric methods when the sub-Gaussian assumptions for transformation functions are mildly violated.

5.2 Large-scale Genomic Data

In this section we investigate the performance of the CODA methods compared with the others using one of the largest microarray data sets (McCall et al., 2010). In summary, we collect in all 13,182 publicly available microarray samples from Affymetrixs HGU133a platform. The raw data contain 20,248 probes and 13,182 samples belonging to 2,711 tissue types (e.g., lung cancers, prostate cancer, brain tumor etc.). There are at most 1599 samples and at least 1 sample belonging to each tissue type. We merge the probes corresponding to the same gene. There are remaining 12,713 genes and 13,182 samples. The main purpose of this experiment is to compare the performance of different methods in classifying tissues.

We adopt the same idea of data preprocessing as in Liu et al. (2012). In particular, we remove the batch effect by applying the surrogate variable analysis proposed by Leek and Storey (2007). There are, accordingly, 12,713 genes left and the data matrix we are focusing is $12,713 \times 13,182$.

We then explore several tissue types with the largest sample size:

- Breast tumor, which has 1599 samples;

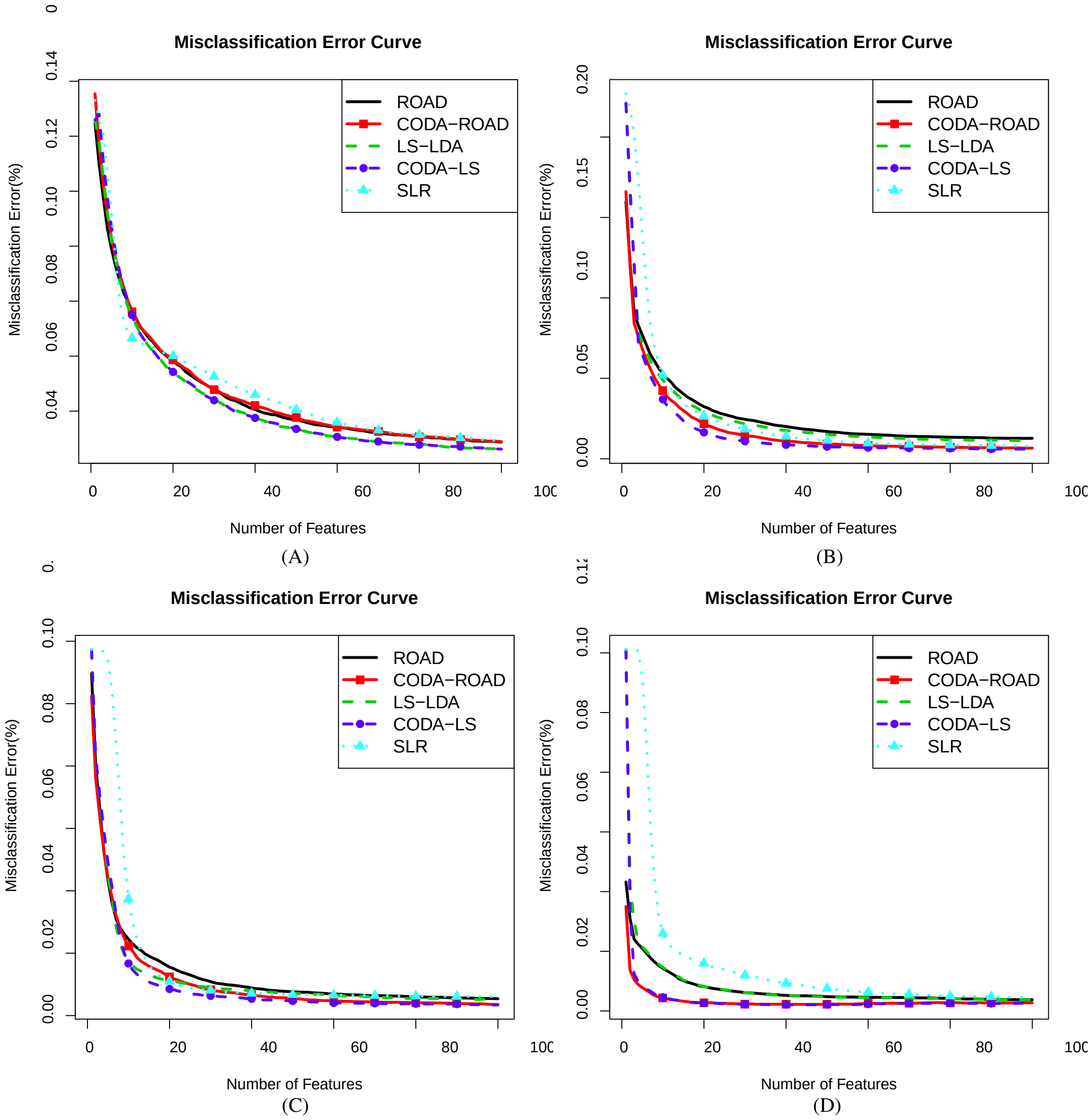
Scheme	f	Bayes(%)	ROAD	CODA-ROAD	LS-LDA	CODA-LS	SLR
Scheme 1	f_{linear}	10.00	16.86(0.77)	16.99(0.94)	**15.31(0.54)**	15.41(0.49)	15.44(0.51)
	f_{CDF}	10.00	18.86(0.83)	17.19(0.79)	17.39(0.68)	**16.16(0.63)**	17.42(0.64)
	f_{power}	10.00	19.13(0.91)	16.84(0.90)	17.91(0.61)	**15.92(0.66)**	18.13(0.62)
	$f_{complex}$	10.00	18.81(0.93)	17.73(0.89)	17.42(0.63)	**15.89(0.62)**	17.94(0.66)
Scheme 2	f_{linear}	10.00	12.58(0.52)	12.59(0.47)	**11.59(0.33)**	11.70(0.38)	12.19(0.29)
	f_{cdf}	10.00	13.97(0.74)	12.86(0.76)	13.36(1.05)	**12.08(1.03)**	13.03(0.32)
	f_{power}	10.00	18.23(0.76)	17.48(0.73)	17.11(0.77)	**16.85(0.59)**	16.86(0.37)
	$f_{complex}$	10.00	16.96(1.59)	16.74(0.61)	17.47(1.99)	**16.80(0.49)**	17.06(0.55)
Scheme 3	f_{linear}	20.00	26.83(0.88)	27.23(0.77)	**25.62(0.64)**	25.74(0.71)	26.21(0.63)
	f_{cdf}	20.00	27.13(0.81)	26.21(0.85)	26.76(0.64)	**25.23(0.61)**	26.43(0.69)
	f_{power}	20.00	30.17(0.85)	26.79(1.00)	29.03(0.73)	**25.15(0.78)**	29.85(0.63)
	$f_{complex}$	20.00	28.43(0.91)	26.82(0.77)	26.74(0.60)	**25.88(0.68)**	27.27(0.71)

Table 3: Quantitative comparisons on different models with linear, Gaussian CDF, power, and complex transformations using the cross validation based penalty parameter selection criterion. The methods compared here are ROAD,CODA-ROAD,LS-LDA,CODA-LS and SLR. Here Bayes is the Bayes risk and the winning methods are in bold. The means of misclassification errors in percentage with their standard deviations in parentheses are presented. The results are based on 3,000 replications.

- B cell lymphoma, which has 213 samples;
- Prostate tumor, which has 148 samples;
- Wilms tumor, which has 143 samples.

Different tissues have been believed to be associated with different sets of genes and microarray data have been heavily used to classify tissue types. See for example, Hans et al. (2004), Wang et al. (2008) and Huang and Chang (2007), among others. For each tissue type listed above, our target is to classify it from all the other tissue types. To this end, each time we randomly split the whole data to three parts: (i) the training set with 200 samples (equal size of case and control); (ii) the testing set with 1000 samples; (iii) the rest. We then run ROAD,CODA-ROAD,LS-LDA,CODA-LS on the training set and applying the learned parameters on the testing set. We repeat this for 1,000 times. The averaged misclassification errors in percentage versus the numbers of extracted features are illustrated in Figure 4. Quantitative results, with penalty parameter selected using the cross validation criterion, are presented in Table 4.

It can be observed that CODA-ROAD and CODA-LS have the best overall performance. Some biological discoveries have also been verified in this process. For example, the MYC gene has been discovered to be relevant to the b cell lymphoma (Lovec et al., 1994; Smith and Wickstrom, 1998) and has recently been found to be associated with the Wilms tumor (Ji et al., 2011). This gene is also constantly selected by the CODA methods in classifying b cell lymphoma and Wilms tumor with the rest.



(A)

(B)

(C)

(D)

Figure 4: Misclassification error curves on the GPL96 data set. (A) Breast tumor; (B) B cell lymphoma; (C) Prostate tumor; (D) Wilms tumor. The *x*-axis represents the numbers of features extracted by different methods; the *y*-axis represents the averaged misclassification errors in percentage of the methods on the testing data set based on 1,000 replications.

