Affine Feature Extraction: A Generalization of the Fukunaga-Koontz Transformation

Wenbo Cao and Robert Haralick

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

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*), 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 for data observations. Our method can be understood as a generalization of the Fukunaga-Koontz Transformation. We show that the proposed method has a closed-form solution and thus can be solved very efficiently. Also we investigated the information-theoretical properties of the new method and study the relationship of our method with other methods. The experimental results show that our method, as PCA and FDA, 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, dimension reduction methods are often used in machine learning and data mining problems. Examples are face recognition in computer vision [3, 20], electroencephalogram (*EEG*) signal classification in Brain-Computer Interface (*BCI*)[5, 16] and microarray data analysis [4]. Linear subspace methods have been widely used for the purpose of dimension reduction. We give a brief review of the most commonly used ones.

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 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*) finds linear subspaces in which the distance between the means of classes is maximized and the variance of each class is minimized at the 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 of 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 [2].

Recently there has been some interest in partial least squares (*PLS*) [18]. Only recently, it has been shown that PLS has a close connection with FDA [1]. PLS finds linear subspaces by iteratively maximizing the covariance of 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 observation matrix by its rank-one estimation found in the previous iteration, and generate deflated observation vectors.

Linear subspaces are specific instances of affine subspaces. In this study, we propose 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*) [9]. 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 present the main result of our work: the motivation of the study, the AFE method and its closed-form solutions. We investigated the information-theoretical properties of AFE and the relationship of AFE with other linear subspace dimension reduction methods in section 3. Experimental results are presented in section 4. We concludes this study with the summary of our work, and possible future directions in section 5.

2 Affine Feature Extraction

Consider a binary classification problem, which is also called *discriminant analysis* in statistics. Let $\{(\mathbf{x}_j, g_j) \in \mathbb{R}^m \times \{1,2\} | j = 1, 2, \cdots, N\}$ 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

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

where $\mathbf{X}_i = (\mathbf{x}_k, \mathbf{x}_{k+1}, \cdots, \mathbf{x}_{k+N_i})$, and $k = \sum_{j=1}^{i-1} N_j$. 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 $\mathbf{Y}_i^T = (\mathbf{X}_i^T, \mathbf{1})$. 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) $\square^T$ is the transpose of a vector or matrix $\square$; and (5) $\mathrm{tr}(\square)$ is the trace of a matrix $\square$.

2.1 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 [11] 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-moment* for class i as follows:

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

$$\text{covariance} \quad \hat{\mathbf{\Sigma}}_i = \frac{1}{N_i}\mathbf{X}_i(\mathbf{I} - \frac{1}{N_i}\mathbf{1}\mathbf{1}^T)^2\mathbf{X}_i^T, \tag{3}$$

$$\text{second-moment} \quad \hat{\mathbf{M}}_i = \frac{1}{N_i}\mathbf{X}_i\mathbf{X}_i^T. \tag{4}$$

One essential assumption of classical discriminant analysis is that the probability density for each class can be modeled as a multivariate normal distribution, i.e. $\mathcal{N}(\mu_i, \mathbf{\Sigma}_i)$ $(i = 1, 2)$. Equations 2 and 3 can be seen as the empirical estimations of classical density parameters μ_i and $\mathbf{\Sigma}_i$, 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:

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

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

When the two classes have a common covariance, i.e. $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, the problem is relatively easy. It is not hard to show that the optimal $\mathbf{w}^*$ is the eigenvector of $\mathbf{\Sigma}^{-1}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$. FDA essentially capture this situation by solving the following problem:

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

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

When $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, to find an optimal linear subspace is hard. The only known closed-form solution is that $\mathbf{w}^*$ is the eigenvector of $\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2 + \mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1$, which has the largest eigenvalue. It can be shown that, when $\mu_1 = \mu_2 = 0$, the solution optimizes the Kullback-Leibler *KL* divergence and the Bhattacharyya distance, (c.f. Section 10.2 of [8]). 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. Widely used in EEG classification problems, common spatial pattern (*CSP*) solves the following problem:

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

Therefore CSP only works well when the difference of class means is small, i.e. $|\mu_2 - \mu_1| \approx 0$. 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 $\hat{\mu}_i$ to zero. It can be seen as a rough shrinkage estimation of the mean for high dimensional data. FKT solves the following problem:

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

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

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

Note $\mathbf{w}^T \hat{\mathbf{M}}_i \mathbf{w} = \frac{1}{N_i} \sum_{j=k+1}^{k+N_i} z_j^2$, where $k = \sum_{j=1}^{i-1} N_j$ and $i = 1, 2$. That is: $\mathbf{w}^T \hat{\mathbf{M}}_i \mathbf{w}$ 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 [7].

