Journal of Multivariate Analysis 135 (2015) 175–188

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Sparse semiparametric discriminant analysis

Qing Mai [a], Hui Zou [b,*]

[a] Department of Statistics, Florida State University, Tallahassee, FL, 32306-4330, USA
[b] School of Statistics, University of Minnesota, Minneapolis, MN, 55455, USA

ARTICLE INFO

Article history:
Received 18 February 2014
Available online 8 January 2015

AMS subject classification:
62H30

Keywords:
Gaussian copulas
Linear discriminant analysis
High-dimension asymptotics
Semiparametric model

ABSTRACT

In recent years, a considerable amount of work has been devoted to generalizing linear discriminant analysis to overcome its incompetence for high-dimensional classification (Witten and Tibshirani, 2011, Cai and Liu, 2011, Mai et al., 2012 and Fan et al., 2012). In this paper, we develop high-dimensional sparse semiparametric discriminant analysis (SSDA) that generalizes the normal-theory discriminant analysis in two ways: it relaxes the Gaussian assumptions and can handle ultra-high dimensional classification problems. If the underlying Bayes rule is sparse, SSDA can estimate the Bayes rule and select the true features simultaneously with overwhelming probability, as long as the logarithm of dimension grows slower than the cube root of sample size. Simulated and real examples are used to demonstrate the finite sample performance of SSDA. At the core of the theory is a new exponential concentration bound for semiparametric Gaussian copulas, which is of independent interest.

1. Introduction

Despite its simplicity, linear discriminant analysis (LDA) has proved to be a valuable classifier in many applications [22,12]. Let $X = (x_1, \ldots, x_p)$ denote the predictor vector and $Y \in \{+1, -1\}$ be the class label. The LDA model states that $X \mid Y \sim N(\mu_Y, \Sigma)$, yielding the Bayes rule

$$\hat{Y}^{\text{Bayes}} = \text{sign}\left[\{X - (\mu_+ + \mu_-)/2\}^{\mathrm{T}} \Sigma^{-1}(\mu_+ - \mu_-) + \log(\pi_+/\pi_-)\right],$$

where $\pi_y = \text{pr}(Y = y)$. Given n observations (Y^i, X^i), $1 \le i \le n$, the classical LDA classifier estimates the Bayes rule by substituting Σ, μ_y and π_y with their sample estimates. As is well known, the classical LDA fails to cope with high-dimensional data where the dimension, p, can be much larger than the sample size, n. A considerable amount of work has been devoted to generalizing LDA to meet the high-dimensional challenges. It is generally agreed that effectively exploiting sparsity is a key to the success of a generalized LDA classifier for high-dimensional data. Early attempts include the nearest shrunken centroids classifier (NSC) [28] and later the features annealed independence rule (FAIR) [8]. These two methods basically follow the diagonal LDA paradigm with an added variable selection component, where correlations among variable are completely ignored. Recently, more sophisticated sparse LDA proposals have been proposed; see Trendafilov and Jolliffe [29], Wu et al. [34], Clemmensen et al. [6], Witten and Tibshirani [33], Mai et al. [20], Shao et al. [25], Cai and Liu [3] and Fan et al. [9]. In these papers, a lot of empirical and theoretical results have been provided to demonstrate the competitive performance of sparse LDA for high-dimensional classification. These research efforts are rejuvenating discriminant analysis.

However, the existing sparse LDA methods become ineffective for non-normal data, which is easy to see from the theoretical viewpoint. See also empirical evidence given in Section 5.1. In the lower dimensional classification problems, some researchers have considered ways to relax the Gaussian distribution assumption. For example, Hastie and Tibshirani [13]

* Corresponding author.
E-mail addresses: mai@stat.fsu.edu (Q. Mai), zouxx019@umn.edu (H. Zou).

http://dx.doi.org/10.1016/j.jmva.2014.12.009
0047-259X/

proposed the mixture discriminant analysis (MDA) that uses a mixture of Gaussian distributions to model the conditional densities of variables given the class label. MDA is estimated by the Expectation-Maximization algorithm. MDA is a nonparametric generalization of LDA, but it is not clear how to further extend MDA to the high-dimensional classification setting with the ability to do variable selection. Lin and Jeon [16] proposed an interesting semiparametric linear discriminant analysis (SeLDA) model. Their model assumes that after a set of unknown monotone univariate transformations the observed data follow the classical LDA model. Lin and Jeon [16] further showed that the SeLDA model can be accurately estimated when p is fixed and n goes to infinity. However, the estimator in Lin and Jeon [16] cannot handle high-dimensional classification problems, especially when p exceeds n.

In this paper, we develop high-dimensional sparse semiparametric discriminant analysis (SSDA), a generalization of SeLDA for high-dimensional classification and variable selection. In particular, we propose a new estimator for the transformation function and establish its uniform consistency property as long as the logarithm of p is smaller than the cube root of n. With the new transformation estimator, we can transform the data and fit a sparse LDA classifier. In this work we use the direct sparse discriminant analysis (DSDA) developed by Mai et al. [20]. SSDA enjoys great computational efficiency: its computational complexity grows linearly with p. We show that, if the Bayes rule of the SeLDA model is sparse, then SSDA can consistently select the important variables and estimate the Bayes rule. At the core of the theory is an exponential concentration bound for semiparametric Gaussian copulas, which is of independent interest.

2. Semiparametric LDA model

Consider the binary classification problem where we have observed n random pairs (Y^i, X^i), $1 \le i \le n$ and wish to classify Y using a function of X. Lin and Jeon [16] proposed the following semiparametric LDA (SeLDA) model that assumes that

$$\big(h_1(X_1), \ldots, h_p(X_p)\big) \mid Y \sim N(\mu_Y, \Sigma), \tag{1}$$

where $h = (h_1, \ldots, h_p)$ is a set of strictly monotone univariate transformations. It is important to note that the SeLDA model does not assume that these univariate transformations are known or have any parametric forms. By properties of the Gaussian distribution, h is only unique up to location and scale shifts. Therefore, for identifiability, assume that $\mu_+ = 0$, $\Sigma_{jj} = 1$, $1 \le j \le p$. The Bayes rule of the SeLDA model is

$$\hat{Y}^{\text{Bayes}} = \text{sign}\left[\{h(X) - (\mu_+ + \mu_-)/2\}^{\mathrm{T}} \Sigma^{-1}(\mu_+ - \mu_-) + \log(\pi_+/\pi_-)\right].$$

The SeLDA model is a very natural generalization of the LDA model. It is equivalent to modelling the within-group distributions with semiparametric Gaussian copulas. For any continuous univariate random variable, W, we have

$$\Phi^{-1} \circ F(W) \sim N(0, 1), \tag{2}$$

where F is the cumulative probability function (CDF) of W and Φ is the CDF of the standard normal distribution. Gaussian copula is a multivariate generalization of that simple fact of univariate case. Semiparametric Gaussian copula has generated a lot of research interests in recent years; see Klaassen and Wellner [15], Song [26], Tsukahara [30], Chen and Fan [4] and Chen et al. [5]. The SeLDA model is the first application of semiparametric Gaussian copula in the context of classification.

The following lemma relates the univariate transformation function to the univariate marginal CDF of each predictor.

Lemma 1. *Consider a random vector $(X_1, \ldots, X_p)$ with strictly increasing marginal CDFs $F_1, \ldots, F_p$. If there exists a set of strictly increasing univariate functions $h = (h_1, \ldots, h_p)$ such that $h(X) \sim N(0, \Sigma)$, we must have $h_j = \Phi^{-1} \circ F_j$.*

In light of Lemma 1, the SeLDA model can be estimated in the low-dimensional setting. The basic idea is straightforward: we first find $\hat{h}_j(\cdot)$ as good estimates of these univariate transformation functions and then fit the LDA model on the "pseudo data" $\left\{Y^i, \hat{h}(X^i)\right\}$, $1 \le i \le n$. To be more specific, in seek of $\hat{h}_j$, we let F_{+j}, F_{-j} be the CDF of X_j conditional on $Y = +1$ and $Y = -1$, respectively, and then we have

$$h_j = \Phi^{-1} \circ F_{+j} = \Phi^{-1} \circ F_{-j} + \mu_-.$$

It can be seen that we only need an estimate of F_{+j} or F_{-j}. Denote n_+, n_- as the sample size within the positive and the negative class, respectively. For convenience, we let $n_+ \ge n_-$ throughout this paper. In other words, we code the class label of the majority class as "+1" and the minority class as "−1".

Denote X_{yj} as the jth entry of an observation X belonging to the group $Y = y$, and $\tilde{F}_{+j}$ as the empirical CDF of X_{+j}. Note that, we cannot directly plug in $\tilde{F}_{+j}$ so that $\hat{h}_j = \Phi^{-1} \circ \tilde{F}_{+j}$, because infinite values would occur at tails. Instead, $\tilde{F}_{+j}$ is Winsorized at a predefined pair of numbers (a, b) to obtain $\hat{F}^{a,b}_{+j}$

$$\hat{F}^{a,b}_{+j}(x) = \begin{cases} b & \text{if } \tilde{F}_{+j}(x) > b; \\ \tilde{F}_{+j}(x) & \text{if } a \le \tilde{F}_{+j}(x) \le b; \\ a & \text{if } \tilde{F}_{+j}(x) < a. \end{cases} \tag{3}$$

Then

$$\hat{h}_j = \Phi^{-1} \circ \hat{F}^{a,b}_{+j}. \tag{4}$$

The Winsorization can be viewed as a bias–variance trade-off.

