https://doi.org/10.1016/j.spl.2015.11.012

High dimensional discrimination analysis via a semiparametric model

Binyan Jiang

Department of Applied Mathematics, Hong Kong Polytechnic University

Chenlei Leng

Department of Statistics, University of Warwick

Abstract

We propose a semiparametric linear programming discriminant (SLPD) rule for high dimensional discriminant analysis under a semiparametric model. As an extension, we further propose a two-stage SLPD (TSLPD) rule, which can have better classification performance under mild sparsity assumptions.

Keywords: Bayes rule, Linear discrimination analysis, Monotone transformation, Semiparametric discriminant analysis, Sparsity.

1. Introduction

High dimension low sample size data sets are frequently encountered nowadays in different fields. However it is known that the statistical analysis of these data sets is very challenging and possibly intractable in some instances. For example, in high dimensional classification, the classical linear discriminant analysis is asymptotically equivalent to random guess even when the Gaussian assumptions are satisfied (Bickel & Levina, 2004). Fortunately, in many situations the data can be assumed to be sparse in that many parameters are close or equal to zero. Motivated by this observation, many approaches are proposed to exploit this sparsity assumption.

Let $X = (x_1, \ldots, x_p)^T$ and $Y = (y_1, \ldots, y_p)^T$ be random variables from two different classes. We shall call these two classes class X and class Y

Email address: by.jiang@polyu.edu.hk (Binyan Jiang)

throughout this paper. Assume the Gaussian model where $X \sim N(\mu_x, \Sigma)$ and $Y \sim N(\mu_y, \Sigma)$. Given a random observation Z from class X or class Y, the well known Bayes rule classifies Z into class X if $[Z-(\mu_x+\mu_y)/2]\Sigma^{-1}(\mu_y-\mu_x) \leq 0$ and into class Y otherwise.

Practically, μ_x, μ_y and Σ are unknown and it is a standard technique to separately estimate μ_x, μ_y and Σ or Σ^{-1} from the sample and plug them into the above Bayes rule. Assuming that both Σ and $\mu = \mu_y - \mu_x$ are sparse, Shao et al. (2011) used thresholding procedures for estimating Σ and μ. By noticing that the Bayes rule depends on Σ and μ only through $\beta = \Sigma^{-1}\mu$, instead of estimating Σ^{-1} and μ separately, Cai & Liu (2011) obtained sparse estimators for β directly. Other approaches for sparse linear discriminant analysis under multivariate normal assumptions can be found in Fan et al. (2012), Mai et al. (2012) and the references therein.

A limitation of the linear discriminant rules is the normality assumption. When p is fixed, Lin & Jeon (2003) considered the so-called transnormal or nonparanormal distribution to allow the marginal distributions unspecified, as discussed in the next subsection; see also Kon & Nikolaev (2011). In this paper, we consider discriminant analysis under this generalized distribution when the dimension p far exceeds the sample size n but grows slower than $\exp(n^{1/2})$. We derive the Bayes rule under this semiparametric model and propose estimators for its components. We show that the risk of our classification rule tends to the Bayes risk in probability.

1.1. A semiparametric model

We begin by introducing some notations. For any matrix M, write M^T as the transpose of M. Let $v = (v_1, \ldots, v_p)^T \in \mathcal{R}^p$ be a p-dimensional vector. Define $|v|_0 = \sum_{i=1}^p I_{\{v_i \neq 0\}}$ and $|v|_\infty = \max_{1 \leq i \leq p} |v_i|$. For any $1 \leq q < \infty$, the l_q norm of v is defined as $|v|_q = (\sum_{i=1}^q |v_i|^q)^{1/q}$. We denote the p-dimensional vector of ones as 1_p and the p-dimensional vector of zeros as 0_p.

Following Lin & Jeon (2003), we say a random vector $V = (V_1, \ldots, V_p)^T$ has a transnormal distribution $TN(h, \mu, 1_p, \Gamma)$ if there exists a set of univariate strictly monotone and differentiable functions $h = (h_1, \ldots, h_p)^T$ such that $h(V) = (h_1(V_1), \ldots, h_p(V_p))^T$ is multivariate normal with mean $\mu = (\mu_1, \ldots, \mu_p)^T$ and correlation matrix $\Gamma = (\gamma_{ij})_{p \times p}$.

The transnormal distribution is also called the nonparanormal distribution in some recent literature and is also related to the Gaussian copula model; see for example Liu et al. (2009). Denote the density functions of X and Y as f_X and g_Y respectively. In this paper we assume that

$X \sim TN(h, \mu_x, 1_p, \Gamma)$ and $Y \sim TN(h, \mu_y, 1_p, \Gamma)$. Without loss of generality we assume that $\mu_x = (0, \ldots, 0)^T$, $\mu_y = \mu = (\mu_1, \ldots, \mu_p)^T$. Therefore $h_i(x_i) \sim N(0,1)$, $h_i(y_i) \sim N(\mu_i, 1)$, and we immediately have

$$h_i = \Phi^{-1} \circ F_i = (\Phi^{-1} \circ G_i) + \mu_i, \quad 1 \leq i \leq p, \tag{1}$$

where $\circ$ denotes the composition of functions, Φ is the univariate standard Gaussian cumulative distribution function, F_i is the cumulative distribution function of x_i and G_i is the cumulative distribution function of y_i. This is a sub model of the functional analysis of variance model; see for example Lin & Jeon (2003) for more discussion. In addition, when $X \sim N(\mu_x, \Sigma)$ and $Y \sim N(\mu_y, \Sigma)$, model (1) is satisfied with $\mu = \mu_y - \mu_x$.