2.2 Method

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

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

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

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

where $0 \leq \xi \leq 1$. We use the sum of ratios to measure the importance of $\mathbf{w}$ instead of two separated optimization problems in FKT. And 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 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d) \in \mathbb{R}^{(m+1)\times d}$ 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 propose the following optimization problem to find $\mathbf{W}$:

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

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

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

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:

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

where $0 \leq \xi \leq 1$, d $(1 \leq d \leq m)$ 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 [11]. 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. $d^* = \inf\{d | C(\mathbf{W}; \xi, d) \geq \beta C_0\}$, where β is a constant.

Constraint $\mathbf{W}^T\hat{\mathbf{\Xi}}_t\mathbf{W} = \mathbf{I}$ is necessary in our generalization from one dimensional to 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 [6] 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. $\hat{\mathbf{\Xi}}_t$-orthogonal constraint: $\mathbf{W}^T\hat{\mathbf{\Xi}}_t\mathbf{W} = \mathbf{I}$;
2. Orthogonal constraint: $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

2.3 Algorithms

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

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

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

Define $\mathbf{\Lambda} = \mathsf{diag}(\lambda_1, \lambda_2, \cdots, \lambda_{m+1})$, and λ_i's are the eigenvalues of $(\hat{\mathbf{\Xi}}_1, \hat{\mathbf{\Xi}}_2)$ $(i = 1, 2, \cdots, m+1)$, i.e. $\hat{\mathbf{\Xi}}_1\mathbf{u}_i = \lambda_i\hat{\mathbf{\Xi}}_2\mathbf{u}_i$. Let $\lambda_i(\xi)$'s be the ordered eigenvalues of $(\hat{\mathbf{\Xi}}_1, \hat{\mathbf{\Xi}}_2)$ with respect to $f(\lambda;\xi)$. That is: define $f_i(\xi) = f(\lambda_i(\xi);\xi)$, then we have $f_1(\xi) \ge f_2(\xi) \ge \cdots \ge f_{m+1}(\xi)$. The following lemma for nonsingular symmetric $\hat{\mathbf{\Xi}}_1$ and $\hat{\mathbf{\Xi}}_2$ can be found in [10]:

Lemma 1. *There exists nonsingular matrix* $\mathbf{U} \in \mathbb{R}^{(m+1)\times(m+1)}$ *such that*

$$\mathbf{U}^T\hat{\mathbf{\Xi}}_2\mathbf{U} = \mathbf{I}\,, \;\; \mathbf{U}^T\hat{\mathbf{\Xi}}_1\mathbf{U} = \mathbf{\Lambda}.$$

In Appendix C, we show that:

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

Note that: 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 1. *Let* $\mathbf{U}_\xi = (\mathbf{u}_1^\xi, \mathbf{u}_2^\xi, \cdots, \mathbf{u}_d^\xi)$*, where* $\mathbf{u}_i^\xi$ *is the eigenvector of* $(\hat{\mathbf{\Xi}}_1, \hat{\mathbf{\Xi}}_2)$ *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 prove $\mathbf{U}_\xi$ maximizes $C(\mathbf{W};\xi,d)$. Note $\mathbf{U}_\xi^T\hat{\mathbf{\Xi}}_2\mathbf{U}_\xi = \mathbf{I}$ and $\mathbf{U}_\xi^T\hat{\mathbf{\Xi}}_1\mathbf{U}_\xi = \mathsf{diag}(\lambda_1(\xi), \lambda_2(\xi), \cdots, \lambda_d(\xi))$. Then it is easy to affirm the proposition.