With $\hat{h}_j$, the covariance matrix Σ is estimated by the pooled sample covariance matrix of $\hat{h}(X^i)$ and μ_{-j} is estimated by

$$\hat{\mu}_{-j} = q^{-1}\left[n_-^{-1}\sum_{i=1}^{n_-}\hat{h}(X_{-j}^i)1_{\tilde{F}(X_{-j}^i)\in(a,b)} + \phi\{\Phi^{-1}\circ\tilde{F}_{-j}\circ\tilde{F}_{+j}^{-1}(b)\} - \phi\{\Phi^{-1}\circ\tilde{F}_{-j}\circ\tilde{F}_{+j}^{-1}(a)\}\right]$$

where ϕ is the density function for a standard normal random variable and

$$q = n_-^{-1}\sum_{i=1}^{n_-}1_{\tilde{F}_{+j}(X^i)\in(a,b)}.$$

$\hat{\mu}_{-j}$ has this complicated form because of the Winsorization. Lin and Jeon [16] showed that when p is fixed and n tends to infinity, $\hat{\Sigma}$, $\hat{\mu}_-$ are consistent.

3. Estimation of the high-dimensional semiparametric LDA model

We need to address two technical problems when applying the SeLDA model to high-dimensional classification. First, we must modify the estimator in Lin and Jeon [16] to achieve consistency under ultra-high dimensions. Second, the SeLDA model is not estimable with the large-p-small-n data, even when we know the true transformation functions. To overcome this difficulty, we propose to fit a sparse SeLDA model by exploiting a sparsity assumption on the underlying Bayes rule. For the sake of presentation, we first discuss how to fit a sparse SeLDA model, provided that good estimators of $h_j(\cdot)$, $1 \le j \le p$, are already obtained. After introducing the sparse SeLDA, we focus on a new strategy to estimate $h_j(\cdot)$, $1 \le j \le p$.

3.1. Exploiting sparsity

We assume that the Bayes rule of the SeLDA model only involves a small number of predictors. To be more specific, let $\beta^{\rm Bayes} = \Sigma^{-1}(\mu_+ - \mu_-)$ and define $A = \{j : \beta_j^{\rm Bayes} \ne 0\}$. Sparsity means that $|A| \ll p$. An elegant feature of SeLDA is that it keeps the interpretation of LDA, that is, variable j is irrelevant if and only if $\beta_j^{\rm Bayes} = 0$.

Suppose that we have obtained $\hat{h}_j(\cdot)$ as a good estimate of $h_j(\cdot)$, $1 \le j \le p$, we focus on estimating the sparse LDA model using the "pseudo data" $\{Y^i, \hat{h}(X^i)\}$, $1 \le i \le n$. Among the previously mentioned sparse LDA proposals in the literature, only Fan and Fan [8], Shao et al. [25], Cai and Liu [3], Mai et al. [20] and Fan et al. [9] provided theoretical analysis of their methods. Fan and Fan [8]'s theory assumes that Σ is a diagonal matrix. Shao et al. [25]'s method works well only under some strong sparsity assumptions on the covariance matrix Σ and $\mu_+ - \mu_-$. The sparse LDA methods proposed in Cai and Liu [3], Mai [20] and Fan et al. [9] are shown to work well under general correlation structures. From the computational perspective, the method in Mai et al. [20] is most computationally efficient. Therefore, it is the method used here to exploit sparsity.

The proposal in Mai et al. [20], which is referred to as DSDA, begins with the observation that the classical LDA direction can be exactly recovered by doing linear regression of Y on $h(X)$ [14] where Y is treated as a numeric variable. Define $\Omega = \mathrm{Cov}\{h(X)\}$, and $\beta^\star = \Omega^{-1}(\mu_+ - \mu_-)$, $\beta^{\rm Bayes} = \Sigma^{-1}(\mu_+ - \mu_-)$. It can be shown that $\beta^\star$ and $\beta^{\rm Bayes}$ have the same direction. For variable selection and classification, it suffices to estimate $\beta^\star$. DSDA aims at estimating $\beta^\star$ by the following penalized least squares approach:

$$\hat{\beta}^{\rm DSDA} = \arg\min_{\beta}\left[n^{-1}\sum_{i=1}^{n}\left\{Y^i - \beta_0 - h(X^i)^{\rm T}\beta\right\}^2 + \sum_{j=1}^{p}P_\lambda(|\beta_j|)\right], \tag{5}$$

$$\hat{\beta}_0^{\rm DSDA} = -(\hat{\mu}_+ + \hat{\mu}_-)^{\rm T}\hat{\beta}^{\rm DSDA}/2 + \log(\hat{\pi}_+/\hat{\pi}_-)\cdot(\hat{\beta}^{\rm DSDA})^{\rm T}\hat{\Sigma}\hat{\beta}^{\rm DSDA}/\{(\hat{\mu}_+ - \hat{\mu}_-)^{\rm T}\hat{\beta}^{\rm DSDA}\}$$

where, under the LDA model, h is known to be $h(X) = X$, and $P_\lambda(\cdot)$ is a sparsity-inducing penalty, such as Lasso [27] or SCAD [10]. Then the DSDA classifier is $\mathrm{sign}\left\{\hat{\beta}_0^{\rm DSDA} + h(X)^{\rm T}\hat{\beta}^{\rm DSDA}\right\}$. There are many other penalty functions proposed for sparse regression, including the elastic net [39], the adaptive lasso [38], SICA [19] and the MCP [36], among others. All these penalties can be used in DSDA. The original paper [20] used the Lasso penalty where $P_\lambda(t) = \lambda t$ for $t > 0$. One could use either `lars` [7] or `glmnet` [11] to efficiently implement DSDA.

If we knew these transformation functions h in the SeLDA model, (5) could be directly used to estimate the Bayes rule of SeLDA. In SSDA we substitute h_j with its estimator $\hat{h}_j$ and apply sparse LDA methods to $(Y, \hat{h}(X))$. For example, to use DSDA in the SeLDA model, we solve for

$$\hat{\beta} = \arg\min_{\beta}\left[n^{-1}\sum_{i=1}^{n}\left\{Y^i - \beta_0 - \hat{h}(X^i)^{\rm T}\beta\right\}^2 + \sum_{j=1}^{p}P_\lambda(|\beta_j|)\right], \tag{6}$$

$$\hat{\beta}_0 = -(\hat{\mu}_+ + \hat{\mu}_-)^{\rm T}\hat{\beta}/2 + \log(\hat{\pi}_+/\hat{\pi}_-)\cdot\hat{\beta}^{\rm T}\hat{\Sigma}\hat{\beta}/\{(\hat{\mu}_+ - \hat{\mu}_-)^{\rm T}\hat{\beta}\}$$

Then (6) yields the SSDA classification rule: $\mathrm{sign}\left\{\hat{\beta}_0 + \hat{h}(X)^{\rm T}\hat{\beta}\right\}$.

3.2. Uniform estimation of transformation functions

We propose a high-quality estimator of the monotone transformation function. In order to establish the theoretical property of SSDA, we need all p estimators of the transformation function to uniformly converge to the truth at a certain fast rate, even when p is much larger than n. Our estimator is defined as

$$\hat{F}_{+j}(x) = \begin{cases} 1 - 1/n_+^2 & \text{if } \tilde{F}_{+j}(x) > 1 - 1/n_+^2 \\ \tilde{F}_{+j}(x) & \text{if } 1/n_+^2 \le \tilde{F}_{+j}(x) \le 1 - 1/n_+^2 \\ 1/n_+^2 & \text{if } \tilde{F}_{+j}(x) < 1/n_+^2 \end{cases} \tag{7}$$

and then

$$\hat{h}_j = \Phi^{-1} \circ \hat{F}_{+j}.$$

Note that the class with a bigger size is coded as "+" as mentioned in Section 2.

In other words, instead of fixing the Winsorization parameters a, b as in (3), we let

$$(a, b) = (a_n, b_n) = (1/n_+^2, 1 - 1/n_+^2). \tag{8}$$

With the presence of Φ^{-1}, it is necessary to choose $a_n > 0, b_n < 1$ to avoid extreme values at tails. On the other hand, $a_n \to 0, b_n \to 1$ so that the bias will automatically vanish as $n \to \infty$. To further see that (8) are proper choices of a_n, b_n, see the theory developed in Section 3 for mathematical justification.

Other estimators have been proposed. For example, Liu et al. [18] considered a one-class problem with Gaussian copulas, which essentially states $h(X) \sim N(0, \Sigma)$, and aims to estimate Σ^{-1}. In their paper, h_j is estimated by $\hat{h}_j = \Phi^{-1} \circ \hat{F}^{a_n,b_n}$, where $a_n = 1 - b_n = (4n^{1/4}\sqrt{\pi \log n})^{-1}$. Liu et al. [18] showed that this estimator is consistent when p is smaller than any polynomial order of n, but it is not clear whether the final SSDA can handle non-polynomial high dimensions.

Remark 1. Rank-based estimators were independently proposed by Liu et al. [17], Xue and Zou [35] for estimating Σ^{-1} without estimating the transformation functions. However, in the discriminant analysis problem considered here we need to estimate both Σ^{-1} and the mean vectors. The estimation of the mean vectors requires to estimate the transformation functions.

3.3. The pooled transformation estimator

We now consider an estimator that pools information from both classes. According to (2), we can find the estimated transformation functions by choosing proper $\hat{F}_{+j}$ and/or $\hat{F}_{-j}$, $\hat{\mu}_{-j}$. The naive estimate only uses the data from the positive class because of the difficulty in estimating μ_{-j}. However, we have the following lemma that will assist us in developing a more sophisticated transformation estimation utilizing all the data points.

