On the optimal class representation in Linear Discriminant Analysis

Alexandros Iosifidis, Anastasios Tefas, Member, IEEE, and Ioannis Pitas, Fellow, IEEE

***Abstract*—Linear Discriminant Analysis (LDA) is a widely used technique for supervised feature extraction and dimensionality reduction. LDA determines an optimal discriminant space for linear data projection based on certain assumptions, e.g. on using normal distributions for each class and employing class representation by the mean class vectors. However, there might be other vectors that can represent each class, in order to increase class discrimination. In this paper, we propose an optimization scheme aiming at the optimal class representation, in terms of Fisher ratio maximization, for LDA-based data projection. Compared to the standard LDA approach, the proposed optimization scheme increases class discrimination in the reduced dimensionality space and achieves higher classification rates in publicly available data sets.**

***Index Terms*—Linear Discriminant Analysis, Class representation, Subspace learning, Data projection.**

I. Introduction

Linear Discriminant Analysis (LDA) is a well-known algorithm for feature extraction and dimensionality reduction, aiming at finding an optimal reduced dimensionality space for data projection, in which the classes are better discriminated. The adopted criterion is the ratio of the between-class scatter to the within-class scatter in the projection space, which is, usually, referred to as Fisher ratio. By maximizing this criterion, maximal class discrimination is achieved. The main idea in standard LDA is that in the reduced dimensionality space the samples belonging to different classes should be as far from one another and that the within class dispersion from their mean should be as small as possible. LDA optimality is based on the assumptions that: a) all classes follow normal distributions having the same covariance structure and b) each class is represented by the mean class vector. Although relying on rather strong assumptions which do not hold in many applications, it has proven very powerful and it has been widely used in many applications, such as facial expression recognition [1], human action recognition [2] and person identification [3].

This work is motivated by the observation that other class representation in the input space than the class mean, result in different scatter matrices and, finally, in a different projection space that could provide superior class discrimination. Consider the example illustrated in Figure 1. Figure 1a illustrates 2D data resulted by applying Principal Component Analysis (PCA) on 10-dimensional (10D) data forming three classes following normal distributions. Let $\mathbf{p}_i$, $i = 1, 2, 3$ and $\mathbf{q}_i$, $i = 1, 2, 3$ represent random vectors that can be used as class representatives instead of the sample mean vectors $\boldsymbol{\mu}_i$, $i = 1, 2, 3$ in the LDA optimization procedure. Figure 1b illustrates the projection space obtained by applying LDA on the 10D data using the mean class vectors $\boldsymbol{\mu}_i$ for class representation, while Figures 1c and 1d illustrate the projection spaces obtained by applying LDA using vectors $\mathbf{p}_i$ and $\mathbf{q}_i$ for class representation, respectively.

As can be seen, the three class data projections obtained by using three different class representations in the input space are quite different. That is, various discriminant LDA spaces can be formed, depending on the class representation choice in the input space, which offer different discrimination ability and recognition performance. Since LDA optimality is based on Fisher ratio maximization, one may think that the optimal class representation can be obtained by an optimization procedure maximizing Fisher ratio. Therefore, we are interested in finding the class representative vectors in the reduced dimensionality space that are as far from one another and as close to the respective class samples as possible. As we shall show in next sections, this can be done by performing an iterative optimization procedure with respect to both the projection matrix and the chosen class representation.

The paper is structured as follows. Section II presents an overview of the standard LDA algorithm. In Section III, we present the proposed optimization scheme. Experimental results assessing its performance are illustrated in Section V and conclusions are drawn in Section VII.

II. Standard LDA

Given a set of D-dimensional data belonging to C classes, $\mathbf{x}_{ij} \in \mathbb{R}^D$, $i = 1, ..., C$, $j = 1, ..., N_i$, $\sum_{i=1}^{C} N_i = N$, and their class labels $l_{ij} = i$, standard LDA aims to find a projection matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$, such that $\mathbf{y}_{ij} = \mathbf{W}^T \mathbf{x}_{ij} \in \mathbb{R}^d$ is the image of $\mathbf{x}_{ij}$ in a d-dimensional feature space, where classes achieve maximal compactness and discrimination.

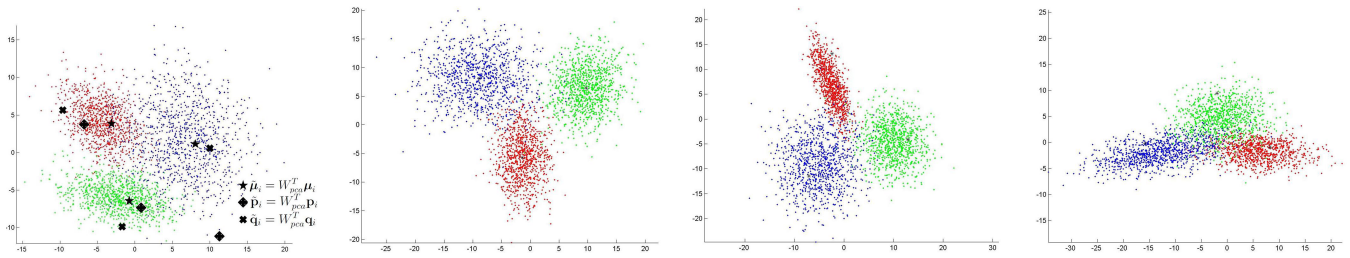
Let us assume that the data are centered to $\mathbf{0}$, i.e., $\frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N_i} \mathbf{x}_{ij} = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^D$ is a vector of zeros[1]. The optimal projection matrix $\mathbf{W}^*$ is obtained by maximizing the ratio of the between-class scatter matrix $\mathbf{S}_b$ to that of the within-class scatter matrix $\mathbf{S}_w$ in the projection space. $\mathbf{S}_b$, $\mathbf{S}_w$ are defined as follows:

$$\mathbf{S}_b = \sum_{i=1}^{C} N_i \, \mathbf{m}_i \mathbf{m}_i^T, \tag{1}$$

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (\mathbf{y}_{ij} - \mathbf{m}_i)(\mathbf{y}_{ij} - \mathbf{m}_i)^T, \tag{2}$$

