



Canonical correlation analysis using within-class coupling[☆]

Olcay Kursun^{a,*}, Ethem Alpaydin^b, Oleg V. Favorov^c

^a Department of Computer Engineering, Istanbul University, Avcilar, Istanbul 34320, Turkey

^b Department of Computer Engineering, Bogazici University, Bebek, Istanbul 34342, Turkey

^c Biomedical Engineering Department, University of North Carolina, Chapel Hill, NC 27599-7575, USA

ARTICLE INFO

Article history:

Received 27 November 2009

Available online 31 October 2010

Communicated by W. Pedrycz

Keywords:

Temporal contextual guidance

Linear discriminant analysis (LDA)

Samples versus samples canonical

correlation analysis (CCA)

Feature extraction

ABSTRACT

Fisher's linear discriminant analysis (LDA) is one of the most popular supervised linear dimensionality reduction methods. Unfortunately, LDA is not suitable for problems where the class labels are not available and only the spatial or temporal association of data samples is implicitly indicative of class membership. In this study, a new strategy for reducing LDA to Hotelling's canonical correlation analysis (CCA) is proposed. CCA seeks prominently correlated projections between two views of data and it has been long known to be equivalent to LDA when the data features are used in one view and the class labels are used in the other view. The basic idea of the new equivalence between LDA and CCA, which we call within-class coupling CCA (WCCCA), is to apply CCA to pairs of data samples that are most likely to belong to the same class. We prove the equivalence between LDA and such an application of CCA. With such an implicit representation of the class labels, WCCCA is applicable both to regular LDA problems and to problems in which only spatial and/or temporal continuity provides clues to the class labels.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Fisher's linear discriminant analysis (LDA; Fisher, 1936) and Hotelling's canonical correlation analysis (CCA; Hotelling, 1936) are among the oldest, yet the most powerful multivariate data analysis techniques. LDA is one of the most popular supervised dimensionality reduction methods incorporating the categorical class labels of the data samples into a search for linear projections of the data that maximize the between-class variance while minimizing the within-class variance (Rencher, 1997; Alpaydin, 2004; Izenman, 2008).

On the other hand, CCA works with two sets of (related) variables and its goal is to find a linear projection of the first set of variables that maximally correlates with a linear projection of the second set of variables. These sets have recently been also referred to as *views* or *representations* (Hardoon et al., 2004). Finding correlated functions (covariates) of the two views of the same phenomenon by discarding the representation-specific details (noise) is expected to reveal the underlying hidden yet influential semantic factors responsible for the correlation (Hardoon et al., 2004; Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003).

Both LDA and CCA have been proposed in 1936, and shortly after, a direct link between them has been shown by Bartlett (1938) as fol-

lows. Given a dataset of samples and their class labels, if we consider the features given for the data samples as one view, versus the class labels as the other view (a single binary variable works for the two-class problem but a form of 1-of-C coding scheme is typically used for multi-class categorical class labels), this CCA setup is known to be equivalent to LDA (Bartlett, 1938; Hastie et al., 1995). In other words, LDA can be simply said to be a special case of CCA.

The knowledge of this insightful equivalence between LDA and CCA enabled the researchers attempt the use of CCA to surpass the quality of the LDA projections. These attempts used samples versus their class labels using several other forms of representations for the labels (Loog et al., 2005; Barker and Rayens, 2003; Gestel et al., 2001; Johansson, 2001; Sun and Chen, 2007). An interesting example of such a label transformation is by replacing hard categorical labels by soft-labels; in (Sun and Chen, 2007), similar to the support vector idea, the aim was to put more weight on the samples near the class boundaries rather than using a common label for all the samples of a class; thus, more useful projections were found as more focus was placed on the problematic regions in the input space rather than the high-density regions with class centers. Another example is the study on an image segmentation task presented in (Loog et al., 2005), which uses image-pixel features and their associated class labels for learning to classify pixels. Their CCA-based method incorporates the class labels of the neighboring pixels as well, which can naturally be expected to yield LDA-like (but possibly more informative) projections. The method can be applied to other forms of, non-image, data by accounting for the spatial class label configuration in the vicinity of every feature vector (Loog et al., 2005).

[☆] The work of O. Kursun was supported by Scientific Research Projects Coordination Unit of Istanbul University under the grant YADOP-5323.

* Corresponding author. Tel.: +90 212 473 7070/17913; fax: +90 212 473 7180.

E-mail addresses: okursun@istanbul.edu.tr (O. Kursun), alpaydin@boun.edu.tr (E. Alpaydin), favorov@bme.unc.edu (O.V. Favorov).

In this paper, we present another extension of CCA to LDA along with its equivalence proof. The main idea is to transform the class label of a sample such that it is represented, in a distributed manner, by all the samples in that same class. In other words, CCA is asked to produce correlated outputs (projections) for any pair of samples that belong to the same class, which we called WCCCA that stands for within-class coupling CCA. This extension of CCA to LDA has various advantages despite its increased complexity (see Section 4.2 for a detailed list). One important advantage of the WCCCA idea of using samples versus samples, as the two views, is in its ability to perform a form of implicitly-supervised LDA (see Section 5.2) as sometimes the class labels may be embedded in the patterns of the data rather than being explicitly available, for example, in the patterns of spatial and temporal continuity (Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003; Borga and Knutsson, 2001; Stone, 1996). Among exemplary applications on such data, the tasks of division of a video into sequences of relevant frames (scenes), segmentation of an image into image regions sharing certain visual characteristics, identifying sequences of acoustic frames belonging to the same word in speech analysis, or finding sequences of base pairs or amino acids belonging to the same protein in biological sequence analysis can be mentioned. In such settings, the use of LDA is uneasy, if not impossible.

The idea of applying CCA or other forms of mutual information maximization models, for example, between the consecutive frames of a video or between the neighboring image patches for finding correlated functions is not a new one (Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003; Borga and Knutsson, 2001; Borga, 1998; Stone, 1996; Kording and Konig, 2000; Phillips et al., 1995; Phillips and Singer, 1997). Many of these attempts are inspired by the learning mechanisms hypothesized to be used by neurons in the cerebral cortex. For example, cortical neurons might tune to correlated functions between their own afferent inputs and the lateral inputs they receive from other neurons with different but functionally related afferent inputs. Thus, groups of neurons receiving such different but related afferent inputs can learn to produce correlated outputs under the contextual guidance of each other (Phillips et al., 1995; Phillips and Singer, 1997). However, it is not mathematically justified whether these correlated functions are good for discrimination. Would the covariates found this way be suitable projections for clustering the frames into scenes or for image segmentation? The results of our study show that such a CCA application would be comparable to performing LDA; and as LDA projections maximize the between-class variance and minimize the within-class variance, the covariates found this way would be useful, for example, to cluster the frames into scenes.

This paper is organized as follows. In Sections 2 and 3, we review the CCA and LDA techniques, respectively. In Section 4, we present the WCCCA idea of using CCA on a samples versus samples basis and provide the proof for its equivalence to LDA; we also show that the theoretically derived equivalence holds also practically on a toy example. In Section 4, we also discuss the advantages and disadvantages of this way of performing LDA; and finally, finish this section by showing the nonlinear kernel extension of WCCCA. In Section 5, we present the experimental results on a face database and show that WCCCA can perform the task of LDA even when the images are made into a movie and the class label information is kept only implicitly through the temporal continuity of the identity of the individual seen in contiguous frames. We conclude in Section 6.

2. Canonical correlation analysis (CCA)

Canonical correlation analysis (CCA) is introduced by Hotelling (1936) to describe the linear relation between two multidimen-

sional (or two sets of) variables as the problem of finding basis vectors for each set such that the projections of the two variables on their respective basis vectors are maximally correlated (Hotelling, 1936; Rencher, 1997; Hardoon et al., 2004; Izenman, 2008). These two sets of variables, for example, may correspond to different views of the same semantic object (e.g. audio versus video of a person speaking, two cameras viewing the same object as in binocular vision, text versus links or images in webpages, etc). Let u -dimensional X and v -dimensional Y denote corresponding two sets of real-valued random variables (i.e., $X \in \mathbb{R}^u$ and $Y \in \mathbb{R}^v$), the canonical correlation is defined as:

$$\rho(X; Y) = \sup_{f, g} \text{corr}(f^T X; g^T Y), \quad (1)$$

where, $\text{corr}(X; Y)$ stands for Pearson's correlation coefficient. Let u -dimensional column vector $X = x_i$ denote the i th sample (row) of the first view (dataset), v -dimensional column vector $Y = y_i$ denote the i th sample of the second dataset, and N denote the total number of samples. Then, the first dataset D_1 is an $N \times u$ matrix that can be expressed as:

$$D_1 = [x_1, x_2, \dots, x_N]^T \quad (2)$$

and similarly, the $N \times v$ matrix for the second dataset D_2 can be written as:

$$D_2 = [y_1, y_2, \dots, y_N]^T. \quad (3)$$

Assuming that each dataset has zero mean, the total covariance matrix of (X, Y) can be written as a block matrix:

$$C(X, Y) = E \left\{ \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}^T \right\} = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}, \quad (4)$$

where the within-sets covariance matrices are given as:

$$\begin{aligned} C_{XX} &= E\{XX^T\}, \\ C_{YY} &= E\{YY^T\} \end{aligned} \quad (5)$$

and the between-sets covariance matrix is given as:

$$C_{XY} = E\{XY^T\} = C_{YX}^T \quad (6)$$

and now the canonical correlation is the maximum of ρ with respect to f and g :

$$\rho(X; Y) = \sup_{f, g} \frac{f^T C_{XY} g}{\sqrt{f^T C_{XX} f g^T C_{YY} g}}. \quad (7)$$

