Hong Hyokyoung ORCID iD: 0000-0002-4280-6243
Li Yi ORCID iD: 0000-0003-1720-2760

Multiclass Linear Discriminant Analysis with Ultrahigh-Dimensional Features

Yanming Li[1], Hyokyoung G. Hong[2] and Yi Li[1]

[1] Department of Biostatistics, University of Michigan,
Ann Arbor, MI 48109, U.S.A.
[2] Department of Statistics and Probability, Michigan State University,
East Lansing, MI 48824, U.S.A.

April 5, 2019

SUMMARY: Within the framework of Fisher's discriminant analysis, we propose a multiclass classification method which embeds variable screening for ultrahigh-dimensional predictors. Leveraging inter-feature correlations, we show that the proposed linear classifier recovers informative features with probability tending to one and can asymptotically achieve a zero misclassification rate. We evaluate the finite sample performance of the method via extensive simulations and use this method to classify post-transplantation rejection types based on patients' gene expressions.

KEY WORDS: Fisher's multiclass discriminant analysis; jointly informative features; marginally informative features; multivariate screening; ultrahigh-dimensional classification.

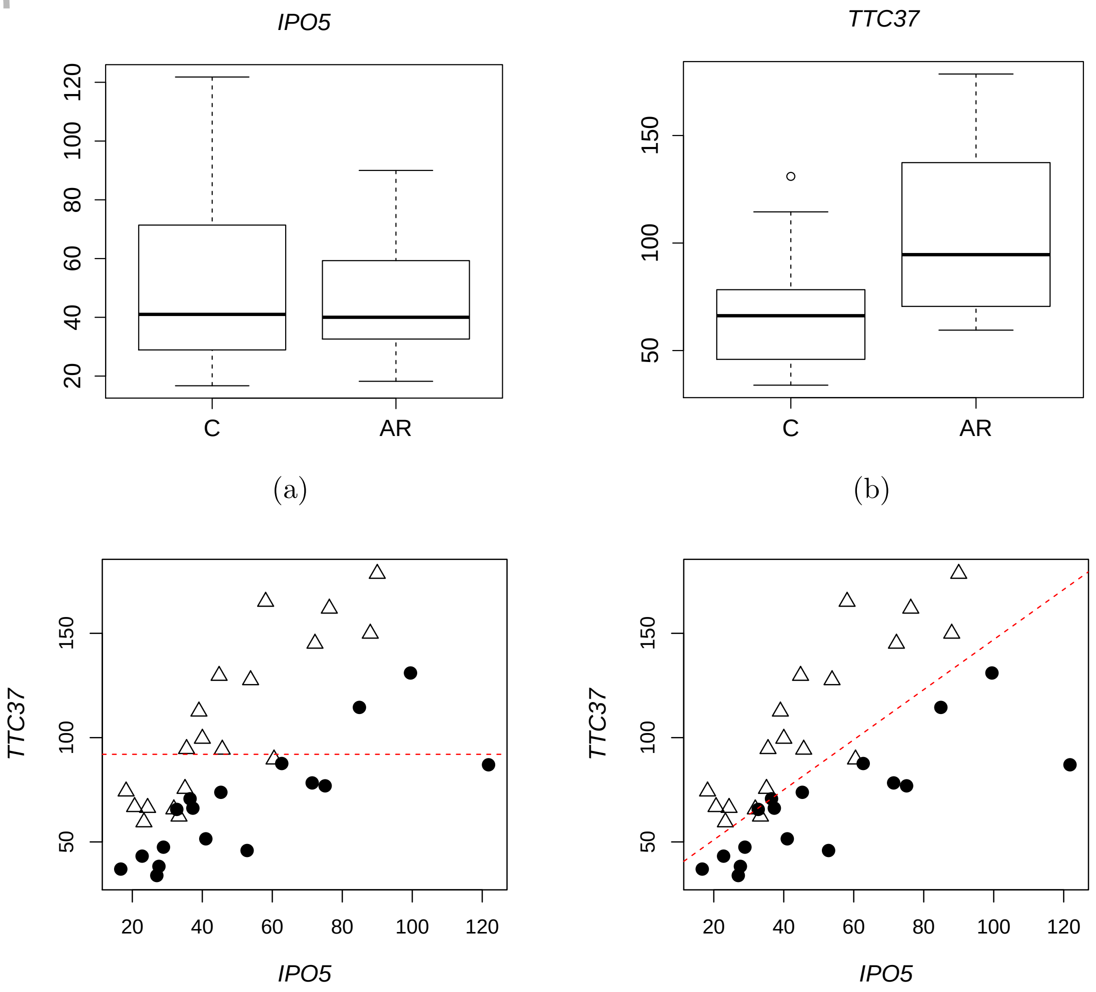
1 Introduction

Ultrahigh-dimensional data, wherein the number of features p is in the exponential order of the sample size n, have now been routinely collected. For example, in the motivating kidney transplant study (Flencher et al., 2004), 62 post-transplant kidney tissue samples have been assayed on 12,625 genes. Distinguishing four types of tissues, namely, those from normal donors (C), well-functioning kidneys (TX), kidneys with acute rejection (AR), and kidneys with acute dysfunction but no rejection (NR), based on their molecular biomarkers is important in balancing the need for immunosuppression to prevent rejection and in minimizing drug-induced toxicities.

Linear discriminant analysis (LDA) is a widely used classification method with ready implementability and close relationships with many modern machine learning techniques (Dorfer et al., 2016; Gorban et al., 2018; Cai et al., 2018). In high-dimensional settings, LDA using all features leads to poor results (Fan and Fan, 2008), and high-dimensional LDA is often preceded by variable selection procedures. Many variable selection methods are based on regularization approaches (Guo, 2010; Witten and Tibshirani, 2011; Xu et al., 2014; Fan et al., 2012; Mai et al., 2012; Cai and Liu, 2011; Gaynanova et al., 2016; Safo and Ahn, 2016), which require iterative estimation of high-dimensional parameters, including computation of a $p \times p$ precision matrix (Xu et al., 2014). It is unclear whether these regularization methods can be directly applied to ultrahigh-dimensional classification. Furthermore, the conditions that guarantee selection consistency may fail to hold for ultrahigh-dimensional cases.

Computationally more efficient screening methods (Fan and Fan, 2008; Fan and Lv, 2008; Pan et al., 2016; Yu et al., 2016) and Bayesian methods (Johnson, 2013; Johnson and Rossell, 2012; Nikooienejad et al., 2016; Rossell and Rubio, 2018) have also been developed for (ultra)high-dimensional variable selection. However, most of the screening methods in literature require the informative features to have strong marginal discriminant effects and ignore the inter-feature correlations, therefore are not designed for weak signal selection. While in ultrahigh-dimensional settings, many marginally weak signals have strong predictive effects on the outcome classes. As shown in Figure 1, gene *IPO5* does not have a sufficient power to distinguish tissues with C and AR rejection types. However, jointly with gene *TTC37*, a marginally informative (MI) feature, the classification accuracy can be much improved. In this case, we call gene *IPO5* a marginally weak but jointly informative (JI) feature.