1.2. Discriminant analysis through the semiparametric model

Suppose h, μ and Γ are known and let $Z = (z_1, \ldots, z_p)^T$ be an independent observation from class X or class Y. Under the semiparametric model introduced in the last subsection, the well known Bayes procedure yields a classification rule that classifies Z to class X if and only if $D_L(Z) \leq 0$ where

$$D_L(Z) = \{h(Z) - \mu/2\}^T \Gamma^{-1} \mu. \tag{2}$$

This is in fact equivalent to applying Fisher's LDA to the transformed data $h(Z), h(X)$ and $h(Y)$ and the misclassification rate of this rule is seen as

$$R = \Phi(-\Delta_p/2), \quad \text{where } \Delta_p = \sqrt{\mu^T \Gamma^{-1} \mu}. \tag{3}$$

When p is bounded, what we introduced above is similar to Case 1 in Lin & Jeon (2003). We now discuss the estimation of the components in $D_L(Z)$ when p is very large. Noting that the discrimination rule $D_L(Z)$ depends on Γ and μ only through the product $\Gamma^{-1}\mu$, we propose to estimate $\beta := \Gamma^{-1}\mu$ by the Dantzig selector in Candes & Tao (2007) and Cai & Liu (2011) as

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{R}^p} \{|\beta|_1 \ \text{ subject to } \ |\hat{\Gamma}\beta - \hat{\mu}|_\infty \leq \lambda_n\}, \tag{4}$$

where λ_n is a tuning parameter, $\hat{\Gamma}$ and $\hat{\mu}$ are estimators of Γ and μ defined in Section 2. On the other hand, we estimate $h(Z) - \mu/2$ using $\tilde{h}_Z$ as in (7). We then classify Z to class X if $\tilde{h}_Z^T \hat{\beta} \leq 0$, and to class Y if $\tilde{h}_Z^T \hat{\beta} > 0$. We shall call this the *Semiparametric Linear Programming Discriminant* (SLPD) rule. Note from (3) that the Bayes risk is independent of h. Consistent to this, the SLPD rule is invariant about h; see Proposition 1.

While we are finishing this paper, we found that the semiparametric model in this paper is also studied in Han et al. (2013) and Mai & Zou (2013), but with key differences. Our method and assumptions are different from those in Han et al. (2013) and Mai & Zou (2013). Under the semiparametric model, we directly estimate the Bayes rule, while Mai & Zou (2013) made use of an equivalent least square formulation for estimating β and Han et al. (2013) is based on the regularized optimal affine discriminant analysis in Fan et al. (2012). In terms of estimation method, we use median in estimating μ and use Dantzig selector in estimating β. In terms of assumptions, we do not assume the irrepresentable condition (Zhao & Yu, 2006); see for example Definition 8 of Han et al. (2013) and (18) of Mai & Zou (2013). This condition is known to be sufficient for selecting the zero entries in β consistently in theory, but can be easily violated in practice (Zhao & Yu, 2006). What is more, our sparsity assumption on β is more general; see (12) in Theorem 1. More specifically, we do not require the number of nonzero elements of β to be relatively small, while Han et al. (2013) and Mai & Zou (2013) considered the case that the number of nonzero elements of β is much smaller than n. Last but not least, we allow the logarithm of the dimension to grow slower than the square root of sample size while Mai & Zou (2013) requires that the logarithm dimension grows slower than the cube root of n.

2. Estimation method

Assume that $X_i = (X_{i1}, \ldots, X_{ip})^T, 1 \leq i \leq n_1$ and $Y_i = (Y_{i1}, \ldots, Y_{ip})^T, 1 \leq i \leq n_2$ are independently identically distributed random vectors from class X and class Y respectively. Denote the total sample size as $n = n_1 + n_2$. Throughout this paper we make the following assumptions.

Assumption 1. There exists a constant $\lambda > 0$ such that $\lambda^{-1} \leq \lambda_1(\Gamma) \leq \lambda_p(\Gamma) \leq \lambda$, where $\lambda_1(\Gamma)$ and $\lambda_p(\Gamma)$ are the smallest and largest eigenvalues of Γ respectively. In addition, there exists a constant $B > 0$ such that $\Delta_p > B$.

Assumption 2. The sample size satisfies $n_1 \asymp n_2$ and $\log p = o(n^{1/2})$.