Lemma 2. *Consider the model in (1). Then we have*

1. *Conditional on $Y = -1$, we have*
$$E(\Phi^{-1} \circ F_{+j}(X_j)) = \mu_{-j}.$$
2. *Conditional on $Y = +1$, we have*
$$E(\Phi^{-1} \circ F_{-j}(X_j)) = -\mu_{-j}.$$

Set $\hat{F}_{+j}$ as defined in (7) and $\hat{F}_{-j}$ as the empirical CDF for X_j conditional on $Y = -1$ Winsorized at $(a_{-n}, b_{-n}) = (1/n_-^2, 1 - 1/n_-^2)$. Then by Lemma 2, we can define a pooled estimator of μ_{-j}:

$$\hat{\mu}_{-j}^{(\text{pool})} = \hat{\pi}_+ \hat{\mu}_{-j}^{(+)} + \hat{\pi}_- \hat{\mu}_{-j}^{(-)},$$

where

$$\hat{\mu}_{-j}^{(+)} = \frac{1}{n_-} \sum_{Y^i=-1} \Phi^{-1} \circ \hat{F}_{+j}(X_j^i)$$

$$\hat{\mu}_{-j}^{(-)} = -\frac{1}{n_+} \sum_{Y^i=+1} \Phi^{-1} \circ \hat{F}_{-j}(X_j^i)$$

Then

$$\hat{h}_j^{(\text{pool})} = \hat{\pi}_+ \hat{h}_j^{(+)} + \hat{\pi}_- \hat{h}_j^{(-)},$$

where

$$\hat{h}_j^{(+)} = \Phi^{-1} \circ \hat{F}_{+j}$$

$$\hat{h}_j^{(-)} = \Phi^{-1} \circ \hat{F}_{-j} + \hat{\mu}_{-j}^{(\text{pool})}.$$

This estimator utilizes all the data points. We refer to this estimator as the pooled estimator. In Section 5 we will present numerical evidence that the pooled estimator does improve over the naive estimator in many cases.

4. Theoretical results

4.1. Estimation of transformation functions

To explore the consistency property of SSDA, we first study the estimation accuracy of semiparametric Gaussian copulas. The results in this subsection are applicable to any statistical model using semiparametric Gaussian copulas, which is of independent interest itself. Consider the one-class estimation case first. Assume that X is a p-dimensional random variable such that $h(X) \sim N(0_p, \Sigma)$ with $h_j = \Phi^{-1} \circ F_j$ and $\hat{h}_j = \Phi^{-1} \circ \hat{F}_j$, where $\hat{F}_j$ is defined as in (7). Denote $\hat{\mu}_j$ and $\hat{\sigma}_{jk}$ as the sample mean and sample covariance for corresponding features. We establish exponential concentration bounds for $\hat{\mu}_j$ and $\hat{\sigma}_{jk}$. For writing convenience, we use c to denote generic constants throughout.

Theorem 1. *Define*

$$\zeta_1^*(\epsilon) = 2\exp(-cn\epsilon^2) + 4\exp(-cn^{1-2\rho}\epsilon^2/\rho) + 4\exp(-cn^{\frac{1}{2}-\rho}),$$
$$\zeta_2^*(\epsilon) = c\exp(-cn\epsilon^2) + c\exp(-cn^{\frac{1}{3}-\rho}) + c\exp(-cn^{1-\rho}) + c\exp\{-c(\log^2 n)n^{1-2\rho}\epsilon^2/\rho^2\}.$$

For sufficiently large n and any $0 < \rho < \frac{1}{3}$, *there exists a positive constant* ϵ_0 *such that, for any* $0 < \epsilon < \epsilon_0$, *we have*

$$\mathrm{pr}(|\hat{\mu}_j - \mu_j| > \epsilon) \le \zeta_1^*(\epsilon) \tag{9}$$
$$\mathrm{pr}(|\hat{\sigma}_{jk} - \sigma_{jk}| > \epsilon) \le \zeta_2^*(\epsilon) \tag{10}$$

For the two-class SeLDA model, we can easily obtain the following corollary from Theorem 1.

Corollary 1. *Define*

$$\zeta_1(\epsilon) = \zeta_1^*(\pi_+^{1/2}\epsilon/2) + \zeta_1^*(\pi_-^{1/2}\epsilon/2) + 4\exp(-cn) \tag{11}$$
$$\zeta_2(\epsilon) = \zeta_2^*(\pi_+^{1/2}\epsilon/2) + \zeta_2^*(\pi_-^{1/2}\epsilon/2) + 4\exp(-cn) + 2\zeta_1(\epsilon) \tag{12}$$

Then there exists a positive constant ϵ_0 *such that, for any* $0 < \epsilon < \epsilon_0$, *we have*

$$\mathrm{pr}(|(\hat{\mu}_{+j} - \hat{\mu}_{-j}) - (\mu_{+j} - \mu_{-j})| > \epsilon) \le \zeta_1(\epsilon)$$
$$\mathrm{pr}(|\hat{\sigma}_{jk} - \sigma_{jk}| > \epsilon) \le \zeta_2(\epsilon)$$

Remark 2. Theorem 1 and Corollary 1 can be used for other high-dimensional statistical problems involving semiparametric Gaussian copulas.

4.2. Consistency of SSDA

In this section, we study the theoretical results for SSDA. For simplicity, we focus on SSDA with the naive estimator of transformation functions. The theoretical properties for SSDA combined with the pooled transformation estimator can be derived similarly with more lengthy calculation of the probability bounds.

With the results in Section 4.1, we are ready to prove the rate of convergence of SSDA. We first define necessary notation. Define $\beta^\star = \Omega^{-1}(\mu_+ - \mu_-)$, where Ω is the covariance of X. Recall that $\beta^\star$ is equal to $c\Sigma^{-1}(\mu_+ - \mu_-) = c\beta^{\mathrm{Bayes}}$ for some positive constant [20]. Then we can write $A = \{j : \beta_j^\star \ne 0\}$. Let s be the cardinality of A. In addition, for an $m_1 \times m_2$ matrix M, denote $\|M\|_\infty = \max_{i=1,\dots,m_1} \sum_{j=1}^{m_2} |M_{ij}|$, and, for a vector u, $\|u\|_\infty = \max|u_j|$. Throughout the proof, we assume that $s \ll n^{1/4}$. Define the following quantities that are repeatedly used:

$$\kappa = \|\Omega_{A^cA}(\Omega_{AA})^{-1}\|_\infty, \qquad \varphi = \|(\Omega_{AA})^{-1}\|_\infty, \Delta = \|\mu_{+A} - \mu_{-A}\|_\infty,$$
$$\Delta_1 = \|\mu_{+A} - \mu_{-A}\|_1, \qquad \Delta_2 = \|\mu_{+A} + \mu_{-A}\|_\infty \qquad \nu = \min_{j\in A}|\beta_j|/\Delta\varphi.$$

Suppose that the lasso estimator correctly shrinks $\hat{\beta}_{A^c}$ to zero, then SSDA should be equivalent to performing SeLDA on X_A. Therefore, define the hypothetical estimator

$$\hat{\beta}_A^{\mathrm{hyp}} = \arg\min_{\beta,\beta_0}\left[n^{-1}\sum_{i=1}^n\left\{Y^i - \beta_0 - \sum_{j\in A}\hat{h}_j(X_j^i)\beta_j\right\}^2 + \sum_{j\in A}\lambda|\beta_j|\right].$$

Then, we wish that $\hat{\beta} = (\hat{\beta}_A^{\mathrm{hyp}}, 0_{A^c})$ with $\hat{\beta}_j^{\mathrm{hyp}} \ne 0$ for $j \in A$. To ensure the consistency of SSDA, we further require the following condition:

$$\kappa = \|\Omega_{A^cA}(\Omega_{AA})^{-1}\|_\infty < 1. \tag{13}$$

The condition in (13) is an analogue of the ir-representable condition for the lasso penalized linear regression model [21,38,37,32]. Weaker conditions exist if one is only concerned with oracle inequalities for the coefficients under the regression model, such as the restricted eigenvalue condition [2,31,24]. It would be interesting to investigate if similar conditions can be extended to the framework of SSDA. We leave this as a future project. On the other hand, if one is reluctant to assume (13), the use of a concave penalty, such as SCAD [10], can remove this condition; see the discussion in Mai et al. [20].

Theorem 2. *Define ζ_1, ζ_2 as in Corollary 1. Pick any λ such that $\lambda < \min\{\min_{j\in A}|\beta_j|/(2\varphi), \Delta\}$. Then for any $\epsilon > 0$ and sufficiently large n such that $\epsilon > csn^{-\rho/2}$, where c does not depend on (n, p, s), we have*

1. *Assuming the condition in (13), with probability at least $1 - \psi_1$, $\hat{\beta}_A = \hat{\beta}_A^{\text{hyp}}$ and $\hat{\beta}_{A^c} = 0$, where*
$$\psi_1 = 2ps\zeta_2(\epsilon/s) + 2p\zeta_1[\lambda(1-\kappa-2\epsilon\varphi)/\{4(1+\kappa)\}]$$
and ϵ is any positive constant less than $\min(\epsilon_0, \lambda(1-\kappa)/[4\varphi\{\lambda/2 + (1+\kappa)\}\Delta])$.
2. *With probability at least $1 - \psi_2$, none of the elements of $\hat{\beta}_A$ is zero, where*
$$\psi_2 = 2s^2\zeta_2(\epsilon/s) + 2s\zeta_1(\epsilon)$$
and ϵ is any positive constant less than $\min\{\epsilon_0, \nu/\{(3+\nu)\varphi\}, \Delta\nu/(6+2\nu)\}$.
3. *For any positive ϵ satisfying $\epsilon < \min\{\epsilon_0, \lambda/(2\varphi\Delta), \lambda\}$, we have*
$$\mathrm{pr}(\|\hat{\beta}_A - \beta_A\|_\infty \le 4\varphi\lambda) \ge 1 - 2s^2\zeta_2(\epsilon/s) - 2s\zeta_1(\epsilon).$$

Theorem 2 provides the foundation for asymptotic results. Assume the following two regularity conditions.

(C1) $n, p \to \infty$ and $s^2\log(ps)/n^{1/3-\rho} \to 0$, for some ρ in $(0, 1/3)$;
(C2) $\min_{j\in A}|\beta_j| \gg \max[sn^{-\rho/2}, s\{\log(ps)/n^{1/3-\rho}\}^{1/2}]$ for some ρ in $(0, 1/3)$.

Condition (C1) restricts that p, s should not grow too fast comparing to n. However, p is allowed to grow faster than any polynomial order of n. Condition (C2) states that the important features should be sufficiently large such that we can separate them from the noises, which is a standard assumption in the literature of sparse recovery. The next theorem shows that SSDA consistently recovers the Bayes rule of the SeLDA model.

