Statistica Sinica Preprint No: SS-2021-0047	
Title	An Efficient Convex Formulation for Reduced-Rank Linear Discriminant Analysis in High Dimensions
Manuscript ID	SS-2021-0047
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0047
Complete List of Authors	Jing Zeng, Xin Zhang and Qing Mai
Corresponding Author	Qing Mai
E-mail	qmai@fsu.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

An Efficient Convex Formulation for Reduced-rank Linear Discriminant Analysis in High Dimensions

Jing Zeng, Xin Zhang and Qing Mai

Florida State University

Abstract: In this paper, we propose a parsimonious reduced-rank linear discriminant analysis model for high-dimensional sparse multi-class discriminant analysis. A sparse dimension reduction subspace is constructed to contain all the necessary information for linear discriminant analysis. We show explicitly the connections between our model and two well-studied models in the literature: the principal fitted component model in sufficient dimension reduction and the multivariate reduced-rank regression model. The likelihood-inspired efficient estimator is then recast from a convex optimization perspective. A doubly penalized convex optimization is proposed to unite sparsity and low-rankness in high dimensions, and is then solved efficiently by a three-operator splitting algorithm. We establish the rank selection consistency and the classification error consistency of proposed method when the number of variable grows very fast with the sample size. The effectiveness of the proposed method is further demonstrated by extensive simulation studies and facial recognition data sets.

Key words and phrases: Dimension reduction, discriminant analysis, linear discriminant analysis, nuclear norm penalty, variable selection.

1. Introduction

High-dimensional linear discriminant analysis (LDA) methods are widely studied and applied (e.g.,Bickel and Levina 2004; Cai and Liu 2011, Shao et al. 2011; Mai et al. 2012) We consider multi-category classification with $K \geq 2$ classes, where linear discriminant analysis can identify at most $K-1$ linearly independent discriminant directions. When the dimension of the subspace spanned by all discriminant directions is less than $K-1$, this is known as the reduced-rank LDA problem (Hastie et al. 2009, Chapter 4.3.3). There are two popular approaches to this problem. The first approach includes methods such as penalized linear discriminant analysis (Witten and Tibshirani 2011) and sparse optimal scoring (Clemmensen et al. 2011). These methods are high-dimensional extensions of Fisher's view of LDA and optimal scoring formulation of LDA. Specifically, such methods implicitly handle the low-rankness by sequential estimation of sparse discriminant directions. The second class of methods, such as Hao et al. (2015) and Niu et al. (2018), rely on principal component analysis. The low-rankness is achieved by selecting the first several principal directions as the discriminant directions (Niu et al. 2018) or by rotation of the data (Hao et al. 2015). However, these methods do not impose sparsity on the original predictors. In addition to these statistical approaches, reduced-rank LDA

methods and algorithms are gaining substantial attention in engineering applications (e.g., Ye and Li 2005), where a probabilistic explanation is greatly desirable.

In this paper, we first introduce a model-based interpretation for the reduced-rank LDA problem. The low-rankness is formally stated as a unique low-dimensional subspace, whose maximum likelihood estimator motivates our re-parameterization of the target parameters, leading to the efficient convex formulation. We then solve a penalized quadratic convex optimization by a three-splitter operator algorithm, which is guaranteed to reach the global minimum. To gain further insights on reduced-rank discriminant analysis, we discuss how low-rankness arises naturally in the settings of ordinal classification (McCullagh 1980, da Costa et al. 2008, 2010, Qiao 2015) and response category combination (Price et al. 2019, Wen and Koppelman 2001).

The model-based interpretation and the maximum likelihood estimator of the low-dimensional subspace are connected to the principal fitted components model (Cook and Forzani 2008) in sufficient dimension reduction and the reduced-rank regression (Anderson 1951, Izenman 1975, Stoica and Viberg 1996) in multivariate linear model. By exploiting such a connection, we can easily derive the maximum likelihood estimator of the

low-dimensional subspace under the LDA model when the dimension of predictor p is smaller than the sample size n. Given the true rank d, the maximum likelihood estimator is obtained from the first d eigenvectors of a symmetric $p \times p$ matrix with rank at most $(K-1)$. Based on such an observation, we augment the low-dimensional subspace parameter into an overparameterized and rank-deficient matrix of dimension $p \times K$. Without pre-specifying the rank, we estimate this rank deficient matrix parameter in high dimensions via nuclear norm penalization.

Convex formulation and convex relaxation of classical multivariate analysis and dimension reduction methods prevail in high-dimensional settings. Our approach is very different from the convex relaxation of sparse principal component analysis (Vu et al. 2013), sparse canonical correlation analysis (Gao et al. 2017), or sparse sliced inverse regression (Tan et al. 2018; Tan et al. 2020). In these convex relaxation approaches, the rank or dimensionality is pre-specified and incorporated into the constraints of optimization. Then the optimization is over $p \times p$ symmetric matrices subject to constraints (e.g. the parameter space of optimization would include projection matrices onto d-dimensional subspaces). Unlike these approaches that augment the d-dimensional subspace as $p \times p$ dimensional matrices, our approach is much more direct. Instead of optimizing over subspaces,

orthogonal basis matrices, or projection matrices, we directly optimize over an unconstrained $p \times K$ dimensional matrix parameter. This leads to much cheaper computation that scales better with large p.

Our approach is also an extension of the multi-class sparse discriminant analysis method by Mai et al. (2019), which does not account for the potential low-rankness and is thus less effective when the number of classes is big. Importantly, although our quadratic objective function is similar to the one in Mai et al. (2019), the new maximum likelihood and least squares estimation naturally leads to different weights for different discriminant direction that is not accounted for in Mai et al. (2019). Moreover, the doubly penalized estimation in our model is more challenging and requires a new algorithm. Our unified approach of deriving the quadratic objective function also extends the scope of multi-class sparse discriminant analysis from the one-versus-all parameterization to one-versus-one parameterization.

We adopt the following notations throughout the paper. For a vector $v = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, we define the L_q-norm as $\|v\|_q = (\sum_{j=1}^p v_j^q)^{1/q}$ for $1 \leq q < \infty$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$, let $\sigma_1 \geq \ldots \geq \sigma_{\min\{p,q\}}$ denote its singular values, define the $L_{2,1}$ norm and the nuclear norm as $\|A\|_{2,1} = \sum_{i=1}^p (\sum_{j=1}^q a_{ij}^2)^{1/2}$ and $\|A\|_\star = \sum_{i=1}^{\min\{p,q\}} \sigma_i$ respectively. The span of A, denoted as span(A) or $\mathcal{S}_A$, is the subspace spanned by the column

vectors of A. Let $\beta \in \mathbb{R}^{p\times r}$ be the orthonormal basis of the subspace $\mathcal{S} \subseteq \mathbb{R}^p$, i.e., $\beta^\top\beta = I_r$, we use $P_{\mathcal{S}} \equiv P_\beta = \beta\beta^\top$ to denote the projection matrix onto the subspace $\mathcal{S}$.

2. Reduced-rank linear discriminant analysis

2.1 Model-based interpretation

We consider the multi-class classification problem for the response $Y \in \{1, \ldots, K\}$ and the predictor $X \in \mathbb{R}^p$. In linear discriminant analysis, within each class k, the predictor is assumed to have mean $\mu_k \in \mathbb{R}^p$ and the common non-singular covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. Let $\pi_k = \Pr(Y = k)$ and $\mu \equiv E(X) = \sum_{k=1}^K \pi_k\mu_k$. The Bayes' rule, $\phi(X) : \mathbb{R}^p \mapsto \{1, \ldots, K\}$, is the optimal classification rule in population and has the following form if we assume that $X \mid Y$ is normally distributed:

$$\phi(X) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \left\{ \left(X - \frac{\mu_k + \mu}{2} \right)^\top \Sigma^{-1}(\mu_k - \mu) + \log \pi_k \right\}. \tag{2.1}$$

From (2.1), it is clear that the K directions $\Sigma^{-1}(\mu_k - \mu)$, $k = 1, \ldots, K$, preserve all the information of X relevant to classification. These K directions are not linearly independent because $\sum_k \pi_k(\mu_k - \mu) = 0$. In this paper, we explicitly state the low-rankness condition as follows.

Low-rankness condition. Let $\mathcal{S} \subseteq \mathbb{R}^p$ be the subspace spanned by the

K discriminant directions $\Sigma^{-1}(\mu_k - \mu)$, $k = 1, \ldots, K$, then its dimension $\dim(\mathcal{S}) = d < K - 1$.

