Statistica Sinica Preprint No: SS-2015-0387R2	
Title	A New Reduced-Rank Linear Discriminant Analysis Method and Its Applications
Manuscript ID	SS-2015-0387R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0387
Complete List of Authors	Yue Selena Niu Ning Hao and Bin Dong
Corresponding Author	Ning Hao
E-mail	nhao@math.arizona.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

A New Reduced-Rank Linear Discriminant Analysis Method and Its Applications

Yue Selena Niu, Ning Hao, and Bin Dong

University of Arizona, University of Arizona and Peking University

Abstract: We consider multi-class classification problems for high-dimensional data. Following the idea of reduced-rank linear discriminant analysis (LDA), we introduce a new dimension reduction tool with a flavor of supervised principal component analysis (PCA). The proposed method is computationally efficient and can incorporate the correlation structure among the features. Besides the theoretical insights, we show that our method is a competitive classification tool by simulated and real data examples.

Key words and phrases: Dimension reduction, Gene expression data, High-dimensional data, Multi-class classification, Supervised principal component analysis.

1. Introduction

Targeting on cancer classification and other modern applications, many high-dimensional classification techniques have been studied recently; see Hastie, Tibshirani and Friedman (2009) for an extensive introduction, and Witten and Tibshirani (2011), Cai and Liu (2011), Fan, Feng and Tong (2012), and Mai, Zou

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

and Yuan (2012) for some recent developments. Although these contemporary classification tools can be applied to high-dimensional data, most of them rely on strong assumptions. For example, many methods assume that the features are independent of each other; other methods assume sparsity conditions. These assumptions make the model simple and robust against growing dimensionality, so classification accuracy and computational efficiency can be achieved. However, they may be too restrictive, and when violated, lead to information loss in data analysis. Moreover, many methods target the binary classification case and are not straightforward to use if more than two classes are present. Convenient and efficient classification tools for multi-class data are quite limited. Therefore, it is desirable to develop new classification techniques that can handle high-dimensional, multi-class data, and also take into account the correlation among the features.

Many linear classification rules depend on the Mahalanobis distance. But it cannot be well-estimated for high-dimensional data when the number of features is greater than the sample size, as the sample covariance is singular. Under the assumption that features are independent, the sample covariance matrix is diagonal and strictly positive definite, so the Mahalanobis distance can be calculated. That is one of the main reasons that the independence assumption is crucial in many classification methods. For example, the nearest shrunken centroids (NSC,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

Tibshirani et al. (2002)), independence rule (IR, Bickel and Levina (2004)), features annealed independence rule (FAIR, Fan and Fan (2008)) all assume that the features are independent of each other. Moreover, some other methods such as regularized discriminant analysis (RDA, Guo, Hastie and Tibshirani (2007)) use a covariance estimator the sample covariance regularized towards a diagonal matrix. Recently, new classification tools have been developed, including penalized linear discriminant analysis (PLDA, Witten and Tibshirani (2011)), linear programming discriminant rule (LPD, Cai and Liu (2011)), regularized optimal affine discriminant rule (ROAD, Fan, Feng and Tong (2012)), direct sparse discriminant analysis (DSDA, Mai, Zou and Yuan (2012)), sparse discriminant analysis (SDA, Clemmensen et al. (2011)), multi-class sparse discriminant analysis (MSDA, Mai, Yang and Zou (2015)). Roughly speaking, these sparse methods obtain sparse models by solving penalized or constrained optimization problems, and their efficiency relies on the sparsity level of the normal vectors to the optimal discriminant boundaries.

Reduced-rank LDA is a classical approach to classification. It conducts dimension reduction by projecting the data to the centroid-spanning space and classifies the data based on nearest centroid. Another commonly used dimension reduction tool is PCA, which projects the data to the space spanned by the top principal components of the total sample covariance matrix. Reduced-rank LDA

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

makes use of the label information (through centroids) but ignores the (within class) covariance information. On the other hand, PCA relies on the covariance information only and is mainly regarded as an unsupervised learning tool.

We propose a new reduced-rank LDA method combining the advantages of the classical reduced-rank LDA and PCA. The principal components of a weighted sum of the sample within class and between class covariance matrices are used for dimension reduction, and standard LDA is employed to the projected data for classification. In this dimension reduction process, both label and covariance information can be taken into account, through between class and within class covariance, respectively. We regard it as a version of supervised PCA. This method does not rely on the aforementioned sparsity or independence assumptions and offers an alternative classification tool for various applications.

For insight on our method, we consider spiked structure of the covariance (Johnstone (2001)). Roughly speaking, a symmetric positive definite matrix is called spiked if all of its eigenvalues are equal except for a few large ones. In other words, it is a sum of a scalar matrix and a low rank matrix. Intuitively, the spiked structure might be a better model than the diagonal one to approximate the true covariance, as it can take into account strong correlation among the features, which is not uncommon in applications.