A. Iosifidis, A. Tefas and I. Pitas are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece. e-mail: {aiosif,tefas,pitas}@aiia.csd.auth.gr.

[1]This can always be done by using $\tilde{\mathbf{x}}_{ij} = \mathbf{x}_{ij} - \boldsymbol{\mu}$, $i = 1, \ldots, C$, $j = 1, \ldots, N_i$, where $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$.



(a) (b) (c) (d)

Fig. 1. *a) PCA space for data forming three classes following normal distributions and LDA spaces obtained by representing classes with b) the mean class vectors $\boldsymbol{\mu}_i$, c) vectors $\mathbf{p}_i$ and d) vectors $\mathbf{q}_i$.*

where $\mathbf{m}_i \in \mathbb{R}^d$ is the mean vector of class i in the projection space, i.e., $\mathbf{m}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{y}_{ij}$. Since $\mathbf{y}_{ij}$ are not a priori known, it is convenient to express $\mathbf{S}_b$, $\mathbf{S}_w$ by using $\mathbf{x}_{ij}$. It can be shown that $\mathbf{S}_b = \mathbf{W}^T \ddot{\mathbf{S}}_b \mathbf{W}$ and $\mathbf{S}_w = \mathbf{W}^T \ddot{\mathbf{S}}_b \mathbf{W}$, where:

$$\ddot{\mathbf{S}}_b = \sum_{i=1}^{C} N_i \, \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T, \tag{3}$$

$$\ddot{\mathbf{S}}_w = \sum_{i=1}^{C}\sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \boldsymbol{\mu}_i)(\mathbf{x}_{ij} - \boldsymbol{\mu}_i)^T. \tag{4}$$

$\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is the mean vector of class i in the input space.

By using the traces of $\mathbf{S}_b$, $\mathbf{S}_w$ to measure the between-class and within-class scatter, the optimal projection matrix $\mathbf{W}^*$ can be obtained by maximizing Fisher ratio:

$$\mathbf{W}^* = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}}{argmax}\ \mathcal{J}(\mathbf{W}), \tag{5}$$

$$\mathcal{J}(\mathbf{W}) = \frac{Tr(\mathbf{W}^T \ddot{\mathbf{S}}_b \mathbf{W})}{Tr(\mathbf{W}^T \ddot{\mathbf{S}}_w \mathbf{W})}. \tag{6}$$

$Tr(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. The constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ is conventionally added in order to obtain a set of orthogonal and normalized projection vectors. Another version of LDA [1] exploits the equality $\ddot{\mathbf{S}}_T = \ddot{\mathbf{S}}_b + \ddot{\mathbf{S}}_w$ in order to maximize an equivalent to (6) criterion, i.e., $\mathcal{J}(\mathbf{W}) = \frac{Tr(\mathbf{W}^T \ddot{\mathbf{S}}_b \mathbf{W})}{Tr(\mathbf{W}^T \ddot{\mathbf{S}}_T \mathbf{W})}$. $\ddot{\mathbf{S}}_T = \sum_{i=1}^{C}\sum_{j=1}^{N_i} \mathbf{x}_{ij}\mathbf{x}_{ij}^T$ is the total data scatter matrix.

That is, the optimal projection matrix $\mathbf{W}^*$ is obtained by solving the so-called *trace ratio* optimization problem, leading to the so-called *Trace Ratio LDA*, which has been used by a number of dimensionality reduction algorithms [4], [5], [6], [7], [8], [9]. However, the trace ratio problem does not have a direct closed-form globally optimal solution. Therefore, it is conventionally approximated by solving the *ratio trace* problem, i.e., $\tilde{\mathcal{J}}(\mathbf{W}) = Tr\left[(\mathbf{W}^T \ddot{\mathbf{S}}_b \mathbf{W})^{-1}(\mathbf{W}^T \ddot{\mathbf{S}}_w \mathbf{W})\right]$, which is equivalent to the optimization problem $\mathbf{S}_w\mathbf{v} = \lambda \mathbf{S}_b\mathbf{v}$, $\lambda \neq 0$, and can be solved by performing eigenanalysis to the matrix $\mathbf{S} = \mathbf{S}_b^{-1}\mathbf{S}_w$ in the case where $\mathbf{S}_b$ is invertible, or $\mathbf{S} = \mathbf{S}_w^{-1}\mathbf{S}_b$ in the case where $\mathbf{S}_w$ is invertible.