The problem of finding the orthogonal projections that achieve the top correlations reduces to a generalized eigenproblem, where the projection f (and the projection g can be solved for similarly) corresponds to the top eigenvector of the following (Hardoon et al., 2004):

$$C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX} f = \lambda_{CCA} f \quad (8)$$

and

$$\rho(X; Y) = \sqrt{\lambda_{CCA}}, \quad (9)$$

where λ_{CCA} corresponds to the largest eigenvalue of Eq. (8).

3. Fisher linear discriminant analysis (LDA)

Fisher linear discriminant analysis (LDA) is a variance preserving approach with the goal of finding the optimal linear discriminant function (Fisher, 1936; Rencher, 1997; Raudys and Duin, 1998; Alpaydin, 2004; Izenman, 2008). As opposed to unsupervised methods such as principal component analysis (PCA), independent component analysis (ICA), or the two view counterpart

CCA, to utilize the categorical class label information in finding informative projections, LDA considers maximizing an objective function that involves the scatter properties of every class as well as the total scatter. The objective function is designed to be maximized by a projection that maximizes the between class (or equivalently total scatter as in PCA) and minimize the within class scatter. Let d -dimensional column vector z_i^c denote the i th sample of class c , N_c denote the number of samples in class c , $m \geq 2$ denote the total number of classes, and finally N denote the total number of samples. Then, the $N \times d$ data matrix \mathbf{D} can be written as:

$$\mathbf{D} = [z_1^1, \dots, z_{N_1}^1, z_1^2, \dots, z_{N_2}^2, \dots, z_1^m, \dots, z_{N_m}^m]^T. \quad (10)$$

Assuming that the dataset \mathbf{D} is centered (i.e. the data is normalized to zero-mean), the overall scatter (covariance) matrix \mathbf{S}_T of the dataset \mathbf{D} is given by:

$$\mathbf{S}_T = \text{cov}(\mathbf{D}) = \frac{1}{N} \sum_c \sum_{i=1}^{N_c} z_i^c z_i^{cT} \quad (11)$$

and the within-class scatter matrix is defined as:

$$\mathbf{S}_W = \sum_c \left[\frac{N_c}{N} \text{cov}(\mathbf{D}^c) \right] = \frac{1}{N} \sum_c \sum_{i=1}^{N_c} (z_i^c - \mu_c)(z_i^c - \mu_c)^T, \quad (12)$$

where $\mu_c \in \mathbb{R}^d$ denotes the mean of \mathbf{D}^c (the samples of class c). The between-class scatter matrix can be formulated as:

$$\mathbf{S}_B = \sum_c \frac{N_c}{N} \mu_c \mu_c^T. \quad (13)$$

In fact, the overall scatter matrix \mathbf{S}_T can be expressed as the sum of the within-class and between-class scatter matrices:

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B. \quad (14)$$

Finally, the LDA objective function for finding the most discriminative projection vector h (and other orthogonal projection vectors) is given by:

$$\lambda_{LDA} = \sup \frac{h^T \mathbf{S}_B h}{h^T \mathbf{S}_W h} = \sup \frac{h^T (\mathbf{S}_T - \mathbf{S}_W) h}{h^T \mathbf{S}_W h} = \sup \frac{h^T \mathbf{S}_T h}{h^T \mathbf{S}_W h} - 1. \quad (15)$$

The optimization can be shown to be accomplished by computing the solution of the following generalized eigenproblem for the eigenvectors corresponding to the largest eigenvalues:

$$\mathbf{S}_B h = \lambda_{LDA} \mathbf{S}_W h. \quad (16)$$

An already established direct connection between LDA and CCA, which we refer to as *samples versus labels* CCA (SLCCA), was first given in (Bartlett, 1938) by showing that LDA is exactly what is accomplished by applying CCA between the data matrix \mathbf{D} and the class label matrix \mathbf{L} , a dummy matrix that carries the class label information using the 1-of-C, or the more compact 1-of-(C-1)) coding, defined as:

$$\mathbf{L} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{0}_{N_1} & \mathbf{0}_{N_1} & \cdots & \mathbf{0}_{N_1} \\ \mathbf{0}_{N_2} & \mathbf{1}_{N_2} & \mathbf{0}_{N_2} & \cdots & \mathbf{0}_{N_2} \\ \mathbf{0}_{N_3} & \mathbf{0}_{N_3} & \mathbf{1}_{N_3} & \cdots & \mathbf{0}_{N_3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{N_m} & \mathbf{0}_{N_m} & \mathbf{0}_{N_m} & \cdots & \mathbf{1}_{N_m} \end{bmatrix}_{N \times m}, \quad (17)$$

where $\mathbf{1}_j$ is a column vector of j ones and $\mathbf{0}_j$ is a column vector of j zeros, and m is the number of classes. Searching for the maximal correlation between the rows of \mathbf{D} and \mathbf{L} matrices via CCA yields the LDA projection as the solution (Bartlett, 1938; Loog et al., 2005; Barker and Rayens, 2003; Sun and Chen, 2007). It is straightforward to show that for the SLCCA setup, the CCA eigenproblem in Eq. (8) reduces to that of LDA in Eq. (16) with:

$$\lambda_{SLCCA} = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1}. \quad (18)$$

Taking advantage of this equivalence, there are more recent references utilizing CCA on the basis of samples versus their class labels to extract more useful projections than LDA using other forms of class label representations; the work in (Sun and Chen, 2007), for example, uses soft labeling of the samples to deem the samples near other classes more important (similar to the support vector idea) rather than using the actual hard labels. In another recent work (Loog et al., 2005), the class labels, used as the second view, were augmented by the class labels of the neighboring pixels (i.e. by taking advantage of the spatial context) for an image segmentation task. Moreover, Kursun and Alpaydin (2010) offer an idea based on the equivalence on the multiview semisupervised learning problem, where a CCA-based setup utilizes the unlabelled examples as well as the labelled ones in learning discriminants.

4. Within-class coupling CCA (WCCCA)

Clearly, for CCA to be applicable to a dataset \mathbf{D} , two views are necessary, denoted as X and Y in Eq. (1). However, constructing a form of the dummy class label matrix \mathbf{L} in Eq. (17) as the second one of the two views is not the only way to create these views. We prove that CCA can be used to perform LDA using a different method of incorporating the class labels of the data samples. Let us create the two views by coupling a pair of samples from the same classes (one for each view). For an example, consider $a, b, c, d \in \mathbb{R}^d$ are our four training examples, with d features each, and also let a and b belong to class 1 and c and d belong to class 2. Then we create two datasets such that the samples belonging to the same classes correspond to each other in the subsequent CCA analysis: a versus a , a versus b , b versus a , b versus b , c versus c , c versus d , d versus c , d versus d . In other words, when the feature vector X represents a sample a (a row) in \mathbf{D} (given in Eq. (10)) that belongs to class c , then the feature vector Y will correspond to another sample that also belongs to class c . The LDA problem can be shown to be polynomially reducible to CCA using this change of representation, as we show below. For the sake of simplicity, let us assume $N_i = n$ for all $1 \leq i \leq m$ (when the classes are of different cardinalities, the number of pairs produced by each class must be adjusted to preserve the prior distribution of the classes, see Section 4.3). As there are n^2 pairs for each class, for the full set of pairs to be presented, $mn^2 \times d$ matrices for the two views can be obtained from the $mn \times d$ data matrix \mathbf{D} as:

$$\mathbf{D}_1 = [z_1^1, \dots, z_n^1, z_1^1, \dots, z_n^1, z_1^1, \dots, z_n^1, \dots, z_1^1, \dots, z_n^1, \dots, z_1^m, \dots, z_n^m]^T, \quad (19)$$

$$\mathbf{D}_2 = [z_1^1, \dots, z_1^1, z_2^1, \dots, z_2^1, z_3^1, \dots, z_3^1, \dots, z_1^1, \dots, z_1^m, \dots, z_1^m, z_2^m, \dots, z_2^m, z_3^m, \dots, z_3^m, \dots, z_1^m, \dots, z_1^m]^T, \quad (20)$$

