HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2017 May 05.

Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(513): 169–179. doi:10.1080/01621459.2014.998760.

Ultrahigh-Dimensional Multiclass Linear Discriminant Analysis by Pairwise Sure Independence Screening

Rui Pan,

Assistant Professor, School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 100081

Hansheng Wang, and

Professor, Guanghua School of Management, Peking University, Beijing, 100871, P.R. China

Runze Li

Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111

Abstract

This paper is concerned with the problem of feature screening for multi-class linear discriminant analysis under ultrahigh dimensional setting. We allow the number of classes to be relatively large. As a result, the total number of relevant features is larger than usual. This makes the related classification problem much more challenging than the conventional one, where the number of classes is small (very often two). To solve the problem, we propose a novel pairwise sure independence screening method for linear discriminant analysis with an ultrahigh dimensional predictor. The proposed procedure is directly applicable to the situation with many classes. We further prove that the proposed method is screening consistent. Simulation studies are conducted to assess the finite sample performance of the new procedure. We also demonstrate the proposed methodology via an empirical analysis of a real life example on handwritten Chinese character recognition.

Keywords

Multi-class Linear Discriminant Analysis; Pairwise Sure Independence Screening; Sure Independence Screening; Strong Screening Consistency

1. INTRODUCTION

Linear discriminant analysis (LDA) has often been used for classification in practice. With the rapid advance in technology, ultrahigh dimensional data become increasingly available and LDA for these data has drawn many attentions. In the ultrahigh dimensional setting, the sample size is typically much smaller than the predictor (or feature) dimension. Accordingly, the inverse of the predictor covariance matrix, which plays an essential role in LDA, is hard to be estimated accurately. What's more, the prediction accuracy of the ordinary LDA can be as poor as random guessing when the predictor dimension is sufficiently high (Bickel and Levina, 2004). As a result, Bickel and Levina (2004) advocated the use of the independence classification rule, which corresponds to the LDA with a diagonally estimated sample

covariance matrix. Nevertheless, the asymptotic performance of the independence rule still can be as poor as random guessing in the presence of ultrahigh dimensional predictors (Fan and Fan, 2008). In order to solve the problem of ultrahigh dimensional LDA, feature screening becomes necessary for practical implementation.

Researches exist on feature screening in LDA with ultrahigh dimensional predictors. Fan and Fan (2008) argued that features used in the final LDA should be carefully selected, and proposed the method of feature annealed independence rule (FAIR). A similar problem was further investigated by Shao et al. (2011) from the perspective of risk optimality. They proposed a method of sparse LDA, which leads to sparse estimates for both the mean and covariance parameters, and showed that the resulting prediction accuracy is asymptotically optimal. For better interpretability and computationally efficiency, a linear programming discriminant (LPD) rule was studied by Cai and Liu (2011), a regularized optimal affine discriminant (ROAD) method was developed by Fan et al. (2012), and a direct approach was proposed by Mai et al. (2012). All those methods are computationally efficient, have excellent theoretical properties, and thus lead to much improved prediction performance.

All aforementioned methods were mainly developed for LDA with two classes, and their extension to multi-class problem is not immediately clear. For example, Fan et al. (2012) has pointed out that the implementation and theoretical properties of ROAD under multi-class setting are interesting topics for future research. In fact, multi-class LDA is important in practice and we find that the number of classes could be relatively large. For instance, in the context of text mining (Weiss et al., 2005), each sample corresponds to one text document and the sample size is the total number of documents. When a large number of documents are available, they can be classified into many categories according to the topics (e.g., news, business, sports, entertainment, and so forth). What's more, the predictors in text mining can be ultrahigh dimensional since the amount of keywords (or features) used in the documents is huge. Most keywords may not be relevant to the class membership. As a result, document classification can be formulated as a classification problem with ultrahigh dimensional feature and a large number of classes.

For multi-class LDA problem, Tibshirani et al. (2003) proposed the nearest shrunken centroid method for class prediction in DNA microarray studies. Witten and Tibshirani (2011) further investigated the problem using penalized LDA (PLDA), which leads to interpretable discriminant vectors. At the same time, Clemmensen et al. (2011) developed a sparse version of LDA using an ℓ_1 penalty, which allows classification and feature selection to be performed simultaneously. When the number of classes is relatively large, the total number of two-class pairs (i.e., two-class LDA problems) could be substantial. Thus, the total number of relevant features diverges to infinity at a rate faster than usual, even if only a few of relevant features contribute to each two-class pair. This motivates us to propose a novel variable screening method, which makes a good use of the pairwise sparsity structure.

Our proposed method is distinctive from the existing methods in the following respects: (a) We propose to solve the ultrahigh dimensional multi-class LDA problem by pairwise LDA, and (b) We propose a feature screening method for pairwise LDA, and establish the strong screening consistency of the proposed procedure under appropriate conditions. Results from

(a) enable us to decompose the multi-class LDA problem into many two-class LDA problems. The feature screening method in (b) enables us to conduct two-class LDA with ultrahigh dimensional features. Furthermore, The strong screening consistency in (b) ensures the proposed screening procedure enjoys the sure screening property in the terminology of Fan and Lv (2008). It further guarantees that overfitting effect can be well controlled. In this paper, we further study the post-screening estimation problem and show that the proposed post-screening estimator for the coefficients in LDA is uniformly consistent over all possible pairs. We also investigate data-driven methods to automatically select the tuning parameters involved in the proposed screening procedures. Our numerical studies show that an EBIC type criterion (Chen and Chen, 2008) performs quite well with a moderate sample size. To further support the usefulness of this method, its screening consistency property is rigorously established.

The rest of the article is organized as follows. In Section 2, we propose a pairwise sure independence screening procedure and establish its theoretical properties. Simulation studies and a real data example are presented in Section 3. A concluding discussion is given in Section 4. All technical proofs are presented in the Appendix.

2. PAIRWISE SURE INDEPENDENCE SCREENING

Let (Y_i, X_i) be the observation collected from the ith $(1 \le i \le n)$ subject, where n is the total sample size and Y_i is the class label taking values in $\{1, 2, \ldots, K\}$. Y_is are assumed to be independent and identically distributed according to probability $P(Y_i = k) = \pi_k > 0$ for every $1 \le k \le K$. For the sake of notation simplicity, it is assumed throughout the rest of this paper that $\pi_k = 1/K$. As a result, the sample size for different classes are likely to be comparable but unlikely to be identical. Furthermore, $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional feature vector. Conditional on $Y_i = k$, X_i follows a multivariate normal distribution with mean $\mu_k = (\mu_{k1}, \ldots \mu_{kp})^\top \in \mathbb{R}^p$ and covariance $\sum = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p \times p}$. Without loss of generality, we assume that the p-dimensional feature vector has been standardized so that $\sigma_{jj} = 1$. Let (Y_0, X_0) be an independent copy of (Y_i, X_i). Then, conditioning on $X_0 = (X_{01}, \ldots, X_{0p})^\top \in \mathbb{R}^p$, the posterior probability of Y_0 is

$$P(Y_0 = k | X_0) \propto \pi_k exp\left\{-2^{-1}(X_0 - \mu_k)^\top \sum^{-1} (X_0 - \mu_k)\right\},$$

where the constants independent of k are ignored. If X_0 is known but Y_0 is not observed, we can then predict Y_0 by maximizing the posterior probability. That is,

$$k^* = arg\max_{1 \le k \le K} exp\left\{-2^{-1}(X_0 - \mu_k)^\top \sum^{-1} (X_0 - \mu_k)\right\}. \quad (2.1)$$

Theoretically, prediction or classification by (2.1) is a very natural choice. However, it is practically infeasible if X is ultrahigh dimensional. This is because estimating the inverse of the covariance matrix (i.e., Σ^{-1}) is challenging (Shao et al., 2011). We are thus motivated to search for alternative solutions.

2.1. Pairwise LDA

Since the exponential function is strictly monotone increasing, it follows by the definition of k^* in (2.1) that for $k \neq k^*$,

$$(X_0 - \mu_k)^\top \sum^{-1} (X_0 - \mu_k) > (X_0 - \mu_{k^*})^\top \sum^{-1} (X_0 - \mu_{k^*}).$$

Then the optimal class prediction k^* can be equivalently expressed by the following set of pairwise inequalities

$$\{X_0 - (\mu_{k^*} + \mu_k)/2\}^\top \beta_{k^*k} > 0 \text{ for every } k \neq k^*. \quad (2.2)$$

where $\beta_{k^*k} = (\beta_{k^*k,1}, \cdots, \beta_{k^*k,p})^\top = \sum^{-1} (\mu_{k^*} - \mu_k) \in \mathbb{R}^p$ for every $1 \leq k^*, k \leq K$. In fact, (2.2) is nothing but a standard linear discriminant function, as defined for an usual two-class LDA problem. Consequently, k^* can be equivalently defined as

$$k^* = arg\max_{k'} \sum_{k \neq k'} I\left(\{X_0 - (\mu_{k'} + \mu_k)/2\}^\top \beta_{k'k} > 0\right). \quad (2.3)$$

With ultrahigh dimensional predictors, the dimension of the coefficient vector β_{k^*k} is also ultrahigh but expected to enjoy certain sparse structure. We can define $\mathcal{M}_{k^*k} = \{j : \beta_{k^*k,j} \neq 0\}$ to collect those indices associated with nonzero coefficients for class pair (k, k^*). Denote by $|\mathcal{M}_{k^*k}|$ the size of $\mathcal{M}_{k^*k}$ (i.e., the number of features contained in $\mathcal{M}_{k^*k}$). Accordingly, the original classification function (2.3) can be re-written as

$$k^* = arg\max_{k'} \sum_{k \neq k'} I\left(\left\{X_{0(\mathcal{M}_{kk'})} - (\mu_{k'(\mathcal{M}_{kk'})} + \mu_{k(\mathcal{M}_{kk'})})/2\right\}^\top \beta_{k'k(\mathcal{M}_{kk'})} > 0\right), \quad (2.4)$$

where $X_{0(\mathcal{M}_{kk'})} = (X_{0_j} : j \in \mathcal{M}_{kk'})^\top \in \mathbb{R}^{|\mathcal{M}_{kk'}|}$ is the subvector of X_0 according to $\mathcal{M}_{kk'}$, while $\mu_{k'(\mathcal{M}_{kk'})}$ and $\beta_{k'k(\mathcal{M}_{kk'})}$ are defined similarly. Operationally, we need to estimate $\mathcal{M}_{kk'}$ first and then conduct LDA based on $\mathcal{M}_{kk'}$. As to be demonstrated in the next section, $\mathcal{M}_{kk'}$ can be estimated by thresholding the estimate of $\mu_k - \mu_{k'}$. The key difference between (2.1) and (2.4) is that the problem involved in (2.4) is low dimensional, while the one in (2.1) is ultrahigh. In other words, the original ultrahigh dimensional LDA problem with many classes is decomposed into many low dimensional ones. This makes the problem computationally feasible.

