Two-Stage Regularized Linear Discriminant Analysis for 2-D Data

Jianhua Zhao, *Member, IEEE*, Lei Shi, and Ji Zhu

***Abstract*— Fisher linear discriminant analysis (LDA) involves within-class and between-class covariance matrices. For 2-D data such as images, regularized LDA (RLDA) can improve LDA due to the regularized eigenvalues of the estimated within-class matrix. However, it fails to consider the eigenvectors and the estimated between-class matrix. To improve these two matrices simultaneously, we propose in this paper a new two-stage method for 2-D data, namely a bidirectional LDA (BLDA) in the first stage and the RLDA in the second stage, where both BLDA and RLDA are based on the Fisher criterion that tackles correlation. BLDA performs the LDA under special separable covariance constraints that incorporate the row and column correlations inherent in 2-D data. The main novelty is that we propose a simple but effective statistical test to determine the subspace dimensionality in the first stage. As a result, the first stage reduces the dimensionality substantially while keeping the significant discriminant information in the data. This enables the second stage to perform RLDA in a much lower dimensional subspace, and thus improves the two estimated matrices simultaneously. Experiments on a number of 2-D synthetic and real-world data sets show that BLDA+RLDA outperforms several closely related competitors.**

***Index Terms*— 2-D data, dimension reduction, linear discriminant analysis (LDA), regularization, separable covariance.**

I. Introduction

FISHER linear discriminant analysis (LDA) is a well-known supervised subspace learning technique for 1-D data, where observations are vectors. To apply LDA to 2-D data such as images, where observations are matrices, one common solution is to first vectorize the 2-D data and then apply LDA to the resulting 1-D data. However: 1) the vectorized 1-D data are often of very high dimension (typically over tens of thousands of pixels), on which LDA suffers from poor performance due to the (approximately) singular within-class sample covariance matrix [1] and 2) even if sufficient number of samples are available, the variables (i.e., pixels of images) are often highly correlated, which could deteriorate the performance of LDA [2].

To deal with problems 1) and 2) simultaneously, several solutions have been proposed in the literature. The regularized LDA (RLDA) proposed in [3] has been proven effective [4], [5]. Moreover, Ji and Ye [6] show that RLDA is comparable with or even better than the state-of-art classifier support vector machine while being computationally more efficient. Despite the success of RLDA, RLDA simply regularizes the eigenvalues of the estimated within-class covariance matrix. It fails to consider the eigenvectors and the estimated between-class covariance matrix. It is thus of interest to investigate whether the performance of RLDA could be improved if some sort of regularization is applied to these two matrices simultaneously.

A simple yet effective way toward this end is to use the idea of two-stage methods presented in [7]: in the first stage, one reduces the data dimensionality to a moderate size (by discarding the unnecessary variables/features); in the second stage, one then performs RLDA in the lower dimensional space. This idea has been frequently employed in the classification of gene expression data, where a preliminary selection of genes using diagonal-covariance LDA (DLDA) is performed in the first stage and then the selected important variables/genes are fed into a classier in the second stage [8], [9].

The key to DLDA is the independence assumption, which has been proven successful for gene expression data [8], [9], [11]. This could be ascribed into the fact that a large number of genes exhibit nearly constant expression levels across samples [12] and hence DLDA is suitable to screen out unimportant ones. However, the variables (pixels) of image data are highly correlated and ignoring correlation among pixels would be suboptimal. Fan *et al.* [13] show that the gain of incorporating the correlation is substantial even for the classification of gene expression data and the crucial challenge is how to incorporate the correlation appropriately into the analysis.

Several two-stage methods have been proposed for image data. In the first stage, one uses 2-D reduction methods that consider the underlying 2-D data structure to reduce the dimensionality of the data, and in the second stage, one performs LDA in the reduced-dimensional space. For example, an unsupervised method called bidirectional principal component analysis (BPCA) is used in [14]. The shortcoming is that the label information is not used for reducing dimensionality, and thus the discarded subspace may contain useful discriminant

Manuscript received August 30, 2013; revised May 22, 2014; accepted August 11, 2014. Date of publication September 5, 2014; date of current version July 15, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 11361071, Grant 61403337, Grant 11161053, and Grant U1302267, in part by the National Science Foundation under Grant DMS0748389, in part by the Hong Kong Research Grants Councils, General Research Fund, University of Hong Kong, Hong Kong, under Grant 706710P, and in part by the Program for New Century Excellent Talents in University.

J. Zhao and L. Shi are with the School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China (e-mail: jhzhao.ynu@gmail.com; shi_lei65@hotmail.com).

J. Zhu is with the Department of Statistics, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jizhu@umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2350993

information. Supervised methods are also proposed, such as bidirectional maximum margin criterion (BMMC) in [16] and a weighted version of BMMC in [15]. One common drawback is that these maximum margin criterion (MMC)-type methods ignore the within-class correlation altogether [17]. The 2-D LDA (2DLDA) proposed in [18] can deal with the within-class correlation. However, the associated algorithm requires iterations for a given subspace dimensionality and its convergence is not guaranteed [16], [19].

In practice, a challenging issue in applying two-stage methods is how to determine the suitable subspace dimensionality in the first stage. Tao *et al.* [16] propose using the retained eigenvalue ratio above a threshold whose value is chosen with validation data. Usually, to achieve satisfactory performance, a large set of candidate values is considered and thus the computational cost could be very high. Yang and Dai [15] suggest setting the threshold as the positive part of the mean of all eigenvalues, which is denoted as YD's criterion for clarity. In spite of its simplicity, our experiments reveal that its performance is far from satisfactory, e.g., it fails to select any feature on the United States Postal Service (USPS) digit data in Section IV-D.

In this paper, we propose a new two-stage RLDA for 2-D data, namely bidirectional LDA (BLDA) plus RLDA. Unlike BMMC, BLDA utilizes the Fisher criterion that tackles the within-class correlation. Unlike independence-based DLDA, BLDA performs the LDA under special separable covariance constraints that incorporate the correlations among columns and rows in 2-D data. More importantly, we can obtain a simple yet effective statistical test for BLDA to determine the subspace dimensionality. Therefore, the first stage reduces the dimensionality substantially, and a much lower dimensional features than that in DLDA could be used in the subsequent RLDA stage.

The remainder of this paper is organized as follows. Section II gives a brief review of LDA and several related variants. Section III proposes our method. Section IV performs experiments to compare the proposed method with several related competitors. We end this paper with some concluding remarks in Section V.

Notations: The identity matrix is denoted by $\mathbf{I}$, its column j by $\mathbf{e}_j$, the Frobenius norm by $\|\cdot\|_F$, the matrix trace by $\mathrm{tr}(\cdot)$, the vectorization operator by $\mathrm{vec}(\cdot)$, and the Kronecker product by $\otimes$. Moreover, for k-class classification problem, π_j is the prior probability of class j, $\mathcal{C}_j$ is the set of n_j observations in class j, where $j = 1, \ldots, k$, and $n = \sum_j n_j$ is the total number of observations.

II. Review of LDA and Related Variants

A. Linear Discriminant Analysis

Let $\mathbf{x} \in \mathbb{R}^d$ be a random vector drawn from k classes. LDA assumes that all classes follow normal distributions with different means $\boldsymbol{\mu}_j$'s but common within-class covariance matrices Σ_w [20], i.e., $\mathbf{x}|\, j \sim \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_w)$, $j = 1, \ldots, k$. The global population mean $\boldsymbol{\mu} = \sum_j \pi_j \boldsymbol{\mu}_j$, and the within-class covariance matrix is defined as

$$\Sigma_w = \sum_{j=1}^{k} \pi_j \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)' | \mathbf{x} \in \mathcal{C}_j]. \tag{1}$$

The between-class covariance matrix is defined as

$$\Sigma_b = \sum_{j=1}^{k} \pi_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})' \tag{2}$$

which measures the scatter of class means $\boldsymbol{\mu}_j$'s around the global mean $\boldsymbol{\mu}$. Consider a linear transformation $\mathbf{y} = \mathbf{V}'\mathbf{x}$, where $\mathbf{V} \in \mathbb{R}^{d \times q}$ and $q < d$. The within-class and between-class covariance matrices in $\mathbf{y}$-space are $\mathbf{V}'\Sigma_w\mathbf{V}$ and $\mathbf{V}'\Sigma_b\mathbf{V}$, respectively. The Fisher criterion aims to find $\mathbf{V}$ that maximizes the between-class covariance in terms of whitened $\mathbf{y}$, i.e., $(\mathbf{V}'\Sigma_w\mathbf{V})^{-1/2}\mathbf{y}$. This is formulated in [21] as

$$\mathcal{F} = \max_{\mathbf{V}} \mathrm{tr}\big\{(\mathbf{V}'\Sigma_w\mathbf{V})^{-1}(\mathbf{V}'\Sigma_b\mathbf{V})\big\}. \tag{3}$$

Let $(\lambda_j, \mathbf{v}_j)$ be the jth eigenvalue–eigenvector pair of $\Sigma_w^{-1}\Sigma_b$. As shown in [21], the closed-form solution $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q]$ to (3) is given by

$$\mathbf{V} \longleftarrow \text{top } q \text{ eigenvectors of } \Sigma_w^{-1}\Sigma_b. \tag{4}$$

Then, the Fisher criterion in (3) can be rewritten as

$$\mathcal{F} = \sum\nolimits_{j=1}^{q} \lambda_j.$$

As $\mathbf{v}_j$ is the eigenvector of $\Sigma_w^{-1}\Sigma_b$, we have $\Sigma_b\mathbf{v}_j = \lambda_j\Sigma_w\mathbf{v}_j$ and

$$\lambda_j = \frac{\mathbf{v}_j'\Sigma_b\mathbf{v}_j}{\mathbf{v}_j'\Sigma_w\mathbf{v}_j} \tag{5}$$

which measures the discriminant capability of direction $\mathbf{v}_j$. Given data $\{\mathbf{x}_i\}_{i=1}^n$ as independent identically distributed (i.i.d.) realizations of $\mathbf{x}$, Σ_b and Σ_w can be estimated by

$$\hat{\Sigma}_b = \frac{1}{n}\sum\nolimits_{j=1}^{k} n_j(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}})' \tag{6}$$

$$\hat{\Sigma}_w = \frac{1}{n}\sum\nolimits_{j=1}^{k}\sum\nolimits_{i\in\mathcal{C}_j}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)' \tag{7}$$

where $\hat{\boldsymbol{\mu}}_j = (1/n_j)\sum_{i\in\mathcal{C}_j}\mathbf{x}_i$ is the sample mean of class j and $\hat{\boldsymbol{\mu}} = (1/n)\sum_i \mathbf{x}_i$ is the global sample mean.

B. Regularized LDA

The RLDA proposed in [3] is a popular method to tackle problems 1) and 2) mentioned in Section I simultaneously. It regularizes $\hat{\Sigma}_w$ in (7) as

$$\hat{\Sigma}_w(\gamma) = \gamma\,\hat{\Sigma}_w + (1-\gamma)\hat{\sigma}^2\mathbf{I} \tag{8}$$

where $\hat{\sigma}^2 = \mathrm{tr}(\hat{\Sigma}_w)/d$ and $\gamma \in [0, 1]$. $\hat{\Sigma}_w$ in (8) is shrunk to a scalar covariance. Recall that ridge regression is a well-known method to deal with the correlation among variables. The relationship between RLDA and ridge regression has been established in [2] and [5], which clarifies why RLDA is effective for image data.