We use the Winsorized estimator (Liu et al., 2009) in estimating F_i:

$$\tilde{F}_i(t) = \begin{cases} \delta_n^x & \text{if } \hat{F}_i(t) < \delta_n^x \\ \hat{F}_i(t) & \text{if } \delta_n^x \leq \hat{F}_i(t) \leq 1 - \delta_n^x \\ 1 - \delta_n^x & \text{if } \hat{F}_i(t) > 1 - \delta_n^x, \end{cases} \tag{5}$$

where $\hat{F}_i(t) = \frac{1}{n_1} \sum_{j=1}^{n_1} \mathcal{I}_{\{X_{ji} \leq t\}}$, and δ_n^x is a truncation parameter. Here the truncation is used to avoid infinity value and clearly there is a bias-variance

tradeoff in choosing δ_n^x. We then define $\hat{h}_i^x = \Phi^{-1} \circ \tilde{F}_i, i = 1, \ldots, p$. Similarly, let $\hat{G}_i = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathcal{I}_{\{Y_{ji} \le t\}}$. We estimate G_i by

$$\tilde{G}_i(t) = \begin{cases} \delta_n^y & \text{if } \hat{G}_i(t) < \delta_n^y \\ \hat{G}_i(t) & \text{if } \delta_n^y \le \hat{G}_i(t) \le 1 - \delta_n^y \\ 1 - \delta_n^y & \text{if } \hat{G}_i(t) > 1 - \delta_n^y, \end{cases} \tag{6}$$

and we define $\hat{h}_i^y = \Phi^{-1} \circ \tilde{G}_i, i = 1, \ldots, p$. Let $\tilde{F}_i^*(t) = \tilde{F}_i(t)$ with $\delta_n^x = 1/(2n_1)$ and $\tilde{G}_i^*(t) = \tilde{G}_i(t)$ with $\delta_n^y = 1/(2n_2)$. Define

$$\hat{\mu}^x = (\hat{\mu}_1^x, \ldots, \hat{\mu}_p^x)^T, \quad \hat{\mu}_i^x = \text{median}\{\Phi^{-1}(\hat{F}_i^*(Y_{ji})), j = 1, \ldots, n_2\},$$
$$\hat{\mu}^y = (\hat{\mu}_1^y, \ldots, \hat{\mu}_p^y)^T, \quad \hat{\mu}_i^y = \text{median}\{\Phi^{-1}(\hat{G}_i^*(X_{ji})), j = 1, \ldots, n_1\}.$$

We then estimate $h(z) - \mu/2$ in (2) by $\tilde{h}_Z = (\tilde{h}_Z^1, \ldots, \tilde{h}_Z^p)^T$ where

$$\tilde{h}_Z^i := \alpha(\hat{h}_i^x(z_i) - \hat{\mu}_i^x/2) + (1 - \alpha)(\hat{h}_i^y(z_i) - \hat{\mu}_i^y/2), \quad i = 1, \ldots, p. \tag{7}$$

Here α can be any arbitrary constant in $[0, 1]$ and a natural choice of α is $\alpha = n_1/n$. We estimate μ by $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^T$ where

$$\hat{\mu}_i = \alpha\hat{\mu}_i^x - (1 - \alpha)\hat{\mu}_i^y, \quad i = 1, \ldots, p. \tag{8}$$

Let $\hat{r}_{ij}^x$ be the Spearman's rank correlations between $(X_{1i}, \ldots, X_{n_1 i})$ and $(X_{1j}, \ldots, X_{n_1 j})$ and let $\hat{r}_{ij}^y$ be the Spearman's rank correlations between $(Y_{1i}, \ldots, Y_{n_2 i})$ and $(Y_{1j}, \ldots, X_{n_2 j})$. We estimate Γ using the adjusted Spearman's rank correlations similar to Xue & Zou (2012):

$$\hat{\Gamma} = (\hat{\gamma}_{ij})_{1 \le i,j \le p}, \qquad \hat{\gamma}_{ij} = 2\alpha \sin\left(\frac{\pi}{6}\hat{r}_{ij}^x\right) + 2(1 - \alpha) \sin\left(\frac{\pi}{6}\hat{r}_{ij}^y\right). \tag{9}$$

We estimate $\beta = \Gamma^{-1}\mu$ by (4) with $\hat{\mu}$ and $\hat{\Gamma}$ defined as in (8) and (9), and then the SLPD rule can be computed using (4) and (7). Next we provide some theoretical results to evaluate our estimation procedure and the misclassification rate of the proposed discriminant rule. Proofs of these results are provided in the supplementary material. Notice from (3) that the Bayes risk depends on μ and Γ only. Consistent to this, we have

Propositon 1. *Let $h = (h_1, \ldots, h_p)^T$ be a monotone function defined as in the transnormal distribution. Given λ_n, the SLPD rule is invariant to h. More specifically, suppose $U_1, \ldots, U_{n_1}$ are independent samples from $N(0_p, \Gamma)$ and $V_1, \ldots, V_{n_2}$ are independent samples from $N(\mu, \Gamma)$. Given Z, the SLPD rules for the following two cases are the same: (i) $X_i = U_i, i = 1, \ldots, n_1$, $Y_i = V_i, i = 1, \ldots, n_2$; (ii) $X_i = h(U_i), i = 1, \ldots, n_1$, $Y_i = h(V_i), i = 1, \ldots, n_2$.*

Lemmas 1, 2 and Proposition 2 given below indicate that $\hat{h}^x$, $\hat{h}^y$, $\hat{\mu}$ and $\hat{\Gamma}$ can estimate $h, h-\mu, \mu$ and Γ well when n and p are large enough.

