



Sparse linear discriminant analysis via ℓ_0 constraint

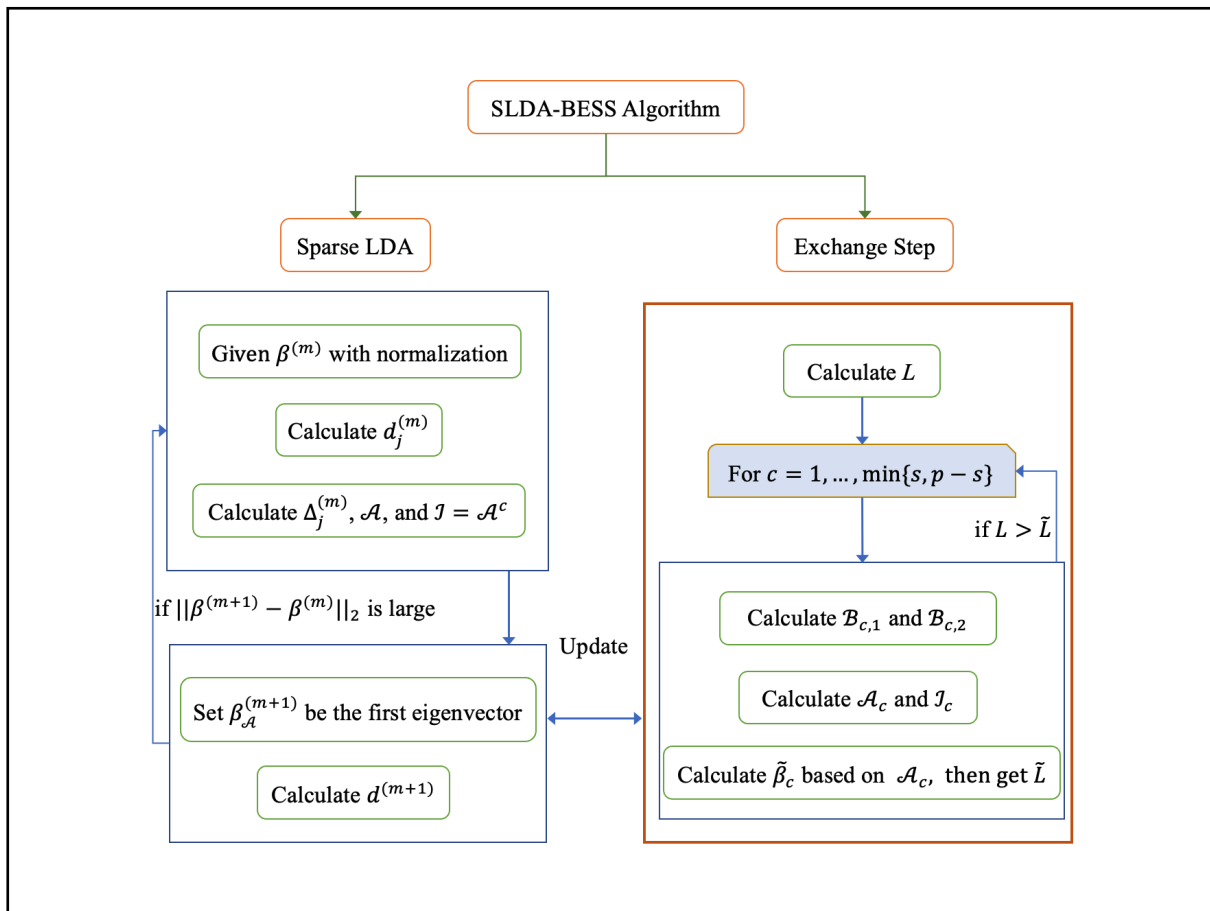
Qi Yin, and Lei Shu 

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

 Correspondence: Lei Shu, E-mail: sl2018@mail.ustc.edu.cn

© 2022 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Graphical abstract




An algorithm for adding the exchange step to improve naive sparse linear discriminant analysis.

Public summary

- This paper extends LDA to a high-dimensional setting by adding an ℓ_0 penalty to produce a discriminant vector involving only a subset of features.
- We propose an SLDA-BASS algorithm to directly solve the estimation of sparse LDA with ℓ_0 constraint, avoiding the unnecessary information loss caused by relaxing the constraint in the conventional algorithm.
- Compared to other estimation methods, the SLDA-BASS algorithm is derived from a natural criterion and is superior in terms of sparsity recovery as well as computational efficiency.

Sparse linear discriminant analysis via ℓ_0 constraint

Qi Yin, and Lei Shu *Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China* Correspondence: Lei Shu, E-mail: sl2018@mail.ustc.edu.cn© 2022 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).Cite This: *JUSTC*, 2022, 52(8): 4 (6pp)

Read Online

Abstract: We consider the problem of interpretable classification in a high-dimensional setting, where the number of features is extremely large and the number of observations is limited. This setting has been extensively studied in the chemometric literature and has recently become pervasive in the biological and medical literature. Linear discriminant analysis (LDA) is a canonical approach for solving this problem. However, in the case of high dimensions, LDA is unsuitable for two reasons. First, the standard estimate of the within-class covariance matrix is singular; therefore, the usual discriminant rule cannot be applied. Second, when p is large, it is difficult to interpret the classification rules obtained from LDA because p features are involved. In this setting, motivated by the success of the primal-dual active set algorithm for best subset selection, we propose a method for sparse linear discriminant analysis via ℓ_0 constraint, which imposes a sparsity criterion when performing linear discriminant analysis, allowing classification and feature selection to be performed simultaneously. Numerical results on synthetic and real data suggest that our method obtains competitive results compared with existing alternative methods.