C. Diagonal-Covariance LDA

The DLDA is the LDA under the diagonal covariance assumption [8]. Under this assumption, both the between-class and within-class covariances are diagonal

$$\Sigma_b^I = \text{diag}\{\phi_{b1}, \phi_{b2}, \ldots, \phi_{bd}\}$$
$$\Sigma_w^I = \text{diag}\{\phi_{w1}, \phi_{w2}, \ldots, \phi_{wd}\}.$$

Substituting Σ_b^I and Σ_w^I into (4), we obtain that $\mathbf{V}$ consists of the q columns of the identity matrix $\mathbf{I}$ corresponding to the leading q largest ratio ϕ_{bj}/ϕ_{wj}'s, and the Fisher criterion in (3) under the diagonal covariance assumption degenerates into a sum of q Fisher subcriterions, each of which is devoted to a variable

$$\max_I \sum_{j\in I} \frac{\phi_{bj}}{\phi_{wj}} = \sum\nolimits_{j\in I} \lambda_j^I \tag{9}$$

where I is the set of indexes of q variables and $\lambda_j^I = \phi_{bj}/\phi_{wj}$ measures the discriminant capability of variable j. Note that Σ_b^I and Σ_w^I can be estimated by the diagonal parts of $\hat{\Sigma}_b$ in (6) and $\hat{\Sigma}_w$ in (7), respectively.

1) Discriminant Direction Screening in DLDA: Consider the following hypothesis tests independently:

$$H_{0j} : \lambda_j^I = 0 \quad \text{versus} \quad H_{1j} : \lambda_j^I \neq 0, \quad j = 1, \ldots, d.$$

Under the normal assumption, we could use the F-test in the classical univariate analysis of variance (ANOVA) to screen out significant variables. Note that in the two-class case, the F-test is equivalent to the two-sample t-test [10], [11].

D. 2-D LDA

Instead of vectorizing 2-D data, 2DLDA performs dimension reduction on 2-D data directly. The row-2DLDA [22] seeks a row linear transformation $\mathbf{Y} = \mathbf{X}\mathbf{U}_r$ to maximize the class separability in the low-dimensional space, where $\mathbf{U}_r = [\mathbf{u}_{r1}, \ldots, \mathbf{u}_{rq_r}] \in \mathbb{R}^{d_r \times q_r}$ and $q_r < d_r$. The relationship between row-2DLDA and LDA is established in the following theorem [19].

Theorem 1: Apart from a constant, the row-2DLDA is equivalent to the LDA under the special separable covariance assumption that the between-class and within-class covariance matrices are $\Sigma_b = \Sigma_b^r \otimes \mathbf{I}$ and $\Sigma_w = \Sigma_w^r \otimes \mathbf{I}$.

By Theorem 1 and the property of Kronecker product that $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC}\otimes\mathbf{BD})$, we get $\Sigma_w^{-1}\Sigma_b = (\Sigma_w^{r\,-1}\Sigma_b^r)\otimes\mathbf{I}$ and substituting Σ_b and Σ_w into (4), we find that the $\mathbf{V}$ solution to Fisher criterion (3) is now factorized into $\mathbf{V} = \mathbf{U}_r \otimes \mathbf{I}$, where

$$\mathbf{U}_r \longleftarrow \text{top } q_r \text{ eigenvectors of } \Sigma_w^{r\,-1}\Sigma_b^r. \tag{10}$$

Using the property $\text{tr}(\mathbf{A}\otimes\mathbf{B}) = \text{tr}(\mathbf{A})\text{tr}(\mathbf{B})$, apart from a constant d_c, the Fisher criterion (3) under this assumption degenerates into a Fisher subcriterion devoted to row directions only

$$\max_{\mathbf{U}_r} \text{tr}\big\{(\mathbf{U}_r'\Sigma_w^r\mathbf{U}_r)^{-1}(\mathbf{U}_r'\Sigma_b^r\mathbf{U}_r)\big\} = \sum_{j=1}^{q_r} \lambda_{rj} \tag{11}$$

where λ_{rj} is the corresponding eigenvalue to $\mathbf{u}_{rj}$, $j = 1, \ldots, q_r$, given by

$$\lambda_{rj} = \frac{\mathbf{u}_{rj}'\Sigma_b^r\mathbf{u}_{rj}}{\mathbf{u}_{rj}'\Sigma_w^r\mathbf{u}_{rj}}. \tag{12}$$

Similar to the Fisher values λ_j (5) in LDA and λ_j^I in DLDA, the value of λ_{rj} in (12) measures the discriminant capability of direction $\mathbf{u}_{rj}$.

Similarly, the column-2DLDA seeks a column linear transformation $\mathbf{Y} = \mathbf{U}_c'\mathbf{X}$, where $\mathbf{U}_c = [\mathbf{u}_{c1}, \ldots, \mathbf{u}_{cq_c}] \in \mathbb{R}^{d_c \times q_c}$ and $q_c < d_c$, to maximize class separation. From [19], we have the following theorem.

Theorem 2: Apart from a constant, the column-2DLDA is equivalent to the LDA under the assumption that the between-class and within-class covariance matrices are $\Sigma_b = \mathbf{I} \otimes \Sigma_b^c$ and $\Sigma_w = \mathbf{I} \otimes \Sigma_w^c$.

By Theorem 2, the $\mathbf{V}$ solution to Fisher criterion (3) is $\mathbf{V} = \mathbf{I} \otimes \mathbf{U}_c$, where

$$\mathbf{U}_c \longleftarrow \text{top } q_c \text{ eigenvectors of } \Sigma_w^{c\,-1}\Sigma_b^c. \tag{13}$$

Apart from a constant d_r, the Fisher criterion (3) under this assumption degenerates into a Fisher subcriterion devoted to column directions

$$\max_{\mathbf{U}_c} \text{tr}\big\{(\mathbf{U}_c'\Sigma_w^c\mathbf{U}_c)^{-1}(\mathbf{U}_c'\Sigma_b^c\mathbf{U}_c)\big\} = \sum_{i=1}^{q_c} \lambda_{ci} \tag{14}$$

where λ_{ci} is the corresponding eigenvalue to $\mathbf{u}_{ci}$, $i = 1, \ldots, q_c$, given by

$$\lambda_{ci} = \frac{\mathbf{u}_{ci}'\Sigma_b^c\mathbf{u}_{ci}}{\mathbf{u}_{ci}'\Sigma_w^c\mathbf{u}_{ci}}. \tag{15}$$

Given a set of 2-D data $\{\mathbf{X}_i\}_{i=1}^n$ consisting of k classes, Σ_b^c, Σ_b^r, Σ_w^c, and Σ_w^r can be estimated by their corresponding sample moments, given by

$$\hat{\Sigma}_b^c = \frac{1}{nd_r}\sum\nolimits_j n_j(\hat{\mathbf{M}}_j - \hat{\mathbf{M}})(\hat{\mathbf{M}}_j - \hat{\mathbf{M}})' \tag{16}$$

$$\hat{\Sigma}_b^r = \frac{1}{nd_c}\sum\nolimits_j n_j(\hat{\mathbf{M}}_j - \hat{\mathbf{M}})'(\hat{\mathbf{M}}_j - \hat{\mathbf{M}}) \tag{17}$$

$$\hat{\Sigma}_w^c = \frac{1}{nd_r}\sum\nolimits_j \sum\nolimits_{i\in\mathcal{C}_j}(\mathbf{X}_i - \hat{\mathbf{M}}_j)(\mathbf{X}_i - \hat{\mathbf{M}}_j)' \tag{18}$$

$$\hat{\Sigma}_w^r = \frac{1}{nd_c}\sum\nolimits_j \sum\nolimits_{n\in\mathcal{C}_j}(\mathbf{X}_i - \hat{\mathbf{M}}_j)'(\mathbf{X}_i - \hat{\mathbf{M}}_j) \tag{19}$$

where the sample mean $\hat{\mathbf{M}}_j$ of class j and global sample mean $\hat{\mathbf{M}}$ are

$$\hat{\mathbf{M}}_j = \frac{1}{n_j}\sum\nolimits_{i\in\mathcal{C}_j}\mathbf{X}_i \quad \text{and} \quad \hat{\mathbf{M}} = \frac{1}{n}\sum\nolimits_i \mathbf{X}_i. \tag{20}$$

E. BLDA

The BLDA proposed in [23] seeks a bilinear transformation $\mathbf{Y} = \mathbf{U}_c'\mathbf{X}\mathbf{U}_r$ that reduces the dimensionality on both column and row directions simultaneously. The $\mathbf{U}_c$ and $\mathbf{U}_r$ solutions are given by (13) and (10), respectively.

1) Justification of BLDA: It can be seen that (3) is invariant for any invertible transformation $\mathbf{R} \in \mathbb{R}^{q\times q}$, i.e., $\mathcal{F}(\mathbf{V}) = \mathcal{F}(\mathbf{VR})$, and thus Fisher criterion finds the important subspace spanned by $\mathbf{V}$, namely span($\mathbf{V}$), or, equivalently, discards the unimportant subspace span($\mathbf{V}^{\perp}$) spanned by its orthogonal complement $\mathbf{V}^{\perp}$.

Let $\mathbf{U}_c^{\perp}$ (respectively $\mathbf{U}_r^{\perp}$) be the orthogonal complement of $\mathbf{U}_c$ (respectively $\mathbf{U}_r$). Similarly, we discard the unimportant subspaces span($\mathbf{U}_r^{\perp}\otimes\mathbf{I}$) by row-2DLDA, and span($\mathbf{I}\otimes\mathbf{U}_c^{\perp}$) by column-2DLDA. Since span($\mathbf{U}_r^{\perp}\otimes\mathbf{I}$) = span($\mathbf{U}_r^{\perp}\otimes[\mathbf{U}_c, \mathbf{U}_c^{\perp}]$) and span($\mathbf{I}\otimes\mathbf{U}_c^{\perp}$) = span($[\mathbf{U}_r, \mathbf{U}_r^{\perp}]\otimes\mathbf{U}_c^{\perp}$), by abandoning these unimportant subspaces, we get the important discriminant subspace span($\mathbf{U}_r\otimes\mathbf{U}_c$). Thus, the $\mathbf{U}_c$ and $\mathbf{U}_r$ can be obtained by (13) and (10), respectively. This provides a justification why the separate solutions from row-2DLDA and column-2DLDA can be taken as the solution to BLDA, which seems missing in the literature.

F. Bidirectional MMC

Different from the Fisher criterion which is based on the ratio of the between-class matrix to the within-class matrix, MMC uses the difference between these two matrices. The BMMC looks for a bilinear transformation $\mathbf{Y} = \mathbf{U}_c'\mathbf{X}\mathbf{U}_r$, where $\mathbf{U}_c \in \mathbb{R}^{d_c\times q_c}$ and $\mathbf{U}_r \in \mathbb{R}^{d_r\times q_r}$ whose columns are orthogonal and $q_c < d_c$, $q_r < d_r$. $\mathbf{U}_c$ and $\mathbf{U}_r$ can be analytically obtained by independently solving

$$\arg\max_{\mathbf{U}_c} \operatorname{tr}\left\{\mathbf{U}_c'(\hat{\Sigma}_b^c - \hat{\Sigma}_w^c)\mathbf{U}_c\right\}$$

and

$$\arg\max_{\mathbf{U}_r} \operatorname{tr}\left\{\mathbf{U}_r'(\hat{\Sigma}_b^r - \hat{\Sigma}_w^r)\mathbf{U}_r\right\}.$$

Let λ_{ci}^m (respectively λ_{rj}^m) be the descending-ordered eigenvalues of $\hat{\Sigma}_b^c - \hat{\Sigma}_w^c$ (respectively $\hat{\Sigma}_b^r - \hat{\Sigma}_w^r$). The YD's criterion [15] determines the subspace dimensionality (q_c, q_r) by

$$q_c = \arg\min_i \lambda_{ci}^m > \max\{\operatorname{tr}(\hat{\Sigma}_b^c - \hat{\Sigma}_w^c)/d_c, 0\}$$
$$q_r = \arg\min_j \lambda_{rj}^m > \max\{\operatorname{tr}(\hat{\Sigma}_b^r - \hat{\Sigma}_w^r)/d_r, 0\}.$$

However, as revealed by our experiments in Sections IV-B and IV-D, the performance of YD's criterion is not satisfactory. It tends to choose too small number of features and even no features on the USPS digit data (Section IV-D). In addition, BMMC-type methods ignore the within-class correlation altogether [19] (namely fail to consider the whitened between-class matrix). Fig. 1 shows the corresponding correlation matrices to $\hat{\Sigma}_b^c$, $\hat{\Sigma}_b^r$, $\hat{\Sigma}_w^c$, and $\hat{\Sigma}_w^r$ in (16)–(19), respectively, obtained on YALE data. It can be observed from Fig. 1 that the row and column within-class correlations are substantial.

