Geometric Mean for Subspace Selection

Dacheng Tao, Member, IEEE, Xuelong Li, Senior Member, IEEE, Xindong Wu, Senior Member, IEEE, and Stephen J. Maybank, Senior Member, IEEE

Abstract—Subspace selection approaches are powerful tools in pattern classification and data visualization. One of the most important subspace approaches is the linear dimensionality reduction step in the Fisher's linear discriminant analysis (FLDA), which has been successfully employed in many fields such as biometrics, bioinformatics, and multimedia information management. However, the linear dimensionality reduction step in FLDA has a critical drawback: for a classification task with \( c \) classes, if the dimension of the projected subspace is strictly lower than \( c - 1 \), the projection to a subspace tends to merge those classes, which are close together in the original feature space. If separate classes are sampled from Gaussian distributions, all with identical covariance matrices, then the linear dimensionality reduction step in FLDA maximizes the mean value of the Kullback-Leibler (KL) divergences between different classes. Based on this viewpoint, the geometric mean for subspace selection is studied in this paper. Three criteria are analyzed: 1) maximization of the geometric mean of the KL divergences, 2) maximization of the geometric mean of the normalized KL divergences, and 3) the combination of 1 and 2. Preliminary experimental results based on synthetic data, UCI Machine Learning Repository, and handwriting digits show that the third criterion is a potential discriminative subspace selection method, which significantly reduces the class separation problem in comparing with the linear dimensionality reduction step in FLDA and its several representative extensions.

Index Terms—Arithmetic mean, Fisher’s linear discriminant analysis (FLDA), geometric mean, Kullback-Leibler (KL) divergence, machine learning, subspace selection (or dimensionality reduction), visualization.

1 Introduction

Fisher's linear discriminant analysis (FLDA) [28], [13], [9], a combination of a linear dimensionality reduction step and a classification step, was first developed by Fisher [14] for binary classification and then extended by Rao [31] to multiclass classification. The advantages of the linear dimensionality reduction step in FLDA have been widely demonstrated in different applications, e.g., biometrics [22], bioinformatics [10], and multimedia information management [6]. In the rest of the paper, to simplify the notation, we refer to the linear dimensionality reduction step in FLDA as FLDA and the classification step in FLDA as the linear discriminant analysis (LDA).

However, for a \( c \)-class classification task, if the dimension of the projected subspace is strictly lower than \( c - 1 \), FLDA merges classes, which are close in the original feature space. This is defined as the class separation problem in this paper. As pointed out by McLachlan in [28], Lotlikar and Kohavi [23], Loog et al. [25], and Lu et al. [27], this merging of classes significantly reduces the recognition accuracy. The example in Fig. 1 shows that FLDA does not always select the optimal subspace for pattern classification. To improve its performance, Lotlikar and Kohavi [23] developed the fractional-step FLDA (FS-FLDA) by introducing a complex weighting function. Loog et al. [25] developed another weighting scheme for FLDA, namely, the approximate pairwise accuracy criterion (aPAC). The advantage of aPAC is that the projection matrix can be obtained by the generalized eigenvalue decomposition. Lu et al. [27] combined the FS-FLDA and the direct FLDA [43] for very high-dimensional problems such as face recognition. Although all existing methods reduce this problem to some extent, there is still some room to obtain a further improvement.

In this paper, aiming to solve the class separation problem, we first generalize FLDA to obtain a general averaged divergence analysis [37]. If different classes are assumed to be sampled from Gaussian densities with different expected values but identical covariances, then FLDA maximizes the arithmetic mean value of the Kullback-Leibler (KL) divergences [5] between the different pairs of densities. Our generalization of the FLDA is based on replacing the arithmetic mean by a general mean function. By choosing different mean functions, a series of subspace selection algorithms is obtained, with FLDA included as a special case.

Under the general averaged divergence analysis, we investigate the effectiveness of the geometric mean-based subspace selection in solving the class separation problem. The geometric mean amplifies the effects of the small divergences and, at the same time, reduces the effects of the large divergences. Next, to further amplify the effects of the small divergences, the maximization of the geometric mean of the normalized divergences is studied. This turns out not to be suitable for subspace selection, because there exist projection matrices, which make all the divergences very small and, at the same time, make all the normalized divergences...
similar in value. We therefore propose a third criterion, which is a combination of the first two, maximization of the geometric mean of all divergences (both the divergences and the normalized divergences) or, briefly, MGMD. Preliminary experiments based on synthetic data, UCI data [29], and handwriting digits show that MGMD achieves much better classification accuracy than FLDA and several representative extensions of FLDA taken from the literature.

The rest of the paper is organized as follows: In Section 2, FLDA is briefly reviewed. In Section 3, we investigate the geometric mean for discriminative subspace selection, propose the MGMD for discriminative subspace selection, and analyze why it helps to reduce the class separation problem. Three examples based on synthetic data are given in Section 4 to demonstrate the advantages of MGMD. A large number of statistical experiments based on synthetic data are given in Section 5 to show the properties of MGMD and demonstrate that MGMD significantly reduces the class separation problem. To show the geometric mean is more reasonable than arithmetic mean in discriminative subspace selection, further empirical studies are given based on the UCI Machine Learning Repository [29] in Section 6. In Section 7, an experimental application to handwriting recognition is given, and Section 8 concludes the paper.

2 Fisher’s Linear Discriminant Analysis

The aim of FLDA [28], [13], [9] is to find in the feature space a low-dimensional subspace, in which the different classes of measurements are well separated for certain cases. The subspace is spanned by a set of vectors, $w_i$, $1 \leq i \leq m$, which form the columns of a matrix $W = [w_1, \ldots, w_m]$. It is assumed that a set of training measurements is available. The training set is divided into $c$ classes. The $i$th class contains $n_i$ measurements $x_{ij}(1 \leq j \leq n_i)$ and has an expected mean of $\mu_i = (1/n_i) \sum_{j=1}^{n_i} x_{ij}$. The between-class scatter matrix $S_b$ and the within-class scatter matrix $S_w$ are respectively defined by

\[ S_b = \frac{1}{n} \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^T, \]
\[ S_w = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T, \]

where $n = \sum_{i=1}^{c} n_i$ is the size of the training set, and $\mu = (1/n) \sum_{i=1}^{c} \sum_{j=1}^{n_i} x_{ij}$ is the expected mean of all training measurements. The projection matrix $W^*$ of FLDA is defined by

\[ W^* = \arg \max_W \text{tr} \left( (W^T S_w W)^{-1} W^T S_b W \right). \]

The projection matrix $W^*$ is computed from the eigenvectors of $S_w^{-1} S_b$, under the assumption that $S_w$ is invertible. If $c$ is equal to 2, FLDA reduces to Fisher’s discriminant analysis [14]; otherwise, FLDA is known as Rao’s multiple discriminant analysis [31]. After the linear subspace selection step, LDA is chosen as the classifier for classification. LDA, a statistical classification method, assumes the probability distribution of each class is Gaussian with different means but identical class covariances. This is consistent with the assumption of FLDA, i.e., if measurements in different classes are randomly drawn from Gaussian densities with different expected means but identical covariances, then FLDA maximizes the mean value of the KL divergences between all pairs of densities. In summary, FLDA maximizes all KL divergences between different pairs of densities with the homoscedastic Gaussian assumption. This is proved in the Observation 1 in Section 3.1.