Theorem 3. *Let $\widehat{A} = \{j : \hat{\beta}_j \ne 0\}$. Under conditions (C1) and (C2), if we choose $\lambda = \lambda_n$ such that $\lambda_n \ll \min_{j\in A}|\beta_j|$ and $\lambda_n \gg s\{\log(ps)/n^{\frac{1}{3}-\rho}\}^{1/2}$, and further assume $\kappa < 1$, then $\mathrm{pr}(\widehat{A} = A) \to 1$ and $\mathrm{pr}\left(\|\hat{\beta}_A - \beta_A\|_\infty \le 4\varphi\lambda_n\right) \to 1$.*

Further, we prove that SSDA is asymptotically equivalent to the Bayes rule in terms of error rate. Note that the Bayes error rate $R = \mathrm{pr}(Y \ne \mathrm{sign}(h(X)^{\mathrm{T}}\beta^* + \beta_0))$ and $R_n = \mathrm{pr}(Y \ne \mathrm{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0))$. We have the following theorem.

Theorem 4. *Define ζ_1, ζ_2 as in Corollary 1. Pick any λ such that $\lambda < \min\{\frac{\min_{j\in A}|\beta_j|}{2\varphi}, \Delta\}$. Then for a sufficiently small constant $\epsilon > 0$ and sufficiently large n such that $\epsilon > csn^{-\rho/2}$, where c does not depend on (n, p, s), with probability no smaller than $1 - \psi_3$, we have $R_n - R < \epsilon$, where*

$$\psi_3 = cs\zeta_1\left(\frac{\epsilon}{s(\phi\Delta_1 + \Delta_2)}\right) + cp\zeta_1\left(\frac{\lambda(1-\kappa+2\epsilon\phi)}{4(1+\kappa)}\right) + 2ps\zeta_2\left(\frac{c\epsilon}{s}\right) + cp\exp\left(-c\frac{n^{1-\rho}}{\rho\log n}\right). \tag{14}$$

Corollary 2. *Under conditions (C1) and (C2), if we choose $\lambda = \lambda_n$ such that $\lambda_n \ll \min_{j\in A}|\beta_j|$ and $\lambda_n \gg \sqrt{\log(ps)\frac{s^2}{n^{\frac{1}{3}-\rho}}}$, and further assume $\kappa < 1$, then*

$$R_n - R \to 0 \quad \text{in probability}. \tag{15}$$

Remark 3. Our results concerning the error rate of SSDA are much more involved than those for sparse LDA algorithms in Cai and Liu [3], Fan et al. [9], because of the semiparametric assumptions. Under the parametric LDA model, the error rate tends to the Bayes error as long as the discriminant direction β is estimated consistently. However, under the SeLDA model, we deal with the extra uncertainty in estimating h and need some uniform convergence results on $\hat{h}(X)$.

5. Numerical results

5.1. Simulation

We examine the finite sample performance of SSDA by simulation. We consider two transformation estimators: the naive estimator and the pooled estimator. The resulting methods are denoted by SSDA(naive) and SSDA(pooled), respectively.

Table 1
Choices of g_j in Models 1b–4b.

$g_j(v)$	Models 1b,2b j	Model 3b j	Model 4b j
v^3	1, 101, ..., 150	1, 201, ..., 300	3, 201, ..., 300
$\exp(v)$	2, 151, ..., 200	2, 301, ..., 400	4, 301, ..., 400
$\arctan(v)$	3, 201, ..., 300	3, 401, ..., 500	5, 401, ..., 500
v^3	4, ..., 50	4, 6, ..., 100	1, 8, ..., 100
$\Phi(v)$	51, ..., 100	5, 101, ..., 200	2, 101, ..., 200
$(v+1)^3$	301, ..., 350	501, ..., 600	6, 501, ..., 600
$\arctan(2v)$	351, ..., 400	601, ..., 800	7, 601, ..., 800

For comparison, in the simulation study we also include DSDA and the sparse LDA algorithm [33] denoted by Witten for presentation purpose. After we apply the estimated transformation to the data, we use Witten's sparse LDA algorithm to fit the classifier. This gives us Se–Witten, another competitor in the simulation study.

Four types of SeLDA models were considered in the study. In each model, we first generated Y with $\pi_+ = \pi_- = 0.5$. For convenience, we say that Σ has AR(ρ) structure if $\Sigma_{ij} = \rho^{|i-j|}$ and Σ has CS(ρ) structure if $\Sigma_{ij} = \rho$ for any $i \neq j$. We fixed $\mu_- = 0$ and $\mu_+ = \Sigma\beta^{\mathrm{Bayes}}$.

Model 1: $n = 150$, $p = 400$. Σ has AR(0.5) structure.

$$\beta^{\mathrm{Bayes}} = 0.556(3, 1.5, 0, 0, 2, 0_{p-5})^{\mathrm{T}}.$$

Model 2: $n = 200$, $p = 400$. Σ has AR(0.5) structure.

$$\beta^{\mathrm{Bayes}} = 0.582(3, 2.5, -2.8, 0_{p-3})^{\mathrm{T}}.$$

Model 3: $n = 400$, $p = 800$. Σ has CS(0.5) structure.

$$\beta^{\mathrm{Bayes}} = 0.395(3, 1.7, -2.2, -2.1, 2.55, 0_{p-5})^{\mathrm{T}}.$$

Model 4: $n = 300$, $p = 800$. Σ is block diagonal with 5 blocks of dimension 160×160. Each block has CS(0.6) structure.

$$\beta^{\mathrm{Bayes}} = 0.916(1.2, -1.4, 1.15, -1.64, 1.5, -1, 2, 0_{p-7})^{\mathrm{T}}.$$

We transform V to X by $X = g(V)$ and the final data to be used are (X, Y). In each type of model, we consider two sets of g. We call the resulting models series a and b. In series a, $X = V$ so that the SeLDA model becomes the LDA model. In series b, we considered some commonly used transformations such that some features become heavily skewed, some heavy-tailed and some bounded. The choices of g are listed in Table 1. In the simulation study we also considered the oracle sparse discriminant classifiers including oracle DSDA and oracle Witten. The idea is to apply the true transformation to variables and then fit a sparse LDA classifier using DSDA or Witten and Tibshirani's method.

The simulation results for Models 1a–4a and Models 1b–4b are reported in Tables 2 and 3, respectively. Note that in Table 2 DSDA and Witten are the oracle DSDA and the oracle Witten. We can draw the following conclusions from Tables 2 and 3.

- Models 1a–4a are actually LDA models. SSDA performs very similarly to DSDA. Although SSDA has slightly higher error rates, this is expected because SSDA does not use the parametric assumption. On the other hand, in Models 1b–4b, SSDA performs much better than DSDA. These results jointly show that SSDA is a much more robust sparse discriminant analysis algorithm than those based on the LDA model.
- In both tables, SSDA is very close to the oracle DSDA, which empirically shows the high quality of the proposed transformation estimator in Section 3.2. In all eight cases, SSDA is a good approximation to the Bayes rule, which is consistent with the theoretical results. On the other hand, in Models 1, 2, 4 (a) & (b) SSDA(pooled) yields slightly lower error rates, which illustrates the advantage of utilizing the information from both classes when estimating the transformation.
- Se–Witten is a different SSDA classifier in which Witten and Tibshirani's method is used to fit the SeLDA model after estimating the transformation functions. Se–Witten performs very well in Models 1a,2a,1b,2b but it performs very poorly in Models 3a, 4a, 3b, 4b. The same is true for the oracle Witten method. By comparing SSDA and Se–Witten, we see that DSDA works better than Witten and Tibshirani's method. In addition to the theory in Section 4, the simulation also supports the use of DSDA in fitting the high-dimensional sparse semiparametric LDA model.

5.2. Malaria data

We further demonstrate SSDA by using the malaria data [23]. This dataset is available at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2362.

Out of 71 samples in the dataset, 49 have been infected with malaria, while 22 are healthy people. The predictors are the expression levels of 22283 genes. The 71 samples were split with a roughly 1:1 ratio to form training and testing sets. We

Table 2
Simulation results for Models 1a–4a. The reported numbers are medians based on 2000 replications. Their standard errors obtained by bootstrap are in parentheses. TRUE selection and FALSE selection denote the numbers of selected important variables and unimportant variables, respectively.

	Bayes	Oracle DSDA	SSDA	SSDA	DSDA	Oracle Witten	Se–Witten		Witten
			(naive)	(pooled)			(naive)	(pooled)	
Model 1 (a)									
Error(%)	10	10.71	11.5	11.11	10.71	11.39	11.56	11.57	11.39
		(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
TRUE selection	3	3	3	3	3	3	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	1	2	2	1	26	26	25	26
		(0.14)	(0.38)	(0.1)	(0.14)	(0.42)	(0.09)	(0)	(0.42)
Model 2 (a)									
Error(%)	10	11.09	11.66	11.57	11.09	13.36	13.46	13.58	13.36
		(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.04)	(0.02)	(0.03)
TRUE selection	3	3	3	3	3	3	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	5	6	6	5	24	24	24	24
		(0.37)	(0.51)	(0.48)	(0.37)	(0)	(0)	(0.5)	(0)
Model 3 (a)									
Error(%)	20	21.93	22.13	22.3	21.93	33.69	34.18	35.05	33.69
		(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0)	(0)	(0.01)
TRUE selection	5	5	5	5	5	3	5	5	3
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	14	13	14	14	419.5	795	795	419.5
		(0.59)	(0.57)	(0.58)	(0.59)	(10.19)	(0)	(0)	(10.19)
Model 4 (a)									
Error(%)	10	12.50	13.20	12.78	12.50	23.90	26.14	26	23.90
		(0.02)	(0.05)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
TRUE selection	7	7	7	7	7	4	5	5	4
		(0)	(0)	(0)	(0)	(0)	(0.02)	(0)	(0)
FALSE selection	0	18	17	17	18	35	153	153	35
		(0.70)	(0.54)	(0.45)	(0.70)	(4.43)	(0)	(0)	(4.43)

Table 3
Simulation results for Models 1b–4b. The reported numbers are medians based on 2000 replications. Their standard errors obtained by bootstrap are in parentheses. TRUE selection and FALSE selection denote the numbers of selected important variables and unimportant variables, respectively.