Although the trace ratio problem does not have a closed form solution, it has been shown in [10], [11] that the original trace ratio problem can be converted to an equivalent *trace difference* problem having the form:

$$\tilde{\mathcal{J}}(\mathbf{W}, \lambda) = Tr\left[\mathbf{W}^T\left(\ddot{\mathbf{S}}_b - \lambda \ddot{\mathbf{S}}_w\right)\mathbf{W}\right], \tag{7}$$

where $\lambda \geq 0$ is the trace ratio $\lambda = \frac{Tr(\mathbf{W}^T \ddot{\mathbf{S}}_b \mathbf{W})}{Tr(\mathbf{W}^T \ddot{\mathbf{S}}_w \mathbf{W})}$. The best trace ratio value λ^* can be calculated by applying an iterative procedure. For more details on the λ^* calculation, please refer to [10], [11]. After obtaining λ^*, the optimal projection matrix $\mathbf{W}^*$ is obtained by:

$$\mathbf{W}^* = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}}{argmax}\ Tr\left[\mathbf{W}^T\left(\ddot{\mathbf{S}}_b - \lambda^* \ddot{\mathbf{S}}_w\right)\mathbf{W}\right]. \tag{8}$$

That is, $\mathbf{W}^*$ is obtained by performing eigenanalysis on the matrix $\ddot{\mathbf{S}} = \ddot{\mathbf{S}}_b - \lambda^* \ddot{\mathbf{S}}_w$. As has been proven in [11], the so-called *Trace Difference LDA* has a closed form solution and the global optimum of the trace ratio problem can be found by applying an efficient algorithm based on the Newton-Raphson method.

III. Optimal LDA class representation

In this paper, we relax the assumption of class representation by the mean class vector and assume that each class can be represented by any vector $\tilde{\boldsymbol{\mu}}_i \in \mathbb{R}^D$, called class representative vector. In order to obtain both the optimal projection matrix $\mathbf{W}^*$ and the optimal class representation $\tilde{\boldsymbol{\mu}}_i^*$, we propose to minimize the following criterion with respect to both $\mathbf{W}$ and $\tilde{\boldsymbol{\mu}}_i$:

$$\mathcal{J}_2(\mathbf{W}, \tilde{\boldsymbol{\mu}}_i) = \frac{Tr(\mathbf{W}^T \tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i)\mathbf{W})}{Tr(\mathbf{W}^T \tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)\mathbf{W})}, \tag{9}$$

where $\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i)$, $\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)$ are defined as:

$$\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i) = \sum_{i=1}^{C} N_i \, \tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_i^T, \tag{10}$$

$$\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i) = \sum_{i=1}^{C}\sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \tilde{\boldsymbol{\mu}}_i)(\mathbf{x}_{ij} - \tilde{\boldsymbol{\mu}}_i)^T. \tag{11}$$

Since both $\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i)$, $\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)$ are positive semidefinite for any $\tilde{\boldsymbol{\mu}}_i$, $i = 1, ..., C$, (9) can be converted to the following equivalent trace difference problem:

$$\tilde{\mathcal{J}}_2(\mathbf{W}, \lambda, \tilde{\boldsymbol{\mu}}_i) = Tr\left[\mathbf{W}^T\left(\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i) - \lambda \tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)\right)\mathbf{W}\right]. \tag{12}$$

The best trace ratio value λ^*, for known $\tilde{\boldsymbol{\mu}}_i$, can be efficiently calculated by applying the iterative procedure described in [11]. The optimal projection matrix $\mathbf{W}^*$ can be, subsequently, calculated by:

$$\mathbf{W}^* = \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}}{argmax}\ Tr\left[\mathbf{W}^T\left(\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i) - \lambda^* \tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)\right)\mathbf{W}\right]. \tag{13}$$

That is, $\mathbf{W}^*$ is obtained by performing eigenanalysis on the matrix $\tilde{\mathbf{S}}(\tilde{\boldsymbol{\mu}}_i) = \tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i) - \lambda^* \tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)$.

Let us denote by $\tilde{\boldsymbol{\mu}}_{i,t}^*$ the vector representing class i that is calculated at the t-th iteration of the proposed optimization

scheme. Clearly, the obtained best trace ratio value λ_t^* and the optimal projection matrix $\mathbf{W}_t^*$ depend on the class representation choice $\tilde{\boldsymbol{\mu}}_{i,t}$. By using λ_t^*, $\mathbf{W}_t^*$, the optimal class representation vectors $\tilde{\boldsymbol{\mu}}_{i,t+1}$, $i = 1, \ldots, C$, which maximize (12), can be obtained by:

$$\tilde{\boldsymbol{\mu}}_{i,t+1} = \alpha_t^* \boldsymbol{\mu}_i, \tag{14}$$

where $\alpha_t^* = \frac{\sum_{i,j=1}^{C,N_i} \mathbf{x}_{ij}^T \mathbf{W}_t^* \mathbf{W}_t^{*T} \mathbf{x}_{ij}}{\sum_{i,j=1}^{C,N_i} \mathbf{x}_{ij}^T \mathbf{W}_t^* \mathbf{W}_t^{*T} \boldsymbol{\mu}_i}$. The derivation of (14), as well as the convergence analysis of the above described optimization scheme are discussed in Appendix A.

In this procedure, the class representative vectors are initialized to the mean class vectors $\tilde{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i$, $i = 1, \ldots, C$, i.e., to the class representation of the standard LDA algorithm. In the general case of class representation using vectors $\tilde{\boldsymbol{\mu}}_{i,t}$, where t denotes the iteration of the proposed optimization scheme, the scatter matrices $\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_{i,t})$, $\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_{i,t})$ are calculated and employed to calculate the corresponding best trace ratio value λ_t^* and the optimal projection matrix $\mathbf{W}_t^*$. The vectors $\tilde{\boldsymbol{\mu}}_{i,t+1}$ are, subsequently calculated using (14). The above described iterative procedure is performed multiple times T, until $\tilde{\mathcal{J}}_2(\mathbf{W}, \lambda, \tilde{\boldsymbol{\mu}}_i)_{t+1} - \tilde{\mathcal{J}}_2(\mathbf{W}, \lambda, \tilde{\boldsymbol{\mu}}_i)_t < \epsilon$, where ϵ is a small positive value. In general, we would expect the class representative vectors to be close to mean class vectors in the cases where the LDA assumptions are met. In these cases, a small number of iterations is required for convergence. In different cases, the class representative vectors may be quite different from the mean class vectors and a higher number of iterations are required.