FLDA is a preprocessing step for LDA to enhance the classification performance. This is not the only way for this objective and related works are listed as below. Friedman [12] has proposed the regularized discriminant analysis (RDA), which is a classification tool to smooth out the effects of ill or poorly conditioned covariance estimates due to the lack of training measurements. RDA is a combination of the ridge shrinkage [28], LDA, and quadratic discriminant analysis (QDA) [13]. It provides many regularization alternatives and is an intermediate classifier between the linear, the quadratic, and the nearest means classifier. It performs well but fails to provide interpretable classification rules. To solve this problem, Benhmail and Celeux [2] proposed the regularized Gaussian discriminant analysis (RGDA) by reparameterizing class covariance matrices, and the optimization stage is based on the eigenvalue decomposition. It provides a clear classification rule and performs at least as well as RDA. By introducing the reduced rank step, both RDA and RGDA can be applied for discriminative subspace selection. However, all extensions will have the class separation problem because they do not consider the effects of different distances between different classes. Bouveyron et al. [3] introduced the high-dimensional discriminant analysis (HDDA) for classification. HDDA reduces dimensions for different classes independently and regularizes class covariance matrices by assuming classes are spherical in their eigenspace. It can be deemed as a generalization of LDA and QDA. In this paper, our focus is the discriminative subspace selection, because it is important for not only classification but also visualization of measurements.
3 GEOMETRIC MEAN FOR SUBSPACE SELECTION

In FLDA, the arithmetic mean of the divergences is used to find a suitable subspace into which to project the feature vectors. The main benefit of using the arithmetic mean is that the projection matrix can be obtained by the generalized eigenvalue decomposition. However, FLDA has the class separation problem defined in Section 1, and it is not optimal for multiclass classification [28, 23, 25, 27] because the effects of small divergences are not emphasized in a proper way. This will merge two classes with a small divergence in the projected subspace so that large divergences can be preserved as much as possible to meet the objective for subspace selection. To reduce this problem, the geometric mean combined with the KL divergence is carefully studied in this Section because this setting emphasizes effects of small divergences and thus reduces the class separation problem significantly in the projected subspace. Moreover, the geometric mean can also be applied to combine with Bregman divergence, as discussed in [37].

For Gaussian probability density functions, \( p_i = N(x; \mu_i, \Sigma_i) \), where \( \mu_i \) is the mean vector of the \( i \)th class measurements, and \( \Sigma_i \) is the within-class covariance matrix of the \( i \)th class, the KL divergence [5] is

\[
D(p_i \parallel p_j) = \frac{1}{2} \ln |\Sigma_j| - \ln |\Sigma_i| + \text{tr}(\Sigma_j^{-1} \Sigma_i) + \text{tr}(\Sigma_j^{-1} D_{ij}),
\]

where \( D_{ij} = (\mu_i - \mu_j) (\mu_i - \mu_j)^T \) and \( |\Sigma| = \text{det}(\Sigma) \). To simplify the notation, we denote the KL divergence between the projected densities \( p(W^T x | y = i) \) and \( p(W^T x | y = j) \):

\[
D_W(p_i \parallel p_j) = D(p(W^T x | y = i)) - p(W^T x | y = j)) = \frac{1}{2} \ln |W^T \Sigma_i W| - \ln |W^T \Sigma_j W| + \text{tr}(W^T \Sigma_j W)^{-1} (W^T (\Sigma_i + D_{ij}) W). \tag{4}
\]

3.1 General Averaged Divergence Analysis

We replace the arithmetic mean by the following general mean,

\[
V_\varphi(W) = \varphi^{-1} \left[ \frac{\sum_{1 \leq i \neq j \leq c} q_i q_j \varphi(D_W(p_i \parallel p_j))}{\sum_{1 \leq i \neq j \leq c} q_i q_j} \right], \tag{5}
\]

where \( \varphi(\cdot) \) is a strict monotone real-valued increasing function defined on \((0, +\infty)\); \( \varphi^{-1}(\cdot) \) is the inverse function of \( \varphi(\cdot) \); \( q_i \) is the prior probability of the \( i \)th class (usually, we can set \( q_i = n_i / n \) or simply set \( q_i = 1 / c \); \( n_i \) is the conditional distribution of the \( i \)th class; \( x \in \mathbb{R}^n \), where \( \mathbb{R}^n \) is the feature space containing the training measurements; and \( W \in \mathbb{R}^{n \times k} \) is the projection matrix. The general averaged divergence function measures the average of all divergences between pairs of classes in the subspace. We obtain the projection matrix \( W^* \) by maximizing the general averaged divergence function \( V_\varphi(W) \) over \( W \) for a fixed \( \varphi(\cdot) \).

On setting \( \varphi(x) = x \) in (5), we obtain the arithmetic mean-based method for choosing a subspace:

\[
W^* = \arg \max_W \sum_{1 \leq i \neq j \leq c} q_i q_j D_W(p_i \parallel p_j) / \sum_{1 \leq i \neq j \leq c} q_i q_j = \arg \max_W \sum_{1 \leq i \neq j \leq c} q_i q_j D_W(p_i \parallel p_j). \tag{6}
\]

Observation 1. FLDA maximizes the arithmetic mean of the KL divergences between all pairs of classes, under the assumption that the Gaussian distributions for the different classes all have the same covariance matrix. The projection matrix \( W^* \) in FLDA can be obtained by maximizing a particular \( V_\varphi(W) \).

Proof. According to (3) and (4), the KL divergence between the \( i \)th class and the \( j \)th class in the projected subspace with the assumption of equal covariance matrices (\( \Sigma_i = \Sigma_j = \Sigma \))

\[
D_W(p_i \parallel p_j) = \frac{1}{2} \text{tr} \left( (W^T \Sigma W)^{-1} W^T D_{ij} W \right) + \text{constant}. \tag{7}
\]

Then, we have

\[
W^* = \arg \max_W \sum_{1 \leq i \neq j \leq c} q_i q_j D_W(p_i \parallel p_j) = \arg \max_W \left( q_i q_j \text{tr} \left( (W^T \Sigma W)^{-1} W^T D_{ij} W \right) \right) = \arg \max_W \left( (W^T \Sigma W)^{-1} W^T \left( \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} q_i q_j D_{ij} \right) W \right). \tag{8}
\]

Because \( S_b = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} q_i q_j D_{ij} \), as proved by Loog in [24], and \( S_i = S_b + S_w = \Sigma \) (see [13]), we have

\[
W^* = \arg \max_W \left( (W^T S_w W)^{-1} W^T S_b W \right). \tag{8}
\]

It follows from (8) that a solution of FLDA can be obtained by the generalized eigenvalue decomposition.

Example. Decell and Mayekar [8] maximize the summation of all symmetric KL divergences between all pairs of classes in the projected subspace. In essence, there is no difference between [8] and maximizing the arithmetic mean of all KL divergences.

3.2 Criterion 1: Maximization of the Geometric Mean of the Divergences

The log function is a suitable choice for \( \varphi \) because it increases the effects of the small divergences and, at the same time, reduces the effects of the large divergences. On setting \( \varphi(x) = \log(x) \) in (5), the generalized geometric mean of the divergences is obtained. The required subspace \( W^* \) is given by
TAO ET AL.: GEOMETRIC MEAN FOR SUBSPACE SELECTION

\[ W^* = \arg \max_w \prod_{1 \leq i \neq j \leq c} \left[ D_W(p_i || p_j) \right]^{\frac{q_i}{\sum_{m \neq n} q_m q_n}}. \]  

(9)

It follows from the mean inequality that the generalized geometric mean is upper bounded by the arithmetic mean of the divergences, i.e.,

\[ \prod_{1 \leq i \neq j \leq c} \left[ D_W(p_i || p_j) \right]^{\frac{q_i}{\sum_{m \neq n} q_m q_n}} \leq \sum_{1 \leq i \neq j \leq c} \left( \frac{q_i q_j}{\sum_{m \neq n} q_m q_n} D_W(p_i || p_j) \right). \]  

(10)