	Bayes	Oracle DSDA	SSDA	SSDA	DSDA	Oracle Witten	Se–Witten		Witten
			(naive)	(pooled)			(naive)	(pooled)	
Model 1 (b)									
Error(%)	10	10.71	11.5	11.11	18.24	11.39	11.56	11.57	16.19
		(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
TRUE selection	3	3	3	3	3	3	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	1	2	2	1	26	26	25	25
		(0.14)	(0.38)	(0.1)	(0)	(0.42)	(0.09)	(0)	(0.5)
Model 2 (b)									
Error(%)	10	11.09	11.66	11.57	19.47	13.36	13.46	13.58	20.16
		(0.02)	(0.03)	(0.03)	(0.09)	(0.03)	(0.04)	(0.02)	(0.04)
TRUE selection	3	3	3	3	3	3	3	3	2
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	5	6	6	5	24	24	24	20
		(0.37)	(0.51)	(0.48)	(0.37)	(0)	(0)	(0.5)	(0.17)
Model 3 (b)									
Error(%)	20	21.93	22.13	22.3	26.76	33.69	34.18	35.05	34.25
		(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0)	(0)	(0)
TRUE selection	5	5	5	5	5	3	5	5	3
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	14	13	14	15	419.5	795	795	795
		(0.59)	(0.57)	(0.58)	(0.67)	(10.19)	(0)	(0)	(10.19)
Model 4 (b)									
Error(%)	10	12.50	13.20	12.78	19.88	23.90	26.14	26	26.83
		(0.02)	(0.05)	(0.03)	(0.04)	(0.01)	(0.01)	(0.01)	(0.01)
TRUE selection	7	7	7	7	6	4	5	5	6
		(0)	(0)	(0)	(0)	(0)	(0.02)	(0)	(0.23)
FALSE selection	0	18	17	17	25	35	153	153	153
		(0.70)	(0.54)	(0.45)	(0.83)	(4.43)	(0)	(0)	(0.09)

Table 4
Comparison of SSDA(Naive), SSDA(pooled), DSDA and ℓ_1 logistic regression on the malaria dataset. The reported numbers are medians of 100 replicates, with standard errors obtained by bootstrap in parentheses.

	SSDA (Naive)	SSDA (Pooled)	DSDA	Logistic
Testing Error	2/35(0.59%)	1/35(1.35%)	6/35(0.99%)	4/35(0.67%)
Fitted Model Size	6(0.4)	6(0.4)	18(1.5)	17(0.6)

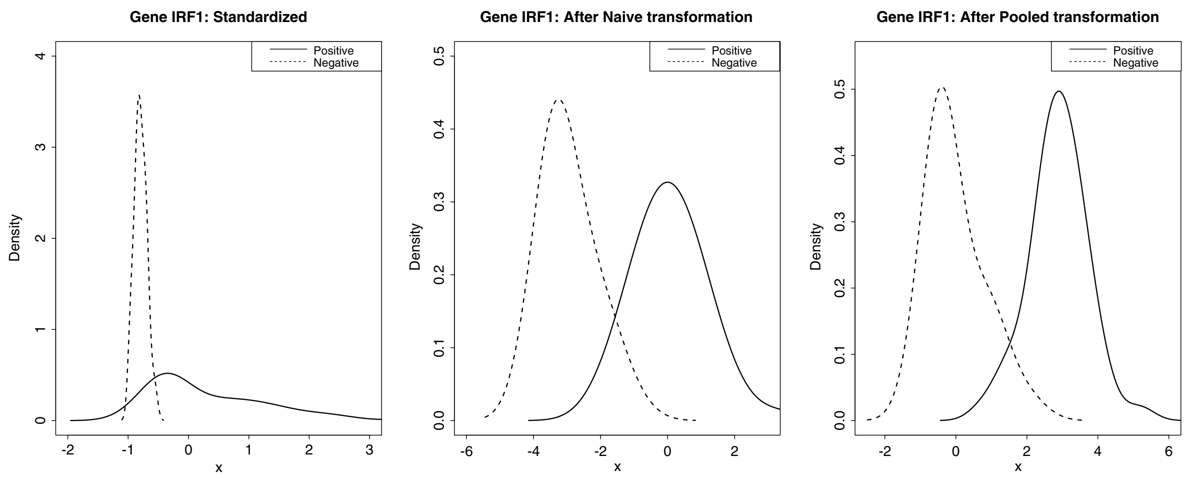

Fig. 1. Density functions of gene IRF1 (the 2059th gene) in the malaria data. From left to right the plots display the density functions of the standardized raw data, naively transformed data as in SSDA(naive) and data after pooled transformation as in SSDA(pooled), respectively.

report the median of 100 replicates in Table 4. Besides DSDA, the ℓ_1 logistic regression [11] was also considered because it is an obvious choice for sparse high-dimensional classification. From Table 4, it can be seen that both the SSDA methods are significantly more accurate than DSDA and the ℓ_1 logistic regression, with SSDA(pooled) yielding the lowest error rate of 1/35. In addition, the two SSDA methods select 6 genes, while the other two methods select more than 17 genes.

To gain more insight, we compared the selected genes by SSDA and those by DSDA or ℓ_1 logistic regression. In those 100 tries the 2059th gene is most frequently selected by SSDA, but seldom by DSDA or ℓ_1 logistic regression. This gene is encoded by IRF1, as it is the first identified interferon regulatory transcription factor (http://en.wikipedia.org/wiki/IRF1). Discovering the role of IRF1 was a major finding in Ockenhouse et al. [23]. Previous studies show that IRF1 influences the immune response. Therefore, healthy and sick people may have different expression levels on this gene. It is very interesting that we can use a pure statistical method like SSDA to select IRF1. We plot in Fig. 1 the within-group density functions of gene IRF1 (the 2059th gene). It can be seen that the raw expression levels of IRF1 are skewed, making linear rules unreliable on this gene. After applying the naive transformation, the distributions of both groups become close to normal, with similar variances. After the pooled transformation, the LDA model becomes even more plausible.

6. Discussion

It has been a hot subject of research in recent years to develop sparse discriminant analysis for high-dimensional classification and feature selection, rejuvenating the traditional discriminant analysis. However, sparse discriminant algorithms based on the LDA model can be very ineffective for non-normal data, as shown in the simulation study. To overcome the normality limitation, we consider the semiparametric discriminant analysis model and propose the SSDA, a high-dimensional semiparametric sparse discriminant classifier. We have justified SSDA both theoretically and empirically. For high-dimensional classification and feature selection, SSDA is more appropriate than the existing sparse discriminant analysis proposals in the literature.

Although we focus on binary classification throughout the paper, a classifier for multiclass problems is easy to obtain under the semiparametric model. Note that our SSDA method contains two independent steps: transforming the data and fitting a sparse LDA classifier. The first step can be carried out for multiclass problem with proper modification of our pooled estimator, as we will discuss in more detail later, while, in the second step, there already exist multiclass sparse LDA methods, such as sparse optimal scoring [6] and ℓ_1-Fisher's discriminant analysis [33]. The combination of the transformation and a multiclass sparse LDA method will yield a high-dimensional semiparametric classifier for multiclass problems. Specifically, consider a multiclass model $h(X) \mid Y \sim N(\mu_Y, \Sigma)$ where $Y = 1, \dots, K$ and $\mu_1 = 0$. Similar to Lemma 2, we can easily show that $E(\Phi^{-1} \circ F_{kj}(X_j) \mid Y = 1) = -\mu_{kj}$ and $E(\Phi^{-1} \circ F_{kj}(X_j) \mid Y = l) = \mu_{lj} - \mu_{kj}$. Define $\hat{F}_{kj}$ as the empirical CDF of X_j within Class k Winsorized at $(1/n_k^2, 1-1/n_k^2)$, where n_k is the sample size within Class k. Then we can find

$$\hat{\mu}_{kj}^{\text{pool}} = \sum_{l=1}^{K} \hat{\pi}_k \hat{\mu}_{kj}^{(l)}, \qquad \hat{h}_j^{\text{pool}} = \sum_{k=1}^{K} \hat{\pi}_k \hat{h}_j^k$$

where $\hat{\pi}_k = n_k/n$, $\hat{\mu}_{kj}^{(l)} = \frac{1}{n_k}\sum_{Y^i=k} \Phi^{-1} \circ \hat{F}_{lj}(X_j^i) - \frac{1}{n_l}\sum_{Y^i=1} \Phi^{-1} \circ \hat{F}_1(X_j^i)$ and $\hat{h}_j^k = \Phi^{-1} \circ \hat{F}_{kj} + \hat{\mu}_{kj}$. With this estimated transformation, one could apply a multiclass sparse LDA method such as the two mentioned above to the pseudo data $(\hat{h}^{\text{pool}}(X), Y)$.

Acknowledgments

Zou's research is partially supported by NSF grant DMS-08-46068 and a grant from Office of Naval Research (grant number N000141110142). The authors thank the editor, associate editor and two referees for constructive suggestions.