The reduced-rank linear discriminant analysis model is then formally presented as,

$$\begin{aligned} &\Pr(Y = k) = \pi_k > 0, \quad X \mid (Y = k) \sim N(\mu_k, \Sigma), \\ &\mu_k = \mu + \Sigma\beta\eta_k, \quad k = 1, \ldots, K, \end{aligned} \tag{2.2}$$

where $\beta \in \mathbb{R}^{p\times d}$ is a basis matrix of the subspace $\mathcal{S}$ in the low-rankness condition, i.e., $\mathcal{S} = \mathcal{S}_\beta$, and $\eta = (\eta_1, \ldots, \eta_K) \in \mathbb{R}^{d\times K}$ is the corresponding coordinates of the K discriminant directions $\Sigma^{-1}(\mu_k - \mu)$.

Under (2.2), the Bayes' rule becomes

$$\phi(X) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \left\{ \left(X - \frac{\mu_k + \mu}{2} \right)^\top \beta\eta_k + \log \pi_k \right\}, \tag{2.3}$$

which implies that given any observation $x \in \mathbb{R}^p$, $\Pr(Y = k \mid X = x) = \Pr(Y = k \mid \beta^\top X = \beta^\top x)$ for $k = 1, \ldots, K$. In other words, the reduction of data from $X \in \mathbb{R}^p$ to $\beta^\top X \in \mathbb{R}^d$ is without any loss of relevant information for classification under model (2.2). If β is known, we can then replace X with $\beta^\top X$ and apply the classical linear discriminant analysis.

Remark 1. The parameters β and η are not identifiable since the decomposition $\beta\eta$ can be replaced by $\widetilde{\beta}\widetilde{\eta}$, where $\widetilde{\beta} = \beta O$ and $\widetilde{\eta} = O^\top\eta$ for any orthogonal matrix $O \in \mathbb{R}^{d\times d}$. Nevertheless, the subspace $\mathcal{S} = \operatorname{span}(\beta)$ is

identifiable and is the key parameter of interest in model (2.2). In this paper, the subspace $\mathcal{S}$ is called the *discriminant subspace*, and its basis β is called the *discriminant basis*. The dimensionality $\dim(\mathcal{S}) = d$ is called the *discriminant rank*. Any vector in $\mathcal{S}$ is called a *discriminant direction*.

The reduced-rank LDA model is closely connected to the principal fitted component model (Cook and Forzani 2008) in sufficient dimension reduction and the multivariate reduced-rank regression (Izenman 1975). To see this, we rewrite model (2.2) as the following equivalent form,

$$X = \mu + \Sigma\beta\eta\xi_Y + \varepsilon, \quad \varepsilon \sim N(0, \Sigma), \tag{2.4}$$

where $\xi_Y \in \mathbb{R}^K$ is the indicator functions of Y: If $Y = k$, then the kth element of ξ_Y is one and all the other elements are zero. There is also an intrinsic constraint that $\Sigma\beta\eta\mathrm{E}(\xi_Y) = 0$ in (2.4). This model is exactly the principal fitted component model, when the fitting functions are chosen as the indicator functions of Y. Hence, our discriminant subspace $\mathcal{S}$ is also the central subspace in sufficient dimension reduction (Cook 1998). If we treat X as response and ξ_Y as predictor, then (2.4) becomes the multivariate reduced-rank regression model (Izenman 1975) and $\Sigma\beta\eta \in \mathbb{R}^{p\times K}$ is the rank-d regression coefficient matrix. Such connections enable us to easily obtain the maximum likelihood estimator for model (2.2), and further motivates our efficient convex formulation.

2.2 Efficient convex formulation for high-dimensional estimation

As discussed in Remark 1, the discriminant basis β is not identifiable but the discriminant subspace S is identifiable. However, optimization over subspace is non-convex and expensive in general. To facilitate high-dimensional computation, we introduce an alternative target object $B \in \mathbb{R}^{p\times K}$ that is identifiable and replaces β and $\mathcal{S}$ in high-dimensional estimation.

We first consider the maximum likelihood estimator of $\mathcal{S}$, which is summarized in the following lemma. Let $\widehat{\Sigma} = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n} I(Y_i = k)(X_i - \bar{X}_k)(X_i - \bar{X}_k)^\top$ denote the within-class covariance matrix, where $I(Y_i = k)$ takes value 1 if $Y_i = k$ and 0 otherwise, and $\widehat{\Sigma}_b = \sum_{k=1}^{K}(n_k/n)(\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^\top$ denote the between-class covariance matrix, where $\bar{X}_k$ is the sample mean of X in class k, $\bar{X}$ is the sample mean of X, n is the overall sample size and n_k is the sample size for class k.

Lemma 1. *Under model* (2.2)*, the maximum likelihood estimator of* $\mathcal{S} = \text{span}(\beta)$ *is* $\widehat{\Sigma}^{-1/2}\text{span}(\widehat{v}_1, \ldots, \widehat{v}_d)$*, where* $\widehat{v}_i$ *is the* i*th eigenvector of* $\widehat{\Sigma}^{-1/2}\widehat{\Sigma}_b\widehat{\Sigma}^{-1/2}$.

Based on (2.4), we can easily verify the results of Lemma 1 from previous works (Cook and Forzani 2008, Stoica and Viberg 1996). Lemma 1 provides solutions to low-dimensional reduced-rank LDA problem.

Let $\widehat{U}, U \in \mathbb{R}^{p\times K}$, $U = \{\pi_1^{1/2}(\mu_1 - \mu), \ldots, \pi_K^{1/2}(\mu_K - \mu)\}$ and $\widehat{U}$ is

its sample estimator. Then $\widehat{\Sigma}_b$ can be rewritten as $\widehat{\Sigma}_b = \widehat{U}\widehat{U}^\top$; and the maximum likelihood estimator $\widehat{\mathcal{S}} = \widehat{\Sigma}^{-1/2}\text{span}(\widehat{v}_1, \ldots, \widehat{v}_d) \subseteq \text{span}(\widehat{\Sigma}^{-1}\widehat{U})$, while in the population $B \equiv \Sigma^{-1}U$ spans the same subspace as $\mathcal{S}$. As such, we target on B for the estimation of subspace $\mathcal{S}$. Since $\text{rank}(B) = d \leq K-1$, B is overparameterized and is to be estimated with rank regularization.

Following model (2.2), we can write $BW = \beta\eta$, where the matrix $W = \text{diag}(\pi_1^{-1/2}, \ldots, \pi_K^{-1/2}) \in \mathbb{R}^{K\times K}$. Consequently, the inverse regression model (2.4) can be rewritten in terms of B as follows,

$$X = \mu + \Sigma BW\xi_Y + \varepsilon, \quad \varepsilon \sim N(0, \Sigma), \tag{2.5}$$

where $BW = \beta\eta$. To avoid ambiguity of the reference to β due to its non-identifiability, we henceforth refer to β as the matrix composed of top-d left singular vectors of B.

Inspired by the inverse regression reformulation (2.5) of the reduced-rank LDA model, a natural way to estimate B is the least squares estimation by solving the following least squares problem,

$$\underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \sum_{i=1}^{n} \|(X_i - \bar{X}) - \widehat{\Sigma}B\widehat{W}\xi_{Y_i}\|_2^2, \tag{2.6}$$

where μ, W and Σ in the inverse regression model (2.5) are replaced by their sample estimators. Again, since B is identifiable, itself is the target of estimation and the rank constraint on B in (2.5) is yet to be imposed in

the least squares formulation (2.6). An equivalent form of (2.6) is given in the following Lemma.

Lemma 2. *Assume that $\widehat{\Sigma}$ is non-singular, the least squares problem in* (2.6) *is equivalent to*

$$\underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\operatorname{tr}(B^\top\widehat{\Sigma}B) - \operatorname{tr}(B^\top\widehat{U}). \tag{2.7}$$

Based on Lemma 2, the least squares estimator of B is $\widehat{\Sigma}^{-1}\widehat{U}$, which is exactly the plug-in estimator of B defined previously. In high dimensions where $p \gg n$, $\widehat{\Sigma}$ is no longer invertible and the least squares estimator is not well-defined. However, the convex formulation (2.7), to be combined with penalization techniques, will provide a new way for estimating the discriminant subspace in high-dimensional setting.

