Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 3919472, 10 pages
http://dx.doi.org/10.1155/2016/3919472

Research Article

A Complete Subspace Analysis of Linear Discriminant Analysis and Its Robust Implementation

Zhicheng Lu and Zhizheng Liang

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

Correspondence should be addressed to Zhizheng Liang; liang@cumt.edu.cn

Received 21 September 2016; Revised 27 October 2016; Accepted 7 November 2016

Academic Editor: Ping Feng Pai

Linear discriminant analysis has been widely studied in data mining and pattern recognition. However, when performing the eigen-decomposition on the matrix pair (within-class scatter matrix and between-class scatter matrix) in some cases, one can find that there exist some degenerated eigenvalues, thereby resulting in indistinguishability of information from the eigen-subspace corresponding to some degenerated eigenvalue. In order to address this problem, we revisit linear discriminant analysis in this paper and propose a stable and effective algorithm for linear discriminant analysis in terms of an optimization criterion. By discussing the properties of the optimization criterion, we find that the eigenvectors in some eigen-subspaces may be indistinguishable if the degenerated eigenvalue occurs. Inspired from the idea of the maximum margin criterion (MMC), we embed MMC into the eigen-subspace corresponding to the degenerated eigenvalue to exploit discriminability of the eigenvectors in the eigen-subspace. Since the proposed algorithm can deal with the degenerated case of eigenvalues, it not only handles the small-sample-size problem but also enables us to select projection vectors from the null space of the between-class scatter matrix. Extensive experiments on several face images and microarray data sets are conducted to evaluate the proposed algorithm in terms of the classification performance, and experimental results show that our method has smaller standard deviations than other methods in most cases.

1. Introduction

Linear discriminant analysis (LDA) [1–4] plays an important role in data analysis and has been widely used in many fields such as data mining and pattern recognition [5]. The main aim of LDA is to find optimal projection vectors by simultaneously minimizing the within-class distance and maximizing the between-class distance in the projection space and optimal projection vectors can be achieved by solving a generalized eigenvalue problem. In solving classical LDA, the within-class scatter matrix is required to be nonsingular in the general case. However, in many applications such as text classification and face recognition [6], the within-class scatter matrix is often singular since the dimension of data we deal with is much bigger than the number of data points. This is known as the small-sample-size (SSS) problem.

In the past several decades, various variants of LDA [7–10] have been proposed to address the problems of high-dimensional data and the SSS problem. It is noted that most of LDA-based methods are divided into four categories in terms of the combination of spaces of the within-class scatter and between-class scatter matrices [11].

The first category of these methods is to consider the range space of the within-class scatter matrix and the range space of the between-class scatter matrix. The typical algorithm of this category is the Fisherface [1] method where PCA is first employed to reduce the dimension of features to make the within-class scatter matrix be full-rank and then the standard LDA is performed. In the direct LDA method [12], the null space of the between-class scatter matrix is first removed and then the projection vectors are obtained by minimizing the within-class scatter distance in the range space of the between-class scatter matrix. Li et al. [13] proposed an efficient and stable algorithm to extract the discriminant vectors by defining the maximum margin criterion (MMC). The main difference between Fisher's criterion and MMC is that the former is to maximize the Fisher quotient while the latter is to maximize the average distance.

The second category mainly depends on exploiting the null space of within-class scatter matrix and the range space of the between-class scatter matrix. In terms of the null space-based LDA, Chen et al. [14] proposed to maximize the between-class scatter in the null space of the within-class scatter matrix and their method is referred to as the NLDA method. In order to reduce the computational cost of calculating the null space of the within-class scatter matrix, several effective methods have been proposed. Instead of directly obtaining the null space of the within-class scatter matrix, Çevikalp et al. [15] first obtained the range space of the within-class scatter matrix and then defined the scatter matrix of common vectors. Based on this, the projection vectors were obtained from the scatter matrix they defined. They also adopted difference subspaces and the Gram-Schmidt orthogonalization procedure to obtain discriminative common vectors. Chu and Thye [16] adopted the QR factorization on several matrices to exploit a new algorithm for the null space-based LDA method. Sharma and Paliwal [17] proposed an alternative null LDA method and discussed its fast implementation. Paliwal and Sharma [18] also developed a variant of pseudoinverse linear discriminant analysis and this method yields better classification performance.

The third category consists of those methods that make use of the null space of within-class scatter matrix, the range space of between-class scatter matrix, and the range space of within-class scatter matrix. Sharma et al. [19] applied improved RLDA to devise a feature selection method to extract important genes. In order to address the problem of the regularization parameter in RLDA, Sharma and Paliwal [20] applied a deterministic method to estimate the parameter by maximizing modified Fisher's criterion.

The fourth category is made up of those methods that explore all the spaces of the within-class scatter matrix and the between-class scatter matrix. Sharma and Paliwal [11] applied a two-stage technique to regularize both the between-class scatter and within-class scatter matrices to achieve the discriminant information.

In addition, there are other variants of LDA that do not belong to four categories mentioned above. Uncorrelated local Fisher discriminant analysis in terms of manifold learning is devised for ear recognition [21]. An exponential locality preserving projection (ELPP) is presented by introducing the matrix exponential to address the SSS problem. A double shrinking model [22] is constructed for manifold learning and feature selection. Li et al. [23] analyzed linear discriminant analysis in the worst case and reduced this problem to a scalable semidefinite feasibility problem. Zollanvari and Dougherty [24] discussed asymptotic generalization bound of linear discriminant analysis. Lu and Renals [25] used probabilistic linear discriminant analysis to model acoustic data.

In this paper, we revisit the optimization criterion for linear discriminant analysis. We find that there exists the degenerated case for some generalized eigenvalues. In order to deal with the degeneration of eigenvalues, we develop a robust implementation for this criterion in this paper. To be specific, the null space of the total scatter matrix is first removed to remedy the singularity problem. Then the eigen-subspace corresponding to each specific eigenvalue is achieved. Finally, in each eigen-subspace, the discriminability of eigenvectors is measured by the maximum margin criterion and the projection vectors can be achieved by optimizing this criterion. We also conduct extensive experiments to evaluate the proposed method on various well-known data sets such as face images and microarray data sets. Experimental results show that our method is more stable than other methods in most cases.