Remark 1—As one can see in (2.4), we convert a multi-class LDA problem into a set of two-class LDA problems. This makes our method different from Fan and Fan (2008) and Fan et al. (2012), which were solely developed for two-class LDA problem.

Remark 2—Here we follow Cai and Liu (2011), Mai et al. (2012), and Fan et al. (2012) to impose sparsity assumption on $\beta_{k^*k} = \Sigma^{-1}(\mu_{k^*} - \mu_k)$ for the convenience of our classification method (i.e., LDA). This leads to optimal LDA classification rule. In contrast, one can also impose the sparsity assumption on $(\mu_{k^*} - \mu_k)$ as in Fan and Fan (2008), which leads to the independence classification rule.

2.2. Screening Method

Motivated by (2.4), we propose here a novel pairwise variable screening method. Our method is inspired by the seminal theory of sure independence screening (Fan and Lv, 2008, SIS). For convenience, we refer to our method as pairwise sure independence screening (PSIS). To fix the idea, consider a given class pair (k_1,k_2). Then, ideally one should search for important variables by estimating $\beta_{k_1k_2}$. However, this is practically infeasible because

$$\beta_{k_1k_2}=\sum^{-1}(\mu_{k_1}-\mu_{k_2})=\sum^{-1}\gamma_{k_1k_2} \quad (2.5)$$

involves Σ^{-1}, which cannot be estimated accurately when the predictor dimension is ultrahigh. Similar problem was also encountered in usual linear regression by Fan and Lv (2008), where the parameter of interest is the regression coefficient

$$\beta=\left\{cov(\tilde{X})\right\}^{-1}cov(\tilde{X},\tilde{Y}). \quad (2.6)$$

Here $\tilde{Y}\in\mathbb{R}^1$ is the response of interest and $\tilde{X}\in\mathbb{R}^p$ is the predictor. However, if the predictor $\tilde{X}$ is ultrahigh dimensional, $\{cov(\tilde{X})\}^{-1}$ cannot be estimated accurately. To solve the problem, Fan and Lv (2008) creatively replaced $\{cov(\tilde{X})\}^{-1}$ in (2.6) by an identity matrix. This enables them to focus on $cov(\tilde{X},\tilde{Y})$ only. They further showed rigorously that the resulting model estimator is screening consistent. This motivates us to similarly replace Σ^{-1} in (2.5) by an identity matrix. Consequently, this drives us to focus on $\gamma_{k_1k_2}$ for variable screening. We can similarly prove that the resulting model estimator is screening consistent; see Section 2.3.

More specifically, we directly search for promising variables by investigating $\hat{\gamma}_{k_1k_2}=\hat{\mu}_{k_1}-\hat{\mu}_{k_2}$, where $\hat{\mu}_k=n_k^{-1}\sum_i X_iI(Y_i=k)$ and $n_k=\sum_i I(Y_i=k)$. We know immediately that $n=\sum_k n_k$. For convenience, we further write $\gamma_{k_1k_2}=(\gamma_{k_1k_2,1},\cdots,\gamma_{k_1k_2,p})^\top\in\mathbb{R}^p$ and its associated estimator as $\hat{\gamma}_{k_1k_2}=(\hat{\gamma}_{k_1k_2,1},\cdots,\hat{\gamma}_{k_1k_2,p})^\top\in\mathbb{R}^p$. Write $\hat{\mu}_k=(\hat{\mu}_{k1},\cdots,\hat{\mu}_{kp})^\top\in\mathbb{R}^p$. For a given constant $c_{k_1k_2}$, we then estimate $\mathcal{M}_{k_1k_2}$ by

$$\hat{\mathcal{M}}_{k_1k_2}=\left\{j:|\hat{\gamma}_{k_1k_2,j}|>c_{k_1k_2}\right\}. \quad (2.7)$$

As one can see, $\hat{\mathcal{M}}_{k_1k_2}$ depends on both (k_1, k_2) and $c_{k_1k_2}$. For simplicity, we omit the subscript $c_{k_1k_2}$ in $\hat{\mathcal{M}}_{k_1k_2}$. Practically, how to decide the value of $c_{k_1k_2}$ is an important issue and is to be discussed in detail in Section 2.5.

2.3 Theoretical Properties

In this section, we study the theoretical properties of the PSIS method. Obviously, we wish to have $\hat{\mathcal{M}}_{k_1k_2} \supset \mathcal{M}_{k_1k_2}$ for every $1 \le k_1, k_2 \le K$ with large probability. In fact, this can be satisfied trivially if we always define $\hat{\mathcal{M}}_{k_1k_2}=\mathcal{M}_T=\{1,\cdots,p\}$, that is the full model. However, such a solution makes the computation of (2.4) back to an ultrahigh dimensional problem and thus is practically useless. As a result, the size of $\hat{\mathcal{M}}_{k_1k_2}$ need to be simultaneously controlled. Theoretically, this means that a desirable variable screening method should have the following two very nice properties

$$P\left(\hat{\mathcal{M}}_{k_1k_2} \supset \mathcal{M}_{k_1k_2} \text{ for every } 1 \le k_1, k_2 \le K\right) \to 1, \quad (2.8)$$

$$P\left(\max_{k_1,k_2}|\hat{\mathcal{M}}_{k_1k_2}| \le m_{\max}\right) \to 1, \quad (2.9)$$

as $n \to \infty$, where $m_{\max}$ should be a number much smaller than the average class sample size n/K. We refer to both (2.8) and (2.9) as **strong screening consistency**.

To establish (2.8) and (2.9) for the PSIS estimator $\hat{\mathcal{M}}_{k_1k_2}$, the following technical conditions are needed.

- **(C1)** (*Divergence Speed*) Assume that $\log p \le \nu_1 n^{\xi_1}$ for some constant $\nu_1>0$ and $0 < \xi_1 < 1$.
- **(C2)** (*Pairwise Sparsity*) Assume that there exist constants $c_0 > 0$ and $0 \le \xi_0 < 1$, such that $1 \le |\mathcal{M}_{k_1k_2}| \le c_0 n^{\xi_0}$.
- **(C3)** (*Coefficient Regularity*) Define $\gamma_{\min}=\min_{k_1,k_2}\min_{j\in\mathcal{M}_{k_1k_2}}|\gamma_{k_1,k_2,j}|$. Assume $\gamma_{\min} \ge c_1 n^{-\kappa_0}$ for some constant $c_1 > 0$ and $\kappa_0 \ge 0$, where $4\kappa_0 + 3\xi_0 + \xi_1 < 1$.

By (C1) we know that the feature dimension p is allowed to grow exponentially fast with the sample size n. Define $\mathcal{M}_T=\cup_{k_1,k_2}\mathcal{M}_{k_1k_2}$, which collects all the relevant features across all class pairs. Then, by (C2), we know its size $|\mathcal{M}_T$ is bounded by $K^2c_0n^{\xi_0}=o(p)$. This suggests that the overall model structure is relatively sparse as compared with the feature dimension, that is $|\mathcal{M}_T|/p \to 0$. By (C2), we further know that the size of the pairwise model might be even smaller and is $O(n^{\xi_0})$. Condition (C3) requires that the between-class mean difference for those relevant features must stay away from 0 with a good margin. This is a crucial condition that assures the screening consistency property. Under usual regression setups, this is similar to assuming that the marginal correlation coefficient between the

response and a relevant predictor must be well bounded away from 0; see, for example, Fan and Lv (2008), Fan et al. (2011), and Zhu et al. (2011).

Theorem 1—Under Conditions (C1) to (C3), assume that for some positive constant $\tau_{\max}$, $\lambda_{\max}(\Sigma) < \tau_{\max} < \infty$, where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ. Then, as $n \to \infty$, there exists a set of constants $c_{k_1k_2}$ for every $1 \le k_1, k_2 \le K$, such that

$$P\left(\hat{\mathcal{M}}_{k_1k_2} \supset \mathcal{M}_{k_1k_2} \text{ for every } 1 \le k_1, k_2 \le K\right) \to 1, \quad (2.10)$$

$$P\left(\max_{k_1,k_2} |\hat{\mathcal{M}}_{k_1k_2}| \le m_{\max}\right) \to 1, \quad (2.11)$$

where $m_{\max} = m_0 n^{\xi_0 + 2\kappa_0}$ and m_0 is some positive constant.

The proof of Theorem 1 is given in Appendix A. The first conclusion reveals that under appropriate conditions, by pairwise sure independence screening, all the relevant features can be selected consistently and uniformly, that is $\hat{\mathcal{M}}_{k_1k_2} \supset \mathcal{M}_{k_1k_2}$ for every pair (k_1, k_2) with $1 \le k_1, k_2 \le K$. As a result, the proposed method enjoys the so-called screening consistency property. However, it is remarkable that very often we have $\hat{\mathcal{M}}_{k_1k_2} \neq \mathcal{M}_{k_1k_2}$. In fact, the conclusion (2.10) can be satisfied trivially if no constraint is put on $|\hat{\mathcal{M}}_{k_1k_2}|$. For example, as discussed before, we can always set $\hat{\mathcal{M}}_{k_1k_2} = \mathcal{M}_F$. In this case, the true model $\mathcal{M}_{k_1k_2}$ is obviously and seriously overfitted. However, by the second conclusion, we know that the overfitting effect suffered by the PSIS estimator is limited, because the maximal size of $\hat{\mathcal{M}}_{k_1k_2}$ is expected to be much smaller than n/K, if $\lambda_{\max}(\Sigma)$ is bounded. In this case, $m_{\max}$ is of the order $n^{\xi_0+2\kappa_0}$. Then by (C3) we know that $n^{\xi_0+2\kappa_0}/(n/K) \to 0$ as $n \to \infty$. This suggests that maximal size of $\hat{\mathcal{M}}_{k_1k_2}$ should be much smaller than the average class sample size n/K asymptotically.

Remark 3—By Theorem 1, we know that the maximal size of $\hat{\mathcal{M}}_{k_1k_2}$ (i.e., $m_{\max}$) is mainly influenced by two factors. First, it is influenced by the true pairwise model size, that is $|\mathcal{M}_{k_1k_2}| \le cn^{\xi_0}$. As ξ_0 increases, the true pairwise model size gets larger, which inevitably calls for larger sized model estimates for screening consistency. Second, the signal strength $\gamma_{\min}$, as controlled by κ_0, is also important. Larger $\gamma_{\min}$ (i.e., smaller κ_0) makes model identification easier and thus smaller sized model estimate is needed.

Remark 4—Conditions (C1), (C2) and (C3) are not the weakest conditions to establish Theorem 1, but they are used to facilitate the technical proof. One may impose some relaxed conditions to establish results in Theorem 1. For instance, the sparsity assumption may be relaxed to be an approximate one. Specifically, one may re-define $\mathcal{M}_{k_1k_2} = \{j : |\beta_{k_1k_2,j}| > b_n\}$ for some positive sequence $b_n \to 0$ as $n \to \infty$. Imposing Condition (C2) on this newly

defined set $\mathcal{M}_{k_1k_2}$, and assuming that $\sqrt{nb_n} \to \infty$ and $\max_{k_1,k_2}\|\beta_{k_1k_2}\| \le C_{\max}n^{\delta_0}$ for some fixed constants $C_{\max} > 0$ and sufficiently small $\delta_0 > 0$, it can be shown that the screening consistency properties as given in Theorem 1 is still valid.