III. Two-Stage RLDA for 2-D Data

For high-dimensional data, $\hat{\Sigma}_b$ in (6) and $\hat{\Sigma}_w$ (7) are not generally considered as good estimates of Σ_b and Σ_w due to diverging spectra and noise accumulation [13]. It is unlikely that the product in (4) is a good estimate if either one is not. This is why LDA suffers from poor performance. RLDA only improves $\hat{\Sigma}_w$, but not $\hat{\Sigma}_b$, which involves the estimates of means. Even though each component of mean parameters can be estimated with accuracy, the aggregated estimation error can be very large [11]. In addition, it can be observed from (8) that only the eigenvalues of $\hat{\Sigma}_w$ are regularized and its eigenvectors do not change. A simple yet effective way for improving $\hat{\Sigma}_w$ and $\hat{\Sigma}_b$ simultaneously is to use the idea of two-stage methods [7], as detailed in Section III-A.

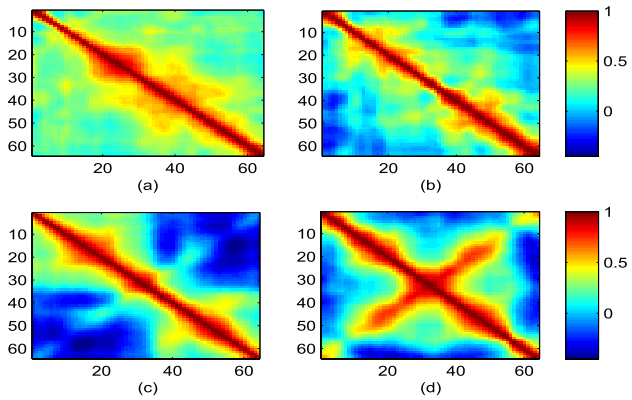


Fig. 1. Color images of correlation matrices on YALE data. (a) Column within-class correlation. (b) Column between-class correlation. (c) Row within-class correlation. (d) Row between-class correlation.

A. Working Principle of Two-Stage RLDA

The idea of applying two-stage methods to LDA is rather simple. If we know in advance which directions $\mathbf{v}_j$'s are not important in terms of Fisher criterion (3), we may just discard these directions in the first stage and project the data onto the important ones. Then, in the second stage, the estimation of $\hat{\Sigma}_w$ and $\hat{\Sigma}_b$ in a lower dimensional space would be more accurate.

In practice, such prior information in the first stage is not available and will need to be estimated using the training data. This could be achieved by applying certain constraints to reduce the degrees of freedom of the transformation $\mathbf{V}$. If $\mathbf{V}$ is more constrained, the estimation will be more accurate but the capability of reducing the dimensionality in the first stage would be weaker. Thus, there exists a tradeoff between the dimensionality of the important subspace in the first stage and the degrees of freedom of $\mathbf{V}$. Typically, the first stage is expected to be:

1) simple and fast to implement;
2) effective to reduce the dimensionality substantially while retaining significant discriminant information in the data.

Using (1) and (2) and substituting the solution $\mathbf{e}_j$ in DLDA into Fisher criterion (3) yields Fisher value $\mathcal{F} = \phi_{bj}/\phi_{wj} = \lambda_j^I$. This means that DLDA performs the Fisher criterion exactly but under the independence constraint that greatly reduces the number of free parameters of $\mathbf{V}$, as shown in Table I. If the test $\lambda_j^I = 0$ is not significant, we could discard $\mathbf{e}_j$. Although DLDA has been proven satisfying 1) and 2) for gene expression data, our experiments in Section IV reveal that DLDA fails to satisfy 2) for image data due to highly correlated variables (pixels).

TABLE I
NUMBER OF FREE PARAMETERS IN DLDA, BLDA, AND LDA ON 2-D DATA

Method	Number of free parameters $(d_c, d_r), (q_c, q_r)$	$(20,20), (5,5)$	$(100,100)$ $(15,15)$
DLDA	$q_c q_r$	25	225
BLDA	$d_c q_c - q_c(q_c+1)/2$ $d_r q_r - q_r(q_r+1)/2$	170	2760
LDA	$d_c d_r q_c q_r - q_c q_r (q_c q_r + 1)/2$	9675	2224575

Again, substituting the solution $\mathbf{u}_{rj} \otimes \mathbf{I}$ in row-2DLDA into Fisher criterion (3) yields Fisher value $\mathcal{F} = \lambda_{rj}$. This means that row-2DLDA also performs the Fisher criterion exactly but under the special separable covariance constraint that also greatly reduces the number of free parameters of $\mathbf{V}$, as shown in Table I. If the test $\lambda_{rj} = 0$ is not significant, we could abandon $\mathbf{u}_{rj}$. Similarly, column-2DLDA also performs the Fisher criterion exactly. If $\lambda_{ci} = 0$ is not significant, we could discard $\mathbf{u}_{ci}$.

BLDA abandons $\mathbf{u}_{ci}$ and $\mathbf{u}_{rj}$ simultaneously. Obviously, BLDA satisfies 1). In addition, BLDA has more advantages than DLDA in that BLDA incorporates the correlations among columns and rows inherent in 2-D data. Despite the BMMC-type methods used in [15] and [16] satisfying 1), they use different cost functions in the first and second stages. Furthermore, the performance of YD's criterion in terms of 2) is not satisfactory, as mentioned in Section II-F. In Section III-B, we develop a simple yet effective statistical test for BLDA to determine the subspace dimensionality. Therefore, BLDA reduces the dimensionality substantially while reserving the significant discriminant information in the data, which makes BLDA satisfy 2) as well. Thus, we obtain a new two-stage RLDA for 2-D data, namely BLDA plus RLDA. The whole algorithm is detailed in Section III-C.

B. New Discriminant Direction Screening Procedure in BLDA

1) Preliminary: Let $\mathcal{W}_d(n, \Sigma)$ denote a Wishart distribution with n degrees of freedom and scale matrix Σ. The following proposition gives a property of Wishart distribution.

Proposition 1: Suppose that $\mathbf{W} = (w_{ij}) \sim \mathcal{W}_d(n, \sigma^2\mathbf{I})$. Then, $w_{ii}, i = 1, 2, \ldots, d$ are i.i.d. and $w_{ii}/\sigma^2 \sim \chi^2(n)$.

Proof: This is the result of [24, Exercise 3.3, p. 86], and hence the proof is omitted. ∎

If we assume that the within-class matrix Σ_w in Section II-A is a scalar covariance, i.e., $\Sigma_w = \sigma^2\mathbf{I}$, then for the sample between-class and within-class covariance $\hat{\Sigma}_b$ and $\hat{\Sigma}_w$ defined in (6) and (7), we have the following proposition.

Proposition 2: $\mathrm{tr}(n\hat{\Sigma}_w)/\sigma^2 \sim \chi^2(d(n-k))$. Under the hypothesis that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k$, $\mathrm{tr}(\hat{\Sigma}_b)$ is independent with $\mathrm{tr}(\hat{\Sigma}_w)$, $\mathrm{tr}(n\hat{\Sigma}_b)/\sigma^2 \sim \chi^2(d(k-1))$, and

$$\frac{\mathrm{tr}(\hat{\Sigma}_b)/(k-1)}{\mathrm{tr}(\hat{\Sigma}_w)/(n-k)} \sim F(d(k-1), d(n-k)). \tag{21}$$

Proof: See Appendix A. ∎

2) Proposed Procedure: In this section, we shall develop a new statistical test to determine whether the discriminant direction $\mathbf{u}_{ci}$ or $\mathbf{u}_{rj}$ is significant or not. The main difference from the tests in Section III-B1 is that our tests are developed for 2-D data, instead of 1-D data. We consider the following hypothesis tests individually:

$$H_{0i}^c : \lambda_{ci} = 0 \quad \text{versus} \quad H_{1i}^c : \lambda_{ci} \neq 0, \quad i = 1, \ldots, d_c$$

and

$$H_{0j}^r : \lambda_{rj} = 0 \quad \text{versus} \quad H_{1j}^r : \lambda_{rj} \neq 0, \quad j = 1, \ldots, d_r.$$

To draw statistical inference for the row direction $\mathbf{u}_{rj}$ in row-2DLDA, the following assumption is made.

Assumption 1: A set of 2-D observations $\{\mathbf{X}_i\}_{i=1}^n$ consists of k classes and $\mathbf{X}_i | \, j \sim \mathcal{N}_{d_c,d_r}(\mathbf{M}_j, \mathbf{I}, \Sigma_w^r)$, $j = 1, \ldots, k$.

In Assumption 1, besides the assumption of normality, we assume the column within-class covariance $\Sigma_w^c = \mathbf{I}$, which means that the with-class covariance $\mathrm{cov}(\mathrm{vec}(\mathbf{X})|j)$ has the restrictive covariance structure $\Sigma_w^r \otimes \mathbf{I}$. Although this assumption seems strong, it is under this assumption that the LDA is equivalent to the row-2DLDA, as shown by Theorem 1. For our test, whether the discriminant direction $\mathbf{u}_{rj}$ is significant or not is determined under the row-2DLDA, we thus make the Assumption 1 required by row-2DLDA. With Assumption 1, we have the following theorem.

Theorem 3: Under Assumption 1 and the null hypothesis $H_0 : \mathbf{M}_1 = \mathbf{M}_2 = \cdots = \mathbf{M}_k$, the F-statistic for row direction $\mathbf{u}_{rj}$

$$\frac{\hat{\mathbf{u}}_{rj}' \hat{\Sigma}_b^r \hat{\mathbf{u}}_{rj}/(k-1)}{\hat{\mathbf{u}}_{rj}' \hat{\Sigma}_w^r \hat{\mathbf{u}}_{rj}/(n-k)} = \hat{\lambda}_{rj} \cdot \frac{n-k}{k-1} \sim F\big(d_c(k-1), d_c(n-k)\big). \tag{22}$$

Proof: See Appendix B. ∎

Similarly, Assumption 2 is made for column direction $\mathbf{u}_{ci}$.

Assumption 2: A set of 2-D observations $\{\mathbf{X}_i\}_{i=1}^n$ consists of k classes and $\mathbf{X}_i \,|\, j \sim \mathcal{N}_{d_c,d_r}(\mathbf{M}_j, \Sigma_w^c, \mathbf{I})$, $j = 1, \ldots, k$.