2. Related Works

Assume that there are a set of n-dimensional data points, denoted by $\{x_1, \ldots, x_N\}$, where $x_i \in R^n$ $(i = 1, \ldots, N)$. When the labels of data points are available, each data point belongs to exactly one of c object classes $\{l_1, \ldots, l_c\}$ and the number of samples in class l_i is n_i. Thus, $N = \sum_{i=1}^{c} n_i$ is the number of all data points. In classical linear discriminant analysis, the between-class scatter matrix, the within-class scatter matrix, and the total scatter matrix are defined as follows:

$$\begin{aligned} S_b &= \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T = H_b H_b^T, \\ S_w &= \sum_{i=1}^{c} \sum_{x \in l_i} (x - m_i)(x - m_i)^T = H_w H_w^T, \\ S_t &= \sum_{i=1}^{N} (x_i - m)(x_i - m)^T = H_t H_t^T, \end{aligned} \tag{1}$$

where m_i is the centroid of the ith class and m is the global centroid of the data set. The precursor matrices are defined as

$$\begin{aligned} H_b &= \left[\sqrt{n_1}(m_1 - m), \ldots, \sqrt{n_c}(m_c - m)\right], \\ H_w &= \left[X_1 - m_1 e_1^T, \ldots, X_c - m_c e_c^T\right], \\ H_t &= \left[x_1 - m, \ldots, x_N - m\right], \end{aligned} \tag{2}$$

where $e_i = (1, \ldots, 1)^T \in \mathfrak{R}^{n_i}$ and X_i is the data matrix that consists of data points from class l_i.

Classical LDA is to find the projection direction by making data points from different classes far from each other and data points from the same class close to each other. To be specific, LDA is to obtain the optimal projection vector by optimizing the following objective function:

$$\max \quad J_1(w) = \max\left(\frac{w^T S_b w}{w^T S_w w}\right). \tag{3}$$

The optimal projection direction w can be achieved by solving the generalized eigenproblem: $S_b w = \lambda S_w w$. In general, there are at most $c - 1$ eigenvectors corresponding to nonzero generalized eigenvalues since the rank of the matrix S_b is not bigger than $c - 1$. When S_w is singular, some methods including PCA plus LDA [1], LDA/GSVD [7], and LDA/QR [26] can be used to deal with this problem.

3. Optimization Criterion and Its Robust Implementation

In this section, we revisit an optimization criterion for linear discriminant analysis and its properties are analyzed in detail. Finally, we discuss its robust implementation.

Note that if the matrix S_w is singular, the optimal function value of (3) will take the positive infinity. There are several variants of the model in (3) that can be found [27]. In fact, when the matrix S_w is nonsingular, it is not difficult to verify that these variants of (3) are equivalent [27]. For convenience, we adopt the following optimization criterion to give a stable and efficient algorithm for linear discriminant analysis, denoted by

$$\min \quad J_2(w) = \min\left(\frac{w^T S_w w}{w^T S_t w}\right). \tag{4}$$

The main aim for adopting (4) is based on the following reasons. First, the objective function is a bounded function in the general case, which avoids the case that the objective function takes the infinity. Second, since the null space of S_w plays an important role in some cases, especially in the small-sample-size problem, the optimization criterion of (4) also provides convenience for analyzing the null space of S_w. In fact, it is straightforward to verify that (4) and (3) are equivalent under some conditions. Most importantly, (4) can produce more generalized eigenvalues than (3) since the rank of S_t is not smaller than the rank of S_w. In addition, from the viewpoint of optimization, the objective function we optimize is usually bounded. Thus, (4) is more preferred than (3) in some cases.

It is obvious that the optimal projection w of (4) can be achieved by solving the generalized eigenproblem: $S_w w = \lambda S_t w$ when the matrix S_t is nonsingular. Later we will note that the generalized eigenvalue λ will take values in the interval of 0 and 1. Different from classical LDA, we extract the discriminant vectors which are composed of the first q eigenvectors of $S_t^{-1}S_w$ corresponding to the first q smallest eigenvalues if S_t is nonsingular. In such a case, we can avoid the singularity problem of the matrix S_w. Before giving an explicit implementation of the optimization criterion of (4), we start by giving the definitions of some subspaces [28].

Definition 1. Let A be an $n \times n$ positive semidefinite matrix and λ be an eigenvalue of A. The set of all eigenvectors of A corresponding to the eigenvalue λ, together with the zero vector, forms a subspace. This subspace is referred to as the eigen-subspace of A with λ.

Definition 2. The null space of the matrix A is the set of all eigenvectors of A with $\lambda = 0$.

Definition 3. The range space of the matrix is the set of all eigenvectors of A corresponding to nonzero eigenvalues.

In the case of the positive semidefinite matrix, the number of repeated roots of the characteristic equation det $|\lambda I - A| = 0$ determines the dimension of the eigen-subspace of A with λ. If the dimension of the eigen-subspace of A with λ is bigger than 1, the eigenvalue λ is degenerative since the number of repeated roots of the characteristic equation is bigger than 1. It is observed from (1) that the matrices S_b, S_w, and S_t are positive semidefinite. According to the above definitions, we can obtain the following four subspaces from S_b and S_w [20]:

(a) The null space of S_b is denoted by null (S_b).

(b) The null space of S_w is denoted by null (S_w).

(c) The range space of S_b is denoted by span (S_b).

(d) The range space of S_w is denoted by span (S_w).

Based on these four subspaces, we can construct another four subspaces.

(e) Subspace A is defined as the intersection of span (S_b) and null (S_w).

(f) Subspace B is defined as the intersection of span (S_b) and span (S_w).

(g) Subspace C is defined as the intersection of null (S_b) and span (S_w).

(h) Subspace D is defined as the intersection of null (S_b) and null (S_w).