We propose here a novel dimension reduction and classification tool that in-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

corporates covariance among features and works well for high-dimensional multi-class data. It illustrates a new supervised way to conduct PCA and is generally applicable to both classification and regression models. Importantly, the proposed method is computationally efficient and can be applied directly to such data as gene expression data in cancer research.

The rest of the paper is organized as follows. Section 2 introduces notation and reviews some linear classification methods from the perspective of dimension reduction. In Section 3, we study a new reduced-rank LDA method for classification and offer some insights about it. Numerical studies and data applications are illustrated in Section 4, followed by a short discussion in Section 5. Proofs are given in the Supplementary Material.

2. Linear methods for classification

2.1 Notation

We consider a standard setup for classification. Let $\mathcal{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^\top$ be a $n \times p$ matrix with each $\mathbf{X}_i$ is a p-dimensional vector. Let $\mathbf{Y} = (Y_1, ..., Y_n)^\top$ be a response vector with $Y_i \in \{1, ..., K\}$, $1 \leq i \leq n$, with the interpretation that $\mathbf{X}_i$ belongs to group k if $Y_i = k$. Denote the index set of group k by $C_k = \{i : Y_i = k\}$ and its cardinality by $n_k = |C_k|$, where $1 \leq k \leq K$. The goal of classification is to establish a classification rule that labels a new observation $\mathbf{X}^*$ based on training data.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

The Gaussian assumption is often used to facilitate statistical analysis of various methods. In the simplest setting, the data from all groups share a common covariance matrix $\mathbf{\Sigma}_w$, $(\mathbf{X}|Y = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_w)$, $1 \leq k \leq K$. For easy presentation, we assume that the prior probabilities $\pi_k = \mathbf{P}(Y = k)$ are equal for all k. In practice, the prior probability can be estimated and taken into account for most methods considered in this paper.

For a specified strictly positive definite symmetric matrix $\mathbf{S}$, the Mahalanobis distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$d_M(\mathbf{u}, \mathbf{v}) = \left\{(\mathbf{u} - \mathbf{v})^\top \mathbf{S}^{-1}(\mathbf{u} - \mathbf{v})\right\}^{\frac{1}{2}}.$$

Under the Gaussian assumption, the classification rule minimizing the expected classification error is called the Bayes rule, which simply classifies a data point to a group with the nearest centroid in terms of Mahalanobis distance with $\mathbf{S} = \mathbf{\Sigma}_w$,

$$Y = \underset{1\leq k\leq K}{\operatorname{argmin}} (\mathbf{X} - \boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_w^{-1}(\mathbf{X} - \boldsymbol{\mu}_k) = \underset{1\leq k\leq K}{\operatorname{argmin}} \left\|\mathbf{\Sigma}_w^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_k)\right\|_2^2. \tag{2.1}$$

A key observation from (2.1) is that if we rotate the sample space by $\mathbf{\Sigma}_w^{-1/2}$ first, then the Bayes rule is equivalent to a nearest centroid classifier with standard Euclidian distance. It then follows that the optimal decision boundary separating groups k and ℓ is the affine space given by $\{\mathbf{X} - (\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell)/2\}\mathbf{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

0. The normal vector to this affine space is $\mathbf{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$. Therefore, the decision boundary of the Bayes rule to the whole classification problem is a subset of the union of these affine spaces, whose normal vectors span a vector space $\mathbf{\Sigma}_w^{-1}\mathbf{C}$, where $\mathbf{C}$ is the vector space spanned by $\{\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell\}_{1\le k<\ell\le K}$. Here $\dim \mathbf{C} = \dim(\text{span}\{\boldsymbol{\mu}_k - \boldsymbol{\mu}_K\}_{k=1}^{K-1}) \le K-1$, with equality when the set of centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$ is in general linear position. By Lemma 2 in the Supplementary Material, when p is larger than K, we lose no information in projecting the data from $\mathbb{R}^p$ to a small subspace $\mathbf{\Sigma}_w^{-1}\mathbf{C}$ for classification. Applying the Bayes rule to the projected data and the original data are equivalent. In practice, when p is large, it is extremely helpful to find a reasonable approximation subspace to $\mathbf{\Sigma}_w^{-1}\mathbf{C}$ to reduce the dimensionality.

