Seventh IEEE International Conference on Data Mining

General Averaged Divergence Analysis

Dacheng Tao [1], Xuelong Li [2], Xindong Wu [3,1], and Stephen J. Maybank [2]
1. Department of Computing, Hong Kong Polytechnic University, Hong Kong
2. Sch. Computer Science & Information Systems, Birkbeck, University of London, London, UK
3. Department of Computer Science, University of Vermont, USA

csdct@comp.polyu.edu.hk; {xuelong, sjmaybank}@dcs.bbk.ac.uk; xwu@cs.uvm.edu

Abstract

Subspace selection is a powerful tool in data mining. An important subspace method is the Fisher–Rao linear discriminant analysis (LDA)*, which has been successfully applied in many fields such as biometrics, bioinformatics, and multimedia retrieval. However, LDA has a critical drawback: the projection to a subspace tends to merge those classes that are close together in the original feature space. If the separated classes are sampled from Gaussian distributions, all with identical covariance matrices, then LDA maximizes the mean value of the Kullback–Leibler* (KL) *divergences between the different classes. We generalize this point of view to obtain a framework for choosing a subspace by 1) generalizing the KL divergence to the Bregman divergence and 2) generalizing the arithmetic mean to a general mean. The framework is named the general averaged divergence analysis* (GADA). *Under this GADA framework, a geometric mean divergence analysis* (GMDA) *method based on the geometric mean is studied. A large number of experiments based on synthetic data show that our method significantly outperforms LDA and several representative LDA extensions.*

1. Introduction

The Fisher-Rao linear discriminant analysis (LDA) has a problem in merging classes that are close together in the original feature space, as shown in Fig. 1. This is referred to as the class separation problem in this paper. As pointed out by McLachlan [14], Loog et al. [11], and Lu et al. [13], this merging of classes significantly reduces the recognition rate. Fig.1 shows an example in which LDA does not select the optimal subspace for pattern classification. To improve its performance, a weighted LDA (WLDA) [8] is introduced. However, the recognition rate of WLDA is sensitive to the selection of the weighting function. Loog et al. [11] developed another weighting method for LDA, namely the approximate pairwise accuracy criterion (aPAC). The advantage of aPAC is that the projection matrix can be obtained by the eigenvalue decomposition.

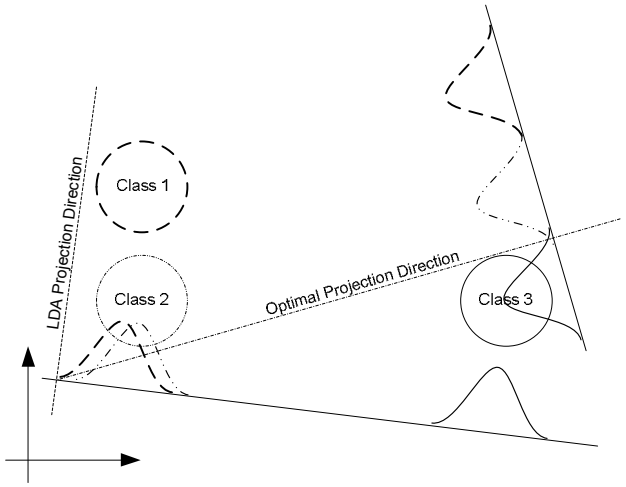

Fig. 1. There are three classes (named 1, 2, and 3) of samples, which are drawn from a Gaussian distribution in each class. LDA finds a projection direction, and merges class 1 and class 2. One of the reasonable projection directions for classification trades the distance between class 1 and class 2 off the distance between the classes 1, 2 and class 3.

In this paper, to further reduce the class separation problem, we first generalize LDA to obtain a general averaged divergence analysis (GADA). If different classes are assumed to be sampled from Gaussian densities with different expected values but identical covariances, then LDA maximizes the mean value of the Kullback–Leibler (KL) divergences [3] between the different pairs of densities. Our generalization of LDA has two aspects: 1) the KL divergence is replaced by the Bregman divergence [2]; and 2) the arithmetic mean is replaced by a general mean function. By choosing different options in 1) and 2) a series of subspace selection algorithms is obtained, with LDA included as a special case.

Under the general averaged divergence analysis, we investigate the effectiveness of the geometric mean and KL divergence based subspace selection for 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. The method is named the geometric mean divergence analysis (GMDA).

1550-4786/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDM.2007.105

2. Linear Discriminant Analysis

The aim of LDA [8] is to find in the feature space a low dimensional subspace in which the different classes of measurements are well separated. The subspace is spanned by a set of vectors, $\mathbf{w}_i$, $1 \le i \le m$, which forms the columns of a matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_m]$. It is assumed that a training set of measurements is available. The training set is divided into c classes. The i^{th} class contains n_i measurements $\mathbf{x}_{i;j}$ $(1 \le j \le n_i)$, and has a mean value of $\boldsymbol{\mu}_i = (1/n_i)\sum_{j=1}^{n_i} \mathbf{x}_{i;j}$. The *between–class scatter* matrix $\mathbf{S}_b$ and the *within–class scatter* matrix $\mathbf{S}_w$ are defined by

$$\begin{cases} \mathbf{S}_b = \dfrac{1}{n}\sum_{i=1}^{c} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \\ \mathbf{S}_w = \dfrac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} (\mathbf{x}_{i;j} - \boldsymbol{\mu}_i)(\mathbf{x}_{i;j} - \boldsymbol{\mu}_i)^T \end{cases} \tag{1}$$

where $n = \sum_{i=1}^{c} n_i$ is the size of the training set and $\boldsymbol{\mu} = (1/n)\sum_{i=1}^{c}\sum_{j=1}^{n_i} \mathbf{x}_{i;j}$ is the mean vector of the total training set. The projection matrix $\mathbf{W}^*$ of LDA is defined by

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \operatorname{tr}\left(\left(\mathbf{W}^T \mathbf{S}_w \mathbf{W}\right)^{-1} \mathbf{W}^T \mathbf{S}_b \mathbf{W} \right). \tag{2}$$

The projection matrix $\mathbf{W}^*$ is computed from the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$, under the assumption that $\mathbf{S}_w$ is invertible. If c equals to 2, LDA reduces to Fisher discriminant analysis [9]; otherwise LDA is known as Rao discriminant analysis [16].