From Subspaces A, B, C, and D, we find that the objective function $J_2(w)$ of (4) satisfies the following equation:

$$J_2(w) = \begin{cases} 0 & w \in \text{Subspace } A \\ (0,1) & w \in \text{Subspace } B \\ 1 & w \in \text{Subspace } C \\ \dfrac{0}{0} & w \in \text{Subspace } D. \end{cases} \tag{5}$$

From (5), one can see that if w is taken from Subspace A, Subspace B, or Subspace C, the objective function $J_2(w)$ is bounded. If w belongs to Subspace D, the objective function $J_2(w)$ takes the indefinite value. It is of interest to note that the null space of S_t is the intersection of the null space of S_b and the null space of S_w. It has been proved that the null space of S_t does not contain any discriminant information [29]. Thus, Subspace D does not contain any discriminant information and this also shows that part of the null space of S_w does not contain discriminant information. Therefore, Subspace D can be removed without losing any information and this can be done by removing the null space of S_t. An effective method to remove the null space of S_t is to perform the singular value decomposition (SVD) [28] on H_t, denoted by $H_t = U_t \Sigma_t V_t^T$, where U_t consists of the left singular vectors corresponding to the nonzero singular values of H_t. In such a case, we do not lose any information of data. By doing so, we also remove part of the null space of S_w that does not contain discriminant information. Since we focus on (4), the range space of S_t must be considered. If the null space of S_t is removed, it is necessary to consider three subspaces in the case of (4): the null space of S_w, the range space of S_w, and the range space of S_t. For these three subspaces, we also give their relations with Subspace A, Subspace B, and Subspace C. It is not difficult to verify that the intersection of the null space of S_w and the range space of S_t is equivalent to Subspace A, and the intersection of the range space of S_w and the range space of S_t contains Subspaces B and C. This shows that we do not lose any discriminant information from Subspace A, Subspace B, and Subspace C if we solve (4). In such a case, we first remove the null space of S_t. That is, we consider the following optimization function in the case of the range space of S_t,

$$\min \quad J_3(a) = \min\left(\frac{a^T\overline{S}_w a}{a^T\overline{S}_t a}\right), \tag{6}$$

where $\overline{S}_t = U_t^T S_t U_t$, $\overline{S}_w = U_t^T S_w U_t$.

It is evident that $\overline{S}_t$ in (6) is nonsingular when the null space of S_t is removed. In such a case, we obtain the projection vectors which are composed of t eigenvectors of $\overline{S}_t^{-1}\overline{S}_w$ corresponding to t eigenvalues. From (6), we can see that $J_3(a)$ takes values in the interval of 0 and 1. In fact, the value $J_3(a)$ gives an indicator of choosing the effective subspace. According to the definition of the optimization criterion, we have the following conclusions: the subspace corresponding to $J_3(a) = 0$ is the most important; the subspace corresponding to $J_3(a) \in (0,1)$ is the second important; the subspace corresponding to $J_3(a) = 1$ is the least important.

By solving the generalized eigenproblem, $\overline{S}_w a = \lambda\overline{S}_t a$, we can obtain $t(= \operatorname{rank}(H_t))$ eigenvalues, which produces t eigenvectors. In some cases some of these t eigenvalues may be equal. In other words, some eigenvalues degenerate into the same eigenvalue, which may affect the performance of some algorithms. Assume that these t eigenvalues consist of d ($d \leq t = \operatorname{rank}(S_t)$) different values λ_i ($i = 1,\ldots, d$) in an increasing order and have multiplicities s_i ($i = 1,\ldots,d$), where s_i denotes the algebraic multiplicities of the eigenvalue λ_i and $\sum_{i=1}^{d} s_i = t$. In some situations, it is useful to work with the set of all eigenvectors associated with a specific value λ_i. Let us define the following set:

$$S(\lambda_i) = \left\{a : a \in R^t \text{ and } \overline{S}_w a = \lambda_i\overline{S}_t a\right\}. \tag{7}$$

The dimension of $S(\lambda_i)$ is in general equal to the algebraic multiplicities of λ_i since $\overline{S}_w$ and $\overline{S}_t$ are symmetric real matrices. The set $S(\lambda_i)$ forms the eigen-subspace of the matrix pairs $(\overline{S}_w, \overline{S}_t)$ corresponding to the generalized eigenvalue λ_i. When the dimension of $S(\lambda_i)$ is equal to 1, it is not necessary to deal with this subspace since it only contains an eigenvector. However, when the dimension of $S(\lambda_i)$ is bigger than 1, it is impossible to determine which eigenvector in this eigen-subspace is the most important since all the eigenvectors correspond to the same eigenvalue. The case often occurs in the small-sample-size problem where the dimension of the eigen-subspace of $S(\lambda_i = 0)$ is relatively high. In such a case, it is infeasible to determine which projection vector in the eigen-subspace of $S(\lambda_i = 0)$ is the most important if we only use (7). For some nonzero generalized eigenvalues from the matrix pair $(\overline{S}_w, \overline{S}_t)$, the dimension of $S(\lambda_i \neq 0)$ may be bigger than 1. For example, $S(\lambda_i = 1)$ shows that the eigenvector is taken from the null space of $\overline{S}_b = \overline{S}_t - \overline{S}_w$. Generally speaking, the dimension of the null space $\overline{S}_b$ is bigger than 1 and this makes the dimension of $S(\lambda_i = 1)$ be bigger than 1. So it is necessary to use an additional strategy to determine the importance of eigenvectors if the dimension of $S(\lambda_i)$ is bigger than 1. For the subspace $S(\lambda_i)$, we can obtain a matrix whose columns consist of the eigenvectors of the generalized eigenvalue λ_i, denoted by P_{λ_i}. Obviously the dimension of $S(\lambda_i)$ is equal to the number of the columns of P_{λ_i}. If this matrix is provided, it is straightforward to obtain an orthogonal basis by performing the QR decomposition on P_{λ_i} and the orthogonal basis can be expressed in the matrix form: Q_{λ_i}. Note that the space spanned by the column vectors of P_{λ_i} is equivalent to the space spanned by the column vectors of Q_{λ_i}. Thus, in the space spanned by the column vectors of Q_{λ_i}, we formulate the following objective function based on the maximum margin criterion.

$$\begin{gathered}\max \quad J_4(g) = \max\left(g^T\widehat{S}_b g - g^T\widehat{S}_w g\right), \\ g^T g = 1,\end{gathered} \tag{8}$$

where $\widehat{S}_b = Q_{\lambda_i}^T U_t^T S_b U_t Q_{\lambda_i}$, $\widehat{S}_w = Q_{\lambda_i}^T U_t^T S_w U_t Q_{\lambda_i}$.