2.4. Post Screening Estimation

In this section, we investigate how the strong screening consistency property about $\hat{\mathcal{M}}_{k_1k_2}$, that is (2.10) and (2.11), can be reflected in the estimation of $\beta_{k_1k_2}$ after feature screening. More specifically, for a given $\hat{\mathcal{M}}_{k_1k_2}$, we can estimate $\beta_{k_1k_2}$ by $\hat{\beta}_{k_1k_2}=(\hat{\beta}_{k_1k_2,j}:1 \le j \le p)^\top \in \mathbb{R}^p$, where $\hat{\beta}_{k_1k_2,j}=0$ for any $j \notin \hat{\mathcal{M}}_{k_1k_2}$ and $\hat{\beta}_{k_1k_2(\hat{\mathcal{M}}_{k_1k_2})}=(\hat{\beta}_{k_1k_2,j}:j \in \hat{\mathcal{M}}_{k_1k_2})^\top=\hat{\sum}^{-1}_{(\hat{\mathcal{M}}_{k_1k_2})}\hat{\gamma}_{k_1k_2(\hat{\mathcal{M}}_{k_1k_2})} \in \mathbb{R}^{|\hat{\mathcal{M}}_{k_1k_2}}$. Note that $\hat{\gamma}_{k_1k_2(\hat{\mathcal{M}}_{k_1k_2})}=(\hat{\gamma}_{k_1k_2,j}:j \in \hat{\mathcal{M}}_{k_1k_2})^\top \in \mathbb{R}^{|\hat{\mathcal{M}}_{k_1k_2}}$ and $\hat{\sum}^{-1}_{(\hat{\mathcal{M}}_{k_1k_2})}$ is the inverse of $\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})}$, which is a submatrix of the covariance of the estimated covariance $\hat{\sum}$ corresponding to $\hat{\mathcal{M}}_{k_1k_2}$. Here, the estimator $\hat{\sum}=(\hat{\sigma}_{j_1j_2}) \in \mathbb{R}^{p\times p}$ is given by

$$\hat{\sigma}_{j_1j_2}=n^{-1}\sum_{k=1}^{K}n_k\hat{\sigma}_{j_1j_2,k}, \tag{2.12}$$

where $\hat{\sigma}_{j_1j_2,k}=n_k^{-1}\sum_{i=1}^{n}(X_{ij_1}-\hat{\mu}_{kj_1})(X_{ij_2}-\hat{\mu}_{kj_2})I(Y_i=k)$ is the corresponding estimator based on data from the kth class.

Remark 5—It is remarkable that in order to have $\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})}$ invertible we need to have $|\hat{\mathcal{M}}_{k_1k_2}| \le n-K$. On the other hand, with a finite data, it is possible to have $|\hat{\mathcal{M}}_{k_1k_2}|>n-K$. This is particularly true if the sample size is small. If that happens, we would include only the top (n–K) relevant features in $\hat{\mathcal{M}}_{k_1k_2}$ to facilitate computation. According to Theorem 1, we know that this is extremely unlikely to happen if the sample size is reasonably large. Our extensive simulation experience corroborates this theoretical finding quite well.

Recall $\lambda_{\max}(A)$ stands for the largest eigenvalue of an arbitrary semipositive definite matrix A. Similarly, define $\lambda_{\min}(A)$ to be the smallest one. We then have the following theorem about the the asymptotic property of the post screening estimator.

Theorem 2—Assume (C1)–(C3). Further assume that $0 < \tau_{\min} < \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) < \tau_{\max} < \infty$ for some constants $\tau_{\min}$ and $\tau_{\max}$. Then $\max_{k_1,k_2}\|\hat{\beta}_{k_1k_2}-\beta_{k_1k_2}\|=o_p(1)$.

The proof of Theorem 2 is given in Appendix B. By Theorem 2 we know that the post-screening estimator $\hat{\beta}_{k_1k_2}$ is consistent for $\beta_{k_1k_2}$, uniformly over $1 \le k_1, k_2 \le K$. This

property ensures the outstanding performance of the subsequent classification. More specifically, for a new observation X_0, we can predict Y_0 by $\hat{k}$, where

$$\hat{k}=arg\max_{k'}\sum_{k\neq k'} I\left(\left\{X_{0(\hat{\mathscr{M}}_{kk'})}-(\hat{\mu}_{k'(\hat{\mathscr{M}}_{kk'})}+\hat{\mu}_{k(\hat{\mathscr{M}}_{kk'})})/2\right\}^{\top}\hat{\beta}_{k'k(\hat{\mathscr{M}}_{kk'})}>0\right). \quad (2.13)$$

As we shall demonstrate in Section 3, the prediction accuracy of $\hat{k}$ is very comparable to that of k^* in (2.4), that is the theoretically optimal prediction result obtained under the true parameters (that is μ_{k_1}, μ_{k_2}, and $\sum_{k_1k_2}$).

2.5. Tuning Parameter Selection

For practical implementation, the selection of the tuning parameter $c_{k_1k_2}$ is important. Different $c_{k_1k_2}$ might lead to different $\hat{\mathscr{M}}_{k_1k_2}$. When every possible value of $c_{k_1k_2}$ is considered, a solution path $\mathscr{F}_{k_1k_2}=\{\hat{\mathscr{M}}_{k_1k_2}:0\leq c_{k_1k_2}<\infty\}$ is generated. Subsequently, the problem of tuning parameter selection about $c_{k_1k_2}$ is converted into a model selection problem about $\hat{\mathscr{M}}_{k_1k_2}\in\mathscr{F}_{k_1k_2}$. Under a classical regression setup with a fixed predictor dimension, this has been extensively studied. See, for example, AIC (Akaike, 1973), BIC (Schwarz, 1978), EBIC (Chen and Chen, 2008), and recent work (Wang, 2009).

However, it is not immediately clear how to apply those selection criteria (e.g., AIC, BIC, and EBIC) here, because the likelihood function of our problem involves Σ^{-1}, which cannot be estimated accurately due to high dimensinoality. To solve the problem, we enforce a working independence structure on Σ. Specifically, we temporarily assume within this subsection that Σ is diagonal, which is equivalent to assuming that different X_{ij}s are independent with each other for a fixed i but different j. It is remarkable that such an assumption is made here solely for the sake of computational convenience and has nothing to do with the true predictor covariance. Under this assumption, we can easily extend those classical model selection criteria (e.g., AIC, BIC, EBIC) to our situation. Extensive simulation experiments suggest that the resulting EBIC criterion of Chen and Chen (2008) leads to excellent finite sample performance. Furthermore, EBIC's screening consistency property (Fan and Lv, 2008) can be established rigorously for a general Σ matrix, even though it was developed under a working independence assumption. Subsequently, we would focus on the EBIC criterion only.

For simplicity, we consider two arbitrary classes, that is $k\in\{k_1, k_2\}$. We further use $n_{k_1k_2}=n_{k_1}+n_{k_2}$ to denote the sample size from class k_1 and k_2. Recall that $\mathscr{M}_{k_1k_2}$ is the true model which collects indices for nonzero components in $\beta_{k_1k_2}$. We denote $\mathscr{M}$ as an arbitrary candidate model with size $|\mathscr{M}|$. Next, we impose the aforementioned working independence assumption. Accordingly, the negative two times maximum log likelihood function is given by

$$\ell_{k_1k_2}(\mathscr{M})=n_{k_1k_2}\left(\sum_{j\in\mathscr{M}}log\hat{\sigma}_{jj,k_1k_2}+\sum_{j\notin\mathscr{M}}log\tilde{\sigma}_{jj,k_1k_2}\right),$$

where some irrelevant constants are omitted and $\hat{\sigma}_{jj,k_1k_2}=n_{k_1k_2}^{-1}(n_{k_1}\hat{\sigma}_{jj,k_1}+n_{k_2}\hat{\sigma}_{jj,k_2})$, $\tilde{\sigma}_{jj,k_1k_2}=n_{k_1k_2}^{-1}\sum_{i=1}^{n}(X_{ij}-\tilde{\mu}_{j,k_1k_2})^2I(Y_i\in\{k_1,k_2\})$, where $\hat{\sigma}_{jj,k_1}$ and $\hat{\sigma}_{jj,k_2}$ are defined in (2.12), $\tilde{\mu}_{j,k_1k_2}=n_{k_1k_2}^{-1}\sum_{i=1}^{n}X_{ij}I(Y_i\in\{k_1,k_2\})$. Note that for each $j\in\mathscr{M}$, we need two different mean parameters (i.e., μ_{k_1j} and μ_{k_2j}). Thus, each $j\in\mathscr{M}$ consumes 2 degrees of freedom. In contrast, for each $j\notin\mathscr{M}$, only 1 degree of freedom is needed. Thus, the overall degrees of freedom for $\ell_{k_1k_2}(\mathscr{M})$ is given by $2|\mathscr{M}|+|\mathscr{M}^c|=|\mathscr{M}|+p$, because $|\mathscr{M}|+|\mathscr{M}^c|=p$, where $\mathscr{M}^c=\mathscr{M}_F/\mathscr{M}$. We then follow the idea of EBIC (Chen and Chen, 2008) and propose the following model selection criterion as

$$\mathrm{EBIC}_{k_1k_2}(\mathscr{M})=\ell_{k_1k_2}(\mathscr{M})+(logn_{k_1k_2}+2logp)(|\mathscr{M}|+p).$$

We then select the optimal model as $\hat{\mathscr{M}}_{k_1k_2}^{ebic}=arg\min_{\mathscr{M}\in\mathscr{F}_{k_1k_2}}\mathrm{EBIC}_{k_1k_2}(\mathscr{M})$. Its screening consistency property is given by the following theorem.

Theorem 3—Assume conditions (C1)–(C3) hold and $\lambda_{\max}(\Sigma)$ is finite. It follows that

$$P(\hat{\mathscr{M}}_{k_1k_2}^{ebic}\supset\mathscr{M}_{k_1k_2},\ for\ every\ k_1,k_2)\to 1\ as\ n\to\infty.$$

The proof of Theorem 3 is given in Appendix C. Our extensive numerical experiences suggest that $\hat{\mathscr{M}}_{k_1k_2}^{ebic}$ performs quite well. See next section for numerical evidence.

3. NUMERICAL STUDIES

3.1. Simulation Models

We present five examples here. The first three examples examine the proposed methods with different covariance structures. The fourth example investigates its sensitivity towards the normality assumption. The last example considers a case with very unbalanced π_k distribution.