Remark 2. Our convex formulation of (2.7) is similar to the optimization in Mai et al. (2019), but is motivated from an efficient likelihood-based perspective. If we replace $U = \{\pi_1^{1/2}(\mu_1-\mu), \ldots, \pi_K^{1/2}(\mu_K-\mu)\} \in \mathbb{R}^{p\times K}$ with an unweighted one-versus-others version $\{(\mu_2-\mu_1), \ldots, (\mu_K-\mu_1)\} \in \mathbb{R}^{p\times(K-1)}$, then (2.7) reproduces the objective function in Mai et al. (2019), which lacks likelihood or least squares interpretation. Moreover, because of our rank regularization introduced later, our method allows more flexible modifications than Mai et al. (2019). For example, we can also use one-versus-one

parameterization to replace $\widehat{U}$ with the $p\times K(K-1)/2$ dimensional pairwise mean difference matrix.

In high-dimensional statistics, the sparsity assumption is commonly imposed such that only a small number of variables are active in the model. Based on Bayes' rule (2.1), the j-th variable X_j makes no contribution to the classification if and only if $b_{j1} = \cdots = b_{jK} = 0$, where b_{jk} is the (j,k)-th element in matrix B. Let $\mathcal{A}$ denote the index set of all the active variables, then $\mathcal{A} = \{j \mid \text{there exists } k \text{ such that } b_{jk} \neq 0\}$, and the sparsity level is denoted as $s = |\mathcal{A}|$.

For simultaneous variable selection and rank shrinkage, we propose the following doubly penalized convex optimization,

$$\widehat{B} = \underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\operatorname{tr}(B^\top\widehat{\Sigma}B) - \operatorname{tr}(B^\top\widehat{U}) + \lambda_1||B||_{2,1} + \lambda_2||B||_\star, \tag{2.8}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters. The $L_{2,1}$ norm penalty $\|B\|_{2,1}$ (Yuan and Lin 2006) and the nuclear norm penalty $||B||_\star$ have been ubiquitously applied in many regularized regression or classification problems (see Roth and Fischer 2008; Meier, Van De Geer and Bühlmann 2008; Yuan et al. 2007; Zhou and Li 2014). After we obtain $\widehat{B}$ from (2.8), the estimated discriminant rank $\widehat{d}$ directly follows from Algorithm 1 to be introduced in the next section. Then by singular value decomposition of $\widehat{B}$,

the discriminant basis estimator is defined as $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{\widehat{d}})$, where $\widehat{\beta}_k$ is the left singular vector of $\widehat{B}$ corresponding to the kth largest singular value. And the active set can be estimated as $\widehat{\mathcal{A}} = \{j \mid \text{there exists } k \text{ such that } \widehat{b}_{jk} \neq 0\}$. Once the estimated discriminant basis $\widehat{\beta}$ is obtained, the classification is performed on the reduced $\widehat{d}$-dimensional data $\widehat{\beta}^\top X$, i.e. (2.3).

2.3 The algorithm

One common way to solve the doubly penalized convex optimization problem (2.8) is to impose an equality constraint and implement the alternating direction method of multipliers algorithm (see, Boyd, Parikh and Chu 2011) by iteratively solving two simpler convex optimization problems, each with only one penalty term. However, such an algorithm introduces an augmented term from the equality constraint and an extra tuning parameter is involved, which makes the tuning procedure more tricky. Instead, we adopt a simpler and more efficient three-operator splitting scheme recently proposed by Davis and Yin (2017). In its application to problem (2.8), the three operators are $\widehat{\Sigma}B - \widehat{U}$, $\lambda_1 \partial \|B\|_{2,1}$ and $\lambda_2 \partial \|B\|_\star$ respectively, where ∂ denotes the subdifferentials. From the implementation of the three-operator splitting algorithm, it can be seen that the algorithm introduces no additional tuning parameters, has easy-to-implement iteration, and is more efficient

than our alternating direction method of multipliers algorithm that we also provided and compared to in the Supplementary Materials.

Following Davis and Yin (2017), the iteration of solving (2.8) is implemented as follows:

(1) Proximal mapping of $L_{2,1}$ norm:

$$B^{(t)} = \underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\|B - A^{(t)}\|_F^2 + \gamma\lambda_1\|B\|_{2,1} \tag{2.9}$$

(2) Proximal mapping of nuclear norm:

$$C^{(t)} = \underset{C\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\|C - \{2B^{(t)} - A^{(t)} - \gamma(\widehat{\Sigma}B^{(t)} - \widehat{U})\}\|_F^2 + \gamma\lambda_2\|C\|_\star. \tag{2.10}$$

(3) Update $A^{(t+1)}$: $A^{(t+1)} = A^{(t)} + \alpha_t(C^{(t)} - B^{(t)})$.

As suggested in Davis and Yin (2017), for simplicity, we fix the constant $\alpha_t = 1$ for $t \geq 0$, and $\gamma = 1.99/\lambda_{\max}(\widehat{\Sigma})$, where $\lambda_{\max}(\widehat{\Sigma})$ is the largest eigenvalue of $\widehat{\Sigma}$. Interested readers are referred to Davis and Yin (2017) for more details on these constants. The updates of $B^{(t)}$ and $C^{(t)}$ in (2.9) and (2.10) are simply the proximal mapping of $L_{2,1}$-norm and nuclear norm, whose solutions are commonly known in many penalization problems (Mai, Yang and Zou 2019, Zhou and Li 2014). We summarize the explicit forms of $B^{(t)}$ and $C^{(t)}$ in the following lemma. Define the positive part function $x_+ = \max\{0, x\}$ for any $x \in \mathbb{R}$.

Lemma 3. *Let* $(a_i^{(t)})^\top$ *denote the* i*th row vector of* $A^{(t)}$*, then the solution* $B^{(t)}$ *in* (2.9) *is* $((b_1^{(t)})^\top, \ldots, (b_p^{(t)})^\top)^\top$ *where* $b_i^{(t)} = a_i^{(t)}(1 - \gamma\lambda_1/\|a_i^{(t)}\|_2)_+$*,* $i = 1, \ldots, p$*. Let* $M^{(t)}$ *denote* $2B^{(t)} - A^{(t)} - \gamma(\widehat{\Sigma}B^{(t)} - \widehat{U})$*, and* $\sum_{i=1}^{\min\{p,K\}} \sigma_i u_i v_i^\top$ *denote the singular value decomposition of* $M^{(t)}$*, the solution* $C^{(t)}$ *in* (2.10) *is* $C^{(t)} = \sum_{i=1}^{\min\{p,K\}} (\sigma_i - \gamma\lambda_2)_+ u_i v_i^\top$.

After enough iterations, the sequences $(B^{(t)})_{t\geq 0}$ and $(C^{(t)})_{t\geq 0}$ converge weakly to the stationary point of the objective function (Davis and Yin 2017). In our problem (2.8), which is convex, the stationary point is hence the global minimizer. Specifically, we have the following results.

Lemma 4. *For problem* (2.8)*, by fixing* $\gamma < 2/\lambda_{\max}(\widehat{\Sigma})$ *and* $\alpha_t = 1$ *for* $t \geq 0$*, as* $t \to \infty$*, both* $(B^{(t)})_{t\geq 0}$ *and* $(C^{(t)})_{t\geq 0}$ *converge weakly to the global minimizer of problem* (2.8).

We summarize our estimation procedure in Algorithm 1. The algorithm requires the input of the sample matrices $\widehat{\Sigma}$ and $\widehat{U}$, the tuning parameters λ_1 and λ_2 and the thresholding value δ. The thresholding value δ is used in rank selection, which is set as 10^{-3} by default. Then we initialize the matrix $A^{(0)} = 0$ and update $B^{(t)}$, $C^{(t)}$ and $A^{(t)}$ iteratively, which can be solved efficiently by Lemma 3. The update of $B^{(t)}$ in (2.9) introduces the group-structure sparsity and the update of $C^{(t)}$ in (2.10) introduces the low-rank structure. In iterations, we use the relative change

Algorithm 1 LSLDA Algorithm

Input: $\widehat{\Sigma}$, $\widehat{U}$, the tuning parameters λ_1, λ_2 and the thresholding value δ.

Initialization: $A^{(0)} = 0$.

repeat

Step 1: Update $B^{(t)}$: the ith row vector of $B^{(t)}$ is $(b_i^{(t)})^\top = (a_i^{(t)})^\top(1 - \gamma\lambda_1/\|a_i^{(t)}\|_2)_+$, where $(a_i^{(t)})^\top$ is the ith row vector of $A^{(t)}$.

Step 2: Update $C^{(t)}$: calculate $M^{(t)} = 2B^{(t)} - A^{(t)} - \gamma(\widehat{\Sigma}B^{(t)} - \widehat{U})$, then $C^{(t)} = \sum_{i=1}^{\min\{p,K\}}(\sigma_i - \gamma\lambda_2)_+ u_i v_i^\top$, where σ_i, u_i and v_i are defined in Lemma 3.

