Mixture subclass discriminant analysis

Nikolaos Gkalelis, Vasileios Mezaris, Ioannis Kompatsiaris

***Abstract*—In this letter, mixture subclass discriminant analysis (MSDA) that alleviates two shortcomings of subclass discriminant analysis (SDA) is proposed. In particular, it is shown that for data with Gaussian homoscedastic subclass structure a) SDA does not guarantee to provide the discriminant subspace that minimizes the Bayes error, and, b) the sample covariance matrix can not be used as the minimization metric of the discriminant analysis stability criterion (DSC). Based on this analysis MSDA modifies the objective function of SDA and utilizes a novel partitioning procedure to aid discrimination of data with Gaussian homoscedastic subclass structure. Experimental results confirm the improved classification performance of MSDA.**

EDICS Category: IMD-PATT

I. Introduction

Linear discriminant analysis (LDA) is one of the most popular techniques in statistical pattern recognition [1]. However, there are three major drawbacks restricting its use: i) The so-called small sample size problem (SSS) [2], [3], ii) The (common) situation that real-world data have heteroscedastic class distributions, which violates the fundamental homoscedasticity assumption of LDA [3]–[5], and, iii) The instability of the LDA criterion in cases when the metric to be minimized and the metric to be maximized are in "conflict" [6]. Subclass discriminant analysis (SDA) [5] overcomes the above limitations. However, as we show, it presents two shortcomings with respect to its use on data with Gaussian homoscedastic subclass structure, which may be the case even if the class distributions are heteroscedastic. In this work, mixture SDA (MSDA) is proposed to alleviate the latter shortcomings.

This letter is organized as follows. Section II reviews discriminant analysis (DA) methods and in section III we present the proposed method. Experiments are reported in section IV and conclusions are drawn in section V.

II. Discriminant analysis

The problem of discriminant analysis can be generally stated as follows [1]. Given a training set of N labelled samples $\mathcal{U} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ belonging to one of C classes $\{\mathcal{X}_1, \ldots, \mathcal{X}_C\}$, find a singular transformation matrix $\tilde{\mathbf{\Psi}} = [\tilde{\boldsymbol{\psi}}_1, \ldots, \tilde{\boldsymbol{\psi}}_D], \tilde{\boldsymbol{\psi}}_d \in \mathbb{R}^F, D << F$, for mapping the F-dimensional sample $\mathbf{x}$ onto a D-dimensional discriminant subspace spanned by the column vectors of $\tilde{\mathbf{\Psi}}$. The transformation matrix is usually identified by maximizing the objective function $J(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{A} \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{B} \mathbf{\Psi})}$ subject to constraints imposed in the properties of $\mathbf{\Psi}$, where $\mathbf{A}$, $\mathbf{B}$ are metric matrices and $\mathrm{tr}()$ is the trace of a matrix.

The authors are with the Informatics and Telematics Institute / Centre for Research and Technology Hellas (CERTH), Thermi 57001, Greece (email: {*gkalelis, bmezaris, ikom*}*@iti.gr*). N. Gkalelis is also with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, UK.

A. Linear discriminant analysis

LDA seeks directions efficient for class separability. For C Gaussian homoscedastic class distributions, LDA provides the $(C-1)$-dimensional subspace that minimizes the Bayes error [1]. The objective function of LDA is $J_{lda}(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_w \mathbf{\Psi})}$, defining the between-class scatter matrix as $\mathbf{S}_b = \sum_{i=1}^{C} p_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$ and the within-class scatter matrix as $\mathbf{S}_w = \sum_{i=1}^{C} p_i \mathbf{\Sigma}_i$, where, N_i, $\mathbf{\Sigma}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{X}_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T$, $p_i = N_i / N$ and $\boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{X}_i} \mathbf{x}$ are the number of samples, the sample covariance matrix, the prior and the sample mean of i-th class, respectively.