3. General Averaged Divergence Analysis

If the different classes are assumed to be sampled from Gaussian densities with different expected values but identical covariances, then LDA maximizes the mean value of the KL divergences between the different pairs of densities. We propose a framework, the *General Averaged Divergence Analysis*, for choosing a discriminative subspace by: 1) generalizing the distortion measure from the KL divergence to the Bregman divergence, and 2) generalizing the arithmetic mean to a general mean function.

3.1. Bregman Divergence [2]

Definition 1 (Bregman Divergence): Let $U : S \to R$ be a C^1 convex function defined on a closed convex set $S \subseteq R^+$. The first derivative of U is U', which is a monotone function. The inverse function of U' is $\xi = (U')^{-1}$. The sample probability of the i^{th} class is $p_i = p(\mathbf{x} \mid y = i)$. The difference at $\xi(p_j)$ between the function U and the tangent line to U at $(\xi(p_i), U(\xi(p_i)))$ is given by:

$$\begin{aligned} & d(\xi(p_i), \xi(p_j)) \\ &= \left\{ U(\xi(p_j)) - U(\xi(p_i)) \right\} - p_i \left\{ \xi(p_j) - \xi(p_i) \right\}. \end{aligned} \tag{3}$$

Based on (3), the *Bregman divergence* for p_i and p_j is

$$D(p_i \,\|\, p_j) = \int d(\xi(p_i), \xi(p_j))\, d\mu, \tag{4}$$

where $d\mu$ is the Lebesgue measure. The right-hand side of (4) is also called the *U–divergence* [15]. Because U is a convex function, $d(\xi(p), \xi(q))$ is non–negative. Consequently, the Bregman divergence is non–negative. Because $d(\xi(p), \xi(q))$ is in general not symmetric, the Bregman divergence is also not symmetric. Detailed information about the Bregman divergence can be found in [15].

If $U(x) = \exp(x)$, then the Bregman divergence reduces to the usual KL-divergence,

$$\begin{aligned} D(p \,\|\, q) &= \int \left(p_j - p_i - p_i \log \frac{p_j}{p_i} \right) d\mu \\ &= \int p_i \log \frac{p_i}{p_j}\, d\mu = KL(p \,\|\, q). \end{aligned} \tag{5}$$

Further examples can be found in [15].

For Gaussian probability density functions, $p_i \sim N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ is the mean vector of the i^{th} class samples and $\boldsymbol{\Sigma}_i$ is the within–class covariance matrix of the i^{th} class, the KL divergence [3] is

$$\begin{aligned} KL(p_i \,\|\, p_j) &= \int d\mathbf{x} N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \ln \frac{N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{N(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\ &= \ln |\boldsymbol{\Sigma}_j| - \ln |\boldsymbol{\Sigma}_i| + \operatorname{tr}\left(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}_j^{-1}\mathbf{D}_{ij}\right), \end{aligned} \tag{6}$$

where $\mathbf{D}_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \otimes (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ and $|\boldsymbol{\Sigma}| \triangleq \det(\boldsymbol{\Sigma})$.

To simplify the notation we denote the KL divergence between the projected densities $p(\mathbf{W}^T\mathbf{x} \mid y = i)$ and $p(\mathbf{W}^T\mathbf{x} \mid y = j)$ by

$$D_{\mathbf{W}}(p_i \,\|\, p_j) = D\left(p(\mathbf{W}^T\mathbf{x} \mid y = i) \,\|\, p(\mathbf{W}^T\mathbf{x} \mid y = j)\right) \tag{7}$$

3.2. General Averaged Diveregnces Analysis

We replace the arithmetic mean by the following general mean,

$$V_\varphi(\mathbf{W}) = \varphi^{-1}\left[\frac{\sum_{1\le i\ne j\le c} q_i q_j \varphi\left(D_{\mathbf{W}}\left(p_i \,\|\, p_j\right)\right)}{\sum_{1\le m\ne n\le c} q_m q_n}\right] \quad (8)$$

where $\varphi(\cdot)$ is a strict monotonic 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 $q_i = 1/c$); p_i is the conditional distribution of the i^{th} class; $\mathbf{x} \in R^n$ where R^n is the feature space containing the training samples; and $\mathbf{W} \in R^{n\times k}$ ($n \ge k$) 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 $\mathbf{W}^*$ by maximizing the general averaged divergence function $V_\varphi(\mathbf{W})$ over $\mathbf{W}$ for a fixed $\varphi(\cdot)$. The general optimization algorithm for subspace selection based on (8) is given in Table 1. Note that, usually, the concavity of $V_\varphi(\mathbf{W})$ cannot be guaranteed. To reduce the effects of local maxima [1], we choose a number of different initial projection matrices, perform optimizations on them, and then select the best one.

If $V_\varphi(\mathbf{W})$ depends only on the subspace of R^n spanned by the columns of $\mathbf{W}$, then $\mathbf{W}$ can be replaced by $\mathbf{WM}$ where $\mathbf{M}$ is an $m\times m$ matrix in which the columns of $\mathbf{WM}$ are orthogonal.

On setting $\varphi(x) = x$, we use the following arithmetic mean based method for choosing a subspace,

$$\begin{aligned}\mathbf{W}^* &= \arg\max_{\mathbf{W}} \sum_{1\le i\ne j\le c} \frac{q_i q_j D_{\mathbf{W}}\left(p_i \,\|\, p_j\right)}{\sum_{1\le m\ne n\le c} q_m q_n} \\ &= \arg\max_{\mathbf{W}} \sum_{1\le i\ne j\le c} q_i q_j D_{\mathbf{W}}\left(p_i \,\|\, p_j\right).\end{aligned} \quad (9)$$

Observation 1: LDA 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 $\mathbf{W}^*$ in LDA can be obtained by maximizing a particular $V_\varphi(\mathbf{W})$.

Proof.