Without loss of generality, we assume that the columns of $\mathcal{X}$ are centered to have mean zero as the methods considered here are translation invariant. The within-class sample covariance matrix is

$$\mathbf{W} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)^\top,$$

where $\hat{\boldsymbol{\mu}}_k = n_k^{-1}\sum_{i\in C_k}\mathbf{X}_i$. The between-class sample covariance matrix is

$$\mathbf{B} = \frac{1}{n}\sum_{k=1}^{K} n_k(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^\top = \frac{1}{n}\sum_{k=1}^{K} n_k\hat{\boldsymbol{\mu}}_k\hat{\boldsymbol{\mu}}_k^\top,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

where $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{k=1}^{K} n_k\hat{\boldsymbol{\mu}}_k = 0$. The total sample covariance matrix is

$$\mathbf{T} = n^{-1}\mathcal{X}^\top\mathcal{X} = \mathbf{W} + \mathbf{B}.$$

2.2 Simple reduced-rank linear discriminant analysis

A reduced-rank LDA (Hastie, Tibshirani and Friedman (2009)) projects the data to the centroid-spanning subspace $\hat{\mathbf{C}} = \text{span}\{\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_\ell\}_{1\le k<\ell\le K}$. The idea is that, when calculating the (Euclidean) distances to find the closest centroid, one can ignore the distances orthogonal to $\hat{\mathbf{C}}$ which contribute equally to all groups. This simple method reduces the dimensionality remarkably. The main drawback is that it does not incorporate the covariance structure and one can lose much information if $\boldsymbol{\Sigma}_w$ is far from a scalar matrix.

2.3 Fisher's approach and the standard LDA

Fisher's approach is to find a subspace so the projected centroids are spread out as much as possible with respect to the covariance. It finds the first direction by solving

$$\mathbf{v}_1 = \underset{\mathbf{v}}{\text{argmax}}\, \mathbf{v}^\top\mathbf{B}\mathbf{v} \text{ subject to } \mathbf{v}^\top\mathbf{W}\mathbf{v} = 1, \tag{2.2}$$

provided $\mathbf{W}$ is not singular. When $K = 2$, $\mathbf{v}_1$ is the same as the normal vector (up to a scalar) of the decision boundary separating two groups obtained by standard

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

LDA. When $K > 2$, one can continue to solve this generalized eigenvalue problem until step $\text{rank}(\mathbf{B})$, as

$$\begin{aligned}
\mathbf{v}_2 &= \underset{\mathbf{v}}{\text{argmax}}\, \mathbf{v}^\top \mathbf{B}\mathbf{v} \text{ subject to } \mathbf{v}^\top \mathbf{W}\mathbf{v} = 1;\ \mathbf{v}^\top \mathbf{W}\mathbf{v}_1 = 0, \\
&\cdots \\
\mathbf{v}_k &= \underset{\mathbf{v}}{\text{argmax}}\, \mathbf{v}^\top \mathbf{B}\mathbf{v} \text{ subject to } \mathbf{v}^\top \mathbf{W}\mathbf{v} = 1;\ \mathbf{v}^\top \mathbf{W}\mathbf{v}_\ell = 0,\ \ell < k, \quad (2.3) \\
&\cdots
\end{aligned}$$

The covariance plays a role here through the sample pooled covariance $\mathbf{W}$ and the dimension of the subspace can be pre-specified or chosen data-adaptively.

The standard LDA can be viewed as a dimension reduction technique. Roughly speaking, it mimics the Bayes rule by plugging in estimators of the common covariance and centroids. It labels an observation $\mathbf{X}$ by $\hat{Y} = \text{argmin}_{1 \le k \le K} \|\mathbf{W}^{-1/2}(\mathbf{X} - \hat{\boldsymbol{\mu}}_k)\|_2$. Similar to the analysis of the Bayes rule, the normal vectors of the decision boundaries of standard LDA span a subspace $\mathbf{W}^{-1}\hat{\mathbf{C}} \subset \mathbb{R}^p$. It is equivalent to apply standard LDA to, instead of the original data, the projected data onto subspace $\mathbf{W}^{-1}\hat{\mathbf{C}}$.

Proposition 1 *If* $\mathbf{W}$ *is nonsingular, then* $\dim \mathbf{W}^{-1}\hat{\mathbf{C}} = \text{rank}(\mathbf{B})$, *and* $\mathbf{W}^{-1}\hat{\mathbf{C}} = \text{span}\{\mathbf{v}_k\}_{k=1}^r$ *where* $r = \dim \hat{\mathbf{C}}$, $\mathbf{v}_k$ *is as defined in (2.3).*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

The standard LDA performs well only when the sample size is large enough so $\mathbf{W}^{-1}\hat{\mathbf{C}}$ is a good approximation to $\mathbf{\Sigma}_w^{-1}\mathbf{C}$.