where each one of the samples of a class is paired with every other samples of that class (located on the same rows of \mathbf{D}_1 and \mathbf{D}_2 , respectively). CCA is, then, asked to find the maximally correlated functions of the rows of \mathbf{D}_1 with those of \mathbf{D}_2 . These functions are indirectly forced to produce the same output for the samples of the same class. What CCA produces as the maximally correlated projections f and g are both the same projection h that LDA would also find; because in this setup, the eigenproblems for LDA and CCA can be shown to be equivalent as follows. Recall that CCA solves the eigenproblem given in Eq. (8):

$$\mathbf{C}_{XX}^{-1} \mathbf{C}_{XY} \mathbf{C}_{YY}^{-1} \mathbf{C}_{YX} f = \lambda_{CCAF} f, \quad (8 \text{ revisited})$$

where for this particular case, we have:

$$\mathbf{C}_{XX} = \text{cov}(\mathbf{D}_1) = \text{cov}(\mathbf{D}) = \text{cov}(\mathbf{D}_2) = \mathbf{C}_{YY} = \mathbf{S}_T, \quad (20)$$

$$\mathbf{C}_{XY} = \mathbf{C}_{YX} = \mathbf{S}_B. \quad (21)$$

Eq. (20) easily follows from the fact that the samples of \mathbf{D}_1 (and similarly \mathbf{D}_2) are basically the samples in the dataset \mathbf{D} repeated n times. Therefore, the covariance matrix will not be altered at all. To show the validity of Eq. (21), consider that X and Y stand for all the pairs of the samples in the same classes. Therefore,

$$\begin{aligned} \mathbf{C}_{XY} &= E\left[z_i^c \cdot z_j^{cT}\right] = \sum_c \left[\frac{N_c}{N} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \frac{1}{N_c} z_i^c \cdot z_j^{cT} \right] \\ &= \sum_c \left[\frac{N_c}{N} \sum_{i=1}^{N_c} \frac{z_i^c}{N_c} \left[\sum_{j=1}^{N_c} \frac{z_j^{cT}}{N_c} \right] \right] = \sum_c \left[\frac{N_c}{N} \sum_{i=1}^{N_c} \frac{z_i^c}{N_c} \mu_c^T \right] \\ &= \sum_c \left[\frac{N_c}{N} \mu_c \mu_c^T \right] = \mathbf{S}_B. \end{aligned} \quad (22)$$

Thus, Eq. (8) is equivalent to:

$$\mathbf{S}_T^{-1} \mathbf{S}_B \mathbf{S}_T^{-1} \mathbf{S}_B f = \lambda_{WCCCA} f. \quad (23)$$

On the other hand, it was mentioned that LDA solves the eigenproblem:

$$\mathbf{S}_B h = \lambda_{LDA} \mathbf{S}_W h. \quad (16 \text{ revised})$$

Equivalently,

$$\mathbf{S}_B h = \lambda_{LDA} (\mathbf{S}_T - \mathbf{S}_B) h,$$

$$\mathbf{S}_B h = \lambda_{LDA} \mathbf{S}_T h - \lambda_{LDA} \mathbf{S}_B h,$$

$$(\lambda_{LDA} + 1) \mathbf{S}_B h = \lambda_{LDA} \mathbf{S}_T h,$$

$$\mathbf{S}_B h = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1} \mathbf{S}_T h,$$

$$\mathbf{S}_T^{-1} \mathbf{S}_B h = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1} h. \quad (24)$$

Let

$$\lambda^* = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1} \quad (25)$$

and

$$\mathbf{A} = \mathbf{S}_T^{-1} \mathbf{S}_B. \quad (26)$$

Then, we get from the last line of Eq. (24):

$$\mathbf{A} h = \lambda^* h. \quad (27)$$

Thus, rewriting Eq. (23), we get:

$$\mathbf{S}_T^{-1} \mathbf{S}_B \mathbf{S}_T^{-1} \mathbf{S}_B h = \mathbf{A} \mathbf{A} h = \mathbf{A} (\mathbf{A} h) = \mathbf{A} (\lambda^* h) = \lambda^* (\mathbf{A} h) = \lambda^* (\lambda^* h) = (\lambda^*)^2 h. \quad (28)$$

Therefore, $f = h$ is a solution of the eigenproblem (Eq. (23)) of WCCCA with:

$$\lambda_{WCCCA} = (\lambda^*)^2, \quad (29)$$

from which, it immediately follows that

$$\rho_{WCCCA}(X; Y) = \sqrt{\lambda_{WCCCA}} = \sqrt{(\lambda^*)^2} = \lambda^* = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1}. \quad (30)$$

Similarly, it can be shown that $g = h$. Therefore, we have:

$$f = h = g. \quad (31)$$

This shows that the LDA and WCCCA projections are exactly the same. In fact, for every eigenvector of the LDA eigenproblem, that eigenvector must be also a solution for the WCCCA eigenproblem with the same ordering of the eigenvalues. That is because the order of the eigenvalues will not change when they are squared.

Concisely, the projections found by LDA, the existing samples versus class labels CCA setup (SLCCA), and the proposed samples versus within-class samples CCA setup (WCCCA) are all identical, with the following relationships among the eigenvalues of their eigenproblems:

$$\sqrt{\lambda_{WCCCA}} = \lambda_{SLCCA} = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1}, \quad (32)$$

from which the relation between the correlation coefficients of WCCCA and SLCCA is seen to be:

$$\rho_{WCCCA} = \rho_{SLCCA}^2. \quad (33)$$

Alternatively, a different way of seeing the LDA and WCCCA equivalence follows from the fact that the WCCCA algorithm is expected to yield identical projections, f and g , as \mathbf{D}_1 and \mathbf{D}_2 are practically the same dataset but only shuffled whilst preserving the class correspondence. Thus, the objective function of CCA given in Eq. (7) could be rewritten by substituting f for g as:

$$\begin{aligned} \rho_{WCCCA}(X; Y) &= \sup_{f, g} \frac{f^T \mathbf{C}_{XY} g}{\sqrt{f^T \mathbf{C}_{XX} f} \sqrt{g^T \mathbf{C}_{YY} g}} = \sup_f \frac{f^T \mathbf{S}_B f}{\sqrt{f^T \mathbf{S}_T f} \sqrt{f^T \mathbf{S}_T f}} \\ &= \sup_f \frac{f^T \mathbf{S}_B f}{f^T \mathbf{S}_T f} = \sup_f \frac{f^T (\mathbf{S}_T - \mathbf{S}_W) f}{f^T \mathbf{S}_T f} \\ &= \sup_f \left(1 - \frac{f^T \mathbf{S}_W f}{f^T \mathbf{S}_T f} \right) = 1 - \inf_f \left(\frac{f^T \mathbf{S}_W f}{f^T \mathbf{S}_T f} \right) \\ &= 1 - \frac{1}{\sup_f \left(\frac{f^T \mathbf{S}_T f}{f^T \mathbf{S}_W f} \right)} = 1 - \frac{1}{\sup_f \left(\frac{f^T \mathbf{S}_T f}{f^T \mathbf{S}_W f} - 1 \right) + 1} \\ &= 1 - \frac{1}{\lambda_{LDA} + 1} = \frac{\lambda_{LDA}}{\lambda_{LDA} + 1}. \end{aligned} \quad (34)$$

It is clear from Eq. (34) as well as Eq. (30) that the objective function for CCA for the WCCCA setup maximizes the objective function of LDA given in Eq. (15). An interesting way to state this is that using the best ratio, λ_{LDA} , for the Rayleigh quotient that LDA maximizes in Eq. (15), we can directly calculate the canonical correlation maximized in Eq. (7). The same conclusion can be drawn from Eq. (32). This is important in that it gives us a mechanism to judge the quality of a projection LDA finds from a correlation coefficient scale (see Section 4.2).

4.1. Toy problem demonstration

For a simple demonstration, we created a dataset with two variables ($d = 2$) shown in Fig. 1(a). The dataset has two classes ($m = 2$) with 100 samples in both classes ($N_1 = N_2 = 100$). The samples in the classes are generated with multivariate random distributions both with covariance identity and with class 1 having its mean at $\mu_1 = (0, 0)$ and class 2 at $\mu_2 = (2, 2)$. The full set of $2 \times 100^2 = 20,000$ pairs of samples are generated. CCA is then asked to produce maximally correlated functions of these two ‘‘views’’.

Shown in Fig. 1(a) is the discriminant learnt by WCCCA, which is identical to that by LDA. We see in Fig. 1(b) that the samples of classes 1 and 2 are clearly separated when projected on the dimension that WCCCA (and also LDA) found.

4.2. Some advantages of such an indirect computation of LDA using CCA

There are many advantages of the WCCCA way of computing LDA that we could identify, which might not be limited to the following.