Under Assumption 2, it follows that $\mathrm{cov}\big(\mathrm{vec}(\mathbf{X})|j\big) = \mathbf{I} \otimes \Sigma_w^c$, which is consistent with the condition in Theorem 2. With Assumption 2, we have the following theorem.

Theorem 4: Under Assumption 2 and the null hypothesis $H_0 : \mathbf{M}_1 = \cdots = \mathbf{M}_k$, the F-statistic for column direction $\mathbf{u}_{ci}$

$$\frac{\hat{\mathbf{u}}_{ci}' \hat{\Sigma}_b^c \hat{\mathbf{u}}_{ci}/(k-1)}{\hat{\mathbf{u}}_{ci}' \hat{\Sigma}_w^c \hat{\mathbf{u}}_{ci}/(n-k)} = \hat{\lambda}_{ci} \cdot \frac{n-k}{k-1} \sim F\big(d_r(k-1), d_r(n-k)\big). \tag{23}$$

Proof: See Appendix C. ∎

Furthermore, we have the following proposition.

Proposition 3: 1) Under Assumption 1 of row-2DLDA

$$\mathbb{E}(\hat{\Sigma}_b^r) = \Sigma_b^r + \frac{k-1}{n}\Sigma_w^r \tag{24}$$

$$\mathbb{E}(\hat{\Sigma}_w^r) = \frac{n-k}{n}\Sigma_w^r \tag{25}$$

where

$$\Sigma_b^r = \frac{1}{nd_c}\sum\nolimits_j n_j(\mathbf{M}_j - \mathbf{M})'(\mathbf{M}_j - \mathbf{M}). \tag{26}$$

2) Under Assumption 2 of column-2DLDA

$$\mathbb{E}(\hat{\Sigma}_b^c) = \Sigma_b^c + \frac{k-1}{n}\Sigma_w^c$$
$$\mathbb{E}(\hat{\Sigma}_w^c) = \frac{n-k}{n}\Sigma_w^c$$

where

$$\Sigma_b^c = \frac{1}{nd_r}\sum_j n_j(\mathbf{M}_j - \mathbf{M})(\mathbf{M}_j - \mathbf{M})'.$$

Proof: See Appendix D. ∎

Let α be the significance level, say, 0.05, and $F_\alpha(a, b)$ be the upper critical value of the F distribution with a and b degrees of freedom. Assume that $\{\hat{\lambda}_{ci}\}_{i=1}^{d_c}$ and $\{\hat{\lambda}_{rj}\}_{j=1}^{d_r}$ have been in descending order. By Theorems 3 and 4, the subspace dimensionality (q_c, q_r) is determined by

$$q_c = \arg\min_i \hat{\lambda}_{ci} > \frac{k-1}{n-k}F_\alpha(d_c(k-1), d_c(n-k)) \quad (27)$$
$$q_r = \arg\min_j \hat{\lambda}_{rj} > \frac{k-1}{n-k}F_\alpha(d_r(k-1), d_r(n-k)) \quad (28)$$

and the desired column and row discriminant transformations are given by $\tilde{\mathbf{U}}_c = [\hat{\mathbf{u}}_{c1}, \ldots, \hat{\mathbf{u}}_{cq_c}]$ and $\tilde{\mathbf{U}}_r = [\hat{\mathbf{u}}_{r1}, \ldots, \hat{\mathbf{u}}_{rq_r}]$.

C. Algorithm

We now briefly summarize the proposed two-stage RLDA as follows. In the first stage, BLDA is performed and the significant discriminant subspace is determined using the F-tests proposed in Section III-B. In the second stage, RLDA is performed in the obtained BLDA subspace.

Specifically, Tibshirani *et al.* [10] perform a regularized DLDA at the DLDA stage of DLDA+RLDA, namely they add the median of diagonal entries of $\hat{\Sigma}_w$ in (7) into the denominator $\hat{\phi}_{wj}$ in (9) to guard against the possibility of large $\hat{\lambda}_j^I$ simply caused by very small values of $\hat{\phi}_{wj}$. Similarly, we also perform a regularized BLDA at the BLDA stage of BLDA+RLDA, that is, we replace $\hat{\Sigma}_w^c$ in (18) and $\hat{\Sigma}_w^r$ in (19) by $\hat{\Sigma}_w^c(\gamma_1)$ and $\hat{\Sigma}_w^r(\gamma_1)$ via (8). The number of features to be used in subsequent RLDA is determined by the proposed F-tests in Section III-B2 at 0.05 level. The regularization parameter in the RLDA stage is denoted by γ_2. For clarity, the whole algorithm of BLDA plus RLDA is given in Algorithm 1.

1) Computational Complexity Analysis: In this section, we compare the time complexity of BLDA+RLDA and RLDA under the assumption that $d_c d_r > N > \max(d_c, d_r, q_c q_r)$, which covers many real applications including all the four face data sets used in our experiments. The cost of Algorithm 1 is mainly in Lines 1 and 6. Line 1 takes $O(N[d_c d_r(d_c + d_r)])$ [19] for the formation of within-class and between-class matrices. Line 6 takes $O(N(q_c q_r)^2)$ for the RLDA in reduced-dimensional space. Thus, the total computational cost of Algorithm 1 is $O(N[d_c d_r(d_c+d_r)+(q_c q_r)^2])$. For RLDA, the time complexity is $O(N^2 d_c d_r)$ [6]. Therefore, when sample size N is small, e.g., $N < \max(d_c, d_r)$, RLDA could be more efficient than BLDA+RLDA. However, when sample size N is not small and BLDA reduces the dimensionality substantially, BLDA+LDA could be more efficient.

Algorithm 1 BLDA+RLDA for 2-D Data

Input: Data $\{\mathbf{X}_i\}_{i=1}^n$, partition $\cup_{j=1}^k \mathcal{C}_j$ and (γ_1, γ_2).

1: Compute sample mean of each class $\hat{\mathbf{M}}_j$, global sample mean $\hat{\mathbf{M}}$ via (20) and $\hat{\Sigma}_b^c$, $\hat{\Sigma}_b^r$, $\hat{\Sigma}_w^c$, $\hat{\Sigma}_w^r$ via (16)-(19).
2: Replace $\hat{\Sigma}_w^c$ and $\hat{\Sigma}_w^r$ by $\hat{\Sigma}_w^c(\gamma_1)$ and $\hat{\Sigma}_w^r(\gamma_1)$ via (8).
3: Compute $\hat{\mathbf{U}}_c = [\hat{\mathbf{u}}_{c1}, \ldots, \hat{\mathbf{u}}_{cd_c}]$ and $\hat{\mathbf{U}}_r = [\hat{\mathbf{u}}_{r1}, \ldots, \hat{\mathbf{u}}_{rd_r}]$ via (13) and (10) and their corresponding eigenvalues $\{\hat{\lambda}_{ci}\}_{i=1}^{d_c}$ and $\{\hat{\lambda}_{rj}\}_{j=1}^{d_r}$.
4: Determine significant q_c (respectively q_r) column (respectively row) directions via (27) (respectively (28)). Set $\tilde{\mathbf{U}}_c = [\hat{\mathbf{u}}_{c1}, \ldots, \hat{\mathbf{u}}_{cq_c}]$ and $\tilde{\mathbf{U}}_r = [\hat{\mathbf{u}}_{r1}, \ldots, \hat{\mathbf{u}}_{rq_r}]$.
5: Arrange $\mathbf{x}_i = \text{vec}(\tilde{\mathbf{U}}_c'\mathbf{X}_i\tilde{\mathbf{U}}_r), i = 1, \ldots, n$.
6: Perform RLDA [6] with γ_2 on $\{\mathbf{x}_i\}_{i=1}^n$ and compute $\mathbf{y}_i = \mathbf{V}'\mathbf{x}_i, i = 1, \ldots, n$.

Output: $\tilde{\mathbf{U}}_c$, $\tilde{\mathbf{U}}_r$, $\mathbf{V}$ and $\{\mathbf{y}_i\}_{i=1}^n$.

IV. Experiments

In this section, we perform experiments on a number of synthetic and real-world data sets. For the DLDA stage of DLDA+RLDA, we follow [10] to perform a regularized DLDA, as mentioned in Section III-C. For BLDA+RLDA, we use Algorithm 1 in Section III-C, where, unless else stated, we set $\gamma_1 = 0.5$. The number of features used in subsequent RLDA for DLDA+RLDA and BLDA+RLDA is determined by the respective F-tests at 0.05 level.

A. Synthetic Data

In this section, we investigate the classification performance of RLDA, DLDA+RLDA, and BLDA+RLDA.

Let $\mathbf{1}_d$ and $\mathbf{0}_{d_c,d_r}$ stand for the $d \times d$ and $d_c \times d_r$ matrices whose entries all equal 1 and 0, respectively. $\mathbf{I}_2$ is the 2×2 identity matrix. Let $\mathbf{B}$ be a zero matrix except for all the elements of the upper left 2×2 submatrix being 1

$$\mathbf{B} = \begin{pmatrix} \mathbf{1}_2 & \mathbf{0}_{2,d-2} \\ \mathbf{0}_{d-2,2} & \mathbf{0}_{d-2,d-2} \end{pmatrix}. \quad (29)$$

Define $c = d/2 - 2$ and $\mathbf{A} = 1/\sqrt{c}[\mathbf{I}_2; \mathbf{I}_2; \ldots; \mathbf{I}_2; \mathbf{0}_2; \mathbf{0}_2]$, where c is the number of 1's in each column of $\mathbf{A}$. Obviously, $\mathbf{A}$ is a $d \times 2$ matrix whose two columns are orthogonal. The data setting is as follows.

1) *Simulation 1:* 1-D sparse mean shift with independent variables. The observation $\mathbf{X}$ from class j follows $\mathcal{N}(\mathbf{M}_j, \mathbf{I}, \mathbf{I})$, where $\mathbf{M}_j = 2j \cdot \mathbf{B}$. By (3), the theoretical Fisher value of this problem is 20 and each variable has Fisher value 5. Note that the means are sparse and only four variables are useful for classification.
2) *Simulation 2:* 1-D nonsparse mean shift with independent variables. The observation from class j follows $\mathcal{N}(\mathbf{M}_j, \mathbf{I}, \mathbf{I})$, where $\mathbf{M}_j = \mathbf{A}(2j \cdot \mathbf{1}_2)\mathbf{A}'$. The theoretical Fisher value of this problem is still 20, but each variable has Fisher value $5/c$ only. The difference from Simulation 1 is that the means here are not sparse and most variables are useful for classification, though their individual discriminant capability is weak.

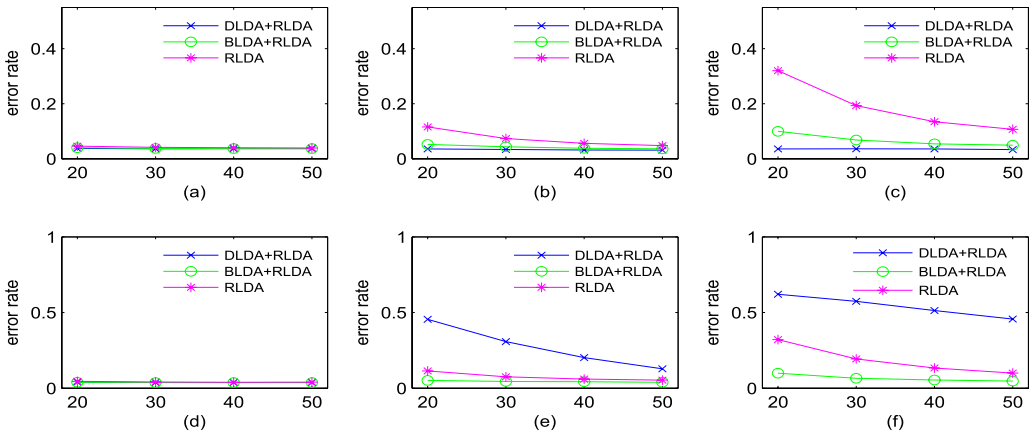


Fig. 2. Results on synthetic data sets generated from Data 1 (row 1) and 2 (row 2) with different data dimensionalities. (a) and (d) $d_c = d_r = 10$. (b) and (e) $d_c = d_r = 20$. (c) and (f) $d_c = d_r = 30$.