2.4 The independence rule and related approaches

LDA does not work well when $p \sim n$ and $p > n$. In the context of dimension reduction, the reason is that $\mathbf{W}^{-1}\hat{\mathbf{C}}$ or $\mathbf{W}^{-}\hat{\mathbf{C}}$ is no longer a good approximation to $\mathbf{\Sigma}_w^{-1}\mathbf{C}$ for high-dimensional data, where $\mathbf{W}^{-}$ is a pseudo-inverse of $\mathbf{W}$. A remedy is to assume that the features are independent, which leads to the independence rule or diagonal LDA. To apply diagonal LDA, one just uses the diagonal part $\hat{\mathbf{D}}_w = diag(\mathbf{W})$ instead of $\mathbf{W}$ in the standard LDA. The IR or diagonal LDA usually outperforms standard LDA when $p > n$ (Bickel and Levina (2004)).

In the spirit of Proposition 1, one can see the equivalence between the IR and Fisher's approach with $\mathbf{W}$ replaced by $\hat{\mathbf{D}}_w$ in (2.3), as stated in Corollary 1 in the Supplementary Material. Witten and Tibshirani (2011) imposed some sparse assumptions on the $\mathbf{v}_k$'s to derive a penalized LDA (PLDA). One can also conduct dimension reduction based on the rank of marginal discriminant power. Two well-known approaches are the NSC (Tibshirani et al. (2002)) and the FAIR (Fan and Fan (2008)).

2.5 Principal component analysis

PCA has been used to solve supervised learning problems, e.g., principal

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

component regression (Jolliffe (2002)), supervised PCA (Bair et al. (2006)), etc. In our context, standard PCA ignores the label information and keeps the eigenvectors corresponding to q top eigenvalues of $\mathbf{T}$, where q can be pre-specified or chosen data adaptively. There is no guarantee that the top principal components have good discrimination power. Bair et al. (2006) proposed a variant of supervised PCA, a two-stage procedure in which marginal statistics are used to reduce dimension before applying standard PCA. It seems that the label information is used only in the first stage.

3. A New Reduced-Rank Linear Discriminant Analysis Method

3.1 Method

To take advantage of existing methods and study the multi-class classification problem in a unified manner, we consider $\mathbf{T}_\gamma = \mathbf{W} + \gamma\mathbf{B}$ with $\gamma > 0$, with $\mathbf{W}$ and $\mathbf{B}$ the within class and between class sample covariance matrices, respectively, and γ is a tuning parameter. If $\gamma = 1$, our proposed procedure is equivalent to the standard PCA; and if $\gamma \to \infty$, the procedure is equivalent to the simple reduced-rank LDA.

Consider the eigenvalue decomposition

$$\mathbf{U}_\gamma^\top \mathbf{T}_\gamma \mathbf{U}_\gamma = \mathbf{D}_\gamma \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

where $\mathbf{D}_\gamma$ is a diagonal matrix with diagonal entries ranked in a descending order and $\mathbf{U}_\gamma$ is an orthogonal matrix. Our reduced-rank LDA procedure based on the first q principal components of $\mathbf{T}_\gamma$ is carried out as follows.

1. Calculate $\mathbf{T}_\gamma$ and $\mathbf{U}_\gamma$ and project the data from $\mathbb{R}^p$ to the linear subspace spanned by the first q columns of $\mathbf{U}_\gamma$.

2. Apply the standard LDA to the projected data.

This procedure allows a varying parameter γ which, along with q, can be chosen adaptively. The label information is taken into account through γ in dimension reduction, and so we call it supervised PCA-based LDA (SPCALDA). The parameter γ makes our procedure more flexible. For example, the qualities of $\mathbf{W}$ and $\mathbf{B}$ to approximate their population counterparts are usually not equally good, and γ can serve as a weight to balance them.

3.2 Theory

To understand the proposed method, we consider a population version. Denote by $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_t$ the population versions of between-class and total covariance matrix, respectively. Take $\mathbf{\Sigma}_\gamma = \mathbf{\Sigma}_w + \gamma\mathbf{\Sigma}_b$, $\gamma > 0$ with eigenvalue decomposition

$$\mathbf{U}_O^\top \mathbf{\Sigma}_\gamma \mathbf{U}_O = \mathbf{D}_O,$$

where $\mathbf{D}_O$ is a diagonal matrix with diagonal entries ranked in a descending order

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

and $\mathbf{U}_O$ is orthogonal. Because γ plays only a minor role here, we drop it from $\mathbf{U}_O$, $\mathbf{D}_O$, etc. We ask when we can project the data by oracle procedure without information loss.

Let $\{\lambda_j\}_{j=1}^p$ be eigenvalues of $\mathbf{\Sigma}_w$ in a descending order, and consider a spiked covariance structure (Johnstone (2001)).

Spiked Condition: $\lambda_1 \geq \cdots \geq \lambda_s > \lambda_{s+1} = \cdots = \lambda_p$ for some integer $s < p$.

