Engineering Applications of Artificial Intelligence 22 (2009) 40–47

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Affine feature extraction: A generalization of the Fukunaga–Koontz transformation ☆

Wenbo Cao *, Robert Haralick

Pattern Recognition Laboratory, CS Department, The Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY 10016, USA

ARTICLE INFO

Article history:
Received 19 April 2008
Accepted 26 April 2008
Available online 25 June 2008

Keywords:
Feature extraction
Affine transformation
Fisher's discriminant analysis
Fukunaga–Koontz transformation
Kullback–Leibler divergence

ABSTRACT

Dimension reduction methods are often applied in machine learning and data mining problems. Linear subspace methods are the commonly used ones, such as principal component analysis (PCA), Fisher's linear discriminant analysis (FDA), common spatial pattern (CSP), et al. In this paper, we describe a novel feature extraction method for binary classification problems. Instead of finding linear subspaces, our method finds lower-dimensional affine subspaces satisfying a generalization of the Fukunaga–Koontz transformation (FKT). The proposed method has a closed-form solution and thus can be solved very efficiently. Under normality assumption, our method can be seen as finding an optimal truncated spectrum of the Kullback–Leibler divergence. Also we show that FDA and CSP are special cases of our proposed method under normality assumption. Experiments on simulated data show that our method performs better than PCA and FDA on data that is distributed on two cylinders, even one within the other. We also show that, on several real data sets, our method provides statistically significant improvement on test set accuracy over FDA, CSP and FKT. Therefore the proposed method can be used as another preliminary data-exploring tool to help solve machine learning and data mining problems.

1. Introduction

Because of the curse of dimensionality and the concern of computational efficiency, dimensionality reduction methods are often used in machine learning and data mining problems. Examples are face recognition in computer vision (Belhumeur et al., 1997; Yang et al., 2002), electroencephalogram (EEG) signal classification in brain–computer interface (BCI) (Dornhege et al., 2004; Ramoser et al., 2000) and microarray data analysis (Dai et al., 2006). Linear subspace methods have been widely used for the purpose of dimension reduction.

Linear subspaces are affine spaces that contain the origin. In this study, we discuss a novel affine feature extraction (AFE) method to find affine subspaces for classification. Our method can be seen as a generalization of the Funkunaga–Koontz transformation (FKT) (Fukunaga and Koontz, 1970). We investigate the information–theoretical properties of our method and study the relationship of AFE and other similar feature extraction methods.

Our paper is organized as follows: in Section 2, we briefly review some subspace methods. In Section 3, we present the main result of our work: the motivation of the study, the AFE method and its closed-form solutions. We investigate the information–theoretical properties of AFE and the relationship of AFE with other linear subspace dimension reduction methods in Section 4. We present experimental results in Section 5, and conclude the study with the summary of our work, and possible future directions in Section 6.

2. Subspace methods

Principal component analysis (PCA) and independent component analysis (ICA) are unsupervised linear subspace methods for dimension reduction. PCA tries to find linear subspaces such that the variance of the projected data are maximally preserved. ICA is a way of finding linear subspaces in which the second- and higher-order statistical dependencies of the data are minimized; that is the transformed variables are as statistically independent from each other as possible. Note that, as unsupervised methods, neither PCA nor ICA use label information, which is crucial for classification problems. Consequently, PCA and ICA are optimal for pattern description, but not optimal for pattern discrimination.

Fisher's discriminant analysis (FDA) determines linear subspaces in which the distance between the means of the classes is maximized and the variance of each class is minimized at the

☆ A preliminary version of this paper appeared in MLDM 2007 (Cao and Haralick, 2007).

* Corresponding author.
E-mail addresses: wcao@gc.cuny.edu (W. Cao), haralick@ptah.gc.cuny.edu (R. Haralick).

0952-1976/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2008.04.008

same time. An important drawback of FDA is that, for K-class classification problems, it can only find $K-1$ dimensional subspaces. This becomes more serious when binary classification problems are considered, for which FDA can only extract one optimal feature. Canonical correlation analysis (CCA) is a method for finding linear subspaces to maximize the correlation of the observation vectors and their labels. It has been known for a long time that FDA and CCA indeed give identical subspaces for the dimension reduction purpose (Bartlett, 1938).

Recently there has been some interest in partial least squares (PLS) (Rosipal and Krämer, 2005). Only recently, it has been shown that PLS has a close connection with FDA (Barker and Rayens, 2003). PLS finds linear subspaces by iteratively maximizing the covariance of the deflated observation vectors and their labels. In one mode, PLS can be used to extract more than one feature for binary classification. The main concern in PLS is the efficiency issue, since in each iteration one has to subtract the observation matrix by its rank-one estimation found in the previous iteration, and generate deflated observation vectors.

3. Affine feature extraction

Consider a binary classification problem, which is also called *discriminant analysis* in statistics. Let $\{(\mathbf{x}_j, g_j) \in \mathbb{R}^m \times \{1,2\} | j = 1,2,\ldots,N\}$ be a training set. $\mathbf{x}_j$ and g_j are the *observation vector* and the corresponding *class label*. For simplicity, we assume the training set is permuted such that observations 1 to N_1 have label 1, and observations N_1+1 to N_1+N_2 have label 2. Define a *data matrix* as

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = (\mathbf{X}_1, \mathbf{X}_2),$$

where $\mathbf{X}_1 = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_1})$, and $\mathbf{X}_2 = (\mathbf{x}_{N_1+1}, \mathbf{x}_{N_1+2}, \ldots, \mathbf{x}_N)$. For the convenience of future discussion, we define *augmented observation vectors* as

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix}. \tag{1}$$

We can similarly define an *augmented data matrix* $\mathbf{Y}_i$ for class i as $\mathbf{Y}_i^{\mathrm{T}} = (\mathbf{X}_i^{\mathrm{T}}, \mathbf{1})$. Throughout this paper, we use the following conventions: (1) vectors are column vectors; (2) $\mathbf{1}$ is a vector of all ones; (3) $\mathbf{I}$ is an identity matrix; (4) $\square^{\mathrm{T}}$ is the transpose of a vector or matrix $\square$; and (5) $\mathrm{tr}(\square)$ is the trace of a matrix $\square$.

3.1. Background

In this subsection, we give a brief introduction of dimension reduction for classical discriminant analysis. Due to the limitation of space, we cannot provide complete details for classical discriminant analysis. We refer to Section 4.3 of Hastie et al. (2001) for a nice treatment on this topic. This subsection also serves as our motivation to carry on this study.