5.3 Brain Imaging Data

In this section we investigate the performance of several methods on a brain imaging data set, the ADHD 200 data set (Eloyan et al., 2012). The ADHD 200 data set is a landmark study compiling over 1,000 functional and structural scans including subjects with and without attention deficit hyperactive disorder (ADHD). The current releases data are from 776 subjects: 491 controls and 285 children diagnosed with ADHD. Each has structural blood oxygen level dependent (BOLD)

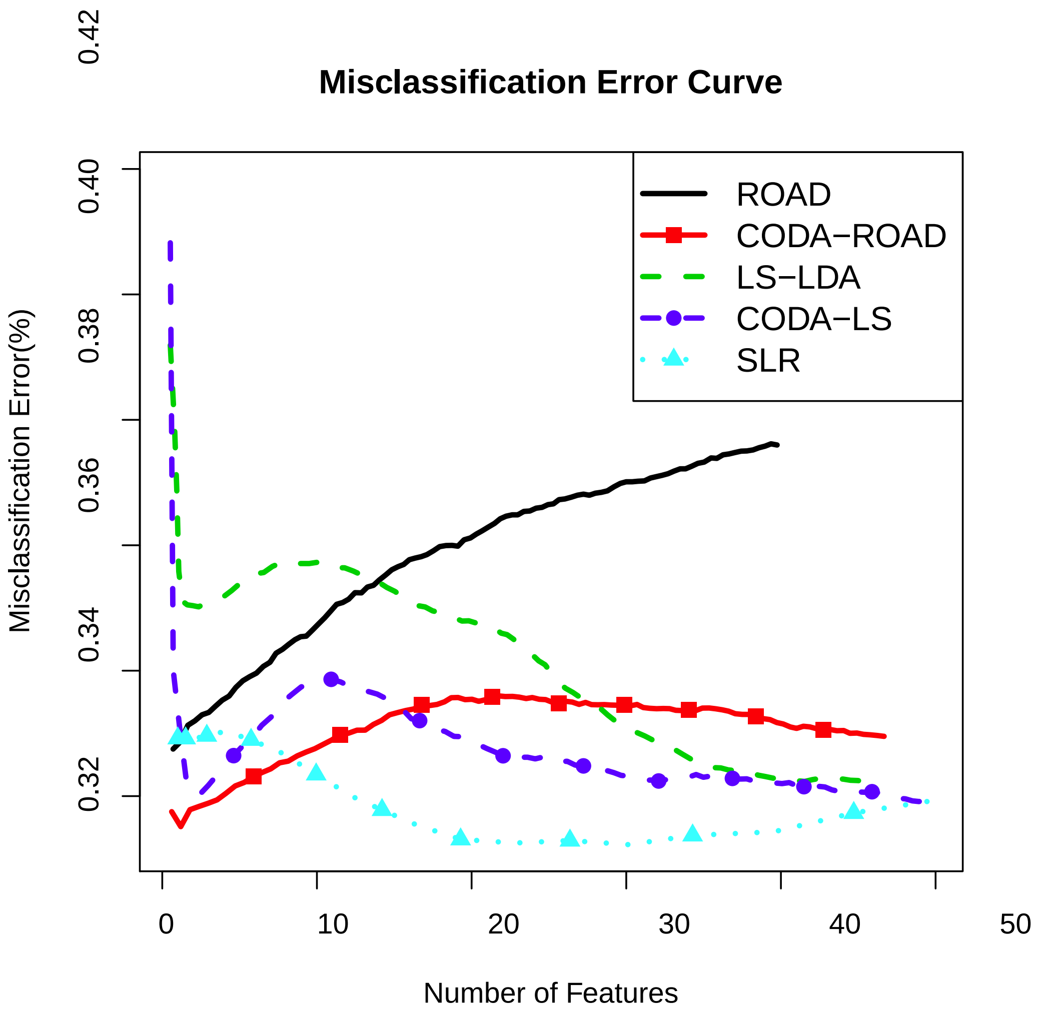


Figure 5: Misclassification error curves on the ADHD data set. The *x*-axis represents the numbers of features extracted by different methods; the *y*-axis represents the averaged misclassification errors in percentage of the methods on the testing data set based on 1,000 replications.

functional MRI scans. The data also include demographic variables as predictors. These include age, IQ, gender and handedness. We refer to Eloyan et al. (2012) for detailed data preprocessing procedures.

We construct our predictors by extracting voxels that broadly cover major functional regions of the cerebral cortex and cerebellum following Power et al. (2011). We also combine the information of the demographic variables, resulting to the final data matrix we will use with the dimension 268×776. The target is to differentiate the subjects with ADHD from those without ADHD.

To evaluate the performance of different methods, each time we randomly sample 155 data points unrepeatedly from the whole data. We then gather them together as the training set. The

Data	ROAD(%)	CODA-ROAD	LS-LDA	CODA-LS	SLR
Genomic (A)	0.29(0.17)	0.29(0.17)	0.26(0.16)	**0.25(0.15)**	0.29(0.18)
Genomic (B)	1.31(0.26)	0.69(0.15)	1.16(0.20)	**0.63(0.13)**	0.82(0.18)
Genomic (C)	0.56(0.13)	0.39(0.11)	0.55(0.15)	**0.37(0.12)**	0.62(0.17)
Genomic (D)	0.38(0.16)	0.23(0.08)	0.38(0.09)	**0.22(0.10)**	0.48(0.12)
ADHD	33.20(0.26)	31.89(0.27)	32.66(0.24)	32.25(0.24)	**31.73(0.21)**

Table 4: Quantitative comparisons on genomic and brain imaging data using the cross validation based penalty parameter selection criterion. The methods compared here are ROAD,CODA-ROAD,LS-LDA,CODA-LS and SLR. Here the winning methods are in bold. The means of misclassification errors in percentage with their standard deviations in parentheses are presented. Here "Genomic (A)" to "Genomic (D)" denote the breast tumor, b cell lymphoma, prostate tumor and Wilms tumor, 'ADHD' denotes the results in brain imaging data analysis.

rest are left as the testing set. We then run ROAD,CODA-ROAD,LS-LDA,CODA-LS on the training set and applying the learned parameters on the testing set. This is repeated for 1,000 times and the averaged misclassification errors in percentage versus the numbers of extracted features are illustrated in Figure 5. Quantitative results, with penalty parameter selected using the cross validation criterion, are presented in Table 4. In this data set, SLR performs the best, followed by CODA-LS and CODA-ROAD. Moreover, the CODA methods beat their corresponding parametric methods in this experiment. It can be observed in Table 4 that there is no significant difference between SLR and CODA-ROAD.

6. Discussions

In this paper a high dimensional classification method named the CODA (Copula Discriminant Analysis) is proposed. The main contributions of this paper include: (i) We relax the normality assumption of linear discriminant analysis through the nonparanormal (or Gaussian copula) modeling; (ii) We use the nonparanormal SKEPTIC procedure proposed by Liu et al. (2012) to efficiently estimate the model parameters; (iii) We build a connection of the ROAD and lasso and provide an approach to solve the problem that the rank-based covariance matrix may not be positive semidefinite; (iv) We provide sufficient conditions to secure the variable selection consistency with the parametric rate, and the expected misclassification error is consistent to the Bayes risk; (v) Careful experiments on synthetic and real data sets are conducted to support the theoretical claims.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their comments and suggestions which have led to valuable improvements of this paper. We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-1116730).

Appendix A. Proof of Theorem 2

To show that Theorem 2 holds, we need to provide several important lemmas using results of large deviation and empirical process. First, define $\phi(\cdot)$ and $\Phi(\cdot)$ to be the probability density function and cumulative distribution function of the standard Gaussian distribution. For any $x \in \mathbb{R}$, we denote by $x^+ = x \cdot I(x > 0)$ and $x^- = -x \cdot I(x < 0)$. By definition, $f_j(t) = \Phi^{-1}(F_j(t))$ and $g_j(u) := f_j^{-1}(u) = F_j^{-1}(\Phi(u))$. Here for notation simplicity, let $\widetilde{F}_j(t)$ and $\widehat{F}_j(t)$ be the abbreviations of $\widetilde{F}_j(t; 1/(2n), \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\widehat{F}_j(t; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ defined in Section 2.2.

The following lemma quantifies the region of the value $\widetilde{F}_j$ in I_n and shows that $\widetilde{F}_j$ is not truncated in I_n almost surely.

Lemma 20 (Liu et al., 2012) *We have for large enough n,*

$$\mathbb{P}\left(\frac{1}{n} \le \widetilde{F}_j(t) \le 1 - \frac{1}{n}, \text{ for all } t \in I_n\right) = 1.$$

With Lemma 20, we can now prove the following key lemma, which provides an uniform convergence rate on $\widetilde{F}_j(t)$ to $F_j(t)$. This result is mentioned in Liu et al. (2012), but without proof.