When the dimension of the set $S(\lambda_i)$ is 1, it is easy to prove that $g = \pm 1$. When the dimension of the set $S(\lambda_i)$ is bigger than 1, it is necessary to obtain s_i eigenvectors of $\widehat{S}_b - \widehat{S}_w$ corresponding to s_i eigenvalues in a decreasing order. These s_i eigenvectors form the matrix $G_{\lambda_i}(= [g_1,\ldots,g_{s_i}])$. Thus, the discriminability of eigenvectors in the eigen-subspace of $S(\lambda_i)$ can be measured by the eigenvalues of $\widehat{S}_b - \widehat{S}_w$. This gives suggestions on how to choose effective discriminant vectors in the eigen-subspace $S(\lambda_i)$, which solves the degenerated case of eigenvalues. In classical LDA, the discriminability of eigenvectors in the eigen-subspace is sometimes neglected.

Note that, in the small-sample-size problem, the dimension of the eigen-subspace of $S(\lambda_i = 0)$ is relatively high. In such a case, we need to obtain this eigen-subspace. In fact, it is noted that the eigen-subspace $S(\lambda_i = 0)$ is the null space of $\overline{S}_w$ and obtaining the null space $\overline{S}_w$ may be time consuming when the dimension of the null space of $\overline{S}_w$ is high. Fortunately, several effective methods have been proposed to obtain the null space of $\overline{S}_w$. Çevikalp et al. [15] have proposed an effective algorithm to avoid computing the null space of $\overline{S}_w$ by finding the range space of $\overline{S}_w$. Note that the dimension of the range space of $\overline{S}_w$ is equal to the rank of the matrix $\overline{S}_w$. Based on the range space of $\overline{S}_w$, we can obtain common vectors for each class and construct the scatter matrix of the common vectors as done in [15]. Finally, the projection vectors can be obtained by performing eigen-decomposition on the scatter matrix of the common vectors.

As a summary of the above discussion, we list the detailed steps for solving linear discriminant analysis in Algorithm 4.

Algorithm 4. It is a stable and efficient algorithm for solving linear discriminant analysis.

Step 1. Construct H_w, H_b, and H_t, and compute the left singular matrix U_t of H_t by performing the SVD on $H_t = U_t \Sigma_t V_t^T$, where U_t consists of singular vectors corresponding to the nonzero singular values of H_t; obtain $\overline{H}_w = (U_t \Sigma_t^{-1})^T H_w$.

Step 2. Obtain the range space of $\overline{H}_w$, denoted by $\overline{U}_w$ whose column vectors are orthogonal; perform the SVD on $(\overline{U}_w)^T \overline{H}_w = U_w \Sigma_w V_w^T$ and assign σ_i in an increasing order from the diagonal elements of Σ_w.

Step 3. Let $Y = I_{t \times t} - \overline{U}_w \overline{U}_w^T$. If Y is not a zero matrix, perform Step 4; otherwise, go to Step 5.

Step 4. Based on Y, obtain the common vectors of each class, compute the scatter matrix of common vectors, and perform the eigen-decomposition on the scatter matrix of common vectors to obtain projection vectors, denoted by Q_0.

Step 5. For each nonzero σ_i, do the following.

Step 5(a). Obtain the singular submatrix U_{σ_i} by searching the column vectors of U_w corresponding to the singular value σ_i; let $P_{\sigma_i} = U_t \Sigma_t^{-1} U_{\sigma_i}$; apply the QR decomposition on P_{σ_i} to obtain the matrix Q_{σ_i} whose column vectors are orthogonal.

Step 5(b). Let $\widehat{S}_b = ((Q_{\sigma_i})^T H_b)(H_b^T Q_{\sigma_i})$ and $\widehat{S}_w = ((Q_{\sigma_i})^T H_w)(H_w^T Q_{\sigma_i})$; compute all discriminant vectors which are the eigenvectors of $\widehat{S}_b - \widehat{S}_w$; sort the eigenvectors according to the eigenvalues of $\widehat{S}_b - \widehat{S}_w$ in a decreasing order and form the matrix G_{σ_i}.

Step 6. Obtain the transformation matrix $T = [U_t Q_0, Q_{\sigma_1} G_{\sigma_1}, \ldots, Q_{\sigma_d} G_{\sigma_d}]$.

Note that, in Step 2 of Algorithm 4, we only need to obtain the range space of $\overline{H}_w$, that is, an orthonormal basis of $\overline{H}_w$. There are some effective methods for obtaining the range space of $\overline{H}_w$. For example, the range space of $\overline{H}_w$ can be achieved by finding the left singular matrix U_t of $\overline{H}_w$ corresponding to nonzero singular values. It is pointed out in [28] that computing left singular vectors of $\overline{H}_w$ corresponding to nonzero singular values is more efficient than finding left singular vectors of $\overline{H}_w$ corresponding to all singular values including zeros. In addition, one may resort to difference subspaces and the Gram-Schmidt orthogonalization procedure [15] to obtain the range space of $\overline{H}_w$. Note that, in Step 3 of Algorithm 4, we use a criterion to judge whether the null space of $\overline{H}_w \overline{H}_w^T$ exists. If $Y = I_{t \times t} - \overline{U}_w \overline{U}_w^T$ is not a zero matrix, this shows that there exists the null space of $\overline{H}_w \overline{H}_w^T$. In such a case, one may use the method (Step 4 of Algorithm 4) proposed in [15] to further deal with the null space of $\overline{H}_w \overline{H}_w^T$. It is observed from Algorithm 4 that we need to perform Step 5 of Algorithm 4 regardless of the existence of the null space of $\overline{H}_w \overline{H}_w^T$. In such a case, we can see that t $(= s_1 + \cdots + s_d)$ eigenvectors can be ordered in terms of their importance. By performing Algorithm 4, we can evaluate the projection vectors from Subspace C which is often neglected in the previous literature. It is obvious that the above method can provide t discriminant vectors because the rank of H_t is t which is much bigger than $c - 1$. As a result, this method may be helpful when the number of classes is relatively small. Note that we use the eigenvalue λ_i in (7) and it is not difficult to verify that $\lambda_i = (\sigma_i)^2$. If the singular value σ_i occurs only once in the diagonal elements of Σ_w, we do not need to perform Step 5(b) in real applications.