Furthermore, many of the high-dimensional classification methods aforementioned are designed for binary classifications. Multiclass classification is more challenging than binary cases (Hastie et al., 2009; Gaynanova et al., 2016). Most multiclass classification methods rely on sequential binary classifications by way of one-versus-the rest (Bishop, 2006), direct pairwise comparison (Bishop, 2006), direct graph traversal (Platt et al., 2003), error-correcting output coding (Allwein et al., 2000), multiclass objective functions (Weston and Watkins, 1998), sequential approaches (Cai and Liu, 2011; Mai et al., 2012; Witten and Tibshirani, 2011), or simultaneous canonical vector estimation (Mai et al., 2017; Gaynanova et al., 2016). However, the choice of reduction method from multiclass to binary is on a case-by-case basis and is not a trivial task (Allwein et al., 2000). In particular, commonly used pairwise comparisons are involved with a large number of individual classifiers, which is likely to incur misclassification error and numerical instability with small sample sizes (Wu et al., 2004). On the other hand, LDA can perform multiclass classification without resorting to pairwise comparisons.



(a)

(b)

(c)

(d)

Figure 1: The roles of marginally informative (MI) and jointly informative (JI) features on classification. (a) the marginal scores of gene *IPO5* between C and AR in the kidney transplant data are similar and thus *IPO5* is not MI; (b) the marginal scores of a MI gene *TTC37*; (c) classification of C (circles) and AR (triangles) based on gene *TTC37*; (d) classification based on both *TTC37* and *IPO5* gives the better performance than *TTC37* only. This figure appears in color in the electronic version of this article.

The covariance-enhanced discriminant analysis method proposed by Xu et al. (2014) requires estimating the $p \times p$-dimensional precision matrix of the covariates and is not computationally feasible in ultrahigh-dimensional settings. The pairwise sure independent screening for multiclass LDA (pairwiseLDA) proposed by Pan et al. (2016) uses the independence rule and ignores the inter-feature correlations. It cannot detect marginally weak signals. Furthermore, the marginal sliced inverse regression (SIR) for model-free feature selection and multiclass classification proposed by Yu et al. (2016) selects the linear combinations of features, and cannot select individual features. It therefore lacks interpretation for the selected features. The general sparse multi-class LDA proposed by Safo and Ahn (2016) projects the original feature space to a low-dimensional canonical subspace, and therefore cannot select individual features either. The work of Cai et al. (2018) is designed for analyzing dependent data with a large number of samples, but not particularly for high-dimensional feature selection.

We propose an ultrahigh-dimensional multiclass classification method within the framework of Fisher's LDA. Our proposal, termed multiclass LDA (mLDA), embeds a computationally feasible screening procedure, specially designed for detection of weak signals by accounting for inter-feature correlations. We show that the proposed method can recover all the informative features, including both MI and JI features, with probability tending to one and can achieve an asymptotically negligible misclassification rate.

The rest of the paper is organized as follows. Section 2 introduces mLDA and Section 3 develops its theoretical properties. In Section 4, the performance of the proposed method is evaluated using simulation studies. We apply the proposed procedure to analyze the renal transplantation data in Section 5 and conclude the paper with a discussion in Section 6. Technical details are provided in the online supplemental materials.

2 Ultrahigh Dimensional Multiclass Classification

2.1 Notation

Denote by $\mathbf{A}'$ the transpose of a $p \times p$ matrix $\mathbf{A}$ and by A_{jk} the (j,k)th entry of $\mathbf{A}$, where $1 \le j, k \le p$. Let $|\mathcal{S}|$ be the cardinality of a set $\mathcal{S}$ and $\mathcal{S}^c$ be the complement of $\mathcal{S}$. We denote the trace of $\mathbf{A}$ as $tr(\mathbf{A})$, the minimum and maximum eigenvalues of $\mathbf{A}$ as $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, and the operator norm and the Frobenius norm as $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$ and $\|\mathbf{A}\|_F = tr(\mathbf{A}'\mathbf{A})^{1/2}$, respectively. Let $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be a p-dimensional vector of features. We refer to X_j as feature j for short, $1 \le j \le p$. Denote by $\boldsymbol{\Sigma}$ the covariance matrix of $\boldsymbol{X}$ and by $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ the precision matrix. Let $\mathcal{G}(\mathcal{V}, \mathcal{E}; \boldsymbol{\Omega})$ be the graph induced by $\boldsymbol{\Omega}$, where $\mathcal{V} = \{1, \ldots, p\}$ is the vertex set and $\mathcal{E}$ is the edge set. An edge refers to a pair of two vertices, j and j', which satisfies $\Omega_{jj'} \neq 0$. For a subset $\mathcal{V}_l \subset \mathcal{V}$, denote by $\boldsymbol{\Omega}_l$ the principal submatrix of $\boldsymbol{\Omega}$ with its row and column indices restricted to $\mathcal{V}_l$. Denote by $\mathcal{E}_l$ the corresponding edge set. A subgraph $\mathcal{G}(\mathcal{V}_l, \mathcal{E}_l, \boldsymbol{\Omega}_l)$ is a connected component in $\boldsymbol{\Omega}$ if any two vertices in $\mathcal{V}_l$ are connected, and for

$j \in \mathcal{V}_l^c$, then $\Omega_{jj'} = 0$ for any $j' \in \mathcal{V}_l$. We write $\mathcal{G}(\boldsymbol{\Omega}) = \mathcal{G}(\mathcal{V}, \mathcal{E}; \boldsymbol{\Omega})$ for short when there is no confusion.

2.2 Marginally Informative Features and Jointly Informative Features

Consider a K-class classification problem, where $K \geq 2$. Denote by Y the class membership and assume that the covariate vector $\boldsymbol{X}$ satisfies

$$\mathbf{X} | \{Y = k\} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k = 1, \dots, K,$$

where $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kp})'$ is a p-dimensional mean vector of class k and $\boldsymbol{\Sigma}$ is the common covariance matrix for all K classes. Let $(Y_1, \boldsymbol{X}_1), \dots, (Y_n, \boldsymbol{X}_n)$ be n independent observations of $(Y, \boldsymbol{X})$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})'$, $i = 1, \dots, n$. Denote by n_k the number of observations in class k such that $\sum_{k=1}^K n_k = n$. For a pair of classes k and k', Fisher's rule, which can also be considered as a Bayes rule with equal prior probabilities, assigns an observation to class k over class k' if $(\mathbf{X} - \boldsymbol{\mu}_k/2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k > (\mathbf{X} - \boldsymbol{\mu}_{k'}/2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{k'}$. This naturally leads to the following classification rule:

$$\widehat{Y} = \arg\max_{1 \leq k \leq K} \{(\mathbf{X} - \boldsymbol{\mu}_k/2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\}. \tag{1}$$

When $p < n$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ can be estimated by $\widehat{\boldsymbol{\mu}}_k = \sum_{i:Y_i=k} \mathbf{X}_i/n_k$ and $\widehat{\boldsymbol{\Sigma}} = \sum_{k=1}^K \sum_{i:Y_i=k} (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)'/(n-K)$. However, when $p > n$, (1) is ill-posed as $\widehat{\boldsymbol{\Sigma}}$ is singular. Hence, a variable selection procedure is usually required to precede classification under some sparsity assumption. It can be shown that a sufficient and necessary condition for feature j, where $1 \leq j \leq p$, to be informative is

$$\mathcal{S}_0 = \left\{1 \leq j \leq p : \sum\nolimits_{j'=1}^p \Omega_{jj'}(\mu_{kj'} - \mu_{k'j'}) \neq 0, \text{ for some } 1 \leq k < k' \leq K\right\}.$$

Under the faithfulness condition that any MI features must belong to $\mathcal{S}_0$, the informative features consist of the following two mutually exclusive sets:

$$\mathcal{S}_1 = \left\{1 \leq m \leq p : \mu_{km} - \mu_{k'm} \neq 0 \text{ for some } 1 \leq k < k' \leq K\right\}$$

and

$$\mathcal{S}_2 = \left\{j \in \mathcal{S}_1^c : \sum\nolimits_{m \in \mathcal{S}_1} \Omega_{jm}(\mu_{km} - \mu_{k'm}) \neq 0 \text{ for some } 1 \leq k < k' \leq K\right\}, \tag{2}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ contain the MI and JI features, respectively.