Lemma 21 *Consider a sequence of sub-intervals $[L_n^{(j)}, U_n^{(j)}]$ with both $L_n^{(j)} := g_j(\sqrt{\alpha \log n})$ and $U_n^{(j)} := g_j(\sqrt{\beta \log n}) \uparrow \infty$, then for any $0 < \alpha < \beta < 2$, for large enough n,*

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}} \sup_{L_n^{(j)} < t < U_n^{(j)}} \left| \frac{\widetilde{F}_j(t) - F_j(t)}{\sqrt{F_j(t)(1 - F_j(t))}} \right| = C \text{ a.s.},$$

where $0 < C < 2\sqrt{2}$ is a constant.

Proof By Lemma 20, for large enough n,

$$\widetilde{F}_j(t) = \widehat{F}_j(t), \quad \text{for all } t \in I_n, \quad \text{almost surely.} \tag{22}$$

Given $\xi_1, \ldots, \xi_n$ a series of i.i.d random variables from Unif$(0,1)$ and define $\mathbb{G}_n(t) := \frac{1}{n}\sum I(\xi_i < t)$, it is easy to see that

$$\widehat{F}_j(t) = \mathbb{G}_n(F_j(t)) \text{ a.s..} \tag{23}$$

Define

$$\mathbb{U}_n(u) := \frac{\mathbb{G}_n(u) - u}{\sqrt{u(1-u)}}.$$

By Equation (22) and (23), it is easy to see that

$$\mathbb{U}_n(F_j(t)) = \frac{\widetilde{F}_j(t) - F_j(t)}{\sqrt{F_j(t)(1 - F_j(t))}} \quad \text{a.s..} \tag{24}$$

By Theorem 1 in Section 2 (Chapter 16) of Shorack and Wellner (1986), we know that

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}} \sup_{0 \le u \le 1/2} (\mathbb{U}_n(u))^- = \sqrt{2} \text{ a.s..} \tag{25}$$

And by Theorem 2 in Section 3 (Chapter 16) of Shorack and Wellner (1986), for $a_n \to 0$ such that $\frac{\log\log(1/a_n)}{\log\log n} \to 1$, we have

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}} \sup_{a_n \le u \le 1/2} (\mathbb{U}_n(u))^+ = \sqrt{2} \text{ a.s..} \tag{26}$$

Combining Equation (25) and (26) together, we have

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}} \sup_{a_n \le u \le 1/2} |\mathbb{U}_n(u)| \le 2\sqrt{2} \text{ a.s..} \tag{27}$$

Furthermore, for any $u \in [0,1]$,

$$\mathbb{G}_n(1-u) = \frac{1}{n}\sum I(\xi_i < 1-u) = \frac{1}{n}\sum I(1-\xi_i \ge u) = 1 - \mathbb{G}_n(u),$$

which implies that

$$\mathbb{U}_n(1-u) = -\mathbb{U}_n(u).$$

Therefore, by Equation (27), for $a_n \downarrow 0$ such that $\frac{\log\log(1/a_n)}{\log\log n} \to 1$, we have

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}} \sup_{1/2 \le u \le 1-a_n} |\mathbb{U}_n(u)| \le 2\sqrt{2} \quad \text{a.s..} \tag{28}$$

Finally, choosing $a_n = 1 - F_j(U_n^{(j)})$, we have

$$a_n = 1 - \Phi(\sqrt{\beta \log n}) \approx n^{-\beta/2} \quad \text{and} \quad \frac{\log\log n^{\beta/2}}{\log\log n} \to 1,$$

so taking $a_n = 1 - F_j(U_n^{(j)})$ into Equation (28), the result follows by using Equation (24). ■

Proof [Proof of the Theorem 2] Finally, we prove the Theorem 2. By symmetry, we only need to conduct analysis on a sub-interval of $I_n^s \subset I_n$:

$$I_n^s := \left[g_j(0), g_j\left(\sqrt{2(1-\gamma)\log n}\right)\right].$$

We define a series $0 < \alpha < 1 < \beta_1 < \beta_2 < \ldots < \beta_\kappa$ and denote by $\beta_0 := \alpha$,

$$I_{0n} := \left[g_j(0), g_j(\sqrt{\alpha \log n})\right],$$

$$I_{1n} := \left[g_j(\sqrt{\alpha\log n}), g_j(\sqrt{\beta_1 \log n})\right], \ \ldots \ , I_{\kappa n} := \left[g_j(\sqrt{\beta_{\kappa-1}\log n}), g_j(\sqrt{\beta_\kappa \log n})\right].$$

For $i = 0, \ldots, \kappa$, we can rewrite

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

By the mean value theorem, for some ξ_n such that

$$\xi_n \in \left[\min\{\widetilde{F}_j(g_j(\sqrt{\beta_{i-1}\log n})), F_j(g_j(\sqrt{\beta_{i-1}\log n}))\}, \max\{\widetilde{F}_j(g_j(\sqrt{\beta_i \log n})), F_j(g_j(\sqrt{\beta_i \log n}))\}\right],$$

we have

$$\sup_{t\in I_{in}} \left|\Phi^{-1}\left(\widetilde{F}_j(t)\right) - \Phi^{-1}\left(F_j(t)\right)\right| = \sup_{t\in I_{in}} \left|(\Phi^{-1})'(\xi_n)\left(\widetilde{F}_j(t) - F_j(t)\right)\right|. \tag{29}$$

Because Φ and Φ^{-1} are strictly increasing function, for large enough n, we have

$$(\Phi^{-1})'(\xi_n) \leq (\Phi^{-1})'\left(\max\left\{F_j\left(g_j\left(\sqrt{\beta_i \log n}\right)\right), \widetilde{F}_j\left(g_j\left(\sqrt{\beta_i \log n}\right)\right)\right\}\right). \tag{30}$$

From Lemma 21, for large enough n, we have

$$\widetilde{F}_j(t) \leq F_j(t) + 4\sqrt{\frac{\log\log n}{n}} \cdot \sqrt{1 - F_j(t)}.$$

In special, using the fact that $F_j(g_j(t)) = \Phi(t)$, we have

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

The last inequality holds given Equation (B.4) to (B.12) in Liu et al. (2012).

Therefore,

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

Returning to Equation (30), we have

$$(\Phi^{-1})'(\xi_n) \leq C(\Phi^{-1})'(F_j(g_j(\sqrt{\beta_i \log n}))) = \frac{C}{\phi(\sqrt{\beta_i \log n})} \leq c_1 n^{\beta_i/2}, \tag{31}$$

where $C > 1$ and c_1 are generic constants. Specifically, when $i = 0$, using the Dvoretzky-Kiefer-Wolfowitz inequality (Massart, 1990; Dvoretzky et al., 1956), from Equation (29), we have

$$\sup_{t\in I_{0n}} \left|\Phi^{-1}(\widetilde{F}_j(t)) - \Phi^{-1}(F_j(t))\right| = O_P\left(\sqrt{\frac{\log\log n}{n^{1-\alpha}}}\right).$$

For any $i \in \{1, \ldots, \kappa\}$, using Lemma 21, for large enough n,

$$\begin{aligned} \sup_{t\in I_{in}} \left|\widetilde{F}_j(t) - F_j(t)\right| &= O_P\left(\sqrt{\frac{\log\log n}{n}} \cdot \sqrt{1 - F_j\left(g_j(\sqrt{\beta_{i-1}\log n})\right)}\right) \\ &= O_P\left(\sqrt{\frac{\log\log n}{n}} \cdot \sqrt{\frac{n^{-\beta_{i-1}/2}}{\sqrt{\alpha \log n}}}\right) \\ &= O_P\left(\sqrt{\frac{\log\log n}{n^{\beta_{i-1}/2+1}}}\right). \end{aligned} \tag{32}$$

Again, using Equation (31), we have

$$(\Phi^{-1})'(\xi_n) \leq C(\Phi^{-1})'(F_j(g_j(\sqrt{\beta_i \log n}))) = \frac{C}{\phi(\sqrt{\beta_i \log n})} \leq c_1 n^{\beta_i/2},$$

and applying Equation (32), we have

$$\sup_{t \in I_{in}} \left| \Phi^{-1}(\widetilde{F}_j(t)) - \Phi^{-1}(F_j(t)) \right| = O_P\left(\sqrt{\frac{\log\log n}{n^{1+\beta_{i-1}/2-\beta_i}}} \right).$$

Chaining the inequalities together and choose

$$\beta_i = (2-(1/2)^i)(1-\gamma), \quad i \in \{0,1,\ldots,\kappa\},$$

we have for any $i \in \{0,1,\ldots,\kappa\}$,

$$\begin{aligned} 1-\alpha &= 1-(1-\gamma) = \gamma \text{ and} \\ 1+\beta_{i-1}/2-\beta_i &= 1+\left(1-\frac{1}{2^i}\right)(1-\gamma)-\left(2-\frac{1}{2^i}\right)(1-\gamma) = \gamma. \end{aligned}$$

And therefore, we have

$$\sup_{I_{0n}\cup\ldots\cup I_{\kappa n}} \left| \Phi^{-1}(\widetilde{F}_j(t)) - \Phi^{-1}(F_j(t)) \right| = O_P\left(\sqrt{\frac{\log\log n}{n^{\gamma}}} \right),$$

while

$$I_{0n}\cup\ldots\cup I_{\kappa n} = \left[g_j(0), g_j\left(\sqrt{(2-2^{-\kappa})(1-\gamma)} \right) \right].$$

Taking $\kappa \uparrow \infty$, we have the result. ■

Appendix B. Proof of Theorem 11

To prove Theorem 11, we need the following three key lemmas. Lemma 22 claims that, under certain constraints on the transformation functions, there exist fast rates for the sample means and projected Spearman's rho/Kendall's tau covariance matrices converging to the population means and covariance matrix for the nonparanormal. Lemma 23 provides exponential inequalities for two estimators we are most interested in in analyzing the theoretical performance of the CODA. Lemma 25 claims that $\widetilde{\mathbf{\Sigma}}_{SS}$ is invertible with high probability.