Theorem 1 *Suppose $p > s+K-1$, $s > 1$, with $\mathbf{U}_O = (\mathbf{U}_{O1}, \mathbf{U}_{O2})$, $\mathbf{U}_{O1}$ and $\mathbf{U}_{O2}$ $p \times (s+K-1)$ and $p \times (p-s-K+1)$ matrices, respectively. Under spiked condition, we have*

$$\mathbf{U}_{O2}^\top \mathbf{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) = 0, \text{ for all } 1 \leq k < \ell \leq K.$$

Thus we lose no discriminant power by projecting the data to a $s+K-1$ dimensional subspace spanned by the columns of $\mathbf{U}_{O1}$. We can generalize this result further as follows. Without loss of generality, assume $\boldsymbol{\mu} = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k = 0$. From the proof of Theorem 1 in the Supplementary Material, we see the conclusion of Theorem 1 holds for principal components of $\mathbf{\Sigma}_{\boldsymbol{\rho}} = \mathbf{\Sigma}_w + \sum_{k=1}^K \rho_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$, where $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)^\top$ with $\rho_k > 0$, $1 \leq k \leq K$. Still, when K is more than three, it is complicated to tune K parameters.

A more general model than the Gaussian is the mixture Gaussian model

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

that allows each group to be distributed as mixture Gaussian with the same covariance; see e.g. Hastie, Tibshirani and Friedman (2009), Section 12.7. Let $(\mathbf{X}|Y=k) \sim \sum_{t=1}^{R_k} \pi_{kt}\mathcal{N}(\boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_w)$, where $1 \le k \le K$, $1 \le t \le R_k$, $\sum_{t=1}^{R_k} \pi_{kt} = 1$.

Theorem 2 *Let $R = \sum_{k=1}^{K} R_k$. Then Theorem 1 holds if we replace each K by R.*

Remark 1. The spiked condition is crucial in Theorems 1 and 2. It is employed by Hao, Dong and Fan (2015), which aimed to sparsify the normal vector of optimal discriminant boundary for binary classification problems. In applications, the spiked condition may not hold exactly, but our numerical studies show that the procedure performs very well.

Remark 2. In practice, we work with $\mathbf{U}$ instead of its population version $\mathbf{U}_O$. Although $\mathbf{U}_O$ may be quit different from $\mathbf{U}$ when $n \ll p$, $\mathbf{U}_1$ can be similar to $\mathbf{U}_{O1}$ under some conditions. For example, when the leading eigenvalues are large enough or their corresponding eigenvectors are sparse, $\mathbf{U}_{O1}$ can be well-estimated by $\mathbf{U}_1$ or its sparse counterpart (Johnstone and Lu (2009)).

3.3 Computation

In many applications, p is much larger than n. For example, in some gene expression data sets, p is a few thousands or more, and n is a few hundreds or less. So it is time-consuming to calculate the $p \times p$ matrix $\mathbf{T}_\gamma$ and its eigenvalue

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

decomposition directly. The following lemma offers a shortcut to finding $\mathbf{U}_1$.

Lemma 1 *We can write* $\mathbf{T}_\gamma = n^{-1}\mathbf{A}_\gamma^\top \mathbf{A}_\gamma$, *where*

$$\mathbf{A}_\gamma = \Big(\mathbf{X}_1 - \hat{\boldsymbol{\mu}}_{Y_1}, ..., \mathbf{X}_n - \hat{\boldsymbol{\mu}}_{Y_n}, (\gamma n_1)^{1/2}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}), ..., (\gamma n_G)^{1/2}(\hat{\boldsymbol{\mu}}_K - \hat{\boldsymbol{\mu}})\Big)^\top$$

is an $(n+K)\times p$ *matrix. Note that* $\hat{\boldsymbol{\mu}}_{Y_i} = \hat{\boldsymbol{\mu}}_k$ *when* $Y_i = k$, $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{k=1}^K n_k\hat{\boldsymbol{\mu}}_k = 0$ *by our convention.*

When $p > n + K$ we can conduct the eigenvalue decomposition for the $(n + K)\times(n+K)$ matrix $\mathbf{A}_\gamma\mathbf{A}_\gamma^\top$ instead of the $p\times p$ matrix $\mathbf{A}_\gamma^\top\mathbf{A}_\gamma$. Thus, by singular decomposition, $\mathbf{A}_\gamma = \mathbf{V}_\gamma\boldsymbol{\Gamma}\mathbf{U}_\gamma^\top$ where $\boldsymbol{\Gamma}$ is diagonal, $\mathbf{V}_\gamma$ is an $(n + K) \times (n + K)$ orthogonal matrix, and $\mathbf{U}_\gamma$ is $p \times p$ orthogonal matrix identical to $\mathbf{U}_\gamma$ in (3.1). For $n + K$ small or moderate, it is easy to find $\mathbf{V}_\gamma$ which consists of eigenvectors of $\mathbf{A}_\gamma\mathbf{A}_\gamma^\top$, and the first $(n+K)$ columns of $\mathbf{U}_\gamma$ can be obtained by standardizing $\mathbf{A}_\gamma^\top\mathbf{V}_\gamma$ column-wise, as $\mathbf{A}_\gamma^\top\mathbf{V}_\gamma = \mathbf{U}_\gamma\boldsymbol{\Gamma}^\top$. Here, it is sufficient to consider only the first $(n+K)$ columns of $\mathbf{U}_\gamma$ because the other columns correspond to eigenvalue 0 and contain little information.