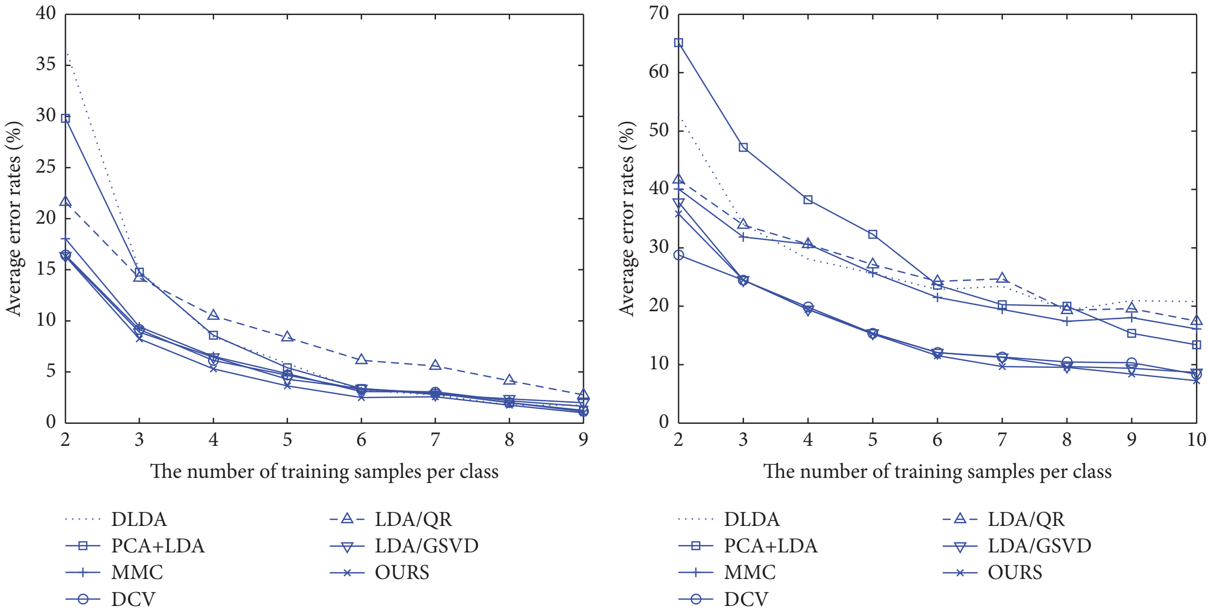
Table 1: Statistics of the data sets we use.

Datasets	Number of samples	Number of dimensions	Number of classes
ORL	400	112 ∗ 92	40
Yale	165	112 ∗ 92	15
ALLAML	72	7129	2
Duke-Breast	42	7129	2
Colon	62	2000	2
Prostate	102	5966	2
Leukemia	72	7129	2
MLL	72	5848	3

4. Experiments Results

In our experiments, we use the ORL face database, the Yale face database, and microarray data sets to evaluate the performance of Algorithm 4. The ORL face database consists of 40 distinct persons, with each containing 10 different images with variations in poses, illumination, facial expressions, and facial details. The original face images are resized to 112 × 92 pixels with 256-level gray scales. The Yale face database contains 165 gray-scale images of 15 individuals. The images demonstrate variations in lighting condition and facial expressions. All of these face images are aligned based on eye coordinates and are resized to 112 × 92 pixels. Six microarray data sets including ALLMLL [30], Duke-Breast [31], Colon [32], Prostate [32], Leukemia [32], and MLL [32] are used to test the proposed method. Table 1 lists the statistics of the data sets we use. It is observed that the dimensions of features of samples on these data sets are much higher than the number of samples. The experiments are performed on a PC with the operating system of Windows 8.1, an i3 CPU (3.30 GHz) and an 8 G memory. The programming environment is MATLAB 2014a.

4.1. Face Recognition. In this set of experiments, the number of each individual in the training set varies from 2 to $\min\{n_1, \ldots, n_c\} - 1$, and the remaining images in the data set are used to form the testing set. To reduce the variations of the accuracies from randomness, the classification performance we report in the experiments is achieved over twenty runs. That is, there are twenty different training and testing sets used for evaluating the classification performance. We compare the proposed method with some



(a) ORL

(b) Yale

Figure 1: The error rates of each algorithm on two face databases.