Furthermore, (9) emphasizes the total volume of all divergences, e.g., in the special case \( q_i = q_j \) for all \( i, j \):

\[ \arg \max_w \prod_{1 \leq i \neq j \leq c} \left[ D_W(p_i || p_j) \right]^{\frac{q_i}{\sum_{m \neq n} q_m q_n}} = \arg \max_w \prod_{1 \leq i \neq j \leq c} \left[ D_W(p_i || p_j) \right]^{q_i} = \arg \max_w \prod_{1 \leq i \neq j \leq c} D_W(p_i || p_j). \]  

(11)

3.3 Criterion 2: Maximization of the Geometric Mean of the Normalized Divergences

We can further strengthen the effects of the small divergences on the selected subspace by maximizing the geometric mean of all normalized divergences in the projected subspace, i.e.,

\[ W^* = \arg \max_w \left[ \prod_{1 \leq i \neq j \leq c} E_W(p_i || p_j) \right]^{\frac{1}{c(c-1)}}. \]  

(12)

where the normalized divergence \( E_W(p_i || p_j) \) between the \( i \)th class and the \( j \)th class is defined by

\[ E_W(p_i || p_j) = \frac{q_i q_j D_W(p_i || p_j)}{\sum_{1 \leq m \neq n \leq c} q_m q_n D_W(p_m || p_n)}. \]  

(13)

The intuition behind (12) is that the product of the normalized divergences achieves its maximum value if and when the normalized divergences are equal to each other. Therefore, maximizing the geometric mean of the normalized divergences tends to make the normalized divergences as similar as possible. And, thus, the effects of the small divergences are further emphasized.

3.4 Criterion 3: Maximization of the Geometric Mean of All Divergences

Although criterion 2 emphasizes the small divergences during optimization, direct use of this criterion is not desirable for subspace selection. This is because experiments in Section 4.1 show that there exist \( W \) for which all the divergences become small, but all the normalized divergences are comparable in value. In such cases, the projection matrix \( W \) is unsuitable for classification, because several classes may be severely overlapped.

To reduce this problem, we combine criterion 2 with criterion 1. The new criterion maximizes the linear combination of 1) the log of the geometric mean of the divergences and 2) the log of the geometric mean of the normalized divergences. This criterion is named the MGMD:

\[ W^* = \arg \max_w \left\{ \alpha \log \left[ \prod_{1 \leq i \neq j \leq c} E_W(p_i || p_j) \right]^{\frac{1}{c(c-1)}} - (1 - \alpha) \log \left[ \prod_{1 \leq i \neq j \leq c} D_W(p_i || p_j) \right]^{\frac{q_i}{\sum_{m \neq n} q_m q_n}} \right\}. \]  

(14)

where the supremum of \( \alpha \) is 1, and the infimum of \( \alpha \) is 0. When \( \alpha = 0 \), (14) reduces to (9); and when \( \alpha = 1 \), (14) reduces to (12). By setting \( q_i = 1/c \), we can simplify the above formula as

\[ W^* = \arg \max_w \left\{ \sum_{1 \leq i \neq j \leq c} \log D_W(p_i || p_j) - \alpha c(c-1) \log \left( \sum_{1 \leq i \neq j \leq c} D_W(p_i || p_j) \right) \right\}. \]

Based on (14), we define the value of the objective function as

\[ L(W) = \frac{1}{c(c-1)} \sum_{1 \leq i \neq j \leq c} \log D_W(p_i || p_j) - \log \left( \sum_{1 \leq i \neq j \leq c} q_i q_j D_W(p_i || p_j) \right) + \frac{1 - \alpha}{\alpha} \sum_{1 \leq m \neq n \leq c} q_m q_n \log D_W(p_m || p_n). \]  

(15)

Therefore, \( W^* = \arg \max_w L(W) \). The objective function \( L(W) \) depends only on the subspace spanned by the columns of \( W \) or \( L(W) = L(WQ) \) when \( Q \) is an orthogonal \( r \times r \) matrix. This is because \( D_W(p_i || p_j) = D_{WQ}(p_i || p_j) \) according to [5] and [13]. In Section 3.5, we discuss how to obtain solutions for (14) based on the gradient steepest ascent optimization procedure.

3.5 Optimization Procedure of MGMD

The gradient steepest ascent could be applied here to obtain a solution of MGMD. That is, the projection matrix can be updated iteratively according to

\[ W \leftarrow W + \kappa \cdot \partial_W L(W), \]
where $\kappa$ is the learning rate parameter, and it affects the learning procedure. If $\kappa$ is very small, it takes a large number of iterations to obtain a local optimal solution for the projection matrix $W$. This means the training procedure will be time consuming. Otherwise, the training procedure fails to converge to a local optimal solution for the projection matrix $W$. To find a suitable $\kappa$ is not easy. In this paper, we just set it as a small value as 0.001 to guarantee the convergence of the training stage. To obtain the optimization procedure for MGMD, we need the first-order derivative of $L(W)$,

$$
\partial_W L(W) = \left( \sum_{1 \leq i \neq j \leq c} \left( \frac{1}{c(c-1)} + \frac{(1-\alpha)q_i q_j}{\sum_{1 \leq m \neq n \leq c} q_m q_n} \right) D^1_W(p_i \parallel p_j) \partial_W D_W(p_i \parallel p_j) \right) - \left( \sum_{1 \leq m \neq n \leq c} q_m q_n D_W(p_m \parallel p_n) \right)^{-1} \left( \sum_{1 \leq i \neq j \leq c} q_i q_j \partial_W D_W(p_i \parallel p_j) \right)
$$

(16)

and

$$
\partial_W D_W(p_i \parallel p_j) = \left( \Sigma_i W(W^T \Sigma_i W)^{-1} - \Sigma_j W(W^T \Sigma_j W)^{-1} \right) + \left( \Sigma_i + D_{ij} \right) W(W^T \Sigma_j W)^{-1} W^T (\Sigma_i + D_{ij}) W(W^T \Sigma_j W)^{-1}
$$

(17)

With (16) and (17), the gradient steepest ascent-based optimization procedure of MGMD is given in Table 1. The time complexity of MGMD is around $c(c-1)\left| 110(n^2k) + 100(k^3) + 60(nk^2) \right| MT$ for $M$ different initializations and $T$ iterations in gradient steepest ascent optimization. This cost is approximated based on the Schoolbook matrix multiplication [33]. By applying Coppersmith-Winograd algorithm [33], the cost can be further reduced.

Note that the concavity of $L(W)$ cannot be guaranteed. To reduce the effects of local maxima [1], a number of different initial projection matrices can be generated, and we can carry out independent optimizations and then select the best one. However, empirical study in Section 5.3 shows that we may not meet this problem for some cases.

### 3.6 Related Works

FLDA does not fully utilize the discriminative information contained in the covariances of different classes; it models each class with a single Gaussian density; and it fairly considers all KL divergences between different pairs of classes. Therefore, FLDA has the corresponding three problems: 1) heteroscedastic problem, 2) multimodal problem, and 3) class separation problem. Moreover, in practical applications, e.g., biometrics research, FLDA encounters the undersampled problem [41], because the number of training measurements is less than the dimension of the feature space. To reduce these problems, a number of extensions have been developed.

**Heteroscedastic problem.** Decell and Mayekar [8] proposed a method to obtain a subspace to maximize the average interclass divergences, which measure the separations between the classes. This criterion takes into account the discriminative information preserved in the covariances of different classes. De la Torre and Kanade [7] developed the oriented discriminant analysis (ODA) based on the same objective function used in [8] but used iterative majorization to obtain a solution. Iterative majorization speeds up the training stage. Each class is modeled by a GMM. Jelinek [21] selected discriminative subspace by modeling each class covariance separately. This resulted in the heteroscedastic discriminant analysis (HDA). The solution is given by maximizing $J(W) = n \log \left( \frac{1}{n} \sum_{i=1}^{c} \frac{1}{n_i} \log \left| W^T S_i W \right| \right)$, where $n = \sum_{i=1}^{c} n_i$. Loog and Duin [26] introduced the Chernoff criterion to heteroscedasticity for FLDA, i.e., the heteroscedastic extension of LDA (HFLDA). HFLDA takes the class covariance information into account, so it can deal with the heteroscedastic problem. HFLDA works more efficiently than HDA because its solution is given by the eigenvalue decomposition. The main drawback of the abovementioned variants is they fairly deal with all interclass divergences, so they have the class separation problem.