For a fixed K, suppose $n < p$. The computational complexities of finding $\mathbf{A}_\gamma\mathbf{A}_\gamma^\top$ and conducting its singular decomposition are $O(n^2p)$ and $O(n^3)$, respectively. The computational complexity in finding and standardizing $\mathbf{A}_\gamma^\top\mathbf{V}_\gamma$ is $O(n^2p)$. The overall computational complexity is then $O(n^2p)$. Our method is

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

computationally efficient for analyzing high-dimensional data.

4. Numerical studies

4.1 Simulated data examples

We compared the SPCALDA method with some other classification tools in simulations. In particular, we considered simple reduced-rank LDA (SRRLDA), LDA after standard PCA (PCALDA), a special case of SPCALDA with fixed $\gamma = 1$, and the independence rule (IR). We added the Bayes rule as an oracle benchmark for comparison.

Six scenarios are reported here. For each scenario, 200 observations were generated and equally split between four classes. Among the 200 observations, 100 were assigned to the training set, and the other 100 served as test data. There were $p = 500$ features. For each class k, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_w)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_w$ were as follows.

Scenario 1. The covariances were $\boldsymbol{\Sigma}_w = \mathbf{I}_p$. The mean vectors were given by $\mu_{1j} = 0.3 * \mathbb{I}_{1\le j\le 125}$, $\mu_{2j} = 0.3 * \mathbb{I}_{126\le j\le 250}$, $\mu_{3j} = 0.3 * \mathbb{I}_{251\le j\le 375}$, $\mu_{4j} = 0.3 * \mathbb{I}_{376\le j\le 500}$, with $\mathbb{I}_{\mathcal{S}}$ a vector with entries 1 for indices in $\mathcal{S}$ and 0 elsewhere.

Scenario 2. Again $\boldsymbol{\Sigma}_w = \mathbf{I}_p$, now with $\mu_{1j} \sim \mathcal{N}(0, 0.3^2)$ when $1 \le j \le 125$, and $\mu_{1j} = 0$ otherwise, $\mu_{2j} \sim \mathcal{N}(0, 0.3^2)$ when $126 \le j \le 250$, and $\mu_{2j} = 0$ otherwise, $\mu_{3j} \sim \mathcal{N}(0, 0.3^2)$ when $251 \le j \le 375$ and $\mu_{3j} = 0$ otherwise,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

$\mu_{4j} \sim \mathcal{N}(0, 0.3^2)$ when $376 \leq j \leq 500$, and $\mu_{4j} = 0$ otherwise.

Scenario 3. $\mathbf{\Sigma}_w = (\sigma_{ij})$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.5$ for $i \neq j$. $\mu_{1j} = 0.21 * \mathbb{I}_{1 \leq j \leq 125}$, $\mu_{2j} = 0.21 * \mathbb{I}_{126 \leq j \leq 250}$, $\mu_{3j} = 0.21 * \mathbb{I}_{251 \leq j \leq 375}$, $\mu_{4j} = 0.21 * \mathbb{I}_{376 \leq j \leq 500}$.

Scenario 4. $\mathbf{\Sigma}_w$ is the same as in Scenario 3. $\mu_{1j} \sim \mathcal{N}(0, 0.21^2)$ when $1 \leq j \leq 125$, and $\mu_{1j} = 0$ otherwise, $\mu_{2j} \sim \mathcal{N}(0, 0.21^2)$ when $126 \leq j \leq 250$, and $\mu_{2j} = 0$ otherwise, $\mu_{3j} \sim \mathcal{N}(0, 0.21^2)$ when $251 \leq j \leq 375$, and $\mu_{3j} = 0$ otherwise, $\mu_{4j} \sim \mathcal{N}(0, 0.21^2)$ when $376 \leq j \leq 500$, and $\mu_{4j} = 0$ otherwise.

To investigate the robustness of the proposed method against departures from the Gaussian and the equal covariance assumptions, we included two more scenarios. Scenario 5 considers a case where the data are contaminated by a random heavy-tailed noise, and in Scenario 6 the observations from different classes do not share a common covariance structure.