1) Setup: For each data setting, we generate 50 data sets. Each data set consists of 400 observations of four classes, and each class contains 100 observations. In each class, the first t observations are chosen for training and the remaining $100-t$ are used for testing. To investigate the performance for different training set sizes, we vary t in the set $\{20, 30, 40, 50\}$. For RLDA and the RLDA stages of DLDA+RLDA and BLDA+RLDA, we enumerate the regularization parameter γ / γ_2 in the set $\{0, 0.01, 0.1, 0.5, 0.9, 0.99\}$ and report the best average test error rates obtained by 1-nearest-neighbor classifier in the 1-D space (as the true data discriminant dimensionality is 1).

2) Results: Fig. 2 shows the classification error rates versus the size of training data per class t. The main observations include the following.

1) When the data dimensionality is not high [Fig. 2(a) and (d)], all methods tend to perform similarly. This is also the case for the higher data dimensionalities [Fig. 2(b), (c), (e), and (f)] as the training number per class t increases.
2) With limited sample size of Simulation 1 [Fig. 2(b) and (c)], DLDA+RLDA performs the best (as expected), which is then followed by BLDA+RLDA, and RLDA is the worst.
3) However, with limited sample size of Simulation 2 [Fig. 2(e) and (f)], DLDA+RLDA now becomes the worst and BLDA+RLDA is the best, which is then followed by RLDA.

Clearly, for both Simulations 1 and 2, BLDA+RLDA performs better than the pure RLDA. This indicates that the utilization of the BLDA stage is successful in improving the pure RLDA. In particular, considering the fact that image data have column and row correlations, as shown in Fig. 1, which appears to be closer to Simulation 2, it would be expected that BLDA+RLDA is more useful than RLDA and DLDA+RLDA for image data. In Sections IV-B and IV-D, we use several image data sets to further examine whether this is the case.

B. Performance on YALE, CMU PIE, and XM2VTS Data Sets

In this section, we compare the proposed BLDA+RLDA with DLDA+RLDA and BMMC+RLDA. The performance of some closely related methods is also included, such as BLDA [23], BPCA [25], principal component analysis (PCA)+LDA [26], BPCA+LDA [14], BMMC+LDA [15], and multiple rank regression (MRR) [27]. Note that MRR has been extended to the support vector machine [28]. The following three real-world data sets are used.

TABLE II
SEVERAL STATISTICS FOR THE USED DATA SETS

Dataset	Total number of images	Image size	Number of classes
YALE	165	64x64	11
CMU PIE	2924	64x64	68
XM2VTS	2360	51x55	295
FERET Fa/Fb	2390	180x170	1195
USPS	11000	16x16	10

1) YALE database[1] contains 165 images of 15 individuals. Each person has 11 images captured under different facial expressions or configurations (e.g., center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink). We subsample the images to the size 64×64.
2) CMU PIE[2] face database contains 41 368 face images of 68 subjects. The facial images for each subject were captured under 13 different poses, 43 different illumination conditions, and with four different expressions. In our experiments, we use the same subdatabase as in [13]. Each person has 43 images, consisting of the frontal face images under different lighting conditions (without expression variations). We subsample the images to the size 64×64.
3) XM2VTS database[3] contains images of 295 individuals. Each individual has eight images taken over a period of four months. The image size is 51×55.

All three face data sets are preprocessed with histogram equilibrium. Several statistics of these data sets are summarized in Table II, and several examples are shown in Fig. 3.

[1] Available from http://cvc.yale.edu/projects/yalefaces/yalefaces.html
[2] Available from http://www.ri.cmu.edu/projects/project_418.html
[3] Available from http://www.face-rec.org/databases/

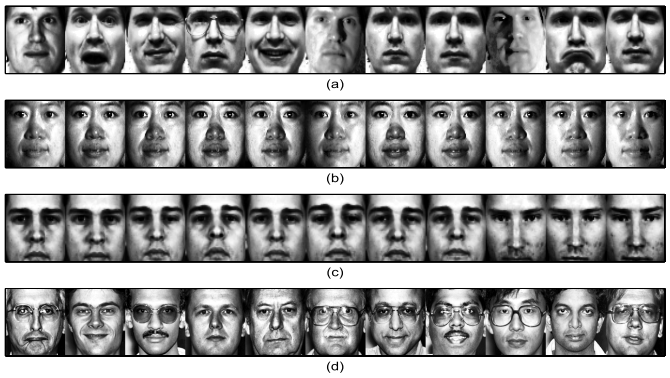


Fig. 3. Example images from four face data sets. (a) YALE. (b) CMU PIE. (c) XM2VTS. (d) FERET.

1) Setup: To measure the misclassification rate, the data are randomly split into a training set containing r samples per class and a test set containing the remaining samples. Several values of r for each data set are investigated. The test classification error rate, averaged over 50 such repetitions, will be reported. For all methods except for MRR, the 1-nearest-neighbor classifier in the lower dimensional space is used to obtain the error rates.

In the PCA stage of PCA+LDA, since the sample covariance is $\hat{\Sigma}_t = \hat{\Sigma}_b + \hat{\Sigma}_w$, with $\hat{\Sigma}_b$ and $\hat{\Sigma}_w$ given by (6) and (7), we use the principal components of $\hat{\Sigma}_t$ whose eigenvalues are larger than $\eta\hat{\sigma}^2$, where $\hat{\sigma}^2 = \mathrm{tr}(\hat{\Sigma}_t)/d$ measures the average variance across all dimensions, and we set $\eta \in \{1, 0.5, 0.1\}$. For the BPCA stage of BPCA+LDA, a similar strategy as in PCA+LDA is applied to the column and row covariance matrices, respectively. In the BMMC stage of BMMC+RLDA (respectively BMMC+LDA), the number of features in RLDA (respectively LDA) is chosen by YD's criterion [15].

The parameter in PCA+LDA and BPCA+LDA is (η, q) (where $\eta \in \{1, 0.5, 0.1\}$ and $q \in \{1, 2, \ldots, k-1\}$), the parameter in BMMC+LDA is q (where $q \in \{1, 2, \ldots, k-1\}$) and the parameter in BLDA and BPCA is (q_c, q_r) (where $q_c \in \{1, 2, \ldots, d_c\}$ and $q_r \in \{1, 2, \ldots, d_r\}$). For BLDA/BPCA, PCA+LDA/BPCA+LDA, and BMMC+LDA, we calculate the average test error rates $e(q_c, q_r)$, $e(\eta, q)$, and $e(q)$, respectively, on the 50 test data sets from 50 random splittings and then report the lowest ones, which is similar as in [9] and [15]. Note that if we choose the parameter by cross validation, the corresponding average test error rate might not be the lowest one.

In contrast, for RLDA and the RLDA stages of DLDA+RLDA, BMMC+RLDA, and BLDA+RLDA, we simply report the results with a fixed parameter setting, namely we set the regularization parameter in (8) to be 0.1 and use the maximal number of discriminant features $q_{\max}(\leq k-1)$, since we find that this simple choice usually yields good overall results, as detailed in Section IV-C. For MRR, we use the MATLAB code[4], which implements the algorithm in [27]. Similar as in [27], the regularization parameter in MRR is chosen in the set $\{1, 10, 50, 100, 1000, 10\,000\}$ by fivefold cross validation.

[4]https://sites.google.com/site/feipingnie/file/TIP-MRR.rar?attredirects=0

2) Results: Fig. 4 shows the typical decreasing-ordered Fisher values λ_j^I's in DLDA, λ_{ci}'s, λ_{rj}'s in BLDA, and eigenvalues λ_{ci}^m's, λ_{rj}^m's in BMMC on the three image data sets, including the number of features determined by their respective criteria. Fig. 5 plots typical evolvements of test error rates versus the number of features used in RLDA, including the error rate corresponding to the selected number of features. Tables III–V summarize the results for all methods, respectively, where the best method is in bold face and marked * if statistically significantly better than the other ones via the paired-sample one-tailed t-test. The - sign in Table IV implies that the DLDA stage fails to select any feature in some repetitions, and hence the average error rate cannot be computed. Table VI collects several computational time results by DLDA+RLDA, BLDA+RLDA, RLDA, and MRR, where the time for MRR is obtained with only a regularization parameter 10. The main observations include the following.

1) BLDA+RLDA versus DLDA/BMMC+RLDA and RLDA.
   a) *Feature Selection:* 1) DLDA usually requires much more number of features than BLDA for recognition, as it ignores the high correlation among pixels of images; 2) BMMC usually selects less number of features than BLDA, though this does not hold when $r = 2, 3$ on CMU PIE (where BMMC includes more number of features, as shown in Fig. 5 and Table IV). Unfortunately, it can be seen clearly from Fig. 5 that this choice is far from the best in terms of recognition performance; and 3) in contrast, the choice in BLDA is found consistently around the best recognition performance.
   b) *Recognition Performance:* BLDA+RLDA performs the best on all three data sets. This reveals that reducing dimensionality for 2-D data by incorporating the correlation among columns and rows has advantages and our proposed feature selection procedure is effective.
   c) *Computational Efficiency:* When the training sample size is small, e.g., $r = 2$ on YALE and CMU PIE, RLDA is computationally more efficient than BLDA+RLDA. However, when the training sample size is not small, e.g., $r = 5$ on XM2VTS and CMU PIE, BLDA+RLDA is more efficient. In contrast, DLDA generally fails to reduce the dimensionality substantially, and hence DLDA+RLDA is comparable with RLDA in terms of computational efficiency.
2) BLDA+RLDA versus PCA/BPCA/BMMC+LDA and BLDA/BPCA. BLDA+RLDA outperforms BLDA/BPCA and PCA/BPCA/BMMC+LDA substantially. This shows the superiority of BLDA+RLDA for image data.
3) BLDA+RLDA versus MRR. BLDA+RLDA performs significantly better than MRR while being much more computationally efficient.
4) BMMC+RLDA versus BMMC+LDA. BMMC+RLDA often outperforms BMMC+LDA substantially, except for the case when a small number of

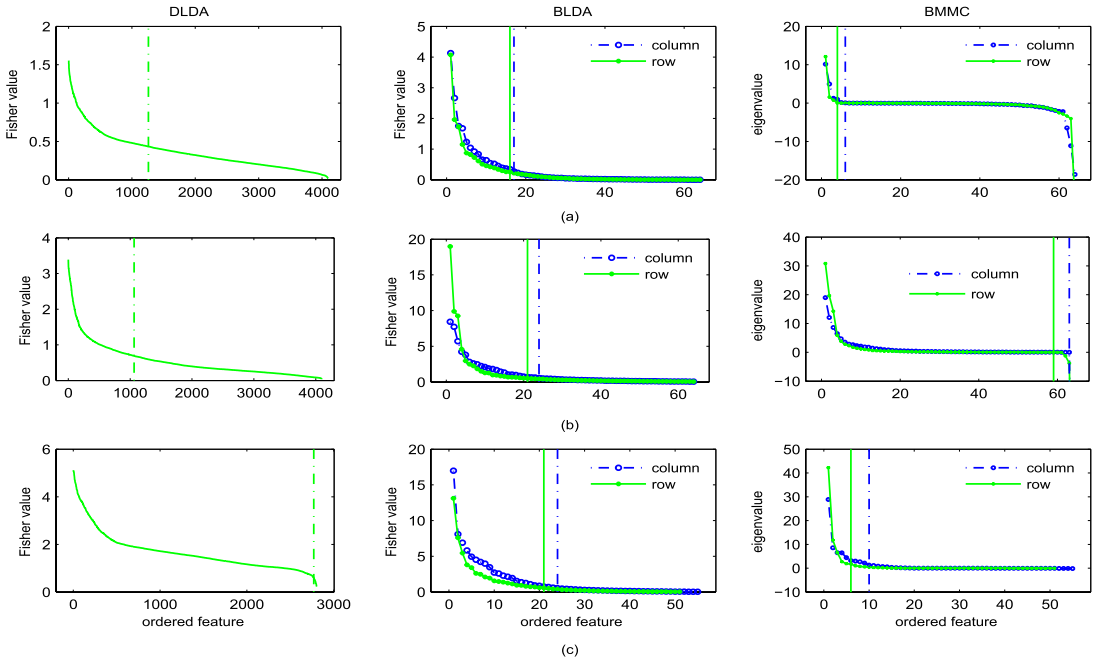


Fig. 4. Distribution of Fisher values/eigenvalues on real data sets: DLDA (column 1), BLDA (column 2), and BMMC (column 3) on (a) YALE with $r = 5$ (row 1), (b) CMU PIE with $r = 2$ (row 2), and (c) XM2VTS with $r = 3$ (row 3). The vertical lines signal the automatically chosen dimensionalities.