Example 1. (Independent Covariance Structure)—We first consider a simple example with independent features. The simulation setup is similar to the second example in Guo et al. (2007). More specifically, we first generate $Y_i\in\{1,\ldots,K\}$ according to $P(Y_i=k)=1/K$. Given $Y_i=k$, X_i is generated from a multivariate normal distribution with $E(X_i|Y_i=k)=\mu_k$, where $\mu_k=(0,\ldots,0,\mu_{kk},0,\ldots,0)^\top\in\mathbb{R}^p$ is a p-dimensional vector with $\mu_{kk}=\mu$. Furthermore, the conditional covariance is given by $\mathrm{cov}(X_i|Y_i=k)=\Sigma=I_p$, where I_p is a $p\times p$ identity matrix. It is easily verified that $\mathscr{M}_{k_1k_2}=\{k_1,k_2\}$, and thus $\mathscr{M}_T=\cup\mathscr{M}_{k_1k_2}=\{j:1\le j\le K\}$.

Example 2. (Autoregressive Covariance Structure)—In this example, we consider an autoregressive covariance structure. The data are generated in a similar manner as Example 1 but with two differences. Firstly, Σ is fixed as $\sigma_{j_1 j_2}=0.5^{|j_1-j_2|}$. Note that $\sum^{-1}=(\omega_{j_1 j_2}) \in \mathbb{R}^{p\times p}$ is very sparse with $\omega_{11} = \omega_{pp} = 4/3$, $\omega_{jj} = 5/3$ for $1 < j < p$, $\omega_{j(j+1)} = \omega_{(j+1)j} = -2/3$ for $1 \leq j < p$, and $\omega_{j_1 j_2}=0$ whenever $|j_1 - j_2| > 1$. Secondly, the mean vector μ_k is set to be $\mu_k = \mu\Sigma_{(k)}$, where $\Sigma_{(k)}$ stands for the kth column vector of Σ. Accordingly, it follows that $\mathcal{M}_{k_1 k_2}=\{k_1, k_2\}$ with $\mathcal{M}_\tau=\{j:1 \leq j \leq K\}$.

Example 3. (Compound Symmetric Covariance Structure)—In this example, we explore another important covariance structure. The data are generated in a similar manner as in Example 1. However, the covariance is changed to $\sigma_{j_1 j_2}=0.5+0.5I(j_1=j_2)$, which is a compound symmetric structure with diagonal components being 1 but all others being 0.5. One can verify that $\mathcal{M}_{k_1 k_2}=\{k_1, k_2\}$ with $\mathcal{M}_\tau=\{j:1 \leq j \leq K\}$. Furthermore, we know that $\lambda_{\max}(\Sigma) = 0.5(p+1)$. It is then of great interest to examine how the proposed procedure is sensitive when the largest eigenvalue of the covariance matrix is diverging.

Example 4. (Normality Assumption)—In this example, we investigate the sensitivity of the proposed procedure to the normality assumption. To this end, $Y_i \in \{1, \ldots, K\}$ is generated according to $P(Y_i = k) = 1/K$. Given $Y_i = k$, we then generate the predictors as $X_i = \mu_k I(Y_i = k) + Z_i$, where $\mu_k = (0, \ldots, 0, \mu_{kk}, 0, \ldots, 0)^\top \in \mathbb{R}^p$ with $\mu_{kk} = \mu$. Furthermore, each component of the random vector $Z_i = (Z_{i1}, \cdots, Z_{ip})^\top \in \mathbb{R}^p$ is independently simulated from a centralized standard exponential distribution, that is exp(1) − 1. Again, we have $\mathcal{M}_{k_1 k_2}=\{k_1, k_2\}$ and $\mathcal{M}_\tau=\{j:1 \leq j \leq K\}$.

Example 5. (Unbalanced Case)—In the last example, we consider a case with very unbalanced π_k distribution. Specifically, we fix $\pi_1 = 1/5$ and $\pi_k = 4/\{5(K-1)\}$ for $k = 2, \ldots, K$. As one can see, $\pi_1 = P(Y_i = 1)$ is the dominating case and it remains as a constant 1/5 as $K \to \infty$. In contrast, $\pi_k \to 0$ as $K \to \infty$ for every $k > 1$. Given the class label Y_i, X_i is then generated similarly as in Example 1. Once again, $\mathcal{M}_{k_1 k_2}=\{k_1, k_2\}$ and $\mathcal{M}_\tau=\{j:1 \leq j \leq K\}$.

3.2. Assessment Criteria

For each simulation model, we fix $\mu = 5$ and $p = 10,000$. Various (n, K) combinations are considered. They are particularly selected so that the average class sample size (n/K) increases as $n \to \infty$. For each (n, K) combination, the experiment is randomly replicated (or repeated) 1,000 times. For each replication, two independent datasets are generated. One is used for training while the other is reserved for testing. The sample size of the training dataset is given by n, as mentioned previously. That of the testing is fixed to be 500. Let $\hat{\mathcal{M}}^{(r)}_{k_1 k_2}$ be the model estimator obtained in the rth replication. We then consider the following measures to gauge the performances (Wang et al., 2007; Wang, 2009, 2012).

First, for each class pair (k_1, k_2) and each simulation replication r, define, respectively, the model size (MS), percentage of correct zeros (CZ), incorrect zeros (IZ), coverage probability (CP), and the root of the sum squared error (RSSE) as follows:

$$\begin{aligned}
\mathrm{MS}_{k_1k_2}^{(r)} &= |\hat{\mathscr{M}}_{k_1k_2}^{(r)}|, \\
\mathrm{CZ}_{k_1k_2}^{(r)} &= \frac{|\mathscr{M}_F \setminus (\mathscr{M}_{k_1k_2} \cup \hat{\mathscr{M}}_{k_1k_2}^{(r)})|}{|\mathscr{M}_F \setminus \mathscr{M}_{k_1k_2}|}, \\
\mathrm{IZ}_{k_1k_2}^{(r)} &= \frac{|(\mathscr{M}_F \setminus \hat{\mathscr{M}}_{k_1k_2}^{(r)}) \cap \mathscr{M}_{k_1k_2}|}{|\mathscr{M}_{k_1k_2}|}, \\
\mathrm{CP}_{k_1k_2}^{(r)} &= I(\hat{\mathscr{M}}_{k_1k_2} \supset \mathscr{M}_{k_1k_2}), \\
\mathrm{RSSE}_{k_1k_2}^{(r)} &= \|\hat{\beta}_{k_1k_2}^{(r)} - \beta_{k_1k_2}\|,
\end{aligned}$$

where $\hat{\beta}_{k_1k_2}^{(r)}$ is the post-screening estimator obtained in the rth simulation replication according to Section 2.4. We then average the above performance measures across not only every simulation replication (that is $1 \le r \le 1,000$) but also every class pair (that is $1 \le k_1, k_2 \le K$). This leads to the final performance measures, denoted MS, CZ, IZ, CP, and RSSE, respectively.

In addition, we consider the maximum MS (MMS), uniform CP (UCP), and maximum RSSE (MRSSE) measures as follows:

$$\begin{aligned}
\mathrm{MMS} &= 1000^{-1} \sum_{r=1}^{1000} \max_{k_1,k_2} \mathrm{MS}_{k_1,k_2}^{(r)}, \\
\mathrm{UCP} &= 1000^{-1} \sum_{r=1}^{1000} \min_{k_1,k_2} \mathrm{CP}_{k_1,k_2}^{(r)}, \\
\mathrm{MRSSE} &= 1000^{-1} \sum_{r=1}^{1000} \max_{k_1,k_2} \mathrm{RSSE}_{k_1,k_2}^{(r)},
\end{aligned}$$

Lastly, the classification accuracy (i.e., CA = the percentage of the samples satisfying $\hat{k}=k$) of the post-screening estimator is evaluated according to (2.13) on the testing dataset. It is calculated for each experiment and then averaged across the 1,000 random replications. For comparison, the ideal classifier, that is k^* as given in (2.1), is also evaluated. The ratio of the averaged CA values with $\hat{k}$ over k^* is computed and is referred to as the relative classification accuracy (RCA). Lastly, the averaged CPU time (CPU) consumed by our method is also reported in seconds. The detailed simulation results are summarized in Table 1.

3.3. Simulation Results

Because the patterns are qualitatively similar across different models, we only focus on Model 1 for interpretation. First of all, we find that most MS and MMS are bounded below 5 and some could be very close to the target $|\mathscr{M}_{k_1k_2}|=2$. Accordingly, the estimated models demonstrated excellent ability to produce correct sparse solutions with CZ values always staying at 100%. As a result, the uniform bounded property of (2.11) should hold for the estimated models. Meanwhile, the uniform coverage property of (2.10) should also hold

because the reported CP and UCP values quickly increase towards 100% as n increases. Accordingly, the strong screening consistency property, as defined by (2.10) and (2.11), should hold for the estimated models. Lastly, the corresponding post screening estimator is uniformly consistent, with both RSSE and MRSSE steadily decreasing towards 0, as n increases. This leads to excellent classification performance, with RCA values above 99% in most cases. To gauge computational speed, the average CPU time used by each PSIS replication is also reported in the last column. We find it increases quickly as K increases. This is expected because the the number of pairs need to be classified increases as a quadratic function in K.

For the sake of comparison, the methods of neural network (NNet), support vector machine (SVM), and the penalized linear discriminant analysis (Witten and Tibshirani, 2011, PLDA) are also evaluated by CA in a similar manner as for PSIS. They are implemented by the existing R-packages (i.e., *nnet, e1701*, and *penalizedLDA*). To save computational time in R, they are replicated 100 times for each example with $p = 2,000$ and $(n, K) = (400, 10)$. The detailed results are given in Table 2. We find that the performance of NNet and SVM are quite poor with classification accuracy no more than 30%. In contrast, that of the PLDA is much better. In some cases (i.e., Models 1, 4 and 5), its performance is extremely comparable with that of the PSIS. However, for other cases, it is clearly outperformed by PSIS.

3.4. Handwritten Chinese Characters

To illustrate the practical usefulness of the proposed method, we present here a real example about automatic recognition of handwritten Chinese characters. For illustration, we consider a total of $K = 10$ frequently used Chinese characters, with each character representing one class. We hire some volunteers to write down these characters. For each Chinese character, a total of 35 handwritten samples are obtained. We randomly split the 35 samples into two sets. The first set contains 30 samples for training and the other contains the remaining 5 for testing. Furthermore, each sample has been converted into 25×25 pixel data in grayscale, which results in a high dimensional predictor with $p = 25 \times 25 = 625$. See Figure 1 for some handwritten samples. Accordingly, the effective sample size used for model training is $n = 10 \times 30 = 300$, which is much smaller than the predictor dimension $p = 625$.

We then apply the proposed PSIS method in conjunction with the EBIC criterion to the training dataset. This leads to the model estimate $\hat{\mathscr{M}}_{k_1 k_2}^{ebic}$ for every $1 \leq k_1, k_2 \leq K$. Accordingly, the post screening estimator $\hat{\beta}_{k_1 k_2}$ can be obtained, whose forecasting accuracy is evaluated on the testing data. We randomly repeat this procedure 1,000 times, which leads to 1,000 different partition of the training and testing samples. For every random replication, various performance measures are computed. Those measures are then averaged over 1,000 experiments and reported in Table 3. For the sake comparison, the methods of NNet, SVM and PLDA are also evaluated. The detailed results are given in Table 3. We find that for PSIS, the average model size is 15.61. The total number of selected features is 60.69. In contrast, that of the PLDA is 173.41. In terms of classification accuracy, PSIS performs best with CA=93.86%, followed by 87.68% of PLDA, 85.41% of NNet and 78.40% of SVM.