Step 3: Update $A^{(t+1)}$: $A^{(t+1)} = A^{(t)} + \alpha_t(C^{(t)} - B^{(t)})$.

until some stopping criterion is met.

Output: Let $\widehat{B}$ be the solution at termination. The discriminant rank is estimated by $\widehat{d} = \sum_{i=1}^{\min\{p,K\}} I(\sigma_i(\widehat{B}) \geq \delta)$, where $\sigma_i(\widehat{B})$ is the ith singular value of $\widehat{B}$. Let $\widehat{\beta}_k$ be the left singular vector of $\widehat{B}$ corresponding to the kth largest singular value. The estimated discriminant basis $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{\widehat{d}})$.

$\|B^{(t)} - C^{(t)}\|_F/(1 + \|A^{(t+1)}\|_F) \leq \delta$ as the convergence criterion, where the tolerance is set as δ, the same as the thresholding value. We count the number of non-zero singular values of $\widehat{B}$ after thresholding with value δ as the estimated rank. Finally, the top-$\widehat{d}$ left singular vectors of $\widehat{B}$ is returned as the discriminant basis estimator $\widehat{\beta}$. We select the tuning parameters λ_1 and λ_2 by cross-validation. More details of the tuning procedure is contained in the Supplementary Materials.

3. Other applications of reduced-rank LDA model

When the number of classes K is large, low-rankness can be a useful approximation. The low-rankness condition may also appear naturally in other situations such as ordinal response and indistinguishable classification.

The first application is the ordinal classification. Many works have been devoted to solving the ordinal classification problems with accounting for the order relation.In particular, the unimodality condition in the following is well-justified in ordinal response (e.g., da Costa et al. 2008, 2010).

Unimodality condition. The ordinal response $Y \in \{1, \ldots, K\}$. For any $x \in \mathbb{R}^p$, $\Pr(Y = k \mid X = x) > \Pr(Y = k+1 \mid X = x)$ for $k \geq \text{mode}(Y)$ and $\Pr(Y = k \mid X = x) > \Pr(Y = k-1 \mid X = x)$ for $k \leq \text{mode}(Y)$.

The unimodality condition arises naturally in many real applications. We take the employee selection dataset from da Costa, Alonso and Cardoso (2008) as an example. Each observation in the dataset consists of four covariates $X = (X_1, X_2, X_3, X_4)$ from psychometric tests and an ordered response $Y \in \{1, \ldots, 9\}$ representing the overall score of the candidate. Intuitively, for a given covariate, if the score is known most likely to be 6, there is no reason to believe that the score is more likely to be 4 than to be 5. To demonstrate the reasoning, we plot the sample distributions of Y

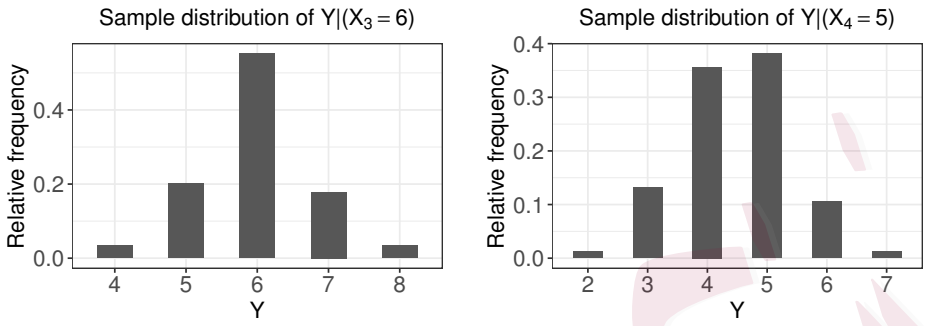

Figure 1: (Left) the sample distribution plot of $Y \mid (X_3 = 6)$ in the employee selection dataset; (Right) the sample distribution plot of $Y \mid (X_4 = 5)$ in the employee selection dataset.

given $X_3 = 6$ and $X_4 = 5$ in Fig. 1. From these two plots, we observe the unimodal distributions with modes 5 and 6 respectively.

Now we consider the ordinal classification under the linear discriminant analysis model, i.e., $X \mid (Y = k) \sim N(\mu_k, \Sigma)$. When the unimodality condition holds, the following lemma shows that there exists an intrinsic low-rank structure in the model.

Lemma 5. *For $Y \in \{1, \ldots, K\}$ and $X \mid (Y = k) \sim N(\mu_k, \Sigma)$, under the unimodality condition, let k' denote the smallest k such that $\mu_{k+1} - \mu_k \neq 0$, assume that k' exists and*

$$\log(\pi_{k+1}/\pi_k) < (\mu_{k+1} - \mu_k)^\top \Sigma^{-1} (\mu_{k+1} - \mu_k)/2, \quad k = 1, \ldots, K-1, \quad (3.1)$$

then $\mu_{k+1} - \mu_k = \alpha_k(\mu_{k'+1} - \mu_{k'})$, *where constants* $\alpha_k \geq 0$ *for* $k \leq K-1$. *Consequently,* $\mathcal{S} = \text{span}\{\Sigma^{-1}(\mu_{k'+1} - \mu_{k'})\}$ *and* $d = 1$.

The assumption that k' exists rules out the trivial case that $\mathcal{S} = \{0\}$, and (3.1) guarantees that the priors do not dominate the classification rule.

By formulating the LDA model into the multinomial logistic regression model, one can show that $\Sigma^{-1}(\mu_{k+1}-\mu_k)$ is the normal vector of the splitting hyperplane separating the consecutive classes k and $k+1$. By Lemma 5, all the splitting hyperplanes are parallel to each other. This observation complies with the parallel splitting hyperplane assumption, which is widely adopted in ordinal classification methods under the support vector machines framework (Shashua and Levin 2002, Wang et al. 2016).

The following lemma provides an intuitive example of unimodal $Y \mid X$.

Lemma 6. *For* $Y \in \{1, \ldots, K\}$ *and* $X \mid (Y = k) \sim N(\mu_k, \Sigma)$, *assume that* $\Pr(Y = k) = 1/K$ *for* $k = 1, \ldots, K$ *and* $\mu_k - \mu_{k-1} = \cdots = \mu_2 - \mu_1 \neq 0$, *the conditional distribution* $Y \mid (X = x)$ *is unimodal for any* $x \in \mathbb{R}^p$.

The second application is the response category combination problems arising in marketing and political polling, where the product preference of customers or the political stand of voters sometimes are not distinct enough to be easily differentiated by statistical models. Price et al. (2019) studied the response category combination problems by adopting the fused lasso

penalty under the multinomial logistic regression model. The indistinguishable classes condition applied is stated as follows:

Indistinguishable classes condition. The response Y takes value in $\{1, \ldots, K\}$. Assume that there exist some $k, j \in \{1, \ldots, K\}$ such that $\Pr(Y = j \mid X = x) = \Pr(Y = k \mid X = x)$ for any $x \in \mathbb{R}^p$.

Under indistinguishable classes condition, there is no clear guidance on how to make the prediction among the classes with the same posterior probability. Therefore, the indistinguishable categories are suggested to be combined. We consider the indistinguishable classes condition under the LDA model, which naturally brings the low-rank structure to discriminant subspace.The following lemma illustrate that our method is suitable for problems with intrinsically (but unknown) indistinguishable classes.

Lemma 7. *For $Y \in \{1, \ldots, K\}$ and $X \mid (Y = k) \sim N(\mu_k, \Sigma)$, under indistinguishable classes condition, the discriminant rank $d < K - 1$.*

4. Theoretical properties

We establish both the non-asymptotic and the asymptotic results for the rank determination, the subspace parameter estimation, and the classification error. For a new observation $(X^\star, Y^\star)$, let R denote the Bayes error

$\Pr(\phi(X^\star) \neq Y^\star)$, where $\phi(\cdot)$ is the Bayes rule (2.1), and conditioning on the training data, let R_n denote the empirical classification error from our estimator $\Pr(\widehat{\phi}(X^\star) \neq Y^\star \mid \widehat{\phi})$, where $\widehat{\phi}(\cdot)$ is the prediction by our method based on the n training samples. Recall that $\beta \in \mathbb{R}^{p\times d}$ and $\widehat{\beta} \in \mathbb{R}^{p\times \widehat{d}}$ are composed of the top-d left singular vectors of B and the top-$\widehat{d}$ left singular vectors of $\widehat{B}$, respectively. We consider the following subspace distance, which is bounded between 0 and 1 if $\widehat{d} = d$,

$$D(\mathcal{S}_\beta, \mathcal{S}_{\widehat{\beta}}) = D(\beta, \widehat{\beta}) = (2d)^{-1/2}\|P_\beta - P_{\widehat{\beta}}\|_F. \tag{4.1}$$

We consider the following three mild assumptions of bounded eigenvalues, bounded prior probabilities, and separable classes, respectively.