Keywords: best subset selection; linear discriminant analysis; ℓ_0 constraint; projection

CLC number: O212.4 **Document code:** A

2020 Mathematics Subject Classification: 62H30

1 Introduction

Linear discriminant analysis (LDA) is a prevalent supervised classification tool in many applications, owing to its simplicity, robustness, and predictive accuracy. LDA uses label information to learn discriminant projections, which can dramatically maximize the between-class distance and decrease the within-class distance, thus improving classification accuracy. Simultaneously, low-dimensional data projections in most discriminant directions are valuable for data interpretation. LDA classifiers can be constructed in three different ways: the multivariate Gaussian model, optimal scoring problem, and Fisher's discriminant problem (see, for example, Hastie et al.^[1]).

LDA is effective and asymptotically optimal when the dimension p is fixed and the number of observations n is large; that is, its misclassification rate converges to 0 over the optimal rule as n increases to infinity. Shao et al.^[2] indicated that LDA remains asymptotic when p diverges to infinity at a rate slower than \sqrt{n} . However, with tremendous advances in data collection, high-dimensional data with dimension p potentially larger than the number of observations n are now frequently encountered in a wide range of applications, and the classification of these data has recently attracted considerable attention. Common applications include genomics, functional magnetic resonance imaging, risk management, and web searches.

In high-dimensional settings, standard LDA performs poorly and may even fail completely. For example, Bickel and Levina^[3] indicated that LDA could not perform better

than random guessing when $p > n$. In addition, in this case, the sample covariance matrix is singular, and its inverse matrix is not well identified. Consequently, it is challenging to select and extract the most discriminative features for supervised classification. A natural solution is to replace the inverse matrix with the generalized inverse matrix of the sample covariance matrix. However, such an estimation is highly biased, unrobust, and can contribute to the poor performance of the classifier. When p is large, the resulting classifier is difficult to interpret because the classification rules involve a linear combination of all p features. Thus, when $p \gg n$, one may desire a classifier with parsimonious features, that is, a classifier that involves only a subset of p features. Such a sparse classifier ensures that the model is easier to interpret and can reduce the overfitting of training data.

Recently, several studies have extended LDA to high-dimensional settings. Some of this literature addressed non-sparse classifiers. For example, in multivariate Gaussian models of LDA, Dudoit et al.^[4] and Bickel and Levina^[3] assumed the independence of features (naive Bayes) and Friedman^[5] suggested applying a ridge penalty to the within-class covariance matrix. Xu et al.^[6] considered other positive definite estimates of the within-class covariance matrix. In addition, some research on sparse classifiers has been conducted. Tibshirani et al.^[7] adapted a naive Bayesian classifier by soft thresholding the mean vector, and Guo et al.^[8] combined a ridge-type penalty on the within-class covariance matrix with a soft thresholding operation. Witten and Tibshirani^[9] employed an ℓ_1 penalty for Fisher's discriminant problem to obtain sparse discriminant vectors, but this approach does not

generalize to the Gaussian mixture setting and lacks simplicity if it is in regression-based optimal scoring problems.

Motivated by the Fisher's discriminant framework of Witten and Tibshirani^[9], we develop a sparse version of LDA with the ℓ_0 constraint. Previous sparse LDAs were typically implemented by imposing penalties of ℓ_1 , ℓ_2 , or a mixture of ℓ_1 and ℓ_2 , because ℓ_0 regularization is non-convex and NP-hard. Whether the optimization problem is more correct is based on the prediction effectiveness, false discovery rate, and sparsity interpretation. ℓ_0 penalized methods have lower error bounds than ℓ_1 methods. Mathematically, for a pre-specified degree of sparsity s , the discriminant vector can be determined by the following optimization problem:

$$\begin{aligned} & \text{maximize}_{\beta_k} && \beta_k^\top \Sigma_b \beta_k \\ & \text{subject to} && \beta_k^\top \Sigma_w \beta_k = 1, \quad \beta_k^\top \Sigma_w \beta_l = 0 \quad \forall l < k, \\ & && \|\beta_k\|_0 \leq s, \end{aligned}$$

where $\|\beta\|_0$ denotes the number of nonzero elements of β , Σ_b is the between-class covariance matrix, and Σ_w is the within-class covariance matrix. Further details are introduced in Section 2.

This study presents a new iterative thresholding algorithm to estimate linear discriminant vectors, which is a development of the primal-dual active-set algorithm proposed by Zhu et al.^[10] to solve the best subset selection problem in regression. The contribution of this study is two-fold. First, we consider addressing the estimation problem of sparse LDA by directly solving the ℓ_0 constrained optimization problem, which avoids unnecessary information loss owing to relaxing the constraints. Second, we constructed a polynomial algorithm for estimating sparse linear discriminant vectors, which is computationally efficient and easy to implement.