In the above described procedure we assume that the rank of $\tilde{\mathbf{S}}(\tilde{\boldsymbol{\mu}}_i)$ is larger than $D - d$, i.e., the dimensionality of the null space of $\tilde{\mathbf{S}}(\tilde{\boldsymbol{\mu}}_i)$ is smaller than d, similar to [11]. This is to make the trace ratio value finite. When the dimensionality of the null space of $\tilde{\mathbf{S}}(\tilde{\boldsymbol{\mu}}_i)$ is greater than d, i.e., greater than the dimensionality of the resulted space, the optimal trace ratio value goes to infinity. In this case a natural alternative solution is to maximize the trace of the between scatter matrix, i.e., to solve for $max\ Tr\left[\mathbf{W}^T \tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_i)\mathbf{W}\right]$, to find the appropriate transform matrix in the null space of $\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_i)$ [12].

IV. Time Complexity

Clearly, in the test phase, the time complexity of the proposed approach is equal to that of the standard LDA- based data projection, i.e., equal to $\mathcal{O}(dD)$. In order to calculate the computational complexity of the proposed optimization scheme in the training phase, we should consider the following facts:

- The iterative optimization scheme is performed for T iterations, until convergence.
- Each of these T iteration involves the following two steps:
  - Solution of the Trace Difference optimization problem.
  - Adaptation of $\tilde{\boldsymbol{\mu}}_i$, $i = 1, \ldots, C$.

Since the complexity of $\tilde{\boldsymbol{\mu}}_i$ adaptation is lower compared to that of the Trace Difference optimization problem solution, we can conclude that the complexity of the proposed optimization scheme is equal to T times the complexity of the trace difference algorithm, i.e., $\mathcal{O}(TD^3)$.

V. Experiments

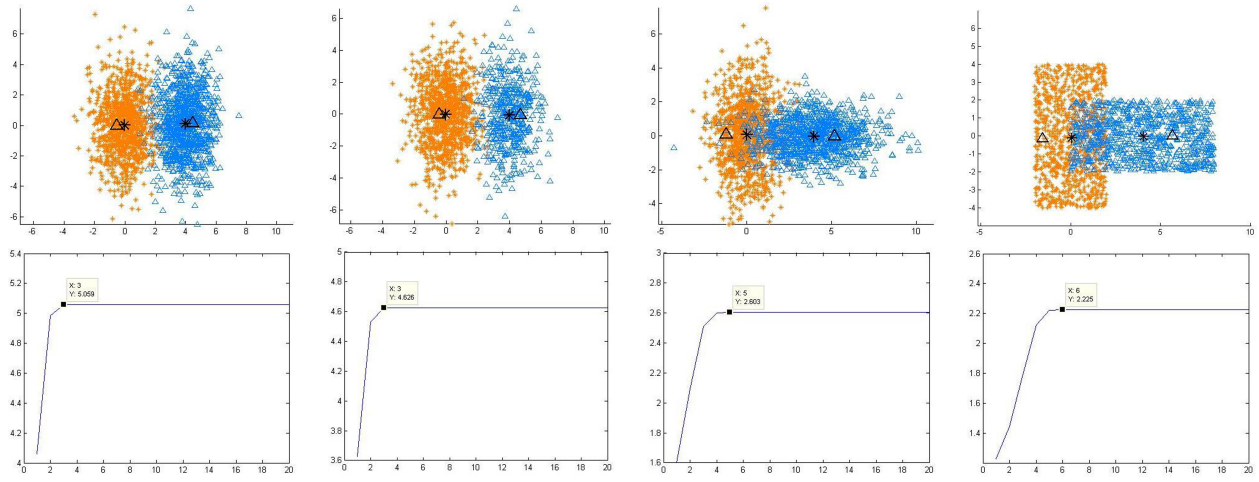
In this section, we present experiments conducted in order to evaluate the proposed LDA optimization scheme. We have conducted experiments on synthetic data, as well as on publicly available data sets. In all these experiments, we compare the performance of the proposed Reference Vector LDA (RV-LDA) algorithm to that of the Ratio Trace LDA (RT-LDA) and Trace Difference LDA (TD-LDA) algorithms. In all the cases, classification is performed by employing a modified nearest class centroid classification algorithm. That is, a given test sample $\mathbf{x}_{test}$ is projected to the decision space by applying $\mathbf{y}_{test} = \mathbf{W}^{*T}\mathbf{x}_{test}$ and is assigned to the class of the nearest vector $\tilde{\mathbf{m}}_i = \mathbf{W}^{*T}\tilde{\boldsymbol{\mu}}_i$ by using the Euclidean distance $l_{test} = \arg\min_i \|\tilde{\mathbf{m}}_i - \mathbf{y}_{test}\|_2$, $i = 1, ..., C$. A parameter value $\epsilon = 0.001$ has been used for both the Trace Difference LDA and the proposed Reference Vector LDA algorithms. All the experiments have been run on a 32bit, 2.40GHz, 3.48GB PC, using a Matlab implementation.