Let $\mathbf{U}_\xi = \mathbf{QR}$, 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 $\mathbf{W}_2 = \mathbf{U}_\xi\mathbf{\Gamma}^{-\frac{1}{2}}$, where

$$\mathbf{\Gamma} = \frac{1}{N}\mathrm{diag}(N_1\lambda_1(\xi) + N_2, N_1\lambda_2(\xi) + N2, \cdots, N_1\lambda_d(\xi) + N_2). \tag{13}$$

It can be easily shown that $\mathbf{W}_2$ maximize $C(\mathbf{W};\xi,d)$ and satisfies the $\hat{\mathbf{\Xi}}_t$-orthogonal constraint. In practice, we only need to check the largest d and the smallest d eigenvalues and eigenvectors of $(\hat{\mathbf{\Xi}}_1, \hat{\mathbf{\Xi}}_2)$ in order to generate d features. The pseudo-code of the algorithm is given in Table 1. Practically, we may need to let $\hat{\mathbf{\Xi}}_i \leftarrow \hat{\mathbf{\Xi}}_i + \alpha_i\mathbf{I}$ to guarantee the positive definiteness of $\hat{\mathbf{\Xi}}_i$, where α_i is a small positive constant.

Algorithm for feature extraction

Input: Data sample $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$
Output: Transformation matrix $\mathbf{W}$

1. Calculate the augment second moment matrices $\hat{\mathbf{\Xi}}_1$, and $\hat{\mathbf{\Xi}}_2$;
2. Compute the largest d and the smallest d eigenvalues and eigenvectors of $(\hat{\mathbf{\Xi}}_1, \hat{\mathbf{\Xi}}_2)$;
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 $\hat{\mathbf{\Xi}}_t$-orthogonal constraint), calculate $\mathbf{\Gamma}$ as Eq. 13;
7∗∗. (For $\hat{\mathbf{\Xi}}_t$-orthogonal constraint), Let $\mathbf{W} = \mathbf{U}_\xi \mathbf{\Gamma}^{-\frac{1}{2}}$;
6. Return $\mathbf{W}$.

Table 1. Pseudo-code for feature extraction

3 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 $\hat{\mathbf{\Xi}}_i$'s are reliably estimated. Therefore we shall use $\mathbf{\Xi}_i$ in our discussion directly.

3.1 Information theoretical property of the criterion

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

$$L_{ij} = \frac{1}{2}\{\log(|\mathbf{\Sigma}_i^{-1}\mathbf{\Sigma}_j|) + \mathrm{tr}(\mathbf{\Sigma}_i\mathbf{\Sigma}_j^{-1}) + (\mu_i - \mu_j)^T\mathbf{\Sigma}_j^{-1}(\mu_i - \mu_j) - m\}; \quad (14)$$

where $p_i = \mathcal{N}(\mu_i, \mathbf{\Sigma}_i)$. The symmetric KL divergence is defined as $J_{ij} = L_{ij} + L_{ji}$. It is easy to obtain

$$J_{12} = \frac{1}{2}\{\mathrm{tr}(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1) + \mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2) + \mathrm{tr}[(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})(\mu_2 - \mu_1)(\mu_2 - \mu_1)^T] - 2m\}. \quad (15)$$

Using formulas in Appendix A, one can easily get that

$$J_{12} = C_0(\frac{1}{2}) - 1; \quad (16)$$

That is, when ξ is $1/2$, 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 [8] for an introduction and [14] for the recent development. It is easy to show that maximizing the lower-dimensional KL divergence in [14] is equivalent to our proposed problem with an additional constraint

$$\mathbf{W}^T = (\mathbf{V}^T, \mathbf{e}) \tag{17}$$

where $\mathbf{V} \in \mathbb{R}^{m\times d}$, and $\mathbf{e}^T = (0, 0, \cdots, 1)$. 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.

3.2 Connection to FDA and CSP

Without losing generality, let us consider the one dimensional case in this subsection. Let $\mathbf{w}^T = (\mathbf{v}_1^T, v_0)$. Then we have $Z = \mathbf{v}_1^T X + v_0$, 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 $C(\mathbf{W}; 1/2, d)$. We know that $\mathbf{w}^*$ is the eigenvector of $\mathbf{\Xi}_1^{-1}\mathbf{\Xi}_2 + \mathbf{\Xi}_2^{-1}\mathbf{\Xi}_1$ with the largest eigenvalue.