B. Robust linear discriminant analysis

One of the major drawbacks of LDA is the so-called small sample size (SSS) problem, i.e., the situation that the number of training samples N is small compared with their dimensionality F. In this case the class covariance matrix estimates $\mathbf{\Sigma}_i$ and equivalently the within-class scatter matrix $\mathbf{S}_w$ are highly unreliable [3]. To alleviate this, taking into account that $\mathbf{\Sigma}_\mathrm{X} = \mathbf{S}_b + \mathbf{S}_w$ and that $\mathbf{\Sigma}_\mathrm{X}$ is generally a more stable estimate than $\mathbf{S}_w$, a robust LDA criterion has been proposed $J'_{lda}(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{\Sigma}_\mathrm{X} \mathbf{\Psi})}$. Considering that $\mathrm{tr}(\mathbf{\Psi}^T \mathbf{\Sigma}_\mathrm{X} \mathbf{\Psi}) = \mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi}) + \mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_w \mathbf{\Psi})$, $J'_{lda}(\mathbf{\Psi})$ and $J_{lda}(\mathbf{\Psi})$ have the same maximizer according to the following theorem (e.g. see [3]):

Theorem 2.1: Suppose that $\forall \boldsymbol{\psi} \in \mathbb{R}^F, u(\boldsymbol{\psi}) \geq 0, v(\boldsymbol{\psi}) \geq 0, u(\boldsymbol{\psi}) + v(\boldsymbol{\psi}) > 0$. Let $h_1(\boldsymbol{\psi}) = \frac{u(\boldsymbol{\psi})}{v(\boldsymbol{\psi})}$ and $h_2(\boldsymbol{\psi}) = \frac{u(\boldsymbol{\psi})}{u(\boldsymbol{\psi}) + v(\boldsymbol{\psi})}$. Then $h_1(\boldsymbol{\psi})$ has the maximum (including positive infinity) at point $\tilde{\boldsymbol{\psi}}$ iff $h_2(\boldsymbol{\psi})$ has the maximum at the same point.

C. Mixture discriminant analysis

A fundamental assumption of LDA is that class distributions are homoscedastic, which is rarely true in practice. A more realistic strategy is to assume that there exists a subclass homoscedastic partition of the data, $\{\mathcal{X}_{1,1}, \ldots, \mathcal{X}_{C,H_C}\}$, where $\mathcal{X}_{i,j}$ denotes the j-th subclass of the i-th class, H_i is the number of subclasses in the i-th class, and H is the total number of subclasses ($H = \sum_{i=1}^{C} H_i$). Upon this assumption, mixture discriminant analysis (MDA) [4] models classes as mixtures of Gaussian subclasses and the following objective function is utilized $J_{mda}(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_{bs} \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_{ws} \mathbf{\Psi})}$, where the between-subclass

scatter matrix $\mathbf{S}_{bs}$ is defined as

$$\mathbf{S}_{bs} = \sum_{i=1}^{C}\sum_{j=1}^{H_i} p_{i,j}(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu})(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu})^T , \quad (1)$$

the within-subclass scatter matrix $\mathbf{S}_{ws}$ as $\mathbf{S}_{ws} = \sum_{i=1}^{C}\sum_{j=1}^{H_i} p_{i,j}\boldsymbol{\Sigma}_{i,j}$, and $p_{i,j}, \boldsymbol{\mu}_{i,j}, \boldsymbol{\Sigma}_{i,j}$ are the prior, sample mean and sample covariance matrix of $\mathcal{X}_{i,j}$ subclass respectively. Using theorem 2.1 and considering that $\boldsymbol{\Sigma}_{\mathrm{X}} = \mathbf{S}_{bs} + \mathbf{S}_{ws}$, a more stable criterion can be formed

$$J'_{mda}(\boldsymbol{\Psi}) = \frac{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bs}\boldsymbol{\Psi})}{\mathrm{tr}(\boldsymbol{\Psi}^T\boldsymbol{\Sigma}_{\mathrm{X}}\boldsymbol{\Psi})} . \quad (2)$$

D. Discriminant analysis stability criterion

The discriminant analysis stability criterion (DSC) [6], summarized in Theorem 2.2, has been formulated to detect cases where DA does not work.