A. Experiments on synthetic data

In our first set of experiments, in order to qualitatively assess the impact of the proposed optimization scheme on the representative class vector choice, we created two data classes centered at $\boldsymbol{\mu}_1 = [0, 0]^T$, $\boldsymbol{\mu}_2 = [0, 4]^T$, following normal and uniform distributions, as illustrated in Figure 2. As can be seen in this Figure, in the case where the two classes follow normal distributions having the same covariance structure (Figures 2a,b) only three iterations are sufficient for the convergence of the proposed iterative optimization scheme. The obtained representative class vectors $\tilde{\boldsymbol{\mu}}_i$, $i = 1, 2$ are quite similar to the mean class vectors. In the cases where the LDA assumptions are not met, i.e., in the case where the two classes do not have the same covariance structure (Figure 2c) or the two classes follow uniform distributions (Figure 2d) the representative class vectors are quite different from the mean class vectors, while a higher number of iterations, 5 and 6 respectively, are required for convergence.

B. Experiments on Standard Classification Problems

We have conducted experiments on publicly available classification data sets coming from the machine learning repository of University of California Irvine (UCI) [13]. Table I provides information concerning the data sets used. This table, also, includes two factors R, F related to the LDA assumptions for each data set. We have used the Shapiro-Wilk parametric hypothesis test of composite normality, in order to determine if the null hypothesis of composite normality of data along each class principal direction is a reasonable assumption, according to a significance level $\alpha = 0.05$. R is the mean ratio of the number of class principal directions following normal distribution to the total number of class principal directions. A value $R \simeq 1$ denotes that the class data follow normal distributions, while a value $R \simeq 0$ denotes that the data along most class principal directions do not follow a normal distribution. Furthermore, we calculated the Frobenius norm F of the difference of covariance matrices referring to different



(a) (b) (c) (d)

Fig. 2. *2D data forming two classes, mean class vectors $\boldsymbol{\mu}_i$ (asterisks), obtained representative class vectors $\tilde{\boldsymbol{\mu}}_i$ (triangles) and $\mathcal{J}_2$ values obtained by applying the proposed optimization scheme: a) normal distributions of* 1000 *samples each, b) normal distributions of* 1000 *and* 500 *samples, c) normal distributions of* 1000 *samples each with different covariance structure and d) uniform distributions of* 1000 *samples each.*

TABLE I
UCI DATA SETS DETAILS.

Data set	Samples	Dimensions	Classes	R	F
Australian	690	14	2	0.5714	2.0726
German	1000	24	2	0.1667	14.3162
Heart	270	13	2	0.5	26.1194
Indians	768	8	2	0.125	591.145
Ionosphere	351	34	2	0.6764	0.049
Iris	150	4	3	1	0.0653
Letter	20000	16	26	0.1995	1.3529
Madelon	2600	500	2	0.941	15.7274
Relax	182	12	2	0.7917	0.0157
Sat	6435	64	6	0.8796	64.6407
Spect	267	22	2	0.2955	0.0234
Spectf	267	44	2	0.7614	22.0091
Tic-tac-toe	958	9	2	0.2222	0.0479
Vertebral2c	310	6	2	0.25	160.149
Wine	178	13	3	0.7436	1.2096

classes, normalized with respect to the number of classes and data dimensions. A value $F \simeq 0$ denotes that the classes forming a data set have the same covariance structure, while a value $F \gg 0$ denotes that the class covariance structure is quite different.

The 5-fold cross-validation procedure has been performed for the standard RT-LDA, the TD-LDA and the proposed RV-LDA algorithm. The mean classification rate over all folds has been used to measure the performance of each algorithm in one experiment. 100 experiments have been performed for each data set. The mean classification rates, the observed standard deviations over all experiments and the mean training time for all three methods, as well as the mean optimal Fisher ratio values for the standard LDA (6) and the proposed RV-LDA (9) algorithms are illustrated in Table II. By observing this Table, it can be seen that the proposed RV-LDA algorithm outperforms the TD-LDA algorithm in all the cases and outperforms the RT-LDA algorithm in all but three cases. Furthermore, it can be seen that, in the cases where the Fisher ratio is low, meaning that the classes are not well separated in the projection subspace, the proposed RV-LDA algorithm supplies consistently better results. Finally, we see that the optimal choice of the representative class vectors leads to a much bigger trace ratio, and, hence to increased class separability in the projection space. In order to compare the computational cost of the TD-LDA and the proposed RV-LDA algorithms, the mean number of iterations T required for the convergence of the RV-LDA algorithm is, also, provided.

C. Experiments on Human Action Recognition

We have conducted experiments on a publicly available human action recognition database, named AIIA-MOBISERV eating and drinking database [14]. 12 persons were recorded during four meals, each in a different day. The persons eat using spoon, cutlery and fork and drink from a cup or a glass. Several actions, such as slicing food and rest appear between eating and drinking periods. That is, a three-class classification problem can be formulated containing the action classes: 'eat', 'drink' and 'apraxia'. We have employed the dyneme-based action representation [15] for action description.