Before going on further, let us define the sample *mean*, *covariance* and *second order moment* for class i as follows:

$$\text{mean } \hat{\boldsymbol{\mu}}_i = \frac{1}{N_i}\mathbf{X}_i\mathbf{1}, \tag{2}$$

$$\text{moment } \hat{\mathbf{M}}_i = \frac{1}{N_i}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}, \tag{3}$$

$$\text{covariance } \hat{\boldsymbol{\Sigma}}_i = \frac{1}{N_i}\mathbf{X}_i\mathbf{P}\mathbf{X}_i^{\mathrm{T}}, \tag{4}$$

where

$$\mathbf{P} = \mathbf{I} - \frac{1}{N_i}\mathbf{1}\mathbf{1}^{\mathrm{T}}.$$

In classical discriminant analysis, the probability density for each class are usually modeled as multivariate normal distributions, i.e. $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ $(i = 1,2)$. It is also well known that, more generally, elliptically contoured distributions also lead to linear or quadratic decision surfaces (Haralick, 1977). Eqs. (2) and (4) can be seen as the (pseudo-)maximum likelihood estimations of class density parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, respectively. Without losing generality, let us consider how to find a one-dimensional linear subspace for classical discriminant analysis; that is to find a linear transformation for observations:

$$z_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i,$$

where $\mathbf{w}^{\mathrm{T}}$ is a m-dimensional vector.

When the two classes have a common covariance, i.e. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the problem is relatively easy. It is not hard to show that the optimal $\mathbf{w}^*$ is the eigenvector of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}$. FDA essentially capture this situation by solving the following problem:

$$\max \frac{\mathbf{w}^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\mathbf{w}}, \tag{5}$$

where $N\hat{\boldsymbol{\Sigma}} = N_1\hat{\boldsymbol{\Sigma}}_1 + N_2\hat{\boldsymbol{\Sigma}}_2$.

When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, finding an optimal linear subspace is harder. The only known closed-form solution is that $\mathbf{w}^*$ is the eigenvector of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1$, which has the largest eigenvalue. It can be shown that, when $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$, the solution optimizes the Kullback–Leibler KL divergence and the Bhattacharyya distance, (cf. Section 10.2 of Fukunaga, 1990). The KL distance and the Bhattacharyya distance are approximations of the Chernoff distance, which is the best asymptotic error exponent of a Bayesian approach. Therefore the optimizing of these distances serves as the theoretical support to use it as a dimension reduction method. The approach has been widely used in EEG classification problems, namely the common spatial pattern (CSP). Formally speaking, CSP solves the following problem:

$$\max \frac{\mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_1\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_2\mathbf{w}} \quad \text{or} \quad \max \frac{\mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_2\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_1\mathbf{w}}. \tag{6}$$

Therefore, CSP only works well when the difference between the class means is small, i.e. $|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1| \approx 0$. For many classification problems, this restriction is unrealistic. Furthermore, unlike FDA, CSP has no natural geometrical interpretation.

The FKT method can be seen as an extension of CSP by shrinking $\hat{\boldsymbol{\mu}}_i$ to zero. It can be seen as a rough shrinkage estimation of the mean for high dimensional data. FKT solves the following problem:

$$\max \frac{\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_1\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_2\mathbf{w}} \quad \text{or} \quad \max \frac{\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_2\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_1\mathbf{w}}. \tag{7}$$

Taking a closer look at the criterion of FKT, we note that the criterion $\max(\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_1\mathbf{w}/\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_2\mathbf{w})$ can be written as

$$\begin{aligned} &\min \quad \mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_2\mathbf{w} \\ &\text{s.t.} \quad \mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_1\mathbf{w} = 1. \end{aligned}$$

Note

$$\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_i\mathbf{w} = \frac{1}{N_i}\sum_{j=k_i+1}^{k_i+N_i} z_j^2,$$

where $k_1 = 1$, $k_2 = N_1$ and $i = 1,2$. That is: $\mathbf{w}^{\mathrm{T}}\hat{\mathbf{M}}_i\mathbf{w}$ is the mean of square transformed observations, i.e. z_j^2, of class i. Therefore, FKT can be interpreted as finding a linear subspace in which one can maximize the distance of the means of square transformed observations. However, FKT may ignore important discriminant information for some cases, for example, the one proposed in Foley and Sammon (1975).

3.2. Problem formulation

Let $z_i = v_0 + \mathbf{v}_1^T \mathbf{x}_i$ be an affine transformation for observations $\mathbf{x}_i$, where $\mathbf{v}_1$ is a m dimensional vector. Linear transformations are a special form of affine transformations, where $v_0 = 0$. Now denoting $\mathbf{w}^T = (\mathbf{v}_1^T, v_0)$, we have $z_i = \mathbf{w}^T \mathbf{y}_i$. Note that we have abused the notation of $\mathbf{w}$. From now on, we shall use $\mathbf{w}$ for affine transformations unless specified otherwise. Define a sample *augmented second moment* matrix as

$$\hat{\Xi}_i = \frac{1}{N_i} \mathbf{Y}_i \mathbf{Y}_i^T. \tag{8}$$

The relation of augmented second moment matrix, covariance matrix and mean can be found in Appendix A. Motivated by FKT, we use the following objective function to find the optimal one-dimensional affine subspace

$$\max \xi \frac{\mathbf{w}^T \hat{\Xi}_1 \mathbf{w}}{\mathbf{w}^T \hat{\Xi}_2 \mathbf{w}} + (1-\xi) \frac{\mathbf{w}^T \hat{\Xi}_2 \mathbf{w}}{\mathbf{w}^T \hat{\Xi}_1 \mathbf{w}}, \tag{9}$$

where $0 \leqslant \xi \leqslant 1$. We use the sum of ratios to measure the importance of $\mathbf{w}$ instead of two separated optimization problems in FKT. The parameter ξ can be used to balance the importance of different classes and thus is useful for asymmetric learning problems.

Now let us consider how to find higher dimensional affine subspaces. Let $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d)$ be a low-rank affine transformation matrix. Let $\mathbf{z}_i$ be the lower-dimensional representation of $\mathbf{x}_i$, i.e. $\mathbf{z}_i = \mathbf{W}^T \mathbf{y}_i$. We define the following optimization problem to find $\mathbf{W}$:

$$\begin{aligned} \max \xi \quad & \sum_{i=1}^{d} \frac{\mathbf{w}_i^T \hat{\Xi}_1 \mathbf{w}_i}{\mathbf{w}_i^T \hat{\Xi}_2 \mathbf{w}_i} + (1-\xi) \sum_{i=1}^{d} \frac{\mathbf{w}_i^T \hat{\Xi}_2 \mathbf{w}_i}{\mathbf{w}_i^T \hat{\Xi}_1 \mathbf{w}_i} \\ \text{s.t.} \quad & \mathbf{w}_i^T \hat{\Xi}_t \mathbf{w}_j = \delta_{ij}, \end{aligned}$$

where $N\hat{\Xi}_t = N_1 \hat{\Xi}_1 + N_2 \hat{\Xi}_2$, and δ_{ij} is 1 if $i = j$, and 0 otherwise. Let $\hat{\Pi}_i = \mathbf{W}^T \hat{\Xi}_i \mathbf{W}$. It is easy to recognize that $\hat{\Pi}_i$'s are indeed the second moment matrices in the lower dimensional space. Now we can write the problem more compactly: find $\mathbf{W}$ to

$$\begin{aligned} \max \xi \quad & \operatorname{tr}(\hat{\Pi}_1^{-1} \hat{\Pi}_2) + (1-\xi) \operatorname{tr}(\hat{\Pi}_2^{-1} \hat{\Pi}_1) \\ \text{s.t.} \quad & \mathbf{W}^T \hat{\Xi}_t \mathbf{W} = \mathbf{I}. \end{aligned}$$