Lemma 1. *Assume that $|\mu|_\infty < U < \infty$ for some constant U. Suppose $Z = (z_1, \ldots, z_p)$ is either a random sample from the X class or the Y class. By choosing $\delta_n^x = \sqrt{\frac{M \log p}{2n_1}} + \Phi(-\sqrt{2M \log p})$ for some constant $M > 1$, we have that, when n_1 and p are large enough, for any $1 \le i \le p$, there exists a constant H large enough such that*

$$E|\hat{h}_i^x(z_i) - h_i(z_i)| I_{\{|h_i(z_i)-u_i| \le \sqrt{2M \log p},\ |\hat{F}_i(z_i)-F_i(z_i)| \le \sqrt{\frac{M \log p}{2n_1}}\}} \le \frac{H \log p}{\sqrt{n}}, \quad (10)$$

where $I_{\{\cdot\}}$ is the indicator function and u_i equals 0 if Z is from the X class and equals μ_i if Z is from the Y class. Similarly, by choosing $\delta_n^y = \sqrt{\frac{M \log p}{2n_2}} + \Phi(-\sqrt{2M \log p})$ for some constant $M > 1$, we have

$$E|\hat{h}_i^y(z_i) - h_i(z_i) - \mu_i| I_{\{|h_i(z_i)-u_i| \le \sqrt{2M \log p},\ |\hat{G}_i(z_i)-G_i(z_i)| \le \sqrt{\frac{M \log p}{2n_2}}\}} \le \frac{H \log p}{\sqrt{n}}. \quad (11)$$

From Lemma 1 of Xue & Zou (2012), the following can be easily shown.

Lemma 2. *Let $\hat{\Gamma}$ be defined as in (9). For any $0 < \epsilon < 1$, there exists a positive constant c_0 such that when $n \ge 24\pi/\epsilon$,*

$$P(|\hat{\gamma}_{ij} - \gamma_{ij}| > \epsilon) \le 2 \exp(-c_0 n \epsilon^2).$$

Propositon 2. *Let $\hat{\mu}$ be defined as in (8). Assume that $|\mu|_\infty < U < \infty$ for some constant U. Then for any $1 \le i \le p$ and $\epsilon > \max\{(2n_1)^{-1}, (2n_2)^{-1}\}$ such that $\epsilon \to 0$ as $n, p \to \infty$, there exist constants $c_1 > 0, c_2 > 0$ such that*

$$P(|\hat{\mu}_i - \mu_i| > \epsilon) \le c_1 n \exp(-c_2 n \epsilon^2).$$

Proposition 2 implies that our estimator $\hat{\mu}$ can estimate μ very well even when $p \to \infty$. The reason of using median (see the definition of $\hat{\mu}^x$ and $\hat{\mu}^y$) instead of mean as in Mai et al. (2012) is that the function $\Phi^{-1}(t)$ is very unstable in that a small change in the value of t when t is close to 0 or 1 would cause a large change in the value of $\Phi^{-1}(t)$. Known as a robust estimator, median is a natural choice here. In terms of theoretical results, to estimate the p elements in μ well simultaneously, Mai et al. (2012) assume that $\log p = o(n^{1/3})$ while from Proposition 2 we can have $\log p = o(n)$.

The next lemma shows that the true $\beta = \Gamma^{-1}\mu$ belongs to the feasible set of (4) with overwhelming probability.

Lemma 3. *Under the assumptions of Lemma 2 and Proposition 2, for any constant $M > 0$, by choosing $\lambda_n = C|\beta|_1\sqrt{(\log p + \log n)/n}$ for some constant C large enough, we have with probability greater than $1 - \mathrm{O}(p^{-M})$,*

$$|\hat{\Gamma}\beta - \hat{\mu}|_\infty \le \lambda_n.$$

Given the sample $\{X_i, Y_j : 1 \le i \le n_1, 1 \le j \le n_2\}$, the conditional misclassification rate of SLPD is seen as

$$\begin{aligned} R_n \;=\; & \frac{1}{2}P\Big\{\frac{(h(Z)-\mu)^T\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}} \le \alpha\Big[\frac{(\hat{\mu}^x/2-\mu)^T\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}} - \frac{\{\hat{h}^x(Z)-h(Z)\}^T\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}}\Big] \\ & +(1-\alpha)\Big[\frac{\hat{\beta}^T\hat{\mu}^y/2}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}} - \frac{\{\hat{h}^y(Z)-h(Z)+\mu\}^T\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}}\Big]\Big|h(Z)\sim N(\mu,\Gamma)\Big\} \\ & +\frac{1}{2}P\Big\{\frac{h(Z)^T\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}} > \alpha\Big[\frac{\hat{\beta}^T\hat{\mu}^x/2}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}} - \frac{\{\hat{h}^x(Z)-h(Z)\}\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}}\Big] \\ & (1-\alpha)\Big[\frac{\hat{\beta}^T(\hat{\mu}^y/2+\mu)}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}} - \frac{\{\hat{h}^y(Z)-h(Z)+\mu\}\hat{\beta}}{\sqrt{\hat{\beta}^T\Gamma\hat{\beta}}}\Big]\Big|h(Z)\sim N(0,\Gamma)\Big\}, \end{aligned}$$

where $\hat{h}^x = (\hat{h}^x_1, \ldots, \hat{h}^x_p)^T$ and $\hat{h}^y = (\hat{h}^y_1, \ldots, \hat{h}^y_p)^T$. The following theorem shows that the conditional misclassification rate R_n tends to R in probability.