According to (6) and (7), the KL divergence between the i^{th} class and the j^{th} class in the projected subspace with the assumption of equal covariance matrices ($\mathbf{\Sigma}_i = \mathbf{\Sigma}_j = \mathbf{\Sigma}$) is as follows:

$$D_{\mathbf{W}}\left(p_i \,\|\, p_j\right) = \operatorname{tr}\left(\left(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{D}_{ij}\mathbf{W}\right) + \text{constant} \quad (10)$$

Then, we have

Table 1. General Averaged Divergence Maximization for Subspace Selection

Input: Training samples $\mathbf{x}_{i;j}$, where i denotes the i^{th} class ($1 \le i \le c$) and j is the j^{th} sample in the i^{th} class ($1 \le j \le n_i$), the dimension of selected features $k < n$ (n is the dimension of $\mathbf{x}_{i;j}$), and M is the maximum number of different initial values for the projection matrix.

Output: Optimal linear projection matrix $\mathbf{W}^*$.

1. *for* $m = 1:M$ {
2. Randomly initialize $\mathbf{W}_t^m$ ($t = 1$), i.e., all entries of $\mathbf{W}_1^m$ are random numbers.
3. *while* $\left|V_\varphi\left(\mathbf{W}_t^m\right) - V_\varphi\left(\mathbf{W}_{t-1}^m\right)\right| > \varepsilon$, *do*{
4. Conduct the gradient steepest ascent algorithm to maximize the averaged divergences defined in (8): $\mathbf{W}_t^m \leftarrow \mathbf{W}_{t-1}^m + \kappa\cdot\partial_{\mathbf{W}} V_\varphi\left(\mathbf{W}_{t-1}^m\right)$ where κ is a small value (e.g., 0.001).
5. $t \leftarrow t+1$
6. }//*while* on line 3
7. }//*for* on line 1
8. $\mathbf{W}^* \leftarrow \arg\max_m V_\varphi\left(\mathbf{W}_t^m\right)$.

$$\begin{aligned}\mathbf{W}^* &= \arg\max_{\mathbf{W}} \sum_{1\le i\ne j\le c} q_i q_j D_{\mathbf{W}}\left(p_i \,\|\, p_j\right) \\ &= \arg\max_{\mathbf{W}} \sum_{1\le i\ne j\le c} \left(q_i q_j \operatorname{tr}\left(\left(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{D}_{ij}\mathbf{W}\right)\right) \\ &= \arg\max_{\mathbf{W}} \operatorname{tr}\left(\left(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\right)^{-1}\mathbf{W}^T\left(\sum_{1\le i\ne j\le c} q_i q_j \mathbf{D}_{ij}\right)\mathbf{W}\right) \\ &= \arg\max_{\mathbf{W}} \operatorname{tr}\left(\left(\mathbf{W}^T\mathbf{\Sigma}\mathbf{W}\right)^{-1}\mathbf{W}^T\left(\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} q_i q_j \mathbf{D}_{ij}\right)\mathbf{W}\right).\end{aligned}$$

Because $\mathbf{S}_b = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} q_i q_j \mathbf{D}_{ij}$, as proved by Loog [10], and $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \mathbf{\Sigma}$ (see [8]), we have

$$\begin{aligned}&\arg\max_{\mathbf{W}} \sum_{1\le i\ne j\le c} q_i q_j D_{\mathbf{W}}\left(p_i \,\|\, p_j\right) \\ &= \arg\max_{\mathbf{W}} \operatorname{tr}\left(\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right).\end{aligned} \quad (11)$$

It follows from (11) that a solution of LDA can be obtained by the generalized eigenvalue decomposition.

Example: Decell and Mayekar [5] maximized the arithmetic mean of all symmetric KL divergences between all pairs of classes in the projected subspace. The symmetric KL divergences are given by

$$SKL\left(p_i \parallel p_j\right) = \frac{1}{2} KL\left(p_i \parallel p_j\right) + \frac{1}{2} KL\left(p_j \parallel p_i\right)$$
$$= \mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j\right) + \mathrm{tr}\left(\left(\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\Sigma}_i^{-1}\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T\right). \quad (12)$$

In essence, there is no difference between [5] and maximizing the arithmetic mean of all KL divergences.

De la Torre and Kanade [4] developed the oriented discriminant analysis (ODA) based on the same objective function used in [5], but used iterative majorization to obtain a solution. Iterative majorization speeds up the training stage. Furthermore, they generalized ODA for a multimodal case as the multimodal ODA (MODA) by combining it with Gaussian Mixture Models (GMM) learnt by a normalized cut for the multimodal case. Each class is modelled by a GMM.

3.3. How to Deal with Multimodal Case

Up to this point it has been assumed that the measurement vectors in a given class are sampled from a single Gaussian distribution. This assumption often fails in large real–world data sets, such as those used for multi–view face/gait recognition, natural image classification or texture classification.

To overcome this limitation, each class can be modeled by a GMM. Many methods for obtaining GMMs are described in the literature. Examples include KMeans [6], GMM with expectation–maximization (EM) [6], graph–cut [17], and spectrum clustering. Unfortunately, these methods are not adaptive, in that the number of subclusters must be specified, and some of them (e.g., EM and KMeans) are sensitive to the initial values. In our algorithm we use a recently introduced GMM–EM like algorithm proposed by Figueiredo and Jain [7], which was named the GMM–FJ method. The reasons for choosing GMM-FJ are as follows: it finds the number of subclusters; it is less sensitive to the choice of the initial values of the parameters than EM; and it can avoid the boundary of the parameter space. We assume that the measurements in each class are sampled from a GMM and the projection matrix $\mathbf{W}$ can be obtained by maximizing the general averaged divergences, which measure the averaged distortion between any pair of subclusters in different classes, i.e.,

$$V_\varphi(\mathbf{W}) = \varphi^{-1}\left[\frac{\sum_{1\le i\ne j\le c}\sum_{1\le k\le C_i}\sum_{1\le l\le C_j} q_i^k q_j^l \varphi\left(D_{\mathbf{W}}\left(p_i^k \parallel p_j^l\right)\right)}{\sum_{1\le m\ne n\le c}\sum_{1\le s\le C_i}\sum_{1\le t\le C_j} q_m^s q_n^t}\right], \quad (13)$$

where q_i^k is the prior probability of the k^{th} subcluster of the i^{th} class; p_i^k is the sample probability of the k^{th} subcluster in the i^{th} class; $D_{\mathbf{W}}\left(p_i^k \parallel p_j^l\right)$ is the divergence between the k^{th} subcluster in the i^{th} class and the l^{th} subcluster in the j^{th} class.