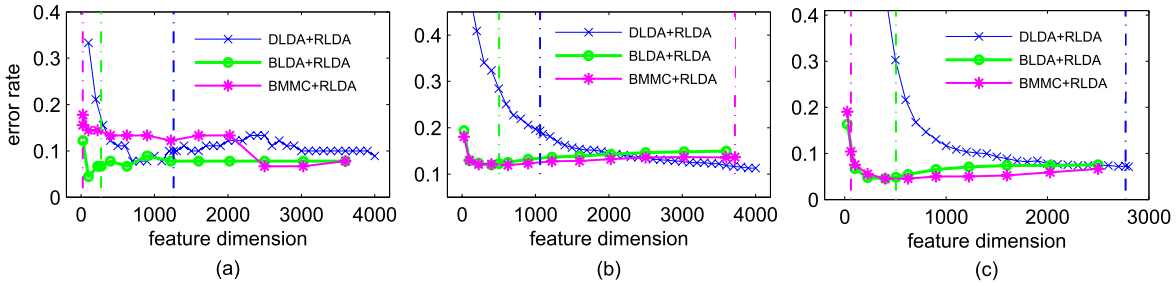


Fig. 5. Typical evolvements of test error rates versus number of used features in RLDA on real data sets. The vertical lines signal the automatically chosen dimensionalities used in the subsequent RLDA stage. (a) YALE. (b) CMU PIE. (c) XM2VTS.

TABLE III
MEAN (AND STANDARD DEVIATION) OF ERROR RATES AND NUMBER OF SELECTED FEATURES ON YALE

	Method	Training number per individual 3	5	7
error rate	DLDA+RLDA	23.7±8.0	14.0±4.7	10.3±5.0
	BMMC+RLDA	18.8±3.5	19.7±5.1	26.6±6.7
	BLDA+RLDA	16.4±5.8	*__10.9__±5.9	*__7.4__±5.5
	MRR	27.2±5.4	20.7±6.2	17.3±7.2
	RLDA	__16.3__±4.2	12.3±5.3	9.5±5.6
	BLDA	35.0±7.0	22.2±6.0	15.3±7.9
	BPCA	21.2±2.9	18.4±4.5	14.7±6.6
	PCA+LDA	17.5±4.2	13.7±4.9	10.4±5.4
	BPCA+LDA	17.9±4.4	14.0±5.1	11.2±6.1
	BMMC+LDA	32.5±7.8	24.2±6.6	25.7±6.0
number of features	DLDA+RLDA	803± 586	1429± 624	2035± 488
	BMMC+RLDA	89± 46	29± 11	15± 7
	BLDA+RLDA	215± 82	247± 63	280± 50

TABLE IV
MEAN (AND STANDARD DEVIATION) OF ERROR RATES AND NUMBER OF SELECTED FEATURES ON CMU PIE

Method	Training number per individual error rate 2	3	4	5
DLDA+RLDA	-	20.4±15.1	6.4±4.4	3.6±3.6
BMMC+RLDA	16.8±14.2	8.5±7.7	3.1±2.7	2.8±2.7
BLDA+RLDA	__15.8__±9.5	__7.5__±6.3	__2.6__±2.7	__2.3__±2.9
MRR	47.0±9.3	22.1±11.4	12.6±10.5	7.7±5.8
RLDA	18.3±13.1	9.2±8.5	2.8±3.1	__2.3__±2.9
BLDA	20.4±14.5	10.9±10.8	4.4±4.5	3.9±4.2
BPCA	29.7±9.7	22.0±8.5	16.0±7.4	12.0±4.8
PCA+LDA	17.9±9.9	9.7±7.8	3.4±3.2	3.0±3.2
BPCA+LDA	16.7±10.4	8.7±7.4	3.3±2.8	2.8±2.7
BMMC+LDA	24.4±20.0	13.6±12.2	7.2±4.8	6.1±5.0
	number of features			
DLDA+RLDA	-	1116±1176	1153±941	1576±840
BMMC+RLDA	1107±1161	835±1002	383±430	390±452
BLDA+RLDA	538±993	472±603	396±236	531±274

features are selected in the BMMC stage (e.g., on XM2VTS).

C. BLDA+RLDA Under Various Parameter Settings

As shown in Algorithm 1, two regularization parameters in BLDA+RLDA have to be predetermined, namely γ_1 and γ_2. In addition, the number of discriminant features q in the RLDA stage can also be viewed as a tuning parameter. In our experiments in Section IV-B, we simply report the results with a fixed parameter setting ($\gamma_1 = 0.5$,

TABLE V
MEAN (AND STANDARD DEVIATION) OF ERROR RATES AND NUMBER OF SELECTED FEATURES ON XM2VTS

Method	Training number per individual error rate 2	3	4	5
DLDA+RLDA	11.3±4.0	5.0±1.7	3.0±1.3	2.1±1.1
BMMC+RLDA	12.8±4.4	8.3±2.1	5.5±2.1	4.1±2.0
MRR	36.1±3.6	20.3±3.7	14.8±3.4	11.6±2.9
BLDA+RLDA	*__8.5__±4.4	*__3.7__±1.3	*__2.3__±1.0	*__1.6__±0.8
RLDA	10.6±4.2	4.9±1.7	3.0±1.3	2.1±1.1
BLDA	12.6±5.0	6.9±1.9	4.8±1.7	3.6±1.4
BPCA	21.7±3.5	15.5±2.4	11.4±2.5	8.9±2.7
PCA+LDA	12.6±5.3	5.4±1.7	3.0±1.4	2.0±1.0
BPCA+LDA	11.4±4.8	5.3±1.5	2.9±1.2	1.9±1.0
BMMC+LDA	10.8±4.6	5.9±1.7	3.7±1.5	2.7±1.3
	number of features			
DLDA+RLDA	2343±310	2739±31	2777± 7	2790± 2
BMMC+RLDA	64± 7	53± 7	46± 3	45± 1
BLDA+RLDA	383± 70	461± 45	508± 33	549± 21

TABLE VI
AVERAGE TRAINING TIME IN SECONDS. r IS THE NUMBER OF TRAINING IMAGES PER PERSON

Dataset	r	BLDA+RLDA	DLDA+RLDA	RLDA	MRR
YALE	3	0.04	0.02	0.02	6.22
	7	0.05	0.04	0.04	30.34
CMU PIE	2	0.64	0.44	0.59	38.12
	5	0.78	0.85	1.06	340.17
XM2VTS	2	3.68	3.88	4.04	548.30
	5	5.77	9.71	9.49	7907.60

$\gamma_2 = 0.1$, and $q = q_{\max}$). In this section, we examine how these parameters affect the performance of BLDA+RLDA. We use the 50 random splittings of YALE, CMU PIE, and XM2VTS in Section IV-B with the training persons per individual $r = 3$, on which we perform the following three experiments.

1) Variations Over q: In this experiment, we fix $\gamma_1 = 0.5$, $\gamma_2 = 0.1$, and let q vary in the set $\{1, 2, \ldots, q_{\max}\}$. For comparison, we also include the performance of RLDA with $\gamma = 0.1$ and variations over q. Fig. 6 shows the average test error rates versus various values of q. The following can be observed from Fig. 6.

1) For YALE and CMU PIE, the larger the value of q, the better both BLDA+RLDA and RLDA perform.
2) For XM2VTS, the performance of BLDA+RLDA is stable as long as q is large enough, while the performance of RLDA for the maximum $q_{\max}$ is slightly worse than the best.

Overall, the maximum $q_{\max}$ yields satisfactory performance for both BLDA+RLDA and RLDA.

2) Variations Over γ_2: In this experiment, we fix $\gamma_1 = 0.5$, $q = q_{\max}$, and let γ_2 vary in the set $\{0, 0.01, 0.1, 0.5, 0.9, 0.99\}$. Fig. 7 shows the average test error rates versus various values of γ_2. Looking at Fig. 7, although γ_2 has a significant impact on the classification performance, the choice $\gamma_2 = 0.1$ or near 0.1 generally leads to good performance for BLDA+RLDA. This is also the case for RLDA with $\gamma = 0.1$ or around 0.1.

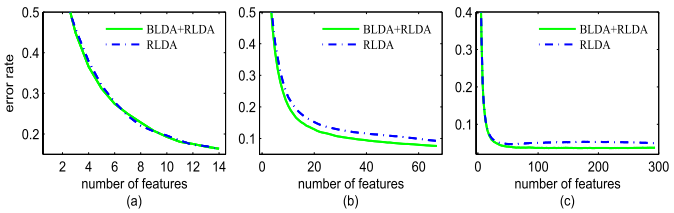


Fig. 6. Test error rates versus various values of q by BLDA+RLDA and RLDA on (a) YALE, (b) CMU PIE, and (c) XM2VTS data sets.

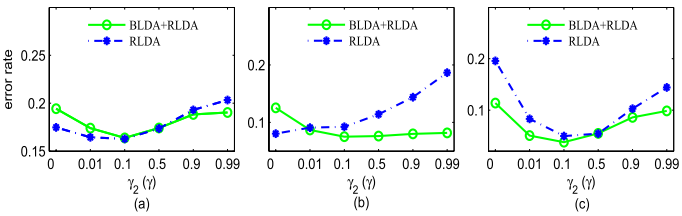


Fig. 7. Test error rates versus various values of γ_2 (respectively γ) for BLDA+RLDA (respectively RLDA) on (a) YALE, (b) CMU PIE, and (c) XM2VTS data sets.