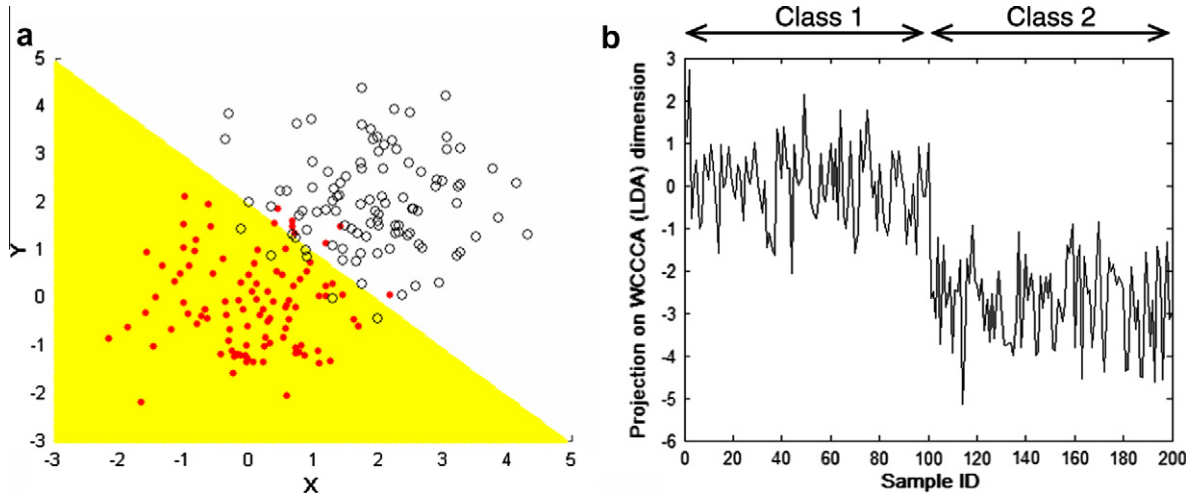


Fig. 1. (a) The found LDA and WCCCA dimensions perfectly coincide. (b) The projection of the samples on the found WCCCA dimension; the first 100 (class 1) samples are generated with $\mu_1 = (0, 0)$ and the second 100 (class 2) samples are generated with $\mu_2 = (2, 2)$.

4.2.1. Computational flexibility

CCA gives an alternative way of performing LDA, which might be a more plausible way of reaching the goal of LDA under some architectures. For example, in the field of neuroscience, a number of researchers have proposed that cerebral cortical neurons perform mutual information extraction from different, but etiologically related sets of inputs, which might be akin to CCA (Favorov and Ryder, 2004; Kording and Konig, 2000; Phillips and Singer, 1997). That is, neural processing units (such as individual dendrites, single neurons, or local neuronal populations) might take advantage of contextual guidance they receive from their lateral inputs to learn to extract such features from their sensory spatio-temporal input patterns that will possess deeper, more inferentially powerful meaning. In other words, neurons will learn to discriminate classes of sensory events that are causally significant. Such an LDA analysis could be accomplished by neurons using a form of the proposed *samples versus samples* CCA.

As another example, taking advantage of robust CCA implementations, robust LDA computations can also be achieved; as a specific illustration, consider SVM-2K (Farquhar et al., 2005). SVM-2K is a method that combines kernel CCA (KCCA) followed by an SVM classification into a single optimization problem and accomplishes a robust form of KCCA. SVM-2K can be given the views \mathbf{D}_1 and \mathbf{D}_2 to get a robust kernel-LDA followed by an SVM classification at one-shot. Yet another extension of WCCCA could be for achieving a form of local LDA (Sugiyama, 2007). This can be done by creating pairs of samples from the same classes in some neighborhood of each other.

4.2.2. Ease of representation

The class labels (or the dummy matrix) could be difficult to represent or store under certain architectures but the data samples could be more readily available in the input channels (for example the observation at time $t - 1$ could be paired with the current observation at time t). Moreover, having a dataset with thousands of different classes would make the dummy matrix with 1-of- C coding impractical. However, the samples versus samples idea, using an online implementation of CCA (or mutual information maximization models) (Becker, 1999; Favorov and Ryder, 2004; Favorov et al., 2003; Fyfe, 2005; Lai and Fyfe, 1999), could accomplish the task of LDA easier because, in fact, not all but some of the pairs of within-class samples (or estimates for the class centers) will be sufficient to learn a good approximation to LDA as shown in Section 4.3.

4.2.3. Independence from class label information

The categorical class labels may simply be unavailable explicitly. The class labels might be embedded into the spatial and/or temporal patterns of the data such as in video and speech processing, in image segmentation, etc. For example, consider a movie of faces such that the consecutive frames are more likely to be of the same individual unless a “scene change” occurs. There are no class labels given explicitly, therefore it is impossible to apply LDA but CCA can be applied (see Section 5).

4.2.4. Correlation coefficient as a tool for evaluation of the LDA projections

The (maximal) correlation coefficient calculated by the CCA algorithm from a training set is typically an overestimate and it is traditional to verify its dependability in terms of how correlated the found covariates on a validation or a test set. However, when using the LDA transform, its projections are given to a subsequent classifier and LDA’s quality is measured in terms of correct classification rate of the classifier applied on the data set. Here, in this section, we show that the quality of LDA projections can be measured without a need for a subsequent classifier (see Section 5.1 and Table 1), again in terms of the correlation coefficient borrowed from the WCCCA equivalence.

It follows from Eqs. (32) and (33) that the LDA correlation is also the *variance explained* between the class labels and the input variables because it is the square of the correlation coefficient produced by SLCCA (the samples versus class labels CCA):

$$\rho_{LDA} = \rho_{WCCCA} = \rho_{SLCCA}^2. \quad (35)$$

Thus, the quality of an LDA projection h on the training set can be calculated using \mathbf{D}_1 and \mathbf{D}_2 sets obtained from the training set \mathbf{D} (as in Eq. (20)) by:

$$\rho_{LDA} = \text{corr}(\mathbf{D}_1 h, \mathbf{D}_2 h). \quad (36)$$

When applied to a test set, Eq. (36) may also be used as an *approximation* to the LDA quality on the test set. However, more generally, on any given dataset \mathbf{D} , whether it is a training or a test set, with the given labels \mathbf{L} (1-of- C coded as in Eq. (17)), the quality of an LDA projection h is given by:

$$\rho_{LDA} = \text{corr}(\mathbf{D}h, \mathbf{L}Mh), \quad (37)$$

where \mathbf{M} is a $m \times d$ matrix that holds the d -dimensional class center vectors, μ_c , of the training set for all the m classes.

Table 1

The quality of the WCCCA projections extracted using 20 random pairs from each individual.

Projection #	Training set		Test set of known individuals		Test set of unknown individuals	
	Canonical correlation	LDA correlation	Canonical correlation	LDA correlation	Canonical correlation	LDA correlation
1	0.97	0.97	0.95	0.96	0.92	0.87
2	0.97	0.97	0.92	0.92	0.47	0.37
3	0.96	0.96	0.89	0.91	0.69	0.80
4	0.95	0.94	0.83	0.83	0.77	0.80
5	0.93	0.92	0.90	0.92	0.75	0.70
6	0.90	0.89	0.88	0.87	0.76	0.72
7	0.90	0.89	0.88	0.88	0.87	0.78
8	0.89	0.87	0.78	0.81	0.63	0.57
9	0.87	0.86	0.77	0.82	0.73	0.86
10	0.84	0.83	0.73	0.57	0.76	0.78
11	0.84	0.83	0.78	0.81	0.93	0.91
12	0.82	0.81	0.55	0.74	0.35	0.42
13	0.81	0.79	0.68	0.67	0.18	0.05
14	0.79	0.76	0.79	0.71	0.82	0.76
15	0.77	0.71	0.61	0.68	0.38	0.75
16	0.74	0.72	0.65	0.63	0.61	0.59
17	0.70	0.70	0.63	0.66	0.57	0.03
18	0.69	0.66	0.70	0.72	0.54	0.25
19	0.65	0.62	0.61	0.65	0.36	0.39
20	0.63	0.59	0.49	0.54	0.06	0.32
21	0.60	0.57	0.56	0.48	0.27	0.21
22	0.55	0.49	0.05	0.07	0.26	0.51
23	0.50	0.47	0.22	0.31	0.40	0.58
24	0.49	0.47	0.22	0.25	0.42	0.58
25	0.44	0.37	0.47	0.46	0.46	0.38

4.3. The resolutions for the disadvantages of WCCCA

Obviously, there are some limitations and disadvantages associated with the WCCCA method of computing LDA. As for its (implicit) limitations, the assumptions of CCA, namely multivariate normality and outliers, must hold in the data; otherwise, violations of these assumptions would bias the algorithm. However, multivariate normality assumption can be overcome using kernels (see Section 4.4) and there are regularization and robust estimation techniques that can apply to CCA and give us flexibility in handling outliers (Hastie et al., 1995; Glendinning and Herbert, 2003; Kursun and Favorov, 2010). In the following, we discuss, in detail, some other implementational disadvantages of WCCCA and provide procedures to overcome them.