(A1) There exists constant $M > 0$ such that $M \geq \varphi_1(\Sigma) \geq \cdots \geq \varphi_p(\Sigma) \geq 1/M > 0$, where $\varphi_k(\Sigma)$ is the kth largest eigenvalue of Σ.

(A2) There exists constant $T > 0$ such that $1/(TK) \leq \pi_k \leq T/K$ for all k.

(A3) There exists constant $Q > 0$ such that $1/Q \leq (\mu_k - \mu_j)^\top \Sigma^{-1}(\mu_k - \mu_j) \leq Q$ for all $k \neq j$.

Assumption (A1) is a commonly used assumption for high-dimensional estimation (e.g., Cai et al. 2010). Assumption (A2) implies that the class size π_k is bounded away from 0 and 1. Assumption (A3) guarantees that the classes are separable in finite Mahalanobis distance.

We present the non-asymptotic results in the following theorem. Let $\sigma_{\min}$ denote the smallest non-zero singular value of B, $\varphi_{\min} \equiv \varphi_p(\Sigma)$ and $\tau = \max\{\|B\|_{2,1} + \|B\|_\star, 2K^{1/2}\}$. For ease of presentation, we assume that $\sigma_{\min}, \varphi_{\min}, d$ and K are fixed. Then τ can be interpreted as the sparsity level of B because the dominating term in τ would be the $L_{2,1}$ norm as p goes to infinity. As p diverges with n, the sparsity level is allowed to diverge with p. Thus, we allow τ to diverge with n in our theoretical study. For notational simplicity, we use C and C' to denote some generic positive constants that could vary from line to line.

Theorem 1. *Under model* (2.2) *and Assumptions (A1)–(A3), for any* ε *such that* $0 < \varepsilon \le C\tau^{-2}$, *and* λ_1, λ_2, δ *satisfying* $5\varepsilon\tau < \lambda_1 \le 6\varepsilon\tau$, $0 < \lambda_2 \le \lambda_1$ *and* $(22\varepsilon/\varphi_{\min})^{1/2}\tau < \delta \le 2(22\varepsilon/\varphi_{\min})^{1/2}\tau$, *with probability at least* $1 - C'p^2\exp(-Cn\varepsilon^2)$ *we have (i)* $\widehat{d} = d$; *(ii)* $D^2(\beta, \widehat{\beta}) \le C'\varepsilon\tau^2$; *(iii)* $|R_n - R| \le C'(\varepsilon\tau^2)^{1/3}$, *for some constants* $C, C' > 0$.

With the proper selections of the thresholding value δ and the tuning parameters λ_1 and λ_2, Theorem 1 shows that with high probability the discriminant rank d and the subspace $\mathcal{S}_\beta$ are estimated accurately, and the classification error is close to Bayes error rate. If we further assume that $\log p = o(n\tau^{-4})$, by letting $n \to \infty$, we obtain the asymptotic results.

Corollary 1. *Under the same conditions as in Theorem 1, and* $\log p = o(n\tau^{-4})$, *for* λ_1, λ_2 *and* δ *satisfying* $5C_1\tau(\log p/n)^{1/2} < \lambda_1 \leq 6C_1\tau(\log p/n)^{1/2}$, $0 < \lambda_2 \leq \lambda_1$ *and* $C_2\tau(\log p/n)^{1/4} < \delta \leq 2C_2\tau(\log p/n)^{1/4}$ *for some positive constants* C_1 *and* C_2, *as* $n, p \to \infty$, *we have (i)* $\Pr(\widehat{d} = d) \to 1$; *(ii)* $D(\beta, \widehat{\beta}) \to 0$ *in probability; (iii)* $|R_n - R| \to 0$ *in probability.*

Corollary 1 shows that the rank determination, the subspace estimation and the Bayes' classification error are consistent as $n, p \to \infty$, where p is allowed to grow with n at an exponential rate.

5. Simulations

To demonstrate the effectiveness of our proposed LSLDA, we conduct simulations from the reduced-rank LDA model (2.2) under high-dimensional sparse settings. In models (M1) and (M2), we vary the predictor correlation from mild and strong. In model (M3), we have unbalanced classes. In models (M4) and (M5), K is relatively large, where (M5) is near full-rank $d = K - 2$. In model (M6), we construct the unimodal distribution of $Y \mid X$ according to Lemma 6. In model (M7), we concur the indistinguishable classes condition, where the posterior probability of classes 2, 3 and 4 are the same. Finally, in model (M8), we vary the parameters s, p, n, K one at a time to illustrate a wide range of settings.

We set $n_k = 30$, $s = 10$, and $p = 3000$ unless otherwise specified. Let n denote the total training sample size. For all models, we generate a separate validation set of size n for parameter tuning and a test set of size $5n$ for model evaluation. We set Σ as a block-diagonal matrix of blocks $\widetilde{\Sigma}$ and I_{2500}), where $\widetilde{\Sigma} \in \mathbb{R}^{500\times 500}$ is positive-definite. Recall that $X \mid (Y = k) \sim N(\mu_k, \Sigma)$ and $\mathcal{S} = \Sigma^{-1}\text{span}(\mu_2 - \mu_1, \ldots, \mu_K - \mu_1)$. We fix $\mu_1 = 0$ and define $\theta_k = \Sigma^{-1}\mu_{k+1}$ for $k = 1, \ldots, K-1$. Then, we generate the discriminant basis $\beta \in \mathbb{R}^{p\times d}$ by taking the top-d left singular vectors of $\theta = (\theta_1, \ldots, \theta_{K-1}) \in \mathbb{R}^{p\times(K-1)}$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times p}$, we call it has the AR$(r, p)$ structure if $a_{ij} = r^{|i-j|}$ for $i, j = 1, \ldots, p$, and the CS(r, p) structure if $a_{ii} = 1$ for $i = 1, \ldots, p$ and $a_{ij} = r$ for $i \neq j$. For each model, the number of classes K, the vectors θ_k, the matrix $\widetilde{\Sigma}$ and the discriminant rank d, are listed as follows, where θ_{kj} denotes the j-th element of θ_k. The vectors θ_k in each model are designed to keep the Bayes error less than 20%.

(M1) (Mild correlation) $K = 4$, $d = 2$, θ_{1i} takes the value 0.8 for $i = 1, \ldots, 5$ and 0 otherwise, θ_{2i} takes value 0.8 for $i = 6, \ldots, 10$ and 0 otherwise, and $\theta_3 = \theta_1 + \theta_2$. The matrix $\widetilde{\Sigma} = \text{AR}(0.5, 500)$.

(M2) (Strong correlation) Same as (M1) except $\theta_3 = 1.5\theta_1 + 1.5\theta_2$ and $\widetilde{\Sigma} = I_{10} \otimes \text{CS}(0.3, 50)$.

(M3) (Unbalanced data) Same as (M2), except that the class sizes (in the training set) are now 10, 10, 50 and 50.

(M4) (Large K) $K = 7$, $d = 2$, $\theta_{1,2i-1} = 2$ and $\theta_{2,2i} = -4$ for $i = 1, \ldots, 5$. For $k = 3, \ldots, K-1$, $\theta_k = (k/2-1)(\theta_1 + \theta_2)$. And $\widetilde{\Sigma} = \mathrm{AR}(0.5, 500)$.

(M5) (Near full-rank basis) $K = 7$, $d = 5$, θ_{ki} takes the value 2 for $i = 2k-1, 2k$ and $k = 1, \ldots, 5$, and 0 otherwise, and $\theta_6 = 0.5 \sum_{k=1}^{5} \theta_k$. And $\widetilde{\Sigma} = \mathrm{AR}(0.5, 500)$.

(M6) (Unimodality) $K = 4$, $d = 1$, $\theta_2 = 2\theta_1$, $\theta_3 = 3\theta_1$, where θ_{1i} takes the value 1 for $i = 1, \ldots, 5$, the value -1 for $i = 6, \ldots, 10$, and 0 otherwise. And $\widetilde{\Sigma} = I_{10} \otimes \mathrm{CS}(0.3, 50)$.

(M7) (Indistinguishable classes) $K = 4$, $d = 1$, $\theta_1 = \theta_2 = \theta_3$, where θ_{1i} takes the value 1 for $i = 1, \ldots, 5$, the value -1 for $i = 6, \ldots, 10$, and 0 otherwise. And $\widetilde{\Sigma} = I_{10} \otimes \mathrm{CS}(0.3, 50)$.