3.4. Geometric Mean based Subspace Selection

In LDA and ODA the arithmetic mean of the divergences is used to find a suitable subspace 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, LDA is not optimal for multiclass classification [14] because of the *class separation* problem mentioned in Section I. Therefore, it is useful to investigate other choices of φ in (8).

The log function is a suitable choice for φ 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 (9) the generalized geometric mean of the divergences is obtained. The required subspace $\mathbf{W}^*$ is given by

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \prod_{1\le i\ne j\le c}\left[D_{\mathbf{W}}\left(p_i \parallel p_j\right)\right]^{\frac{q_i q_j}{\sum_{1\le m\ne n\le c} q_m q_n}}. \quad (14)$$

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

$$\prod_{1\le i\ne j\le c}\left[D_{\mathbf{W}}\left(p_i \parallel p_j\right)\right]^{\frac{q_i q_j}{\sum_{1\le m\ne n\le c} q_m q_n}} \le \sum_{1\le i\ne j\le c}\left(\frac{q_i q_j}{\sum_{1\le m\ne n\le c} q_m q_n} D_{\mathbf{W}}\left(p_i \parallel p_j\right)\right).$$

Furthermore, (14) emphasizes the total volume of all divergences. For example, in the special case of $q_i = q_j$ for all i, j,

$$\arg\max_{\mathbf{W}} \prod_{1\le i\ne j\le c}\left[D_{\mathbf{W}}\left(p_i \parallel p_j\right)\right]^{\frac{q_i q_j}{\sum_{1\le m\ne n\le c} q_m q_n}} = \arg\max_{\mathbf{W}} \prod_{1\le i\ne j\le c}\left[D_{\mathbf{W}}\left(p_i \parallel p_j\right)\right]^{q_i q_j} = \arg\max_{\mathbf{W}} \prod_{1\le i\ne j\le c} D_{\mathbf{W}}\left(p_i \parallel p_j\right).$$

3.5. KL Divergence

In this paper, we combine the KL divergence and the geometric mean as an example for practical applications.

Replacing $D_{\mathbf{W}}\left(p_i \parallel p_j\right)$ with the KL divergence and optimizing the logarithm of (14), we have

$$\begin{aligned}\mathbf{W}^* &= \arg\max_{\mathbf{W}} L(\mathbf{W}) \\ &= \arg\max_{\mathbf{W}} \sum_{1\le i\ne j\le c} \log KL_{\mathbf{W}}\left(p_i \,\|\, p_j\right),\end{aligned} \tag{15}$$

and $KL_{\mathbf{W}}\left(p_i \,\|\, p_j\right)$ is the KL divergence between the i^{th} class and the j^{th} class in the projected subspace,

$$\begin{aligned}KL_{\mathbf{W}}\left(p_i \,\|\, p_j\right) = \frac{1}{2}\log\left|\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right| - \log\left|\mathbf{W}^T\boldsymbol{\Sigma}_i\mathbf{W}\right| \\ + \mathrm{tr}\left(\left(\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right)^{-1}\left(\mathbf{W}^T\boldsymbol{\Sigma}_i\mathbf{W}\right)\right) \\ + \mathrm{tr}\left(\left(\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{D}_{ij}\mathbf{W}\right).\end{aligned} \tag{16}$$

To obtain the optimization procedure for the geometric mean and the KL divergence based subspace selection algorithm based on Table 1, we need the first order derivative of $L(\mathbf{W})$,

$$\partial_{\mathbf{W}} L(\mathbf{W}) = \sum_{1\le i\ne j\le c} KL_{\mathbf{W}}^{-1}\left(p_i \,\|\, p_j\right)\partial_{\mathbf{W}} KL_{\mathbf{W}}\left(p_i \,\|\, p_j\right) \tag{17}$$

and

$$\begin{aligned}\partial_{\mathbf{W}} KL_{\mathbf{W}}\left(p_i \,\|\, p_j\right) = \boldsymbol{\Sigma}_j\mathbf{W}\left(\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right)^{-1} - \boldsymbol{\Sigma}_i\mathbf{W}\left(\mathbf{W}^T\boldsymbol{\Sigma}_i\mathbf{W}\right)^{-1} \\ + \left(\boldsymbol{\Sigma}_i + \mathbf{D}_{ij}\right)\mathbf{W}\left(\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right)^{-1} \\ - \boldsymbol{\Sigma}_j\mathbf{W}\left(\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right)^{-1}\mathbf{W}^T\left(\boldsymbol{\Sigma}_i + \mathbf{D}_{ij}\right)\mathbf{W}\left(\mathbf{W}^T\boldsymbol{\Sigma}_j\mathbf{W}\right)^{-1}.\end{aligned} \tag{18}$$

The multimodal extension of (15) can be directly obtained from (13).

4. Comparative Studies Using Synthetic Data

In this section, we denote the proposed method as the geometric mean divergence analysis (GMDA), and compare GMDA with LDA [8], aPAC [11], WLDA (similar to aPAC, but with a different weighting function), HLDA [12], ODA [4], and MODA [4]. We use the weighting function d^{-3} for WLDA.

4.1. Heteroscedastic Problem

To examine the classification ability of these subspace selection methods for the heteroscedastic problem [12], we generate two classes such that each class has 500 samples, drawn from a Gaussian distribution. The two classes have identical mean values but different covariances. As shown in Fig. 2, LDA, aPAC, and WLDA separate class means without taking the differences between covariances into account. In contrast, HLDA, ODA, and GMDA consider both the differences between class means and the differences between class covariances, so they have less training errors, as shown in Fig. 2.

4.2. Multimodal Problem

In many applications it is useful to model the distribution of a class using a GMM, because samples in the class may be drawn from a multimodal distribution. To demonstrate the classification ability of the multimodal extension of GMDA, we generate two classes; each class has two subclusters, and samples in each subcluster are drawn from a Gaussian. Fig. 3 shows the selected subspaces of different methods. In this case LDA, WLDA, and aPAC do not select the suitable subspace for classification. However, the multimodal extensions of ODA and GMDA can find the suitable subspace. Furthermore, although HLDA does not take account of multimodal classes, it 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. For complex cases, e.g., when each class consists of more than 3 subclusters, HLDA will fail to find the optimal subspace for classification.