The remainder of this paper is organized as follows. In Section 2, we review LDA and classical solution procedures and then propose a new solution for sparse linear discriminant analysis. Sections 3 and 4 compare the results of our proposed method and existing methods in simulated experiments and applications to real data, respectively. Section 5 presents our conclusions.

2 Methodology

2.1 A review of linear discriminant analysis

Let \mathbf{X} be an $n \times p$ data matrix with p features measured on n observations and suppose that each of the n observations belongs to one of the K classes. In addition, we assume that each of the p features is centered to satisfy the zero mean and is normalized to have an identical variance if they are not measured on the same scale. Let \mathbf{x}_i denote the i th observation and C_k denote the index set of observations in the k th class.

Consider a simple multivariate Gaussian data generation process in which the distribution of observations in class k is $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_w)$, where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the mean vector of the k th class and Σ_w is a $p \times p$ within-class covariance matrix over all K classes. Here, $|C_k|^{-1} \sum_{i \in C_k} \mathbf{x}_i$ is the estimate of $\boldsymbol{\mu}_k$ and $n^{-1} \sum_{k=1}^K \sum_{i \in C_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$ is the estimate of Σ_w , where $|\cdot|$ denotes the cardinality of the index set. The LDA classification rule then results in the application of the Bayesian rule to estimate the most likely class for a test observation.

LDA can also be argued to arise from Fisher's discriminant problem. We define the between-class covariance matrix $\Sigma_b = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$, where π_k is the prior probability of class k . Fisher's discriminant problem involves solving for the discriminant vectors $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_{K-1}$, which sequentially maximizes the vector. The corresponding optimization problem is

$$\begin{aligned} & \text{maximize}_{\beta_k} && \beta_k^\top \Sigma_b \beta_k \\ & \text{subject to} && \beta_k^\top \Sigma_w \beta_k = 1, \quad \beta_k^\top \Sigma_w \beta_l = 0 \quad \forall l < k. \end{aligned}$$

Because the rank of Σ_b is at most $K - 1$, the above-generalized eigenproblem has at most $K - 1$ nontrivial solutions and, therefore, at most $K - 1$ discriminant vectors. These solutions are the directions in which the data have the maximal between-class covariance relative to their within-class covariance. Moreover, it has been demonstrated that classification based on the nearest centroid of matrix $(\mathbf{X}\boldsymbol{\beta}_1, \dots, \mathbf{X}\boldsymbol{\beta}_{K-1})$ produces the same LDA classification rule as the multivariate Gaussian model described previously (see Hastie et al.^[11]). One advantage of Fisher's discriminant problem over the multivariate Gaussian model of LDA is that it allows for reduced-rank classification by performing nearest centroid classification on the matrix $(\mathbf{X}\boldsymbol{\beta}_1, \dots, \mathbf{X}\boldsymbol{\beta}_q)$ with $q < K - 1$. It can be proven that performing nearest centroid classification on this $n \times q$ matrix is exactly equivalent to conducting a full-rank LDA on the $n \times q$ matrix.

The standard estimate of the within-class covariance matrix Σ_w is:

$$\widehat{\Sigma}_w = \frac{1}{n} \sum_{k=1}^K \sum_{i \in C_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^\top, \quad (1)$$

where $\hat{\boldsymbol{\mu}}_k$ denotes the sample mean vector for class k . In this subsection, we assume that $\widehat{\Sigma}_w$ is nonsingular. Furthermore, the standard estimate for the between-class covariance matrix Σ_b is given by:

$$\widehat{\Sigma}_b = \frac{1}{n} \mathbf{X}^\top \mathbf{X} - \widehat{\Sigma}_w = \frac{1}{n} \sum_{k=1}^K n_k \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^\top,$$

where $n_k = |C_k|$.

In later subsections, we will make use of the fact that

$$\widehat{\Sigma}_b = \frac{1}{n} \mathbf{X}^\top \mathbf{Y} (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{X},$$

where \mathbf{Y} is an $n \times K$ matrix and Y_{ik} is an indicator of whether observation i is in the k th class. Then, the empirical version of LDA can be written as

$$\begin{aligned} & \text{maximize}_{\beta_k} && \beta_k^\top \widehat{\Sigma}_b \beta_k \\ & \text{subject to} && \beta_k^\top \widehat{\Sigma}_w \beta_k \leq 1, \quad \beta_k^\top \widehat{\Sigma}_w \beta_l = 0 \quad \forall l < k. \end{aligned} \quad (2)$$

Problem (2) is commonly written in terms of the equality constraint instead of an inequality constraint, but the two are equivalent when $\widehat{\Sigma}_w$ is full rank, as detailed in Ref. [9, Appendix A]. The solution $\hat{\boldsymbol{\beta}}_k$ to problem (2) is generally referred to as the k th discriminant vector. Problem (2) can be solved by the variable substitution $\tilde{\boldsymbol{\beta}}_k = \widehat{\Sigma}_w^{1/2} \boldsymbol{\beta}_k$, where $\widehat{\Sigma}_w^{1/2}$ is the symmetric matrix square root of $\widehat{\Sigma}_w$. The Fisher's discriminant problem was then reduced to a standard eigen-

problem. However, when $p > n$, $\widehat{\Sigma}_w$ becomes singular. Any discriminant vector in the null space of $\widehat{\Sigma}_w$ but not in the null space of $\widehat{\Sigma}_b$ can lead to an arbitrarily large value of the objective function.