Though identifying marginally weak features is challenging in general, the JI features can be found by searching the connected components in $\boldsymbol{\Omega}$ which contain at least one feature in $\mathcal{S}_1$. Furthermore, Theorem 1 in Section 3 shows that the connected components in $\boldsymbol{\Omega}$ can be accurately recovered by thresholding the corresponding sample covariance matrix.

2.3 Algorithm of Multiclass Linear Discriminant Analysis (mLDA)

Given the training dataset $\{Y_i, \mathbf{X}_i\}_{i=1}^n$, denote by $\overline{X}_{\cdot j}^{(k)} = n_k^{-1}\sum_{i:Y_i=k} X_{ij}$ the sample mean of feature j within class $k \in \{1, \ldots, K\}$. Denote by $\widetilde{\mathbf{\Sigma}}$ the thresholded sample covariance matrix. That is, $\widetilde{\Sigma}_{jj'} = \widehat{\Sigma}_{jj'} 1(|\widehat{\Sigma}_{jj'}| \geq \alpha)$, $1 \leq j, j' \leq p$, where $\widehat{\Sigma}_{jj'}$ is the (j, j')th entry of $\widehat{\mathbf{\Sigma}}$, $1(\cdot)$ is the indicator function, and α is a threshold. For a pair of classes (k, k'), $1 \leq k < k' \leq K$, select the set $\widehat{\mathcal{S}}_1(k, k')$ containing indices m which satisfy

$$\left|\overline{X}_{\cdot m}^{(k)} - \overline{X}_{\cdot m}^{(k')}\right| > \tau,$$

where $\tau > 0$ is a thresholding parameter controlling the size of $\widehat{\mathcal{S}}_1(k, k')$. Denote by $\widehat{\mathcal{S}}_1 = \bigcup_{1 \leq k < k' \leq K} \widehat{\mathcal{S}}_1(k, k')$ the set containing all of the MI features.

With $\widehat{\mathcal{S}}_1$, we use the recursive labeling algorithm (Shapiro and Stockman, 2002) to identify the connected components in $\mathcal{G}(\widetilde{\mathbf{\Sigma}})$, the graph introduced by $\widetilde{\mathbf{\Sigma}}$, that contain any features in $\widehat{\mathcal{S}}_1$. Suppose there are $B \leq |\widehat{\mathcal{S}}_1|$ such connected components, say, $\widehat{\mathcal{C}}_l$, $l = 1, \ldots, B$, each containing at least one MI feature. Let $\mathcal{U} = \bigcup_{l=1}^B \widehat{\mathcal{C}}_l$ with $u = |\mathcal{U}|$. Notice that $\widehat{\mathcal{S}}_1 \subseteq \mathcal{U}$.

Let $\widetilde{\mathbf{\Sigma}}_l$ be the principal submatrix of $\widetilde{\mathbf{\Sigma}}$ with the row and column indices restricted to $\widehat{\mathcal{C}}_l$, and compute $\widehat{\mathbf{\Omega}}_l = (\widetilde{\mathbf{\Sigma}}_l)^{-1}$. Let $\widehat{\mathbf{\Omega}}^u = \operatorname{diag}\big(\widehat{\mathbf{\Omega}}_1, \ldots, \widehat{\mathbf{\Omega}}_B\big)$ be a block diagonal matrix of dimension $u \times u$. Under the sparsity assumption, $\widehat{\mathcal{C}}_l$ are of small sizes and u is much smaller than p. To detect the JI features, we only need to consider $\mathcal{U}$ as the candidate set for them. Thus, $\mathcal{S}_2$ can be estimated by:

$$\widehat{\mathcal{S}}_2 = \left\{ j \in \mathcal{U} \cap \widehat{\mathcal{S}}_1^c : \big|\sum\nolimits_{j' \in \widehat{\mathcal{S}}_1} \widehat{\Omega}_{jj'}^u (\overline{X}_{\cdot j'}^{(k)} - \overline{X}_{\cdot j'}^{(k')})\big| \geq \nu_n \text{ for some } 1 \leq k < k' \leq K \right\}, \tag{3}$$

where $\nu_n > 0$ is a thresholding parameter controlling the size of selected JI features.

We denote by $\widehat{\mathcal{S}}_0 = \widehat{\mathcal{S}}_1 \cup \widehat{\mathcal{S}}_2$ the set containing all the informative features. Finally, for a new observation with covariate vector $\mathbf{X}_{\text{new}}$, we determine the class membership by

$$\underset{1 \leq k \leq K}{\arg\max}\, d_k,$$

where d_k is the Fisher discriminant statistics for class k, defined by

$$d_k = (\mathbf{X}_{\text{new}}^s - \widehat{\boldsymbol{\mu}}_k^s/2)' \widehat{\mathbf{\Omega}}^s \widehat{\boldsymbol{\mu}}_k^s. \tag{4}$$

Here $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\mathbf{\Omega}}^u$ are estimated from the training data and $\mathbf{X}_{\text{new}}^s$, $\widehat{\boldsymbol{\mu}}_k^s$ and $\widehat{\mathbf{\Omega}}^s$, respectively, are subvectors or submatrices of $\mathbf{X}_{\text{new}}$, $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\mathbf{\Omega}}^u$ with the elements indexed by $\widehat{\mathcal{S}}_0$.

For ease of understanding, Figure 2 depicts the flowchart of the mLDA algorithm. The proposed mLDA algorithm utilizes the dependence between the MI and JI features, and when the informative features are sparse, the proposed screening procedure to identify $\widehat{\mathcal{S}}_0$ is computationally feasible, as we will demonstrate in Section 4. Moreover, as shown in Section 3, setting the tuning parameters as $\tau = O((r \log p)^s)$, $\alpha = O(n^{(\xi-1)/2})$ and $\nu_n = O\left((r(\log p)\exp(n^{\xi}))^{s'}\right)$, for some $0 < r < 1$, $0 < s < 1/2$, $0 < \xi < 1$ and $0 < s' \leq 1/2$ guarantees that mLDA has selection consistency and a zero asymptotic misclassification rate.

3 Theoretical properties

Under regularity conditions (A1) - (A11) listed in the Appendix, mLDA possesses theoretical properties, such as the sure screening property and asymptotic vanishing post-screening misclassification rate.

For feature j, denote by $\mathcal{C}_{[j]}$ and $\widehat{\mathcal{C}}_{[j]}$ the vertex sets of the connected component containing j in the graphs induced by $\mathbf{\Omega}$ and $\widetilde{\mathbf{\Sigma}}$, respectively.