Theorem 2.2: Let $\boldsymbol{\Psi}_A = [\boldsymbol{\psi}_{A_1}, \ldots, \boldsymbol{\psi}_{A_p}]$ and $\boldsymbol{\Lambda}_A = \mathrm{diag}(\lambda_{A_1}, \ldots, \lambda_{A_p})$ be the eigenvector and eigenvalue matrices of the metric $\mathbf{A}$ to be maximized, i.e., $\mathbf{A}\boldsymbol{\Psi}_A = \boldsymbol{\Psi}_A\boldsymbol{\Lambda}_A$, and, $\boldsymbol{\Psi}_B = [\boldsymbol{\psi}_{B_1}, \ldots, \boldsymbol{\psi}_{B_q}]$ and $\boldsymbol{\Lambda}_B = \mathrm{diag}(\lambda_{B_1}, \ldots, \lambda_{B_q})$ be the eigenvector and eigenvalue matrices of the metric $\mathbf{B}$ to be minimized , i.e., $\mathbf{B}\boldsymbol{\Psi}_B = \boldsymbol{\Psi}_B\boldsymbol{\Lambda}_B$, where p and q are the ranks of $\mathbf{A}$ and $\mathbf{B}$ respectively, $\lambda_{A_1} \geq \cdots \geq \lambda_{A_p}$, $\lambda_{B_1} \geq \cdots \geq \lambda_{B_q}$, and $q \geq p$. Define the discriminant analysis stability criterion

$$\Theta = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{i}(\cos\theta_{i,j})^2 = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{i}(\boldsymbol{\psi}_{A_i}^T\boldsymbol{\psi}_{B_j})^2 \geq 0 , \quad (3)$$

where $r \leq p$, and $\theta_{i,j}$ is the angle between the eigenvectors $\boldsymbol{\psi}_{A_i}$ and $\boldsymbol{\psi}_{B_j}$. Then if $\Theta \neq 0$ the basis vectors given by maximizing the DA criterion will not guarantee to minimize the Bayes error for the given data distribution.

We should note that a large Θ indicates a severe "conflict" between DA metrics. Therefore, the design of algorithms minimizing Θ may have a beneficial effect on DA methods.

E. Subclass discriminant analysis

The between-subclass matrix (1) measures the scatter between all subclasses. Therefore, the overall solution provided by the MDA criterion (2) may be biased towards the directions minimizing the distance between subclasses of the same class. In subclass discriminant analysis (SDA) [5], a more useful objective function is used to emphasize the separation of subclasses belonging to different classes

$$J_{sda}(\boldsymbol{\Psi}) = \frac{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bsb}\boldsymbol{\Psi})}{\mathrm{tr}(\boldsymbol{\Psi}^T\boldsymbol{\Sigma}_{\mathrm{X}}\boldsymbol{\Psi})} , \quad (4)$$

where $\mathbf{S}_{bs}$ in (2) has been replaced by $\mathbf{S}_{bsb}$, measuring only the scatter between subclasses of different classes

$$\mathbf{S}_{bsb} = \sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{i,j}p_{k,l}(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{k,l})(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{k,l})^T . \quad (5)$$

The optimization of (4) is done using an iterative procedure, where at the r-th iteration a nearest neighbor based (NN-based) clustering algorithm is used to provide a new subclass partition of the data $\{\mathcal{X}_{1,1}^{(r)}, \ldots, \mathcal{X}_{C,H_C}^{(r)}\}$. At each iteration the number of the subclasses referring to the i-th class is increased by one, $H_i^{(r)} = H_i^{(r-1)} + 1$, and, therefore, the total number of subclasses is increased by C, i.e., $H^{(r)} = \sum_{i=1}^{C} H_i^{(r)} = H^{(r-1)} + C$. Each subclass partition is evaluated using either a leave-one-out-cross-validation based (LOOCV-based) criterion, or the DSC criterion (3) setting $\mathbf{A} = \mathbf{S}_{bsb}$ and $\mathbf{B} = \boldsymbol{\Sigma}_{\mathrm{X}}$, and the best subclass partition is chosen as the one that optimizes the respective criterion.

III. Mixture Subclass Discriminant Analysis

For data with a Gaussian homoscedastic subclass structure and under stable situations (according to theorem 2.2) we propose the following mixture-based subclass objective function

$$J_{msda}(\boldsymbol{\Psi}) = \frac{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bsb}\boldsymbol{\Psi})}{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{ws}\boldsymbol{\Psi})} . \quad (6)$$

This provides a discriminant subspace that minimizes the Bayes error, as can be easily proven by treating the Gaussian homoscedastic subclasses as the main classes and constructing linear likelihood classification rules [1].