Remark 6—It is remarkable that the Chinese characters considered in this case are similar to each other. That makes the coefficient $\beta_{k_1k_2}=\sum^{-1}\gamma_{k_1k_2}$ very sparse for every class pair. However, if two Chinese characters are drastically different, the number of relevant pixels could be much larger. This makes $\beta_{k_1k_2}$ less sparse and thus the proposed PSIS method becomes less effective. As a result, it is important to first cluster a huge number of Chinese characters into different groups, so that the characters within each group are sufficiently similar to each other. However, how to conduct this preliminary clustering analysis effectively and correctly is a very interesting research topic currently under investigation. Our preliminary numerical experiences seem to be extremely encouraging.

4. CONCLUDING REMARKS

In this paper, we developed a new feature screening method for ultrahigh dimensional LDA with many classes. We propose a pairwise method for variable screening. This leads to pairwise classification with competitive performances. We rigorously show that the proposed procedure enjoys the strong screening consistency property, which implies that screening consistency holds uniformly over a large number of classes. It is remarkable that this property is established with uniformly bounded model sizes across every class pair under appropriate conditions. We further show that the post-screening estimator is uniformly consistent.

The newly proposed method is based on normality assumption. The traditional multivariate skewness and kurtosis (Mardia, 1970) involves the inverse of sample covariance matrix, and therefore they cannot be directly applied for test of ultrahigh dimensional normality. Multivariate Ghosh's T_3-plot developed in Fang et al. (1998) provides a graphical tool for detecting the ultrahigh non-multinormality. The test of multinormality based on low-dimensional projection (Liang et al., 2000) may be used for test of ultrahigh dimensional normality.

In order to solve a multi-class LDA problem, there exist at least two competing choices. One is to solve the problem jointly, i.e., evaluate the posterior probability for each class according to the joint likelihood. As an alternative, one can also solve the problem in a pairwise manner as what has been proposed here. Then, it is natural to ask which solution is better under an ultrahigh dimensional setup? Such an important theoretical question was never rigorously addressed in the past literature, according to our best knowledge. We then fulfil this gap by demonstrating both theoretically and numerically that pairwise classification is optimal under appropriate conditions. With this solid theoretical foundation, one can imagine immediately that essentially any well developed two-class sparse LDA methods (Fan and Fan, 2008; Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012) can be readily extended to multi-class LDA problems. The resulting classification performances are also likely to be optimal. We regard that this is one of our major contributions in this work.

To conclude this article, we would like to discuss here some interesting topics for future research. First, as demonstrated by the simulation study, as K increases, the demanded CPU time also increases at a quadratic rate. This places a serious challenge for PSIS with very large K. Then, how to reduce the computational complexity without sacrificing forecasting

accuracy is our first important topic for future research. Second, when K is large, the common covariance assumption becomes questionable. It is more likely to have unequal covariances. That leads to pairwise quadratic discriminant analysis (QDA). Then, how to conduct consistent variable screening for pairwise QDA is another interesting topic. Numerically, it seems that our current pairwise screening method can be directly applied. However, theoretically whether this can be justified would be the key research question. Third, our current theoretical results, given in Theorems 1 and 2, are all based on asymptotic analysis without explicit convergence rate for $\hat{\beta}_{k_1k_2}$. However, as pointed out by one referee, the convergence rate should be useful if one wants to understand the asymptotic behavior of the risk of the Bayes classifier. This is particularly true if $\|\beta_{k_1k_2}\|=\|\sum^{-1}(\mu_{k_1}-\mu_{k_2})\|\to\infty$. In this case the risk of the Bayes classifier goes to 0. Lastly, another referee pointed out that it may be of interest to investigate finite sample theoretical properties. Theoretical study on the risk of the Bayes classifier and finite sample properties of the proposed procedure is out of the scope of this paper. They both are interesting topics for future research.

Acknowledgments

Rui Pan and Hansheng Wang's research was supported in part by National Natural Science Foundation of China (NSFC, 11131002, 11271032), Fox Ying Tong Education Foundation, the Center for Statistical Science at Peking University, and the Business Intelligence Research Center at Peking University. Wang's research was also supported in part by the Methodology Center at Pennsylvania State University. Runze Li's research was supported by National Institute of Health grants P50-DA10075, R01 CA168676, R01 MH096711 and NSFC 11028103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Appendix A. Proof of Theorem 1

The proof of Theorem 1 consists of four steps. In Step 1, we show that the sample size of the kth class n_k is of order $O_p(n/K)$. In Step 2, we show that $\hat{\gamma}_{k_1k_2}$ is a uniformly consistent estimator for $\gamma_{k_1k_2}$. In Step 3, we show that with probability tending to one, $\mathscr{M}_{k_1k_2}\subset\hat{\mathscr{M}}_{k_1k_2}$ uniformly over every $1\le k_1, k_2\le K$. In Step 4, we further show that the size of $\hat{\mathscr{M}}_{k_1k_2}$ can be uniformly bounded.

Step 1

For every $1\le k\le K$, define $R_k=(n_k-nK^{-1})/(nK^{-1})$. We then have $n_k=nK^{-1}(1+R_k)$. Define $R=\max_k|R_k|$ and we then have

$$n_k\ge nK^{-1}(1-R). \quad \text{(A.1)}$$

Subsequently, we will show R is an $o_p(1)$. Specifically, recall that $n_k=\sum_i I(Y_i=k)$ and define $Z_i=I(Y_i=k)-\pi_k$. We then have $EZ_i=0$, $EZ_i^2=\pi_k-\pi_k^2$, and $|Z_i|\le M$ with $M=1$. As a result, by Bernstein's inequality, it follows that

$$P\left(\sum_i Z_i > \varepsilon\right) \leq exp\left\{\frac{-3\varepsilon^2}{2M\varepsilon + 6\sum_i EZ_i^2}\right\},$$

where $\varepsilon > 0$ is an arbitrary positive constant. Similar inequality also can be obtained for $-Z_i$. Note that $\Sigma_i Z_i = n_k - n\pi_k$, and $\pi_k = 1/K$. We then have that

$$P\left(|n_k - n/K| > \varepsilon\right) \leq 2exp\left\{\frac{-3\varepsilon^2}{2M\varepsilon + 6n/K - 6n/K^2}\right\}.$$

Recall that $R = \max_k |R_k| = \max_k (n_k - n/K)/(nK^{-1})$, we then have

$$\begin{aligned} P(R > \varepsilon) = P(\max_k |R_k| > \varepsilon) &= P(\max_k |n_k - n/K| > \varepsilon n/K) \\ &\leq \sum_k P(|n_k - n/K| > \varepsilon n/K) \\ &\leq 2Kexp\left\{\frac{-3\varepsilon^2 n^2/K^2}{2M\varepsilon n/K + 6n/K - 6n/K^2}\right\} \\ &= 2Kexp\left\{\frac{-3\varepsilon^2 n/K}{2M\varepsilon + 6 - 6/K^2}\right\} \to 0 \end{aligned}$$

as $n \to \infty$. As a result, we have $R = o_p(1)$. This accomplishes the first step.

Step 2

Since $\hat{\mu}_{kj} = \sum_i X_{ij} I(Y_i = k)/n_k$, we know that $\hat{\mu}_{kj}$ is distributed as a normal variable with mean μ_{kj} and variance σ_{jj}/n_k. For the normal distribution, we have the following tail probability inequality (Fan and Fan, 2008): $1 - \Phi(x) \leq (\sqrt{2\pi}x)^{-1} exp(-x^2/2)$, where $\Phi(\cdot)$ is the cumulative function of a standard normal distribution. For simplicity, from now on, we use $\mathbb{P}(\cdot)$ to denote the conditional probability with n_ks given. Next, let ν_n be a positive constant of order $n^{-\kappa_0}$, we have

$$\mathbb{P}(|\hat{\mu}_{kj} - \mu_{kj}| > \nu_n) = \mathbb{P}\left(\left|\frac{\hat{\mu}_{kj} - \mu_{kj}}{\sqrt{\sigma_{jj}}/\sqrt{n_k}}\right| > \frac{\sqrt{n_k}\nu_n}{\sqrt{\sigma_{jj}}}\right) = 2\left\{1 - \Phi\left(\frac{\sqrt{n_k}\nu_n}{\sqrt{\sigma_{jj}}}\right)\right\}.$$

The above quantity is further bounded by

$M_1 exp(-M_2 n_k \nu_n^2) \leq M_1 exp\{-M_2 nK^{-1}(1-R)\nu_n^2\}$, where M_1 and M_2 are some positive constants depending on ν_n. The last inequality is due to (A.1) in the first step. By definition, we know that $\hat{\gamma}_{k_1 k_{2,j}} = \hat{\mu}_{k_1 j} - \hat{\mu}_{k_2 j} = \sum_i X_{ij} I(Y_i = k_1)/n_{k_1} - \sum_i X_{ij} I(Y_i = k_2)/n_{k_2}$. As a result,

$$\begin{aligned} \mathbb{P}(|\hat{\gamma}_{k_1 k_2, j} - \gamma_{k_1 k_2, j}| > \nu_n) &= \mathbb{P}(|\hat{\mu}_{k_1 j} - \mu_{k_1 j} - \hat{\mu}_{k_2 j} + \mu_{k_2 j}| > \nu_n) \\ &\leq \mathbb{P}(|\hat{\mu}_{k_1 j} - \mu_{k_1 j}| > \nu_n/2) + P(|\hat{\mu}_{k_2 j} - \mu_{k_2 j}| > \nu_n/2) \\ &\leq C_1 exp\{-C_2 nK^{-1}(1-R)\nu_n^2\}, \end{aligned} \quad \text{(A.2)}$$

for some constants C_1 and C_2. Thus, it follows that

$$\begin{aligned}&\mathbb{P}(\max_{k_1,k_2,j}|\hat{\gamma}_{k_1k_2,j}-\gamma_{k_1k_2,j}|>\nu_n)\\ \le &\sum_{k_1,k_2,j}\mathbb{P}(|\hat{\gamma}_{k_1k_2,j}-\gamma_{k_1k_2,j}|>\nu_n)\le K^2pC_1exp\{-C_2nK^{-1}(1-R)\nu_n^2\}\\ =&C_1exp\{2logK+logp-C_2nK^{-1}(1-R)\nu_n^2\}\\ \le &C_0exp\{2logK+\nu_1n^{\xi_1}-C_2K^{-1}n^{1-2\kappa_0}(1-R)\}\to_p0\end{aligned} \tag{A.3}$$

as $n \to \infty$, where C_0 is some constant and the last convergence result is due to Condition (C3) and also the fact that $R = o_p(1)$ in Step 1. Consequently, we have that

$$P\left(\max_{k_1,k_2,j}|\hat{\gamma}_{k_1k_2,j}-\gamma_{k_1k_2,j}|>\nu_n\right)=E\left\{\mathbb{P}\left(\max_{k_1,k_2,j}|\hat{\gamma}_{k_1k_2,j}-\gamma_{k_1k_2,j}|>\nu_n\right)\right\}\to_p0 \tag{A.4}$$

as $n \to \infty$. This accomplishes the second step.