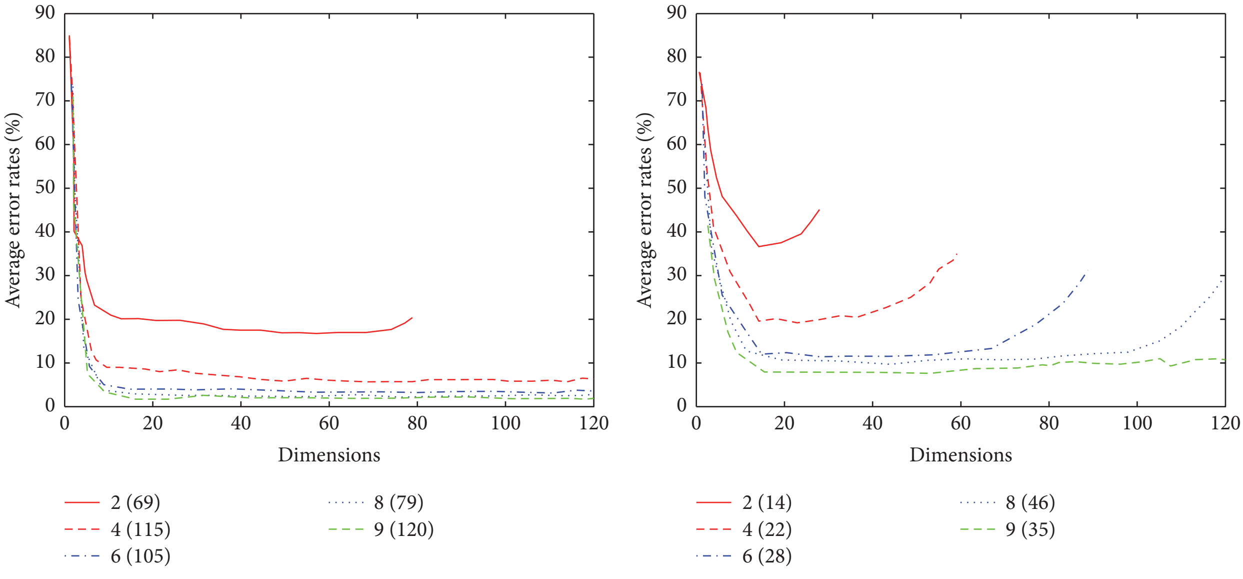
Table 2: Performance comparisons (%) of some methods on face databases.

Databases	TS	DLDA	PCA+LDA	MMC	DCV	LDA/QR	LDA/GSVD	Ours
ORL	2	36.73 (4.05)	29.82 (3.08)	18.04 (2.35)	16.43 (1.95)	21.62 (3.08)	19.92 (2.08)	**16.32** (1.91)
	4	8.45 (1.45)	8.60 (1.83)	6.52 (1.60)	6.12 (1.95)	8.45 (1.45)	10.45 (1.72)	**5.31** (1.70)
	6	3.21 (1.31)	3.37 (1.62)	2.96 (1.31)	3.26 (1.14)	3.21 (1.31)	7.09 (1.34)	**2.50** (1.11)
	8	**1.75** (2.99)	2.00 (1.68)	2.18 (2.29)	2.00 (2.20)	**1.75** (2.29)	4.12 (1.76)	**1.75** (1.33)
Yale	2	52.88 (4.68)	65.14 (8.70)	40.07 (4.35)	**28.77** (15.25)	41.66 (5.23)	37.81 (4.45)	35.81 (3.85)
	4	28.09 (4.51)	38.23 (3.68)	30.61 (4.80)	19.85 (4.18)	30.61 (4.80)	**19.47** (4.02)	**19.47** (3.06)
	6	22.86 (4.07)	23.60 (5.24)	21.53 (4.56)	12.06 (3.76)	24.26 (6.08)	12.06 (3.87)	**11.53** (3.27)
	8	19.33 (4.45)	20.00 (5.67)	17.44 (6.29)	10.44 (3.80)	19.33 (4.45)	9.66 (3.61)	**9.55** (3.25)

previous methods including LDA/GSVD [7], LDA/QR [26], DLDA [12], PCA+LDA [1], MMC [13], and the discriminant common vector (DCV) approach [15] which is an effective approach for solving NLDA [14]. Note that these methods are designed to solve the small-sample-size problem when linear discriminant analysis is used. Subspace A or Subspace B are considered in LDA/QR, DLDA, PCA+LDA, and DCV. Although LDA/GSVD makes use of three subspaces (Subspace A, Subspace B, and Subspace C), the importance of projection vectors in some eigen-subspaces is not effectively measured in some cases. In this paper we do not compare other discriminant methods since the main objective of paper is to provide a stable and efficient algorithm for solving the degenerated eigenvalue of LDA. Note that we do not give the running time of algorithms we test since some methods only make use of part of subspaces in (5). Generally speaking, the performance of each algorithm varies with the change of the dimension of features. For comparisons, we try to search for the performance on all the feature dimensions and list the best one.

Figure 1 shows the error rate of each method we test with different training images in each class on the ORL and Yale face databases. For clarity, we also show the mean and standard deviation in the parentheses of the error rates of each method in Table 2. Note that the best performance of each method in each line is highlighted in bold and we show the results of 2, 4, 6, and 8 training images per class.

From Figure 1 and Table 2, one can see that the error rate of each algorithm decreases as the number of the training samples in each class increases in most cases. It is observed from Table 2 that the standard deviation of our method is smaller than that of the other methods in most cases. On the ORL face database, the error rate of our method decreases from 16.32% with 2 training samples per class to 1% with 9 training samples per class, while the error rates of DLDA, PCA+LDA, MMC, DCV, LDA/QR, and LDA/GSVD decrease from 36.73%, 29.82%, 18.04%, 16.43%, 21.62%, and 19.92% with 2 training samples per class to 1.75%, 1.25%, 1.625%, 1.125%, 2.75%, and 3% with 9 training samples per class, respectively. The results show that our method outperforms



(a) ORL

(b) Yale

Figure 2: The error rate of the proposed method with the change of features.

other methods in most cases. On the Yale face database, although the DCV method gives the best result in the case of 2 training samples per class, it obtains the biggest standard deviation. It is also observed that our method is superior to other methods in terms of the classification performance with the increase of training samples.

Since the number of the extracted features of samples by using the proposed method is not limited by the number of classes and only limited by the rank of S_t, we can project the samples onto the space whose dimension is greater than the number of classes. Figure 2 shows a plot of the error rate versus dimensionality. The numbers in the parentheses denote the optimal dimension corresponding to the best classification performance. As can be seen from Figure 2, the error rate of the proposed method decreases with the increase of training samples per class. It is also found from Figure 2 that the classification performance may be improved when the dimension of the reduced space is bigger than the number of classes. On the Yale face database, it is observed that the error rate of the proposed method first decreases and then rises with the increase of dimensions, which shows that choosing too many features yields the overfitting phenomenon in the classification task. On the ORL face database, the error rate of the proposed method first decreases drastically and then becomes flat when the number of training samples is bigger than 2. It is found that the best performance of our method is achieved when the number of extracted features is much bigger than the number of classes. In short, these experimental results show that Subspace C which is often neglected in classical LDA in (5) may play a role in face recognition in some cases.