**Multimodal problem.** Hastie and Tibshirani [19] combined GMM with FLDA based on the fact that FLDA is equivalent to the maximum likelihood classification when each class is modeled by a single Gaussian distribution. The extension directly replaces the original single Gaussian in

<table>
<thead>
<tr>
<th>Input</th>
<th>Training measurements $x_{i,j}$, where $i$ denotes the $i^{th}$ class ($1 \leq i \leq c$) and $j$ denotes the $j^{th}$ measurement in the $i^{th}$ class ($1 \leq j \leq n_i$), the dimensionality of selected features $k &lt; n$ (is the dimensionality of $x_{i,j}$), the maximum number $M$ of different initial values for the projection matrix $W$, the learning rate $\kappa$ (a small value), the combination factor $\eta$ , and a small value $\epsilon$ as the convergence condition.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>An estimate $W^*$ of the projection matrix.</td>
</tr>
<tr>
<td>1.</td>
<td><strong>for</strong> $m = 1 : M$</td>
</tr>
<tr>
<td>2.</td>
<td><strong>Initialize</strong> $W^w_m$ ($t = 1$) randomly.</td>
</tr>
<tr>
<td>3.</td>
<td><strong>while</strong> $L(W_{m}^w) - L(W_{m-1}^w) &gt; \epsilon$</td>
</tr>
<tr>
<td>4.</td>
<td><strong>do</strong></td>
</tr>
<tr>
<td>5.</td>
<td><strong>while</strong> on line 3</td>
</tr>
<tr>
<td>6.</td>
<td><strong>t</strong> $\leftarrow$ $t$ $+ 1$</td>
</tr>
<tr>
<td>7.</td>
<td><strong>if</strong> on line 1</td>
</tr>
<tr>
<td>8.</td>
<td>$W^w$ $\leftarrow$ $W_t^w$</td>
</tr>
<tr>
<td>9.</td>
<td><strong>else</strong></td>
</tr>
<tr>
<td>10.</td>
<td>$W$ $\leftarrow$ arg max$_w L(W_w)$</td>
</tr>
<tr>
<td>11.</td>
<td><strong>Orth-normalization Step.</strong> $W^*$ $\leftarrow$ orth-normalize($W$).</td>
</tr>
</tbody>
</table>
each class by a Gaussian mixture model (GMM) in Campbell's result [4]. De la Torre and Kanade [7] generalized ODA for a multimodal case as the multimodal ODA (MODA) by combining it with the GMMs learned by normalized cut [34].

**Class separation problem.** Lotlikar and Kothari [23] found for discriminative subspace selection that it is important to consider the nonuniform distances between classes, i.e., some distances between different classes are small, while others are large. To take this information into account, weightings are updated iteratively. According to the analysis for the class separation problem, weightings are helpful to reduce the class separation problem. It is worth noting that Lotlikar and Kothari did not mention this problem formally, and polynomial functions over distances between different classes are applied as weightings in a heuristic way. Unlike [23], Loog et al. [25] claimed that \( (1/2x^2) \text{erf}(x/\sqrt{2}) \) is a suitable weighting function, and \( x \) is the distance between two classes weighted by the inverse of the total covariance matrix. In Loog's weighting scheme, iterative optimization procedure is avoided and eigenvalue decomposition is applied to have the projection matrix for subspace selection. Moreover, Loog et al. proved that their proposed criterion approximates the mean classification accuracy. However, both methods have the class separation problem with the heteroscadastic setting, as demonstrated in Section 4.3, and it is difficult to extend them directly for heteroscadastic setting.

**Undersampled problem.** Friedman [12] proposed the RDA, which is a combination of ridge shrinkage, LDA, and QDA. Hastie et al. [17] viewed FLDA as multivariate linear regression and used the penalized least squares regression to reduce the SSS problem. Swets and Weng [36] introduced PCA as the preprocessing step in FLDA for face recognition. Raudys and Duin [32] applied the pseudoinverse to the covariance matrix in FLDA to reduce this problem. Ye et al. [41] introduced the generalized singular value decomposition (GSVD) to avoid this problem in FLDA. Ye and Li [42] combined the orthogonal triangular decomposition with FLDA to reduce the SSS problem.

**Nonparametric models.** Recently, to take the non-linearity of the measurement distribution into account, classification driven nonparametric models have been developed and achieved improvements. Hastie and Tibshirani [20] realized that the nearest neighbor classifier suffers from bias in high-dimensional space, and this affects the classification accuracy. To reduce this problem, LDA is applied locally to estimate the local distance metric to estimate neighborhood measurements for future classification. By reformulating the within-class scatter matrix and the between-class scatter matrix defined in FLDA in a pairwise manner, Sugiyama [35] proposed the local FLDA (LFLDA). The original objective of LFLDA is to reduce the multimodal problem in FLDA. Torkkola [40] applied the nonparametric Renyi entropy to model the mutual information between measurements in the selected subspace and corresponding labels. Because the mutual information is relevant to the upper bound of the Bayes error rate, it can be applied to select discriminative subspaces. Fukumizu et al. [15] developed the kernel dimensionality reduction (KDR) based on the estimation and optimization of a particular class of operators on the reproducing kernel Hilbert space, which provides characterizations of general notions of independence. These characterizations are helpful to design objective functions for discriminative subspace selection. Peltonen and Kaski [30] generalized FLDA through a probabilistic model by maximizing mutual information. Unlike [40], they applied Shannon entropy for probabilistic inference. This model is equivalent to maximizing the mutual information with classes asymptotically. Zhu and Hastie [44] developed a general model for discriminative subspace selection without assuming class densities belong to a particular family. With nonparametric density estimators, this model maximizes the likelihood ratio between class-specific and class-independent models. Goldberger et al. [16] developed the neighborhood components analysis (NCA) for learning a distance metric for the \( k \)-nearest-neighbor (\( k \text{NN} \)) classification. It maximizes a stochastic variant of the leave-one-out \( k \text{NN} \) score over training measurements, and it is a nonparametric classification driven subspace selection method. Although nonparametric models achieve improvements for classification and do not assume class densities belong to any particular family, it is still important to study models with specific prior probability models for understanding statistical properties of measurements. In addition, the new method could in principle be faster than several nonparametric methods since it only uses the covariance and \( D_2 \) matrices for optimization, not the individual data point locations.

## 4 A COMPARATIVE STUDY USING SYNTHETIC DATA

In this section, we compare the effectiveness of MGMD with FLDA [13], HDA [21], aPAC [25], weighted FLDA (WFLDA) [13], FS-FLDA [23], HFLDA [26], ODA [7], and MODA [7] in handling three typical problems (heteroscedastic problem, multimodal problem, and class separation problem) appeared in FLDA. WFLDA [13] is similar to aPAC, but the weighting function is \( d^{-5} \), where \( d \) is the original-space distance between a pair of class centers. In FS-FLDA, the weighting function is \( d^{-5} \), and the number of fractional steps is 30. We denote the proposed method as MGMD(\( \alpha \)), where \( \alpha \) is the combination factor defined in (14). The empirical studies in this section demonstrate that MGMD has no heteroscedastic problem; the multimodal extension of MGMD has no multimodal problem, and MGMD can significantly reduce the class separation problem. To better understand differences of these algorithms and to visualize measurements conveniently, measurements are generated in two-dimensional (2D) spaces. In the following experiments, projections with less classification errors are better choices as a classification goal because they better separate classes in the projected subspace. We choose the nearest class for each measurement by Mahalanobis distance to the class center for classification.