Generally speaking, we want to generate compact representations of the original observations. Therefore it is desirable to encourage finding lower dimensional affine subspaces. Motivated by the Akaike information criterion and Bayesian information criterion, we propose the following objective function that is to be maximized:

$$C(\mathbf{W}; \xi, d) = (1-\xi) \operatorname{tr}(\hat{\Pi}_2^{-1} \hat{\Pi}_1) + \xi \operatorname{tr}(\hat{\Pi}_1^{-1} \hat{\Pi}_2) - d, \tag{10}$$

where $0 \leqslant \xi \leqslant 1$, d ($1 \leqslant d \leqslant m$) is the number of features we want to generate. We see that high dimensional solutions are penalized by the term $-d$. Hyperparameter ξ may be tuned via standard cross-validation methods (Hastie et al., 2001). In principal, the optimum d can also be determined by cross-validation procedures. However, such a procedure is often computationally expensive. Therefore, we propose the following alternative: define $C_0(\xi) = C(\mathbf{I}; \xi, m)$; we select the smallest d such that C is large enough, i.e. $d^* = \inf\{d | C(\mathbf{W}; \xi, d) \geqslant \gamma C_0\}$, where γ is a constant.

The constraint $\mathbf{W}^T \hat{\Xi}_t \mathbf{W} = \mathbf{I}$ is necessary in our generalization from the one dimensional to the high dimensional formulation, but it does not generate mutually orthogonal discriminant vectors. Obtaining orthogonal discriminant vectors basis is geometrically desirable. Therefore, we introduce another orthogonality constraint $\mathbf{W}^T \mathbf{W} = \mathbf{I}$. We refer to Edelman et al. (1999) for a geometrical view of the roles of the two constraints in optimization problems. To summarize, we are interested in two different kinds of constraints as follows:

(1) $\hat{\Xi}_t$-orthogonal: $\mathbf{W}^T \hat{\Xi}_t \mathbf{W} = \mathbf{I}$;
(2) orthogonal: $\mathbf{W}^T \mathbf{W} = \mathbf{I}$.

3.3. Basic algorithms

In this subsection, we show how to solve the proposed optimization problems. Define the function f as

$$f(x; \xi) = \xi x + (1-\xi) \frac{1}{x}. \tag{11}$$

Let $0 < a \leqslant x \leqslant b$. Note that f is a convex function, and thus achieves its maximum at the boundary of x, i.e. either a or b.

Define $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{m+1})$, and λ_i's are the eigenvalues of $(\hat{\Xi}_1, \hat{\Xi}_2)$ $(i = 1, 2, \ldots, m+1)$, i.e. $\hat{\Xi}_1 \mathbf{u}_i = \lambda_i \hat{\Xi}_2 \mathbf{u}_i$. Let $\lambda_i(\xi)$'s be the ordered eigenvalues of $(\hat{\Xi}_1, \hat{\Xi}_2)$ with respect to $f(\lambda; \xi)$. That is: define $f_i(\xi) = f(\lambda_i(\xi); \xi)$, then we have $f_1(\xi) \geqslant f_2(\xi) \geqslant \cdots \geqslant f_{m+1}(\xi)$. The following lemma for nonsingular symmetric $\hat{\Xi}_1$ and $\hat{\Xi}_2$ can be found in Golub and Van Loan (1996):

Lemma 1. *If $\mathbf{A} \in \mathbb{R}^{k \times k}$ is symmetric, and $\mathbf{B} \in \mathbb{R}^{k \times k}$ is symmetric positive definite, then there exists a nonsingular matrix $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k) \in \mathbb{R}^{k \times k}$ such that $\mathbf{U}^T \mathbf{B} \mathbf{U} = \mathbf{I}$ and $\mathbf{U}^T \mathbf{A} \mathbf{U} = \Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{m+1})$. Moreover, $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{B}\mathbf{u}_i$, i.e. λ_i and $\mathbf{u}_i$ are the generalized eigenvalue and eigenvector of $(\mathbf{A}, \mathbf{B})$. Furthermore, if $\mathbf{A}$ is also positive definite, then $\lambda_i > 0$.*

In Appendix C, we show that:

$$C(\mathbf{W}; \xi, d) \leqslant \sum_{i=1}^{d} f_i(\xi) - d. \tag{12}$$

Remark 2. If $\mathbf{W}_1$ maximizes $C(\mathbf{W}; \xi, d)$, then $\mathbf{W}_1 \mathbf{R}$ also maximizes $C(\mathbf{W}; \xi, d)$, where $\mathbf{R}$ is a nonsingular matrix. The proof is straight forward and therefore is omitted.

Proposition 3. *Let $\mathbf{U}_\xi = (\mathbf{u}_1^\xi, \mathbf{u}_2^\xi, \ldots, \mathbf{u}_d^\xi)$, where $\mathbf{u}_i^\xi$ is the eigenvector of $(\hat{\Xi}_1, \hat{\Xi}_2)$ and has eigenvalue $\lambda_i(\xi)$. Let $\mathbf{R}$ be a nonsingular matrix. Then $\mathbf{W} = \mathbf{U}_\xi \mathbf{R}$ maximize $C(\mathbf{W}; \xi, d)$.*

Proof. It is enough to show $\mathbf{U}_\xi$ maximizes $C(\mathbf{W}; \xi, d)$. Note $\mathbf{U}_\xi^T \hat{\Xi}_2 \mathbf{U}_\xi = \mathbf{I}$ and $\mathbf{U}_\xi^T \hat{\Xi}_1 \mathbf{U}_\xi = \operatorname{diag}(\lambda_1(\xi), \lambda_2(\xi), \ldots, \lambda_d(\xi))$. Then it is easy to affirm the proposition. □

Remark 4. Let $\mathbf{U}_\xi = (\mathbf{u}_1^\xi, \mathbf{u}_2^\xi, \ldots, \mathbf{u}_d^\xi)$ maximize $C(\mathbf{W}; \xi, d)$; let $\mathbf{u}_{d+1}^\xi$ be an eigenvector of $(\hat{\Xi}_1, \hat{\Xi}_2)$ whose eigenvalue is 1. Then it is straightforward to show that $C(\mathbf{U}_\xi; \xi, d) = C((\mathbf{U}_\xi, \mathbf{u}_{d+1}^\xi); \xi, d+1)$. We prefer $\mathbf{U}_\xi$ to $(\mathbf{U}_\xi, \mathbf{u}_{d+1}^\xi)$, because of the lower dimensionality. In other words, we can safely ignore the eigenvectors of $(\hat{\Xi}_1, \hat{\Xi}_2)$ whose eigenvalues are 1.