A. Shortcomings of SDA

Here we show that the SDA objective function (4) proposed in [5] is not equivalent with (6), i.e., SDA does not necessarily minimize the Bayes error under the conditions identified at the beginning of section III. The between-class scatter matrix can be rewritten as $\mathbf{S}_b = \sum_{i=1}^{C-1}\sum_{k=i+1}^{C} p_i p_k(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)^T$, (e.g., see [7]–[9]). Similarly to this, we can express the between-subclass scatter matrix (1) as

$$\mathbf{S}_{bs} = \sum_{i=1}^{C}\sum_{j=1}^{H_i}\sum_{k=i}^{C}\sum_{l=\Gamma}^{H_k} p_{i,j}p_{k,l}(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{k,l})(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{k,l})^T ,$$

where, $\Gamma = \begin{cases} j+1 & \text{if } k = i; \\ 1 & \text{if } k > i. \end{cases}$. We can rewrite the above as

$$\begin{aligned}\mathbf{S}_{bs} &= \sum_{i=1}^{C}\sum_{j=1}^{H_i}\sum_{l=j+1}^{H_i} p_{i,j}p_{i,l}(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{i,l})(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{i,l})^T \\ &\quad + \sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{i,j}p_{k,l}(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{k,l})(\boldsymbol{\mu}_{i,j}-\boldsymbol{\mu}_{k,l})^T \\ &= \mathbf{S}_{bsw} + \mathbf{S}_{bsb} , \end{aligned} \quad (7)$$

where $\mathbf{S}_{bsb}$ (5) is the scatter of means of subclasses between different classes (inter-subclass scatter of means) and $\mathbf{S}_{bsw}$ is the scatter of means of subclasses within the same classes (intra-subclass scatter of means). Therefore, the sample covariance matrix can be expressed as

$$\boldsymbol{\Sigma}_{\mathrm{X}} = \mathbf{S}_{bs} + \mathbf{S}_{ws} = \mathbf{S}_{bsb} + \mathbf{S}_{bsw} + \mathbf{S}_{ws} . \quad (8)$$

Replacing this expression in the optimization criterion of SDA (4), and assuming that $\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{ws}\boldsymbol{\Psi}) > 0$ we get

$$\begin{aligned} J_{sda}(\boldsymbol{\Psi}) &= \frac{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bsb}\boldsymbol{\Psi})}{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bsb}\boldsymbol{\Psi})+\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bsw}\boldsymbol{\Psi})+\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{ws}\boldsymbol{\Psi})} \\ &= \frac{J_{msda}(\boldsymbol{\Psi})}{J_{msda}(\boldsymbol{\Psi})+h(\boldsymbol{\Psi})+1} , \end{aligned} \quad (9)$$

where $h(\boldsymbol{\Psi}) = \frac{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{bsw}\boldsymbol{\Psi})}{\mathrm{tr}(\boldsymbol{\Psi}^T\mathbf{S}_{ws}\boldsymbol{\Psi})}$ is a function that varies independently with $J_{msda}(\boldsymbol{\Psi})$. Due to this fact, theorem 2.1 can not

be used to show that $J_{sda}(\mathbf{\Psi})$ and $J_{msda}(\mathbf{\Psi})$ have the same maximum. Furthermore, according to (8) minimizing $\mathbf{\Sigma}_{\mathrm{X}}$ has the desired effect of minimizing $\mathbf{S}_{ws}$, as well as the undesired effect of minimizing $\mathbf{S}_{bsb}$. This conclusion is important as it reveals a second drawback of SDA, i.e., $\mathbf{\Sigma}_{\mathrm{X}}$ can not be used as the minimization metric $\mathbf{B}$ in theorem 2.2.