Lemma 22 *For any $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ i.i.d drawn from $\boldsymbol{X}$, where $\boldsymbol{X} \sim NPN(\boldsymbol{\mu},\mathbf{\Sigma},f)$, $0 < 1/\sigma_{\max} < \min_j\{\sigma_j\} < \max_j\{\sigma_j\} < \sigma_{\max} < \infty$, $\max_j |\mu_j| \leq \sigma_{\max}$ and $g := f^{-1}$ satisfies $g_j^2 \in TF(K)$, $j = 1,\ldots d$ for some constant $K < \infty$, we have for any $t \geq 32\pi\sqrt{\frac{\log d}{n \log 2}}$,*

$$\begin{aligned} \mathbb{P}\left(|\widetilde{S}_{jk} - \Sigma_{jk}| > t \right) &\leq 2\exp(-c_0' n t^2), & (33) \\ \mathbb{P}\left(|\widehat{\mu}_j - \mu_j| > t \right) &\leq 2\exp(-c_1' n t^2), & (34) \end{aligned}$$

where c_0' and c_1' are two constants only depending on the choice of $\{g_j\}_{j=1}^d$.

Proof Because $\sigma_{\max}$ is a constant which does not scale with (n,d,s), without loss of generality we can assume that $K \geq 1$, $\boldsymbol{\mu} = \mathbf{0}$ and $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{1}$. The key is to prove that the high order moments of each X_j and X_j^2 will not grow very fast.

We only focus on $j = 1$ and the results can be generalized to $j = 2,3,\ldots,d$. Define $Z := f_1(X_1) \sim N(0,1)$. We have $\forall\ m \in \mathbb{Z}^+$, because $g_1^2 \in \mathrm{TF}(K)$ for some constant K,

$$\mathbb{E}|X_1^2|^m = \mathbb{E}|g_1^2(Z)|^m \leq \frac{m!}{2}K^m.$$

Therefore, by Lemma 5.7 of van de Geer (2000), $\widehat{\sigma}_1^2$ goes to σ_1^2 exponentially fast. To show that the Equation (34) holds, we have

$$\begin{aligned}
\mathbb{E}|X_1|^m &= \mathbb{E}|X_1^2|^{m/2} \leq \frac{(m/2)!}{2}K^{m/2} < \frac{m!}{2}K^m, \quad \text{if } m \text{ is even},\\
\mathbb{E}|X_1|^m &\leq 1 + \mathbb{E}|X_1|^m I(|X_1| \geq 1) \leq 1 + \mathbb{E}(|X_1|^{m+1} I(|X_1| \geq 1))\\
&\leq 1 + \mathbb{E}|X_1|^{m+1} \leq 1 + \frac{\left(\frac{m+1}{2}\right)!}{2}K^{\frac{m+1}{2}} < \frac{m!}{2}(2K+2)^m, \quad \text{if } m \text{ is odd}.
\end{aligned}$$

Therefore, again by Lemma 5.7 of van de Geer (2000), $\widehat{\mu}_1$ goes to μ_1 exponentially fast.

Similarly we can prove that $\mathbb{P}(|\widehat{\sigma}_j - \sigma_j| \geq t) = O(\exp(-cnt^2))$ for the generic constant c. Therefore, to prove that Equation (33) holds, the only thing left is to show that combining $\widehat{\sigma}_j, \widehat{\sigma}_k$ with $\widetilde{R}_{jk}$ does not change the rate. Actually, suppose that

$$\begin{aligned}
\mathbb{P}\left(\left|\widehat{\sigma}_j - \sigma_j\right| > \varepsilon\right) &\leq \eta_1(n,\varepsilon),\\
\mathbb{P}\left(\left|\widetilde{R}_{jk} - \Sigma_{jk}^0\right| > \varepsilon\right) &\leq \eta_2(n,\varepsilon),
\end{aligned}$$

then we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\widetilde{S}_{jk} - \Sigma_{jk}\right| > \varepsilon\right)\\
= &\ \mathbb{P}\left(\left|(\widehat{\sigma}_j\widehat{\sigma}_k - \sigma_j\sigma_k)\widetilde{R}_{jk} + \sigma_j\sigma_k\left(\widetilde{R}_{jk} - \Sigma_{jk}^0\right)\right| > \varepsilon\right)\\
\leq &\ \mathbb{P}\left(\left|(\widehat{\sigma}_j\widehat{\sigma}_k - \sigma_j\sigma_k)\widetilde{R}_{jk}\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\sigma_j\sigma_k\left(\widetilde{R}_{jk} - \Sigma_{jk}^0\right)\right| > \frac{\varepsilon}{2}\right)\\
\leq &\ \mathbb{P}\left(\left|\widehat{\sigma}_j\widehat{\sigma}_k - \sigma_j\sigma_k\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\widetilde{R}_{jk} - \Sigma_{jk}^0\right| > \frac{\varepsilon}{2\sigma_{\max}^2}\right)\\
\leq &\ \mathbb{P}\left(\left|(\widehat{\sigma}_j - \sigma_j)(\widehat{\sigma}_k - \sigma_k) + \sigma_j(\widehat{\sigma}_k - \sigma_k) + \sigma_k(\widehat{\sigma}_j - \sigma_j)\right| > \frac{\varepsilon}{2}\right) + \eta_2\left(n, \frac{\varepsilon}{2\sigma_{\max}^2}\right)\\
\leq &\ \mathbb{P}\left(\left|(\widehat{\sigma}_j - \sigma_j)(\widehat{\sigma}_k - \sigma_k)\right| > \frac{\varepsilon}{6}\right) + \mathbb{P}\left(\left|\sigma_j(\widehat{\sigma}_k - \sigma_k)\right| > \frac{\varepsilon}{6}\right)\\
&+\mathbb{P}\left(\left|\sigma_k(\widehat{\sigma}_j - \sigma_j)\right| > \frac{\varepsilon}{6}\right) + \eta_2\left(n, \frac{\varepsilon}{2\sigma_{max}^2}\right)\\
\leq &\ \mathbb{P}\left(\left|\widehat{\sigma}_j - \sigma_j\right| > \sqrt{\frac{\varepsilon}{6}}\right) + \mathbb{P}\left(\left|\widehat{\sigma}_k - \sigma_k\right| > \sqrt{\frac{\varepsilon}{6}}\right)\\
&+\mathbb{P}\left(\left|\widehat{\sigma}_k - \sigma_k\right| > \frac{\varepsilon}{6\sigma_{max}}\right) + \mathbb{P}\left(\left|\widehat{\sigma}_j - \sigma_j\right| > \frac{\varepsilon}{6\sigma_{max}}\right) + \eta_2\left(n, \frac{\varepsilon}{2\sigma_{max}^2}\right)\\
\leq &\ 2\eta_1\left(n, \sqrt{\frac{\varepsilon}{6}}\right) + 2\eta_1\left(n, \frac{\varepsilon}{6\sigma_{max}}\right) + \eta_2\left(n, \frac{\varepsilon}{2\sigma_{max}^2}\right).
\end{aligned}$$

Due to Lemma 5, we have for all $t \geq 32\pi\sqrt{\frac{\log d}{n \log 2}}$

$$\mathbb{P}(|\widetilde{R}_{jk} - \Sigma^0_{jk}| > t) \leq 2\exp(-cnt^2),$$

for some generic constant c. It means that η_1 and η_2 are both of parametric exponential decay rate. we complete the proof. ■

Lemma 23 *If n_0 and n_1 are deterministic, then there exists a constant c_0 such that for any $\varepsilon \geq 32\pi\sqrt{\frac{\log d}{(n_0 \wedge n_1)\log 2}}$, we have*

$$\mathbb{P}\left(\left|\widetilde{\Sigma}_{jk} - C_{jk}\right| > \varepsilon\right) \leq 2\exp\left(-c_0 n \varepsilon^2\right), \quad \forall\, j,k = 1,\ldots,d;$$
$$\mathbb{P}(||(\widehat{\mu}_1 - \widehat{\mu}_0) - (\mu_1 - \mu_0)||_\infty > \varepsilon) \leq 2d\exp(-c_1 n\varepsilon^2).$$

Proof Using Lemma 22 and the fact that $\mathbb{P}(|n_j - \frac{n}{2}| \geq n\varepsilon) \leq 2\exp(-2n\varepsilon^2)$ for $j = 0, 1$, we have the result. ■