To address this singularity problem, some modifications to the Fisher's discriminant problem have been proposed. Krzanowski et al.^[11] have considered a modification of problem (2) by trying to find a unit vector β that maximizes $\beta^T \widehat{\Sigma}_b \beta$ subject to $\beta^T \widehat{\Sigma}_w \beta = 0$. Tebbens and Schlesinger^[12] further required that the solution does not lie in the null space of $\widehat{\Sigma}_b$. It has also been proposed to modify problem (2) using positive definite estimates of Σ_w . For example, Friedman^[5], Dudoit et al.^[4], and Bickel and Levina^[3] consider the use of a diagonal estimate

$$\widetilde{\Sigma}_w = \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2),$$

where $\hat{\sigma}_j^2$ is the j th diagonal element of $\widehat{\Sigma}_w$ in Eq. (1). There are of course some other forms of positive definite estimates for Σ_w suggested in Xu et al.^[6]. Given a positive definite estimate $\widetilde{\Sigma}_w$, the resulting optimization problem is

$$\begin{aligned} & \text{maximize}_{\beta_k} && \beta_k^T \widetilde{\Sigma}_b \beta_k \\ & \text{subject to} && \beta_k^T \widetilde{\Sigma}_w \beta_k \leq 1, \quad \beta_l^T \widetilde{\Sigma}_w \beta_l = 0 \quad \forall l < k. \end{aligned} \quad (3)$$

The new optimization problem (3) addresses the singularity issue but not the interpretability issue. At this point, we extend problem (3) such that the resulting discriminant vector is interpretable. We use Lemma 2.1, which provides a reformulation of problem (3) to obtain the same solution.

Lemma 2.1. The solution $\hat{\beta}_k$ to problem (3) is equivalent to the solution to the following problem:

$$\text{maximize}_{\beta_k} \{\beta_k^T \widehat{\Sigma}_b^k \beta_k\} \quad \text{subject to} \quad \beta_k^T \widetilde{\Sigma}_w \beta_k \leq 1,$$

where

$$\widehat{\Sigma}_b^k = \frac{1}{n} X^T Y (Y^T Y)^{-1/2} P_k^\perp (Y^T Y)^{-1/2} Y^T X.$$

P_k^\perp is defined as follows: $P_1^\perp = I$, and for $k > 1$, P_k^\perp is an orthogonal projection matrix in a space orthogonal to $(Y^T Y)^{-1/2} Y^T X \hat{\beta}_i$ for all $i < k$.

The proof of Lemma 2.1 can be found in Ref. [9]. When we obtain the discriminant vector $\hat{\beta}_1, \dots, \hat{\beta}_k$, we use $\widehat{\Sigma}_b^{k+1}$ to replace $\widehat{\Sigma}_b^k$ and repeat the same procedure for computing $\hat{\beta}_i$ until we obtain all the discriminant vectors.

In this study, we modify problem (3) by imposing ℓ_0 constraint on the discriminant vectors. Given a pre-specified degree of sparsity s , we define the k th sparse discriminant vector $\hat{\beta}_k$ as the solution to the following optimization problem:

$$\text{maximize}_{\beta_k} \{\beta_k^T \widehat{\Sigma}_b^k \beta_k\} \quad \text{subject to} \quad \beta_k^T \widetilde{\Sigma}_w \beta_k = 1, \|\beta_k\|_0 \leq s. \quad (4)$$

2.2 Sparse linear discriminant analysis

Recall that the Lagrangian form of (4) is as follows:

$$\text{minimize}_{\beta \in \mathbb{R}^p} L(\beta) = -\beta^T \widehat{\Sigma}_b^k \beta + \lambda (\beta^T \widetilde{\Sigma}_w \beta - 1) \quad \text{subject to} \quad \|\beta\|_0 \leq s, \quad (5)$$

where $\lambda > 0$ is a parameter that controls the normalization to $\widetilde{\Sigma}_w$ of the discriminant vector β . We denote β^* as a coordinate-

wise minimizer of problem (5). If the active set $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ is known, then by disregarding the constraint $\|\beta\|_0 = s$, we can minimize the objective function $L(\beta)$ without ℓ_0 constraint.

To determine a valid active set \mathcal{A} , we introduce the definition of sacrifice $\Delta = \{\Delta_j : j = 1, \dots, p\}$, which is similar to that in Ref. [10]. Specifically, the j th sacrifice Δ_j measures the increase or decrease in the value of the objective function $L(\beta)$ with the current $\hat{\beta}_j$ set to 0 during each round of iterations. Consider the m th iterative process, by fixing the other coordinates to their current global optimum, the new marginal optimal of β_j is given by $\beta_j^{(m+1)} = \beta_j^{(m)} + d_j^{(m)}$, where $d_j^{(m)} = -\frac{\partial L(\beta)/\partial \beta_j}{\partial^2 L(\beta)/\partial \beta_j^2} \Big|_{\beta=\beta^{(m)}} = (\widehat{\Sigma}_j \beta^{(m)} - \lambda \hat{\sigma}_j^2 \beta_j^{(m)}) / (\lambda \hat{\sigma}_j^2 - \widehat{\Sigma}_{jj})$ denote the dual variable of $\beta_j^{(m)}$, $\widehat{\Sigma}_j$ denote the j th row of $\widehat{\Sigma}_b^k$ and $\widehat{\Sigma}_{jj}$ denote the (j, j) -element of $\widehat{\Sigma}_b^k$. Here, we add the symbol “ (m) ” in the upper-right corner of each variable to indicate that this is the value taken during the m th iteration. Then, the sacrifice Δ_j is defined as $\Delta_j^{(m)} = (\lambda \hat{\sigma}_j^2 - \widehat{\Sigma}_{jj}) (\beta_j^{(m)} + d_j^{(m)})^2$, which indicates an approximation of the loss change. Denote the inactive set $\mathcal{I}^c = \mathcal{A}^c = \{j : \beta_j^* = 0\}$. With the primal-dual condition, we have