First, let us consider $\mu_1 = \mu_2 = \mu$. Using formulas in Appendix A, we can simplify $\mathbf{\Xi}_1^{-1}\mathbf{\Xi}_2 + \mathbf{\Xi}_2^{-1}\mathbf{\Xi}_1$ as

$$\mathbf{\Xi}_1^{-1}\mathbf{\Xi}_2 + \mathbf{\Xi}_2^{-1}\mathbf{\Xi}_1 = \begin{pmatrix} \mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2 + \mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1 & 0 \\ 2\mu^T - \mu^T(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2 + \mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1) & 1 \end{pmatrix}$$

Then by simple linear algebra, we can show that $\mathbf{v}_1^*$ is also the eigenvector of $\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2 + \mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1$ with the largest eigenvalue.

Second, let us consider $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$. In this case, it is easy to verify the following:

$$\mathbf{\Xi}_2^{-1}\mathbf{\Xi}_1 - \mathbf{I} = \begin{pmatrix} \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)\mu_1^T & \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \\ \mu_1^T - \mu_2^T - \mu_2^T\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)\mu_1^T & -\mu_2^T\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2) \end{pmatrix}$$
$$\mathbf{\Xi}_1^{-1}\mathbf{\Xi}_2 - \mathbf{I} = \begin{pmatrix} \mathbf{\Sigma}^{-1}(\mu_2 - \mu_1)\mu_2^T & \mathbf{\Sigma}^{-1}(\mu_2 - \mu_1) \\ \mu_2^T - \mu_1^T - \mu_1^T\mathbf{\Sigma}^{-1}(\mu_2 - \mu_1)\mu_2^T & -\mu_1^T\mathbf{\Sigma}^{-1}(\mu_2 - \mu_1) \end{pmatrix}$$

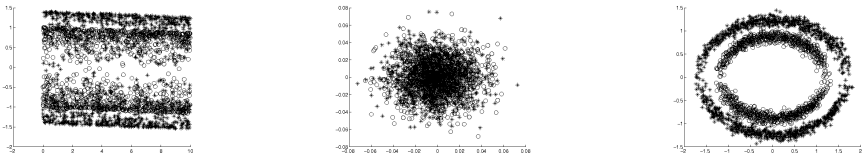
Then we have

$$\mathbf{\Xi}_1^{-1}\mathbf{\Xi}_2 + \mathbf{\Xi}_2^{-1}\mathbf{\Xi}_1 = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix} + 2\mathbf{I},$$

where $\mathbf{A} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $\mathbf{B} = (\mu_1 - \mu_2)^T\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)^T$. 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 distribute data. Therefore, theoretically speaking AFE is more flexible than FDA and CSP.

4 Experiments



(a) PCA (b) FDA (c) AFE

Fig. 1. Comparision of features found by PCA, FDA, and Our method. Star and circle points belong to different classes.

In order to compare our method with PCA and FDA, a 7-dimensional toy data set has been generated. The toy data set contains 3-dimensional relevant components, while the others are merely random noise. The 3 relevant components form two concentric cylinders. The generated data are spread along the surfaces of the cylinders. Figure 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 reflect the separation of the data (Figure 1(a)). Figure 1(b) shows that FDA fails to separate the two classes. On the other hand, Figure 1(c) shows that our method correctly captures the discriminant information in the data.

We selected three benchmark data sets: German, Diabetes and Waveform. The dimensionality of these data sets are 20, 8, and 21 respectively. They can be freely downloaded from http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm. 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 [15], kernel PLS [19] and soft-margin AdaBoost [17].

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 Figures 2, 3 and 4 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 discrimi-

nant features, for example, the average dimensionality of extracted features for German, Diabetes and Waveform are 8.16, 3.18 and 1.2, respectively.