The Leave-One-Person-Out cross-validation procedure has been performed for the RT-LDA, the TD-LDA and the proposed RV-LDA algorithms. That is, the algorithms have been trained multiple times (folds) by using the action videos depicting all but one persons in the database and tested by using the action videos of the remaining person. Twelve folds, one for each test person, have been performed in order to complete an experiment. The mean correct classification rate over all folds has been used to measure the performance of each algorithm in one experiment. 100 experiments have been performed, in order to assess the performance of the three LDA-based classification schemes. This procedure has been repeated 10 times, for different numbers of dynemes $K = 5k$, $k = 1, \ldots, 10$, as illustrated in Figure 3.

As can be seen, the proposed RV-LDA algorithm outperforms both the RT- and the TD-LDA algorithms in most cases. The best action classification rate (88.34%) has been obtained

TABLE II
COMPARISON RESULTS ON STANDARD CLASSIFICATION PROBLEMS.

Data set	$\frac{Tr[W^T S_b W]}{Tr[W^T S_w W]}$	$\frac{Tr[W^{*T} S_b W^*]}{Tr[W^{*T} S_w W^*]}$	T	Classification Rate (%)			Computation Time		
				RT-LDA	TD-LDA.	RV-LDA	RT-LDA	TD-LDA.	RV-LDA
Australian	1.4866	2.4259	4.28	**85.92 (±0.14)**	84.05 (±2.46)	84.81 (±2.42)	$0.24ms$	$0.49ms$	$2.21ms$
German	0.3501	1.501	4.31	72.09 (±0.53)	72.09 (±0.53)	**76.53 (±0.53)**	$0.46ms$	$0.77ms$	$3.32ms$
Heart	1.2343	2.2343	2.84	83.68 (±0.67)	83.68 (±0.67)	**83.81 (±0.67)**	$0.12ms$	$0.26ms$	$1ms$
Indians	0.4402	1.4402	5.09	75.87 (±0.52)	75.87 (±0.53)	**77 (±0.5)**	$0.21ms$	$0.32ms$	$1.63ms$
Ionosphere	1.764	2.764	3.57	86.5 (±0.94)	86.5 (±0.94)	**86.51 (±0.86)**	$0.47ms$	$0.7ms$	$2.5ms$
Iris	16.522	25.037	4	97.85 (±0.39)	97.39 (±0.5)	**98.11 (±0.6)**	$0.07ms$	$0.11ms$	$0.44ms$
Letter	0.7446	1.4793	5.66	**70.19 (±0.06)**	57.5 (±0.07)	62.03 (±0.08)	$65ms$	$68ms$	$38.54ms$
Madelon	0.4253	1.4253	3.03	55.52 (±0.78)	55.52 (±0.78)	**55.53 (±0.78)**	$11.12s$	$14.28s$	$43.34s$
Relax	0.0415	1.414	3.73	44.43 (±3.28)	44.41 (±3.2)	**69.9 (±1.26)**	$0.01ms$	$0.19ms$	$0.71ms$
Sat	3.1324	7.9127	2.83	**84.12 (±0.13)**	73.41 (±0.5)	75.36 (±0.49)	$4.29ms$	$7.77ms$	$22.05ms$
Spect	0.4167	1.4167	3.77	72.56 (±1.28)	72.56 (±1.28)	**81.77 (±1.67)**	$0.17ms$	$0.31ms$	$1.17ms$
Spectf	0.4201	1.4201	3.27	66.11 (±2.01)	66.11 (±2.01)	**76.72 (±0.13)**	$0.42ms$	$1.2ms$	$3.93ms$
Tic-tac-toe	0.069	1.07	5.89	57.72 (±0.77)	57.72 (±0.77)	**68 (±0.7)**	$0.26ms$	$0.29ms$	$1.71ms$
Vertebral2c	0.5437	1.544	4.66	79.32 (±0.97)	79.29 (±0.85)	**83.32 (±0.93)**	$0.1ms$	$0.15ms$	$0.7ms$
Wine	6.8345	9.8594	3.34	98.3 (±0.6)	92.37 (±1.03)	**98.34 (±0.86)**	$0.11ms$	$0.32ms$	$1.07ms$

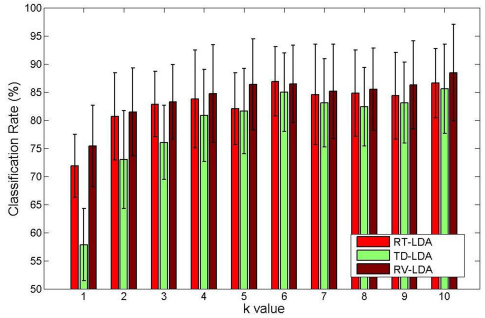


Fig. 3. *Action classification rates on the AIIA-MOBISERV database for different values of k.*

by using 50 dynemes and the RV-LDA algorithm. The best action classification rates for the RT-LDA and the TD-LDA algorithms have been obtained by using 30 and 50 dynemes and are equal to 86.88% and 85.56%, respectively. Furthermore, it can be seen that the RV-LDA algorithm outperforms both the RT-LDA and the TD-LDA algorithms in most of the classification problems illustrated in Figure 3.

D. Experiments on Face Recognition

We have conducted experiments on a publicly available face recognition database, namely ORL Database of Faces [16]. It contains 400 images from 40 persons, each person is depicted in 10 images. Each image has been downsized to 40×30 pixels for computation speed consideration and vectorized to produce a 1200 dimensional facial vector. The dimensionality of the facial vectors has been reduced by applying PCA so that 90% of the energy is preserved. The 5-fold cross-validation procedure has been performed for the RT-LDA, the TD-LDA and the RV-LDA algorithms. The mean classification rate over all folds has been used to measure the performance of each algorithm in one experiment. 100 experiments have been performed in total. The obtained mean classification rates and the observed standard deviations over all experiments are illustrated in Table III. As can be seen, the proposed RV-LDA algorithm outperforms both the competing LDA algorithms.