### 4.1 Heteroscedastic Example

To examine the classification ability of these subspace selection methods for the heteroscedastic problem [26], we generate two classes such that each class has 500 measurements, drawn from a Gaussian distribution. The two classes have identical mean values but different covariances. As shown in Fig. 2, FLDA, aPAC, WFLDA, and FS-FLDA separate class means without taking the differences
4.2 Multimodal Example

In many applications, it is useful to model the distribution of a class using a GMM, because measurements in the class may be drawn from a non-Gaussian distribution. The multimodal extension of MGMD (M-MGMD) is defined as

\[ W^* = \arg \max_W \]

\[
\begin{align*}
&\frac{1}{\alpha(n-1)} \sum_{i \leq j < n} \sum_{1 \leq C_i \leq C_j} \sum_{1 \leq i_1 \leq C_i} \log \left[ D_W \left( p_i^{(i)} \parallel p_j^{(j)} \right) \right] \\
&- \log \left[ \sum_{1 \leq i_1 \leq C_i} \sum_{1 \leq C_j} \sum_{1 \leq i_1 \leq C_i} \sum_{1 \leq j \leq C_j} q_i q_j d^2 \left[ D_W \left( p_i^{(i)} \parallel p_j^{(j)} \right) \right] \right] \\
&+ \alpha \sum_{1 \leq i_1 \leq C_i} \sum_{1 \leq C_j} \sum_{1 \leq i_1 \leq C_i} \sum_{1 \leq j \leq C_j} q_i q_j d^2 \left[ \log D_W \left( p_i^{(i)} \parallel p_j^{(j)} \right) \right]
\end{align*}
\]

where \( q_i \) is the prior probability of the \( i \)th subcluster of the \( i \)th class, \( p_i^{(i)} \) is measurement probability of the \( i \)th subcluster in the \( i \)th class, and \( D_W(p_i^{(i)} \parallel p_j^{(j)}) \) is the divergence between the \( j \)th subcluster in the \( i \)th class and the \( i \)th subcluster in the \( j \)th class. The parameters for these subclusters can be obtained from the GMM-FJ method [11], which is a GMM-EM-like algorithm, proposed by Figureiredo and Jain. That is, we first run the GMM-FJ to obtain relevant parameters and then apply (18) for discriminative subspace selection. The merits of the incorporation of MGMD into GMM-FJ are given as follows:

1. It reduces the class separation problem, defined in Section 1.
2. It inherits the merits of GMM-FJ [11]. In detail, it determines the number of subclusters in each class automatically; it is less sensitive to the choice of initial values of the parameters than EM; and it avoids the boundary of the parameter space.
3. The objective function \( L_{GMM}(W) \) is invariant to rotation transformation.

To demonstrate the classification ability of M-MGMD, we generate two classes; each class has two subclusters; and measurements in each subcluster are drawn from a Gaussian distribution. Fig. 4 shows the selected subspaces of different methods. In this case, FLDA, WFLDA, FS-FLDA, and aPAC do not select the suitable subspace for classification. However, the multimodal extensions of ODA and MGMD can find the suitable subspace. Furthermore, although HDA and HFLDA do not take account of multimodal classes, they can select the suitable subspace. This is because in this case, the two classes have similar class means but significantly different class covariance matrices when each class is modeled by a single Gaussian
distribution. For complex cases, e.g., when each class consists of more than three subclusters, HDA and HFLDA will fail to find the optimal subspace for classification.

### 4.3 Class Separation Problem

The most prominent advantage of MGMD is that it can significantly reduce the classification errors caused by the strong effects of the large divergences between certain classes. To demonstrate this point, we generate three classes, and measurements in each class are drawn from a Gaussian distribution. Two classes are close together and the third is further away.

In this case, MGMD(0) can also reduce the problem, but it cannot work as well as MGMD(5/6). In Fig. 5, MGMD(5/6) shows a good ability of separating the last two classes of measurements. Furthermore, in MGMD, we achieve the same projection direction when setting $\alpha$ to 1/6, 2/6, 3/6, 4/6, and 5/6 for this data set. However, FLDA, HDA, WFLDA, HFLDA, and ODA do not give good results. FS-FLDA and aPAC algorithms are better than FLDA, but neither of them gives the best projection direction.

In Fig. 5, different Gaussians have identical class covariances. In this case, FS-FLDA works better than aPAC. In Fig. 6, different Gaussians have different class covariances. In this case, aPAC works better than FS-FLDA. Moreover, in both cases, the proposed MGMD(5/6) achieves the best performance. Therefore, it is important to consider the heteroscedastic problem and the class separation problem simultaneously.

### 5 Statistical Experiments and Analysis

In this section, we utilize a synthetic data model, which is a generalization of the data generation model used by Torre and Kanade [7] to evaluate MGMD in terms of accuracy and robustness. The accuracy is measured by the average error rate and the robustness is measured by the standard deviation of the classification error rates. In this data generation model, there are five classes, which are represented by the symbols $\times$, $\circ$, $+$, $\Box$, and $\diamond$ in Fig. 11. In our experiments, for each of the training and testing sets, the data generator gives 200 measurements for each of the five classes (therefore, there are 1,000 measurements in total). Moreover, the measurements in each class are obtained from a single Gaussian. Each Gaussian density is a linear transformation of a “standard normal distribution”. The linear transformations
are defined by $x_{ij} = T x_i + n_j$, where $x_{ij} \in \mathbb{R}^{20}$, $T_i \in \mathbb{R}^{20 \times 20}$, $z \sim N(0, 1) \in \mathbb{R}^7$, $n \sim N(0, 2i) \in \mathbb{R}^{20}$, $i$ denotes the $ith$ class, $j$ denotes the $jth$ measurement in this class, and $\mu_i$ is the mean value of the corresponding normal distribution. The $\mu_i$ are assigned as $\mu_1 = (2N(0, 1) + 4)1_{20}$, $\mu_2 = 0_{20}$, $\mu_3 = (2N(0, 1) - 4)1_{10}$, $\mu_4 = (2N(0, 1) + 4)1_{10}$, $\mu_5 = (2N(0, 1) + 4)1_{10}$, $\mu_6 = (2N(0, 1) + 4)1_{10}$. The projection matrix $T_i$ is a random matrix. Each of its elements is sampled from $N(0, 5)$. Based on this data generation model, 800 groups (each group consists of the training and testing measurements) of synthetic data are generated.

For our comparative study, subspace selection methods, e.g., MGMD, are first utilized to select a given number of features. Then, the nearest neighbor rule [9] and the standard deviations of error rates are computed on different feature dimensionalities from 1 to 6. Correspondingly, the standard deviations under each condition, which measure the robustness of the classifiers, are given in Tables 3 and 5.

We have 20 feature dimensions for each measurement, and all the measurements are divided into five classes, and therefore, the maximal feature number for FLDA, HDA, aPAC, WFLDA, and FS-FLDA is $5 - 1 = 4$; in contrast, LFLDA, HFLDA, ODA, and MGMD can extract more features than FLDA and HDA. From Tables 2, 3, 4, and 5, it can be concluded that MGMD usually outperforms FLDA, HDA, aPAC, WFLDA, FS-FLDA, LFLDA, HFLDA, ODA, and MGMD, consistently. Finally, the Mahalanobis distance outperforms the nearest neighbor rule when the data are strictly sampled from Gaussian distributions.