B. Mixture subclass discriminant analysis

Based on the above analysis and according to theorem 2.1 we further propose the following robust mixture SDA (MSDA) objective function

$$J'_{msda}(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_{bsb} \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \breve{\mathbf{\Sigma}}_{\mathrm{X}} \mathbf{\Psi})} , \tag{10}$$

where $\breve{\mathbf{\Sigma}}_{\mathrm{X}}$ is defined as $\breve{\mathbf{\Sigma}}_{\mathrm{X}} \equiv \mathbf{S}_{bsb} + \mathbf{S}_{ws}$. The optimization of (6) or (10) is performed using an iterative procedure similar to SDA, and each subclass partition is evaluated using the LOOCV-based criterion or the DSC criterion (3) setting $\mathbf{A} = \mathbf{S}_{bsb}$ and $\mathbf{B} = \mathbf{S}_{ws}$.

Moreover, in contrary to SDA, at each iteration a specific class is selected and only the number of subclasses of this class is increased by one, i.e., only one additional subclass is introduced at each iteration ($H^{(r)} = H^{(r-1)}+1$). The selection of the class to be re-partitioned is done using a nongaussianity criterion based on the skewness and kurtosis. Estimates of the standardized skewness and kurtosis of the $\mathcal{X}_{i,j}$ subclass along the k-th dimension can be computed as follows,

$$\gamma^{(n)}_{i,j,k} = \frac{\frac{1}{N_{i,j}} \sum_{x_k \in \mathcal{X}_{i,j}} (x_k - \mu_{i,j,k})^n}{\sigma^n_{i,j,k}} , \tag{11}$$

setting $n = 3$ and $n = 4$ respectively, where x_k is the k-th element of sample $\mathbf{x}$, and $\mu_{i,j,k}, \sigma_{i,j,k}$ are the sample mean and standard deviation of $\mathcal{X}_{i,j}$ subclass along the k-th dimension. Then, an estimate of the skewness $\gamma^{(3)}_{i,j}$ and kurtosis $\gamma^{(4)}_{i,j}$ of the $\mathcal{X}_{i,j}$ subclass can be obtained by averaging along all dimensions

$$\gamma^{(3)}_{i,j} = \frac{1}{F} \sum_{k=1}^{F} |\gamma^{(3)}_{i,j,k}| , \quad \gamma^{(4)}_{i,j} = \frac{1}{F} \sum_{k=1}^{F} |\gamma^{(4)}_{i,j,k} - 3| , \tag{12}$$

where $|\beta|$ denotes absolute value of β. Skewness and kurtosis measure the deviation of a probability density from the Gaussian density in terms of asymmetry and peakedness respectively, and their estimates in (12) will be zero if $\mathcal{X}_{i,j}$ subclass has a Gaussian distribution, and deviate from zero the more the subclass distribution deviates from a Gaussian distribution. Thus, a measure of nongaussianity of $\mathcal{X}_{i,j}$ subclass can be defined as $\Phi_{i,j} = \gamma^{(3)}_{i,j} + \gamma^{(4)}_{i,j}$ and similarly a measure of nongaussianity of $\mathcal{X}_i$ class with respect to its subclasses can be defined as

$$\Phi_i = \frac{1}{H_i} \sum_{j=1}^{H_i} \Phi_{i,j} = \frac{1}{H_i} \sum_{j=1}^{H_i} (\gamma^{(3)}_{i,j} + \gamma^{(4)}_{i,j}) . \tag{13}$$

Therefore, at the r-th iteration $\Phi^{(r)}_i$ is computed for each class, and the class $\mathcal{X}_y$ to be re-partitioned is selected according to the following rule