Remark 24 *Here n_0 and n_1 are "pretended" to be deterministic but not random variables. Later we will see that because $n_0 \wedge n_1 > \frac{n}{4}$ with an overwhelming probability, we can easily rewrite the condition $\varepsilon \geq 32\pi\sqrt{\frac{\log d}{(n_0\wedge n_1)\log 2}}$ to be a deterministic one: $\varepsilon \geq 64\pi\sqrt{\frac{\log d}{n\log 2}}$ in the final presentation.*

Lemma 25 *Let $\lambda_{\min}(\mathbf{C}_{SS}) = \delta$. If $\delta \geq 64\pi s\sqrt{\frac{\log d}{(n_0\wedge n_1)\log 2}}$, we have*

$$\mathbb{P}(\widetilde{\mathbf{\Sigma}}_{SS} \succ 0) \geq 1 - 2s^2\exp\left(-\frac{c_0 n\delta^2}{4s^2}\right). \tag{35}$$

Proof Let $\widehat{\mathbf{\Delta}} = \widetilde{\mathbf{\Sigma}}_{SS} - \mathbf{C}_{SS}$. Using Lemma 23, in probability $1 - 2s^2\exp(-c_0 nt^2)$, $\|\widehat{\mathbf{\Delta}}\|_{\max} \leq t$. Therefore, for any $\boldsymbol{v} \in \mathbb{R}^s$,

$$\boldsymbol{v}^T\widetilde{\mathbf{\Sigma}}_{SS}\boldsymbol{v} = \boldsymbol{v}^T\mathbf{C}_{SS}\boldsymbol{v} + \boldsymbol{v}^T\widehat{\mathbf{\Delta}}\boldsymbol{v} \geq \delta\|\boldsymbol{v}\|_2^2 + \lambda_{\min}(\widehat{\mathbf{\Delta}})\|\boldsymbol{v}\|_2^2,$$

where $\delta = \lambda_{\min}(\mathbf{C}_{SS})$. By the norm equivalence, we have

$$\|\widehat{\mathbf{\Delta}}\|_{\text{op}} \leq s\|\widehat{\mathbf{\Delta}}\|_{\max} \leq st,$$

where $||\cdot||_{\text{op}}$ is the matrix operator norm. Since

$$||-\widehat{\mathbf{\Delta}}||_{\text{op}} = \lambda_{\max}(-\widehat{\mathbf{\Delta}}) = \lambda_{\max}(\mathbf{C}_{SS} - \widetilde{\mathbf{\Sigma}}_{SS}) = -\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_{SS} - \mathbf{C}_{SS}) = -\lambda_{\min}(\widehat{\mathbf{\Delta}}),$$

and

$$||-\widehat{\mathbf{\Delta}}||_{\text{op}} \leq s||-\widehat{\mathbf{\Delta}}||_{\max} = s||\widehat{\mathbf{\Delta}}||_{\max} \leq st,$$

we can further have

$$\lambda_{\min}(\widehat{\boldsymbol{\Delta}}) \geq -\|\widehat{\boldsymbol{\Delta}}\|_{\text{op}} \geq -st.$$

Therefore we have

$$\boldsymbol{v}^T \widetilde{\boldsymbol{\Sigma}}_{SS} \boldsymbol{v} \geq (\delta - st)\,\|\boldsymbol{v}\|_2^2, \quad \text{i.e., } \lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{SS}) \geq \delta - st.$$

In other words, for all $t < \delta/s$, we have $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{SS}) > 0$. In particular, choosing $t = \delta/(2s)$, we have $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{SS}) = \delta/2 > 0$. This proves that Equation (35) holds with high probability. ■

Using Lemma 22, Lemma 23 and Lemma 25, Theorem 11 can be obtained using a similar proof structure of Mai et al. (2012). For concreteness and self-containedness, we provide a proof of the remaining part in the last section of the appendix.

Appendix C. Proof of Theorem 15

To prove Theorem 15, we first need to quantify the convergence rate of $\widehat{f}$ to f, or equivalently, $\widehat{f}_0 := \{\widehat{f}_{0j}\}_{j=1}^d$ and $\widehat{f}_1 := \{\widehat{f}_{1j}\}_{j=1}^d$'s convergence rates to f. By symmetry, we can focus on $\widehat{f}_0$.

Lemma 26 *Let $g_j := f_j^{-1}$ be the inverse function of f_j. We define*

$$I_n := \left[g_j\left(-\sqrt{2(1-\gamma)\log n}\right), g_j\left(\sqrt{2(1-\gamma)\log n}\right)\right],$$

then $\sup_{t \in I_n} |\widehat{f}_{0j}(t) - f_j(t)| = O_P\left(\sqrt{\frac{\log\log n}{n^{\gamma}}}\right)$.

Proof Using Lemma 22, a similar proof as Theorem 2 can be applied. ■

Then we can proceed to proof of Theorem 15:

Proof We define $\{j_1, \ldots, j_s\} = S$ to be the indices of the s discriminant features, that is,

$$\beta^*_{j_k} \neq 0, \quad k = 1, \ldots, s.$$

In this way, we can further define

$$T_n = \left[g_{j_1}(-\sqrt{b\log n}), g_{j_1}(\sqrt{b\log n})\right] \times \ldots, \times \left[g_{j_s}(-\sqrt{b\log n}), g_{j_s}(\sqrt{b\log n})\right],$$

for some $0 < b < 1$. Moreover, an event M_n is defined as

$$M_n := \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}_S \in T_n\}.$$

Then we have

$$\begin{aligned}
&\mathbb{P}\left(|\widehat{G}\left(\boldsymbol{X}, \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}, \widehat{f}\right) - G(\boldsymbol{X}, \boldsymbol{\beta}^{**}, f)| > t\right) \\
&\leq \mathbb{P}\left(|\widehat{G}\left(\boldsymbol{X}, \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}, \widehat{f}\right) - G(\boldsymbol{X}, \boldsymbol{\beta}^{**}, f)| > t \,|\, \mathcal{R}(\mathbf{X}, \boldsymbol{\beta}^*, \lambda_n), M_n\right) \\
&\quad + \mathbb{P}(M_n^c) + \mathbb{P}(\mathcal{R}(\mathbf{X}, \boldsymbol{\beta}^*, \lambda_n)^c).
\end{aligned}$$

Given $\mathcal{R}(\mathbf{X}, \boldsymbol{\beta}^*, \lambda_n)$ and M_n hold, we have

$$\begin{aligned}\left|\widehat{G}\left(\boldsymbol{X}, \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}, \widehat{f}\right) - G(\boldsymbol{X}, \boldsymbol{\beta}^{**}, f)\right| \leq & \left|(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^T \boldsymbol{\beta}^{**}\right| + \left|\widehat{f}(\boldsymbol{X})^T\left(\boldsymbol{\beta}^{**} - \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}\right)\right| \\ & + \left|\widehat{\boldsymbol{\mu}}^T\left(\boldsymbol{\beta}^{**} - \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}\right)\right| + \left|(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_a)^T \boldsymbol{\beta}^{**}\right| \\ \leq & \beta^{**}_{\max}||(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))_S||_1 + ||(\widehat{f}(\boldsymbol{X}))_S||_1||\boldsymbol{\beta}^{**} - \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}||_\infty \\ & + ||\widehat{\boldsymbol{\mu}}_S||_1||\boldsymbol{\beta}^{**} - \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}||_\infty + ||\boldsymbol{\beta}^{**}||_\infty||(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_a)_S||_1.\end{aligned}$$

Using Theorem 2,

$$\sup_{\boldsymbol{X}\in M_n} ||(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))_S||_1 = O_P\left(s\sqrt{\frac{\log\log n}{n^{1-b/2}}}\right), \tag{36}$$

and by Lemma 26,

$$\sup_{\boldsymbol{X}\in M_n} ||(\widehat{f}(\boldsymbol{X}))_S||_1 = O_P\left(s\sqrt{2\log n}\right).$$

Using the Gaussian tail inequality,

$$\mathbb{P}\left(f_j(X_j) \geq \sqrt{b\log n}\right) = O\left(n^{-c_2\cdot b}\right),$$

so $\mathbb{P}(M_n^c) = O\left(sn^{-c_2\cdot b}\right)$. Using Lemma 23,

$$||\widehat{\boldsymbol{\mu}}_S||_1 = O_P(s\Delta_{\max}) \quad \text{and} \quad ||(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_a)_S||_1 = O_P(sn^{-1/2}).$$

Using the assumption that $A+B+C \to 0$ and $B \to 0$ in the Theorem 15, we have

$$||\boldsymbol{\beta}^{**} - \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}||_\infty = O_P\left(D_{\max}\lambda_n\right) \quad \text{and} \quad \mathbb{P}\left(\mathcal{R}(\mathbf{X}, \boldsymbol{\beta}^*, \lambda_n)^c\right) = o(1). \tag{37}$$