Appendix. Proofs

Proof of Lemma 1. First, note that we must have $h_j(X_j) \sim N(0, 1)$. Now we show the uniqueness of h_i. Suppose $h_i^{(1)}$ and $h_i^{(2)}$ are two strictly increasing transformations such that $h_j^{(1)}(X_j) \sim N(0, 1)$, $h_j^{(2)}(X_j) \sim N(0, 1)$. Then for any $t \in \mathbb{R}$, we have

$$\begin{aligned} F_j[\{h_j^{(1)}\}^{-1}(t)] &= \text{pr}[X_j < \{h_j^{(1)}\}^{-1}(t)] = \text{pr}\{h_j^{(1)}(X_j) < t\} = \Phi(t) \\ &= \text{pr}\{h_j^{(2)}(X_i) < t\} = \text{pr}[X_j < \{h_j^{(2)}\}^{-1}(t)) = F_j[\{h_j^{(2)}\}^{-1}(t)]. \end{aligned}$$

Because F_i is strictly monotone, we have that $\{h_j^{(1)}\}^{-1}(t) = \{h_j^{(2)}\}^{-1}(t)$ for all t, which implies $h_j^{(1)} = h_j^{(2)}$. Now note that $\Phi^{-1} \circ F_j$ is a strictly monotone function that transforms X_i to a standard normal random variable, and the conclusion follows. □

Proof of Lemma 2. By (3), conditional on $Y = -1$, we have $\Phi^{-1} \circ F_{+j}(X_j) \sim N(\mu_{-j}, 1)$, while, conditional on $Y = +1$, we have $\Phi^{-1} \circ F_{-j}(X_j) + \mu_{-j} \sim N(0, 1)$. Hence, the conclusions follow. □

The following properties of the normal distribution are repeatedly used in our proof [18,1].

Proposition 1. *Let $\phi(t)$ and $\Phi(t)$ be the pdf and CDF of $N(0, 1)$, respectively.*
1. *For $t \geq 1$, $(2t)^{-1}\phi(t) \leq 1 - \Phi(t) \leq t^{-1}\phi(t)$,*
2. *For $t \geq 0.99$, $\Phi^{-1}(t) \leq [2\log\{(1-t)^{-1}\}]^{1/2}$.*

Define

$$A_n = [-(\gamma_1 \log n)^{1/2}, (\gamma_1 \log n)^{1/2}], \tag{16}$$

where $0 < \gamma_1 < 1$ is a fixed number and n is the sample size. The following lemma shows that $\hat{h}_j(x)$ is an accurate estimator of $h_j(x)$ for $h_j(x) \in A_n$.

Lemma 3. *For sufficiently large n and $0 < \gamma_1 < 1$, we have*

$$\text{pr}\{\sup_{h_j(x)\in A_n} |\hat{h}_j(x) - h_j(x)| \geq \epsilon\} \leq 2\exp\{-n^{1-\gamma_1}\epsilon^2/(32\pi^2\gamma_1 \log n)\} + 2\exp\{-n^{1-\gamma_1}/(16\pi\gamma_1 \log n)\}.$$

Proof of Lemma 3. By mean value theorem,

$$\hat{h}_j(x) - h_j(x) = (\Phi^{-1})'(\xi)\{\hat{F}_j(x) - F_j(x)\},$$

for some $\xi \in [\min\{\hat{F}_j(x), F_j(x)\}, \max\{\hat{F}_j(x), F_j(x)\}]$.

First, we bound $|(\Phi^{-1})'(\xi)|$. This is achieved by bounding $F_j(x)$ and $\hat{F}_j(x)$. By definition, for any $h_j(x) \in A_n$,

$$\begin{aligned} n^{-\gamma_1/2}/\{2(2\pi\gamma_1 \log n)^{1/2}\} &\leq \Phi\{-(\gamma_1 \log n)^{1/2}\} \leq F_j(x) \\ &\leq \Phi\{(\gamma_1 \log n)^{1/2}\} \leq 1 - n^{\gamma_1/2}/\{2(2\pi\gamma_1 \log n)^{1/2}\}. \end{aligned}$$

On the other hand, for x such that $h_j(x) \in A_n$

$$\begin{aligned} \text{pr}[n^{-\gamma_1/2}/\{4(2\pi\gamma_1 \log n)^{1/2}\} &\leq \hat{F}_j(x) \leq 1 - n^{-\gamma_1/2}/\{4(2\pi\gamma_1 \log n)^{1/2}\}] \\ &\geq \text{pr}[\sup_{h_j(x)\in A_n} |\tilde{F}_j(x) - F_j(x)| \leq n^{-\gamma_1/2}/\{4(2\pi\gamma_1 \log n)^{1/2}\}] \\ &\geq 1 - 2\exp\{-n^{1-\gamma_1}/(16\pi\gamma_1 \log n)\}, \end{aligned}$$

where the last inequality follows from Dvoretzky–Kiefer–Wolfowitz (DKW) inequality.

Consequently, with a probability no less than $1 - 2\exp\{-n^{1-\gamma_1}/(16\pi\gamma_1 \log n)\}$,

$$n^{-\gamma_1/2}/\{4(2\pi\gamma_1 \log n)^{1/2}\} \le \xi \le 1 - n^{-\gamma_1/2}/\{4(2\pi\gamma_1 \log n)^{1/2}\},$$

and, combining this fact with Proposition 1, we have

$$\begin{aligned}|(\Phi^{-1})'(\xi)| &= [\phi\{\Phi^{-1}(\xi)\}]^{-1} = (2\pi)^{1/2}\exp\{\Phi^{-1}(\xi)^2/2\}\\ &\le (2\pi)^{1/2}\exp[\log\{4n^{\gamma_1/2}(2\pi\gamma_1 \log n)^{1/2}\}]\\ &= 8\pi n^{\gamma_1/2}(\gamma_1 \log n)^{1/2} \equiv M_n.\end{aligned}$$

Then

$$\mathrm{pr}\{\sup_{h_j(x)\in A_n} |\hat{h}_j(x) - h_j(x)| > \epsilon\} \le \mathrm{pr}\{M_n \sup_{h_j(x)\in A_n} |\hat{F}_j(x) - F_j(x)| > \epsilon\} + 2\exp\{-n^{1-\gamma_1}/(16\pi\gamma_1 \log n)\}.$$

For the first term on the right hand side,

$$\mathrm{pr}\{M_n \sup_{h_j(x)\in A_n} |\hat{F}_j(x) - F_j(x)| > \epsilon\} \le \mathrm{pr}\{M_n \sup_{h_j(x)\in A_n} |\hat{F}_j(x) - \tilde{F}_j(x)| > \epsilon/2\} + \mathrm{pr}\{M_n \sup_{h_j(x)\in A_n} |F_j(x) - \tilde{F}_j(x)| > \epsilon/2\}.$$

Because $\sup_{h_j(x)\in A_n} |\hat{F}_j(x) - \tilde{F}_j(x)| \le \delta_n = 1/n^2$, $\delta_n M_n \to 0$ and so the first term is 0 for sufficiently large n. Apply the DKW inequality to the second term and the conclusion follows. □

The above lemma guarantees that $\hat{h}_j(X_j)$ is very close to $h_j(X_j)$ on A_n. Now we consider observations in A_n^c. Partition A_n^c to three regions:

$$\begin{aligned}B_n &= [-\gamma_2 \log n, -(\gamma_1 \log n)^{1/2}) \cup ((\gamma_1 \log n)^{1/2}, \gamma_2 \log n];\\ C_n &= [-n^{\gamma_3}, -\gamma_2 \log n) \cup (\gamma_2 \log n, n^{\gamma_3}];\\ D_n &= (-\infty, -n^{\gamma_3}) \cup (n^{\gamma_3}, \infty).\end{aligned}$$

Define $\#B_n = \#\{i : h_j(X_j^i) \in B_n\}$ and $\#C_n$, $\#D_n$ analogously.

Lemma 4. *For sufficiently large n and positive constants α_1, α_2 such that $\alpha_1 > 1 - \gamma_1/2$, we have*

$$\sup_{h_j(x)\in B_n} |\hat{h}_j(x) - h_j(x)| \le 2(\log n)^{1/2} + \gamma_2 \log n; \tag{17}$$

$$\sup_{h_j(x)\in C_n} |\hat{h}_j(x) - h_j(x)| \le 2(\log n)^{1/2} + n^{\gamma_3}; \tag{18}$$

$$\mathrm{pr}(\#B_n > n^{\alpha_1}) \le \exp(-n^{2\alpha_1 - 1}/4); \tag{19}$$

$$\mathrm{pr}(\#C_n > n^{\alpha_2}) \le \exp(-n^{2\alpha_2 - 1}/4); \tag{20}$$

$$\mathrm{pr}(\#D_n > 1) \le (2\pi)^{-1/2} 2n^{1-\gamma_3}\exp(-n^{2\gamma_3}/2). \tag{21}$$

Proof of Lemma 4. Eqs. (17)–(18) are direct consequences of the definitions of $\hat{h}$ and B_n, C_n. Indeed, because $\hat{F} < 1 - \delta_n$, by Proposition 1, for $x \in B_n \cup C_n$

$$|\hat{h}_j(x)| \le \Phi^{-1}(1 - \delta_n) \le \{2\log(\delta_n^{-1})\}^{1/2} = 2(\log n)^{1/2}.$$

Combining this bound with the definitions of B_n, C_n, we have the desired conclusions.

For (19), note that, for sufficiently large n,

$$\mathrm{pr}\{h_j(X_j) \in B_n\} \le 2\mathrm{pr}\{h_j(X_j) > (\gamma_1 \log n)^{1/2}\} \le 2^{1/2} n^{-\gamma_1/2}/(\pi\gamma_1 \log n)^{1/2} \le n^{-\gamma_1/2}.$$

Therefore, by Hoeffding's inequality

$$\begin{aligned}\mathrm{pr}(\#B_n > n^{\alpha_1}) &\le \mathrm{pr}\left(\sum_{i=1}^{n}[I\{h_j(X_j^i) \in B_n\} - \mathrm{pr}\{h_j(X_j^i) \in B_n\}] > n^{\alpha_1} - n^{1-\gamma_1/2}\right)\\ &\le \exp\{-n^{2\alpha_1 - 1}(1 - n^{1-\gamma_1/2-\alpha_1})^2/2\} \le \exp(-n^{2\alpha_1 - 1}/4),\end{aligned}$$

for sufficiently large n.