Scenario 5. $\mathbf{X}$ was generated as in Scenario 3. With $\mathbf{Z}$ is a p dimensional random vector with entries IID from the t-distribution with 3 degrees of freedom. We took realizations of $\tilde{\mathbf{X}} = \mathbf{X} + 0.2\mathbf{Z}$ as observed instead of $\mathbf{X}$.

Scenario 6. $\mathbf{X}$ was generated as in Scenario 3. For each class k, we generated a p-dimensional vector $\mathbf{d}_k$ with entries IID from the standard uniform distribution, and fixed a diagonal covariance matrix $\mathbf{\Delta}_k$ =diag($\mathbf{d}_k^2$). For

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

each class k, we had $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Delta}_k)$, and took realizations of $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{Z}$ as observed instead of $\mathbf{X}$.

There are two tuning parameters for SPCALDA, γ and q, and one parameter for the PCALDA method; they were chosen by five-fold cross validation. The fitted models were evaluated using the test set for all methods. We repeated each experiment 100 times. The average and standard deviation of classification error rates for each method are listed in Table 1. The SPCALDA method always outperformed PCALDA, which indicates that it is helpful to tune the parameter γ. In the independence cases, SPCALDA is comparable with SRRLDA and IR, but much better than them in the correlated cases. When the Gaussian assumption or the equal covariance assumption is violated, we see that SPCALDA still performs reasonably well.

Table 1: Mean (and standard errors) of classification error rates (%).

	SPCALDA	PCALDA	SRRLDA	IR	Oracle
Scenario 1	18.93(4)	26.53(4.52)	19.33(3.94)	18.45(3.86)	2.69(1.6)
Scenario 2	19.96(3.91)	27.71(5.1)	20.46(4.7)	19.29(4.03)	2.8(1.75)
Scenario 3	20.73(4.32)	30(5.64)	36.61(10.75)	63.92(5.41)	2.73(1.63)
Scenario 4	22.78(4.4)	32.26(5.82)	38.61(10.31)	64.38(7.92)	3.07(1.69)
Scenario 5	28.8(4.82)	38.42(6.41)	43.52(9.66)	64.38(5.8)	NA
Scenario 6	38.29(5.35)	50.75(6.72)	49.44(8.85)	64.79(6.57)	NA

4.2 Data examples

In this section, we illustrate the performance of our method using six gene expression data sets, that have been studied in the literature. In particular,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

we considered three binary data sets, **Chin** (Chin et al. (2006)), **Chowdary** (Chowdary et al. (2006)), **Gordon** (Gordon et al. (2002)), and three multi-class data sets, **Golub** (Golub et al. (1999)), **Nakayama** (Nakayama et al. (2007)), and **Sun** (Sun et al. (2006)). The three binary data sets are available in R package `datamicroarray`. The data set **Golub** is available in R package `golubEsets`. The original **Nakayama** data set contains 105 samples from 10 types of soft tissue tumors. We considered a subset of 86 samples belonging to 5 tumor types and ignored the other tumor types for which less than 7 samples were available. **Nakayama** and **Sun** are available on Gene Expression Omnibus (Barrett et al. (2005)) with accession numbers GDS2736 and GDS1962, respectively. We list in Table 2 the sample size, number of features, number of classes, data distribution among different classes, and related disease for each data set.

Table 2: Data sets used in this study.

Data set	related disease	# samples	# features	# classes	data distribution
Chin	breast cancer	118	22,215	2	43, 75
Chowdary	breast cancer	104	22,283	2	42, 62
Gordon	lung cancer	181	12,533	2	87, 94
Golub	leukemia	72	7,129	3	9, 25, 38
Nakayama	soft tissue tumor	86	22,283	5	15, 15, 16, 19, 21
Sun	glioma	180	54,613	4	23, 26, 50, 81

Besides the methods considered in Section 4.1, we included the multi-class classification tools NSC (Tibshirani et al. (2002)), RDA (Guo, Hastie and Tibshirani (2007)), PLDA (Witten and Tibshirani (2011)), and SDA (Clemmensen et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

(2011)); these have been implemented by `R` packages `pamr`, `rda`, `penalizedLDA`, and `sparseLDA`, respectively. These methods are based on various sparsity assumptions.

For each of the data sets, we randomly split the data, with a 3 to 1 ratio, into a training set and a test set. Five-fold cross-validation was conducted on the training set to select the tuning parameters for all methods, and the classification error rates using the test sets were recorded. In Table 3 we list the average classification error rates and their standard deviations over 25 random training/test set splits. We omit the results of PCALDA and IR which were dominated by SPCALDA and NSC, respectively. We see that SPCALDA performed best for two data sets, and second best for four data sets. In particular, SPCALDA did the best in pairwise comparisons with other methods. We list in Table 4 the computation time for each method. All methods were reasonably fast in handling contemporary high-dimensional data sets. SPCALDA offers a competitive classification tool for high-dimensional gene expression data.