Theorem 1 *For any feature j, $1 \leq j \leq p$, suppose that $\mathcal{C}_{[j]} = O(\exp(n^{\xi}))$ for the ξ given in Condition (A3), then, together with Conditions (A5) and (A7), we have*

$$P\left(\mathcal{C}_{[j]} = \widehat{\mathcal{C}}_{[j]}\right) \to 1 \text{ as } n \to \infty.$$

Therefore, the principal submatrices of $\mathbf{\Omega}$ corresponding to the relevant connected components can be estimated by inverting the corresponding submatrices of $\widetilde{\mathbf{\Sigma}}$. For a properly chosen thresholding parameter α, Bickel and Levina (2008) and Fan et al. (2011) showed that the estimated precision matrix using $\widetilde{\mathbf{\Sigma}}$ is consistent.

Theorem 2 (Sure screening property)*: Under conditions (A1)-(A9) and (A11),*

$$P(\mathcal{S}_0 \subseteq \widehat{\mathcal{S}}_0) \to 1 \text{ as } n \to \infty.$$

Theorem 3 (False positive control property)*: Under conditions (A1)-(A8) and (A10)-(A11), for any $\zeta = o(n \log p)$, we have*

$$P\left(|\widehat{\mathcal{S}}_0 \cap \mathcal{S}_0^c| \leq \zeta^{-1}|\mathcal{S}_0^c|\right) \to 1 \text{ as } n \to \infty.$$

Remark. Theorems 2 and 3 hold for any distributions satisfying (A1) given in the Appendix. Condition (A1) characterizes a rich family of distributions, including distributions with polynomial tails such as the t distribution.

Given the training samples $\mathcal{D} = \{Y_i, \mathbf{X}_i\}_{i=1}^n$, we can assess the conditional misclassification rate for a class k of mLDA by

$$R_{\text{mLDA}}(k; \mathcal{D}) = P\Big(\underset{1 \leq l \leq K}{\arg\max}\{(\mathbf{X}_{\text{new}}^{\text{s}} - \widehat{\boldsymbol{\mu}}_l^{\text{s}}/2)'\widehat{\mathbf{\Omega}}^{\text{s}}\widehat{\boldsymbol{\mu}}_l^{\text{s}}\} \neq k | Y_{\text{new}} = k; \mathcal{D}\Big),$$

where $\widehat{\boldsymbol{\mu}}_l^{\text{s}}$, and $\widehat{\mathbf{\Omega}}^{\text{s}}$ are estimated from $\mathcal{D}$. As pointed out in Shao et al. (2011), by the dominated convergence theorem, it suffices to focus on the conditional misclassification rate, instead of the

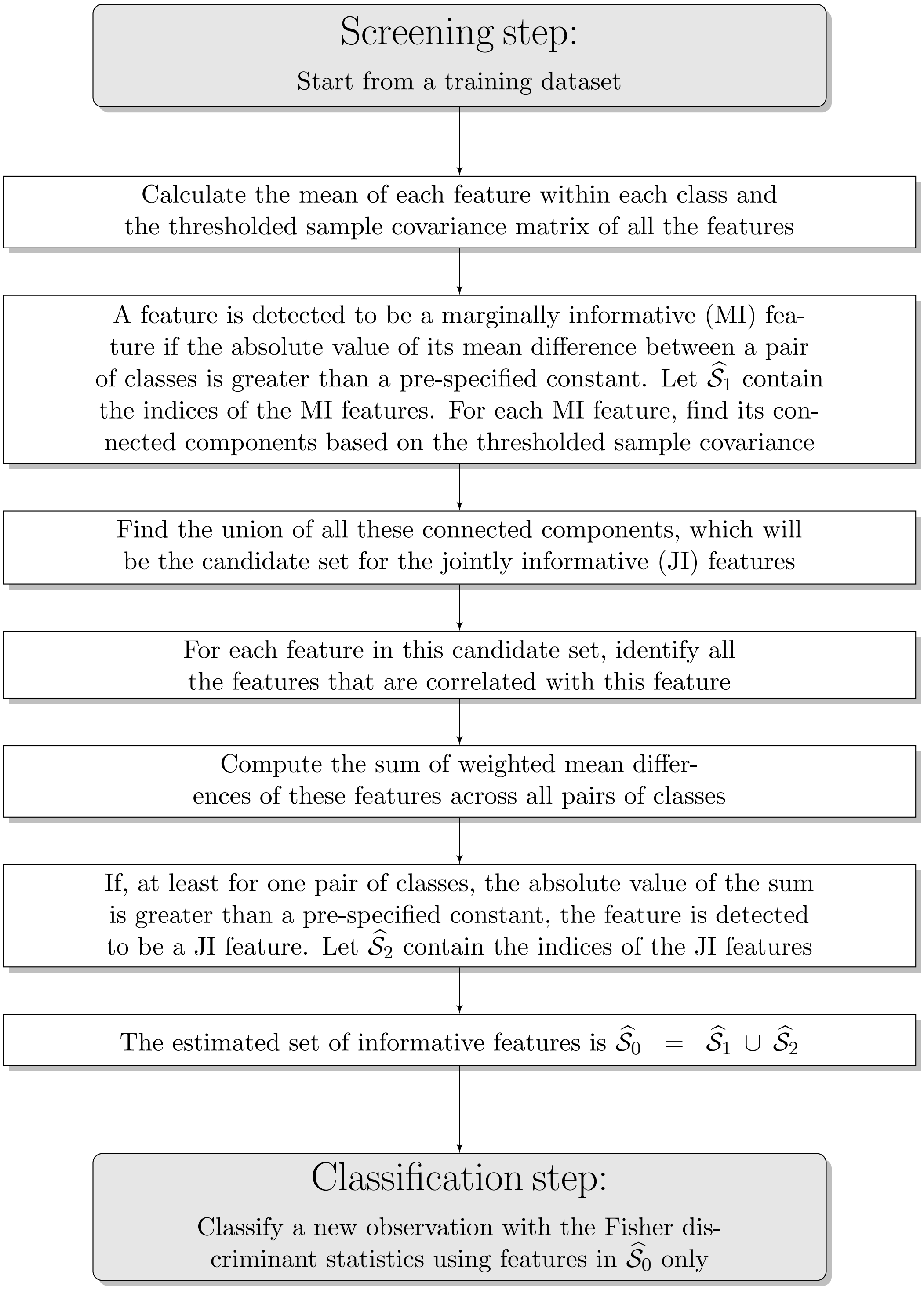

Screening step:
Start from a training dataset
Calculate the mean of each feature within each class and the thresholded sample covariance matrix of all the features
A feature is detected to be a marginally informative (MI) feature if the absolute value of its mean difference between a pair of classes is greater than a pre-specified constant. Let $\widehat{\mathcal{S}}_1$ contain the indices of the MI features. For each MI feature, find its connected components based on the thresholded sample covariance
Find the union of all these connected components, which will be the candidate set for the jointly informative (JI) features
For each feature in this candidate set, identify all the features that are correlated with this feature
Compute the sum of weighted mean differences of these features across all pairs of classes
If, at least for one pair of classes, the absolute value of the sum is greater than a pre-specified constant, the feature is detected to be a JI feature. Let $\widehat{\mathcal{S}}_2$ contain the indices of the JI features
The estimated set of informative features is $\widehat{\mathcal{S}}_0 = \widehat{\mathcal{S}}_1 \cup \widehat{\mathcal{S}}_2$
Classification step:
Classify a new observation with the Fisher discriminant statistics using features in $\widehat{\mathcal{S}}_0$ only

Figure 2: Flowchart of the mLDA procedure.