For (20), note that

$$\mathrm{pr}\{h_j(X_j^i) \in C_n\} \le 2n^{-\gamma_2^2 \log n/2}/\gamma_2 \log n.$$

So (20) can be proven similarly.

For (21),

$$\mathrm{pr}(\#D_n > 1) \le 2n\mathrm{pr}\{h_j(X_j^i) > n^{\gamma_3}\} \le 2n^{1-\gamma_3}(2\pi)^{-1/2}\exp(-n^{2\gamma_3}/2). \quad \square$$

Proof of Theorem 1. We first prove (9).

$$\begin{aligned}\mathrm{pr}(|\hat{\mu}_j - \mu_j| > \epsilon) &\le \mathrm{pr}\left\{n^{-1}\sum_{i=1}^n |\hat{h}_j(X_j^i) - h_j(X_j^i)| > \epsilon/2\right\} + \mathrm{pr}\left\{\left|n^{-1}\sum_{i=1}^n h_j(X_j^i) - \mu_j\right| > \epsilon/2\right\}\\ &\equiv L_1 + L_2.\end{aligned}$$

By the Chernoff bound, $L_2 \le 2\exp(-cn\epsilon^2)$.

$$\begin{aligned}L_1 \le\ &\mathrm{pr}\{\sup_{h_j(x)\in A_n} |\hat{h}_j(x) - h_j(x)| > \epsilon/8\} + \mathrm{pr}\{n^{-1}(\#B_n)\sup_{h_j(x)\in B_n}|\hat{h}_j(x) - h_j(x)| > \epsilon/8\}\\ &+ \mathrm{pr}\{n^{-1}(\#C_n)\sup_{h_j(x)\in C_n}|\hat{h}_j(x) - h_j(x)| > \epsilon/8\} + \mathrm{pr}\{n^{-1}(\#D_n)\sup_{h_j(x)\in D_n}|\hat{h}_j(x) - h_j(x)| > \epsilon/8\}.\end{aligned}$$

By Lemma 4, it can be checked that, under Condition (C1), if $\#B_n \le n^{\alpha_1}$ and $\#D_n = 0$ then

$$\mathrm{pr}\{n^{-1}(\#B_n)\sup_{h_j(x)\in B_n}|\hat{h}_j(x) - h_j(x)| > \epsilon/8\} = 0,$$

$$\mathrm{pr}\{n^{-1}(\#D_n)\sup_{h_j(x)\in D_n}|\hat{h}_j(x) - h_j(x)| > \epsilon/8\} = 0,$$

for sufficiently large n. If $\gamma_3 + \alpha_2 < 1$, similarly we have

$$\mathrm{pr}\left\{n^{-1}(\#C_n)\sup_{h_j(x)\in C_n}|\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}\right\} = 0.$$

It follows that, if $\alpha_1 < 1$ and $\gamma_3 + \alpha_2 < 1$, then we have

$$L_1 \le 4\exp(-cn^{1-\gamma_1}\epsilon^2/\gamma_1) + \exp(-cn^{2\alpha_1-1}) + \exp(-cn^{2\alpha_2-1}) + (2\pi)^{-1/2}2n^{1-\gamma_3}\exp(-n^{2\gamma_3}/2),$$

Take $\gamma_1 = 2\rho$, $\alpha_1 = 1 - \rho/2$, $\alpha_2 = 3/4 - \rho/2$, $\gamma_3 = 1/4 - \rho/2$ and the conclusion follows.

Now we prove (10). By the proof in Liu et al. [18], it suffices to bound

$$\mathrm{pr}\left[\left|n^{-1}\sum_{i=1}^n h_j(X_j^i)\{\hat{h}_k(X_k^i) - h_k(X_k^i)\}\right| > \epsilon\right].$$

We can decompose the summation into four terms.

$$\begin{aligned}&n^{-1}\sum_{i=1}^n h_j(X_j^i)\{\hat{h}_k(X_k^i) - h_k(X_k^i)\}\\ &= n^{-1}\left(\sum_{h_j(X_j^i)\in D_n \text{ or } h_k(X_k^i)\in D_n} + \sum_{h_j(X_j^i)\notin D_n, h_k(X_k^i)\in C_n} + \sum_{h_j(X_j^i)\in A_n\cup B_n, h_k(X_k^i)\in B_n} + \sum_{h_j(X_j^i)\in A_n, h_k(X_k^i)\in A_n}\right)[h_j(X_j^i)\{\hat{h}_k(X_k^i) - h_k(X_k^i)\}]\\ &\equiv S_1 + S_2 + S_3 + S_4.\end{aligned}$$

Write $\#D_{nj} = \#\{i : h_j(X_j^i) \in D_n\}$. Then

$$\begin{aligned}\mathrm{pr}(|S_1| > \epsilon) &\le \mathrm{pr}(\#D_{nj} > 1) + \mathrm{pr}(\#D_{nk} > 1)\\ &\le 4n^{1-\gamma_3}(2\pi)^{-1/2}\exp(n^{2\gamma_3}/2).\end{aligned}$$

Note that, for a pair of α_2, γ_3, such that $\alpha_2 + 2\gamma_3 - 1 < 0$, we have $n^{\alpha_2+2\gamma_3-1} \to 0$. Therefore, for sufficiently large n,

$$\begin{aligned}\mathrm{pr}(|S_2| > \epsilon) &\le \mathrm{pr}\left(n^{-1}\sum_{h_k(X_k^i)\in C_n}|\hat{h}_k(X_k^i) - h_k(X_k^i)| > \epsilon/n^{\gamma_3}\right)\\ &\le \mathrm{pr}(\#C_n > n^{\alpha_2}) + \mathrm{pr}[n^{\alpha_2-1}\{2(\log n)^{1/2} + n^{\gamma_3}\} > \epsilon/n^{\gamma_3}]\\ &\le \exp(-n^{2\alpha_2-1}/4) + 0,\end{aligned}$$

Similarly, for $0 < \alpha_1 < 1$,

$$\begin{aligned}\mathrm{pr}(|S_3| > \epsilon) &\le \mathrm{pr}(\#B_n > n^{\alpha_1}) + \mathrm{pr}[n^{\alpha_1-1}(\gamma_2 \log n)\{2(\log n)^{1/2} + \gamma_2 \log n\} > \epsilon] \\ &\le \exp(-n^{2\alpha_1-1}/4) + 0,\end{aligned}$$

where $0 < \alpha_1 < 1$. Finally,

$$\begin{aligned}\mathrm{pr}(|S_4| > \epsilon) &\le \mathrm{pr}\{\sup_{h_k(X_k^i)\in A_n} |\hat{h}_k(X_k^i) - h_k(X_k^i)| > \epsilon(\gamma_1 \log n)^{-1/2}\} \\ &\le 4\exp\{-cn^{1-\gamma_1}\epsilon^2/(\gamma_1^2 \log^2 n)\}.\end{aligned}$$

Pick $\gamma_1 = 2\rho, \gamma_3 = 1/6 - \rho, \alpha_2 = 2/3 - \rho/2, \alpha_1 = 1 - \rho/2$ and the conclusion follows. □

Proof of Corollary 1. Note that n_+ is a summation of n i.i.d random variables with distribution Bernoulli$(1, \pi_+)$. Therefore, by Chernoff bound, there exists $c > 0$ such that $\mathrm{pr}(n_+ > \pi_+ n/2) > 1 - 2\exp(-cn)$. Hence, by Theorem 1,

$$\mathrm{pr}(|\hat{\mu}_{+j} - \mu_{+j}| \ge \epsilon/2) < \zeta_1^*(\pi_+^{1/2}\epsilon/2) + 2\exp(-cn).$$

Similarly,

$$\mathrm{pr}(|\hat{\mu}_{-j} - \mu_{-j}| \ge \epsilon/2) < \zeta_1^*(\pi_-^{1/2}\epsilon/2) + 2\exp(-cn).$$

Hence, we have (11). Eq. (12) can be proven similarly. □

Proof of Theorems 2 and 3. By Mai et al. [20], the consistency is implied by accurate estimators of $\hat{\mu}_y, \hat{\sigma}_{ij}$. Therefore, Theorem 2 can be proven by following the proof in their paper and applying Corollary 1. □

Theorem 3 is direct consequence of Theorem 2. Hence, the proof is omitted here for the sake of space.

Lemma 5. *For any $\epsilon < \min\{\epsilon_0, \lambda/(2\phi\Delta_1), \lambda\}$ and large enough n such that $\epsilon > sn^{-1/4}$, we have*

1.

$$\mathrm{pr}(\|\hat{\beta}_A - \beta_A\|_1 \ge \epsilon) \le 2s^2\zeta_2(\epsilon/s) + 2s\zeta_1(\epsilon/s). \tag{22}$$

2. *If we further assume that $\pi_+, \pi_- > c > 0$, then*

$$\begin{aligned}\mathrm{pr}(|\hat{\beta}_0 - \beta_0| \ge c\epsilon) &\le 2\exp(-cn) + cs\zeta_1[\epsilon/\{s(\phi\Delta_1 + \Delta_2)\}] \\ &\quad + 2p\zeta_1\{\lambda(1-\kappa+2\epsilon\phi)/4(1+\kappa)\} + 2s^2\zeta_2\{\epsilon/(s\Delta_2)\} + 2ps\zeta_2(\epsilon/s)\end{aligned} \tag{23}$$

Proof. We first prove (22). Similar to the proof of Conclusion 3, Theorem 1 in Mai et al. [20], we have

$$\|\hat{\beta}_A - \beta_A\|_1 \le (1 - \eta_1\phi)^{-1}\{\lambda/2 + \phi\|(\hat{\mu}_{+A} - \hat{\mu}_{-A}) - (\mu_{+A} - \mu_{-A})\|_1 + \phi^2\eta_1\Delta_1\} \tag{24}$$

where $\eta_1 = \|\Omega_{AA} - \Omega_{AA}^{(n)}\|_\infty$. Under the events $\eta_1 < \epsilon$ and $\|(\hat{\mu}_{+A} - \hat{\mu}_{-A}) - (\mu_{+A} - \mu_{-A})\|_1 < \epsilon$ we have $\|\hat{\beta}_A - \beta_A\|_1 \le \epsilon$. Hence, (22) follows.