4.3.1. Complexity

As opposed to LDA, creating all the within-class pairs of samples results in squaring the number of samples presented subsequently to CCA. However, this complexity can be overcome because WCCCA can work well even when only a small subset of pairs are used. To quantify the quality of the LDA approximation by WCCCA using fewer pairs than all the possible ones, we have ranged the number of pairs used for the experiment given in Section 4.1. We have randomly taken p of all 10,000 pairs available for each class, thus, presenting a total of $2p$ pairs to CCA. The plot of p versus the angle θ between the LDA and the WCCCA projection vectors, h and f , (see Eq. (38)) is shown in Fig. 2 (reported angles in degrees are the averages of 100 runs for each value of p).

$$\theta = \arccos\left(\frac{f^T h}{\|f\| \cdot \|h\|}\right). \quad (38)$$

Fig. 3 shows the average test errors and the standard deviations of the found LDA and WCCCA projections over 100 test runs for each value of p . Each test set is generated the same way that the training data has been generated. We see that WCCCA does not require all the pairs of within-class samples in order to attain the quality of the optimal LDA projection, because the WCCCA error on the test

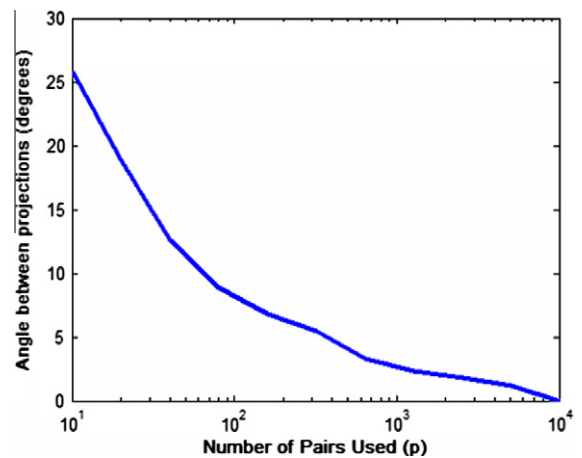


Fig. 2. The number of pairs used for each class p (in log scale) versus the quality of WCCCA approximation to LDA (the angle θ , in degrees, between the found projections by LDA by WCCCA).

set very quickly reduces down to that of LDA as the number of pairs increases.

Moreover, it is straightforward to show that using the class centers (the average of all samples of each class) as the second view in a *samples versus class centers* CCA basis is also equivalent to LDA. For example, a form of *running average* could be used to estimate the class centers in such a setup to reduce the complexity back to that of SLCCA.

4.3.2. Imbalance of class priors

Using the within-class coupling, classes with higher prior probabilities would have many more pairs of their samples than the pairs of classes with lower priors due to the squaring effect. Thus, simply taking all within-class pairs would be an unfair modification of the prior probabilities of the classes that would bias the LDA approximation by WCCCA. This issue can be easily resolved by preserving the prior probabilities of classes also in their pairs

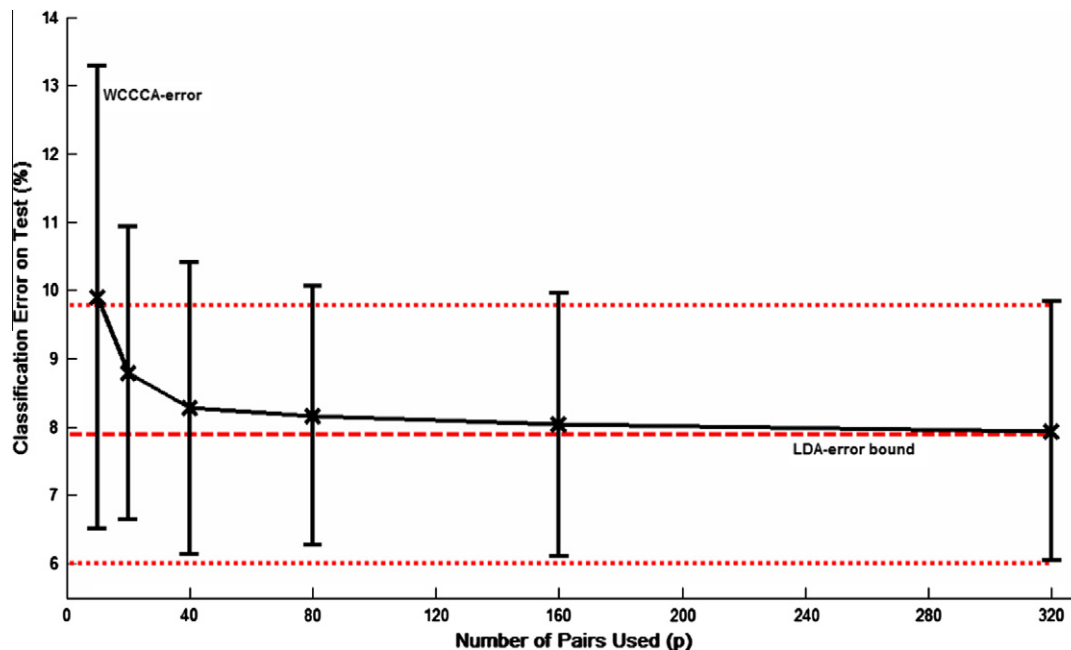


Fig. 3. The number of pairs used for each class p versus the test error of the projections (showing both averages and standard deviations of 100 independent random test sets, each generated with the distribution of the training set). The test error of WCCCA quickly approaches to that of LDA.

created into D_1 and D_2 . While generating all the within-class pairs, the pairs of the classes with lesser prior probabilities must be appropriately proliferated (repeated). Recall the example in Section 4.1 with N_1 equals to $2 \times N_2$ and suppose that the true equivalence to LDA is sought using WCCCA. In this case, as the total number of distinct pairs of the first class would be four times that of the second class, each pair of the second class should be generated twice. Besides, the theoretical basis for this adjustment, our numerical simulations have verified that the WCCCA and LDA give exactly the same projections also in this case.

4.4. Kernel-WCCCA for approximating kernel-LDA

We present a straightforward nonlinear extension of WCCCA using the kernels. When the classes are not linearly separable, the kernels have been used efficiently that enable linear methods calculate nonlinear discriminants (Melzer et al., 2003; Alpaydin, 2004; Haroon et al., 2004; Shawe-Taylor and Cristianini, 2004; Gonen and Alpaydin, 2010). CCA can be asked, as before, to produce correlated outputs using the kernel matrices (whose respective rows preserve the class label information) as the two views (Haroon et al., 2004; Melzer et al., 2003). Fig. 4 shows the results on a known example of a decision boundary using the polynomial kernel of degree two on a dataset with two Gaussians with means $\mu_1 = [0, 0]$ and $\mu_2 = [0, 4]$ and with the covariance matrices $\sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\sigma_2 = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$.

5. Experimental results

In this section, we will present the results obtained using WCCCA on two sets of experiments on the AT&T (ORL) face database, which is composed of 400 grayscale images obtained from 40 different individuals, ten different images per person. The images were taken at different times, varying the lighting, the viewing angle (frontal or more or less semi-frontal view), facial expressions (open or closed eyes, smiling or not, etc.), and other facial details (e.g., with or without glasses). All images were taken

against a dark homogeneous background with the subjects in an upright position, with tolerance for some side movement. Each original image was 92×112 pixels, with 256 gray levels. To reduce the computational load, we down-sampled each image to 23×28 (=644) pixels by bilinear interpolation.

5.1. Supervised LDA by WCCCA on face images

Among the total of 40 individuals (classes), we left out (a random) five of the individuals for testing. This test set that we called the test set of unknown individuals, is a difficult one because it consists of individuals not used for training. We also left out two pictures of each one of the 35 training individuals for a test of known individuals in order to measure the generalization of the methods to the known individuals. Therefore, we used a total of $35 \times 8 = 280$ pictures for learning WCCCA (and LDA) projections. To avoid the computational instability of using 644 dimensional covariance matrices obtained by such a low number of samples, we first performed PCA dimensionality reduction to 50 components that preserved around 91% of the total variation in the original 644 dimensions.