Theorem 1. *Assume that $|\mu|_\infty < U < \infty$ for some constant U and*

$$\frac{|\beta|_1^2}{\Delta_p^2} = o\Big(\frac{\sqrt{n}}{\log p}\Big). \tag{12}$$

By choosing $\delta_n^x = \sqrt{\frac{(M+1)\log p}{2n_1}} + \Phi(-\sqrt{2(M+1)\log p})$, $\delta_n^y = \sqrt{\frac{(M+1)\log p}{2n_2}} + \Phi(-\sqrt{2(M+1)\log p})$ and $\lambda_n = C|\beta|_1\sqrt{(\log p + \log n)/n}$ for some constants $M > 0$ and C large enough, we have with probability tending to 1,

$$R_n - R \to 0, \quad \text{as} \;\; n \to \infty.$$

The sparse assumption (12) here is more general than the assumptions in Han et al. (2013) and Mai et al. (2012) in that it allows the case where β is only approximately sparse. What is more, from Cauchy-Schwarz inequality and

Assumption 1 we have $\frac{|\beta|_1^2}{\Delta_p^2} \leq \frac{|\beta|_0|\beta|_2^2}{\Delta_p^2} \leq \frac{|\beta|_0|\lambda^2|\mu|_2^2}{\lambda^{-2}|\mu|_2^2} = \lambda^4|\beta|_0$. Therefore, $|\beta|_0 = o(\frac{\sqrt{n}}{\log p})$ implies (12). In the case where β_i's are very small, we can allow β to be nonsparse. On the other hand, from Conditions 3-5 and Theorem 15 in Han et al. (2013), we can see that they roughly require $|\beta|_0^2 = o(\sqrt{\frac{n}{\log p}})$; From condition (C1) of Mai et al. (2012) we can see that they require $|\beta|_0^2 = o(\frac{n^{1/3}}{\log p})$.

3. A two-stage procedure

As an extension of our SLPD rule introduced above, we propose a Two-stage SLPD (TSLPD) rule. TSLPD operates by first screening out variables via 4, then retaining the remaining variables that pass a threshold. Then the classification is conducted by using SLPD again. This technique was first pointed out by Candes & Tao (2007) in linear regression. See also Wang et al. (2013). For any vector $v = (v_1, \ldots, v_p)^T$ and an index set S, let v_S denote a new vector such that $v_S = (v_i : i \in S)^T$. Similarly for any matrix $\Gamma = (\gamma_{i,j})_{1\leq i,j\leq p}$ we define $\Gamma_S = (\gamma_{i,j})_{i,j\in S}$.

Throughout this section we assume that $\frac{\beta_i^2}{\Delta_p^2} \gg \frac{|\beta|_1^2}{\Delta_p^2}\sqrt{\frac{n}{\log q+\log n}} \gg \eta_S$ for any $i \in S$ where $\eta_S = \frac{\sum_{i\in S^C}\beta_i^2}{\Delta_p^2} \to 0$. Let p_0 be the size of S. For a new observation Z, the TSLPD rule is given below:

(i) Let $\hat{\beta}$ be the estimator of β obtained using (4) and for any $q \geq p_0$, let $\hat{S}_q = \{i : \sum_{j=1}^p I_{\{|\hat{\beta}_i|>|\hat{\beta}_j|\}} < q\}$ be the set of indices of the q largest $|\hat{\beta}_i|$'s.

(ii) Let $X_{i\hat{S}_q}$ and $Y_{j\hat{S}_q}$ be the sub-vectors of X_i and Y_j with features indexed by $\hat{S}_q$ for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. We then implement our SLPD rule to $Z_{\hat{S}_q}$ based on $X_{1\hat{S}_q}, \ldots, X_{n_1\hat{S}_q}$ and $Y_{1\hat{S}_q}, \ldots, Y_{n_2\hat{S}_q}$.

Propositon 3. *Assume that there exists an index set $S \subset \{1, 2, \ldots, p\}$, such that $\eta_S = \sum_{i\in S^C}\beta_i^2 \to 0$. Under Assumption 1, we have: $\mu_S^T\Gamma_S^{-1}\mu_S = \mu^T\Gamma^{-1}\mu + O(\eta_S\Delta_p^2)$. In particular, when $\eta_S = 0$, we have $\mu_S^T\Gamma_S^{-1}\mu_S = \mu^T\Gamma^{-1}\mu$.*

This proposition is similar to propositions 1 and 2 in Wang et al. (2013). It implies that if we only consider the important features, the change in the Bayes risk is negligible. The following theorem further indicate that stage (i) of our TSLPD rule is able to estimate S consistently under mild conditions.

Theorem 2. *Assume that*

$$\frac{|\beta|_1^2}{\Delta_p^2} = o\Big(\sqrt{\frac{n}{\log q + \log n}}\Big). \tag{13}$$

Let λ_1, λ_2 be the tuning parameters in the two stages of our TSLPD rule. For any constant $M > 0$, by choosing $\lambda_1 = C|\beta|_1\sqrt{(\log p + \log n)/n}$ and $\lambda_2 = C|\beta|_1\sqrt{(\log q + \log n)/n}$ for some constant C large enough, we have $P(S \subseteq \hat{S}_q) = 1 - O(p^{-M})$. In particular, when $q = p_0$, we have $P(\hat{S}_{p_0} = S) = 1 - O(p^{-M})$.