Table 3: Mean (and standard errors) of classification error rates (%).

	SPCALDA	SRRLDA	NSC	PLDA	RDA	SDA
Chin	11.57(6.57)	12.11(6.13)	12.41(7.42)	13.87(6.79)	12.25(5.32)	10.13(4.47)
Chowdary	4.13(3.62)	10.43(5.82)	5.19(5.03)	33.63(9.52)	4.75(3.9)	17.19(8.26)
Gordon	0.62(1.19)	2.29(2.99)	0.79(1.08)	0.53(0.96)	1.4(1.25)	6.02(3.41)
Golub	5.43(5.44)	24.2(12.93)	4.6(4.17)	7.41(5.91)	6.3(3.74)	15.89(11.62)
Nakayama	16.37(7.05)	20.6(8.94)	23.51(6.36)	27.6(8.84)	15.68(7.98)	33.73(6.89)
Sun	30.43(5.73)	31.63(6.89)	33.24(6.03)	33.21(5.89)	33.48(6.97)	33.33(8.78)

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

Table 4: Mean (and standard error) of computation time per replicate (in second).

	SPCA-LDA	SRR-LDA	NSC	PLDA	RDA	SDA
Chin	14.5(1.03)	0.05(0.03)	3.93(0.61)	13.51(0.77)	51.17(3.27)	1.12(0.22)
Chowdary	12.09(0.25)	0.05(0.04)	3.63(0.21)	12.41(0.75)	49.52(0.43)	1.09(0.17)
Gordon	11.68(0.15)	0.04(0.02)	1.95(0.07)	6.82(0.1)	30.2(0.44)	0.53(0.04)
Golub	2.98(0.08)	0.01(0.01)	0.71(0.04)	3.9(0.34)	10.97(0.25)	0.22(0.03)
Nakayama	8.77(0.2)	0.06(0.02)	2.33(0.09)	27.46(0.73)	38.14(0.55)	1.75(0.22)
Sun	55.16(1.19)	0.24(0.02)	9.16(0.28)	56.3(0.94)	162.03(2.46)	6.42(0.75)

5. Discussion

Feature selection and feature extraction are two popular strategies in statistical machine learning. In the context of this paper, the sparse methods such as NSC and SDA can conduct variable selection and model estimation simultaneously, and belong to the first category. Our methods, including classical reduced-rank LDA and PCA as special cases, belong to a second. While sparse methods can achieve model selection consistency and efficiency under various assumptions, they can fail when the true model is far from sparse. Our approach does not depend on sparse assumptions and is robust against the sparsity level of the true model. Our data examples support the robustness of our method. In general, we can not expect a result on model selection consistency or efficiency, but we discuss a spiked covariance condition under which our method may achieve efficiency.

6. Supplementary Materials

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

The Supplementary Material is available on the journal web site. It contains Corollary 1, Lemma 2, and proofs for Theorems 1 and 2.

Acknowledgements

This research was partially supported by grants DMS 1309507 from the National Science Foundation and 11671022 from the National Science Foundation of China. The authors are grateful to Daniela Witten for helpful discussions, and to the Editor, associate editor, and two referees for comments and suggestions that greatly improved the paper.

References

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. and Edgar, R. (2005). Ncbi geo: mining millions of expression profilesdatabase and tools. *Nucleic acids research* **33**, D562–D566.

Bickel, P. and Levina, E. (2004). Some theory for fisher's linear discrimi-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

nant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

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

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T. et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell* **10**, 529–541.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y. and Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative. *The journal of molecular diagnostics* **8**, 31–39.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics* **53**, 406–413.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics* **36**, 2605–2637.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 745–771.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science* **286**, 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research* **62**, 4963–4967.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86–100.

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**, 827–851.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer-Verlag New York.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist* **29**, 295–327.

Johnstone, I. and Lu, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**, 682–693.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer.

Mai, Q., Yang, Y. and Zou, H. (2015). Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845* .

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **102**, 33–45.

Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Seki, K., Yoshida, T., Toyama, Y., Ichikawa, H. and Hasegawa, T. (2007). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Modern pathology* **20**, 749–759.

Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R. et al. (2006).

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0387

Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer cell* **9**, 287–300.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* **99**, 6567–6572.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 753–772.

University of Arizona, Tucson, AZ, 85721, USA

E-mail: yueniu@math.arizona.edu

University of Arizona, Tucson, AZ, 85721, USA

E-mail: nhao@math.arizona.edu

Beijing International Center for Mathematical Research, Peking University, Beijing, China

E-mail: dongbin@math.pku.edu.cn