### 5.2 Justification of MGMD

In this part, we justify why MGMD is a suitable subspace selection method for classification. We first study the relationship between $L(W)$, which is defined in (15), and the training error rate obtained from the Mahalanobis distance-based classification. Experiments are done on a randomly selected data set from 800 data sets generated at the beginning of this section. We set training iterations to be 200 and $\alpha$ to be 0 and 0.75, because MGMD achieves its best performance by setting $\alpha$ as 0.75. By setting $\alpha$ to $2/(c - 1)$ as in Tables 2, 3, 4, and 5, we can have similar curves as those obtained by setting $\alpha$ to 0.75. Fig. 7a shows the
classification error rate trends to decrease with the increasing of the training iterations, and Fig. 7a shows the objective function values $L(W)$ monotonically increase with the increasing of the training iterations. Therefore, the classification error rate trend to decrease with the increasing of $L(W)$. This means maximizing $L(W)$ is useful to achieve a low classification error rate.

It is also important to investigate how KL divergences between different classes changes with the increasing number of training iterations, because it is helpful to deeply understand how and why MGMD reduces the class separation problem. In Figs. 8 and 9, we show how KL divergences change in MGMD(0) and MGMD(0.75) over 1st, 2nd, 5th, 10th, 20th, and 200th training iterations, respectively. There are five classes, so we can map KL divergences to a $5 \times 5$ matrix with zero diagonal values. The entry of the $i$th column and the $j$th row means the KL divergence between the $i$th class and the $j$th class. We denote it as $D_{ij}(W_t)$, where $t$ means the $t$th training iteration. Because KL divergence is not symmetric, the matrix is not symmetric, i.e., $D_{ij}(W_t) \neq D_{ji}(W_t)$.

According to that in Fig. 8, in the first training iteration (the top-left $5 \times 5$ matrix), there are eight values less than 2 in the matrix. In the second iteration (the top-right $5 \times 5$ matrix), there are only four values less than 2. Compared with the first iteration, 6 out of 8 are increased. In the fifth iteration (the middle-left $5 \times 5$ matrix), there are only two values that are less than 2, and they are increased in comparing with values in the first iteration. However, these two divergences are decreased to 1.0439 and 1.0366 in the 200th iteration (the bottom-left $5 \times 5$ matrix) to guarantee the increasing of $L(W)$. This is not suitable to separate classes, because the divergences between them are very small. To further increase these two values, according to Sections 3.3 and 3.4, the normalized divergences will be helpful here.

As shown in Fig. 9, in the first training iteration (the top-left $5 \times 5$ matrix), there are eight values less than 2 in the matrix. In the second iteration (the top-right $5 \times 5$ matrix), there are still eight values less than 2, but 7 out of 8 are increased compared with them in the first iteration. In the fifth iteration (the middle-left $5 \times 5$ matrix), there are only three values that are less than 2, and two of them are increased in comparing with the values from second iteration. In the 10th iteration (the middle-right $5 \times 5$ matrix), one of these three divergences is increased significantly but two of them are reduced slightly. After that, these two smallest divergences are always increasing. In the 200th iteration, they are increased to 1.1480 and 1.1517. In comparing with MGMD(0), which are 1.0439 and 1.0366, they are much larger in MGMD(0.75). Moreover, the classification error rate, obtained from the Mahalanobis distance-based classification, is also reduced from 21 percent to 19.3 percent with 8.1 percent improvement. Therefore, the normalized divergences are helpful to further reduce the class separation problem according to Sections 3.3 and 3.4.

### 5.3 Initial Values

We generate a training set according to the synthetic data model described at the beginning of this section and randomly initialize parameters in MGMD(2/(c-1)) to examine how different initial values affect the solution. Note that we omit lines 1, 2, 7, 8, and 9 from the optimization process for MGMD given in Table 1, because these lines are used to set different initial values. We can see in Fig. 10 that MGMD(2/(c-1)) is insensitive to the choice of initial values in 50 random experiments.

### 5.4 Nest Structure Property

**Definition.** Given a subspace selection method, $W_s = [w_i]_{1 \leq i \leq m}$ and $U_m = [u_i]_{1 \leq i \leq c}$ are projection matrices for selecting different dimensions of subspaces on the same training data. The $U_m$ selects more features than $W_s$, i.e., $m > n$. Let $U_n$ be the first $n$ columns of $U_m$. If $U_n = W_s Q$ and $Q$ is an orthogonal matrix, we say the subspace selection method has the nest structure property. For example, PCA holds the nest structure property.
In other words, it should avoid the nest dimensionality. In other words, it should avoid the nest dimensionality. In other words, it should avoid the nest dimensionality.

A desirable subspace selection method should adapt to the best dimensionality. In other words, it should avoid the nest structure property when the user chooses different dimensions of the subspace by projection from the original high-dimensional space. For instance, in our experiments, we first extract 2D features from the original entire feature set (20 features) based on \( \text{MGMD}(2/(c-1)) \); with these two features, the profile of the five classes of measurements are illustrated in Fig. 11a. We then extract 10 features based on \( \text{MGMD}(2/(c-1)) \) with the same training measurements, but we only show the first two dimensions of the extracted 10-dimensional features in Fig. 11b. Based on this figure, we can see that these first two features in the two cases are different. Moreover, we can obtain the same observation as the observation obtained from the above experiment, when we conduct experiments with the same procedure as the above experiment but with different data sets and different \( c \). Therefore, \( \text{MGMD}(2/(c-1)) \) does not have the nest structure.

Fig. 10. Initial values: the mean value and the corresponding standard deviation of the KL divergence between class \( i \) and class \( j \), over 50 different initial values in the (a) 10th, (b) 50th, (c) 100th, and (d) 1,000th training iterations. Because there are five classes in the training set, there are 20 KL divergences to examine. The circles in each subfigure show the mean values of the KL divergences for 50 different initial values. The error bars show the corresponding standard deviations. For a better visualization, the scale for showing the standard deviations is 10 times larger than the vertical scale in each subfigure. The standard deviations of these 20 KL divergences approach 0 as the number of training iterations increases.

Fig. 11. \( \text{MGMD}(2/(c-1)) \) does not have the nest structure property.

Fig. 12. The statistical averaged effects of the linear combination coefficient \( \alpha \) on classification error rate.

5.5 The Effects of the Linear Combination Coefficient \( \alpha \) on Classification Error Rate

In this experiment, we select six feature dimensions and conduct 800 statistical experiments, as described in Section 5.1, to study the effects of the linear combination coefficient \( \alpha \) on classification error rate obtained from the Mahalanobis distance-based classification. Let \( \alpha \) change from 0/20 to 19/20 with step 1/20, because \( \alpha \) is strictly less than 1, and the maximal value 1 means the algorithm would use the normalized divergences only. The statistical average over all 800 experiments shows the classification error rate changes slightly when \( 0/20 \leq \alpha \leq 13/20 \) (strictly, it is convex, but the curvature is very small), and the classification error rate increases with the increasing of \( \alpha \) when it changes from 14/20 to 19/20. This can be seen in Fig. 12. When \( 0/20 \leq \alpha \leq 13/20 \), the normalized divergences affects the classification error rate slightly because the class separation problem is not serious in many cases.

For practical applications, it is important to use the cross-validation method to select a suitable \( \alpha \) for different data sets, because \( \alpha \) affects the classification error rate to some extent. For data sets with slight class separation problem, we can set \( \alpha \) to be small or even zero. For data sets with severe class separation problem, we need to set \( \alpha \) to be large. This can be seen from subfigures in Fig. 13, which show how \( \alpha \) affects the classification error rates over six different data sets (randomly selected out of the 800 experiments). From Figs. 13a, 13b, 13c, 13d, 13e, and 13f, to achieve the smallest classification error rate over these data sets, we need to set \( \alpha \) as 0/20, 6/20, 7/20, 12/20, 15/20, and 11/20, respectively. For some data sets, identical classification error rates could be obtained with different \( \alpha \), as shown in Fig. 13e.