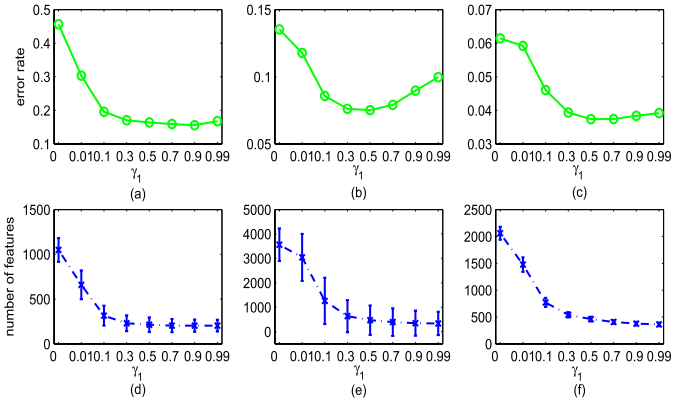


Fig. 8. (a)–(c) Test error rates and (d)–(f) number of chosen features in Stage 1 versus various values of γ_1 by BLDA+RLDA on YALE, CMU PIE, and XM2VTS data sets.

3) Variations Over γ_1: In this experiment, we fix $\gamma_2 = 0.1$, $q = q_{\max}$, and let γ_1 vary in the set $\{0, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. The results for various values of γ_1 are shown in Fig. 8, from which it can be seen that γ_1 substantially affects the classification performance and dimensionality reduction in the first stage. Small values of γ_1 (e.g., 0 and 0.01) are not effective. However, when γ_1 is around 0.5, the classification performance is satisfactory, and when $\gamma_1 \geq 0.3$, the number of chosen features in the first stage tends to be stable.

Overall, the three experiments reveal that the fixed setting $\gamma_1 = 0.5, \gamma_2 = 0.1, q = q_{\max}$ performs satisfactorily on YALE, CMU PIE, and XM2VTS data sets. Due to its simplicity, we use this setting in all our experiments on real data.

D. Further Comparison on FERET and USPS Data Sets

In this section, we further investigate the performance of DLDA+RLDA, BMMC+RLDA, BLDA+RLDA, and RLDA on the FERET and USPS data sets.

TABLE VII
MEAN (AND STANDARD DEVIATION) OF ERROR RATES AND NUMBER OF SELECTED FEATURES ON FERET DATABASE

	Number of training persons			
	error rate			
method	30	50	70	100
DLDA+RLDA	17.5±1.3	12.5±1.1	10.0±0.9	7.9±0.7
BMMC+RLDA	14.6±1.2	10.1±1.0	8.0±0.8	6.5±0.7
BLDA+RLDA	**14.3**±2.0	*__9.6__±1.7	*__7.4__±1.2	*__5.6__±0.8
RLDA	17.7±1.2	12.6±1.1	10.0±0.8	7.9±0.8
BMMC+LDA	19.2±1.5	16.9±1.3	17.6±1.3	19.9±2.1
	number of features			
DLDA+RLDA	27203 ±1367	29071 ±648	29668 ±414	30080 ±240
BMMC+RLDA	176± 14	175± 12	179± 12	179± 12
BLDA+RLDA	1176±129	1235±105	1268± 87	1277± 72

TABLE VIII
MEAN (AND STANDARD DEVIATION) OF ERROR RATES AND NUMBER OF SELECTED FEATURES ON USPS

	Number of training samples per digit			
	error rate			
Method	100	200	300	500
DLDA+RLDA	10.9±0.5	8.8±0.3	8.0±0.3	7.3±0.3
BMMC+RLDA	-	-	-	-
BLDA+RLDA	*__9.9__±0.5	*__8.4__±0.3	*__7.7__±0.3	*__7.1__±0.3
MRR	16.6±0.8	15.6±0.5	15.3±0.5	15.1±0.5
RLDA	10.8±0.5	8.8±0.3	8.0±0.3	7.3±0.3
	number of features			
DLDA+RLDA	239± 2	247± 2	252± 1	253± 1
BMMC+RLDA	-	-	-	-
BLDA+RLDA	82± 7	92± 4	103± 6	110± 2

1) FERET database [29] consists of 14 051 grayscale images taken under different expressions, poses, illuminations, and occlusion. In our experiments, we use the Fa/Fb subdatabase, which contains 1195 individuals. Each person has two images. We subsample the images to the size 180×170.
2) USPS handwritten digit database[5] contains 11 000 images of 10 digits, derived from the well-known USPS set of handwritten digits. Each digit has 1100 grayscale images with size 16×16.

Several statistics of the two data sets are summarized in Table II. The FERET database is preprocessed with histogram equilibrium, and several examples are shown in Fig. 3.

FERET and USPS exhibit very different data characteristics from the three data sets in Section IV-B: FERET contains a large number of people (1195) and USPS has sufficient data samples while its data dimensionality (256) is not very high. The experiment on USPS database is similar as in Section IV-B. For FERET Fa/Fb database, we use Fa as the gallery set and Fb as the probe set. We are interested in investigating, given some people for training, which method has the best generalization performance on new people. In addition, to examine the performance for different training set sizes, r people are randomly chosen for training, and the remaining people (i.e., $1195 - r$) are used for testing. The average test error across 50 such repetitions will be reported. Note that MRR requires using all classes of the data for training while only a few classes are available in this case. Thus, MRR is not used in this experiment.

The results are summarized in Tables VII and VIII, respectively. As in Section IV-B, the best method is in bold face and marked * if statistically significantly better. The - sign in Table VIII indicates that the BMMC stage fails to include any feature in all 50 repetitions, and hence the average error rate cannot be computed. It can be observed from Tables VII and VIII that the results are generally consistent with those in Section IV-B. BLDA+RLDA performs the best. DLDA seems inefficient in terms of initial dimension reduction. BMMC disappointedly cannot screen out any useful feature on USPS since all eigenvalues λ_{ci}^m's and λ_{rj}^m's are negative.

[5]Available from http://www.cs.nyu.edu/~roweis/data.html

E. Dimension Reduction in the First Stage

In this section, we compare in detail our proposed statistical tests with YD's criterion [13] on determining the discriminant subspace dimensionality.

1) Theoretical Analysis: We focus on analyzing row-2DLDA only, but the analysis is applicable to column-2DLDA as well. For row-2DLDA, by (24) and (25) in Proposition 3, we have that $\mathbb{E}(\hat{\Sigma}_b^r) \to \Sigma_b^r$ and $\mathbb{E}(\hat{\Sigma}_w^r) \to \Sigma_w^r$, as n goes to infinity. Thus, YD's criterion approximately uses the positive eigenvalues of $\Sigma_b^r - \Sigma_w^r$.

Let us consider Simulation 1 with $d = 10$. By (26), we have $\Sigma_b^r = \mathbf{B}$, where $\mathbf{B}$ is given by (29). The largest eigenvalue of Σ_b^r is 2, and all the remaining ones are 0. Since $\Sigma_w^r = \mathbf{I}$, $\Sigma_b^r - \Sigma_w^r$ has a positive eigenvalue. Now, consider Simulation 1 with $d = 40$. By (26), we get $\Sigma_b^r = 0.25 \cdot \mathbf{B}$, where $\mathbf{B}$ is given by (29). The largest eigenvalue of Σ_b^r is 0.5, and all the remaining ones are 0. Since $\Sigma_w^r = \mathbf{I}$, all the eigenvalues of $\Sigma_b^r - \Sigma_w^r$ are negative. Therefore, in this case, it is impossible for YD's criterion to include any feature.

With Simulation 1, only four discriminant features are useful for discrimination and all the remaining ones are not useful. Due to the divisor d_c in (26), the largest eigenvalue of Σ_b^r becomes smaller as d_c increases. This is the reason why YD's criterion fails when $d = 40$. In contrast, our proposed test for row directions compares Σ_b^r with Σ_w^r in a completely different manner. It corrects $\hat{\Sigma}_b^r$ in (24) by $k-1$ and $\hat{\Sigma}_w^r$ in (25) by $n - k$. This makes the second term in (24) and the term in (25) the same, and thus makes our proposed test free from the above problem.

2) Empirical Analysis: To examine the above theoretical analysis, we use the 50 data sets from Simulation 1 with training number per class $t = 50$ to compare BMMC+RLDA with BLDA+RLDA. The results are shown in Table IX, from which it can be observed that: 1) when $d = 10$, YD's criterion and our proposed tests perform similarly, and BMMC+RLDA and BLDA+RLDA obtain similar error rates and 2) however, when $d = 40$, BMMC+RLDA fails completely due to the failure of YD's criterion to include any feature, in strong contrast with the satisfactory performance of our proposed tests.

The above theoretical and empirical analyses provide an explanation why BMMC tends to select too small number of

TABLE IX
MEAN (AND STANDARD DEVIATION) OF NUMBER OF SELECTED FEATURES AND ERROR RATES ON DATA 1

Measure	(d_c, d_r)	BMMC+RLDA	BLDA+RLDA
number of features	(10,10)	1±0	8±3
	(40,40)	0±0	185±18
error rate	(10,10)	3.4±1.3	3.4±1.2
	(40,40)	-	7.5±2.1

features on YALE and XM2VTS data sets and even fails to select any feature on USPS data set.

V. CONCLUSION

To improve RLDA on 2-D data such as images, we proposed in this paper a new two-stage method, namely BLDA plus RLDA, in which both the BLDA and RLDA stages are based on the well-known Fisher criterion. The key to our method is that BDLDA can incorporate the correlations among columns and rows into reducing the dimensionality of 2-D data in an appropriate manner via a simple yet effective feature selection procedure. The empirical results on several benchmark image data sets demonstrate that BDLDA can reduce the data dimensionality substantially while keeping the useful discriminant information satisfactorily, and hence the proposed two-stage method, BLDA plus RLDA, is compared favorably with some closely related methods.

Nowadays, the observations in many real-world data are 3-D or higher order tensors. For example, color images and grayscale video sequences are 3-D and color video sequences are 4-D. The BMMC+LDA has been adapted to handle tensor structure data in [14]. It would be interesting to extend our BLDA+RLDA to accommodate such data.

APPENDIX A
PROOF FOR PROPOSITION 2

Proof: When $d = 1$, this is the classical result of univariate ANOVA. When $d > 1$, from [23, Exercise 5.10, p. 222], we have: 1) $\hat{\Sigma}_w$ is independent with $\hat{\Sigma}_b$ and 2) $n\hat{\Sigma}_w \sim \mathcal{W}_d(n-k, \sigma^2\mathbf{I})$ and $n\hat{\Sigma}_b \sim \mathcal{W}_d(k-1, \sigma^2\mathbf{I})$. Using Proposition 1 and the additive property of χ^2 distribution, we have $\mathrm{tr}(n\hat{\Sigma}_w)/\sigma^2 \sim \chi^2(d(n-k))$ and $\mathrm{tr}(n\hat{\Sigma}_b)/\sigma^2 \sim \chi^2(d(k-1))$. By 1), we have that $\mathrm{tr}(\hat{\Sigma}_w)$ is independent with $\mathrm{tr}(\hat{\Sigma}_b)$ and hence (21) holds. This completes the proof. □