We formed 20 random pairs for each class out of the 35 classes in the training set. As each class has eight training samples, the total number of possible pairs is 64, or 56 if excluding the pairs of an image with itself, but we choose using a random subset basis (Lee and Huang, 2007; Lee and Mangasarian, 2001) to lessen the computational loading as described in Section 4.3.1. Then we applied CCA to get the most interesting WCCCA projections. As a demonstration, the top three WCCCA and LDA projections are shown in Fig. 5b–d, respectively. The eigenvectors of WCCCA and LDA are practically identical up to their signs. To quantify the difference caused by the approximation to LDA by WCCCA, in Table 1, we tabulated the correlation coefficient of the covariates on all the datasets (the training set, the test set of known individuals, and the test set of unknown individuals). For a comparison, we also reported the correlation of the LDA projections, ρ_{LDA} , according to Eq. (36) (i.e. refer to the advantage number four in Section 4.2). Although WCCCA used only 20 pairs of pictures from each individual, the correlations on all sets are practically the same.

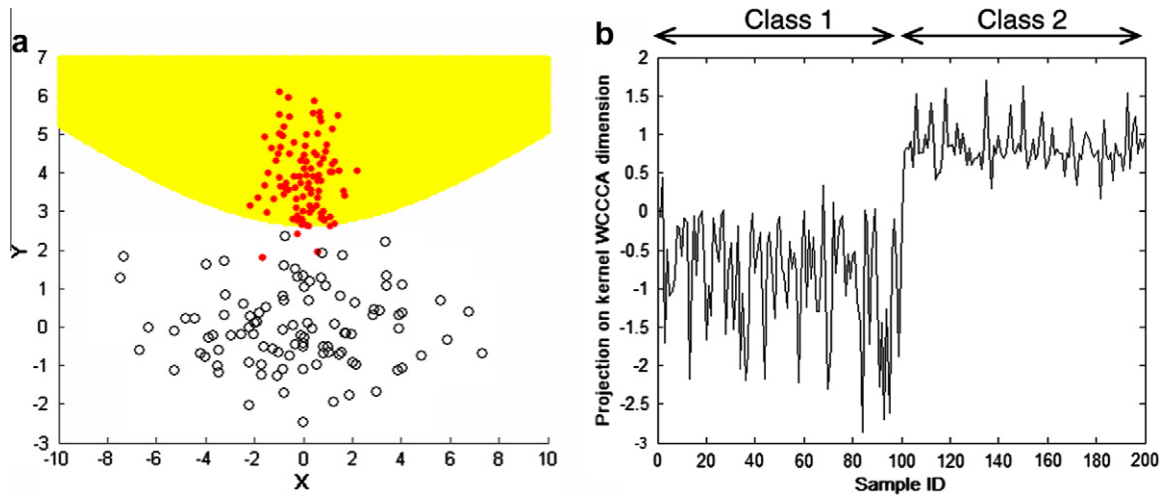


Fig. 4. The nonlinear extension of WCCCA using polynomial kernel of degree two. (a) The found kernel LDA and kernel WCCCA dimensions perfectly coincide. (b) The projection of the samples on the found kernel WCCCA dimension.

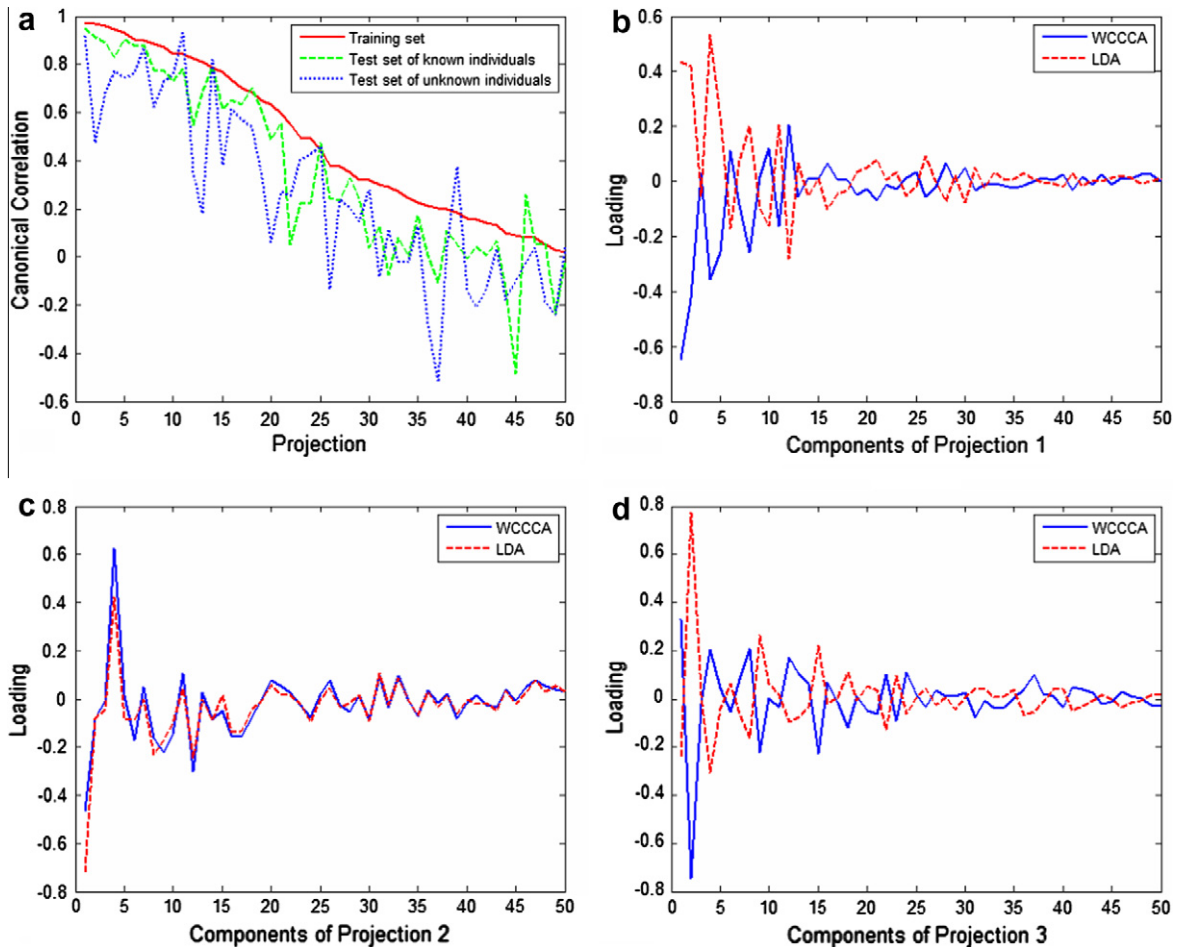


Fig. 5. The results of a representative run of WCCCA on the face images with explicit class labels. (a) Plot of canonical correlations of the found projections (on the training set shown with solid line, on the test set of known individuals with dashed line, and on the test set of unknown individuals with dotted line). (b–d) The top three eigenvectors for WCCCA (shown with solid line) and LDA (shown with dashed line).

When the images are projected on the found WCCCA covariates (and LDA dimensions to compare with), as expected, the images that belong to the same individuals have similar projections but the projections vary from individual to individual (Fig. 6). For

example, when the images are projected on the first covariate, the average of the standard deviations of the projections for pictures of the same individuals is only 0.16 ± 0.07 . For the second and third covariates, these averages are also low, 0.19 ± 0.05 and

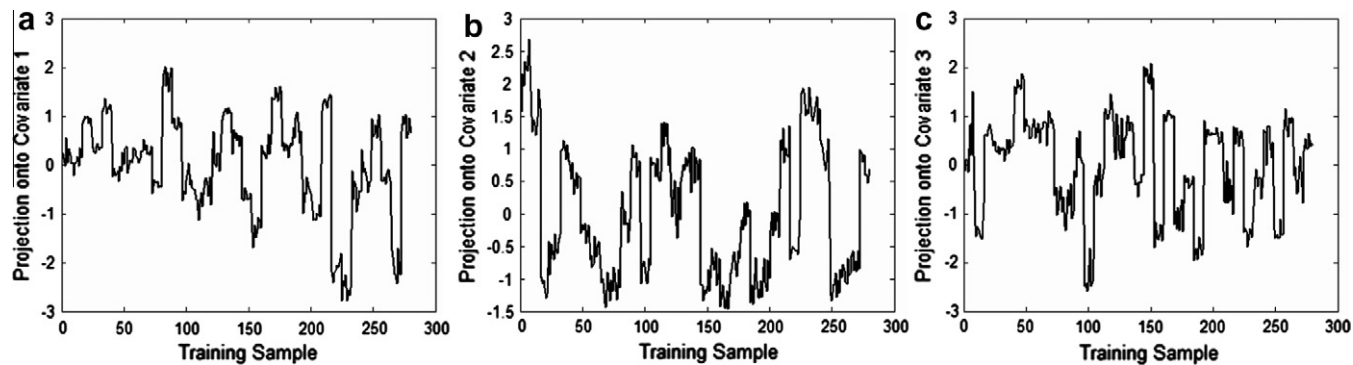


Fig. 6. The projections of the training images (all the eight pictures of each individual are organized to take place consecutively in the dataset, thus on the X-axis) onto the top three covariates found by WCCCA.