For every data set, Hinton windows are applied to show the following:

1. KL divergences between different classes in the original data set without dimension reduction (the top-left window in every subfigure).
2. KL divergences between different classes with six selected features by ODA (the top right window in every subfigure).
3. Differences between KL divergences obtained from MGMD(0) and KL divergences obtained from ODA (the bottom left window in every subfigure).
Fig. 13. The effects of the linear combination coefficient $\alpha$ on the classification error rate over six different data sets based on the Mahalanobis distance. To achieve the smallest classification error rate over these data sets, we need to set $\alpha$ as (a) 0/20, (b) 6/20, (c) 7/20, (d) 12/20, (e) 15/20, and (f) 11/20, respectively.

4. Differences between KL divergences obtained from MGMD($\alpha$) and KL divergences obtained from ODA (the bottom right window in every subfigure).

Because there are five classes, KL divergences are mapped to a $5 \times 5$ matrix with zero diagonal values. The size of a block represents the value of the entry. The larger of the block is, the larger of the absolute value of the corresponding entry is. The green (light) block means the entry is positive, and the red (dark) block means the entry is negative.

Based on these Hinton windows, small divergences in MGMD(0) are always larger than the corresponding divergences in ODA indicated by green (light) color in the bottom Hinton diagrams in each subfigure, and large divergences in MGMD(0) are always smaller than the corresponding divergences in ODA indicated by red (dark) color in the bottom Hinton diagrams in each subfigure. Moreover, the MGMD($\alpha$) helps to further enlarge small divergences in the projected low-dimensional subspace, and the cost is the large divergences are further reduced. This is because the normalized divergences are taken into account. Although large $\alpha$ is helpful to enlarge small divergences in subspace selection, some classes will be merged together, and this enlarges the classification error. Therefore, with suitable $\alpha$ in MGMD, small divergences will be enlarged suitably for discriminative subspace selection.

5.6KL Divergence versus Symmetric KL Divergence

In this part, we compare the KL divergence with the symmetric divergence in classification based on the proposed MGMD. With (3), the symmetric KL divergence is

$$S(p_i \| p_j) = \frac{1}{2} \left[ D(p_i \| p_j) + D(p_j \| p_i) \right]$$

$$= \frac{1}{4} \left[ \text{tr} \left( \Sigma_j^{-1} \Sigma_i \right) + \text{tr} \left( \Sigma_i^{-1} \Sigma_j \right) \right] + \text{tr} \left( \left( \Sigma_j^{-1} + \Sigma_i^{-1} \right) D_{ij} \right).$$

(19)

To simplify the notation, we denote the symmetric KL divergence between the projected densities $p(W^T x | y = i)$ and $p(W^T x | y = j)$ by

$$S_{W}(p_i \| p_j) = S(p(W^T x | y = i) \| p(W^T x | y = j))$$

$$= \frac{1}{4} \left[ \text{tr} \left( (W^T \Sigma_i W)^{-1} (W^T (\Sigma_j + D_{ij}) W) \right) \right] + \text{tr} \left( (W^T \Sigma_j W)^{-1} (W^T (\Sigma_j + D_{ij}) W) \right).$$

(20)

Therefore, (14) changes to

$$W^* = \arg \max_{W} \left\{ \frac{1}{2} \sum_{1 \leq i \neq j \leq c} \log S_{W}(p_i \| p_j) \right\} - \log \left( \sum_{1 \leq m \neq n \leq c} q_m q_n S_{W}(p_m \| p_n) \right) + \frac{(1-\alpha)}{\alpha} \sum_{1 \leq m \neq n \leq c} q_m q_n \sum_{1 \leq i \neq j \leq c} q_i q_j \log S_{W}(p_i \| p_j) \right\},$$

(21)

where the supremum of $\alpha$ is 1, and the infimum of $\alpha$ is 0. When $\alpha = 0$, (14) reduces to (9); and when $\alpha = 1$, (14) reduces to (12).

For empirical justification, experiments are conducted on 800 data sets for both KL divergence-based and symmetric KL divergence-based MGMD, and the linear combination coefficient $\alpha$ changes from 0/20 to 19/20 with step 1/20. Therefore, two matrices $E_1$ and $E_2$ with size $800 \times 20$ are obtained the entries of $E_1$ and $E_2$ store the classification error rate, which is obtained from the Mahalanobis distance-based classification, for a specific data set and a specific $\alpha$ for KL divergence-based and symmetric KL divergence-based MGMD, respectively. Fig. 14a describes the error bar information of the difference between $E_1$ and $E_2$, i.e., $E_1 - E_2$, over different $\alpha$. This figure shows the difference is very small in terms of classification error rates between KL divergence-based and symmetric divergence-based MGMD when $\alpha$ changes from 0/20 to 13/20. When $\alpha$ changes from 14/20 to 19/20, the difference is a little large. These two observations are based on the length of the corresponding error bars. In Fig. 15b, we show this point based on some specific data sets. In this figure, the $x$-coordinate represents the difference of the classification error rates, obtained from the Mahalanobis distance-based classification,
between KL divergence-based and symmetric KL divergence-based MGMD. The error bars in the right subfigure are calculated over different $\alpha$, which changes from 0/20 to 19/20. Usually, the difference of the classification error rates between KL divergence-based and symmetric KL divergence-based MGMD is very small because averaged difference over different $\alpha$ is almost zero, and the corresponding error bar is short for different data sets. However, for the 9th, 18th, 24th, 26th, and 35th data sets here, the averaged differences are not zero, and the corresponding standard deviations are large. Therefore, the cross validation could be also applied to select suitable divergence, i.e., KL divergence or symmetric KL divergence.

6 EXPERIMENTS WITH UCI MACHINE LEARNING REPOSITORY

To justify the significance of the geometric mean for subspace selection, we conduct experiments to compare MGMD($0/\alpha$) with ODA over six data sets in the UCI Machine Learning Repository [29]. In all experiments, we set $\alpha$ to be $2/c$, and the nearest neighbor rule is applied for classification. Variations on the averaged classification errors are obtained from conventional fivefold cross validation over different random runs, i.e., measurements in each data set are divided into five folds, and one fold is used for testing and the rest for training in each turn.

Fig. 15a shows the performance comparison on “Car” data set, which consists of 1,728 measurements from four classes in $R^6$. Fig. 15b shows the results on “Ecoli” data set, which consists of 336 measurements from eight classes in $R^5$. Fig. 15c shows the results on “Flare” data set, which consists of 1,389 measurements from five classes in $R^6$. Fig. 15d shows the results on “Glass” data set, which consists of 214 measurements from seven classes in $R^{16}$. Fig. 15e shows the results on “Pendigit” data set, which consists of 10,992 measurements from 10 classes in $R^{10}$. Fig. 15f shows the results on “Yeast” data set, which consists of 1,484 measurements from seven classes in $R^{17}$. MGMD($\alpha$) outperforms ODA on all but the Yeast data, and MGMD(0) outperforms ODA on all but the Yeast and Flare data. When the number of the selected features is large enough, they perform similarly.

7 EXPERIMENTS WITH THE USPS DATABASE

In this section, we report the experimental results of the proposed algorithms using a well-known character recognition data set, the United States Postal Services (USPS) database [17], in which there are 9,298 handwriting character measurements divided into 10 classes. The entire USPS data set is divided into two parts [17], a training set with 7,291 measurements and a test set with 2,007 measurements. Each measurement is a vector with 256 dimensions. In our experiments, we utilize the entire USPS database to evaluate the performances of FLDA, HDA, aPAC, WFLDA, FS-FLDA, HFLDA, LFLDA, ODA, MODA, MGMD(0), MGMD(0.25), and M-MGMD(0.25).