4.3. Class Separation Problem

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

In Fig. 4., it is demonstrated that GMDA shows a good ability to separate the last two classes of samples. However, LDA, HLDA, and ODA do not give good results. The aPCA and WLDA algorithms are better than LDA but neither of them gives the suitable projection direction. The results obtained from aPCA are better than those obtained from WLDA, because aPAC uses a better weighting strategy than WLDA.

5. Statistical Experiments

In this section, we utilize a synthetic data model, which is a generalization of the data generation model used by Torre and Kanade [4], to evaluate MGMKLD 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. In our experiments, for each of the training/testing sets, the data generator gives 200 samples for each of the five classes (therefore, 1,000 samples in total). Moreover, the samples in each class are obtained from a Gaussian. Each Gaussian density is a linear transformation of a "standard normal distribution". The

linear transformations are defined by $\mathbf{x}_{i;j} = \mathbf{T}_i \mathbf{z}_j + \boldsymbol{\mu}_i + \mathbf{n}_j$, where $\mathbf{x}_{i;j} \in R^{20}$, $\mathbf{T}_i \in R^{20\times 7}$, $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}) \in R^7$, $\mathbf{n} \sim N(\mathbf{0}, 2\mathbf{I}) \in R^{20}$, i denotes the i^{th} class, j denotes the j^{th} sample in this class, and $\boldsymbol{\mu}_i$ is the mean value of the corresponding normal distribution. The $\boldsymbol{\mu}_i$ are assigned as follows: $\boldsymbol{\mu}_1 = (2N(0,1)+4)\mathbf{1}_{20}$, $\boldsymbol{\mu}_2 = \mathbf{0}_{20}$, $\boldsymbol{\mu}_3 = (2N(0,1)-4)[\mathbf{0}_{10}, \mathbf{1}_{10}]^T$, $\boldsymbol{\mu}_4 = (2N(0,1)+4)[\mathbf{1}_{10}, \mathbf{0}_{10}]^T$, and $\boldsymbol{\mu}_5 = (2N(0,1)+4)[\mathbf{1}_5, \mathbf{0}_5, \mathbf{1}_5, \mathbf{0}_5]^T$. The projection matrix $\mathbf{T}_i$ is a random matrix. Each of its elements is sampled from $N(0,5)$. Based on this data generation model, 800 groups (each group with the training and testing samples) of synthetic data are generated from the model.

For comparison, the subspace selection methods are first utilized to select a given number of features. Then the Mahalanobis distance [6] and the nearest neighbour rule are used to examine the accuracy and robustness of GMDA in comparison with LDA and its extensions. The baseline algorithms are LDA, aPAC, WLDA, HLDA, and ODA.

We conducted the above designed experiments 800 times based on randomly generated data sets. The experimental results are reported in Tables 2-5. Tables 2 and 4 show the average error rates of LDA, aPAC, WLDA, HLDA, ODA, and GMDA based on the Mahalanobis distance and the nearest neighbour rule, respectively. Herein arithmetic mean values are computed on different feature dimensions from 1 to 6 (by column). Correspondingly, the standard deviations under each condition, which measure the robustness of the classifiers, are given in Tables 3 and 5. We have twenty feature dimensions for each sample and all the samples are divided into five classes. Therefore, the maximal feature number for LDA, aPAC, and WLDA is 5–1=4; in contrast, HLDA, ODA, and GMDA can extract more features than LDA, aPAC, and WLDA. Based on Tables 2-5, it can be concluded that GMDA outperforms LDA, aPAC, WLDA, HLDA, and ODA, consistently.

We now demonstrate why GMDA is a suitable subspace selection method for classification. Let us first study the relationship between $L(\mathbf{W})$, which is defined in (15), and the training error rate. Experiments are done on a randomly selected data set from 800 data sets generated at the beginning of this section. We set training iterations as 200. In Fig. 5, the left shows that the classification error rates decrease with the increase of the training iterations and the right shows that the objective function values $L(\mathbf{W})$ increase with the increase of the training iterations monotonically. Therefore, the classification error rates decrease with the increase of $L(\mathbf{W})$. This means that maximizing $L(\mathbf{W})$ will be useful to achieve a low classification error rate.

It is also important to investigate how KL divergences between different classes change with the increasing number of training iterations, because it is helpful to deeply understand how and why GMDA reduces the class separation problem. In Fig. 6, we show how KL divergences change in GMDA over the 1st, 2nd, 5th, 10th, 20th, and 200th training iterations, respectively. The small KL divergences, which are less than 2, are marked with rectangles. There are 5 classes, so we can map KL divergences to a 5×5 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 $KL_{ij}(\mathbf{W}_t)$, where t means the t^{th} training iteration. Because the KL divergence is not symmetric, the matrix is not symmetric, i.e., $KL_{ij}(\mathbf{W}_t) \neq KL_{ji}(\mathbf{W}_t)$.

According to Fig. 6, in the 1st training iteration (the top left 5×5 matrix), there are 8 values less than 2. In the 2nd iteration (the top right 5×5 matrix), there are only 4 values less than 2. Compared with the 1st iteration, 6 out of 8 have increased. In the 5th iteration (the middle left 5×5 matrix), there are only 2 values less than 2 and they have increased in comparison with the 2nd iteration. However, these two divergences have decreased to 1.0439 and 1.0366 in the 200th iteration (the bottom right 5×5 matrix) in comparing with the 20th iteration (the bottom left 5×5 matrix) to guarantee the increase of $L(\mathbf{W})$. This is not suitable to separate classes, because the divergences between them are very small.

6. Conclusion

If separate classes are sampled from Gaussian distributions, all with identical covariance matrices, then the Fisher–Rao linear discriminant analysis (LDA) maximizes the mean value of the Kullback–Leibler (KL) divergences between the different classes. We have generalized this point of view to obtain a framework for choosing a subspace by 1) generalizing the KL divergence to the Bregman divergence and 2) generalizing the arithmetic mean to a general mean. The framework is named the general averaged divergence analysis (GADA).