Step 3

Define $\hat{\mathscr{M}}_{k_1k_2}=\{j:|\hat{\gamma}_{k_1k_2,j}|>\gamma_{\min}/2\}$, where $\gamma_{\min}$ is defined in Condition (C3). Accordingly, we want to show that $\hat{\mathscr{M}}_{k_1k_2}$ should uniformly cover $\mathscr{M}_{k_1k_2}$ with probability tending to one. Otherwise, there must exist a pair (k_1, k_2), and at least one $j^* \in \mathscr{M}_{k_1k_2}$ missed by $\hat{\mathscr{M}}_{k_1k_2}$. By the definition of $\hat{\mathscr{M}}_{k_1k_2}$, we must have $|\hat{\gamma}_{k_1k_2,j^*}| \le \gamma_{\min}/2$. However, by Condition (C3), if $j^* \in \mathscr{M}_{k_1k_2}$, then $|\gamma_{k_1k_2,j^*}|>\gamma_{\min}$. These together suggest that $|\hat{\gamma}_{k_1k_2,j^*}-\gamma_{k_1k_2,j^*}|>\gamma_{\min}/2$. As a result, if $\hat{\mathscr{M}}_{k_1k_2} \not\supset \mathscr{M}_{k_1k_2}$, we must have $\max_{k_1k_2,j}|\hat{\gamma}_{k_1k_2,j}-\gamma_{k_1k_2,j}|>\gamma_{\min}/2$. Define $\nu_n=\gamma_{\min}/2$, by (A.4) and Condition (C3) we know that

$P(\hat{\mathscr{M}}_{k_1k_2} \not\supset \mathscr{M}_{k_1k_2} \text{ for some } k_1,k_2) \le P(\max_{k_1,k_2,j}|\hat{\gamma}_{k_1k_2,j}-\gamma_{k_1k_2,j}|>\gamma_{\min}/2)\to_p0$, as $n \to \infty$. This suggests that $P(\hat{\mathscr{M}}_{k_1k_2} \supset \mathscr{M}_{k_1k_2} \text{ for every } k_1,k_2)\to_p1$ as $n \to \infty$. This finishes the third step.

Step 4

We next verify that the size of $\hat{\mathscr{M}}_{k_1k_2}$ can be uniformly bounded. First, by (C1) to (C3), we have $\|\gamma_{k_1k_2}\|^2 \le \lambda_{\max}(\sum{}'\sum)\|\beta_{k_1k_2}\|^2 \le c_0\lambda^2_{\max}(\sum)\beta^2_{\max}n^{\xi_0} \doteq \gamma^2_{\max}$, where $\beta_{\max}=\max_{k_1,k_2,j\in\mathscr{M}_{k_1k_2}}|\beta_{k_1k_2,j}|$ This immediately suggests that

$$\max_{k_1,k_2}\|\gamma_{k_1k_2}\| \le \gamma_{\max}. \tag{A.5}$$

Further define $\mathscr{M}^*_{k_1k_2}=\{j:|\gamma_{k_1k_2,j}|>\gamma_{\min}/4\}$. Then, $\gamma^2_{\max} \ge \|\gamma_{k_1k_2(\mathscr{M}^*_{k_1k_2})}\|^2 \ge |\mathscr{M}^*_{k_1k_2}|\gamma^2_{\min}/16$. By (C3), then

$|\mathcal{M}^*_{k_1k_2}| \le 16\gamma^2_{\max}\gamma^{-2}_{\min} = m_0 n^{\xi_0+2\kappa_0} \doteq m_{\max}$, where m_0 is some fixed constant. Because $m_{\max}$ is independent of k_1 and k_2, we know that $\max_{k_1,k_2}|\mathcal{M}^*_{k_1k_2}| \le m_{\max}$. Thus, Theorem 1 follows by proving that, with probability tending to one, $|\mathcal{M}^*_{k_1k_2}| \ge |\hat{\mathcal{M}}_{k_1k_2}|$ for any pair (k_1, k_2). Assume that there exists a pair (k_1, k_2) such that $|\hat{\mathcal{M}}_{k_1k_2}| > |\mathcal{M}^*_{k_1k_2}|$, we then must have $\mathcal{M}^*_{k_1k_2} \not\supset \hat{\mathcal{M}}_{k_1k_2}$. This means there must exist at least one $\hat{j} \in \hat{\mathcal{M}}_{k_1k_2}$ but $\hat{j} \notin \mathcal{M}^*_{k_1k_2}$. Because $\hat{j} \in \hat{\mathcal{M}}_{k_1k_2}$, we know that $|\hat{\gamma}_{k_1k_2,\hat{j}}| > \gamma_{\min}/2$. On the other hand, because $\hat{j} \notin \hat{\mathcal{M}}^*_{k_1k_2}$, we must have $|\gamma_{k_1k_2,\hat{j}}| \le \gamma_{\min}/4$. We know immediately that $|\hat{\gamma}_{k_1k_2,\hat{j}} - \gamma_{k_1k_2,\hat{j}}| > \gamma_{\min}/4$, and then $\max_{k_1,k_2,j}|\hat{\gamma}_{k_1k_2,j} - \gamma_{k_1k_2,j}| > \gamma_{\min}/4$. By (A.4) with $\nu_n = \gamma_{\min}/4$ and Condition (C3) we have $P(\mathcal{M}^*_{k_1k_2} \not\supset \mathcal{M}_{k_1k_2} \text{ for some } k_1, k_2) \to_p 0$, as $n \to \infty$. This suggests that $P(|\hat{\mathcal{M}}_{k_1k_2}| \le m_{\max}) \to 1$ as $n \to \infty$. This completes the proof of the last step and hence of Theorem 1.

Appendix B. Proof of Theorem 2

Recall that $\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})}$ and $\sum_{(\hat{\mathcal{M}}_{k_1k_2})}$ are the submatrices of $\hat{\sum}$ and Σ, respectively. In order to prove Theorem 2, we need the following lemma.

Lemma 1

Under Conditions of Theorem 2, then with probability tending to one,

$$P\left(\max_{k_1,k_2}\lambda_{\max}\left\{\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})} - \sum_{(\hat{\mathcal{M}}_{k_1k_2})}\right\} > n^{-\xi_0/2}\right) \to 0 \quad \text{(A.6)}$$

$$2^{-1}\tau_{\min} < \min_{k_1,k_2}\lambda_{\min}\left\{\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})}\right\} \le \max_{k_1,k_2}\lambda_{\max}\left\{\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})}\right\} < 2\tau_{\max}. \quad \text{(A.7)}$$

Proof

Note that the second conclusion (A.7) can be easily verified if (A.6) is correct. As a result, we focus on (A.6) only. To this end, let $r = (r_1, \ldots, r_p)^\top$ be an arbitrary p-dimensional vector, and ${}^{r}(\hat{\mathcal{M}}_{k_1k_2})$ be its subvector corresponding to set $\hat{\mathcal{M}}_{k_1k_2}$. Then the desired conclusion (A.6) can be implied by

$$P\left(\max_{k_1,k_2}\sup_{\|r_{(\hat{\mathcal{M}}_{k_1k_2})}\|=1}\left|r^\top_{(\hat{\mathcal{M}}_{k_1k_2})}\left\{\hat{\sum}_{(\hat{\mathcal{M}}_{k_1k_2})} - \sum_{(\hat{\mathcal{M}}_{k_1k_2})}\right\} r_{(\hat{\mathcal{M}}_{k_1k_2})}\right| > \varepsilon_n\right) \to 0, \quad \text{(A.8)}$$

where $\varepsilon_n = n^{-\xi_0/2}$. It follows by Theorem 1 that, with probability tending to one, we must have $\max_{k_1 k_2} |\hat{\mathscr{M}}_{k_1 k_2}| \leq m_{\max}$, where $m_{\max}$ is defined in Appendix A. Then, following the similar arguments in Wang (2009), the left side of the above inequality has an upper bound

$$\sum_{k_1,k_2} \sum_{j_1,j_2 \in \hat{\mathscr{M}}_{k_1 k_2}} P\left(|\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}| > \frac{\varepsilon_n}{m_{\max}} \right). \tag{A.9}$$

Next, by Lemma A3 in Bickel and Levina (2008) we have

$$\begin{aligned} P(|\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}| > \varepsilon_n m_{\max}^{-1}) &= E\left\{ \mathbb{P}\left(|\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}| > \varepsilon_n m_{\max}^{-1} \right) \right\} \\ &= E\left\{ \mathbb{P}\left(|\sum_k n_k (\hat{\sigma}_{j_1 j_2, k} - \sigma_{j_1 j_2})| > n \varepsilon_n m_{\max}^{-1} \right) \right\} \\ &\leq E\left\{ \sum_k \mathbb{P}\left(|n_k (\hat{\sigma}_{j_1 j_2, k} - \sigma_{j_1 j_2})| > n_k \varepsilon_n m_{\max}^{-1} \right) \right\} \\ &= E\left\{ \sum_k \mathbb{P}\left(|(\hat{\sigma}_{j_1 j_2, k} - \sigma_{j_1 j_2})| > \varepsilon_n m_{\max}^{-1} \right) \right\} \\ &\leq K N_1 E\left[exp\left\{ -N_2 K^{-1} n^{1-3\xi_0 - 4\kappa_0} (1-R) \right\} \right], \end{aligned}$$

where N_1 and N_2 are some constants. Applying this inequality to the right hand side of (A. 9), it can be further bounded from above by

$$\begin{aligned} &K^3 p^2 N_1 E\left[exp\left\{ -N_2 K^{-1} n^{1-3\xi_0 - 4\kappa_0} (1-R) \right\} \right] \\ &\leq N_0 E\left[exp\left\{ 3logK + 2\nu_1 n^{\xi_1} - N_2 K^{-1} n^{1-3\xi_0 - 4\kappa_0} (1-R) \right\} \right] \to_p 0 \end{aligned} \tag{A.10}$$

as $n \to \infty$, where N_0 and N_3 are some constants. The last convergence result is due to the second conclusion in Theorem 1 and Condition (C3). This proves (A.8) and completes the proof of Lemma 1.

Lemma 2

For any overfitted model $\mathscr{M} \supset \mathscr{M}_{k_1 k_2}$ with arbitrary two classes $k_1 \neq k_2$, we should have $\beta_{k_1 k_2 (\mathscr{M})} = \sum_{(\mathscr{M})}^{-1} \gamma_{k_1 k_2 (\mathscr{M})}$.