We apply different algorithms to the USPS database. As illustrated in Table 6, either MGMD(0.25) or M-MGMD(0.25) gives the best performance for all cases among all algorithms. When the dimensionality of the required feature number is set higher, e.g., 15 or 20, the MGMD variants are clearly better than the other algorithms. The error rates are roughly around 60 percent of ODA and MODA. Note that in the original 256-dimensional feature space, the number of the classes is 10, and therefore, the maximum dimensionality of the extracted features is nine ($10 − 1 = 9$) for FLDA, aPAC, HDA, WFLDA, and FS-FLDA, while in the other listed algorithms, including the proposed MGMD(0.25), more features can be selected. This is a minor advantage of our algorithms comparing with the conventional FLDA. Moreover, the classification error rates are 0.0563 and 0.5795, respectively, for the nearest neighbor rule-based classification and the support vector machine (SVM)-based classification without any dimensionality reduction. SVM performs poorly in this experiment because the machine parameters (the Gaussian kernel parameter and the SVM regularizer) were not tuned precisely. With the same setting of SVM, we briefly tested that if SVM is used instead of the nearest neighbor classification, the performances improve, but MGMD(0) and MGMD(0.25) remain better than LFLDA and HFLDA.

8 CONCLUSION

The conventional FLDA is one of the most important subspace methods in machine learning research and applications; however, for the $c$-class classification task, it has a tendency to merge together nearby classes under projection to a subspace of the feature space if the dimension of the projected subspace is strictly lower than $c − 1$. To reduce this merging, three new criteria for subspace selection are defined based on the geometric
mean of the divergences between the different pairs of classes, namely, 1) maximization of the geometric mean of the divergences, 2) maximization of the geometric mean of the normalized divergences, and 3) MGMD. The third criterion is a combination of the first two criteria. The new subspace selection algorithm has been tested experimentally using synthetic data, UCI Machine Learning Repository, and handwriting digits from the USPS database. The experiments show that the third criterion, namely, the maximization of the geometric mean of all KL divergences, is more effective than FLDA and its representative extensions.

ACKNOWLEDGMENTS

The authors would like to thank the associate editor Professor Mario Figueiredo and the four anonymous reviewers for their constructive comments on the first two versions of this paper. The research was partially supported by the Start-up Grant of the Nanoyang Technological University (under account number M5802010).

REFERENCES

Dacheng Tao received the BEng degree from the University of Science and Technology of China (USTC), the MPhil degree from the Chinese University of Hong Kong (CUHK), and the PhD degree from the University of London. Currently, he is a Nanyang assistant professor in the School of Computer Engineering at the Nanyang Technological University, a visiting professor at Xi Dian University, a guest professor at Wu Han University, and a visiting research fellow at the University of London. His research is mainly on applying statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and visual surveillance. He has published more than 90 scientific papers in journals including nearly 20 journal publications and more than 20 in IEEE transactions. He is an associate editor of the IEEE Transactions on Pattern Recognition, and current Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the founder and current Honorary Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), the Founding Chair (2002-2006) of the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), and a Series Editor of the Springer Book Series on Advanced Information and Knowledge Processing (Ai&KP). He has served as a chair of more than 100 conferences and a program committee member for more than 80 conferences. He has been a reviewer for more than 100 journals and conferences, including 11 IEEE transactions. He is an academic committee member of the China Society of Image and Graphics, a senior member of the IEEE, the chair of IEEE Systems, Man and Cybernetics Society Technical Committee on Cognitive Computing, and a member of several other technical committees of IEEE Systems, Man and Cybernetics Society and IEEE Signal Processing Society Technical Committee on Machine Learning for Signal Processing. He is a Chapters Coordinator of the IEEE Systems, Man and Cybernetics Society.

Xueling Li holds a senior lectureship at Birkbeck College, University of London and visiting/guest professorships at Tianjin University and the University of Science and Technology of China. He has published more than 90 scientific papers in journals including nearly 90 journal publications and more than 20 in IEEE transactions) and with several Best Paper Awards and finalists. He is an author/editor of four books, an associate editor of the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Systems, Man and Cybernetics Part B, and the IEEE Transactions on Systems, Man and Cybernetics Part C. He is also an associate editor (or editorial board member) of 10 other international journals and a guest coeditor of eight special issues. He has served as a chair of more than 20 conferences and a program committee member for more than 80 conferences. He has been a reviewer for more than 100 journals and conferences, including 11 IEEE transactions. He is an academic committee member of the China Society of Image and Graphics, a senior member of the IEEE, the chair of IEEE Systems, Man and Cybernetics Society Technical Committee on Cognitive Computing, and a member of several other technical committees of IEEE Systems, Man and Cybernetics Society and IEEE Signal Processing Society Technical Committee on Machine Learning for Signal Processing. He is a Chapters Coordinator of the IEEE Systems, Man and Cybernetics Society.

Stephen J. Maybank received the BA degree in mathematics from King's College, Cambridge, in 1976 and the PhD degree in computer science from Birkbeck College, University of London, in 1988. He was a research scientist at GEC from 1980 to 1995, first, at MCCS, Frimley, and, then, in 1989, at the GEC Marconi Hirst Research Centre, London. He became a lecturer in the Department of Computer Science, University of Reading, in 1995, and he became a professor in the School of Computer Science and Information Systems, Birkbeck College, University of London, in 2004. His research interests include camera calibration, visual surveillance, tracking, filtering, applications of projective geometry to computer vision and applications of probability, and statistics and information theory to computer vision. He is the author or coauthor of more than 100 scientific publications and one book. He is a fellow of the Institute of Mathematics and its Applications, a fellow of the Royal Statistical Society, and a senior member of the IEEE. For further information, see http://www.dcs.bbk.ac.uk/~sjmaybank.

Xindong Wu is a professor and the chair of the Computer Science Department at the University of Vermont. He received the PhD degree in artificial intelligence from the University of Edinburgh, Britain. His research interests include data mining, knowledge-based systems, and Web information exploration. He has published more than 170 refereed papers in these areas in various journals and conferences, including the IEEE Transactions on Knowledge and Data Engineering (TKDE), the IEEE Transactions on Pattern Analysis and Machine Intelligence, ACM TOIS, DMKD, KAIS, IJCAI, AAAI, IJML, KDD, ICDM, and WWW, as well as 21 books and conference proceedings. He won the IEEE IJCAI-2005 Best Paper Award and the ICDM-2007 Best Theory/Algorithms Paper Runner Up Award. His research has been supported by the US National Science Foundation (NSF), the National Natural Science Foundation of China (NSFC), the US Department of Defense (DOD), as well as industrial companies including U.S. West Advanced Technologies and Em pact Solutions. Dr. Wu was the Editor-in-Chief of the IEEE TKDE, the founder and current Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the founder and current Honorary Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), the Founding Chair (2002-2006) of the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), and a Series Editor of the Springer Book Series on Advanced Information and Knowledge Processing (Ai&KP). He served as Program Committee Chair for ICDM ’03 (the 2003 IEEE International Conference on Data Mining) and as Program Committee Cochair for KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining). Professor Wu is the 2004 ACM SIGKDD Service Award winner, the 2006 IEEE ICDM Outstanding Service Award winner, and a 2005 Chair Professor in the Changjiang (or Yangtze River) Scholars Programme at the Hefei University of Technology appointed by the Ministry of Education of China. He has been an invited/keynote speaker at numerous international conferences including NSF-NGD M ’07, PAKDD-07, IEEE EDOC ’06, IEEE IJCAI ’04, IEEE/WIC/ACM WI ’04/ IAT ’04, SEKE 2002, and PADD-97. He is a senior member of the IEEE.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.