$$y = \underset{i=1,\ldots,C}{\mathrm{argmax}} (\Phi^{(r)}_i) . \tag{14}$$

IV. Experiments

A. Artificial dataset

An artificial dataset with Gaussian homoscedastic subclass structure is used to justify the theoretical analysis of the proposed method (Figure 1). The dataset consists of two main classes $\mathcal{X}_1, \mathcal{X}_2$, and three subclasses $\mathcal{X}_{1,1}, \mathcal{X}_{1,2}, \mathcal{X}_{2,1}$, i.e., the first class consists of two Gaussian subclasses, whereas the second class is a single Gaussian. The parameters of the Gaussian distributions are: $\boldsymbol{\mu}_{1,1} = [2 \quad 1]^T$, $\boldsymbol{\mu}_{1,2} = [6 \quad -3]^T$, $\boldsymbol{\mu}_{2,1} = [5 \quad 0]^T$, $\mathbf{\Sigma}_{1,1} = \mathbf{\Sigma}_{1,2} = \mathbf{\Sigma}_{2,1} = [1 \quad 0.7 \; ; \; 0.7 \quad 1]$. The true subclass labelling of the data is directly used, and LDA, SDA, and MSDA are applied to derive the one dimensional projection that maximizes their objective function. We should note that, in this example, recovering another 2D subspace would be useless, as this would result in the same computational complexity and classification error as the original space. The derived projection directions, $\boldsymbol{\psi}_{LDA}$, $\boldsymbol{\psi}_{SDA}$ and $\boldsymbol{\psi}_{MSDA}$, are shown in Figure 1. As expected, LDA does not recover the optimal projection as the first class consists of two separate Gaussian distributions. Although the data have a clear subclass homoscedastic structure, SDA also fails to provide the optimal projection. On the other hand, using (10) MSDA correctly identifies the projection that minimizes the Bayes error.

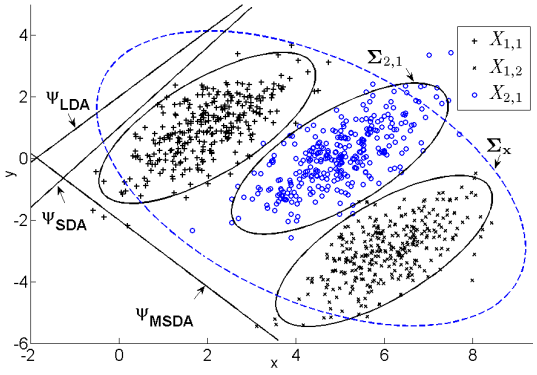

Fig. 1. *Artificial dataset with Gaussian homoscedastic subclass distributions.*

In a second experiment with the same data, we evaluate two different formulations of the DSC criterion, denoted Θ_{X} and Θ_{ws}. In the former, the minimization metric $\mathbf{B}$ is set to $\mathbf{B} = \mathbf{\Sigma}_{\mathrm{X}}$ as in [5], while in the latter the within-subclass scatter matrix is used ($\mathbf{B} = \mathbf{S}_{ws}$). The NN-based algorithm [5] is applied to partition each class to L subclasses, where $L \in [1, 5]$. Table I shows that Θ_{X} is minimized for $L = 1$, i.e., this criterion suggests to preserve the original class structure, which is clearly wrong. In contrary, Θ_{ws} is minimized for $L = 2$, which provides a better subclass division. This shows that formulating the DSC criterion using $\mathbf{B} = \mathbf{S}_{ws}$ (MSDA algorithm) instead of $\mathbf{B} = \mathbf{\Sigma}_{\mathrm{X}}$ (SDA algorithm) is more effective for selecting the suitable subclass division.

B. Real-life datasets

A subset of the MediaMill Challenge dataset is used to evaluate the proposed algorithm for event recognition [10]. This subset consists of 492 shots in total belonging to one of five

TABLE II
Classification rates of various methods.

Dataset \ Method	PCA	LDA	aPAC	FS-LDA	SDA$_S$	MSDA$_S$	MGMD	SDA$_L$	MSDA$_L$
MediaMill	68% (101)	64.9% (4)	63.5% (4)	63.9% (4)	67.2% (44)	68% (46)	69.5% (27)	69.3% (35)	**71.4%** (23)
Sheffield	94.9% (236)	95.5% (19)	96.6% (19)	96.8% (19)	96% (19)	97% (21)	–	97.2% (31)	**97.6%** (24)

TABLE I
Comparison of the two different formulations of the stability criterion.