unconditional misclassification rate. We define the overall conditional misclassification rate of mLDA as

$$R_{\mathrm{mLDA}}(\mathcal{D}) \quad = \quad K^{-1}\sum\nolimits_{k=1}^{K} R_{\mathrm{mLDA}}(k;\mathcal{D}).$$

Theorem 4 (Asymptotic vanishing post − screening misclassification rate)*: For any pair of classes* $1 \leq k < k' \leq K$*, let* $\Delta_p^2(k,k') = \min_{1\leq k<k'\leq K}\{(\boldsymbol{\mu}_k^0 - \boldsymbol{\mu}_{k'}^0)'\boldsymbol{\Omega}^0(\boldsymbol{\mu}_k^0 - \boldsymbol{\mu}_{k'}^0)\}$*, where the superscript "0" denotes subvectors or submatrices with indices restricted to* $\mathcal{S}_0$*. Let* $\Delta_p^2 = \min_{1\leq k<k'\leq K}\Delta_p^2(k,k')$*. Under conditions (A2)-(A11), when classifying* $\mathbf{X}_{new}$ *based on features selected from the screening step, for sufficiently large n, we have*

$$R_{mLDA}(\mathcal{D}) \leq K\Phi\left(-(1+O_P(a_n))^{1/2}(1+O_P(\rho_n))^{1/2}\Delta_p/2\right) + o_P(1),$$

where Φ *is the standard normal distribution function,* ρ_n *is given in (A11) and*

$$a_n \equiv \min_{1\leq k<k'\leq K}\max\left\{\frac{|\mathcal{S}_0|^{1/2}}{n^{1/2}\Delta_p(k,k')}, \frac{|\mathcal{S}_0|}{n\Delta_p^2(k,k')}, \frac{1}{\Delta_p^2(k,k')}\right\}.$$

Furthermore, if $\Delta_p^2\min\{n/|\mathcal{S}_0|,1\} \to \infty$*, then* $R_{mLDA}(\mathcal{D}) \to 0$*.*

Notice that the condition $\Delta_p^2\min\{n/|\mathcal{S}_0|,1\} \to \infty$ is weaker than $n\Delta_p^2/p \to \infty$, which is required for achieving an asymptotic zero misclassification rate under the independent rule (Fan and Fan, 2008).

4 Simulation studies

We compared the finite sample performance of mLDA with that of other ultrahigh-dimensional classification methods, including the penalizedLDA (Witten and Tibshirani, 2011), the regularized risk minimization package (bmrm) (Teo et al., 2010), the multi-group sparse discriminant analysis (MGSDA) (Gaynanova et al., 2016), pairwiseLDA (Pan et al., 2016), SIR (Yu et al., 2016), the feature annealed independence rule (MS) (Fan and Fan, 2008) and the sure independence screening (SIS) (Fan and Lv, 2008). As an oracle benchmark, we also applied Fisher's rule with informative features known *a priori.* We first investigated the cases where the variance-covariance matrices of the features were equal across different classes, and, hence, the classes were linearly separable. We specifically considered the following two models.

Model I (multivariate normal distribution): Set $K = 3$ with class sizes $n_1 = n_2 = n_3 = 100$ and p =10,000. Variables 1–30 were generated from a multivariate normal distribution with the means specified as in Table 1. These 30 features were divided into six independent blocks: X_1–X_5, X_6–X_{10}, X_{11}–X_{15}, X_{16}–X_{20}, X_{21}–X_{25} and X_{26}–X_{30}. Features within the block were governed by the same covariance structures such as compound symmetry (CS),

Table 1: Means of the informative features

Features	Class 1	Class 2	Class 3
X_1–X_4, X_{11}–X_{14}, X_{21}–X_{24}	0	0	0
X_5, X_{15}	0	2.5	0
X_6–X_{10}, X_{16}–X_{20}	1.5	-1.5	-1.5
X_{25}	0	0	2.5
X_{26}–X_{30}	-1.5	-1.5	1.5

first order autocorrelation (AR1), banded, star and "unstructured" (Un). The explicit form of the last three covariance structures was given in (5). When the covariance structure required a correlation coefficient parameter, we used $\rho = 0.7$. In this case, variables 5–10, 15–20, 25–30 were MI features, whereas X_1–X_4, X_{11}–X_{14} were considered JI features for class pair (1,2), X_{21}–X_{24} were considered JI features for class pair (1,3) and X_1–X_4, X_{11}–X_{14}, X_{21}–X_{24} were considered JI features for class pair (2,3). The remaining non-informative 9,970 features were independently generated from $N(0,1)$ and were independent of the first 30 variables.

Model II (multivariate t distribution): It was the same as Model I except that variables 1–30 were generated from the multivariate t distribution with four degrees of freedom and the remaining non-informative 9,970 features were independently generated from the univariate t_4 distribution and were independent of the first 30 features.

$$\text{Banded:} \begin{pmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & \rho & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{pmatrix}, \quad \text{Star:} \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & 0 & 0 & 0 \\ \rho & 0 & 1 & 0 & 0 \\ \rho & 0 & 0 & 1 & 0 \\ \rho & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \text{Un:} \begin{pmatrix} 1 & \rho & 0 & 0 & \rho \\ \rho & 1 & \rho & \rho & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & \rho & \rho & 1 & 0 \\ \rho & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (5)$$

Even though condition (A7) provides the orders of the tuning parameters so that mLDA will render the desired theoretical properties, these orders do not provide specific ranges of the tuning parameters with given p and n in practice. In our numerical studies, we used 5-fold cross-validation to choose the optimal tuning parameters, α, τ and ν_n.

To assess the performance in feature selection, we used false positives (FP), false negatives (FN), and the minimum number of features needed to include all informative features (MMS). To assess the classification performance, we used the number of misclassified cases (ER). The simulation results were reported in Table 2. It appears that mLDA had the lowest FP, FN, ER and MMS under various covariance structures. When features did not follow multivariate Gaussian distributions, ERs tended to be larger across all of the methods, compared to the multivariate Gaussian. However, mLDA consistently outperformed the other methods. More simulation results with different correlations were reported in Table S6 in the Web Appendices. We next investigated the performance of mLDA when the variance-covariance matrices differed across classes so that the classes were not linearly separable. We compared the classification performance of mLDA with various nonlinear classification methods, including the mixed discriminant analysis (Hastie and Tibshirani, 1995), the quadratic discriminant analysis (Ripley, 1996), the regularized discriminant analysis (Hastie et al., 1995), the shrunken-centroids regularized discriminant analysis (Guo et al., 2005), neural network (Ripley, 1996), kernel support vector machine (Hsu and Lin, 2002), k-nearest neighbors (Torgo, 2010) and Naive Bayes (Ng and Jordan, 2001).

Model III (heterogeneous covariance): We set $K = 3$, $n_1 = n_2 = n_3 = 100$ and $p =$10,000. The variables in each class were simulated from multivariate normal distributions with the same mean structure as in Model I and a class-specific covariance matrix. To increase the heterogeneity and the level of nonlinearity of the class boundaries, we let different classes have different correlation coefficients in the covariance matrices. The "unstructured" covariance matrix in (5) was used for each of the six blocks within the first 30 features. Class-specific correlation coefficients were set to be ρ_1, ρ_2 and ρ_3 for the first, second and third class, respectively. The remaining non-informative 9,970 features were independently generated from $N(0, 1)$ and were independent of the first 30 features.