In each model setting, we compare our LSLDA method with several competitors, including the supervised PCA-based LDA (SPCALDA; Niu et al. 2018), the multi-class sparse discriminant analysis (MSDA; Mai et al. 2019), the sparse optimal scoring (SOS; Clemmensen et al. 2011), the penalized LDA (PLDA; Witten and Tibshirani 2011) and the penalized multinomial logistic regression model (Logistic, Friedman et al. 2010). The

Table 1: The means (and the standard errors) of the classification error (%), the subspace distance D, the TPR (%) and the FPR (%) on simulated data generated from Models (M1)–(M6). The results are based on 200 replicates. The standard errors for TPR and FPR are all less than 3.5%, and are thus omitted.

Method	Err(%)	D	TPR(%)	FPR(%)	Err(%)	D	TPR(%)	FPR(%)
		Model (M1)				Model (M2)		
Bayes	17.4(0.1)	–	–	–	14.2(0.1)	–	–	–
LSLDA	**18.9(0.1)**	0.321(0.538)	99.9	0.8	**16.4(0.2)**	0.379(0.975)	99.7	0.6
PP	56.8(0.3)	1.121(0.035)	100.0	100.0	58.7(0.4)	1.170(0.008)	100.0	100.0
SPCALDA	48.2(0.2)	1.296(3.406)	100.0	100.0	33.6(0.1)	1.411(3.328)	100.0	100.0
MSDA	22.4(0.2)	0.809(0.453)	77.5	0.1	**19.9(0.2)**	0.815(0.442)	78.8	0.2
SOS($q = K-1$)	24.3(0.2)	0.656(0.474)	97.3	0.5	35.4(0.2)	0.981(0.261)	66.3	0.8
SOS($q = d$)	**19.7(0.1)**	0.443(0.722)	97.0	0.4	33.2(0.2)	0.843(0.309)	70.5	0.5
PLDA($q = K-1$)	49.2(0.2)	1.056(0.056)	100.0	100.0	32.3(0.1)	1.055(0.424)	100.0	97.0
PLDA($q = d$)	48.8(0.4)	0.931(0.065)	100.0	100.0	33.9(0.2)	0.932(0.331)	99.7	95.5
Logistic	22.1(0.2)	0.799(0.404)	82.8	0.3	24.7(0.2)	0.877(0.428)	74.2	0.3
		Model (M3)				Model (M4)		
Bayes	8.6(0.1)	–	–	–	3.2(0.1)	–	–	–
LSLDA	**10.8(0.2)**	0.525(1.415)	98.9	0.6	**9.0(0.3)**	0.698(1.604)	88.0	2.2
PP	41.6(0.4)	1.168(0.008)	100.0	100.0	45.0(0.4)	1.315(0.030)	100.0	100.0
SPCALDA	25.0(0.7)	0.857(0.646)	100.0	100.0	28.0(0.3)	1.544(7.625)	100.0	100.0
MSDA	**13.3(0.1)**	0.872(0.448)	67.8	0.2	12.3(0.4)	1.207(0.386)	57.2	0.8
SOS($q = K-1$)	19.5(0.1)	0.978(0.273)	64.7	0.8	12.2(0.1)	1.189(0.038)	70.2	1.1
SOS($q = d$)	18.9(0.2)	0.839(0.322)	69.5	0.5	**8.2(0.1)**	0.649(0.082)	69.2	0.3
PLDA($q = K-1$)	17.6(0.1)	1.059(0.336)	100.0	98.0	18.7(0.2)	0.536(0.220)	89.4	0.0
PLDA($q = d$)	19.6(0.6)	0.939(0.216)	100.0	98.0	18.7(0.2)	0.536(0.220)	89.4	0.0
Logistic	46.8(1.3)	0.977(0.134)	22.1	0.6	27.4(0.2)	1.241(0.386)	74.0	0.5
		Model (M5)				Model (M6)		
Bayes	10.1(0.1)	–	–	–	13.9(0.1)	–	–	–
LSLDA	**11.6(0.1)**	0.235(0.298)	100.0	0.2	**15.2(0.1)**	0.219(1.461)	100.0	0.7
PP	60.0(0.2)	0.999(0.019)	100.0	100.0	61.9(0.3)	1.446(0.030)	100.0	100.0
SPCALDA	54.7(0.2)	1.179(1.484)	100.0	100.0	61.1(0.2)	2.269(4.379)	100.0	100.0
MSDA	13.0(0.1)	0.496(0.457)	96.1	0.1	18.4(0.2)	1.059(0.209)	98.8	0.2
SOS($q = K-1$)	14.9(0.1)	0.530(0.420)	100.0	0.9	25.0(0.2)	1.029(0.131)	100.0	0.9
SOS($q = d$)	13.6(0.1)	0.428(0.514)	99.9	0.9	**15.6(0.1)**	0.241(0.475)	100.0	0.1
PLDA($q = K-1$)	50.5(0.4)	0.888(1.186)	86.4	76.5	60.9(0.2)	1.344(0.054)	100.0	100.0
PLDA($q = d$)	56.0(0.7)	0.832(1.081)	81.7	70.5	68.2(0.1)	0.899(0.161)	99.0	99.0
Logistic	**12.4(0.1)**	0.451(0.300)	99.8	0.4	34.5(0.2)	1.138(0.320)	90.0	0.5

five competitors above are implemented by R packages `SPCALDA`, `msda`, `sparseLDA`, `penalizedLDA` and `glmnet`, respectively. We also include a simple projection pursuit method (PP) that first project the data onto $\widehat{U} \in \mathbb{R}^{p\times K}$ to reduce the dimension of X from p to K. The linear discriminant analysis is then performed on the K-dimensional reduced predictor. In addition, we include the Bayes error, i.e., the best possible error rate. The implementations of SOS and PLDA in R packages provide the option to pre-specify the number of discriminant directions, denoted by q. We consider both the full rank option (i.e., specifying $q = K - 1$) and the option of using the true rank (i.e., specifying $q = d$).

We compare different methods by several criteria, including the classification error, the subspace estimation error, the true positive rate (TPR) and the false positive rate (FPR). The subspace estimation error is measured by the subspace distance defined in (4.1). With the true active set $\mathcal{A}$ and the estimated active set $\widehat{\mathcal{A}}$, we obtain the TPR $= |\widehat{\mathcal{A}} \cap \mathcal{A}|/|\mathcal{A}|$ and the FPR $= |\widehat{\mathcal{A}} \cap \mathcal{A}^c|/|\mathcal{A}^c|$. We report these comparison criteria over 200 replicates under Models (M1)–(M6) in Table 1. Due to space limit, the results under model (M7), which is further evaluated under different criteria, and the estimated ranks from LSLDA and SPCALDA (the only two methods that are able to select ranks) are provided in the Supplementary Materials.

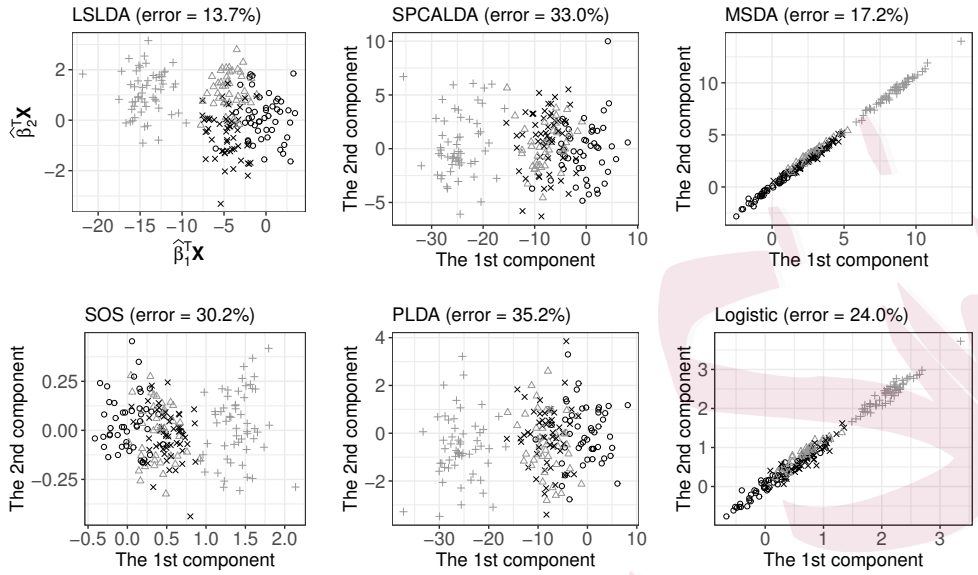

Figure 2: The scatterplots of the first two discriminant components $\widehat{\beta}_1^\top X$ and $\widehat{\beta}_2^\top X$ estimated from LSLDA and the first two components estimated from other competitors. The rank $d = 2$ is given for SOS and PLDA. The plots are based on one replicate in Model (M2) and the samples in each class are represented by different symbols.