- If $j \in \mathcal{A}$, then $\beta_j^* \neq 0, d_j^* = 0$.
- If $j \in \mathcal{I}^c$, then $\beta_j^* = 0, d_j^* = (\widehat{\Sigma}_j \beta^* - \lambda \hat{\sigma}_j^2 \beta_j^*) / (\lambda \hat{\sigma}_j^2 - \widehat{\Sigma}_{jj})$.

The purpose of problem (5) is to minimize the objective function $L(\beta)$, which implies that the sacrifice should be as small as possible. That is, among all candidates, we enforce that the coordinates corresponding to the smaller sacrifice are set to zero. To achieve this, we rearrange $\{\Delta_j, j = 1, \dots, p\}$ in decreasing order; that is, $\Delta_{[1]} \geq \Delta_{[2]} \geq \dots \geq \Delta_{[p]}$ where $\Delta_{[i]}$ denotes the i th largest value among $\{\Delta_j, j = 1, \dots, p\}$. We then truncate the ordered sacrifice vector at position s , that is, set the estimate of the active set $\mathcal{A} = \{j : \Delta_j \geq \Delta_{[s]}\}$.

When \mathcal{A} is given, we can extract the corresponding rows and columns of $\widehat{\Sigma}_b^k$ and $\widetilde{\Sigma}_w$, from which the problem becomes a classical LDA problem. The solution to this problem is then obtained, with $\beta_{\mathcal{A}}$ being the first eigenvector of $(\widetilde{\Sigma}_w^{-1/2} \widehat{\Sigma}_b^k \widetilde{\Sigma}_w^{-1/2})_{\mathcal{A}\mathcal{A}}$ and $\beta_{\mathcal{I}} = \mathbf{0}$. The above discussion is summarized in Algorithm 2.1.

Algorithm 2.1. Naive algorithm for sparse linear discriminant analysis

Input: data matrix X , indicator matrix Y , sparsity s .

Output: $\hat{\beta}_1, \dots, \hat{\beta}_q$.

1: $\widehat{\Sigma}_b = n^{-1} X^T Y (Y^T Y)^{-1} Y^T X$.

2: $\widetilde{\Sigma}_w = X^T X / n - \widehat{\Sigma}_b$ and $\widetilde{\Sigma}_w = \text{diag}(\widetilde{\Sigma}_w)$.

3: **for** $k = 1, \dots, q$ **do**

4: (a) $\widehat{\Sigma}_b^k = n^{-1} X^T Y (Y^T Y)^{-1/2} P_k^\perp (Y^T Y)^{-1/2} Y^T X$.

5: (b) Initialize β_k with normalization $\beta_k^T \widetilde{\Sigma}_w \beta_k = 1$.

6: (c) Iterate until convergence or a maximum number of iteration is reached:

 • $\Delta_j = (\lambda \hat{\sigma}_j^2 - \widehat{\Sigma}_{jj}) (\beta_j + d_j)^2, \mathcal{A} = \{j : \Delta_j \geq \Delta_{[s]}\}$ and $\mathcal{I} = \mathcal{A}^c$.

 • $\beta_{k,\mathcal{A}}$ is the first eigenvector of $(\widetilde{\Sigma}_w^{-1/2} \widehat{\Sigma}_b^k \widetilde{\Sigma}_w^{-1/2})_{\mathcal{A}\mathcal{A}}$ and $\beta_{k,\mathcal{I}} = \mathbf{0}$.

 • Update dual variables $d_j = (\widehat{\Sigma}_j \beta_k - \lambda \hat{\sigma}_j^2 \beta_{k,j}) / (\lambda \hat{\sigma}_j^2 - \widehat{\Sigma}_{jj})$.

7: **end for**

2.3 Best subset selection

Owing to the discontinuity of the ℓ_0 norm, the naive algorithm for sparse linear discriminant analysis may converge to a local minimum and encounter the problem of periodic iterations. To obtain sparse linear discriminant vectors more efficiently, an iterative algorithm based on the primal-dual condition of problem (5) was proposed in this subsection. Motivated by Zhu et al.^[10], who developed a novel algorithm based on exchanging active sets to avoid periodic iterations in solving the best subset selection problem in linear regression models, we extended their work to the sparse LDA problem. An exchanging step is added to Algorithm 2.1 to prevent the algorithm from entering a loop, that is, choosing a subset of the active set to exchange with a subset of the inactive set. We then decide whether to adopt the new candidate solution by comparing the objective function values before and after the exchange.