Theorem 2 indicates that the first step of our TSLPD rule is able to select the important features for $q > p_0$ and $\hat{S}_{p_0}$ is consistent in estimating S. Clearly, when $q = p$, the TSLPD rule reduces to the SLPD rule. In practice, p_0 can be far smaller than the original dimension p. By choosing $q = p_0$ or slightly larger than p_0, the first stage can screen out a large number of irrelevant variables, resulting in better classification results in the second step comparing to the one-stage SLPD rule. In addition, from Theorems 1 and 2, we can also see that theoretically, the TSLPD rule can further improved the diverging rate of $\frac{|\beta|_1^2}{\Delta_p^2}$ to $o\Big(\sqrt{\frac{n}{\log p_0 + \log n}}\Big)$.

4. Numerical study

Let $u = (u_1, \ldots, u_p)^T \in \mathbf{R}^p$ and let γ_i be the i-th row of $\hat{\Gamma}$. The convex optimization problem (4) can be implemented via linear programming as

$$\min \sum_{i=1}^{p} u_i \ \text{ subject to } u_i \le \beta_i \le u_i \text{ and } -\lambda_n \le \gamma_i^T \beta - \hat{\mu}_i \le \lambda_n, \ i = 1, \ldots, p,$$

This is similar to the implementation of the Dantzig selector; see example Candes & Tao (2007) and Cai & Liu (2011). The tuning parameter λ_n in (4) is chosen using K-folder cross validation. More specifically, randomly divide the index sets $\{1, \ldots, n_1\}$ into K subgroups $N_{11}, \ldots, N_{1K}$, and divide $\{1, \ldots, n_2\}$ into K subgroups $N_{21}, \ldots, N_{2K}$. Denote the full sample set as $S = \{X_i, Y_j : 1 \le i \le n_1, 1 \le j \le n_2\}$ and let $S_k = \{X_i, Y_j : i \in N_{1k}, Y \in N_{2k}\}$ for $k = 1, \ldots, K$. For a given λ_n and an observation Z, let $\hat{\beta}^{(k)}$ and $\tilde{h}_Z^{(k)}$ be defined as in (4) and (7) based on $S \setminus S_k$. For each $k = 1, \ldots, K$, let $C_{1k} = \sum_{i \in N_{1k}} I_{\{\tilde{h}_{X_i}^{(k)T} \hat{\beta}^{(k)} \le 0\}}$ and $C_{2k} = \sum_{i \in N_{2k}} I_{\{\tilde{h}_{Y_i}^{(k)T} \hat{\beta}^{(k)} > 0\}}$. We then find λ_n such that it maximizes the averaged correct classification number: $CV(\lambda_n) = \frac{1}{K}\sum_{k=1}^{K}(C_{1k} + C_{2k})$.

In our simulation study, we set α in (7), (8) and (9) to be n_1/n. One can also choose α using cross validation. However, according to the simulation

Table 1: Simulation results under Models 1-3

	Model 1 R=0.101			Model 2 R=0.093			Model 3 R=0.127		
$h^i_{inv}(t)$	R^{slpd}	R^{Tslpd}	R^{lpd}	R^{slpd}	R^{Tslpd}	R^{lpd}	R^{slpd}	R^{Tslpd}	R^{lpd}
t	0.185	0.189	0.178	0.126	0 .118	0.125	0.155	0.152	0.144
t^3	0.182	0.183	0.227	0.126	0.121	0.156	0.173	0.164	0.293
e^t	0.184	0.187	0.232	0.127	0.120	0.197	0.162	0.148	0.233

R: Bayes risk; R^{lpd}: misclassification rates for the LPD rule;
R^{slpd}: misclassification rates for the SLPD rule.
R^{Tslpd}: misclassification rates for the TSLPD rule.

we have done, there is no significant improvement in choosing α using cross validation than simply setting $\alpha = n_1/n$.

Next we provide some numerical results on two simulation studies. In the first simulation study, we compare our SLPD rule to the LPD rule in Cai & Liu (2011). We consider the following models.

MODEL 1. $n_1 = n_2 = 50, p = 100$, $\Sigma = (\sigma_{i,j})_{p\times p}$ where $\sigma_{ij} = 0.6^{|i-j|}, 1 \leq i, j \leq p$ and $\beta = (\beta_1, \ldots, \beta_p)^T$ with $\beta_i = 0.129, 1 \leq i \leq p$.

MODEL 2. $n_1 = n_2 = 100, p = 200$, $\Sigma = (\sigma_{i,j})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}, 1 \leq i, j \leq p$, $\beta = (\beta_1, \ldots, \beta_p)^T$ with $\beta_i = 0.4$ if $1 \leq i \leq 10$, $\beta_i = 0.3$ if $11 \leq i \leq 20$ and $\beta_i = 0$ otherwise.

MODEL 3. $n_1 = n_2 = 50, p = 100$, $\Sigma = (\sigma_{i,j})_{p\times p}$ where $\sigma_{ii} = 1$ for $1 \leq i \leq p$ and $\sigma_{ij} = 0.5$ otherwise, $\beta = (\beta_1, \ldots, \beta_p)^T$ with $\beta_i = 0.3$ if $1 \leq i \leq 10$, $\beta_i = 0.001$ if $11 \leq i \leq p$.