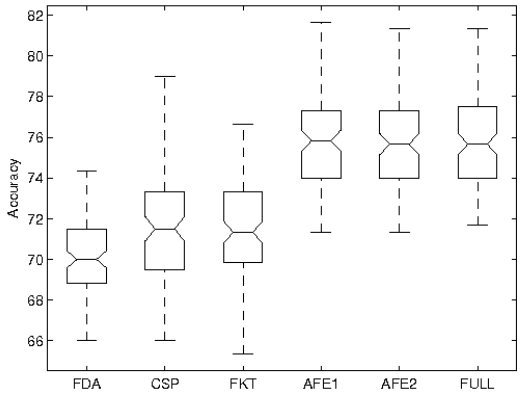


Fig. 2. Test set accuracy for German data set

We conducted preliminary experiments with AFE1 and AFE2 on data sets Tübingen:1a and Berlin:IV from BCI competition 2003 [1]. We used AFE1 and AFE2 to generate low-dimensional representations and then apply logistic regression on the extracted features. For data set Tübingen:1a, we obtained TSA as 77.13% and 85.32% for AFE1+ and AFE2+logistic regression, respectively. The results are comparable with the ones of rank 11 and rank 4 of the competition, correspondingly. For data set Berlin:IV, we obtained TSA 71% for both AFE1+ and AFE2+logistic regression, which are comparable with rank 8 of the competition.

5 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.

[1] see http://ida.first.fraunhofer.de/projects/bci/competition_ii/results/index.html

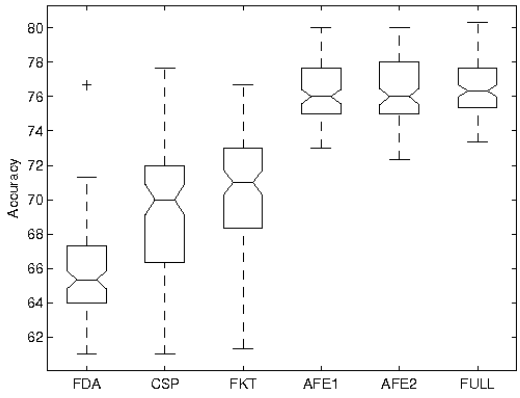


Fig. 3. Test set accuracy for Diabetes data set

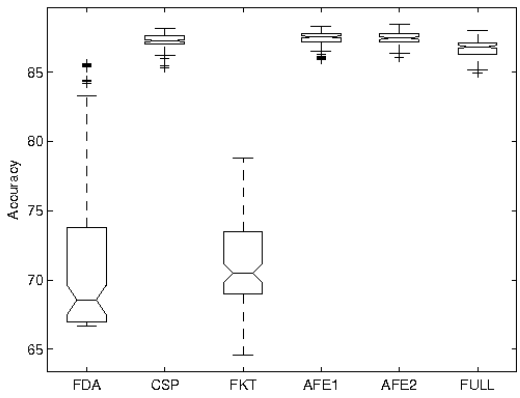


Fig. 4. Test set accuracy for Waveform data set

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 [5]. 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, $(i = 1, 2, \cdots, K)$, can be modeled by multivariate normal distribution. Then we have

$$\sum_{i=1}^{K} \mathbf{\Xi}_i^{-1} \mathbf{\Xi}_t \propto \sum_{i,j=1}^{K} J_{ij},$$

where $\mathbf{\Xi}_i$ is the augmented second moment matrix for class i and $N\mathbf{\Xi}_t = \sum_{i=1}^{K} N_i \mathbf{\Xi}_i$. Therefore we may calculate the truncated spectrum of $\sum_{i=1}^{K} \mathbf{\Xi}_i^{-1} \mathbf{\Xi}_t$ 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) [11]. But there is no well-accepted dimension reduction method for QDA in the literature. Recently, Hou et al. proposed the FKT might be seen as an optimal one for QDA under certain circumstance [13]. Our future work will be dedicated to finding the relationship of AFE and QDA.

Acknowledgement

The authors thank anonymous reviewers for constructive comments on improving the presentation of this work.

Appendix A

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

$$\begin{aligned}
\mu &= E_{X\sim p} X, \\
\mathbf{\Sigma} &= E_{X\sim p} (X-\mu)(X-\mu)^T, \\
\mathbf{\Xi} &= E_{X\sim p} \left\{ \begin{pmatrix} X \\ 1 \end{pmatrix} (X^T, 1) \right\},
\end{aligned}$$

where μ, $\mathbf{\Sigma}$ and $\mathbf{\Xi}$ are, respectively, the mean, covariance and augmented second moment of X. When μ and $\mathbf{\Sigma}$ are finite, i.e. $\|\mu\| < \infty$ and $\|\mathbf{\Sigma}\| < \infty$, we have