Under this framework, the geometric mean and KL divergence based subspace selection is then studied. LDA has a critical drawback in that the projection to a subspace tends to merge those classes that are close together in the original feature space. A large number of experiments based on synthetic data have shown that our method significantly outperforms LDA and several representative LDA extensions in overcoming this drawback.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[2] L.M. Bregman, "The Relaxation Method to Find the Common Points of Convex Sets and Its Application to the Solution of Problems in Convex Programming," *USSR Compt. Math. and Math. Phys.*, no. 7, pp. 200–217, 1967.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[4] F. De la Torre and T. Kanade "Multimodal Oriented Discriminant Analysis," *Int'l Conf. Machine Learning*, 2005.

[5] H. P. Decell and S. M. Mayekar, "Feature Combinations and the Divergence Criterion," *Computers and Math. With Applications*, vol. 3, pp. 71–76, 1977.

[6] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. John Wiley and Sons Inc. 2001.

[7] M. Figueiredo and A.K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[8] K. Fukunaga, *Introduction to statistical pattern recognition* (Second Edition). Academic Press. 1990.

[9] R. A. Fisher, "The Statistical Utilization of Multiple Measurements," *Ann. Eugenics*, vol. 8 pp. 376-386, 1938.

[10] M. Loog, "Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalizations of the Fisher Criterion," Delft Univ. Press, 1999.

[11] M. Loog, R. P.W. Duin, and R. Haeb–Umbach, "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 23, no. 7, pp. 762–766, July 2001.

[12] M. Loog and R. P.W. Duin, "Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 26, no. 6, pp. 732–739, June 2004.

[13] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using LDA Based Algorithms," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.

[14] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

[15] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information Geometry of U–Boost and Bregman Divergence," *Neural Computation*, vol. 16, no. 7, pp. 1,437–1,481, 2004.

[16] C. R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," *J. Royal Statistical Soc.*, B, vol. 10, pp. 159-203, 1948.

[17] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

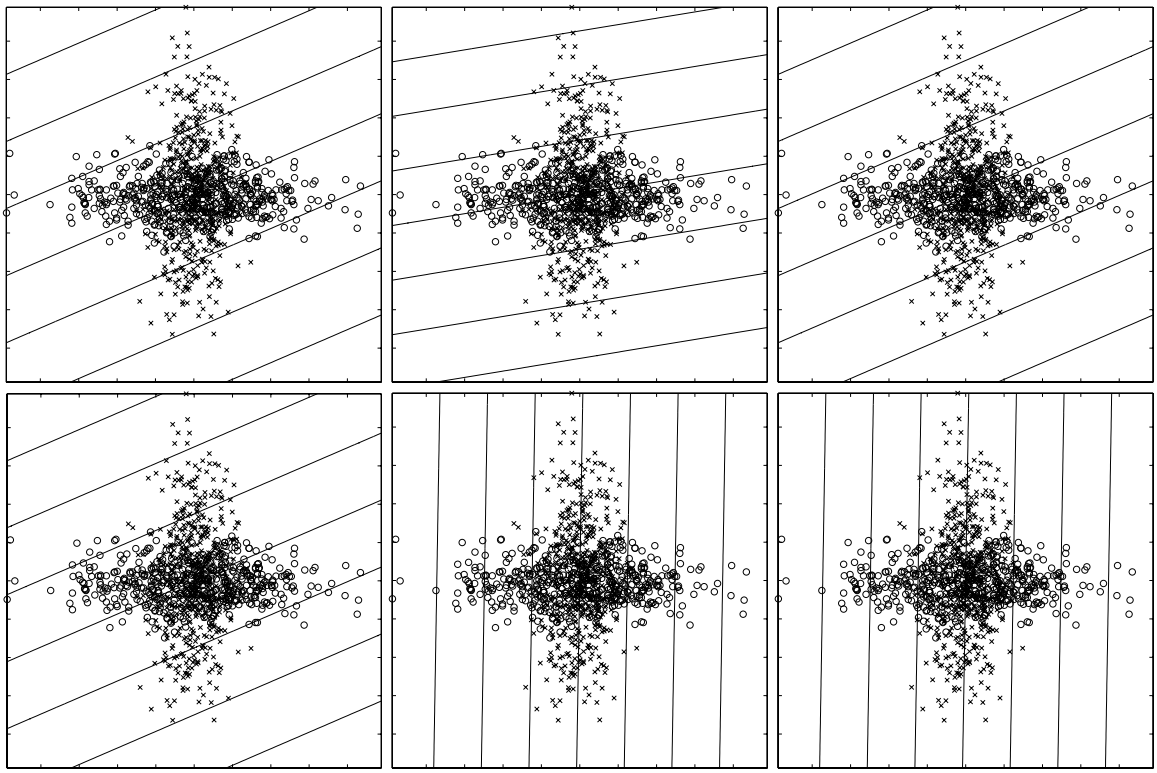

Fig. 2. Heteroscedastic problem: in this figure, from left to right, from top to bottom, there are six subfigures showing the projection directions obtained using LDA, HLDA, aPAC, WLDA, ODA, and GMDA. The training errors of these methods, as measured by Mahanalobis distance, are 0.3410, 0.2880, 0.3410, 0.3410, 0.2390, and 0.2390. ODA and GMDA find the best projection direction for classification.