Let $\mathbf{U}_\xi = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q}$ and $\mathbf{R}$ are the thin QR factorization of $\mathbf{U}_\xi$; then $\mathbf{W}_1 = \mathbf{U}_\xi \mathbf{R}^{-1}$ maximizes $C(\mathbf{W}; \xi, d)$ and satisfies the orthogonal constraint. Let $\mathbf{W}_2 = \mathbf{U}_\xi \Gamma^{-1/2}$, where

$$\begin{aligned} \Gamma = \frac{1}{N} \{ & \operatorname{diag}(N_1 \lambda_1(\xi) + N_2, N_1 \lambda_2(\xi) \\ & + N2, \ldots, N_1 \lambda_d(\xi) + N_2)\}. \end{aligned} \tag{13}$$

It can be easily shown that $\mathbf{W}_2$ maximizes $C(\mathbf{W}; \xi, d)$ and satisfies the $\hat{\Xi}_t$-orthogonal constraint. In practice, we only need to check the largest d and the smallest d eigenvalues and eigenvectors of $(\hat{\Xi}_1, \hat{\Xi}_2)$ in order to generate d features. The pseudo-code of the algorithm is given in Table 1.

Table 1
Pseudo-code for feature extraction

Algorithm for feature extraction
Input: Data sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$
Output: Transformation matrix $\mathbf{W}$
1. Calculate the augment second moment matrices $\hat{\Xi}_1$, and $\hat{\Xi}_2$;
2. Compute the largest d and the smallest d eigenvalues and eigenvectors of $(\hat{\Xi}_1, \hat{\Xi}_2)$;
3. Sort $2d$ eigenvalues and eigenvectors with respect to Eq. (11);
3. Selected the largest d eigenvectors to form $\mathbf{U}_\xi$;
4∗. (For orthogonal constraint) apply the thin QR factorization on $\mathbf{U}_\xi$, i.e. $\mathbf{U}_\xi = \mathbf{QR}$;
5∗. (For orthogonal constraint) Let $\mathbf{W} = \mathbf{Q}$;
6 ∗ ∗. (For $\hat{\Xi}_t$-orthogonal constraint), calculate Γ as Eq. (13);
7 ∗ ∗. (For $\hat{\Xi}_t$-orthogonal constraint), Let $\mathbf{W} = \mathbf{U}_\xi \Gamma^{-1/2}$;

3.4. Computational issues

For convenience, we will not differentiate $\hat{\Xi}_i$'s and Ξ_i's in this subsection unless otherwise specified. For the proposed AFE problems, we need to solve the generalized eigenvalue problem (Ξ_2, Ξ_1). In our derivation, we assume the positive definiteness of Ξ_i's, which may not be satisfied in real applications. The deficiency can be fixed by adding a small regularization matrix to Ξ_i's; that is $\Xi_i \leftarrow \Xi_i + \alpha\mathbf{I}$, where α is a small positive constant. In this paper, we take $\alpha = 10^{-3}$. It is easy to show that, if $\mathbf{w}$ is a generalized eigenvector of (Ξ_2, Ξ_1) with eigenvalue λ, it is also a generalized eigenvector of (Ξ_2, Ξ_t) with eigenvalue β, where

$$\lambda = \frac{N_1\beta}{N - N_2\beta}.$$

Then we can write the Taylor expansion of λ about $\beta = 1$ as follows:

$$\lambda = 1 + \frac{N}{N_1}(\beta - 1) + O((\beta - 1)^2).$$

Hence we have $|\lambda - 1| \approx N|\beta - 1|/N_1$. Solving the generalized eigenvalue problem of (Ξ_2, Ξ_t) is numerically more stable than that of (Ξ_2, Ξ_1). Therefore we recommend solving the generalized eigenvalue problem of (Ξ_2, Ξ_t), and remove eigenvectors whose eigenvalue is near 1, i.e. $|\beta - 1| < \tau'$. From now on, we shall not differentiate β and λ for the sake of simplicity of our argument.

When the dimensionality of observations is high, i.e. $m \gg 1$, solving the generalized eigenvalue problem (Ξ_2, Ξ_t) is not only computationally expensive, but also memory intensive. In the remainder of this subsection, we show an efficient algorithm to overcome the handicap. Let $\mathbf{Y} = \mathbf{U}\,\mathrm{diag}(\Lambda, \mathbf{0})\mathbf{V}^{\mathrm{T}}$ be the SVD of the augmented data matrix $\mathbf{Y}$, where $\Lambda \in \mathbb{R}^{s\times s}$ is a diagonal matrix that contains non-zero singular values of $\mathbf{Y}$, and $\mathbf{U}$ and $\mathbf{V}$ are orthonormal. Then we have $N\Xi_t = \mathbf{U}\,\mathrm{diag}(\Lambda^2, \mathbf{0})\mathbf{U}^{\mathrm{T}}$, and $N\mathbf{U}^{\mathrm{T}}\Xi_t\mathbf{U} = \mathrm{diag}(\Lambda^2, \mathbf{0})$. Let $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{m+1}) = (\mathbf{U}_1, \mathbf{U}_2)$, such that $\mathbf{U}_1 \in \mathbb{R}^{(m+1)\times s}$ contains singular vectors with nonzero singular values, and $\mathbf{U}_2$ be the remaining part of $\mathbf{U}$. Since $N\mathbf{U}^{\mathrm{T}}\Xi_t\mathbf{U} = \mathbf{U}^{\mathrm{T}}(N_1\Xi_1 + N_2\Xi_2)\mathbf{U}$, we know by the positive semidefinite properties of Ξ_i's that $\mathbf{U}^{\mathrm{T}}\Xi_i\mathbf{U} = \mathrm{diag}(\Phi_i, \mathbf{0})$ and $\Xi_i = \mathbf{U}\,\mathrm{diag}(\Phi_i, \mathbf{0})\mathbf{U}^{\mathrm{T}}$, where $\Phi_i = \mathbf{U}_1^{\mathrm{T}}\Xi_i\mathbf{U}_1$, i.e. Φ_i is the second moment of class i in the span of $\mathbf{U}_1$ (see Appendix D). Define $\Phi_t = \mathbf{U}_1^{\mathrm{T}}\Xi_t\mathbf{U}_1$.

Now consider the regularized generalized eigenvalue problem,

$$\mathbf{A}\mathbf{w}_i = \lambda_i\mathbf{B}\mathbf{w}_i, \tag{14}$$

where

$$\mathbf{A} = \mathbf{U}\begin{pmatrix}\Phi_2 + \alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \alpha\mathbf{I}\end{pmatrix}\mathbf{U}^{\mathrm{T}},$$

and

$$\mathbf{B} = \mathbf{U}\begin{pmatrix}\Phi_t + \alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \alpha\mathbf{I}\end{pmatrix}\mathbf{U}^{\mathrm{T}}.$$

Let

$$\mathbf{w}_i = \sum_{j=1}^{m+1} c_{i,j}\mathbf{u}_j = \mathbf{U}\mathbf{c}_i,$$

where $\mathbf{c}_i^{\mathrm{T}} = (c_{i,1}, c_{i,2}, \ldots, c_{i,m+1})^{\mathrm{T}}$. Then the problem can be simplified as

$$\begin{pmatrix}\Phi_2 + \alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \alpha\mathbf{I}\end{pmatrix}\mathbf{c}_i = \lambda_i\begin{pmatrix}\Phi_t + \alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \alpha\mathbf{I}\end{pmatrix}\mathbf{c}_i.$$