L	1	2	3	4	5
Θ_{X}	$\mathbf{3 \cdot 10^{-3}}$	0.99	0.99	0.99	0.99
Θ_{ws}	$2 \cdot 10^{-3}$	$\mathbf{10^{-4}}$	0.99	$7 \cdot 10^{-4}$	0.96

different sport events, namely, baseball, basketball, football, golf and soccer. Each shot is represented by a 101-dimensional model vector, where the κ-th component of this vector is in the range $[0, 1]$, expressing the degree of confidence that the κ-th concept (out of 101 concepts) is present in the shot [10]. Likewise, the multiview Sheffield (previously UMIST) face database is used to perform experiments for face recognition [11]. This database offers 564 gray-scale cropped facial images of 20 individuals. The facial images are scaled to size 32×32 pixels using bicubic interpolation and then scanned column-wise to provide 1024-dimensional feature vectors.

These two datasets are used for comparing the proposed method (MSDA$_S$ or MSDA$_L$) with SDA$_S$, SDA$_L$, principal component analysis (PCA), LDA [2], fractional step LDA (FS-LDA) [8], approximate pairwise accuracy criterion (aPAC) LDA [7], and the method proposed in [9] for the maximization of the geometric mean of divergences, denoted as MGMD(α), where $\alpha \in [0, 1]$ is a combination factor. Subscripts S and L in SDA and MSDA denote that the optimization of the algorithm is performed using the stability criterion or the LOOCV procedure respectively. The latter, in the case of MSDA, takes advantage of the objective function (10) and the subclass partitioning of section III-B, but replaces the criterion of theorem 2.2 with a well-performing but more computationally expensive cross-validation procedure (e.g., see [5]). For FS-LDA we experimented with different weighting functions $d^{-t}, t = 4, 6, 8, 10, 12$, where d is the Euclidean distance between class means, and similarly, for MGMD(α) we used different combination factors $\alpha = 0.25, 0.45, 0.65, 0.85$. The evaluation of the methods is done by applying a 30-fold cross-validation (CV) procedure, where at each validation cycle η% of the samples from each class are removed to form the test set, while the remaining $(100 - \eta)$% of the samples are used to form the training set. The values of η were set to $\eta = 20$ and $\eta = 60$ for event recognition (MediaMill dataset) and face recognition (Sheffield dataset) respectively. Test samples are classified using the NN rule. For aPAC, FS-LDA, MGMD(α), SDA$_L$ and MSDA$_L$, at each CV cycle the maximum correct classification rate (CCR) for the different examined dimensionalities is retained. Consequently, the overall performance of a method is measured using the average CCR (ACCR) along all CV cycles. The ACCR of each algorithm along with the average dimensionality of the discriminant subspace, computed by averaging the dimensionality of the projected samples D along all CV cycles, are presented in Table II. In this table we separate MSDA$_L$, SDA$_L$ and MGMD from the other methods to denote that they require considerably more processing time during optimization. We should note that we do not report results for MGMD on the Sheffield dataset, because in this dataset the classes consist of only a few high dimensional samples and MGMD, which requires the inversion of class covariance matrices, is severely affected by the SSS problem [12]. From these results it is concluded that MSDA$_S$ outperforms SDA$_S$ and different LDA variants, and that MSDA$_L$ provides the highest classification performance.

V. Conclusions

In this letter, two shortcomings of SDA have been presented and upon their analysis MSDA has been proposed. Experimental results showed the effectiveness of the proposed method.

Acknowledgment

This work was supported by the European Commission under contract FP7-248984 GLOCAL.

References

[1] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[3] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, Jan. 2005.

[4] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 155–176, Jul. 1996.

[5] M. Zhu and A. Martinez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, August 2006.

[6] A. M. Martinez and M. Zhu, "Where are linear feature extraction methods applicable?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1934–1944, Dec. 2005.

[7] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.

[8] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 623–627, Jun. 2000.

[9] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[10] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Automatic event-based indexing of multimedia content using a joint content-event model," in *ACM Multimedia 2010, EiMM Workshop*, Firenze, Italy, Oct. 2010.

[11] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications, Computer and Systems Sciences*, H. Wechsler and P. J. Phillips et al., Eds. NATO ASI Series F, 1998, vol. 163, pp. 446–456.

[12] A. K. Qin, P. N. Suganthan, and M. Loog, "Uncorrelated heteroscedastic LDA based on the weighted pairwise Chernoff criterion," *Pattern Recognition*, vol. 38, no. 4, pp. 613–616, Apr. 2005.