The results reported in Tables S4 and S5 in the Web Appendices showed that, in most cases considered, mLDA still outperformed the other linear classification methods, in terms of selecting JI features and classification accuracy. The mLDA procedure also outperformed the nonlinear classification methods in classification accuracy when the classes were nearly linear separable. As expected, the classification performance of mLDA deteriorated as the classes became more linearly inseparable.

5 Classification of Post-transplant Rejection Types

We applied the proposed mLDA to classify post-kidney transplant rejection types based on patients' gene expressions. The kidney transplant study (Flencher et al., 2004) had a total of 62 kidney tissue samples taken from 17 normal donor kidneys (C), 19 well-functioning kidneys more than 1-year post-transplant (TX), 13 biopsy-confirmed acute rejection (AR), and 13 acute dysfunction with no rejection (NR). Each sample was microarrayed (by HG-U95Av2 GeneChips, Affymetix) with 12,625 genes from kidney biopsies and peripheral blood lymphocytes at transplant.

For comparisons, we also considered regularization methods including the regularized optimal affine discriminant (ROAD) by Fan et al. (2012), the linear programming discriminant (LPD) by Cai and Liu (2011), the covariance-enhanced discriminant analysis (CED) by Xu et al. (2014), and screening methods including MS (Fan and Fan, 2008), SIS and the iterative-SIS

Table 2: Comparisons with the competing methods

		Model I					Model II				
		CS	AR1	Band	Star	Un	CS	AR1	Band	Star	Un
FP	mLDA	9.4	9.5	16.2	16.3	17.4	7.5	8.5	15.8	19.6	28.0
		(0.9)	(0.9)	(1.0)	(2.1)	(2.3)	(0.9)	(1.0)	(2.1)	(3.5)	(4.1)
	MS	19.8	24.0	59.4	41.6	59.4	18.2	26.4	59.1	39.3	40.7
		(1.2)	(1.2)	(2.4)	(1.3)	(2.1)	(1.2)	(1.1)	(2.3)	(2.1)	(2.0)
	pairwiseLDA	16.7	18.9	24.1	20.5	25.7	17.1	20.6	25.5	24.1	27.3
		(2.3)	(2.6)	(3.0)	(3.0)	(3.1)	(2.6)	(2.7)	(3.1)	(2.9)	(3.2)
	SIR	34.2	35.6	37.1	38.3	37.8	33.9	36.2	35.1	37.7	39.3
		(5.7)	(6.8)	(5.9)	(6.0)	(7.2)	(6.3)	(7.7)	(6.4)	(6.6)	(7.8)
	bmrm	24.0	26.6	23.9	25.4	27.1	24.3	25.7	25.9	26.1	28.8
		(4.6)	(5.1)	(4.4)	(5.2)	(5.6)	(6.3)	(5.9)	(6.0)	(6.1)	(6.6)
	MGSDA	45.1	43.2	49.0	47.1	47.3	48.8	49.3	50.1	47.4	51.2
		(7.2)	(6.7)	(6.9)	(7.0)	(7.3)	(8.7)	(7.9)	(7.4)	(7.6)	(8.5)
	SIS	0.8	1.8	2.2	0.3	1.1	1.4	2.1	2.2	1.0	1.0
		(0.02)	(0.5)	(0.7)	(0.09)	(0.06)	(0.1)	(0.3)	(0.4)	(0.2)	(0.1)
FN	mLDA	0.3	0.6	0.5	1.0	1.1	1.7	2.3	2.5	4.7	10.3
		(0.9)	(1.2)	(1.5)	(1.4)	(1.5)	(1.7)	(1.9)	(2.1)	(2.1)	(1.1)
	MS	12.3	12.0	12.4	13.1	11.9	11.5	12.4	11.9	11.8	11.8
		(0.9)	(0.9)	(1.2)	(1.0)	(1.4)	(0.9)	(0.8)	(0.8)	(0.9)	(1.4)
	pairwiseLDA	12.2	12.7	12.1	13.0	12.9	12.3	13.1	12.4	13.3	13.4
		(0.9)	(1.1)	(1.0)	(1.0)	(1.1)	(1.0)	(1.0)	(1.0)	(1.1)	(1.1)
	SIR	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
	bmrm	12.0	12.0	12.0	12.0	12.0	12.1	12.0	12.0	12.1	12.1
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.0)	(0.0)	(0.2)	(0.2)
	MGSDA	13.0	12.7	13.4	13.5	13.7	14.1	13.6	13.9	14.2	13.8
		(2.4)	(1.8)	(2.5)	(2.2)	(2.5)	(3.2)	(3.0)	(3.3)	(3.1)	(3.2)
	SIS	20.7	20.9	20.7	20.4	21.0	20.9	21.3	21.6	20.5	21.0
		(1.5)	(1.3)	(1.3)	(1.2)	(1.3)	(1.4)	(1.3)	(1.5)	(1.5)	(1.5)
MMS	mLDA	38.7	36.2	42.1	39.8	44.0	48.9	49.1	44.5	36.3	48.8
		(2.1)	(3.4)	(3.0)	(5.8)	(3.2)	(8.5)	(9.1)	(4.5)	(5.6)	(5.9)
	MS	9856	9811	9773	9832	9804	9733	9864	9770	9699	9634
		(286)	(243)	(275)	(263)	(286)	(293)	(266)	(231)	(269)	(247)
	bmrm	9673	9526	9725	9766	9422	9567	9327	9764	9327	9334
		(223)	(284)	(257)	(279)	(247)	(265)	(251)	(270)	(284)	(255)
	SIS	9463	9634	9721	9644	9579	9842	9756	9688	9720	9591
		(232)	(243)	(257)	(234)	(261)	(270)	(225)	(291)	(256)	(248)
ER	mLDA	6.1	6.5	6.3	2.7	5.0	10.4	10.5	7.6	4.1	9.7
		(2.1)	(2.3)	(2.1)	(1.8)	(2.2)	(3.0)	(3.0)	(2.4)	(2.1)	(2.9)
	MS	10.2	10.3	9.4	3.0	7.8	11.8	11.8	11.0	5.4	10.7
		(2.6)	(2.7)	(2.5)	(2.7)	(2.6)	(3.2)	(3.0)	(3.1)	(2.1)	(3.1)
	pairwiseLDA	8.8	9.2	8.4	8.0	8.3	12.4	11.5	11.7	10.3	11.4
		(3.2)	(3.6)	(3.3)	(3.1)	(3.4)	(3.7)	(3.9)	(3.5)	(3.4)	(3.5)
	SIR	43.2	42.1	39.7	36.5	44.0	42.9	41.8	40.5	41.4	46.7
		(14.6)	(13,2)	(13.5)	(13.8)	(13.9)	(14.3)	(14.5)	(14.3)	(13.6)	(14.7)
	penalizedLDA	47.9	43.3	45.8	47.7	49.3	48.0	49.2	50.4	52.1	51.5
		(12.1)	(14.0)	(13.2)	(12.8)	(11.9)	(14.7)	(15.2)	(15.0)	(14.1)	(13.5)
	bmrm	28.1	24.6	29.3	28.4	33.3	31.6	28.7	30.9	32.4	36.8
		(6.2)	(6.7)	(6.3)	(6.5)	(6.8)	(7.1)	(6.9)	(7.0)	(7.3)	(7.1)
	MGSDA	9.3	8.7	9.9	10.0	7.5	11.3	12.0	10.4	9.7	12.5
		(3.7)	(3.6)	(3.7)	(4.0)	(4.1)	(4.3)	(4.2)	(4.5)	(4.5)	(4.7)
	SIS	17.9	17.3	7.4	5.0	17.7	19.7	19.6	9.3	5.5	20.2
		(3.5)	(2.8)	(3.4)	(3.1)	(3.9)	(3.4)	(3.8)	(3.1)	(3.5)	(4.8)
	Oracle	5.2	6.0	5.8	2.4	4.6	9.7	8.9	6.5	3.2	8.8
		(2.0)	(1.9)	(1.9)	(1.6)	(2.1)	(3.1)	(3.3)	(3.0)	(2.2)	(3.0)

(ISIS) by Fan and Lv (2008). Since ROAD, LPD, CED cannot handle ultrahigh-dimensional data, we performed variable selection using mLDA before applying the corresponding regularization method.