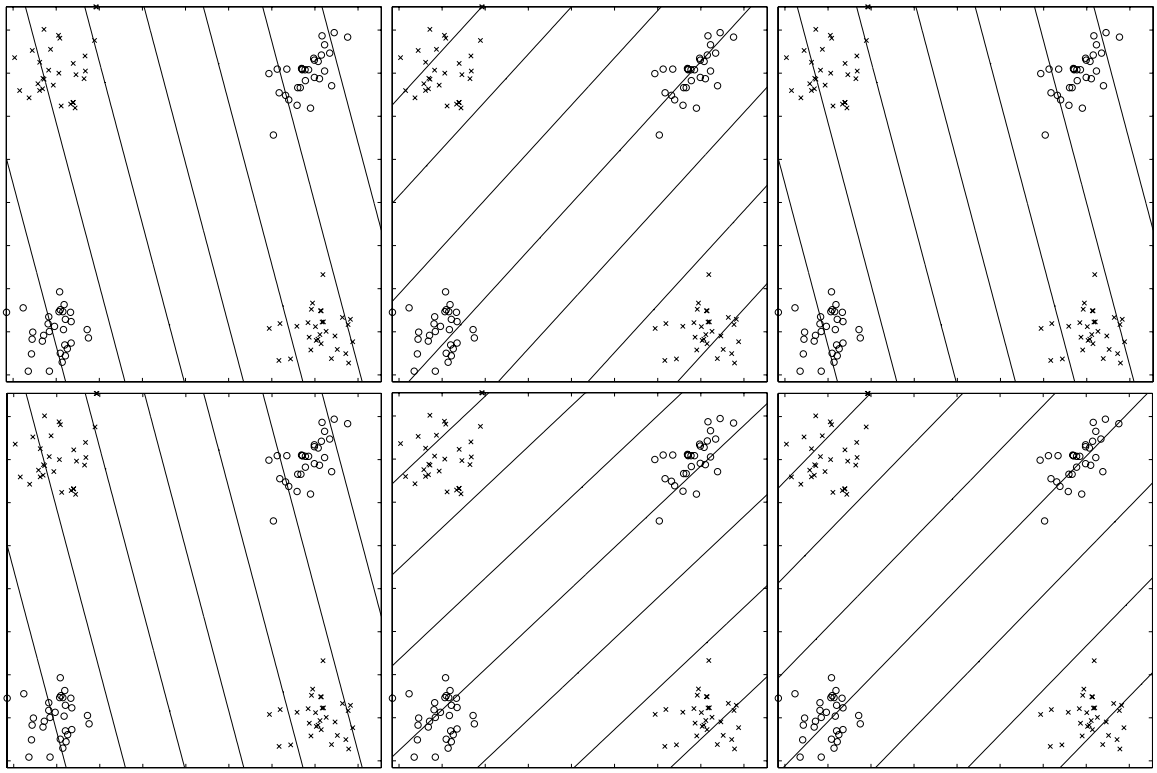

Fig. 3. Multimodal problem: in this figure, from left to right, from top to bottom, there are six subfigures to describe the optimal projection directions by using LDA, HLDA, aPAC, WLDA, MODA, and a multimodal extension of GMDA (M-GMDA). The training errors measured by Mahalanobis distance of these methods are 0.0917, 0.0167, 0.0917, 0.0917, 0.0083, and 0.0083. MODA and M-GMDA find the best projection direction for classification.

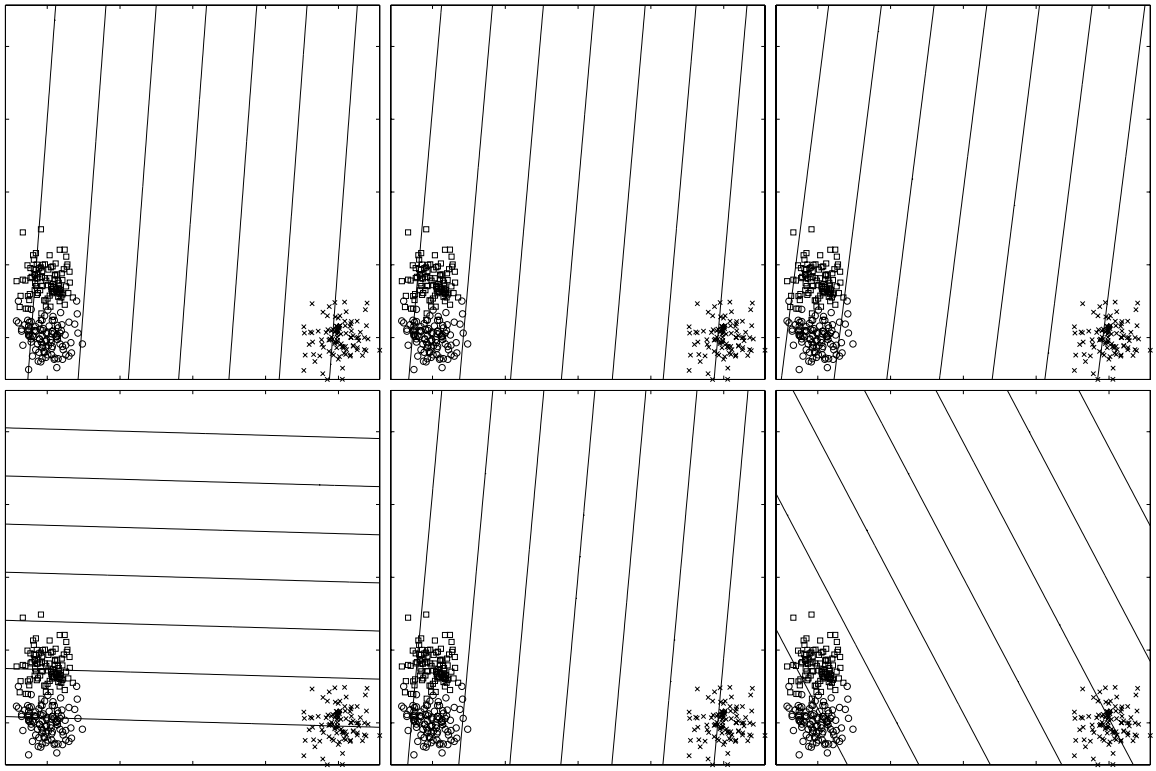

Fig. 4. Large class divergence problem: in this figure, from left to right, from top to bottom, there are nine subfigures to describe the projection directions (indicated by lines in each subfigure) by using LDA, HLDA, aPAC, WLDA, ODA, and GMDA. The training errors measured by Mahalanobis distance of these methods are 0.3100, 0.3100, 0.2900, 0.3033, 0.3100, and 0.1167. GMDA finds the best projection direction for classification.