TABLE III
COMPARISON RESULTS ON THE ORL DATABASE OF FACES.

RT-LDA	TD-LDA	RV-LDA
95.97% (±0.85)	95.43% (±0.77)	**97.00% (±0.76)**

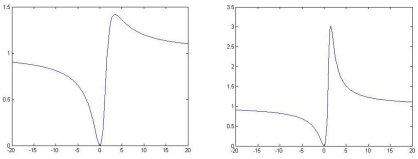
VI. STATISTICAL SIGNIFICANCE ANALYSIS OF EXPERIMENTAL RESULTS

The Friedman test was used, in order to test the null hypothesis that all the three classifiers perform equally well and the observed differences are merely random [17]. After ordering the algorithms according to their performance on each data set, the obtained mean ranks are equal to $R_{RT} = 2.0756$, $R_{TD} = 2.7308$ and $R_{RV} = 1.1932$ for the RT-LDA, the TD-LDA and the RV-LDA algorithms, respectively. The overall mean rank is equal to $R_j = 2$. The Friedman statistic is equal to $\chi_F^2 = 20.213$ and $F_F = 23.4574$. With $k = 3$ classifiers and $N = 17$ data sets, F_F is distributed according to an F distribution with $(3-1) = 2$ and $(3-1)\times(17-1) = 32$ degrees of freedom. The critical value of $F(2, 32)$ for $\alpha = 0.05$ is 3.32, so we reject the null hypothesis that all the classifiers perform the same.

Following the Nemenyi test for pairwise comparisons, we obtain a critical value equal to 2.343 and, thus, the critical difference is equal to $CD = 2.343\sqrt{\frac{k(k+1)}{6N}} = 0.8284$. By calculating the differences between the ranks of the three classifiers, we obtain $P_{RT} - P_{RV} = 0.8824 > CD$, $P_{TD} - P_{RV} = 1.5376 > CD$ and $P_{TD} - P_{RT} = 0.6552 < CD$. Thus, the proposed RV-LDA performs significantly better than the two competing ones, while the Ratio Trace LDA and the Trace Difference LDA perform the same.

VII. CONCLUSION

In this paper, we presented an optimization scheme aiming at the optimal class representation for LDA based data projection. By optimizing the LDA criterion with respect to both the data projection matrix and the class representation in the projection space, the optimal discriminant projection space, in



(a) (b)

Fig. 4. $\mathcal{J}_2(\alpha)$ *for different values of* λ^**: a)* $\lambda^* = 0.4175$ *and b)* $\lambda^* = 2.0281$.

terms of Fisher ratio maximization, is obtained. Experimental results on synthetic and real data show that the proposed optimization scheme increases class discrimination compared to the standard LDA approach.

APPENDIX A

The proposed optimization scheme consists of two steps. The first one, for given class representative vectors $\tilde{\boldsymbol{\mu}}_i$, $i = 1, \ldots, C$, determines the optimal parameter λ^* value and the optimal projection matrix $\mathbf{W}^*$, while the second one updates $\tilde{\boldsymbol{\mu}}_i$ by using the obtained λ^* and $\mathbf{W}^*$. Since the convergence of the first step has been proven in [11], we focus on proving the convergence of the second step.

Let us denote as $\tilde{\mathbf{S}}_{b,t}$ and $\tilde{\mathbf{S}}_{w,t}$ the scatter matrices corresponding to the class representative vectors $\tilde{\boldsymbol{\mu}}_{i,t}$ calculated for the t-th iteration of the proposed optimization scheme. (12) can be written in the form:

$$\tilde{\mathcal{J}}_2(\tilde{\boldsymbol{\mu}}_{i,t}) = Tr\left[\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_{b,t}\mathbf{W}_t^*\right] - \lambda_t^*\, Tr\left[\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_{w,t}\mathbf{W}_t^*\right].$$

Using (10), (11) and (A.1), the first derivative of $\tilde{\mathcal{J}}_2(\tilde{\boldsymbol{\mu}}_{i,t})$ with respect to $\tilde{\boldsymbol{\mu}}_{i,t}$ can be expressed by:

$$\frac{\partial\tilde{\mathcal{J}}_2(\tilde{\boldsymbol{\mu}}_{i,t})}{\partial\tilde{\boldsymbol{\mu}}_{i,t}} = 2N_i\mathbf{W}_t^*\mathbf{W}_t^{*T}\left[(1-\lambda_t^*)\,\tilde{\boldsymbol{\mu}}_{i,t} + \lambda_t^*\boldsymbol{\mu}_i\right]. \qquad \text{(A.1)}$$

When solving for $\frac{\partial\tilde{\mathcal{J}}_2(\tilde{\boldsymbol{\mu}}_{i,t})}{\partial\tilde{\boldsymbol{\mu}}_{i,t}} = 0$ and by exploiting that $\mathbf{W}_t^*$ is formed by the eigenvectors of $\tilde{\mathbf{S}}_t = \tilde{\mathbf{S}}_{b,t} - \lambda_t^*\tilde{\mathbf{S}}_{w,t}$, we obtain:

$$(1-\lambda_t^*)\,\tilde{\boldsymbol{\mu}}_{i,t} + \lambda_t^*\boldsymbol{\mu}_i = 0. \qquad \text{(A.2)}$$