Now let us explain the reason why our method can achieve the good classification performance. The DLDA and LDA/QR methods first remove the null space of S_b. However, removing the null space of S_b will also lose part of the null space of S_w and may result in the loss of important information in the null space of S_w. The PCA+LDA method does not consider the null space of S_w. It has been proved that the null space of S_w will play an important role in the SSS problem [14]. The DCV method does not make use of subspace B in (5) and this subspace may be helpful in obtaining discriminant vectors in the SSS problem. Although the LDA/GSVD method considers three subspaces, the discriminability of each eigen-subspace is not analyzed. In the MMC method, the discriminant vectors in Subspace A and Subspace B in (5) may have the same objective function. This results in the difficulty in determining which discriminant vector is the most important. In fact, Subspace C in (5) is often neglected in LDA-based methods in the previous literature. We give a strategy to measure the importance of each discriminant vector in all subspaces including Subspace C for the first time. As can be seen from Figure 2, Subspace C also plays a role in face recognition. As a result, the proposed method can achieve better classification performance than other methods in the general case.

In the following experiments, we study the effect of image sizes on the classification performance in terms of two face databases. Since the number of face images on these two face databases is relatively small, the leave-one-out method is performed where it takes one image for testing and the remaining images for training. By reducing the image resolution of 112 $*$ 92 pixels, we can obtain 56 $*$ 46 pixels where each pixel value is the average value of a 2 $*$ 2 subimage of the original images. Similarly, we can achieve the images with 28 $*$ 23 pixels. In such cases, there exists the null space of the within-class scatter matrix. Table 3 shows the experimental results of each method in three resolutions on two face databases.

As can be seen from Table 3, the error rate of each method does not always increase with the reduction of image

Table 3: Comparisons of misclassification rates (%) of several methods on face databases.

Databases	Training size	DLDA	PCA+LDA	MMC	DCV	LDA/QR	LDA/GSVD	Ours
ORL	112 ∗ 92	2.25 (1.68)	2.50 (2.68)	2.50 (1.66)	**1.50** (1.74)	3.25 (2.68)	2.50 (2.04)	2.25 (1.84)
	56 ∗ 46	2.00 (2.58)	2.00 (2.83)	2.50 (2.83)	2.00 (2.58)	3.00 (2.35)	2.75 (3.40)	**1.25** (1.31)
	28 ∗ 23	**2.00** (2.29)	4.75 (2.94)	**2.00** (2.58)	5.25 (4.31)	3.75 (2.58)	6.25 (4.75)	**2.00** (2.37)
Yale	112 ∗ 92	21.00 (10.24)	16.36 (8.62)	11.51 (6.72)	12.12 (8.33)	20.60 (9.16)	12.12 (9.34)	**9.09** (6.22)
	56 ∗ 46	20.00 (6.66)	15.15 (7.93)	27.27 (10.21)	10.90 (8.85)	21.21 (8.85)	10.90 (5.39)	**7.87** (5.19)
	28 ∗ 23	19.39 (10.93)	12.72 (8.66)	14.54 (9.34)	9.09 (9.07)	20.00 (11.15)	9.09 (9.07)	**8.48** (9.02)

Table 4: Error rates (%) of each method on microarray data sets.

Datasets	DLDA	PCA+LDA	MMC	DCV	LDA/QR	LDA/GSVD	Ours
ALLMLL	4.27 (3.50)	5.00 (4.27)	4.11 (4.32)	3.99 (4.01)	4.32 (5.05)	4.01 (4.72)	**3.78 (3.23)**
Duke-Breast	14.32 (9.15)	12.67 (8.96)	11.98 (7.69)	12.05 (8.99)	13.03 (9.45)	11.88 (8.22)	**11.56 (8.03)**
Colon	31.11 (18.28)	22.78 (19.24)	23.33 (19.56)	23.33 (19.56)	31.11 (18.92)	23.33 (19.56)	**20.00 (17.21)**
Prostate	6.88 (6.32)	6.94 (7.51)	6.74 (8.02)	6.63 (7.03)	7.88 (6.23)	6.55 (6.21)	**6.32 (5.99)**
Leukemia	12.85 (17.10)	16.28 (17.10)	2.85 (6.02)	2.85 (6.02)	12.95 (6.02)	2.85 (6.20)	**1.42 (4.51)**
MLL	9.88 (5.52)	13.93 (6.02)	9.52 (5.09)	9.34 (5.85)	10.93 (5.42)	9.62 (5.45)	**9.01 (5.01)**

resolutions. On the ORL face database, the DCV method obtains the best classification result on the resolution of 112 ∗ 92 pixels. With the reduction of image resolutions, the performance of NLDA becomes worse since the dimension of the null space of S_w becomes smaller. On the ORL face database, the proposed method is better than LDA/GSVD and has a smaller standard deviation than other methods in most cases. The main reason is that we consider the degenerated case of the eigenvalue. It is noted that our method achieves the best classification result when the resolution of images is 56∗46 pixels. On the Yale face database, the proposed method outperforms other methods in terms of the classification performance. It is also observed that the best recognition rate among all methods is 92.13% and is achieved by the proposed method when the images are 56 ∗ 46 pixels on the Yale face database. From these experiments, we can also notice that it is not necessary to use the large-size images to obtain good classification performance in the classification task.

4.2. Applications to Microarray Data Sets. In this set of experiments, we further validate the proposed method on microarray data sets. In order to evaluate the classification performance of various LDA methods, we adopt the tenfold cross validation on these data sets. In other words, we divide each data set into ten subsets of approximately equal sizes. Then we perform training and testing ten times, each time leaving out one of the subsets for training and the discarded subset for testing. The classification performance is averaged over ten runs. Table 4 shows the mean and the standard deviation of the error rate of each method.