Combining Equation (36) to (37), we have

$$\left|\widehat{G}\left(\boldsymbol{X}, \frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1}, \widehat{f}\right) - G(\boldsymbol{X}, \boldsymbol{\beta}^{**}, f)\right| = O_P\left(s\beta^{**}_{\max}\sqrt{\frac{\log\log n}{n^{1-b/2}}} + sD_{\max}\lambda_n(\sqrt{\log n} + \Delta_{\max}) + \frac{s\beta^{**}_{\max}}{\sqrt{n}}\right).$$

This completes the proof. ■

Appendix D. Proof of Corollary 19

Proof For notation simplicity, we denote by

$$\mathcal{D}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_n,y_n)\} \quad \text{and} \quad G^*:=G(\boldsymbol{X},\boldsymbol{\beta}^{**},f), \quad \widetilde{G}:=\widehat{G}\left(\boldsymbol{X},\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0n_1},\widehat{f}\right).$$

Here we note that $\mathrm{sign}(\widetilde{G})=\mathrm{sign}\left(\widehat{G}\left(\boldsymbol{X},\widehat{\boldsymbol{\beta}},\widehat{f}\right)\right)$. Then we have

$$\begin{aligned}\mathbb{P}\left(Y\cdot\mathrm{sign}(\widetilde{G})<0|\mathcal{D}\right) &= \mathbb{P}\left(Y\cdot\mathrm{sign}(G^*)+Y\cdot(\mathrm{sign}(\widetilde{G})-\mathrm{sign}(G^*))<0 \mid \mathcal{D}\right)\\ &\leq \mathbb{P}(Y\cdot\mathrm{sign}(G^*)<0)+\mathbb{P}\left(Y\cdot(\mathrm{sign}(\widetilde{G})-\mathrm{sign}(G^*))<0 \mid \mathcal{D}\right)\\ &\leq \mathbb{P}(Y\cdot\mathrm{sign}(G^*)<0)+\mathbb{P}\left(\mathrm{sign}(\widetilde{G})\neq\mathrm{sign}(G^*) \mid \mathcal{D}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}(\mathcal{C}(\widehat{g}))-\mathcal{C}(g^*) &\leq \mathbb{E}\left(\mathbb{P}(\mathrm{sign}(\widetilde{G})\neq\mathrm{sign}(G^*) \mid \mathcal{D})\right)\\ &= \mathbb{E}\left(\mathbb{E}(I(\mathrm{sign}(\widetilde{G})\neq\mathrm{sign}(G^*)) \mid \mathcal{D})\right)\\ &= \mathbb{E}\left(I(\mathrm{sign}(\widetilde{G})\neq\mathrm{sign}(G^*))\right)\\ &= \mathbb{P}\left(\mathrm{sign}(\widetilde{G})\neq\mathrm{sign}(G^*))\right).\end{aligned}$$

Given $t_{n,d,s}$ a constant depending only on (d,n,s), we have

$$\begin{aligned}&\mathbb{P}\left(\mathrm{sign}(\widetilde{G})\neq\mathrm{sign}(G^*)\right)\\ =\ &\mathbb{P}\left(\widetilde{G}\cdot G^*<0\right)\\ =\ &\mathbb{P}\left(\widetilde{G}\cdot G^*<0,|\widetilde{G}-G^*|<t_{n,d,s}\right)+\mathbb{P}\left(\widetilde{G}\cdot G^*<0,|\widetilde{G}-G^*|\geq t_{n,d,s}\right)\\ \leq\ &\mathbb{P}\left(|\widetilde{G}-G^*|<t_{n,d,s},\widetilde{G}\cdot G^*<0\right)+\mathbb{P}\left(|\widetilde{G}-G^*|>t_{n,d,s}\right)\\ \leq\ &\mathbb{P}\left(\widetilde{G}\cdot G^*<0 \ \mid\ |\widetilde{G}-G^*|<t_{n,d,s}\right)\mathbb{P}\left(|\widetilde{G}-G^*|<t_{n,d,s}\right)+\mathbb{P}\left(|\widetilde{G}-G^*|>t_{n,d,s}\right)\\ \leq\ &\mathbb{P}\left(|G^*|<t_{n,d,s} \ \mid\ |\widetilde{G}-G^*|<t_{n,d,s}\right)\mathbb{P}\left(|\widetilde{G}-G^*|<t_{n,d,s}\right)+\mathbb{P}\left(|\widetilde{G}-G^*|>t_{n,d,s}\right)\\ \leq\ &\mathbb{P}(|G^*|<t_{n,d,s})+\mathbb{P}\left(|\widetilde{G}-G^*|>t_{n,d,s}\right).\end{aligned}$$

Suppose that the conditions in Corollary 3 hold, then choosing

$$t_{n,d,s}=s^2\log^2 n\cdot\sqrt{\frac{\log d+\log s}{n}},$$

using Corollary 17, we know that

$$\mathbb{P}\left(\left|\widehat{G}\left(\boldsymbol{X},\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0n_1},\widehat{f}\right)-G(\boldsymbol{X},\boldsymbol{\beta}^{**},f)\right|>t_{n,d,s}\right)=o(1).$$

And using Lemma 6, we have

$$G(\boldsymbol{X},\boldsymbol{\beta}^{**},f)=(f(\boldsymbol{X})-\boldsymbol{\mu}_a)^T\boldsymbol{\beta}^{**}\sim N\left(\frac{\tau\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d}{2},\tau^2\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d\right),$$

where $\tau=\dfrac{4}{4+\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d}>0$. Therefore, by simple calculation, we have

$$\mathbb{P}(|G(\boldsymbol{X},\boldsymbol{\beta}^{**},f)|<t_{n,d,s})=\Phi\left(\frac{t_{n,d,s}-\frac{\tau\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d}{2}}{\tau\sqrt{\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d}}\right)-\Phi\left(\frac{-t_{n,d,s}-\frac{\tau\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d}{2}}{\tau\sqrt{\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d}}\right)=o(1),$$

as long as $t_{n,d,s}\to 0$ because of the continuity of Φ. This proves that $\mathbb{P}\left(\text{sign}(\widetilde{G})\neq\text{sign}(G^*)\right)=o(1)$ and completes the proof. The same argument can be generalized to the case where the conditions in Theorem 15 hold. ■

Appendix E. Proof of the Remaining Part of Theorem 11

we now start to prove the remaining part of Theorem 11. In the sequel, all the equalities and inequalities are element-wise. The main structure of the proof is coming from Mai et al. (2012) and we include the proof here only for the paper concreteness and self-containedness.

Lemma 27 *Given $n_j\sim$ Binomial$\left(n,\frac{1}{2}\right)$ for $j=0,1$, we have*

$$\mathbb{P}\left(\frac{3}{16}\leq\frac{n_0n_1}{n^2}\leq\frac{1}{4}\right)\geq 1-2\exp\left(-\frac{n}{8}\right).$$

Proof Using the Hoeffding's inequality, we have

$$\mathbb{P}\left(\frac{3}{16}\leq\frac{n_0n_1}{n^2}\leq\frac{1}{4}\right)=\mathbb{P}\left(\left|n_j-\frac{n}{2}\right|\leq\frac{n}{4}\right)=\mathbb{P}\left(\left|\frac{n_j}{n}-\frac{1}{2}\right|\leq\frac{1}{4}\right)\geq 1-2\exp\left(-\frac{n}{8}\right).$$

This completes the proof. ■

Proof [Proof of the Theorem 11] We define the event

$$E_0:=\left\{\frac{3}{16}\leq\frac{n_0n_1}{n^2}\leq\frac{1}{4}\right\}.$$

Under the event E_0, we consider the optimization problem in Equation (11). We firstly consider an intermediate optimum:

$$\widetilde{\boldsymbol{\beta}}_S:=\underset{\boldsymbol{\beta}_S\in\mathbb{R}^s}{\text{argmin}}\left\{\frac{1}{2n}||\boldsymbol{y}-\widetilde{\mathbf{X}}_S\boldsymbol{\beta}_S||_2^2+\lambda_n||\boldsymbol{\beta}_S||_1\right\}.$$

Reminding that $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}/n$ has been replaced by $\widetilde{\boldsymbol{\Sigma}}$ in calculating $\widehat{\boldsymbol{\beta}}$. Using Lemma 25, $\widetilde{\boldsymbol{\Sigma}}_{SS}$ is invertible with high probability. Then, under the event that $\widetilde{\boldsymbol{\Sigma}}_{SS}$ is invertible, $\widetilde{\boldsymbol{\beta}}_S$ exists and is unique, moreover

$$\widetilde{\boldsymbol{\beta}}_S=(\widetilde{\boldsymbol{\Sigma}}_{SS})^{-1}\left[\frac{n_0n_1}{n^2}(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)_S-\lambda_n\boldsymbol{z}_S\right],$$

where $\boldsymbol{z}_S$ is the subgradient such that $z_j = \text{sign}(\widetilde{\beta}_j) \neq 0$ and $-1 \leq z_j \leq 1$ if $\widetilde{\beta}_j = 0$.

