DE GRUYTER OPEN

DOI: 10.2478/bile-2014-0005

Biometrical Letters
Vol 51 (2014), No. 1, 57-73

A kernel-based learning algorithm combining kernel discriminant coordinates and kernel principal components

Karol Deręgowski[1], Mirosław Krzyśko[1,2]

[1]President Stanislaw Wojciechowski Higher Vocational State School in Kalisz, Institute of Management, Nowy Świat 4, 62-800 Kalisz, Poland, e-mail: k.deregowski@pwsz.kalisz.pl
[2]Adam Mickiewicz University, Faculty of Mathematics and Computer Science, Umultowska 87, 61-614 Poznań, Poland, e-mail: mkrzysko@amu.edu.pl

SUMMARY

Kernel principal components (KPC) and kernel discriminant coordinates (KDC), which are the extensions of principal components and discriminant coordinates, respectively, from a linear domain to a nonlinear domain via the kernel trick, are two very popular nonlinear feature extraction methods. The kernel discriminant coordinates space has proven to be a very powerful space for pattern recognition. However, further study shows that there are still drawbacks in this method. To improve the performance of pattern recognition, we propose a new learning algorithm combining the advantages of KPC and KDC.

Key words: kernel principal components, kernel discriminant coordinates

1. Introduction

Classical principal component analysis (PCA) (Hotelling, 1933) was introduced as a technique for deriving a reduced set of orthogonal linear projections of a single collection of correlated variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, where the projections are ordered by decreasing variance. Principal component analysis is used, for example, in lossy data compression, pattern recognition, and image analysis. In addition to reducing dimensionality, principal component analysis can be used to discover important features of the data. Discovery in principal component analysis takes the form of graphical displays of the principal component scores. The first few principal component scores can reveal whether most of the data actually live on a linear subspace of R^p, and can

be used to identify outliers, distributional peculiarities, and clusters of points. The last few principal component scores show those linear projections of $\mathbf{X} = (X_1, X_2,\ldots, X_p)^T$ that have the smallest variance; any principal component with zero or near-zero variance is virtually constant, and hence can be used to detect collinearity, as well as outliers that affect the perceived dimensionality of the data.

When we have samples originating from *c* groups, we would often like to present them graphically, to see their configuration or to eliminate outlying observations. However it may be difficult to produce such a presentation even if only three features are observed, and with a higher number of features it becomes impossible. A different method must therefore be sought for presenting multidimensional data originating from multiple groups. To make the task easier, in the first step every p-dimensional observation $\mathbf{X} = (X_1, X_2,\ldots, X_p)^T$ can be transformed into a one-dimensional observation $u_1 = \mathbf{a}_1^T\mathbf{X} = a_{11}X_1 + a_{12}X_2 +\ldots+ a_{1p}X_p$, and the resulting one-dimensional observations can be presented graphically as points on a line. In the second step we can define a second linear combination $u_2= \mathbf{a}_2^T\mathbf{X}= a_{21}X_1 + a_{22}X_2 +\ldots+ a_{2p}X_p$, not correlated with the first, and present the observations graphically as points on a plane.

Generally, the aim is to construct new uncorrelated variables $u_1, u_2,\ldots, u_s$, $s \leq p$, which will be linear combinations of the original observations $X_1, X_2,\ldots, X_p$ and which will discriminate the *c* groups to a maximum degree; that is to say, in the new system the centres of the *c* groups will be maximally spaced, and the observations from a given group will be maximally concentrated around its centre. These new variables are called discriminant coordinates (see Seber, 1984, p. 270). They are also sometimes called canonical variates, but this name is misleading, because canonical variables with completely different properties occur in canonical correlation analysis. Another name used is “discriminant functions” – this is inappropriate because discriminant functions are surfaces that separate the *c* groups from one another. The space of discriminant coordinates is a space which is convenient for the

application of various classification methods (methods of discriminant analysis). In the case $c = 2$ we obtain only one discriminant coordinate, coinciding with the well-known linear discriminant function of Fisher (1936).

The linear projection method can be extremely useful in discovering low-dimensional structure when the data actually lie in a linear (or approximately linear) lower-dimensional subspace (called a manifold) M of the input space R^p. But what can we do if we know or suspect that the data actually lie on a low-dimensional nonlinear manifold, whose structure and dimensionality are both assumed unknown? We can then construct the nonlinear principal components and nonlinear discriminant coordinates.

Kernel principal components (KPC) and kernel discriminant coordinates (KDC), which are the extensions of principal components and discriminant coordinates, respectively, from a linear domain to a nonlinear domain via the kernel trick, are two very popular nonlinear feature extraction methods. The kernel discriminant coordinates space has proven to be a very powerful space for pattern recognition. However, further study shows that there are still drawbacks in this method. One of the major drawbacks of the kernel discriminant coordinate method is that it will lose the within-class scatter information, as for so-called “small sample size” problems, because all of the optimal discriminant vectors in this case are limited to the null space of the within-class scatter matrix – and this information is also important for pattern recognition. To improve the performance of pattern recognition, we propose another learning algorithm combining the advantages of KPC and KDC. Our proposed algorithm can be divided into three steps:

(1) compute the optimal discriminant vectors of KDC;

(2) compute the optimal vectors of KPC;

(3) use the two kinds of features for recognition.