As can be seen from Table 4, the classification performance of the proposed method is consistently superior to that of other methods on all the data sets we tested. It is found that our method is more stable than other methods since the standard deviation of our method is smaller than that of other methods on all of data sets we tested. It is noted that PCA+LDA performs poorly on Leukemia and MLL data sets. This may come from the fact that the null space of the within-class scatter matrix is removed and it plays an important role in obtaining discriminant feature vectors. It is also found that DLDA does not give satisfactory results on Duke-Breast and Colon data sets since DLDA may remove the part of the null space of the within-class scatter matrix. One can see from Table 4 that the NLDA method achieves good classification accuracies on these data sets since these data sets are the small-sample-size sets. One can also observe that the LDA/QR method does not perform well on some data sets. This may be explained by the fact that the LDA/QR method may remove part of the range space of S_w and part of the null space of S_b. It is found that LDA/GSVD is not better than our method although LDA/GSVD considers three subspaces. This is possibly because in LDA/GSVD the discriminability of each eigen-subspace is not given. Because the discriminant vectors in Subspace A and Subspace B in the MMC method may correspond to the same objective function, this may lead to the degradation in MMC. Overall, the proposed method is very stable on these data sets due to the fact that we consider the degenerated eigenvalues of scatter matrices, especially for Subspace C which is neglected in previous literature.

5. Conclusions

In this paper, we revisit linear discriminant analysis based on an optimization criterion. Different from the existing LDA-based algorithms, the new algorithm adopts the spirit of the maximum margin criterion (MMC) and applies MMC to the eigen-subspace when the eigenvalue is degenerative. The new implementation avoids the singularity problem in the SSS problem and provides more than $c - 1$ discriminant vectors. We also conduct a series of comparative studies on face images and microarray data sets to evaluate the proposed method. Our experiments on face images and microarray data sets demonstrate that the classification performance achieved by our method is better than that of other LDA-based algorithms in most cases and the proposed method is an effective and stable linear discriminant method for dealing with high-dimensional data.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was partially supported by the Fundamental Research Funds for the Central Universities (2015XKMS084).

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[2] X. Jin, M. Zhao, T. W. S. Chow, and M. Pecht, "Motor bearing fault diagnosis using trace ratio linear discriminant analysis," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 5, pp. 2441–2451, 2014.

[3] M. Kolar and H. Liu, "Optimal feature selection in high-dimensional discriminant analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1063–1083, 2015.

[4] J.-H. Xue and P. Hall, "Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1109–1112, 2015.

[5] C. Hou, C. Zhang, Y. Wu, and Y. Jiao, "Stable local dimensionality reduction approaches," *Pattern Recognition*, vol. 42, no. 9, pp. 2054–2066, 2009.

[6] Y. Zhang and D.-Y. Yeung, "Semisupervised generalized discriminant analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1207–1217, 2011.

[7] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 995–1006, 2004.

[8] W. Bian and D. Tao, "Asymptotic generalization bound of fisher's linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2325–2337, 2014.

[9] D. Chu, L.-Z. Liao, M. K. Ng, and X. Wang, "Incremental linear discriminant analysis: a fast algorithm and comparisons," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2716–2735, 2015.

[10] J. Zhao, L. Shi, and J. Zhu, "Two-stage regularized linear discriminant analysis for 2-D data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 8, pp. 1669–1681, 2015.

[11] A. Sharma and K. K. Paliwal, "Linear discriminant analysis for the small sample size problem: an overview," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 3, pp. 443–454, 2015.

[12] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[13] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, 2006.

[14] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "New LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[15] H. Çevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, 2005.

[16] D. Chu and G. S. Thye, "A new and fast implementation for null space based linear discriminant analysis," *Pattern Recognition*, vol. 43, no. 4, pp. 1373–1379, 2010.

[17] A. Sharma and K. K. Paliwal, "A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices," *Pattern Recognition*, vol. 45, no. 6, pp. 2205–2213, 2012.

[18] K. K. Paliwal and A. Sharma, "Improved pseudoinverse linear discriminant analysis method for dimensionality reduction," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 1, Article ID 1250002, 9 pages, 2012.

[19] A. Sharma, K. K. Paliwal, S. Imoto, and S. Miyano, "A feature selection method using improved regularized linear discriminant analysis," *Machine Vision and Applications*, vol. 25, no. 3, pp. 775–786, 2014.

[20] A. Sharma and K. K. Paliwal, "A deterministic approach to regularized linear discriminant analysis," *Neurocomputing*, vol. 151, no. 1, pp. 207–214, 2015.

[21] H. Huang, J. Liu, H. Feng, and T. He, "Ear recognition based on uncorrelated local Fisher discriminant analysis," *Neurocomputing*, vol. 74, no. 17, pp. 3103–3113, 2011.

[22] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 244–257, 2013.

[23] H. Li, C. Shen, A. van den Hengel, and Q. Shi, "Worst case linear discriminant analysis as scalable semidefinite feasibility problems," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2382–2392, 2015.

[24] A. Zollanvari and E. R. Dougherty, "Generalized consistent error estimator of linear discriminant analysis," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2804–2814, 2015.

[25] L. Lu and S. Renals, "Probabilistic linear discriminant analysis for acoustic modeling," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 702–706, 2014.

[26] J. Ye and Q. Li, "A two-stage linear discriminant analysis via QR-decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 929–941, 2005.

[27] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, NY, USA, 2nd edition, 1990.

[28] G. H. Folub and C. F. Van Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore, Md, USA, 1996.

[29] R. Huang, Q. Liu, H. Lu et al., "Solving the small sample size problem of LDA," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, IEEE Computer Society, August 2002.

[30] S. A. Armstorng, J. E. Staunton, L. B. Silverman et al., "MLL translocation specify a district gene expression profile at distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[31] M. West, C. Blanchette, H. Dressman et al., “Predicting the clinical status of human breast cancer by using gene expression profiles,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 13462–11467, 2011.

[32] C. Blake, E. Keogh, and C. J. Merz, “UCI Repository of machine learning databases,” University of California, Irvine, Calif, USA, 2014, http://www.ics.uci.edu/~mlearn/MLRepository.html.