Overall, the proposed method significantly outperforms all the other competitors. It is almost as good as the Bayes rule in classification and provides the best subspace estimation and variable selection results. The only exception is in model (M4), where SOS with true rank information has an edge over LSLDA, which is still significantly better than all other methods. We note that with the knowledge of true rank d, the results of

SOS can improve substantially over the standard (full-rank) SOS. Both PP and SPCALDA fail in all criteria due to the lack of variable selection. Since PP gives consistently poor performance, we exclude it from all subsequent simulations. From Table S3 in Supplementary Materials, we also show that our method can select the rank consistently, while SPCALDA severely overestimates the rank in most settings.

The classification error of MSDA is usually close to that of our method. But, MSDA fails to estimate the discriminant subspace accurately and tends to miss important variables. Logistic regression performs poorly, because it is expected to lose efficiency comparing to the LDA-based methods. Comparing models (M1) to (M2), LSLDA and MSDA are more robust to strong correlation than other methods. For unbalanced data in model (M3), LSLDA performs well on both majority and minority classes, and the additional results are provided in the Supplementary Materials. The results from Models (M6) and (M7) also confirm the effectiveness of our proposal in the ordinal classification and the response category combination problems.

For Model (M2), we visualize in Fig. 2 the first two discriminant directions/components from each method (based on 1/3 of the test data, and from one replicate). From Fig. 2, the four classes are well separated by our estimator which is the clear winner in this setting.

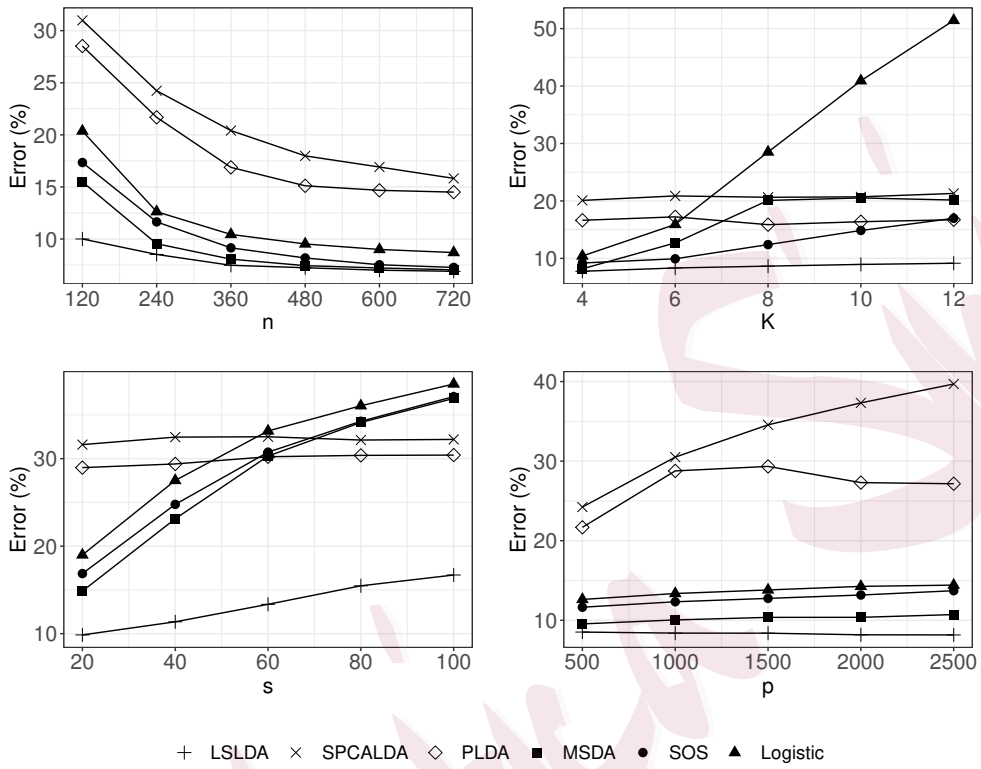

Figure 3: The means of classification error (%) as one of the parameters n, K, s, p varies. The results are based on 200 replicates in Model (M8). The rank $d = 2$ is provided as known for SOS and PLDA. We set $(K, s, n, p) = (4, 20, 120, 500)$ as default for all simulations except for varying K setting we have $n = 360$ and for varying p setting we have $n = 240$.

We construct another model to study the effects of the sample size n, the number of classes K, the sparsity level s, and the total number of predictors p. The covariance matrix Σ and the vectors θ_k, $k = 1, \ldots, K-1$,

are given as follows, where we set $\|\Sigma^{1/2}\theta_1\|_2$ and $\|\Sigma^{1/2}\theta_2\|_2$ as fixed in order to keep a reasonable Bayes error.

(M8) $\theta_{1i} = w$ for $i = 1, \ldots, s$, $\theta_{2,2j-1} = z$ and $\theta_{2,2j} = -z$ for $j = 1, \ldots, s/2$, where the positive constants w and z are selected such that $\|\Sigma^{1/2}\theta_1\|_2 = \|\Sigma^{1/2}\theta_2\|_2 = 5$. For $k = 3, \ldots, K-1$, $\theta_k = (k/2-1)(\theta_1+\theta_2)$. The covariance matrix Σ has the AR$(0.5, p)$ structure. The discriminant rank $d = 2$.

The averaged classification errors over 200 replicates for each method are displayed in Fig. 3. The SOS and PLDA in the comparison use the true rank by specifying $q = d$, which are better than their full rank versions. In general, LSLDA outperforms all the other competitors. As we increase the sample size n, all methods except for PLDA and SPCALDA are converging quickly to the Bayes error. When K increases, the low-rank estimators, LSLDA and SPCALDA are more robust than others. However, for MSDA, SOS and Logistic, since more redundant directions are estimated, their performances are getting worse as K increases. Also, as the sparsity level s increases, the classification errors of MSDA, SOS and Logistic rise up rapidly. This might due to the poor variable selection, as seen in Table 1. When p increases, it is observed that the performances of LSLDA, MSDA, SOS, and Logistic are not affected remarkably. Thus, our proposed method

is effective in a wide-range of parameter settings.

In Section S2.2 of Supplementary Materials, we report the computation time of all the LDA-based methods (LSLDA, SPCALDA, MSDA, SOS, and PLDA). Result suggests that LSLDA is indeed computationally efficient and scalable to very high dimensions.

6. Real data analysis

We study three face image data sets **face94**, **face95** and **grimace** collected by Spacek (2009). For each subject k, $n_k = 20$ images are taken with variation of facial expression, position of face in image, head scale and so on. The task is to classify these images to the corresponding subject. In **face94**, we have $K = 20$ male staffs. In **face95**, we only use face images of the first 15 subjects out of the total 72 subjects, so $K = 15$. Finally, **grimace** contains $K = 18$ subjects. In each data set, greyscale images of size 180×200 are transformed into a vector of dimension $360,000$. Following Mai et al. (2019), we perform the F-test variable screening (designed for multi-category response) on these predictors, and keep $p = 500$ variables.

To compare our method LSLDA with the same competitors in simulations, each data set is randomly split into training and test sets with a $3:1$ ratio, and the tuning parameters are selected by five-fold cross validation

Table 2: The means (and the standard errors) of the classification error (%) and the estimated sparsity level $\widehat{s}$ over 100 training-test set splits.