It is obvious from (A.2) that $\tilde{\boldsymbol{\mu}}_{i,t}^*$ is in the direction of $\boldsymbol{\mu}_i$, i.e, $\tilde{\boldsymbol{\mu}}_{i,t}^* = \alpha_t\boldsymbol{\mu}_i$, $\alpha_t \in \mathbb{R}$. In Figure 4, we illustrate $\mathcal{J}_2(\alpha_t)$ for two choices of λ_t^*. As can be seen, $\mathcal{J}_2(\alpha_t)$, typically, has two stationary points. The first one is its global minimum for $\alpha_t = 0$, i.e., when $\tilde{\boldsymbol{\mu}}_i = \boldsymbol{\mu}$ and, thus, $Tr\left[\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_b\mathbf{W}_t^*\right] = 0$. The second one is the global maximum, obtained by using a value α_t^* that corresponds to the optimal, in terms of $\mathcal{J}_2(\alpha_t)$ maximization, representative class vectors.

By using (3), (4) and setting $b = \sum_{i=1}^{C} N_i\,\boldsymbol{\mu}_i^T\mathbf{W}_t^*\mathbf{W}_t^{*T}\boldsymbol{\mu}_i$, $c = \sum_{i,j=1}^{C,N_i}\mathbf{x}_{ij}^T\mathbf{W}_t^*\,\mathbf{W}_t^{*T}\mathbf{x}_{ij}$ and $d = 2\sum_{i,j=1}^{C,N_i}\mathbf{x}_{ij}^T\mathbf{W}_t^*\mathbf{W}_t^{*T}\boldsymbol{\mu}_i$, $Tr\left[\mathbf{W}_t^{*T}\ddot{\mathbf{S}}_w\mathbf{W}_t^*\right] = c - d + b$ and $Tr\left[\mathbf{W}_t^{*T}\ddot{\mathbf{S}}_b\mathbf{W}_t^*\right] = b$. By substituting $\tilde{\boldsymbol{\mu}}_{i,t} = \alpha_t\boldsymbol{\mu}_i$ in (10), (11), we obtain $Tr\left[\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_w(\alpha_t)\mathbf{W}_t^*\right] = c - \alpha_t d + \alpha_t^2 b$ and $Tr\left[\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_b(\alpha_t)\mathbf{W}_t^*\right] = \alpha_t^2 b$. Thus, the first derivative of $\tilde{\mathcal{J}}_2(\alpha_t)$ with respect to α_t is given by:

$$\frac{\partial\mathcal{J}_2(\alpha_t)}{\partial\alpha_t} = \frac{2\alpha_t bc - \alpha_t^2 bd}{(\alpha_t^2 b - \alpha_t d + c)^2}, \qquad \alpha_t^2 b - \alpha_t d + c > 0. \qquad \text{(A.3)}$$

By solving for $\frac{\partial\mathcal{J}_2(\alpha_t)}{\partial\alpha_t} = 0$, two stationary points are obtained, $\alpha_{t1}^* = 0$, $\alpha_{t2}^* = \frac{2c}{d}$. It is straightforward to show that $\frac{\partial\mathcal{J}_2^2(\alpha_{t1}^*)}{\partial\alpha_t^2} > 0$ and $\frac{\partial\mathcal{J}_2^2(\alpha_{t2}^*)}{\partial\alpha_t^2} < 0$ and that $\mathcal{J}_2(\alpha_t) \to 1$ for $\alpha_t \to \pm\infty$. That is, α_{t1}^*, α_{t2}^* correspond to the global minimum and maximum of $\mathcal{J}_2(\alpha_t)$, respectively. Since we aim at maximizing $\mathcal{J}_2(\alpha_t)$, the representative class vectors are given by $\tilde{\boldsymbol{\mu}}_{i,t+1} = \alpha_{t2}^*\boldsymbol{\mu}_i$.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Collaborative European Project MOBISERV FP7-248434 (http://www.mobiserv.info), An Integrated Intelligent Home Environment for the Provision of Health, Nutrition and Mobility Services to the Elderly.

REFERENCES

[1] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face recognition using lda-based algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.

[2] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, pp. 347–360, 2012.

[3] A. Iosifidis, A. Tefas, and I. Pitas, "Activity based Person Identification using Fuzzy Representation and Discriminant Learning," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 530–542, 2012.

[4] F. Dufrenois and J. Noyer, "Formulating robust linear regression estimation as a one-class lda criterion: Discriminative hat matrix," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 2, pp. 262–273, 2013.

[5] Y. Hou, I. Song, H. Min, and C. Park, "Complexity-reduced scheme for feature extraction with linear discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 1003–1009, 2012.

[6] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Mixture subclass discriminant analysis link to restricted gaussian model and other generalizations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 8–21, 2013.

[7] Y. Huang, D. Xu, and F. Nie, "Semi-supervised dimension reduction using trace ratio criterion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 519–526, 2002.

[8] C. Dhir and S. Lee, "Discriminant independent component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 845–857, 2011.

[9] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1119–1132, 2011.

[10] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," *Conference on Computer Vision and Pattern Recognition - CVPR*, pp. 1–8, 2007.

[11] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[12] L. Chen, M. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system wich can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[13] A. Frank and A. Asuncion, "Uci machine learning repository," 2010.

[14] A. Iosifidis, E. Marami, A. Tefas, and I. Pitas, "Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2201–2204, 2012.

[15] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–425, 2012.

[16] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," *IEEE Workshop on Applications of Computer Vision*, pp. 138–142, 1994.

[17] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.