The parameter β in Model 1 is sparse in a way that every β_i has the same weak strength. β in Model 2 is sparse such that most of the β_i's equal zero. In Model 3, β is sparse in a sense that only a few β_i has strong signal strength while others are (nonzero) weak signals. We generate n_1 independent samples $U_1 \ldots, U_{n_1}$ from $N(0_p, \Sigma)$, and n_2 independent samples $V_1 \ldots, V_{n_2}$ from $N(\mu, \Sigma)$ where $\mu = \Sigma\beta$. We then set $X_i = h_{inv}(U_i), 1 \leq i \leq n_1$ and $Y_i = h_{inv}(V_i), 1 \leq i \leq n_2$, where $h_{inv}(x) = (h^1_{inv}, \ldots, h^p_{inv})^T$ is the set of inverse functions of $h = (h_1, \ldots, h_p)^T$ needed in the transnormal distribution. In this simulation we set $h^1_{inv}(t) = \cdots = h^p_{inv}(t)$ and consider the following three cases: $h^i_{inv}(t) = t, t^3$ or e^t. Note that $h^i_{inv}(t) = t$ implies the Gaussian assumption is satisfied. We then generate a random sample W_1 from $N(0_p, \Sigma)$ and use the SLPD rule and the LPD rule in classifying $Z = h_{inv}(W_1)$. The above procedure is repeated 1000 times and we define C_1^{slpd} as the number of times of classifying Z to Class X using the SLPD rule

and C_1^{lpd} as the number of times of classifying Z to Class X using the LPD rule. Similarly, we generate $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ as above and use the SLPD rule and LPD rule to classify $Z = h_{inv}(W_2)$ where W_2 is a random sample from $N(\mu, \Sigma)$. This again is repeated for 1000 times and we define C_2^{slpd} as the number of times of classifying Z to Class Y using the SLPD rule and C_2^{lpd} as the number of times of classifying Z to Class Y using the LPD rule. The misclassification rates are then given by

$$R^{slpd} = 1 - (C_1^{slpd} + C_2^{slpd})/2000, \qquad R^{lpd} = 1 - (C_1^{lpd} + C_2^{lpd})/2000.$$

R^{Tslpd} is defined similarly for the TSLPD rule. q in the first step and λ_2 in the second step of the TSLPD rule are chosen using the cross validation method described as in the beginning of this section. Tuning parameter λ_1 in the first step is of secondary importance and is simply set it to be $\sqrt{(\log p + \log n)/n}$.

Results of this simulation are given in Table 1, where we can observe that the SLPD rule and the TSLPD rule have very good overall performance. More specifically, when the Gaussian assumption is violated (when $h_{inv}^i(t) = t^3$ or e^t), R^{lpd} is clearly larger than the R^{slpd} under Models 1-3, indicating that the SLPD rule is outperforming the LPD rule in these cases. When the Gaussian assumption is satisfied (when $h_{inv}^i(t) = t$), R^{lpd} and R^{slpd} are comparable. In addition, we see clearly that the misclassification error of SLPD is similar across different h_{inv}, confirming the theoretical result in Proposition 1. These results suggest that the SLPD rule can be an alternative choice besides the LPD rule in high dimensional sparse discriminant analysis. We have also tried the methods in Han et al. (2013) and Mai & Zou (2013) and the results are very similar to those of SLPD. On the other hand, comparing with the SLPD rule, the TSLPD rule does improve the classification performance under Models 2 and 3, where β is sparse. The sparse assumptions for the TSLPD rule are violated in Model 1. However, classification results using the TSLPD rule are still comparable to the results using the SLPD rule.

In the next simulation, we study the asymptotic properties of the SLPD rule as in Theorem 1 under different scenarios. By Proposition 1, we only have to consider the Gaussian case (where $h_{inv}^i(t) = t$). We consider the following three models.

Model 4. $p = 64$, $\Sigma = (\sigma_{i,j})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}, 1 \le i, j \le p$ and $\beta = (\beta_1, \ldots, \beta_p)^T$ with $\beta_i = 0.502$ if $1 \le i \le 10$ and $\beta_i = 0$ otherwise. We set $n_1 = n_2 = 36, 108, 180, 252$.

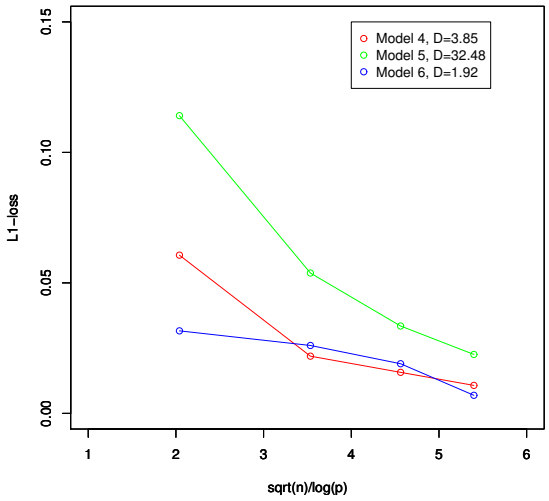

Figure 1: Plot of the L1-loss versus the rescaled sample size $\sqrt{n}/\log p$

MODEL 5. $p = 128$, $\Sigma = (\sigma_{i,j})_{p\times p}$ where $\sigma_{ij} = 0.6^{|i-j|}, 1 \le i, j \le p$ and $\beta = (\beta_1, \ldots, \beta_p)^T$ with $\beta_i = 0.114, 1 \le i \le p$. We set $n_1 = n_2 = 49, 147, 245, 343$.