Specifically, at the m th iteration of computing the k th discriminant vector of Algorithm 2.1, we obtain the sacrifice $\mathcal{A}^{(m)}$. Subsequently, we classified the variables that needed to be exchanged in the active and inactive sets.

- Subset of the active set that need to be exchanged to the inactive set:

$$\mathcal{B}_{c,1}^{(m)} = \{j \in \mathcal{A}^{(m)} : \sum_{i \in \mathcal{A}^{(m)}} I(\Delta_j^{(m)} \geq \Delta_i^{(m)}) \leq c\}.$$

- Subset of the inactive set that need to be exchanged to the active set:

$$\mathcal{B}_{c,2}^{(m)} = \{j \in \mathcal{I}^{(m)} : \sum_{i \in \mathcal{I}^{(m)}} I(\Delta_j^{(m)} \leq \Delta_i^{(m)}) \leq c\},$$

where c is an arbitrary constant ranging from 1 to $\min\{s, p-s\}$.

Then, we updated the active and inactive sets using $\mathcal{A}_c^{(m+1)} = (\mathcal{A}^{(m)} / \mathcal{B}_{c,1}^{(m)}) \cup \mathcal{B}_{c,2}^{(m)}$ and $\mathcal{I}_c^{(m+1)} = (\mathcal{I}^{(m)} / \mathcal{B}_{c,2}^{(m)}) \cup \mathcal{B}_{c,1}^{(m)}$. Given $\mathcal{A}_c^{(m)}$ and $\mathcal{I}_c^{(m)}$, we define $\tilde{\beta}_c^{(m)}$ as the solution to problem (5), which is given by

$$\tilde{\beta}_c^{(m)} = \arg \max_{\beta_{\mathcal{I}^{(m)}} = 0, \beta^\top \tilde{\Sigma}_w \beta = 1} \beta^\top \tilde{\Sigma}_b \beta.$$

Further, for accuracy, we search c over 1 to $\min\{s, p-s\}$ and determine an optimal sparse vector with the smallest value of $L(\beta)$, that is, $\hat{c} = \arg \max_c L(\tilde{\beta}_c^{(m)})$. Algorithm 2.2 displays the

Algorithm 2.2. Exchange step

Input: $\tilde{\Sigma}_b^k, \tilde{\Sigma}_w, \lambda$, sacrifice \mathcal{A} , active set \mathcal{A} , inactive set \mathcal{I} , sparsity s .

Output: discriminant vector $\hat{\beta}$, dual variables \mathbf{d} and sacrifice \mathcal{A} .

- 1: $L = -\beta^\top \tilde{\Sigma}_b^k \beta + \lambda(\beta^\top \tilde{\Sigma}_w \beta - 1)$.
 - 2: **for** $c = 1, \dots, \min\{s, p-s\}$ **do**
 - 3: $\mathcal{B}_{c,1} = \{j \in \mathcal{A} : \sum_{i \in \mathcal{A}} I(\Delta_j \geq \Delta_i) \leq c\}$ and
 $\mathcal{B}_{c,2} = \{j \in \mathcal{I} : \sum_{i \in \mathcal{I}} I(\Delta_j \leq \Delta_i) \leq c\}$.
 - 4: $\mathcal{A}_c = (\mathcal{A} / \mathcal{B}_{c,1}) \cup \mathcal{B}_{c,2}$ and $\mathcal{I}_c = (\mathcal{I} / \mathcal{B}_{c,2}) \cup \mathcal{B}_{c,1}$.
 - 5: Calculate $\tilde{\beta}_c$ based on \mathcal{A}_c , and $\tilde{L} = -\tilde{\beta}_c^\top \tilde{\Sigma}_b^k \tilde{\beta}_c + \lambda(\tilde{\beta}_c^\top \tilde{\Sigma}_w \tilde{\beta}_c - 1)$.
 - 6: If $L > \tilde{L}$, then $L \leftarrow \tilde{L}$ and $(\mathcal{A}, \mathcal{I}, \beta) \leftarrow (\mathcal{A}_c, \mathcal{I}_c, \tilde{\beta}_c)$.
 - 7: **end for**
 - 8: Update dual variables \mathbf{d} and sacrifice \mathcal{A} .
-

Algorithm 2.3. SLDA-BESS for sparse LDA

Input: data matrix \mathbf{X} , indicator matrix \mathbf{Y} , sparsity s, λ .

Output: β_1, \dots, β_q .