Denote the kth canonical vector by $\mathbf{e}_k$; that is the kth component of $\mathbf{e}_k$ is 1 and the others are zero. Note $\mathbf{e}_{s+1}, \mathbf{e}_{s+2}, \ldots, \mathbf{e}_{m+1}$ are eigenvectors with eigenvalue 1, and therefore can be safely removed. Hence we only need consider $\mathbf{c}_i$ with the form of $\mathbf{c}_i^{\mathrm{T}} = (\mathbf{d}_i^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})$. It is easy to verify that $\mathbf{d}_i$ is the generalized eigenvector of $(\Phi_2 + \alpha\mathbf{I}, \Phi_t + \alpha\mathbf{I})$, i.e.

$$(\Phi_2 + \alpha\mathbf{I})\mathbf{d}_i = \lambda(\Phi_t + \alpha\mathbf{I})\mathbf{d}_i. \tag{15}$$

Since $\mathbf{w}_i = \mathbf{U}_1\mathbf{d}_i$, we can get $\mathbf{W}_1$ as

$$\mathbf{W}_1 = \mathbf{U}_1\mathbf{D}, \tag{16}$$

where $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n)$.

To summarize, for data sets with high dimensionality, we can carry on the calculation in two levels. In the first level, we apply SVD on the augmented data matrix $\mathbf{Y}$; we then select singular vectors to form $\mathbf{U}_1$, whose singular values are larger than a predefined threshold value. In the second level, we project data in the span of $\mathbf{U}_1$ and calculate the second moments Φ_i's; finally we solve the generalized eigenvalue problem (15) and obtain the solution as defined in Eq. (16).

4. Discussion

In this section, we investigate the properties of our proposed method, and study the relationship of the new proposed method with other dimension reduction methods. For simplicity, we assume that $\hat{\Xi}_i$'s are reliably estimated. Therefore we shall use Ξ_i in our discussion directly.

4.1. Information theoretical property of the criterion

The KL divergence of two multivariate normal distributions p_i and p_j has a closed expression as:

$$L_{ij} = \tfrac{1}{2}\{\log(|\Sigma_i^{-1}\Sigma_j|) + \mathrm{tr}(\Sigma_i\Sigma_j^{-1}) + (\mu_i - \mu_j)^{\mathrm{T}}\Sigma_j^{-1}(\mu_i - \mu_j) - m\}, \tag{17}$$

where $p_i = \mathcal{N}(\mu_i, \Sigma_i)$. The symmetric KL divergence is defined as $J_{ij} = L_{ij} + L_{ji}$. Using formulas in Appendix A, one can easily get that

$$J_{12} = C_0(\tfrac{1}{2}) - 1, \tag{18}$$

that is, when ξ is $\frac{1}{2}$, C_0 is equivalent to the symmetric KL divergence (up to a constant) of two normal distributions. The solution of maximizing C can be seen as finding an affine subspace that maximally preserves C_0, i.e. an optimal truncated spectrum of J_{12}.

The KL divergence can be seen as a distance measure between two distributions, and therefore a measure of separability of classes. Traditional viewpoints aim at maximizing the KL divergence between classes in lower dimensional linear subspaces, see Fukunaga (1990) for an introduction and la Torre and Kanade (2005) for the recent development. It is easy to show that maximizing the lower-dimensional KL divergence in Fukunaga (1990) and la Torre and Kanade (2005) is equivalent to our

proposed problem with an additional constraint

$$\mathbf{W}^{\mathrm{T}} = (\mathbf{V}^{\mathrm{T}}, \mathbf{e}), \tag{19}$$

where $\mathbf{V} \in \mathbb{R}^{m\times d}$, and $\mathbf{e}^{\mathrm{T}} = (0, 0, \ldots, 1)$. With the additional constraint, a closed-form solution cannot be found. By relaxing $\mathbf{e} \in \mathbb{R}^{m\times 1}$, we can find closed-form solutions.

4.2. Connection to FDA and CSP

Without losing generality, let us consider the one dimensional case in this subsection. Let $\mathbf{w}^{\mathrm{T}} = (\mathbf{v}_1^{\mathrm{T}}, v_0)$. Then we have $Z = \mathbf{v}_1^{\mathrm{T}} X + v_0$, where X and Z are random covariate in higher- and lower-dimensional spaces. Displacement v_0 is the same for both classes, and therefore plays no important role for final classifications. In other words, the effectiveness of the generated feature is solely determined by $\mathbf{v}_1$. Let $\mathbf{v}_1^*$ be an optimal solution.

Consider maximizing $C(\mathbf{W}; \frac{1}{2}, d)$. We know that $\mathbf{w}^*$ is the eigenvector of $\Xi_1^{-1}\Xi_2 + \Xi_2^{-1}\Xi_1$ with the largest eigenvalue.

First, let us consider $\mu_1 = \mu_2 = \mu$. Using formulas in Appendix A, we can simplify $\Xi_1^{-1}\Xi_2 + \Xi_2^{-1}\Xi_1$ as

$$\Xi_1^{-1}\Xi_2 + \Xi_2^{-1}\Xi_1 = \begin{pmatrix} \Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 & 0 \\ 2\mu^{\mathrm{T}} - \mu^{\mathrm{T}}(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1) & 1 \end{pmatrix}.$$

Then by simple linear algebra, we can show that $\mathbf{v}_1^*$ is also the eigenvector of $\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1$ with the largest eigenvalue.

Second, let us consider $\Sigma_1 = \Sigma_2 = \Sigma$. In this case, it is easy to verify the following:

$$\Xi_1^{-1}\Xi_2 + \Xi_2^{-1}\Xi_1 = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix} + 2\mathbf{I},$$

where $\mathbf{A} = \Sigma^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^{\mathrm{T}}$ and $\mathbf{B} = (\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1}(\mu_1 - \mu_2)^{\mathrm{T}}$. It is then not hard to show that $\mathbf{v}_1^*$ is the eigenvector of $\mathbf{A}$ with the largest eigenvalue.

In summary, we show that FDA and CSP are special cases of our proposed AFE for normally distributed data. Therefore, theoretically speaking AFE is more flexible than FDA and CSP.

5. Experiments

5.1. Visualization on simulated data sets

In order to compare our method with PCA and FDA, Several 7-dimensional toy data sets have been generated. The toy data sets contain three-dimensional relevant components, while the others are merely random noise. The three relevant components form two concentric cylinders. The generated data are spread along the surfaces of the cylinders. The cylinders are of elliptic, parabolic and hyperbolic forms. Fig. 1 illustrates the first two features found by PCA, FDA and our new approach AFE. As a result of preserving the variance of data, PCA projects data along the surfaces, and thus does not preserve the separation of the conic cylinders; FDA fails to separate the two classes in most cases; on the other hand, our method correctly captures the discriminant information in the data.