The paper is organized as follows. In section 2, classical principal components are presented. Kernel principal components are presented in section 3. In section 4, classical discriminant coordinates are presented, and kernel discriminant coordinates are described in section 5. The discriminant

algorithm based on a combination of features from the kernel discriminant coordinates space and kernel principal components space is described in section 6. Finally, section 7 examines the quality of the new discriminant algorithm presented in this paper.

2. Classical principal component analysis

Assume that the random p-vector $\mathbf{X} = (X_1, X_2,\ldots, X_p)^T$ has mean $\boldsymbol{\mu}$ and the (p×p) covariance matrix $\boldsymbol{\Sigma}$. PCA seeks to replace the set of p (unordered and correlated) input variables, $X_1, X_2,\ldots, X_p$ by a (potentially smaller) set of t (ordered and uncorrelated) linear projections, $\xi_1,\ldots, \xi_t$ ($t \le p$), of the input variables,

$$\xi_j = \mathbf{b}_j^T \mathbf{X} = b_{j1}X_1 + \ldots + b_{jp}X_p \text{ , } j = 1, 2,\ldots, t; \tag{2.1}$$

where we minimize the loss of information due to the replacement.

In PCA, "information" is interpreted as the "total variation" of the original input variables,

$$\sum_{j=1}^{p} \mathrm{Var}\left(X_j\right) = \mathrm{tr}(\boldsymbol{\Sigma}) .$$

From the spectral decomposition theorem, we can write $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$, where the diagonal matrix $\boldsymbol{\Lambda}$ has as diagonal elements the eigenvalues $\{\lambda_j\}$ of $\boldsymbol{\Sigma}$, and the columns of $\mathbf{U}$ are the eigenvectors of $\boldsymbol{\Sigma}$. Thus the total variation is $\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_{j=1}^{p} \lambda_j$.

The jth coefficient vector, $\mathbf{b}_j = (b_{j1},\ldots,b_{jp})^T$, is chosen so that:

- The first t linear projections ξ_j, j = 1, 2,…, t, of $\mathbf{X}$ are ranked in importance according to their variances $\{\mathrm{Var}(\xi_j)\}$, which are listed in decreasing order of magnitude:

  $$\mathrm{Var}(\xi_1) \ge \mathrm{Var}(\xi_2) \ge \ldots \ge \mathrm{Var}(\xi_t).$$

- ξ_j is uncorrelated with all ξ_k, $k < j$.

The linear projections (1) are then known as the first *t* principal components of X.

In practice, we estimate the principal components using *N* independent observations, $\{\mathbf{X}_i, i = 1, 2,\ldots, N\}$, on $\mathbf{X}$. We estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = N^{-1}\sum_{i=1}^{N}\mathbf{X}_i$ and

$$\hat{\Sigma} = \mathbf{S} = (N-1)^{-1}\sum_{i=1}^{N}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T, \tag{2.2}$$

respectively.

The ordered eigenvalues of $\hat{\Sigma}$ are denoted by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$, and the eigenvector associated with the jth largest sample eigenvalue $\hat{\lambda}_j$ is the jth sample eigenvector $\hat{\mathbf{v}}_j$, j = 1, 2,…, p.

The coordinates of the projection of the ith observation $\mathbf{X}_i$ on the jth principal components are equal to:

$$\hat{\xi}_{ij} = \hat{\mathbf{v}}_j^T\mathbf{X}_i, \tag{2.3}$$

i = 1, 2,…, N, j = 1, 2,…, p.

A sample measure of how well the first *t* principal components represent the *p* original variables is given by the statistic

$$\frac{\hat{\lambda}_1 + \ldots + \hat{\lambda}_t}{\hat{\lambda}_1 + \ldots + \hat{\lambda}_p}$$

which is the proportion of the total sample variation that is explained by the first *t* sample principal components.

It is hoped that the sample variances of the first few sample PCs will be large, and that the remainder will be small enough for the corresponding set of sample PCs to be omitted. A variable that does not change much (relative to other variables) in independent measurements may be treated approximately as a constant, and so omitting such low-variance sample PCs and focusing exclusively on the high-variance sample PCs is a convenient way of reducing the dimensionality of the data set. For diagnostic and data analytic purposes, it

is usual to plot the first sample PC scores against the second sample PC scores, $(\hat{\xi}_{i1}, \hat{\xi}_{i2})$, where $\hat{\xi}_{ij}$ is given by (2.3), for i = 1,…, N, j = 1, 2.

3. Kernel principal component analysis

An approach that generalizes linear PCA is that of kernel PCA (Schölkopf et al., 1998). This is an application of what are called kernel methods.