- 1: $\tilde{\Sigma}_b = n^{-1} \mathbf{X}^\top \mathbf{Y} (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{X}$, $\tilde{\Sigma}_w = \mathbf{X}^\top \mathbf{X} - \tilde{\Sigma}_b$ and $\tilde{\Sigma}_w = \text{diag}(\tilde{\Sigma}_w)$.
 - 2: Initialize $\beta^{(0)}$ with normalization $\beta^{(0)\top} \tilde{\Sigma}_w \beta^{(0)} = 1$.
 - 3: Calculate $d_j^{(0)} = (\tilde{\Sigma}_j \beta^{(0)} - \lambda \hat{\sigma}_j^2 \beta_j^{(0)}) / (\lambda \hat{\sigma}_j^2 - \tilde{\Sigma}_{jj})$.
 - 4: **for** $m = 0, \dots, M$ **do**
 - 5: $\Delta_j = (\lambda \hat{\sigma}_j^2 - \tilde{\Sigma}_{jj})(\beta_j^{(m)} + d_j^{(m)})^2$, $\mathcal{A} = \{j : \Delta_j \geq \Delta_{[s]}\}$ and $\mathcal{I} = \mathcal{A}^c$.
 - 6: Set $\beta_{\mathcal{A}}^{(m+1)}$ be the first eigenvector of $(\tilde{\Sigma}_w^{-1/2} \tilde{\Sigma}_b^k \tilde{\Sigma}_w^{-1/2})_{\mathcal{A}\mathcal{A}}$ and $\beta_{\mathcal{I}}^{(m+1)} = \mathbf{0}$.
 - 7: Update $(\beta^{(m+1)}, \mathbf{d}^{(m+1)})$ by Algorithm 2.2 with $\beta^{(m+1)}, \mathcal{A}^{(m+1)}, \mathcal{A}$ and \mathcal{I} .
 - 8: When $\|\beta^{(m+1)} - \beta^{(m)}\|_2$ is sufficiently small, break.
 - 9: **end for**
 - 10: Repeat the above steps to achieve the rest discriminant vectors.
-

process of exchanging some variables in the active and inactive sets.

Based on the exchange step proposed above, we developed an efficient algorithm to avoid falling into a local minimum and guarantee convergence. Specifically, we added an exchange step after updating the primal and dual variables in Algorithm 2.1. The details are presented in Algorithm 2.3, which is called SLDA-BESS.

The improved Algorithm 2.3 has the theoretical guarantee of Theorem 2.1, which indicates that such a best subset selection algorithm incorporating the exchange step is feasible.

Theorem 2.1. The SLDA-BESS algorithm terminates after a finite number of iterations.

Proof. The objective function always increases, and the choice of the active set is finite. Algorithm 2.3 terminates after a finite number of iterations.

3 Numeric studies

We compared SLDA-BESS with penalized LDA- ℓ_1 ^[9], nearest shrunken centroids (NSC)^[7], and shrunken centroids regularized discriminant analysis (RDA)^[8] in simulation studies. LDA- ℓ_1 is a method that adds a ℓ_1 penalty to the objective function $\beta^\top \Sigma \beta$. NSC is a simple modified version of the nearest centroid method that divides a between-class standard deviation when calculating the centroid distance and is a modified version based on NSC. In each simulation, 1200 observations were set to belong equally to several different classes. Arbitrary 300 of these 1200 observations are set as the training set, and the remaining 900 belong to the test set. Each simulation consisted of measurements of 500 features, i.e., $p = 500$.

Simulation 1. Consider four different classes where the features are independent of each other and the mean value is shifted. Given the set of indicators for four classes, C_1, C_2, C_3 and C_4 , $x_i \sim N(\mu_k, \mathbf{I})$ if $i \in C_k$, where $\mu_{1,j} = 0.7$ for $1 \leq j \leq 25$, $\mu_{2,j} = 0.7$ for $26 \leq j \leq 50$, $\mu_{3,j} = 0.7$ for $51 \leq j \leq 75$, $\mu_{4,j} = 0.7$ for $76 \leq j \leq 100$ and $\mu_{k,j} = 0$ otherwise for $j = 1, \dots, 500$.

Simulation 2. Consider two different classes where the features are dependent of each other and the mean value is

shifted. Given the set of indicators for two classes, C_1 and C_2 , $x_i \sim \mathcal{N}(0, \Sigma)$ for $i \in C_1$ and $x_i \sim \mathcal{N}(\mu, \Sigma)$ for $i \in C_2$, where $\mu_j = 0.6$ if $j \leq 100$ and $\mu_j = 0$ otherwise and $\Sigma = (\Sigma_{ij}) = (0.6^{|i-j|})$. The covariance structure Σ is intended to mimic gene expression data, in which genes are positively correlated within a pathway and independent between pathways.

Simulation 3. Consider four different classes in which the features are independent of each other and have only one dimension, where the mean is shifted. Given four indicator sets, C_1, C_2, C_3 and C_4 , for $i \in C_k, x_{ij} \sim \mathcal{N}((k-1)/3, 1)$ if $j \leq 100$ and $x_{ij} \sim \mathcal{N}(0, 1)$ otherwise. The one-dimensional projection of the data fully captures the structure of the class.

For each method, the models were fitted to the training set using a range of values for the tuning parameters. Tuning parameter values were chosen to minimize errors in the validation set. The parameters estimated for the training set were then evaluated for the test set. After obtaining the estimated discriminant vectors, we applied the KNN method to the new dataset to perform classification. This process was repeated 200 times.

The classification error in the test set and the number of non-zero features used by the discriminant vectors are listed in Table 1. In the “errors” row of each simulation result, the average number of misclassified individuals on the test set with 900 observations is presented, and the standard deviation is in parentheses. Correspondingly, in the “features” row, the average number of non-zero features used in estimating the discriminant vectors is reported, with the standard deviation in parentheses. As shown in Table 1, our method has a smaller error when the number of features used is almost equal. Moreover, it is easier to adjust the parameters using the ℓ_0 penalty method.

4 Real data analysis

This section compares our SLDA-BESS method with three existing methods: penalized LDA- ℓ_1 , NSC, and RDA.