5.2. Experiments with real data sets

We selected four benchmark data sets: German, diabetes, waveform and heart.[1] The dimensionality of these data sets are 20, 8, 21, and 13, respectively. The data sets had been preprocessed and partitioned into 100 training and test sets (about 40%:60%). They have been used to evaluate the performance of kernel FDA (Mika et al., 1999), kernel PLS (Rosipal et al., 2003). We compared our new approach with FDA, CSP, and FKT. For convenience, AFE1 and AFE2 are used for orthogonal and Ξ_t-orthogonal AFE algorithms. We used FDA, CSP, FKT, AFE1 and AFE2 to generate lower-dimensional features; the features are then used by linear support vector machines (SVM) to do classifications. To measure the discriminant information of the data set, we also classified the original data set via linear SVMs, which we denote FULL in the reported figures. Feature extraction and classification are trained on training sets, and test-set accuracy (TSA) are calculated with predictions on test sets. Statistical boxplots of TSAs are shown in Figs. 2–5 for the three chosen data sets. The poor performance of FDA, CSP and FKT affirms that first-order or second-order statistics alone cannot capture discriminant information contained in the data sets. By comparing AFE1 and AFE2 with FULL, we see that AFE1 and AFE2 are capable of extracting the discriminant information of the chosen data. AFE1 and AFE2 can be used to generate much compact discriminant features, for example, the average dimensionality of extracted features for German, diabetes and waveform are 8.16, 3.18, 1.2 and 4.96, respectively.

6. Conclusions

In this study, we proposed a novel dimension reduction method for binary classification problems. Unlike traditional linear subspace methods, the new proposed method finds lower-dimensional affine subspaces for data observations. We presented the closed-form solutions of our new approach, and investigated its information–theoretical properties. We showed that our method has close connections with FDA, CSP and FKT methods in the literature. Numerical experiments show the competitiveness of our method as a preliminary data-exploring tool for data visualization and classification.

Though we focus on binary classification problems in this study, it is always desirable to handle multi-class problems. One can extend AFE to multi-class problems by following the work presented in Dornhege et al. (2004). Here we proposed another way to extend AFE to multi-class. Let J_{ij} be the symmetric KL distance of classes i and j, and assume class i, $(i = 1, 2, \ldots, K)$, can be modeled by multivariate normal distribution. Then we have

$$\sum_{i=1}^{K} \Xi_i^{-1}\Xi_t \propto \sum_{i,j=1}^{K} J_{ij},$$

where Ξ_i is the augmented second moment matrix for class i and

$$N\Xi_t = \sum_{i=1}^{K} N_i \Xi_i.$$

Therefore, we may calculate the truncated spectrum of

$$\sum_{i=1}^{K} \Xi_i^{-1}\Xi_t$$

for the lower-dimensional representations.

Another more important problem is to investigate the relationship of our new proposed method with quadratic discriminant analysis (QDA). It has long been known that FDA is an optimal dimension reduction method for linear discriminant analysis (LDA) (Hastie et al., 2001). But there is no well-accepted dimension reduction method for QDA in the literature. Recently, Huo et al. (2003) proposed that FKT might be seen as an optimal one for QDA under certain circumstance. Our future work will be dedicated to finding the relationship of AFE and QDA.

[1] http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

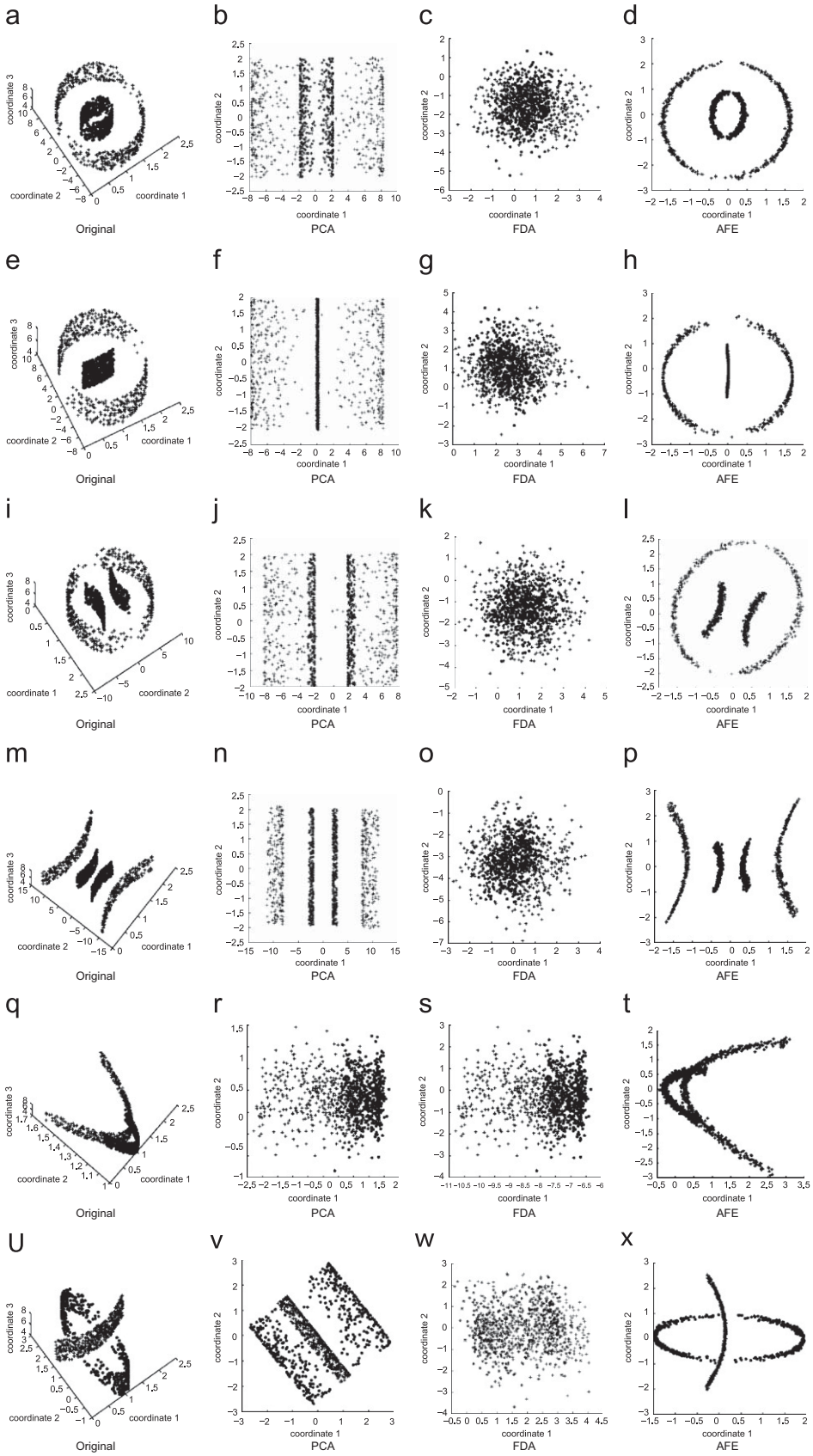

Fig. 1. Comparison of features found by PCA, FDA, and our method. Star and plus points belong to different classes.