For (23), assume that $\hat{\beta}_{A^C} = 0$. Then we have

$$\begin{aligned}|\hat{\beta}_0 - \beta_0| &= |\{\log(n_+/n_-) - \log(\pi_+/\pi_-)\} - (\hat{\mu}_{+A} + \hat{\mu}_{-A})^{\mathrm{T}}\hat{\beta}_A/2 + (\mu_{+A} + \mu_{-A})^{\mathrm{T}}\beta_A/2| \\ &\le |\log\hat{\pi}_+ - \log\pi_+| + |\log\hat{\pi}_- - \log\pi_-| + |\{(\hat{\mu}_{+A} + \hat{\mu}_{-A}) - (\mu_{+A} + \mu_{-A})\}^{\mathrm{T}}(\hat{\beta}_A - \beta_A)|/2 \\ &\quad + |(\mu_{+A} + \mu_{-A})^{\mathrm{T}}(\hat{\beta}_A - \beta_A)| + |(\mu_{+A} + \mu_{-A})^{\mathrm{T}}(\hat{\beta}_A - \beta_A)|/2.\end{aligned}$$

Under the events $|\hat{\pi}_j - \pi_j| \le \min\{c/2, 2\epsilon/c\}$, $\|\hat{\mu}_{jA} - \mu_{jA}\|_1 \le \epsilon/\phi\Delta_1$ and $\|\hat{\beta}_A - \beta_A\|_1 \le \epsilon/\Delta_2$, we have $|\hat{\beta}_0 - \beta_0| \le c\epsilon$. □

Proof of Theorem 4. Note that

$$\begin{aligned}R_n &\le 1 - \mathrm{pr}(Y = \mathrm{sign}(h(X)^{\mathrm{T}}\beta + \beta_0), \mathrm{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) = \mathrm{sign}(h(X)^{\mathrm{T}}\beta + \beta_0)) \\ &\le R + \mathrm{pr}(\mathrm{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) \ne \mathrm{sign}(h(X)^{\mathrm{T}}\beta + \beta_0)).\end{aligned}$$

Therefore,

$$R_n - R \le \mathrm{pr}(\mathrm{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) \ne \mathrm{sign}(h(X)^{\mathrm{T}}\beta + \beta_0)) \tag{25}$$

$$\begin{aligned}&\le \mathrm{pr}(|h(X)^{\mathrm{T}}\beta + \beta_0| \le \epsilon) \\ &\quad + \mathrm{pr}\left(|(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) - (h(X)^{\mathrm{T}}\beta + \beta_0)| \ge \frac{\epsilon}{2}\right).\end{aligned} \tag{26}$$

Now

$$\mathrm{pr}(|h(X)^{\mathrm{T}}\beta + \beta_0| \le \epsilon) \le \frac{c\epsilon}{\sqrt{2\pi}}. \tag{27}$$

For the second term, assume that $\hat{\beta}_{A^C} = 0$, $|\hat{\beta}_0 - \beta_0| \le c\epsilon$, $\|\hat{\beta}_A - \beta_A\|_1 \le \frac{\epsilon}{\sqrt{\log n}}$ and $\sup_{t\in A_n} |\hat{h}_j(t) - h_j(t)| \le c\frac{\epsilon}{\phi\Delta_1}$ for all j, where A_n is defined as in (16). Then

$$|(\hat{h}(X_A)^{\mathrm{T}}\hat{\beta}_A + \hat{\beta}_0) - (h(X_A)^{\mathrm{T}}\beta_A + \beta_0)| \tag{28}$$

$$\le |\hat{\beta}_0 - \beta_0| + \|\hat{h}(X_A)\|_\infty \|\hat{\beta}_A - \beta_A\|_1 + \|\hat{h}(X_A) - h(X_A)\|_\infty \|\beta_A\|_1 \tag{29}$$

$$\le |\hat{\beta}_0 - \beta_0| + 2\sqrt{\log n}\|\hat{\beta}_A - \beta_A\|_1 + \phi\Delta_1\|\hat{h}(X_A) - h(X_A)\|_\infty, \tag{30}$$

which is smaller than ϵ as long as $h_j(X_j) \in A_n$ for all j. Therefore, take $\gamma_1 = 1/2$ in A_n, we have

$$\mathrm{pr}\left(|(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) - (h(X)^{\mathrm{T}}\beta + \beta_0)| \ge \frac{\epsilon}{2}\right) \le \mathrm{pr}(\cup_{j\in A} h_j(X_j) \in A_n) \le \frac{csn^{-1/4}}{\sqrt{\log n}}, \tag{31}$$

which will be smaller than ϵ for sufficiently large n.

Therefore, by Lemma 3, (22), (23), we have the desired conclusion. □

References

[1] F. Abramovich, Y. Benjamini, D.L. Donoho, I.M. Johnstone, Adapting to unknown sparsity by controlling the false discovery rate, Ann. Statist. 34 (2006) 584–653.
[2] P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of lasso and dantzig selector, Ann. Statist. 37 (2009) 1705–1732.
[3] T. Cai, W. Liu, A direct estimation approach to sparse linear discriminant analysis, J. Amer. Statist. Assoc. 106 (2011) 1566–1577.
[4] X. Chen, Y. Fan, Estimation of copula-based semiparametric time series models, J. Econometrics 130 (2006) 307–335.
[5] X. Chen, Y. Fan, V. Tsyrennikov, Efficient estimation of semiparametric multivariate copula models, J. Amer. Statist. Assoc. 101 (2006) 1228–1240.
[6] L. Clemmensen, T. Hastie, B. Ersbøll, Sparse discriminant analysis, Technometrics 53 (2011) 406–413.
[7] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2004) 407–499.
[8] J. Fan, Y. Fan, High dimensional classification using features annealed independence rules, Ann. Statist. 36 (2008) 2605–2637.
[9] J. Fan, Y. Feng, X. Tong, A ROAD to classification in high dimensional space, J. R. Stat. Soc. Ser. B 74 (2012) 745–771.
[10] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[11] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2008) 1–22.
[12] D.J. Hand, Classifier technology and the illusion of progress, Statist. Sci. 21 (2006) 1–14.
[13] T. Hastie, R. Tibshirani, Discriminant analysis by gaussian mixtures, J. R. Stat. Soc. Ser. B 58 (1996) 155–176.
[14] T. Hastie, R. Tibshirani, J.H. Friedman, Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer Verlag, 2008.
[15] C. Klaassen, J. Wellner, Efficient estimation in the bivariate normal copula model: normal margins are least favourable, Bernoulli 3 (1997) 55–77.
[16] Y. Lin, Y. Jeon, Discriminant analysis through a semiparametric model, Biometrika 90 (2) (2003) 379–392.
[17] H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, High dimensional semiparametric gaussian copula graphical models, Ann. Statist. 40 (2012) 2293–2326.
[18] H. Liu, J. Lafferty, L. Wasserman, The nonparanormal: semiparametric estimation of high dimensional undirected graphs, J. Mach. Learn. Res. 10 (2009) 2295–2328.
[19] J. Lv, Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, Ann. Statist. 37 (2009) 3498–3528.
[20] Q. Mai, H. Zou, M. Yuan, A direct approach to sparse discriminant analysis in ultra-high dimensions, Biometrika 99 (2012) 29–42.
[21] N. Meinshausen, P. Bühlmann, High dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (2006) 1436–1462.
[22] D. Michie, D. Spiegelhalter, C. Taylor, Machine Learning, Neural and Statistical Classification, first ed., Ellis Horwood, 1994.
[23] C.F. Ockenhouse, W.C. Hu, K.E. Kester, J.F. Cummings, A. Stewart, D.G. Heppner, A.E. Jedlicka, A.L. Scott, N.D. Wolfe, M. Vahey, D.S. Burke, Common and divergent immune response signaling pathways discovered in peripheral blood mononuclear cell gene expression patterns in presymptomatic and clinically apparent malaria, Infect. Immun. 74 (2006) 5561–5573.
[24] G. Raskutti, M.J. Wainwright, B. Yu, Restricted eigenvalue properties for correlated gaussian designs, J. Mach. Learn. Res. 11 (2010) 2241–2259.
[25] J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis with high dimensional data, Ann. Statist. 39 (2011) 1241–1265.
[26] P. Song, Multivariate dispersion models generated from gaussian copula, Scand. J. Stat. 27 (2000) 305–320.
[27] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[28] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. Natl. Acad. Sci. 99 (2002) 6567–6572.
[29] N.T. Trendafilov, I.T. Jolliffe, DALASS: variable selection in discriminant analysis via the lasso, Comput. Statist. Data Anal. 51 (2007) 3718–3736.
[30] H. Tsukahara, Semiparametric estimation in copula models, Canad. J. Statist. 33 (2005) 357–375.
[31] S.A. van de Geer, P. Bühlmann, On the conditions used to prove oracle results for the lasso, Electron. J. Stat. 3 (2009) 1360–1392.
[32] M. Wainwright, Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso), IEEE Trans. Inform. Theory 55 (2009) 2183–2202.
[33] D. Witten, R. Tibshirani, Penalized classification using Fisher's linear discriminant, J. R. Stat. Soc. Ser. B 73 (2011) 753–772.
[34] M. Wu, L. Zhang, Z. Wang, D. Christiani, X. Lin, Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection, Bioinformatics 25 (2008) 1145–1151.
[35] L. Xue, H. Zou, Regularized rank-based estimation of high-dimensional nonparanormal graphical models, Ann. Statist. 40 (2012).
[36] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2010) 894–942.
[37] P. Zhao, B. Yu, On model selection consistency of lasso, J. Mach. Learn. Res. 7 (2006) 2541–2567.
[38] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.
[39] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.