Proof

Recall $\mathscr{M}_{k_1 k_2}$ is the true model for class pair (k_1, k_2). We know immediately that $\beta_{(\mathscr{M}^c)} = 0$, where $\mathscr{M}^c = \mathscr{M}_F \backslash \mathscr{M}$ and $\mathscr{M}_F = \{1, 2, \cdots, p\}$ is the full model. Furthermore, we decompose Σ according to $\mathscr{M}$ and $\mathscr{M}^c$ as $\Sigma = (\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22})$. We know immediately that $\sum_{(\mathscr{M})} = \sum_{11}$ and $\sum_{(\mathscr{M}^c)} = \sum_{22}$. Similarly, we write $\gamma_{k_1 k_2} = (\gamma_{k_1 k_2 (\mathscr{M})}^\top, \gamma_{k_1 k_2 (\mathscr{M}^c)}^\top)^\top$. Next, write $\Sigma^{-1} = (\Gamma_{11}, \Gamma_{12}; \Gamma_{21}, \Gamma_{22})$ Since $\beta_{k_1 k_2} = \sum^{-1} \gamma_{k_1 k_2}$, we then have

$$\Gamma_{11}\gamma_{k_1k_2(\mathscr{M})}+\Gamma_{12}\gamma_{k_1k_2(\mathscr{M}^c)}=\beta_{k_1k_2(\mathscr{M})}, \quad \text{(A.11)}$$

$$\Gamma_{12}\gamma_{k_1k_2(\mathscr{M})}+\Gamma_{22}\gamma_{k_1k_2(\mathscr{M}^c)}=\beta_{k_1k_2(\mathscr{M}^c)}=0. \quad \text{(A.12)}$$

Furthermore, we have $\Sigma\Sigma^{-1} = I$, where I stands for an identity matrix with a compatible dimension. We then have

$$\sum_{11}\Gamma_{12}+\sum_{12}\Gamma_{22}=0, \quad \text{(A.13)}$$

$$\sum_{11}\Gamma_{11}+\sum_{12}\Gamma_{12}=I, \quad \text{(A.14)}$$

$$\begin{aligned}\sum_{11}\beta_{k_1k_2(\mathscr{M})} &= \sum_{11}\Gamma_{11}\gamma_{k_1k_2(\mathscr{M})} \\ &+ \sum_{11}\Gamma_{12}\gamma_{k_1k_2(\mathscr{M}^c)} \\ &= \sum_{11}\Gamma_{11}\gamma_{k_1k_2(\mathscr{M})} \\ &- \sum_{12}\Gamma_{22}\gamma_{k_1k_2(\mathscr{M}^c)} \\ &= \sum_{11}\Gamma_{11}\gamma_{k_1k_2(\mathscr{M})} \\ &+ \sum_{12}\Gamma_{12}\gamma_{k_1k_2(\mathscr{M})} \\ &= I_{\gamma_{k_1k_2(\mathscr{M})}}\end{aligned}$$

It is interesting to note that the following equality as , where the above four equations are due to equations (A.11), (A.13), (A.12) and (A.14) respectively. As a result, $\beta_{k_1k_2(\mathscr{M})}=\sum_{11}^{-1}\gamma_{k_1k_2}(\mathscr{M})=\sum_{(\mathscr{M})}^{-1}\gamma_{k_1k_2}(\mathscr{M})$ and this completes the proof.

Now we prove Theorem 2. Let $\hat{\mathscr{M}}^c_{k_1k_2}=\mathscr{M}_F\setminus\hat{\mathscr{M}}_{k_1k_2}=\{j:j\notin\hat{\mathscr{M}}_{k_1k_2}\}$. Then, it follows by definition that the subvector $\hat{\beta}_{k_1k_2(\hat{\mathscr{M}}^c_{k_1k_2})}=0$ Using (2.10), we have that, with probability tending to one, $\hat{\mathscr{M}}_{k_1k_2}\supset\mathscr{M}_{k_1k_2}$ for every $k_1 \neq k_2$. This implies that $\beta_{k_1k_2(\hat{\mathscr{M}}^c_{k_1k_2})}=0$ with probability tending to one. Then by Lemma 2, with probability tending to one, the convergence rate of $\max_{k_1,k_2}\|\beta_{k_1k_2}-\hat{\beta}_{k_1k_2}|$ is identical to that of

$$\begin{aligned}
&\max_{k_1,k_2}\|\hat{\beta}_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}-\beta_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\| \\
=&\max_{k_1,k_2}\|\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\hat{\gamma}_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}-\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\| \\
=&\max_{k_1,k_2}\|\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\left\{\hat{\gamma}_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}-\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\right\}+\left\{\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}| \\
\le&\max_{k_1,k_2}\|\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\left\{\hat{\gamma}_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}-\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\right\}\| \\
&+\max_{k_1,k_2}\|\left\{\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\| \\
\le&\max_{k_1k_2}\lambda_{\max}\left\{\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\max_{k_1,k_2}\|\hat{\gamma}_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}-\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\|
\end{aligned} \tag{A.15}$$

$$+\max_{k_1,k_2}\|\left\{\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\|\max_{k_1,k_2}\|\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\|. \tag{A.16}$$

Recall that $\|A\|$ in (A.16) is the spectral norm of matrix A, defined as $\{\lambda_{\max}(A^{'}A)\}^{1/2}$. Subsequently, we want to demonstrate that both quantities in (A.15) and (A.16) are $o_p(1)$. First, by Lemma 1, we know that $\min_{k_1k_2}\lambda_{\min}((\hat{\sum}_{(\hat{\mathscr{M}}_{k_1k_2})}))>\tau_{\min}/2$ with probability tending to one. As a result, we know that $\max_{k_1k_2}\lambda_{\max}((\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}))\le 2\tau^{-1}_{\min}$ in probability. Furthermore, by the result of (A.4) together with (2.11), we have that $\max_{k_1k_2}\|\hat{\gamma}_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}-\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\|=o_p(1)$, under Conditions to (C3). As a result, the quantity involved in (A.15) is of $o_p(1)$. We thus only need to focus on (A.16). To this end, note that by (A.5) we have

$$\max_{k_1,k_2}\|\gamma_{k_1k_2(\hat{\mathscr{M}}_{k_1k_2})}\|\le\max_{k_1,k_2}\|\gamma_{k_1k_2}\|\le\gamma_{\max}=O(n^{\xi_0/2}). \tag{A.17}$$

Furthermore, $\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}=-\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\{\hat{\sum}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum_{(\hat{\mathscr{M}}_{k_1k_2})}\}\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}$. Then

$$\begin{aligned}
&\max_{k_1,k_2}\|\left\{\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\| \\
\le&\max_{k_1,k_2}\|\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\|\max_{k_1,k_2}\|\left\{\hat{\sum}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\|max\|\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\| \\
\le&\max_{k_1,k_2}\lambda_{\max}\left\{\hat{\sum}^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\max_{k_1,k_2}\lambda_{\max}\left\{\hat{\sum}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}\max_{k_1,k_2}\lambda_{\max}\left\{\sum^{-1}_{(\hat{\mathscr{M}}_{k_1k_2})}\right\} \\
\le&2\tau^{-2}_{\min}\max_{k_1,k_2}\lambda_{\max}\left\{\hat{\sum}_{(\hat{\mathscr{M}}_{k_1k_2})}-\sum_{(\hat{\mathscr{M}}_{k_1k_2})}\right\}=o_p(n^{-\xi_0/2}).
\end{aligned}$$

The last equality is due to Lemma 1. Combine this result with (A.17), we know that the quantity in (A.16) is a $o_p(1)$. This together with (A.15) demonstrates that $\max_{k_1,k_2}\|\beta_{k_1k_2}-\hat{\beta}_{k_1k_2}\|=o_p(1)$, and also completes the proof.

Appendix C. Proof of Theorem 3

We want to prove that $\hat{\mathscr{M}}^{ebic}_{k_1k_2}\supset\mathscr{M}_{k_1k_2}$ uniformly over all the possible class pairs with probability tending to one. Otherwise, there must exit one pair (k_1, k_2), such that $\hat{\mathscr{M}}^{ebic}_{k_1k_2}\not\supset\mathscr{M}_{k_1k_2}$. By conditions (C1)–(C3) and Theorem 1, we know that there exists a model $\hat{\mathscr{M}}_{k_1k_2}\in\mathscr{F}_{k_1k_2}$ such that $\hat{\mathscr{M}}_{k_1k_2}\supset\mathscr{M}_{k_1k_2}$ and $|\hat{\mathscr{M}}_{k_1k_2}|\le C_m n^{\xi_0+2\kappa_0}$, where C_m is some positive constant. Because the models contained in the solution path $\mathscr{F}_{k_1k_2}$ is mutually nested with each other. We then must have $\hat{\mathscr{M}}^{ebic}_{k_1k_2}\subset\hat{\mathscr{M}}_{k_1k_2}$. Moreover, by the definition of $\hat{\mathscr{M}}^{ebic}_{k_1k_2}$, we must have $\mathrm{EBIC}(\hat{\mathscr{M}}^{ebic}_{k_1k_2})<\mathrm{EBIC}(\hat{\mathscr{M}}_{k_1k_2})$. Write $\tilde{\mathscr{M}}_{k_1k_2}=\hat{\mathscr{M}}_{k_1k_2}\setminus\hat{\mathscr{M}}^{ebic}_{k_1k_2}$. We then have $n^{-1}_{k_1k_2}\{\mathrm{EBIC}(\hat{\mathscr{M}}^{ebic}_{k_1k_2})<\mathrm{EBIC}(\hat{\mathscr{M}}_{k_1k_2})\}<0$. As a result

$$\begin{aligned}0>&\min_{k_1k_2}n^{-1}_{k_1k_2}\left\{\mathrm{EBIC}(\hat{\mathscr{M}}^{eibc}_{k_1k_2})-\mathrm{EBIC}(\hat{\mathscr{M}}_{k_1k_2})\right\}\\=&\min_{k_1k_2}\left\{\sum_{j\in\tilde{\mathscr{M}}_{k_1k_2}}log\left(\frac{\tilde{\sigma}_{jj,k_1k_2}}{\hat{\sigma}_{jj,k_1k_2}}\right)-|\tilde{\mathscr{M}}_{k_1k_2}|\left(\frac{logn_{k_1k_2}+2logp}{n_{k_1k_2}}\right)\right\}\\\ge&\min_{k_1k_2}log\left(\frac{\tilde{\sigma}_{j*j*,k_1k_2}}{\hat{\sigma}_{j*j*,k_1k_2}}\right)-\max_{k_1k_2}|\tilde{\mathscr{M}}_{k_1k_2}|\left(\frac{logn_{k_1k_2}+2logp}{n_{k_1k_2}}\right)\end{aligned}\tag{A.18}$$