		LSLDA	PP	SPCALDA	MSDA	SOS	PLDA	Logistic
face94	Err(%)	**0.2(0.0)**	**0.0(0.0)**	0.4(0.1)	0.2(0.0)	0.3(0.1)	1.0(0.1)	58.8(0.3)
	$\widehat{s}$	233.2(10.3)	500.0(0.0)	500.0(0.0)	66.5(0.7)	104.5(10.0)	500.0(0.0)	10.9(0.3)
face95	Err(%)	**24.5(0.5)**	**24.6(0.4)**	24.7(0.4)	33.6(0.5)	27.5(0.4)	44.1(0.4)	36.5(0.4)
	$\widehat{s}$	227.6(3.4)	500.0(0.0)	500.0(0.0)	24.3(0.4)	326.8(14.4)	500.0(0.0)	24.1(0.3)
grimace	Err(%)	**0.0(0.0)**	**0.0(0.0)**	0.1(0.1)	0.0(0.0)	0.1(0.0)	1.1(0.1)	0.5(0.1)
	$\widehat{s}$	241.5(3.1)	500.0(0.0)	500.0(0.0)	76.8(1.2)	130.1(0.6)	500.0(0.0)	23.8(0.3)

on the training set. After the model is refitted with the selected tuning parameters, the evaluation on the test set is recorded. The averaged classification error and the estimated sparsity level $\widehat{s}$ over 100 training-test set splits are recorded in Table 2. From Table 2, we can see that our method achieves competitive classification accuracy on all data sets. Compared to PP and SPCALDA, our method produces a sparse estimator. Moreover, although PP is also highly accurate on the real datasets, it produces a $(K-1)$-dimensional reduction of the data, while LSLDA is more aggressive in achieving low-rank data projection. On the other hand, compared to other sparse competitors, our estimator makes use of low-rank structure to attain lower classification error.

The averaged estimated rank $\widehat{d}$ from LSLDA (versus SPCALDA) on

face94, **face95** and **grimace** are 7.7 (versus 3.6), 9.4 (versus 14.5), and 11.5 (versus 6.3), respectively. The standard errors are all less than 0.5. Both methods produce low-rank estimator and the advantage of one method over the other is not clear. We provide the low-dimensional visualization of the data points using the two methods. In Supplementary Materials, we show that LSLDA has better visualization and separation of classes than SPCALDA.

7. Discussion

In this paper, we consider the reduced-rank linear discriminant analysis model in high dimensions. Motivated from low-dimensional likelihood-based dimension reduction approach, we propose a doubly penalized convex optimization and developed a computationally efficient algorithm. Simulations and real data analysis provide two complementary perspectives for LSLDA. Simulations suggest that the proposed LSLDA method is widely applicable provided the sample size is not too small (e.g., $n_k \geq 10$), and the Bayes classifier is reasonably sparse (e.g, $s \leq 100$). We have tested LSLDA on datasets with dimensions up to 25,000, and the algorithm converges within a reasonable amount of time. The low-rank assumption may be especially desirable when the number of classes is large, but the advan-

tage starts to show when K is as small as 4 in the simulations. Thanks to the synergy between low-rank and sparse inducing penalties, our method is generally more accurate and robust than existing sparse LDA methods (such as PLDA and SOS), while the non-sparse projection-based classification methods (such as SPCALDA or PP) clearly fail under sparsity assumptions. However, in real data analysis the non-sparse projection-based methods perform well. LSLDA adapts to these problems by automatically learning a less sparse ($\widehat{s} \geq 200$ from $p = 500$) but low-dimensional ($7 \leq \widehat{d} \leq 11$ from $K = 15$) structure from these data sets, and outperforms most competitors.

Supplementary Materials

Supplementary materials include our alternating direction method of multipliers algorithm that is compared to the proposed three-operator splitting algorithm, additional numerical results and technical proofs.

Acknowledgements

The authors thank the Editor, Associate Editor, and Referee for helpful comments. The authors are also grateful for insightful discussion with Aaron J. Molstad. Research for this paper was supported in part by grants

CCF-1908969, DMS-2053697 and DMS-2113590 from the US National Science Foundation.

References

Anderson, T. W. (1951), 'Estimating linear restrictions on regression coefficients for multivariate normal distributions', *The Annals of Mathematical Statistics* **22**(3), 327–351.

Bickel, P. J. and Levina, E. (2004), 'Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations', *Bernoulli* **10**(6), 989–1010.

Boyd, S., Parikh, N. and Chu, E. (2011), *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Now Publishers Inc.

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, T. T., Zhang, C.-H. and Zhou, H. H. (2010), 'Optimal rates of convergence for covariance matrix estimation', *The Annals of Statistics* **38**(4), 2118–2144.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011), 'Sparse discriminant analysis', *Technometrics* **53**(4), 406–413.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York:John Wiley & Sons.

Cook, R. D. and Forzani, L. (2008), 'Principal fitted components for dimension reduction in

regression', *Statistical Science* **23**(4), 485–501.

da Costa, J. F. P., Alonso, H. and Cardoso, J. S. (2008), 'The unimodal model for the classification of ordinal data', *Neural Networks* **21**(1), 78–91.

da Costa, J. F. P., Sousa, R. and Cardoso, J. S. (2010), An all-at-once unimodal svm approach for ordinal classification, *in* '2010 Ninth International Conference on Machine Learning and Applications', IEEE, pp. 59–64.

Davis, D. and Yin, W. (2017), 'A three-operator splitting scheme and its optimization applications', *Set-valued and variational analysis* **25**(4), 829–858.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1.

Gao, C., Ma, Z. and Zhou, H. H. (2017), 'Sparse cca: Adaptive estimation and computational barriers', *The Annals of Statistics* **45**(5), 2074–2101.

Hao, N., Dong, B. and Fan, J. (2015), 'Sparsifying the fisher linear discriminant by rotation', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(4), 827–851.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Izenman, A. J. (1975), 'Reduced-rank regression for the multivariate linear model', *Journal of multivariate analysis* **5**(2), 248–264.

Mai, Q., Yang, Y. and Zou, H. (2019), 'Multiclass sparse discriminant analysis', *Statistica Sinica*

29(1), 97–111.

Mai, Q., Zou, H. and Yuan, M. (2012), 'A direct approach to sparse discriminant analysis in ultra-high dimensions', *Biometrika* **99**(1), 29–42.

McCullagh, P. (1980), 'Regression models for ordinal data', *Journal of the Royal Statistical Society: Series B (Methodological)* **42**(2), 109–127.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008), 'The group lasso for logistic regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 53–71.

Niu, Y. S., Hao, N. and Dong, B. (2018), 'A new reduced-rank linear discriminant analysis method and its applications', *Statistica Sinica* pp. 189–202.

Price, B. S., Geyer, C. J. and Rothman, A. J. (2019), 'Automatic response category combination in multinomial logistic regression', *Journal of Computational and Graphical Statistics* **28**(3), 758–766.

Qiao, X. (2015), 'Learning ordinal data', *Wiley Interdisciplinary Reviews: Computational Statistics* **7**(5), 341–346.

Roth, V. and Fischer, B. (2008), The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms, *in* 'Proceedings of the 25th international conference on Machine learning', pp. 848–855.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011), 'Sparse linear discriminant analysis by thresholding for high dimensional data', *The Annals of statistics* **39**(2), 1241–1265.

Shashua, A. and Levin, A. (2002), 'Ranking with large margin principle: Two approaches', *Advances in neural information processing systems* **15**, 961–968.

Spacek, L. (2009), 'Facial images databases', `http://cmp.felk.cvut.cz/~spacelib/faces/`.

Stoica, P. and Viberg, M. (1996), 'Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions', *IEEE Transactions on Signal Processing* **44**(12), 3069–3078.

Tan, K. M., Wang, Z., Zhang, T., Liu, H. and Cook, R. D. (2018), 'A convex formulation for high-dimensional sparse sliced inverse regression', *Biometrika* **105**(4), 769–782.

Tan, K., Shi, L. and Yu, Z. (2020), 'Sparse sir: optimal rates and adaptive estimation', *The Annals of Statistics* **48**(1), 64–85.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013), 'Fantope projection and selection: A near-optimal convex relaxation of sparse pca', *Advances in neural information processing systems* **26**, 2670–2678.

Wang, J., Shen, X., Sun, Y. and Qu, A. (2016), 'Classification with unstructured predictors and an application to sentiment analysis', *Journal of the American Statistical Association* **111**(515), 1242–1253.

Wen, C.-H. and Koppelman, F. S. (2001), 'The generalized nested logit model', *Transportation Research Part B: Methodological* **35**(7), 627–641.

Witten, D. M. and Tibshirani, R. (2011), 'Penalized classification using fisher's linear dis-

criminant', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 753–772.

Ye, J. and Li, Q. (2005), 'A two-stage linear discriminant analysis via qr-decomposition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(6), 929–941.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007), 'Dimension reduction and coefficient estimation in multivariate linear regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(3), 329–346.

Yuan, M. and Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Zhou, H. and Li, L. (2014), 'Regularized matrix regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(2), 463–483.

Department of Statistics, Florida State University, Tallahassee, Florida 32306, U.S.A.

E-mails: jing.zeng@stat.fsu.edu; henry@stat.fsu.edu; qmai@fsu.edu