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

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

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

Appendix B

Lemma 2. *Let* $\mathbf{A}$ *be an* $r \times s$ *matrix,* $(r \geq s)$*, and* $\mathbf{A}^T\mathbf{A} = I$*. Let* $\mathbf{\Lambda}$ *be a diagonal matrix. Then*

$$\xi tr(\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}) + (1-\xi)tr([\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}]^{-1}) \leq \sum_{i=1}^{s} f_i(\xi);$$

Proof. By the Poincaré separation theorem (c.f. [12] P190), we know the eigenvalues of $\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}$ interlaces with those of $\mathbf{\Lambda}$. That is, for each integer j, $(1 \leq j \leq s)$, we have

$$\lambda_j \leq \tau_j \leq \lambda_{j+r-s},$$

where τ_j is the eigenvalue of $\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}$. Then it is obvious that

$$\begin{aligned} & \xi \mathrm{tr}(\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}) + (1-\xi)\mathrm{tr}([\mathbf{A}^T\mathbf{\Lambda}\mathbf{A}]^{-1}) \\ = & \textstyle\sum_{i=1}^{s}[\xi\tau_i + (1-\xi)\frac{1}{\tau_i}] \\ \leq & \textstyle\sum_{i=1}^{s} f_i(\xi); \end{aligned}$$

Appendix C

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

$$\begin{aligned} \hat{\mathbf{\Pi}}_2 &= \mathbf{W}^T(\mathbf{U}^{-1})^T\mathbf{U}^T\hat{\mathbf{\Xi}}_2\mathbf{U}\mathbf{U}^{-1}\mathbf{W} = \mathbf{V}^T\mathbf{V} \\ \hat{\mathbf{\Pi}}_1 &= \mathbf{W}^T(\mathbf{U}^{-1})^T\mathbf{U}^T\hat{\mathbf{\Xi}}_1\mathbf{U}\mathbf{U}^{-1}\mathbf{W} = \mathbf{V}^T\mathbf{\Lambda}\mathbf{V}, \end{aligned}$$

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

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

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

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

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

Thus by Lemma 2, we know that

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

References

[1] Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.

[2] M. S. Bartlett. Further aspects of the theory of multiple regression. *Proc. Camb. Phil. Soc.*, 34:33–40, 1938.

[3] Peter N. Belhumeur, Joao Hespanha, and David J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] Jian J Dai, Linh Lieu, and David Rocke. Dimension reduction for classification with gene expression microarray data. *Stat Appl Genet Mol Biol*, 5:Article6, 2006.

[5] Guido Dornhege, Benjamin Blankertz, Gabriel Curio, and Klaus-Robert Müller. Increase information transfer rates in BCI by CSP extension to multi-class. In *Advances in Neural Information Processing Systems 16*, 2004.

[6] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

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

[8] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, 2nd edition, 1990.

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

[10] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, October 1996.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, August 2001.

[12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, February 1990.

[13] Xiaoming Huo, Michael Elad, Ana Georgina Flesia, Bob Muise, Robert Stanfill, Jerome Friedman, Bogdan Popescu, Jihong Chen, Abhijit Mahalanobis, and David L. Donoho. Optimal reduced-rank quadratic classifiers using the Fukunaga-Koontz transform, with applications to automated target recognition. SPIE's 7th International Symposium on Aerospace/Defense Sensing.

[14] Fernando De la Torre and Takeo Kanade. Multimodal oriented discriminant analysis. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 177–184, New York, NY, USA, 2005. ACM Press.

[15] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Proceedings of IEEE Neural Networks for Signal Processing Workshop 1999*, 1999.

[16] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehab. Eng.*, 8:441–446, December 2000.

[17] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Mach. Learn.*, 42(3):287–320, 2001.

[18] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In Craig Saunders, Marko Grobelnik, Steve R. Gunn, and John Shawe-Taylor, editors, *SLSFS*, volume 3940 of *Lecture Notes in Computer Science*, pages 34–51. Springer, 2005.

[19] Roman Rosipal, Leonard J. Trejo, and Bryan Matthews. Kernel PLS-SVM for linear and nonlinear classification. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 640–647. AAAI Press, 2003.

[20] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1):34–58, 2002.