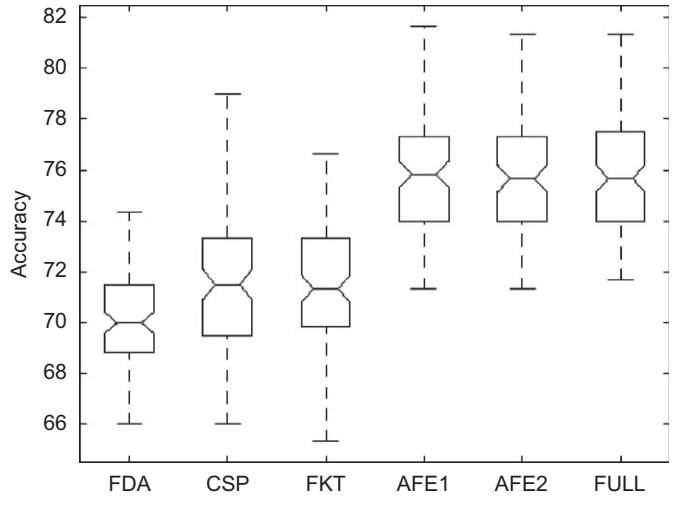

Fig. 2. Test set accuracy for German data sets. See text for notations and details.

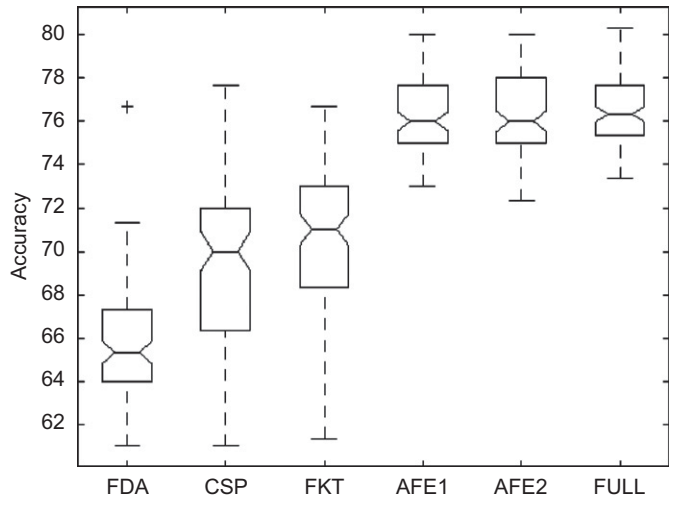

Fig. 3. Test set accuracy for diabetes data set. See text for notations and details.

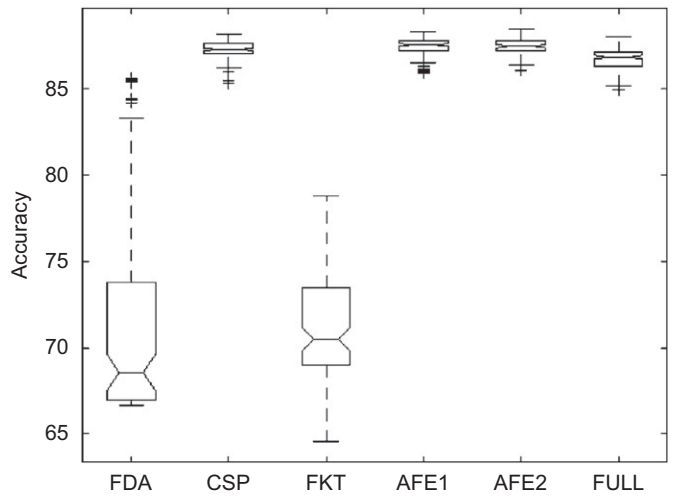

Fig. 4. Test set accuracy for waveform data set. See text for notations and details.

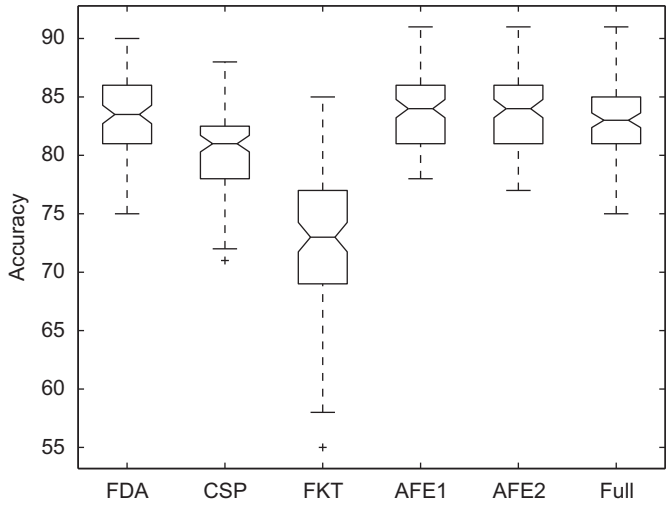

Fig. 5. Test set accuracy for heart data set. See text for notations and details.

Appendix A

Let X be a random covariate which has probability distribution p. So we have

$$\boldsymbol{\mu} = E_{X\sim p}X,$$

$$\boldsymbol{\Sigma} = E_{X\sim p}(X-\mu)(X-\mu)^{\mathrm{T}},$$

$$\boldsymbol{\Xi} = E_{X\sim p}\left\{\begin{pmatrix} X \\ 1 \end{pmatrix}(X^{\mathrm{T}}, 1)\right\},$$

where $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Xi}$ are, respectively, the mean, covariance and augmented second moment of X. When μ and $\boldsymbol{\Sigma}$ are finite, we have

$$\boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{\Sigma}+\mu\mu^{\mathrm{T}} & \mu \\ \mu^{\mathrm{T}} & 1 \end{pmatrix}.$$

Assuming $\boldsymbol{\Sigma}$ is positive definite, we have the inverse of $\boldsymbol{\Xi}$ as follows:

$$\boldsymbol{\Xi}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & -\boldsymbol{\Sigma}^{-1}\mu \\ -\mu^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} & 1+\mu^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mu \end{pmatrix}.$$

Appendix B

Lemma 5. *Let* $\mathbf{A}$ *be an* $r\times s$ *matrix,* $(r\geqslant s)$*, and* $\mathbf{A}^{\mathrm{T}}\mathbf{A}=I$*. Let* $\boldsymbol{\Lambda}$ *be a diagonal matrix. Then*

$$\xi\,\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{A}) + (1-\xi)\,\mathrm{tr}([\mathbf{A}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{A}]^{-1}) \leqslant \sum_{i=1}^{s} f_i(\xi).$$

Proof. By the Poincaré separation theorem (cf. Horn and Johnson, 1990), we know the eigenvalues of $\mathbf{A}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{A}$ interlaces with those of $\boldsymbol{\Lambda}$. That is, for each integer j, $(1\leqslant j\leqslant s)$, we have $\lambda_j\leqslant\tau_j\leqslant\lambda_{j+r-s}$, where τ_j is the eigenvalue of $\mathbf{A}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{A}$. Then it is obvious that

$$\begin{aligned}&\xi\,\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{A}) + (1-\xi)\,\mathrm{tr}([\mathbf{A}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{A}]^{-1}) \\ &\quad= \sum_{i=1}^{s}\left[\xi\tau_i + (1-\xi)\frac{1}{\tau_i}\right] \leqslant \sum_{i=1}^{s} f_i(\xi).\end{aligned}$$