where $j^*\in\tilde{\mathscr{M}}_{k_1k_2}\cap\mathscr{M}_{k_1k_2}$. Because $\tilde{\mathscr{M}}^{ebic}_{k_1k_2}\supset\tilde{\mathscr{M}}_{k_1k_2}$, $\tilde{\mathscr{M}}^{ebic}_{k_1k_2}\not\supset\mathscr{M}_{k_1k_2}$ and $\hat{\mathscr{M}}_{k_1k_2}\supset\mathscr{M}_{k_1k_2}$, we know j^* exists and is well defined. The above inequality is also due to the fact that $\tilde{\sigma}_{jj,k_1k_2}\ge\hat{\sigma}_{jj,k_1k_2}$ by definition. Then further lower bound the right hand side of the above quantity by

$$\begin{aligned}\ge&\min_{k_1,k_2}\min_{j\in\mathscr{M}_{k_1k_2}}log\left(\frac{\tilde{\sigma}_{jj,k_1k_2}}{\hat{\sigma}_{jj,k_1k_2}}\right)-\max_{k_1,k_2}|\widetilde{\mathscr{M}}_{k_1k_2}|\left(\frac{logn_{k_1k_2}+2logp}{n_{k_1k_2}}\right)\\=&\min_{k_1,k_2}\min_{j\in\mathscr{M}_{k_1k_2}}log\left(\frac{\tilde{\sigma}_{jj,k_1k_2}}{\hat{\sigma}_{jj,k_1k_2}}\right)-O(n^{2\kappa_0+\xi_0+\xi_1-1}),\end{aligned}\tag{A.19}$$

because max $\max_{k_1k_2}|\hat{\mathscr{M}}_{k_1k_2}|\le C_m n^{\xi_0+2\kappa_0}$, Condition (C1) and the fact that $n_{k_1k_2}=O(n/K)$. By the definition of $\tilde{\sigma}_{jj,k_1k_2}$ and $\hat{\sigma}_{jj,k_1k_2}$, one can verify that

$$\tilde{\sigma}_{jj,k_1k_2}=\hat{\sigma}_{jj,k_1k_2}+\frac{n_{k_1}n_{k_2}}{(n_{k_1}+n_{k_2})^2}(\hat{\mu}_{k_1j}-\hat{\mu}_{k_2j})^2.$$

Following similar technique in the proof of Theorem 1, one can easily show that $\max_{k_1,k_2,j}|\hat{\sigma}_{jj,k_1k_2} - \sigma_{jj,k_1k_2}| = o_p(1)$ and $\max_{k_1,k_2}|n_{k_1}n_{k_2}(n_{k_1}+n_{k_2})^{-2} - 1/2| = o_p(1)$. All these suggest that the right hand side of (A.19) can be further lower bounded by

$$\geq C^* \min_{k_1,k_2} \min_{j \in \mathscr{M}_{k_1k_2}} (\hat{\mu}_{k_1j} - \hat{\mu}_{k_2j})^2 - O(n^{2\kappa_0+\xi_0+\xi_1-1}) \geq 2^{-1}C^*\gamma_{\min}^2 - O(n^{2\kappa_0+\xi_0+\xi_1-1}), \quad \text{(A.20)}$$

due to the fact that $\hat{\mu}_{kj}$ is uniformly consistent for μ_{kj}; see the proof of Theorem 1. Here C^* is some positive constant. By condition (C3), we know that the first term in (A.20) is a positive quantity with order $n^{-2\kappa_0}$, which dominates the second term in (A.20), that is $O(n^{2\kappa_0+\xi_0+\xi_1-1})$. Once again this is due to the condition (C3). We know then that the right hand side of (A.20) must be positive with probability tending to one. This suggests that (A. 18) should happen with probability tending to zero. As a result, we must have $\hat{\mathscr{M}}_{k_1k_2}^{ebic} \supset \mathscr{M}_{k_1k_2}$ uniformly over all class pairs with probability tending to one. This proves the theorem conclusion.

References

Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, BN., Csaki, F., editors. 2nd International Symposium on Information Theory. Budapest: Akademia Kiado; 1973. p. 267-281.

Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, "naïve Bayes", and some alternatives when there are many more variables than observations. Bernoulli. 2004; 10:989–1010.

Bickel PJ, Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics. 2008; 36:199–227.

Cai TT, Liu W. A direct estimation approach to sparse linear discriminant analysis. Journal of the American Statistical Association. 2011; 106:1566–1577.

Chen J, Chen Z. Extended Bayesian information criterion for model selection with large model spaces. Biometrika. 2008; 95:759–771.

Clemmensen L, Hastie T, Ersbøll. Sparse discriminant analysis. Technometrics. 2011; 53(4):406–413.

Fan J, Fan Y. High dimensional classification using features annealed independence rules. The Annals of Statistics. 2008; 36:2605–2637. [PubMed: 19169416]

Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high dimensional additive models. Journal of the American Statistical Association. 2011; 116:544–557.

Fan J, Feng Y, Tong X. A road to classification in high dimensional space: the regularized optimal affine discriminant. Journal of the Royal Statistical Society: Series B. 2012; 74(4):745–771.

Fan J, Lv J. Sure independence screening for ultra-high dimensional feature space (with discussion). Journal of the Royal Statistical Society, Series B. 2008; 70:849–911.

Fan J, Song R. Sure independent screening in generalized linear models with NP-dimensionality. The Annals of Statistics. 2010; 38:3567–3604.

Fang KT, Li R, Liang J. A multivariate version of Ghosh's MT_3 plot to detect non-multinormality. Computational Statistics and Data Analysis. 1998; 28:371–386.

Guo Y, Hastie T, Tibshirani R. Regularized discriminant analysis and its application in microarrays. Biostatistics. 2007; 1:86–100.

Liang JJ, Li R, Fang KT, Fang HB. Testing multinormality based on low-dimensional projection. Journal of Statistical Planning and Inference. 2000; 86:129–141.

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

Mardia KV. Measures of multivariate skewness and kurtosis with applications. Biometrika. 1970; 69:519–530.

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6:461–464.

Shao J, Wang Y, Deng X, Wang S. Sparse linear disciminant analysis by thresholding for high dimensional data. The Annals of Statistics. 2011; 39:1241–1265.

Tibshirani R, Hastie T, Narashimhan B, Chu G. Class prediction by nearest shrunken centriods with applications to DNA microarrays. Statistical Science. 2003; 18:104–117.

Wang H. Forward regression for ultra-high dimensional variable screening. Journal of the American Statistical Association. 2009; 104:1512–1524.

Wang H. Factor profiled independence screening. Biometrika. 2012; 99:15–28.

Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007; 94:553–568. [PubMed: 19343105]

Weiss, SM., Indurkhya, N., Zhang, T., Damerau, FJ. Text Mining: Predictive Methods for Analyzing Unstructured Information. New York: Springer; 2005.

Witten DM, Tibshirani R. Penalized classification using Fisher's linear discriminant. Journal of the Royal Statistical Society: Series B. 2011; 73(5):753–772.

Zhu LP, Li L, Li R, Zhu LX. Model-free feature screening for ultrahigh dimensional data. Journal of the American Statistical Association. 2011; 106:1464–1475. [PubMed: 22754050]

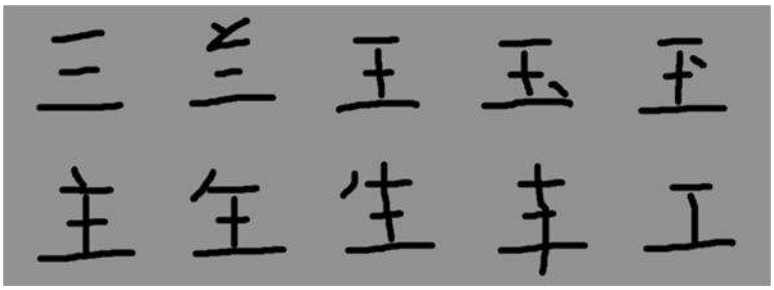

Figure 1.
Handwritten Version of the Ten Chinese Characters

Table 1

Simulation Results with 1,000 Replications

(n,K)	MS	MMS	CZ(%)	IZ(%)	CP(%)	UCP(%)	RSSE	MRSSE	$\hat{k}$(%)	k^*(%)	RCA(%)	CPU
						Model 1						
(100,10)	2.02	2.89	100.00	1.15	97.70	52.50	1.82	5.17	99.75	99.85	99.90	1.73
(400,20)	2.03	3.25	100.00	0.00	100.00	100.00	0.85	2.67	99.71	99.69	100.02	13.25
(1600,40)	2.02	3.36	100.00	0.00	100.00	100.00	0.46	1.67	99.54	99.42	100.12	162.99
						Model 2						
(100,10)	2.49	5.69	99.99	10.42	79.50	0.00	2.73	6.22	87.68	98.67	88.86	1.73
(400,20)	5.22	8.87	99.97	0.33	99.35	28.20	1.46	4.92	97.56	98.48	99.06	17.26
(1600,40)	7.83	11.76	99.94	0.00	100.00	100.00	1.01	2.27	98.44	98.23	100.22	174.78
						Model 3						
(100,10)	1.77	2.45	100.00	12.34	75.32	0.10	5.06	11.93	95.33	99.84	95.49	1.73
(400,20)	2.01	3.25	100.00	0.02	99.95	94.80	3.57	6.01	99.70	99.68	100.01	16.21
(1600,40)	2.01	5.04	100.00	0.00	100.00	100.00	3.87	5.31	99.53	99.42	100.11	171.13
						Model 4						
(100,10)	1.96	2.33	100.00	2.71	94.64	12.40	2.79	7.12	96.06	97.06	98.98	1.76
(400,20)	2.00	2.56	100.00	0.01	99.99	97.30	1.17	3.32	94.03	93.85	100.19	16.21
(1600,40)	2.00	2.98	100.00	0.00	100.00	100.00	0.60	1.83	89.46	87.91	101.76	171.66
						Model 5						
(100,10)	1.99	2.89	100.00	2.34	95.33	27.70	1.93	5.39	99.69	99.85	99.84	1.79
(400,20)	2.03	3.25	100.00	0.00	100.00	99.70	0.87	2.78	99.71	99.68	100.03	16.47
(1600,40)	2.03	3.38	100.00	0.00	100.00	100.00	0.48	1.80	99.48	99.41	100.07	166.92

MS: model size; MMS: maximum model size; CZ: % of correct zeros; IZ: % of incorrect zeros

CP: coverage probability; UCP: uniform coverage probability; RSSE: root of the sum squared error; MRSSE: maximum RSSE

J Am Stat Assoc. Author manuscript; available in PMC 2017 May 05.

Table 2

Different Classification Methods with $(n, K) = (400, 10)$ and $p = 2000$.

Example	SVM	NNet	PLDA	PSIS
1	11.75	14.43	99.64	99.81
2	12.61	15.12	80.10	98.71
3	11.29	13.75	41.03	99.86
4	12.06	14.75	96.44	97.04
5	19.97	22.23	99.70	99.83

Table 3

Detailed Results for Chinese characters data set

Method	Classification Accuracy (%)	Total Number of Selected Features	Average Model Size
PSIS	93.86	60.69	15.61
PLDA	87.68	173.41	–
NNet	85.41	–	–
SVM	78.40	–	–