Table 1. Simulation result.

Simulation		SLDA-BESS	Penalized LDA- ℓ_1	NSC	RDA
1	errors	57.90(10.69)	81.73(13.05)	56.44(18.93)	66.38(12.30)
	features	95.76(4.19)	98.35(3.56)	95.91(3.52)	95.30(6.38)
2	error	46.62(9.09)	66.43(13.61)	56.00(8.75)	81.82(17.59)
	features	100.00(0.00)	111.5(7.02)	107.11(7.53)	107.37(10.80)
3	errors	101.62(13.36)	115.12(15.53)	358.64(26.63)	318.18(14.98)
	features	100.00(0.00)	99.79(0.71)	102.01(1.87)	107.99(14.98)

Table 2. Results obtained on three different gene expression datasets.

		SLDA-BESS	Penalized LDA- ℓ_1	NSC	RDA
Ramaswamy	errors	13.00(1.82)	15.90(2.52)	11.27(2.68)	10.50(2.79)
	features	429.52(10.56)	9426.00(573.78)	2629.66(321.19)	927.16(108.64)
Nakayamy	errors	4.86(1.74)	4.98(1.62)	4.84(1.51)	5.40(1.72)
	features	213.18(5.75)	11044.88(376.34)	1517.92(163.49)	2858.82(278.98)
Sun	errors	14.80(2.58)	16.06(2.29)	13.24(3.34)	24.14(3.39)
	features	215.24(14.03)	20715.10(2396.99)	14453.12(1280.76)	4740.70(670.81)

We applied the four methods to the three gene expression datasets.

Ramaswamy dataset^[13]: A dataset consisting of 16063 gene expression measurements and 198 samples belonging to 14 distinct cancer subtypes.

Nakayama dataset^[14]: A dataset consisting of 105 samples from 10 types of soft tissue tumors, with 22283 gene expression measurements per sample. We shall limit the analysis to five tumor types with at least 15 samples in the data, resulting in a subset of data containing 86 samples.

Sun dataset^[15]: A dataset consisting of 180 samples and 54613 expression measurements. The sample falls into four classes: one non-tumor class and three glioma classes.

Each dataset is split into a training set containing 75% of the samples and a test set containing the remaining 25% of the samples. A total of 100 replications were performed, each with randomly selected training and testing sets. The results are presented in Table 2. The results suggest that the four methods perform roughly equally in terms of error, but our SLDA-BESS method utilizes fewer features. Moreover, our SLDA-BESS method is more suitable for remarkably sparse models, whereas the other three methods may fail when using a few features.

5 Conclusions

Linear discriminant analysis is a commonly used classification method. However, it fails if the number of features is large relative to the number of observations. This study extended LDA to a high-dimensional setting by adding an ℓ_0 penalty to produce a discriminant vector involving only a subset of features. Our extension is based on Fisher’s discriminant problem, such as a generalized eigen problem, and then uses the SLDA-BESS algorithm, which combines the naive algorithm and the exchange step to solve the sparse discriminant vector. Sparse discriminant vectors were generated while

making the classifier more interpretable in practical situations. The results of both numerical simulations and analysis of real data demonstrate the superior performance of our SLDA-BESS method. The convergence rate and complexity of the SLDA-BESS algorithm, as well as the theoretical properties of the minimax lower and upper bounds of the estimator, can be further considered in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (71771203).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Qi Yin is currently a graduate student at the University of Science and Technology of China. His research interests focus on variable selection.

Lei Shu is currently a Ph.D. student at the University of Science and Technology of China. His research interests focus on high-dimensional statistical inference, including variable selection, change point detection, and factor analysis.

References

- [1] Hastie T, Tibshirani R, Friedman J H, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Berlin: Springer, 2009.
- [2] Shao J, Wang Y, Deng X, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, **2011**, 39 (2): 1241–1265.
- [3] Bickel P, Levina E. Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **2004**, 10 (6): 989–1010.
- [4] Dudoit S, Fridlyand J, Speed T P. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **2002**, 97 (457): 77–87.
- [5] Friedman J H. Regularized discriminant analysis. *Journal of the American Statistical Association*, **1989**, 84 (405): 165–175.
- [6] Xu P, Brock G N, Parrish R S. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis*, **2009**, 53 (5): 1674–1687.
- [7] Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **2002**, 99 (10): 6567–6572.
- [8] Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **2007**, 8 (1): 86–100.
- [9] Witten D M, Tibshirani R. Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **2011**, 73 (5): 753–772.
- [10] Zhu J, Wen C, Zhu J, et al. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences of the United States of America*, **2020**, 117 (52): 33117–33123.
- [11] Krzanowski W, Jonathan P, McCarthy W, et al. Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **1995**, 44 (1): 101–115.
- [12] Tebbens J D, Schlesinger P. Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. *Computational Statistics & Data Analysis*, **2007**, 52 (1): 423–437.
- [13] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, **2001**, 98 (26): 15149–15154.
- [14] Nakayama R, Nemoto T, Takahashi H, et al. Gene expression analysis of soft tissue sarcomas: Characterization and reclassification of malignant fibrous histiocytoma. *Modern Pathology*, **2007**, 20 (7): 749–759.
- [15] Sun L, Hui A M, Su Q, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, **2006**, 9 (4): 287–300.