MODEL 6. $p = 256$, $\Sigma = (\sigma_{i,j})_{p\times p}$ where $\sigma_{ii} = 1$ for $1 \le i \le p$ and $\sigma_{ij} = 0.5$ otherwise, $\beta = (\beta_1, \ldots, \beta_p)^T$ with $\beta_i = 0.165$ if $1 \le i \le 20$, $\beta_i = 0.001$ if $21 \le i \le p$. We set $n_1 = n_2 = 64, 192, 320, 448$.

In each model we consider four cases by increasing the sample size n_1 and n_2. Define the rescaled sample size as $\sqrt{n}/\log p$. The n_1, n_2 values in Models 4-6 are chosen such that the $\sqrt{n}/\log p = 2.04, 3.53, 4.56, 5.40$ in each model. For each case in each model, the data generating procedure is the same as the previous simulation except that the repeated times for calculating C_1^{slpd} and C_2^{slpd} are set to be 500. Hence the estimated misclassification rate is given as $R^{slpd} = 1 - (C_1^{slpd} + C_2^{slpd})/1000$. From assumption (12), we define a parameter $D = |\beta|_1^2/\Delta_p^2$. Figure 1 presents the plot of the L1-loss$= |R^{slpd} - R|$, the difference between the SLPD risk and the Bayes risk, versus the rescaled sample size $\sqrt{n}/\log p$. It can be seen that in each model, R^{slpd} converges very fast to R. The D value to some degree describes the difficulty in classifying a new observation using the SLPD rule. For example, Model 6 has a relatively small D value and it can be seen that the L1-loss is already very small even when the sample size is $n_1 = n_2 = 64$, which is

much smaller than the dimension $p = 256$. On the other hand, under Model 5, we have $D = 32.48$. This implies that condition (12) is seriously violated. Although the L1-loss is quite large when $\sqrt{n}/\log p = 2.04$, a fast convergence pattern can be observed when the rescaled sample size is increased.

5. Discussion

To overcome the Gaussian limitation in high dimensional LDA, we propose a semiparametric linear programming discriminant rule under a semiparametric model. We have demonstrated via theoretical results and numerical study that comparing with normality-based discriminant rules our proposed SLPD rule can significantly improve the classification accuracy under the semiparametric model. In this paper we focus on binary classification and linear discriminant analysis only. It will be interesting to extend it to multi-class and quadratic discriminant analysis problems. In addition, as pointed out by a reviewer, it is also interesting to extent the results to functional data where variables are observations on discretized grids of some continuous data (Aneiros & Vieu, 2014, 2015).

References

Aneiros, G. & Vieu, P. (2014). Variable selection in infinite-dimensional problems. *Stat. Probab. Lett.*, **94**, 12-20.

Aneiros, G. & Vieu, P. (2015). Partial linear modeling with multi-functional covariates. *Comput. Stat.*, **30**, 647-671.

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

Cai, T. & Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Assoc.* **106**, 1566-77.

Candes, E. & Tao, T. (2007), The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*

Fan, J., Feng, Y. & Tong, X. (2012). A ROAD to classification in high dimensional space. *J. R. Stat. Soc. Ser.* B **74**, 745-71.

Han, F., Zhao, T. & Liu, H. (2013) CODA: High dimensional copula discriminant analysis. *J. Mach. Learn. Res.* **14**, 629-71.

Kon, A. M. and Nikolaev, N. (2011). Empirical normalization for quadratic discriminant analysis and classifying cancer subtypes. *Machine Learning and Applications and Workshops (ICMLA)*, 10th International Conference on. IEEE, 2011, **2**, 374-79.

Lin, Y. & Jeon, Y. (2003). Discriminant analysis through a semiparametric model. *Biometrika.* **90**, 379-92.

Liu, H., Han, F., Yuan, M., Lafferty, J. & Wasserman, L. (2012). High dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293-326.

Liu, H., Lafferty, J. & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**, 2295-328.

Mai, Q. & Zou, H. (2013). Semiparametric sparse discriminant analysis in ultra-high dimensions. arXiv preprint arXiv:1304.4983.

Mai, Q., Zou, H. & Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultrahigh dimensions. *Biometrika.* **99**, 29-42.

Shao, J., Wang, Y., Deng, X. & Wang, S. (2011). Sparse linear discriminant analysis with high dimensional data. *Ann. Statist.* **39**, 1241-65.

Serfling, R.(1992). Nonparametric confidence intervals for generalized quantile parameters in multi-sample contexts. *Nonparametric Statistics and Related Topics* , pp. 121-39, Elsevier Science Publisher B.V.

Wang, C., Cao, L. & Miao, B. (2013). Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data. *Comput. Stat. Data. An.* **66**, 140-49.

Xue, L. & Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40**, 2541-71.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541-67.