To prove that $\widehat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_S, 0)$ with high probability, it suffices to show that $||\boldsymbol{z}_{S^c}||_\infty \leq 1$, or equivalently $\mathcal{R}_1(\mathbf{X}, \boldsymbol{\beta}^{**}, \lambda_n)$ holds, where

$$\begin{aligned}&\mathcal{R}_1(\mathbf{X}, \boldsymbol{\beta}^{**}, \lambda_n) := \\ &\quad \left\{ || \frac{n_0 n_1}{n^2} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_{S^c} - \widetilde{\boldsymbol{\Sigma}}_{S^c S} (\widetilde{\boldsymbol{\Sigma}}_{SS})^{-1} \left[\frac{n_0 n_1}{n^2} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - \lambda_n \boldsymbol{z}_S \right] ||_\infty \leq \lambda_n \right\}.\end{aligned} \tag{38}$$

Then following Equation (38), with high probability, we now can write

$$\begin{aligned}&\mathbb{P}(\mathcal{R}_1(\boldsymbol{X}, \boldsymbol{\beta}^{**}, \lambda_n)^c) \\ \leq &\mathbb{P}\left(|| \frac{n_0 n_1}{n^2} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_{S^c} - \widetilde{\boldsymbol{\Sigma}}_{S^c S} (\widetilde{\boldsymbol{\Sigma}}_{SS})^{-1} \left(\frac{n_0 n_1}{n^2} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - \lambda_n \boldsymbol{z}_S \right) ||_\infty > \lambda_n \right).\end{aligned}$$

Let $\lambda = \frac{2n^2 \lambda_n}{n_0 n_1}$ and using the matrix norm equivalency, we have

$$\begin{aligned}\mathbb{P}(\mathcal{R}_1(\boldsymbol{X}, \boldsymbol{\beta}^{**}, \lambda_n)^c) &\leq \mathbb{P}\left(||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_{S^c} - \widetilde{\boldsymbol{\Sigma}}_{S^c S} (\widetilde{\boldsymbol{\Sigma}}_{SS})^{-1} \left((\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - \frac{\lambda}{2} \boldsymbol{z}_S) \right) ||_\infty > \lambda/2 \right) \\ &\leq \mathbb{P}\Bigg(\zeta \Delta_{\max} + ||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_{S^c} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_{S^c}||_\infty \\ &\qquad + (\zeta + \psi) \cdot \left(\frac{\lambda}{2} + ||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty \right) > \frac{\lambda}{2} \Bigg),\end{aligned}$$

where

$$\zeta := ||\widetilde{\boldsymbol{\Sigma}}_{S^c S} (\widetilde{\boldsymbol{\Sigma}}_{SS})^{-1} - \mathbf{C}_{S^c S} (\mathbf{C}_{SS})^{-1}||_\infty.$$

.

The key part of the rest of proof is obtain by using the concentration inequalities for several key estimators. In Lemma 28, we give such a result:

Lemma 28 *There exist constants c_0 and c_1 such that, under the event E_0, for any $\varepsilon > 64\pi\sqrt{\frac{\log d}{n \log 2}}$, we have*

$$\begin{aligned}\mathbb{P}\left(\left| \widetilde{\Sigma}_{jk} - C_{jk} \right| > \varepsilon \right) &\leq 2 \exp\left(-n c_0 \varepsilon^2 \right), \quad \forall\, j, k = 1, \ldots, d; &(39) \\ \mathbb{P}\left(||\widetilde{\boldsymbol{\Sigma}}_{SS} - \mathbf{C}_{SS}||_\infty > \varepsilon \right) &\leq 2s^2 \exp\left(-\frac{n c_0 \varepsilon^2}{s^2} \right); &(40) \\ \mathbb{P}\left(||\widetilde{\boldsymbol{\Sigma}}_{S^c S} - \mathbf{C}_{S^c S}||_\infty > \varepsilon \right) &\leq 2(d-s)s \exp\left(-\frac{n c_0 \varepsilon^2}{s^2} \right); &(41) \\ \mathbb{P}(||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)||_\infty > \varepsilon) &\leq 2d \exp(-n c_1 \varepsilon^2). &(42)\end{aligned}$$

And for any $\varepsilon < 1/D_{\max}$, we have

$$\mathbb{P}\left(\zeta > \varepsilon D_{\max} (\psi + 1)(1 - D_{\max} \varepsilon)^{-1} \right) \leq 2ds \exp\left(-\frac{n c_0 \varepsilon^2}{s^2} \right). \tag{43}$$

Proof [Proof of the Lemma 28] Given Lemma 23, Equation (39) and Equation (42) are correct and Equation (40) and (41) are straightforward using Equation (39). To prove that Equation (43) holds, we have the key observation from Mai et al. (2012):

$$||\widetilde{\mathbf{\Sigma}}_{S^cS}(\widetilde{\mathbf{\Sigma}}_{SS})^{-1}-\mathbf{C}_{S^cS}(\mathbf{C}_{SS})^{-1}||_\infty \leq \\ \left(\psi||\widetilde{\mathbf{\Sigma}}_{SS}-\mathbf{C}_{SS}||_\infty+||\widetilde{\mathbf{\Sigma}}_{S^cS}-\mathbf{C}_{S^cS}||_\infty\right)\left(D_{\max}+||(\widetilde{\mathbf{\Sigma}}_{SS})^{-1}-(\mathbf{C}_{SS})^{-1}||_\infty\right).$$

We choose

$$\varepsilon \geq \max\{||\widetilde{\mathbf{\Sigma}}_{SS}-\mathbf{C}_{SS}||_\infty, ||\widetilde{\mathbf{\Sigma}}_{S^cS}-\mathbf{C}_{S^cS}||_\infty\},$$

and substitute $||\widetilde{\mathbf{\Sigma}}_{SS}-\mathbf{C}_{SS}||_\infty$ and $||\widetilde{\mathbf{\Sigma}}_{S^cS}-\mathbf{C}_{S^cS}||_\infty$ with ε, then apply Equation (40) and Equation (41), we have Equation (43). ■

Therefore, using the condition that, under the event E_0, we have

$$\varepsilon \leq \frac{n^2\lambda_n(1-\psi)}{2D_{\max}(n^2\lambda_n+n_0n_1(1+\psi)\Delta_{\max})}=\frac{\lambda(1-\psi)}{4D_{\max}(\lambda/2+(1+\psi)\Delta_{\max})},$$

we have

$$\Delta_{\max} \leq \frac{(1-2\varepsilon D_{\max}-\psi)\lambda}{4(1+\psi)\varepsilon D_{\max}}.$$

Noticing that if

$$\zeta \leq \frac{(\psi+1)\varepsilon D_{\max}}{1-D_{\max}\varepsilon},\quad ||(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)||_\infty \leq \frac{\lambda(1-\psi-2\varepsilon D_{\max})}{4(1+\psi)}, \\ \lambda < \Delta_{\max} \quad \text{and} \quad \Delta_{\max} \leq \frac{(1-2\varepsilon D_{\max}-\psi)\lambda}{4(1+\psi)\varepsilon D_{\max}},$$

then $\zeta\Delta_{\max}+||(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)_{S^c}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)_{S^c}||_\infty+(\zeta+\psi)\cdot\left(\frac{\lambda}{2}+||(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)_S-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)_S||_\infty\right) \leq \frac{\lambda}{2}$. Accordingly, denoting by

$$E_1=\left\{\zeta \geq \frac{(\psi+1)\varepsilon D_{\max}}{1-D_{\max}\varepsilon}\right\}, \\ E_2=\left\{||(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)||_\infty \geq \frac{\lambda(1-\psi-2\varepsilon D_{\max})}{4(1+\psi)}\right\},$$

we have

$$\mathbb{P}(\mathcal{R}_1(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)^c) \leq \mathbb{P}(E_1)+\mathbb{P}(E_2).$$

Using Lemma 28, we have the result that

$$\mathbb{P}(\mathcal{R}_1(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)^c) \leq 2ds\cdot\exp\left(-\frac{c_0n\varepsilon^2}{s^2}\right)+2d\cdot\exp\left(-\frac{4c_1n\lambda_n^2(1-\psi-2\varepsilon D_{\max})^2}{(1+\psi)^2}\right). \tag{44}$$

Then, to prove that $|\widetilde{\boldsymbol{\beta}}_S|>0$ with high probability, we consider the second set:

$$\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)=\left\{\left|\widetilde{\boldsymbol{\beta}}_S\right|>0\right\}=\left\{\left|(\widetilde{\mathbf{\Sigma}}_{SS})^{-1}\left[\frac{n_0n_1}{n^2}(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_0)_S-\lambda_n\boldsymbol{z}_S\right]\right|>0\right\}.$$

Again, denoting by $\lambda = \frac{2n^2\lambda_n}{n_0 n_1}$, we have

$$\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n) = \left\{ \left| (\widetilde{\mathbf{\Sigma}}_{SS})^{-1} \left[(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - \lambda \boldsymbol{z}_S/2 \right] \right| > 0 \right\}.$$