Appendix C

Let $\mathbf{U}$ be a nonsingular matrix such that $\mathbf{U}^{\mathrm{T}}\hat{\boldsymbol{\Xi}}_2\mathbf{U}=\mathbf{I}$ and $\mathbf{U}^{\mathrm{T}}\hat{\boldsymbol{\Xi}}_1\mathbf{U}=\boldsymbol{\Lambda}$. Then we have

$$\hat{\boldsymbol{\Pi}}_2 = \mathbf{W}^{\mathrm{T}}(\mathbf{U}^{-1})^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\hat{\boldsymbol{\Xi}}_2\mathbf{U}\mathbf{U}^{-1}\mathbf{W} = \mathbf{V}^{\mathrm{T}}\mathbf{V},$$

$$\hat{\mathbf{\Pi}}_1 = \mathbf{W}^{\mathrm{T}}(\mathbf{U}^{-1})^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\hat{\mathbf{\Xi}}_1\mathbf{U}\mathbf{U}^{-1}\mathbf{W} = \mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V},$$

where $\mathbf{V} = \mathbf{U}^{-1}\mathbf{W} \in \mathbb{R}^{(m+1)\times k}$. Then we can get

$$C(\mathbf{W};\xi,d) = (1-\xi)\,\mathrm{tr}[(\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}] + \xi\,\mathrm{tr}[(\mathbf{V}^{\mathrm{T}}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V}].$$

Applying SVD on $\mathbf{V}$, we get $\mathbf{V} = \mathbf{A}\mathbf{D}\mathbf{B}^{\mathrm{T}}$. Here $\mathbf{A}$ and $\mathbf{B}$ are $(m+1)\times d$ and $d\times d$ orthogonal matrices, i.e. $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}$, $\mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{I}$, and $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$. $\mathbf{D}$ is a $d\times d$ diagonal matrix. Therefore we have:

$$\begin{aligned}\mathrm{tr}[(\mathbf{V}^{\mathrm{T}}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V}] &= \mathrm{tr}[\mathbf{V}(\mathbf{V}^{\mathrm{T}}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}]\\ &= \mathrm{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{A}),\end{aligned}$$

$$\begin{aligned}\mathrm{tr}[(\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}] &= \mathrm{tr}[\mathbf{V}(\mathbf{V}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}]\\ &= \mathrm{tr}[\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}]\\ &= \mathrm{tr}[(\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{A})^{-1}].\end{aligned}$$

Thus by Lemma 5, we know that

$$\begin{aligned}C(\mathbf{W};\xi,d) &= \mathrm{tr}[\xi\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{A} + (1-\xi)(\mathbf{A}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{A})^{-1}] - d\\ &\leqslant \sum_{i=1}^{d} f_i(\xi) - d.\end{aligned}$$

Appendix D

Since

$$\mathbf{U}^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U} = \begin{pmatrix}\mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_1 & \mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_2\\ \mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_1 & \mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_2\end{pmatrix},$$

we have

$$\frac{N_1}{N}\begin{pmatrix}\mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_1\mathbf{U}_1 & \mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_1\mathbf{U}_2\\ \mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_1\mathbf{U}_1 & \mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_1\mathbf{U}_2\end{pmatrix} + \frac{N_2}{N}\begin{pmatrix}\mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_2\mathbf{U}_1 & \mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_2\mathbf{U}_2\\ \mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_2\mathbf{U}_1 & \mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_2\mathbf{U}_2\end{pmatrix} = \begin{pmatrix}\mathbf{\Delta}^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}.$$

Since $\mathbf{\Xi}_i$'s are positive semidefinite, i.e. $\mathbf{\Xi}_i \succcurlyeq 0$, we have $\mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_2 \succcurlyeq 0$ and $\mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_2 \succcurlyeq 0$. Therefore, we must have $\mathbf{U}_1^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_2 = \mathbf{0}$ and $\mathbf{U}_2^{\mathrm{T}}\mathbf{\Xi}_i\mathbf{U}_2 = \mathbf{0}$, otherwise the above equation will be invalid.

References

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. Journal of Chemometrics 17 (3), 166–173.

Bartlett, M.S., 1938. Further aspects of the theory of multiple regression. Proceedings of the Cambridge Philosophical Society 34, 33–40.

Belhumeur, P.N., Hespanha, J., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. TPAMI 19 (7), 711–720.

Cao, W., Haralick, R.M., 2007. Affine feature extraction: a generalization of the Fukunaga–Koontz transformation. In: Perner, P. (Ed.), Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science, vol. 4571. Springer, Berlin, pp. 160–173.

Dai, J.J., Lieu, L., Rocke, D., 2006. Dimension reduction for classification with gene expression microarray data. Statistical Applications in Genetics and Molecular Biology 5 (Article 6).

Dornhege, G., Blankertz, B., Curio, G., Müller, K.-R., 2004. Increase information transfer rates in BCI by CSP extension to multi-class. In: NIPS2005, pp. 733–740.

Edelman, A., Arias, T.A., Smith, S.T., 1999. The geometry of algorithms with orthogonality constraints. SIAM Journal on Matrix Analysis and Applications 20 (2), 303–353.

Foley, D.H., Sammon, J.W., Jr., 1975. An optimal set of discriminant vectors. IEEE Transactions on Computers C-24, 281–289.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed. Academic Press, New York.

Fukunaga, K., Koontz, W., 1970. Application of the Karhunen–Loéve expansion to feature selection and ordering. IEEE Transactions on Computers C-19, 311–318.

Golub, G.H., Van Loan, C.F., 1996. Matrix Computations. The Johns Hopkins University Press, Baltimore, MD.

Haralick, R., 1977. Pattern discrimination using ellipsoidally symmetric multivariate density functions. Pattern Recognition 9 (2), 89–94.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning. Springer, Berlin.

Horn, R.A., Johnson, C.R., 1990. Matrix Analysis. Cambridge University Press, Cambridge.

Huo, X., et al., 2003. Optimal reduced-rank quadratic classifiers using the Fukunaga–Koontz transform, with applications to automated target recognition. In: Proceedings of SPIE, vol. 5094, pp. 59–72.

la Torre, F.D., Kanade, T., 2005. Multimodal oriented discriminant analysis. In: ICML2005, pp. 177–184.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K., 1999. Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing, vol. IX, pp. 41–48.

Ramoser, H., Müller-Gerking, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Transactions on Rehabilitation Engineering 8, 441–446.

Rosipal, R., Krämer, N., 2005. Overview and recent advances in partial least squares. In: Subspace, Latent Structure and Feature Selection Techniques, pp. 34–51.

Rosipal, R., Trejo, L.J., Matthews, B., 2003. Kernel pls-svc for linear and nonlinear classification. In: ICML2003, pp. 640–647.

Yang, M.-H., Kriegman, D.J., Ahuja, N., 2002. Detecting faces in images: a survey. TPAMI 24 (1), 34–58.