APPENDIX B
PROOF FOR THEOREM 3

Proof: Under Assumption 1, for an arbitrary nonzero vector $\mathbf{u} \in \mathbb{R}^{d_r}$, we have $\mathbf{X}_i\mathbf{u} \sim \mathcal{N}_{d_c}(\mathbf{M}_j\mathbf{u}, (\mathbf{u}'\Sigma_w^r\mathbf{u})\cdot\mathbf{I})$. Under the null hypothesis $H_0 : \mathbf{M}_1 = \mathbf{M}_2 = \cdots = \mathbf{M}_k$, substituting $\mathbf{x}_i = \mathbf{X}_i\mathbf{u}$, $\boldsymbol{\mu}_j = \mathbf{M}_j\mathbf{u}$, and $\sigma^2 = \mathbf{u}'\Sigma_w^r\mathbf{u}$ into Proposition 2, we obtain $\mathrm{tr}(n\hat{\Sigma}_w) = nd_c\mathbf{u}'\hat{\Sigma}_w^r\mathbf{u}$ and $\mathrm{tr}(n\hat{\Sigma}_b) = nd_c\mathbf{u}'\hat{\Sigma}_b^r\mathbf{u}$, $nd_c\mathbf{u}'\hat{\Sigma}_w^r\mathbf{u}/\mathbf{u}'\Sigma_c\mathbf{u} \sim \chi^2(d_c(n-k))$, $nd_c\mathbf{u}'\hat{\Sigma}_b^r\mathbf{u}/\mathbf{u}'\Sigma_c\mathbf{u} \sim \chi^2(d_c(k-1))$, $\mathbf{u}'\hat{\Sigma}_b^r\mathbf{u}$ is independent with $\mathbf{u}'\hat{\Sigma}_w^r\mathbf{u}$, and

$$\frac{\mathbf{u}'\hat{\Sigma}_b^r\mathbf{u}/(k-1)}{\mathbf{u}'\hat{\Sigma}_w^r\mathbf{u}/(n-k)} \sim F\big(d_c(k-1), d_c(n-k)\big).$$

Replacing $\mathbf{u}$ by $\hat{\mathbf{u}}_{rj}$ and using (12), we obtain the F-statistic for row direction $\hat{\mathbf{u}}_{rj}$ in (22). The proof is completed. □

APPENDIX C
PROOF FOR THEOREM 4

Proof: The proof is similar as Theorem 3. Under H_0 and Assumption 2, $\mathbf{u}'\hat{\Sigma}_b^c\mathbf{u}$ is independent with $\mathbf{u}'\hat{\Sigma}_w^c\mathbf{u}$, $nd_r\mathbf{u}'\hat{\Sigma}_b^c\mathbf{u}/\mathbf{u}'\Sigma_r\mathbf{u} \sim \chi^2(d_r(k-1))$, $nd_r\mathbf{u}'\hat{\Sigma}_w^c\mathbf{u}/\mathbf{u}'\Sigma_r\mathbf{u} \sim \chi^2(d_r(n-k))$, and thus we obtain the F-statistic for column direction $\hat{\mathbf{u}}_{ci}$ in (23). The proof is concluded. □

APPENDIX D
PROOF FOR PROPOSITION 3

Proof: 1) For clarity, denote $\mathbf{X}_i|\, j$ by $\mathbf{X}_{ji}$. Let $\mathbf{M} = \sum_{j=1}^k n_j\mathbf{M}_j/n$ and $\mathbf{T}_j = \mathbf{M}_j - \mathbf{M}$, from which we have $\sum_{j=1}^k n_j\mathbf{T}_j = \mathbf{0}$. The assumption that $\mathbf{X}_{ji} \sim \mathcal{N}_{d_c,d_r}(\mathbf{M}_j, \mathbf{I}, \Sigma_w^r)$, $j = 1,\ldots,k, i = 1,\ldots,n_j$, i.i.d. can be rewritten as

$$\begin{cases} \mathbf{X}_{ji} = \mathbf{M} + \mathbf{T}_j + \boldsymbol{\epsilon}_{ji} \\ \boldsymbol{\epsilon}_{ji} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \Sigma_w^r), \quad j = 1,\ldots,k, \quad i = 1,\ldots,n_j, \quad i.i.d. \end{cases} \tag{30}$$

From (30), we have $\hat{\mathbf{M}}_j = \mathbf{M} + \mathbf{T}_j + \bar{\boldsymbol{\epsilon}}_j$ and $\hat{\mathbf{M}} = \mathbf{M} + \bar{\boldsymbol{\epsilon}}$, where $\hat{\mathbf{M}}_j$ and $\hat{\mathbf{M}}$ are given by (20) and, similarly

$$\bar{\boldsymbol{\epsilon}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{X}_{ji}, \quad \bar{\boldsymbol{\epsilon}} = \frac{1}{n}\sum_j\sum_{i=1}^{n_j}\mathbf{X}_{ji}. \tag{31}$$

By (31), we have $\bar{\boldsymbol{\epsilon}}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \Sigma_w^r/n_j)$ and $\bar{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}, \Sigma_w^r/n)$.

Next, we find the expectation of $\hat{\Sigma}_w^r$ in (19)

$$\begin{aligned} \mathbb{E}(\hat{\Sigma}_w^r) &= \frac{1}{nd_c}\sum\nolimits_j\sum\nolimits_{i=1}^{n_j}\mathbb{E}[(\boldsymbol{\epsilon}_{ji} - \bar{\boldsymbol{\epsilon}}_j)'(\boldsymbol{\epsilon}_{ji} - \bar{\boldsymbol{\epsilon}}_j)] \\ &= \frac{1}{nd_c}\sum\nolimits_j\Big[\sum\nolimits_{i=1}^{n_j}\mathbb{E}(\boldsymbol{\epsilon}_{ji}'\boldsymbol{\epsilon}_{ji}) - n_j\mathbb{E}(\bar{\boldsymbol{\epsilon}}_j'\bar{\boldsymbol{\epsilon}}_j)\Big] \\ &= \frac{1}{nd_c}\sum_j\big[n_jd_c\Sigma_w^r - n_jd_c\Sigma_w^r/n_j\big] \\ &= \frac{n-k}{n}\Sigma_w^r. \end{aligned}$$

The expectation of $\hat{\Sigma}_b^r$ in (17) is

$$\begin{aligned} \mathbb{E}(\hat{\Sigma}_b^r) &= \mathbb{E}\Big[\frac{1}{nd_c}\sum\nolimits_j n_j(\mathbf{T}_j + \bar{\boldsymbol{\epsilon}}_j - \bar{\boldsymbol{\epsilon}})'(\mathbf{T}_j + \bar{\boldsymbol{\epsilon}}_j - \bar{\boldsymbol{\epsilon}})\Big] \\ &= \frac{1}{nd_c}\sum\nolimits_j n_j\big[\mathbf{T}_j'\mathbf{T}_j + \mathbb{E}(\bar{\boldsymbol{\epsilon}}_j - \bar{\boldsymbol{\epsilon}})'(\bar{\boldsymbol{\epsilon}}_j - \bar{\boldsymbol{\epsilon}})\big] \\ &= \Sigma_b^r + \frac{1}{nd_c}\Big[\sum\nolimits_j n_j\mathbb{E}(\bar{\boldsymbol{\epsilon}}_j'\bar{\boldsymbol{\epsilon}}_j) - n\mathbb{E}(\bar{\boldsymbol{\epsilon}}'\bar{\boldsymbol{\epsilon}})\Big] \\ &= \Sigma_b^r + \frac{1}{nd_c}\Big[\sum_j n_jd_c\Sigma_w^r/n_j - nd_c\Sigma_w^r/n\Big] \\ &= \Sigma_b^r + \frac{k-1}{n}\Sigma_w^r. \end{aligned}$$

2) The proof is similar as 1) and hence omitted here. □

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the three anonymous reviewers for their valuable comments and suggestions that strengthened this paper substantially.

REFERENCES

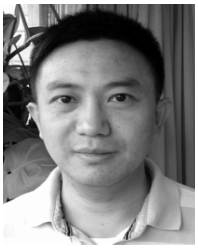
[1] D. M. Witten and R. Tibshirani, "Penalized classification using Fisher's linear discriminant," *J. Roy. Statist. Soc. B*, vol. 73, no. 5, pp. 753–772, 2011.
[2] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Ann. Statist.*, vol. 23, no. 1, pp. 73–102, 1995.
[3] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.
[4] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, Jan. 2005.
[5] Z. Zhang, G. Dai, C. Xu, and M. I. Jordan, "Regularized discriminant analysis, ridge regression and beyond," *J. Mach. Learn. Res.*, vol. 11, pp. 2199–2228, Aug. 2010.
[6] S. Ji and J. Ye, "Generalized linear discriminant analysis: A unified framework and efficient model selection," *IEEE Trans. Neural Netw.*, vol. 19, no. 10, pp. 1768–1782, Oct. 2008.
[7] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Statist. Soc. B*, vol. 70, no. 5, pp. 849–911, 2008.
[8] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Statist. Assoc.*, vol. 97, no. 457, pp. 77–87, 2002.
[9] W.-H. Yang, D.-Q. Dai, and H. Yan, "Feature extraction and uncorrelated discriminant analysis for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 601–614, May 2008.
[10] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 99, no. 10, pp. 6567–6572, 2002.
[11] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.
[12] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2005.
[13] J. Fan, Y. Feng, and X. Tong, "A road to classification in high dimensional space: The regularized optimal affine discriminant," *J. Roy. Statist. Soc. B*, vol. 74, no. 4, pp. 745–771, 2012.
[14] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA: A novel fast feature extraction technique for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 4, pp. 946–953, Aug. 2006.
[15] W.-H. Yang and D.-Q. Dai, "Two-dimensional maximum margin feature extraction for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 4, pp. 1002–1012, Aug. 2009.
[16] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.
[17] X. R. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.
[18] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Proc. 8th Adv. Neural Inf. Process. Syst.*, 2005, pp. 1569–1576.
[19] J. Zhao, P. L. H. Yu, L. Shi, and S. Li, "Separable linear discriminant analysis," *Comput. Statist. Data Anal.*, vol. 56, no. 12, pp. 4290–4300, 2012.
[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.
[21] K. Fukunaga, *Introduction to Statistical Pattern Classification*. New York, NY, USA: Academic, 1990.
[22] H. Xiong, M. N. S. Swamy, and M. O. Ahmad, "Two-dimensional FLD for face recognition," *Pattern Recognit.*, vol. 38, no. 7, pp. 1121–1124, 2005.
[23] S. Noushath, G. H. Kumar, and P. Shivakumara, "$(2D)^2$ LDA: An efficient approach for face recognition," *Pattern Recognit.*, vol. 39, no. 7, pp. 1396–1400, 2006.
[24] J. Wang, *Multivariate Statistical Analysis*, 1st ed. Beijing, China: Science Press, 2008. [Online]. Available: http://www.sciencep.com
[25] D. Zhang and Z.-H. Zhou, "$(2D)^2$ PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, 2005.
[26] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[27] C. Hou, F. Nie, D. Yi, and Y. Wu, "Efficient image classification via multiple rank regression," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 340–352, Jan. 2013.
[28] C. Hou, F. Nie, C. Zhang, D. Yi, and Y. Wu, "Multiple rank multi-linear SVM for matrix data classification," *Pattern Recognit.*, vol. 47, no. 1, pp. 454–469, 2014.
[29] P. J. Philips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

Jianhua Zhao (S'09–M'09) received the Ph.D. degree in statistics from the University of Hong Kong, Hong Kong, in 2009.

He is currently a full-time Professor with the Department of Statistics, Yunnan University of Finance and Economics, Kunming, China. His current research interests include statistical machine learning and data mining and computational statistics.

Lei Shi received the Ph.D. degree in statistics from the University of Calgary, Calgary, AB, Canada.

He joined the Department of Statistics with the Yunnan University of Finance and Economics, Kunming, China, in 2007, as a Chief Professor, and is currently a Cheung Kong Professor with the Ministry of Education of China, Beijing, China. His current research interests include data mining, financial time series, ecology statistics, regression diagnostics, and meta-analysis.



Ji Zhu received the Ph.D. degree in statistics from Stanford University, Stanford, CA, USA, in 2003.

He is currently a full-time Professor with the Department of Statistics, University of Michigan, Ann Arbor, MI, USA. His current research interests include statistical learning and data mining, statistical network analysis, and statistical modeling in computational biology and health sciences.