0.21 ± 0.10 , respectively. Given that each covariate vector is normalized such that the standard deviation of the projections of all the training samples onto the covariate is equal to 1.0, the variation is due to the between-class separation of the data by the found projection vectors.

5.2. Unsupervised LDA by WCCCA on a movie of face images

The ORL face dataset is made into a movie in a way that the consecutive frames are more likely to be of the pictures of the same

individuals rather than different. When a scene change occurs, the movie continues with the pictures of another individual and so on.

With this movie dataset, there are no class labels given explicitly, therefore it is impossible to directly apply LDA but our WCCCA can be applied to get the interesting (individual discriminatory) projections similar to those found in Section 5.1. To accomplish this, every two consecutive frames (frames at time t and $t + 1$) can be used as the respective samples of the two views. That is, we are asked to produce correlated outputs for the different images

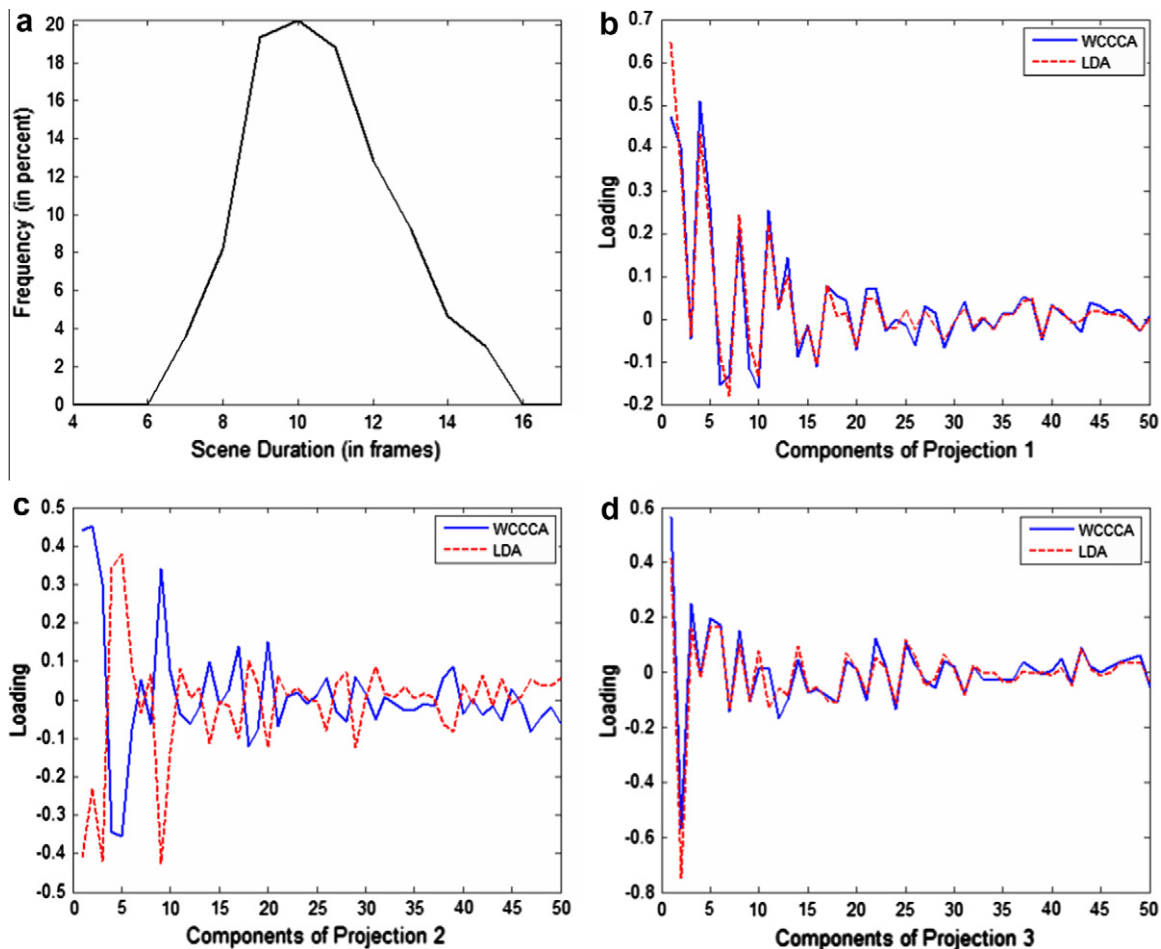


Fig. 7. (a) Distribution of the scene duration in the movie. (b–d) The top three eigenvectors for both methods. WCCCA projections learnt from the movie without explicit class labels (shown with solid lines) are approximately equivalent, up to their signs, to the projections LDA learnt from the labeled image dataset (shown with dashed lines).

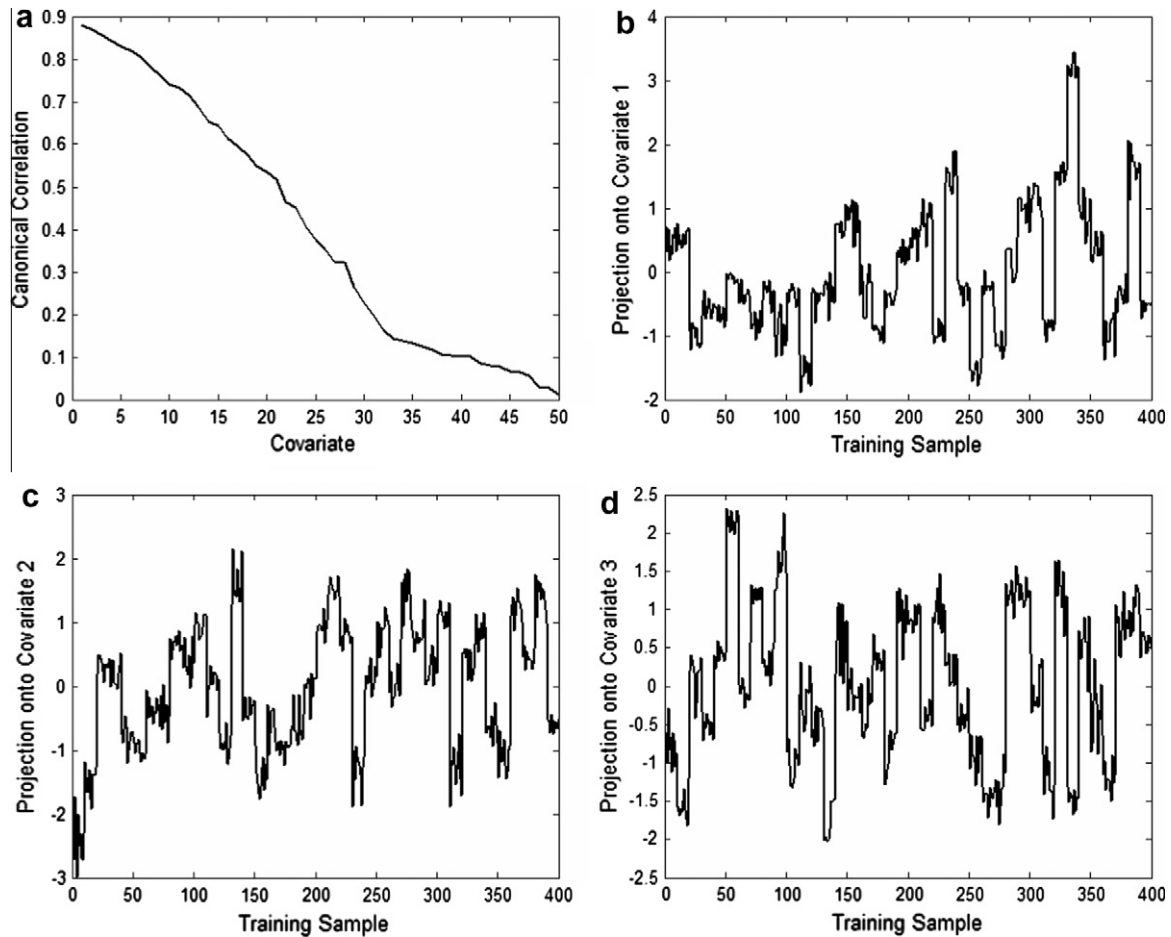


Fig. 8. WCCCA results on the face movie without any explicit class labels. (a) The canonical correlations (between the functions on learnt from consecutive frames in the movie). (b–d) The projections of all the training pictures on the first three covariates found (all the ten pictures of each individual are organized to take place in the dataset consecutively and as a result, the obtained step-like responses are satisfactory).

in consecutive frames. We randomly chose the length of a scene (i.e. the number of frames during which the pictures of the same individual to be played) from a normal-like distribution, shown in Fig. 7a, ranging between 6 and 15 with its mean around 10.5 frames. After a scene completes, we chose another individual to continue the movie for another randomly picked scene duration, and so on. For our experiments, we created a sequence of 8000 frames such that, as in Section 5.1, each image is shown in the movie with an expectation of 20 times. With 8000 frames we had 7999 pairs of consecutive frames; however, as a pair of consecutive frames could be used for both views in a symmetric fashion (i.e. frame t versus frame $t + 1$, and in addition, frame $t + 1$ versus frame t can be used as for the two views), we obtained a total of 15998 training pairs for WCCCA. As in Section 5.1, to avoid computational instability, we performed PCA to reduce dimensionality from 644 pixels down to 50 principal components, which preserved 90.17% of the total variation in the original 644 dimensional data. The WCCCA projections found are practically identical up to their signs with the projections LDA would find only if the class labels were made explicitly available. Fig. 7b–d shows the top three WCCCA projections, respectively.