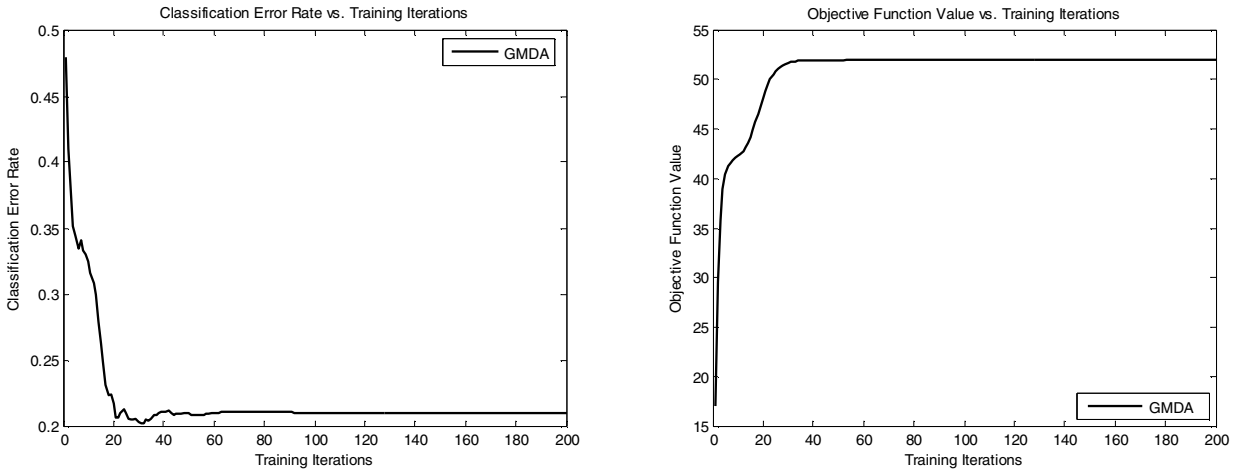

Fig. 5. The consistency of the GMDA objective function $L(\mathbf{W})$ and the classification error rate.

0	3.7880	3.7088	1.1125	1.0817
4.1404	0	1.0427	2.5646	4.5892
4.8868	1.0523	0	2.9725	5.3222
1.1516	2.9784	2.9799	0	1.2638
1.1009	4.6159	4.4595	1.2093	0

$L(\mathbf{W}_1) = 17.0239$

0	20.6130	24.4473	4.7297	1.4808
23.4151	0	1.0228	4.3685	10.5270
24.6297	1.0202	0	4.7669	11.1702
6.9970	5.7420	7.0154	0	2.6469
1.7597	14.9669	17.9361	2.6981	0

$L(\mathbf{W}_2) = 36.1371$

0	31.2804	41.3570	10.7155	3.9663
41.2677	0	1.0294	5.3064	9.0976
41.6566	1.0248	0	5.4603	9.3370
15.6225	5.8584	7.6554	0	2.3828
6.8247	14.7442	19.6081	2.9090	0

$L(\mathbf{W}_5) = 42.3015$

0	32.9683	41.0523	10.3524	10.0145
46.6993	0	1.0459	6.2743	10.6705
46.8454	1.0361	0	6.3327	11.1018
15.8456	7.2997	8.3128	0	8.1064
16.3486	16.4747	23.9465	7.8080	0

$L(\mathbf{W}_{10}) = 47.8674$

0	35.1551	40.8057	11.1944	14.6375
48.2916	0	1.0511	6.4943	18.1236
48.6078	1.0409	0	6.5376	18.5219
15.8031	9.7435	9.4070	0	21.1255
21.3417	17.2344	25.4093	14.4491	0

$L(\mathbf{W}_{20}) = 51.9253$

0	35.3877	40.8188	11.4925	13.9351
48.3992	0	1.0439	6.4784	18.2821
48.7597	1.0366	0	6.5307	18.5347
16.0963	10.0445	9.5688	0	21.7898
20.7283	17.1730	24.4107	14.8540	0

$L(\mathbf{W}_{200}) = 51.9612$

Fig. 6. The KL divergences in GMDA over 1st, 2nd, 5th, 10th, 20th, and 200th training iterations.

Table 2: Average error rates (mean for 800 experiments) of LDA, aPAC, WLDA, HLDA, ODA, and GMDA (Mahalanobis distance).

Basis	1	2	3	4	5	6
LDA	0.2455	0.1199	0.0811	0.0813	—	—
aPAC	0.2626	0.1134	0.0827	0.0813	—	—
WLDA	0.3272	0.1331	0.0833	0.0813	—	—
HLDA	0.2456	0.1216	0.0821	0.0791	0.0764	0.0741
ODA	0.2500	0.1327	0.1037	0.0894	0.0829	0.0796
GMDA	**0.2226**	**0.1099**	**0.0815**	**0.0776**	**0.0751**	**0.0725**

Table 3: Standard deviations of error rates (for 800 experiments) of LDA, aPAC, WLDA, HLDA, ODA, and GMDA (Mahalanobis distance).

Basis	1	2	3	4	5	6
LDA	0.0932	0.0843	0.0792	0.0795	—	—
aPAC	0.1175	0.0987	0.0816	0.0795	—	—
WLDA	0.1305	0.1074	0.0829	0.0795	—	—
HLDA	0.0919	0.0852	0.0799	0.0772	0.0745	0.0725
ODA	0.0923	0.0894	0.0879	0.0810	0.0766	0.0739
GMDA	0.1033	0.0844	0.0788	0.0746	0.0720	0.0695

Table 4: Average error rates (mean for 800 experiments) of LDA, aPAC, WLDA, HLDA, ODA, and GMDA (Nearest neighbor rule).

Basis	1	2	3	4	5	6
LDA	0.2968	0.1552	0.1103	0.1504	—	—
aPAC	0.3206	0.1469	0.1088	0.1324	—	—
WLDA	0.4320	0.1930	0.1126	0.1092	—	—
HLDA	0.2982	0.1561	0.1073	0.1050	0.1043	0.1043
ODA	0.3029	0.1706	0.1370	0.1266	0.1219	0.1206
GMDA	**0.2548**	**0.1397**	**0.1054**	**0.1030**	**0.1024**	**0.1018**

Table 5: Standard deviations of error rates (for 800 experiments) of LDA, aPAC, WLDA, HLDA, ODA, and GMDA (Nearest neighbor rule).

Basis	1	2	3	4	5	6
LDA	0.1002	0.0995	0.0985	0.1077	—	—
aPAC	0.1270	0.1171	0.0979	0.0983	—	—
WLDA	0.1275	0.1239	0.1051	0.0995	—	—
HLDA	0.1001	0.0992	0.0968	0.0942	0.0928	0.0920
ODA	0.1010	0.1037	0.1016	0.0992	0.0965	0.0957
GMDA	0.1094	0.0985	0.0948	0.0919	0.0905	0.0893