The classification performance was assessed by the leave-one-out procedure. The thresholding parameters τ, α and ν_n in (3) were chosen by 5-fold cross-validation. ROAD, LPD, and the `R` package `SIS`, which implements the SIS and ISIS methods, cannot handle categorical outcomes with $K > 2$. When implementing them, we first carried out pairwise comparisons between rejection types, and then used the majority vote to decide the final membership. The binary classification approaches inadvertently produced ties, which made the final class membership assignment difficult. When a tie occurred, we randomly assigned a class membership among the tied rejection types. On the other hand, mLDA, which performed multiclass classification without resorting to pairwise comparisons, did not encounter the tie issue. It turned out that the numbers of misclassified tissues given by mLDA, ROAD, LPD, CED, MS, SIS, and ISIS were 6, 9, 12, 8, 15, 16, and 13, respectively.

For each gene, we computed the frequency of its being selected during the leave-one-out procedure. The top ten genes with the highest selection frequency were given in Table 3. Among them, the JI genes *CEACAM8*, *RNASE3*, *TCN1*, *BPI* and *CRISP3* were all highly correlated with the MI gene *TCF12*, while the JI gene *IGHV3-23* was highly correlated with the MI gene *HLA-G*. Our results have biological explanations. For example, the identified gene *TCN1* encodes a member of the vitamin B12 binding protein family and vitamin B12 inefficiency can cause kidney injury (Gowder, 2014). Gene *BPI* fold-containing family A member 2/parotid secretory protein is associated with acute kidney injury (Kota et al., 2017). Gene *TCF12* is expressed in the forming collecting ducts in the developing kidney as well as in the liver (Lazzaro et al., 1992), while *HLA-G* expression in biliary epithelial cells is associated with allograft acceptance in liver-kidney transplantations (Xiao et al., 2013). Our study has also identified some novel genes, such as genes *CEACAM8* (a carcinoembryonic antigen related to cell adhesion), *RNASE3* (associated with allergic rhinitis) and *CRISP3* (strongly up-regulated in prostate carcinomas), which are all JI features and have not been reported in transplant literature.

6 Discussion

The proposed mLDA can be easily extended to accommodate non-Gaussian covariates. Indeed, the results in Section 4 hinted that mLDA works well for the heavy-tailed t distribution. Our proposed mLDA is based on the assumption of the common covariance matrix across classes. We have examined the performance of our proposal when the covariance matrices vary across classes. Our empirical results show that mLDA performs reasonably well even under the misspecified models, as long as the common covariance assumption is not severely

Table 3: Top ten genes selected from the post-transplant rejection study

Gene	Selection frequency	MI or JI
CEACAM8	.90	JI
TCN1	.85	JI
HLA-G	.84	MI
BPI	.82	JI
GUSBP11	.82	MI
IGHV3-23	.79	JI
TAF6L	.74	MI
RNASE3	.73	JI
CRISP3	.66	JI
TCF12	.58	MI

violated. The performance of mLDA might deteriorate when the covariance matrices do differ much across classes. For these cases, the nonlinear classification methods, such as quadratic discriminant analysis, may be more appropriate.

Algorithms can be developed more efficiently. For instance, when the whole feature space can be divided into uncorrelated subspaces, such as blocks of chromosomes in a genome or different functional regions in a brain, parallel computing on multiple partitioned feature spaces can be implemented. We are also cognizant that there might be some marginally weak signals that may not be detected by our method, such as the marginally weak signals with joint pooled effects (Li and Leal, 2008). In this case, the proposed method may be used with other weak signal detection techniques to boost weak signal selection and classification performance. We will pursue this.

Acknowledgements

The authors would like to thank the co-editor, an associate editor and two reviewers for their insightful comments which have helped substantially improve the manuscript. We also thank Drs. Runzhe Li and Jun Shao for providing the source codes for pairwiseLDA and SIR, respectively. The work was supported by several grants of the National Institutes of Health.

References

Allwein, E. L., Schapire, R. E., and Singer, Y. (2000). Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research* **1,** 113–141.

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36,** 2577–2604.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer Science + Business Media, LLC.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106(496),** 1566–1577.

Cai, W., Guan, G., Pan, R., Zhu, X., and Wang, H. (2018). Network linear discriminant analysis. *Computational Statistics and Data Analysis* **117,** 32–44.

Dorfer, M., Kelz, R., and Wildmer, G. (2016). Deep linear discriminant analysis. *Proc. Int.Conf. Learn. Representations* .

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36,** 2605–2637.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Statist. Soc. B* **74(4),** 745–771.

Fan, J., Liao, Y., and Min, M. (2011). High-dimensional covariance matrix estimation in approxiamte factor models. *Ann. Statist.* **39,** 3320–3356.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B* **70,** 849–911.

Flencher, S. M., Kurian, S. M., Head, S. M., Sharp, S. M., Whisenant, T. C., Zhang, J., et al. (2004). Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. *Am. J. Transplant* **4,** 1475–1489.

Gaynanova, I., Booth, J. G., and Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the $p \gg n$ setting. *Journal of the American Statistical Association* **111,** 696–706.

Gorban, A., Golubkov, A., Grechuk, B., Mirkes, E., and Tyukin, I. (2018). Correction of AI systems by linear discriminants: Probabilistic foundations. *Information Sciences* **466,** 303–322.

Gowder, S. (2014). Renal membrane transport proteins and the transporter genes. *Gene Technology* **3:e109,** 229–234.

Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics* **11,** 599–608.

Guo, Y., Hastie, T., and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics* **8,** 86–100.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23,** 73–102.

Hastie, T. and Tibshirani, R. (1995). Discriminant analysis by Gaussian mixtures. *J. R. Statist. Soc. B* **58,** 155–176.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2 edition.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison on methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* **13,** 415–425.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **106,** 8859–8864.

Johnson, V. E. (2013). On numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings. *Bayesian Analysis* **8,** 741–758.

Johnson, V. E. and Rossell, D. (2012). Bayesian variable selection in high-dimensional settings. *Journal of the American Statistical Association* **107,** 649–660.

Kota, S. K., Pernicone, E., Leaf, D. E., Stillman, I. E., Waikar, S. S., and Kota, S. B. (2017). BPI fold-containing family A member 2/parotid secretory protein is an early biomarker of AKI. *J Am Soc Nephrol.* **28,** 3473–3478.

Lazzaro, D., De Simone, V., De Magistris, L., Lehtonen, E., and Cortese, R. (1992). LFB1 and LFB3 homeoproteins are sequentially expressed during kidney development. *Development* **114,** 469–479.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321.

Mai, Q., Yang, Y., and Zou, H. (2017). Multiclass sparse discriminant analysis. *Statistica Sinica,* in press.