The canonical correlations found by WCCCA are shown in Fig. 8(a). In Fig. 8b–d, we show that, as in Section 5.1, WCCCA learns to produce similar projections for the images of same individuals but different responses for the images of different individuals. Each covariate (projection vector) is scaled so as that the standard deviation of the projections of the whole training set is

1.0. For the first covariate, the average of the standard deviations of the projections of the pictures of the same individuals is 0.17 ± 0.06 ; for the second covariate, this average of the deviations within individuals is 0.21 ± 0.07 ; and it is 0.22 ± 0.09 for the third covariate (these numbers are very close to those found in Section 5.1 on the supervised dataset). These results show that from the movie dataset in which no explicit class labels were presented, individual-discriminatory representations (features) have been extracted by WCCCA. Moreover, using a few of such features, an unsupervised-LDA based coding of the currently viewing frame could be obtained.

6. Conclusions

Fisher's linear discriminant analysis (LDA) has two main goals: (1) minimize the within-class variance, and (2) maximize the between-class variance. LDA has been long known to be a special case of Hotelling's canonical correlation analysis (CCA). That is, CCA can be performed on a view that constitutes of samples (predictive features) versus a second view that is directly made up of the class labels of the samples in order to obtain projections that are identical to those of LDA. In this paper, it has been shown that CCA can perform LDA also when it is applied on a *samples versus samples* basis, which can be viewed as accomplishing LDA through a rather indirect and distributed style of an implicit presentation of the categorical class labels. More specifically, in the proposed WCCCA method,

each one of the samples of a class, serving as the first view, is paired with every other samples of that class serving as the second view. We prove that when CCA is asked to find correlated functions between these two views, it yields the LDA projections as expected because it is the LDA projections that minimize the within-class variance. As LDA has this as its first goal, samples from the same class presented to CCA as pairs would give similar scores when projected on the dimensions found by LDA. Likewise, if samples that belonged to different classes gave different outputs that would also help maximize the canonical correlation. Thus, WCCCA can also be said to aim the maximization of the between-class variance, which is already the second one of the LDA goals. This equivalence and the application of WCCCA can be particularly useful when the class labels, rather than being explicitly available, can be tracked down in the temporal and/or spatial patterns of the data, such as for the tasks of splitting a video into scenes (sequences of relevant frames), segmentation of an image into image regions sharing certain visual characteristics, speech analysis, or biological sequence analysis. In such settings, the use of LDA is not possible forthright. It has been known that it is possible to perform a CCA analysis between, say, the consecutive frames of a video for searching correlations between a frame and the next; the equivalence proof presented in this paper assures that the use of CCA in this manner would yield discriminatory features favorable for the subsequent learning tasks such as classification, segmentation, or clustering, without any need for explicit supervised class memberships. We have also demonstrated that WCCCA can be easily generalized to its nonlinear version via the kernel trick. It is straightforward to extend the equivalence proofs given for the linear case to the kernel version.

References

- Alpaydin, E., 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning Series). The MIT Press.
- Barker, M., Rayens, W., 2003. Partial least squares for discrimination. *J. Chemometr.* 17, 166–173.
- Bartlett, M.S., 1938. Further aspects of the theory of multiple regression. *Proc. Cambridge Philos. Soc.* 34, 33–40.
- Becker, S., 1999. Implicit learning in 3d object recognition: The importance of temporal context. *Neural Comput.* 11, 347–374.
- Borga, M., Knutsson, H., 2001. A Canonical Correlation Approach to Blind Source Separation. Technical Report LiU-IMT-EX-0062. Department of Biomedical Engineering, Linköping University, Sweden.
- Borga, M., 1998. Learning Multidimensional Signal Processing, Ph.D. Thesis. Department of Electrical Engineering, Linköping University, Linköping, Sweden.
- Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J., Szedmak, S., 2005. Two view learning: SVM-2K, theory and practice. In: Proceedings of NIPS.
- Favorov, O.V., Ryder, D., 2004. SINBAD: A neocortical mechanism for discovering environmental variables and regularities hidden in sensory input. *Biological Cybernet.* 90, 191–202.
- Favorov, O.V., Ryder, D., Hester, J.T., Kelly, D.G., Tommerdahl, M., 2003. The cortical pyramidal cell as a set of interacting error backpropagating networks: A mechanism for discovering nature's order. In: Hecht-Nielsen, R., McKenna, T. (Eds.), *Theories of the Cerebral Cortex*. Springer, London, pp. 25–64.
- Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenetic.* 7, 179–188.
- Fyfe, C., 2005. *Hebbian Learning and Negative Feedback Networks*. Springer.
- Gestel, T.V., Suykens, J.A.K., De Brabanter, J., De Moor, B., Vandewalle, J., 2001. Kernel canonical correlation analysis and least squares support vector machines. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN 2001), pp. 384–389.
- Glendinning, R.H., Herbert, R.A., 2003. Shape classification using smooth principal components. *Pattern Recognition Lett.* 24 (12), 2021–2030.
- Gonen, M., Alpaydin, E., 2010. Cost-conscious multiple kernel learning. *Pattern Recognition Lett.* 31 (9), 959–965.
- Hardoon, D., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* 16, 2639–2664.
- Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. *Ann. Statist.* 23 (1), 73–102.
- Hotelling, H., 1936. Relations between two sets of variates. *Biometrika* 28, 321–377.
- Izenman, A.J., 2008. *Modern Multivariate Statistical Techniques*. Springer.
- Johansson, B., 2001. On Classification: Simultaneously Reducing Dimensionality and Finding Automatic Representation Using Canonical Correlation. Technical Report LiTH-ISY-R-2375. ISSN 1400-3902, Linköping University.
- Kording, K.P., Konig, P., 2000. Learning with two sites of synaptic integration. *Network: Comput. Neural Systems* 11, 25–39.
- Kursun, O., Alpaydin, E., 2010. Canonical correlation analysis for multiview semisupervised feature extraction. In: L. Rutkowski et al. (Eds.), Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2010) Part I. Springer, Poland, pp. 216–223.
- Kursun, O., Favorov, O.V., 2010. Feature selection and extraction using an unsupervised biologically-suggested approximation to Gebelein's maximal correlation. *Internat. J. Pattern Recognition Artif. Intell.* 24 (3), 337–358.
- Lai, P.L., Fyfe, C., 1999. A neural network implementation of canonical correlation. *Neural Networks* 12 (10), 1391–1397.
- Lee, Y.J., Huang, S.Y., 2007. Reduced support vector machines: A statistical theory. *IEEE Trans. Neural Networks* 18, 1–13.
- Lee, Y.J., Mangasarian, O.L., 2001. RSVM: Reduced support vector machines. In: *Proceedings of the First SIAM International Conference on Data Mining, Chicago*.
- Loog, M., van Ginneken, B., Duin, R.P.W., 2005. Dimensionality reduction of image features using the canonical contextual correlation projection. *Pattern Recognition* 38, 2409–2418.
- Melzer, T., Reiter, M., Bischof, H., 2003. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition* 36, 1961–1971.
- Phillips, W.A., Singer, W., 1997. In search of common foundations for cortical computation. *Behav. Brain Sci.* 20, 657–722.
- Phillips, W.A., Kay, J., Smyth, D., 1995. The discovery of structure by multi-stream networks of local processors with contextual guidance. *Network: Comput. Neural Systems* 6, 225–246.
- Raudys, S., Duin, R.P.W., 1998. Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Lett.* 19 (5–6), 385–392.
- Rencher, A.C., 1997. *Multivariate Statistical Inference and Applications*. Wiley-Interscience.
- Shawe-Taylor, J., Cristianini, N., 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.
- Stone, J., 1996. Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Comput.* 8, 1463–1492.
- Sugiyama, M., 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *J. Machine Learn. Res.* 8, 1027–1061.
- Sun, T., Chen, S., 2007. Class label versus sample label-based CCA. *Appl. Math. Comput.* 185, 272–283.