Denote by $\zeta_1 := ||\widetilde{\mathbf{\Sigma}}_{SS} - \mathbf{C}_{SS}||_\infty$ and $\zeta_2 := ||(\widetilde{\mathbf{\Sigma}}_{SS})^{-1} - (\mathbf{C}_{SS})^{-1}||_\infty$, we have

$$\begin{aligned}(\widetilde{\mathbf{\Sigma}}_{SS})^{-1} \left[(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - \lambda \boldsymbol{z}_S/2 \right] &= \boldsymbol{\beta}_S^{**} + (\widetilde{\mathbf{\Sigma}}_{SS})^{-1}[(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S] \\ &\quad + [(\widetilde{\mathbf{\Sigma}}_{SS})^{-1} - (\mathbf{C}_{SS})^{-1}](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S - \lambda(\widetilde{\mathbf{\Sigma}}_{SS})^{-1}\boldsymbol{z}_S/2, \end{aligned} \tag{45}$$

where we remind that $\mathbf{C}_{SS}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S = \boldsymbol{\beta}_S^{**}$. Therefore

$$\begin{aligned}\mathbb{P}(\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)) \geq \mathbb{P}\Big(& \beta_{\min}^{**} - (\zeta_2 + D_{\max})(\lambda/2 + ||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty) \\ & - \zeta_2 \Delta_{\max} > 0 \Big).\end{aligned}$$

Moreover, when $\zeta_1 D_{\max} < 1$, we have

$$\zeta_2 \leq ||(\widetilde{\mathbf{\Sigma}}_{SS})^{-1}||_\infty \zeta_1 D_{\max} \leq (D_{\max} + \zeta_2)\zeta_1 D_{\max},$$

and hence $\zeta_2 < D_{\max}^2 \zeta_1/(1 - \zeta_1 D_{\max})$. Therefore

$$\begin{aligned}&\mathbb{P}(\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)) \\ &\geq \mathbb{P}\Big(\omega\Delta_{\max}D_{\max} - (1 - \zeta_1 D_{\max})^{-1}(D_{\max}\lambda/2 + ||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty D_{\max} \\ &\quad + D_{\max}^2 \zeta_1 \Delta_{\max}) > 0 \Big).\end{aligned}$$

Noting that $\omega \leq 1$ because $\Delta_{\max}D_{\max} \geq ||\beta^{**}||_\infty$, we have $\lambda \leq \frac{\beta_{\min}^{**}}{2D_{\max}} \leq \frac{2\beta_{\min}^{**}}{(3+\omega)D_{\max}}$ under event E_0. Therefore, given that $\zeta_1 \leq \varepsilon$ and $||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty \leq \varepsilon$ and $\varepsilon \leq \frac{\omega}{(3+\omega)D_{\max}}$, $\varepsilon \leq \frac{\Delta_{\max}\omega}{2(\omega+3)}$, we have

$$\omega\Delta_{\max}D_{\max} - (1 - \zeta_1 D_{\max})^{-1}(D_{\max}\lambda/2 + ||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty D_{\max} + D_{\max}^2 \zeta_1 \Delta_{\max}) > 0.$$

Therefore

$$\begin{aligned}\mathbb{P}(\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)^c) &\leq \mathbb{P}(\zeta_1 \geq \varepsilon) + \mathbb{P}(||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty \geq \varepsilon) \\ &\leq 2s^2 \exp(-c_0 n\varepsilon^2/s^2) + 2s\exp(-nc_1\varepsilon^2).\end{aligned} \tag{46}$$

Combining Equation (44) and Equation (46), we have that $\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)$ holds with high probability.

Finally, using Equation (45), given that $\mathcal{R}_1(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)$ and $\mathcal{R}_2(\mathbf{X},\boldsymbol{\beta}^{**},\lambda_n)$ hold, we have

$$\begin{aligned}||\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1} - \boldsymbol{\beta}^{**}||_\infty &= ||\frac{n^2\widetilde{\boldsymbol{\beta}}_S}{n_0 n_1} - \boldsymbol{\beta}_S^{**}||_\infty \\ &\leq (1 - \zeta_1 D_{\max})^{-1}(D_{\max}\lambda/2 + ||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty D_{\max} + D_{\max}^2 \zeta_1 \Delta_{\max}).\end{aligned}$$

Using the fact that $\varepsilon \leq \frac{\lambda}{2D_{\max}\Delta_{\max}}$ and $\varepsilon \leq \lambda$, we have that, under the event $\zeta_1 \leq \varepsilon$ and $||(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)_S - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_S||_\infty \leq \varepsilon$,

$$\begin{aligned} ||\frac{n^2\widehat{\boldsymbol{\beta}}}{n_0 n_1} - \boldsymbol{\beta}^{**}||_\infty &\leq (1-\zeta_1 D_{\max})^{-1}(D_{\max}\lambda/2 + \lambda D_{\max} + D_{\max}\lambda/2) \\ &\leq \frac{2D_{\max}\lambda}{1 - \frac{\lambda}{2\Delta_{\max}}} \leq 4D_{\max}\lambda. \end{aligned}$$

We finalize the proof by using Lemma 27 to show that $\mathbb{P}(E_0^c) \leq 2\exp(-n/8)$. This completes the proof. ■

References

P.J. Bickel and E. Levina. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10: 989–1010, 2004.

T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106:1566–1577, 2012.

T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

D. Christensen. Fast algorithms for the calculation of kendall's τ. *Computational Statistics*, 20(1): 51–62, 2005.

R.T. Clemen and R. Reilly. Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224, 1999.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27 (3):642–669, 1956.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

A. Eloyan, J. Muschelli, M.B. Nebel, H. Liu, F. Han, T. Zhao, A. Barber, S. Joel, J.J. Pekar, S. Mostofsky, et al. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. 2012.

J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

J. Fan, Y. Feng, and X. Tong. A road to classification in high dimensional space. *Arxiv preprint arXiv:1011.6095*, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

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

J.H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2007.

C.P. Hans, D.D. Weisenburger, T.C. Greiner, R.D. Gascoyne, J. Delabie, G. Ott, H.K. Müller-Hermelink, E. Campo, R.M. Braziel, E.S. Jaffe, et al. Confirmation of the molecular classification of diffuse large b-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103 (1):275–282, 2004.

T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag. New York, NY, 2001.

H.L. Huang and F.L. Chang. Esvm: Evolutionary support vector machine for automatic feature selection and classification of microarray data. *Biosystems*, 90(2):516–528, 2007.

H. Ji, G. Wu, X. Zhan, A. Nolan, C. Koh, A. De Marzo, H.M. Doan, J. Fan, C. Cheadle, M. Fallahi, et al. Cell-type independent myc target genes reveal a primordial signature involved in biomass accumulation. *PloS one*, 6(10):e26057, 2011.

C. A. J. Klaassen and J. A. Wellner. Efficient estimation in the bivariate normal copula model: Normal margins are least-favorable. *Bernoulli*, 3(1):55–77, 1997.

W.H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53 No. 284.:814–861, 1958.

J.T. Leek and J.D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

Y. Lin and Y. Jeon. Discriminant analysis through a semiparametric model. *Biometrika*, 90(2): 379–392, 2003.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High dimensional semiparametric gaussian copula graphical models. *Annals of Statistics*, 2012.

H. Lovec, A. Grzeschiczek, M.B. Kowalski, and T. Möröy. Cyclin d1/bcl-1 cooperates with myc genes in the generation of b-cell lymphoma in transgenic mice. *The EMBO journal*, 13(15):3487, 1994.

Q. Mai, H. Zou, and M. Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 2012.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

M.N. McCall, B.M. Bolstad, and R.A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3), 2006.

J. Nocedal and S.J. Wright. *Numerical optimization (2nd ed.)*. Springer verlag, 2006.

J.D. Power, A.L. Cohen, S.M. Nelson, G.S. Wig, K.A. Barnes, J.A. Church, A.C. Vogel, T.O. Laumann, F.M. Miezin, B.L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems 22*, Cambridge, MA, 2009. MIT Press.

A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515, 2008.

K. Scheinberg, S. Ma, , and D. Glodfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems (NIPS), 23,*, 2010.

J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Arxiv preprint arXiv:1105.3561*, 2011.

G.R. Shorack and J.A. Wellner. *Empirical Processes With Applications to Statistics*. Wiley, 1986.

J.B. Smith and E. Wickstrom. Antisense c-myc and immunostimulatory oligonucleotide inhibition of tumorigenesis in a murine b-cell lymphoma transplant model. *Journal of the National Cancer Institute*, 90(15):1146–1154, 1998.

G. Strang. *Introduction to Linear Algebra*. Wellesley Cambridge Pr, 2003.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99 (10):6567, 2002.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

S van de Geer. *Empirical Processes in M-estimation*, volume 105. Cambridge university press Cambridge, UK, 2000.

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, May 2009.

L. Wang, J. Zhu, and H. Zou. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24(3):412–419, 2008.

S. Wang and J. Zhu. Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23(8):972, 2007.

D. Witten and R. Tibshirani. Penalized classification using fishers linear discriminant. *Journal of the Royal Statistical Society, Series B*, 2011.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Annals of Statistics*, 2012.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7:2541–2567, 2007.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research*, 98888:1059–1062, 2012.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.