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

Ng, A. Y. and Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *NIPS'01 Proceedings of the 14th International Conference on Neural Information* **14,** 841–848.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics* **32,** 1338–1345.

Pan, R., Wang, H., and Li, R. (2016). Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independent screening. *Journal of American Statistical Association* **111,** 169–179.

Platt, J., Cristianini, N., and Shawe-Taylor, J. (2003). Large margin DAGs for multiclass classification. In Solla, S., Leen, T., and Muller, K.-R., editors, *Extreme Values in Finance, Telecommunications, and the Environment*, volume 12, pages 547–553. MIT Press.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Rossell, D. and Rubio, F. J. (2018). Tractable bayesian variable selection: Beyond normality. *Journal of the American Statistical Association,* in press.

Safo, S. E. and Ahn, J. (2016). General sparse multi-class linear discriminant analysis. *Computational Statistics & Data Analysis* **99,** 81–90.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high-dimensional data. *Ann. Statist.* **39,** 1241–1265.

Shapiro, L. and Stockman, G. (2002). *Computer Vision.* Prentice Hall.

Teo, C. H., Vishwanathan, S. V. N., Smola, A., and Le, Q. V. (2010). Bundle methods for regularized risk minimization. *The Journal of Machine Learning Research* **11,** 311–365.

Torgo, L. (2010). *Data Mining using R: learning with case studies.* CRC Press.

Weston, J. and Watkins, C. (1998). Multi-class support vector machines. Technical report, Department of Computer Science, University of London.

Witten, D. M. and Tibshirani, R. J. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc. B* **73,** 753–772.

Wu, T. F., Lin, C. J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research* **5,** 975–1005.

Xiao, L., Zhou, W., Shi, B., Feng, K., He, X., Wei, Y., et al. (2013). HLA-G expression in the peripheral blood of live kidney transplant recipients. *Chin Med J (Engl)* **126,** 2652–2655.

Xu, P., Zhu, J., Zhu, L., and Li, Y. (2014). Covariance-enhanced discriminant analysis. *Biometrika* **102,** 33–45.

Yu, Z., Dong, Y., and Shao, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann. Statist.* **44,** 2594–2623.

Zhang, T. K., Jin, J., and Fan, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42,** 2202–2242.

Appendix: Technical conditions

To derive the sure screening property and asymptotic vanishing post-screening misclassification rate for mLDA, we assume the following conditions.

(A1) For any positive integer $l \geq 1$, $E|X_j|^l \leq l!C^l$ for some constant $C > 0$, $1 \leq j \leq p$.

(A2) (Faithfulness condition) For any j satisfying $\mu_{kj} - \mu_{k'j} \neq 0$ for some $1 \leq k < k' \leq K$, $\sum_{j'=1}^{p} \Omega_{jj'}(\mu_{kj'} - \mu_{k'j'}) \neq 0$.

(A3) $p = O(\exp(n^{\varsigma}))$ for some $0 < \varsigma < 1$ and $C_{\max} = O(\exp(n^{\xi}))$ for some $0 < \xi < \varsigma$, where $C_{\max}$ is the maximum size of the connected components in $\mathbf{\Omega}$ containing at least one MI signal. For any MI feature $j \in \mathcal{S}_1$, if $\mu_{kj} - \mu_{k'j} \neq 0$ for some pair (k, k'), $1 \leq k < k' \leq K$, we assume that $|\mu_{kj} - \mu_{k'j}| > \sqrt{r \log p}$, where $0 < r < 1$ controls the overall strength of the marginally informative features.

(A4) For any subset $\mathcal{V}_l \subset \{1, \ldots, p\}$ with $|\mathcal{V}_l| = O(n)$,

$$\max_{1 \leq k \leq K} \sup_{j} E|X_{ij} 1(Y_i = k, j \in \mathcal{V}_l)|^{2t} < C_t < \infty$$

for some constants $t > 0$ and $C_t > 0$.

(A5) $\min_{(j,j') \in \mathcal{E}} |\Sigma_{jj'}| \geq C n^{(\xi-1)/2}$ and $\max_{(j,j') \notin \mathcal{E}} |\Sigma_{jj'}| = o(n^{(\xi-1)/2})$ for the ξ in (A3).

(A6) There exist positive constants κ_1 and κ_2 such that

$$0 < \kappa_1 < \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) < \kappa_2 < \infty.$$

(A7) $\tau = O((r \log p)^s)$ for some $0 < s < 1/2$ and r given in (A3), $\alpha = O(n^{(\xi-1)/2})$ and $\nu_n = O\left((r(\log p)\exp(n^{\xi}))^{s'}\right)$ for some $0 < s' \leq 1/2$ and ξ given in (A3).

(A8) There exist positive constants c_1 and c_2 such that

$$0 < c_1 \leq \min_{1 \leq k \leq K} \frac{n_k}{n} \leq \max_{1 \leq k \leq K} \frac{n_k}{n} \leq c_2 < \infty$$

for all n.

(A9) For any pair of classes, the number of informative features that differentiate the pair is of $O(n^{\varrho})$, for some uniform constant $\varrho > 0$ not depending on the pairs.

(A10) There are $p^{-\beta}$, $0 < \beta < 1$, fraction of features that are MI and $p^{-\gamma}$, $0 < \gamma < 1$, fraction of features that are JI.

(A11) $\rho_n = C_t(n^{-1}C_{\max}^{4/t}) \to 0$ as $n \to \infty$, where t and C_t are given in (A4) and $C_{\max}$ is defined in (A3).

(A1) is required to prove that the connected components of $\mathbf{\Omega}$ can be consistently recovered by $\widetilde{\mathbf{\Sigma}}$. Gaussian distributed random variables satisfy (A1). (A2) ensures that all MI features belong to $\mathcal{S}_0$. (A3) implies that for each connected component of a marginally informative feature, its size cannot exceed the order of $\exp(n^{\xi})$ for some $\xi \in (0,1)$. This condition is required for consistently estimating precision matrices by thresholding covariance matrices and is also assumed by Bickel and Levina (2008). (A4) ensures that the connected components in $\mathbf{\Omega}$ can be adequately estimated from the connected components in $\widetilde{\mathbf{\Sigma}}$; see Shao et al. (2011) and Bickel and Levina (2008) for details. (A5) guarantees that the true zero entries in Ω can be reliably separated from the nonzero entries in the covariance matrix, and therefore the connected components containing at least one MI feature in the precision matrix can be detected accurately (Zhang et al., 2014). (A6) is commonly assumed on design matrices in high-dimensional settings (Shao et al., 2011; Fan et al., 2011; Bickel and Levina, 2008). (A7) gives the order of tuning parameters τ, α and ν_n to achieve the desired theoretical properties (Shao et al., 2011; Jin, 2009; Fan et al., 2011; Bickel and Levina, 2008). (A8) implies that the K classes are of comparable sample sizes and the sample size of each class goes to infinity when n goes to infinity (Shao et al., 2011; Fan and Fan, 2008). (A9) requires that the numbers of informative features differentiating different pairs of classes have a uniform lower bound of the order $O(n^{\varrho})$. (A10) indicates the sparsity of the MI and JI features and is required for proving Theorem 3. As β and γ get closer to 1, MI and JI features become sparser, respectively (Jin, 2009). (A11) is used to gauge the covariance and precision estimation errors (Shao et al., 2011; Fan et al., 2011; Bickel and Levina, 2008).