Let $\mathbf{X}_i \in R^p$, i = 1, 2,…, N, be the input data points. We can think of kernel PCA as a two-step process:

- Nonlinearly transform the ith input data point $\mathbf{X}_i \in \mathbf{R}^p$ into a point $\Phi(\mathbf{X}_i)$ in an N_H-dimensional feature space H (the Hilbert space), where

  $$\mathbf{\Phi}(\mathbf{X}_i) = (\Phi_1(\mathbf{X}_i),\ldots, \Phi_{N_H}(\mathbf{X}_i))^T \in H,\ i = 1, 2,\ldots, N.$$

- The transformation $\mathbf{\Phi} : R^p \to H$ is called a feature transformation, and each $\{\Phi_j\}$ is a nonlinear transformation.
- Given $\mathbf{\Phi}(\mathbf{X}_1)$,…, $\mathbf{\Phi}(\mathbf{X}_N) \in H$, solve a linear PCA problem in the feature space H, which has a higher dimensionality than that of the input space (i.e. $N_H > p$).

Consider the data presented in Figure 1. Let $\mathbf{X}_i = (X_{i1}, X_{i2})^T$, and define $\mathbf{\Phi} : R^2 \to R^3$ by $\mathbf{\Phi}(\mathbf{X}_i) = \mathbf{\Phi}(X_{i1}, X_{i2}) = (X_{i1}^2, \sqrt{2}X_{i1}X_{i2}, X_{i2}^2) = (z_{i1}, z_{i2}, z_{i3})^T$.

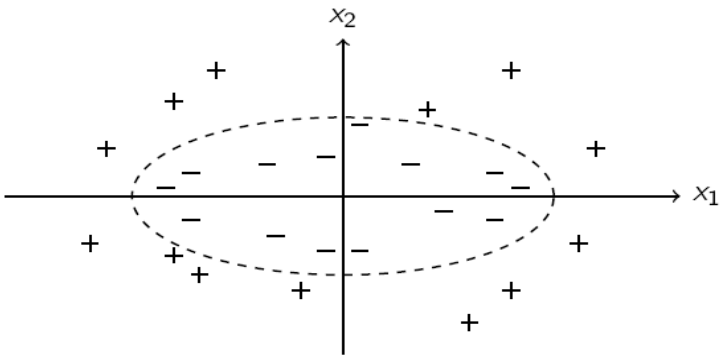

Figure 1. Original data in the plane. The data cannot be separated linearly

With this $\mathbf{\Phi}$, a difficult nonlinear classification problem in R^2 is converted to a standard linear classification task in R^3 (see Figure 2).

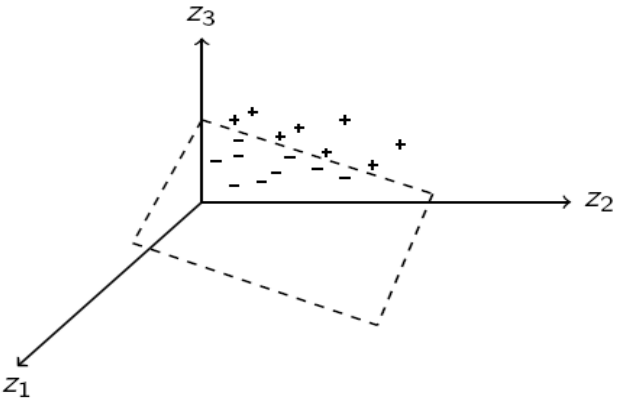

Figure 2. A three-dimensional representation of the pluses and minuses

Let $\mathbf{X}_i = (X_{i1}, X_{i2})^T$ and $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^T$ be two vectors in the input space R^2, and consider the transformation to R^3 used earlier. Let $\mathbf{\Phi}(\mathbf{X}_i)$ and $\mathbf{\Phi}(\mathbf{Y}_i)$ be two feature vectors generated by $\mathbf{X}_i$ and $\mathbf{Y}_i$. Now consider the inner product $\mathbf{\Phi}^T(\mathbf{X}_i)\,\mathbf{\Phi}(\mathbf{Y}_i)$ in the feature space. It takes the form

$$\mathbf{\Phi}^T(\mathbf{X}_i)\,\mathbf{\Phi}(\mathbf{Y}_i) = (\,X_{i1}^2, \sqrt{2}X_{i1}X_{i2}, X_{i2}^2\,)(\,Y_{i1}^2, \sqrt{2}Y_{i1}Y_{i2}, Y_{i2}^2\,)^T =$$

$$(X_{i1}Y_{i1} + X_{i2}Y_{i2})^2 = (\,\mathbf{X}_i^T\mathbf{Y}_i\,)^2 = k(\mathbf{X}_i, \mathbf{Y}_i). \qquad (3.1)$$

Equation (4) shows how an inner product based on $\mathbf{\Phi}$ converts to a function of the two inputs. Since choosing an inner product and performing computations with it in the feature space can quickly become computationally infeasible, it would be desirable to choose a function *k*, called a kernel, so as to summarize the geometry of the feature space vectors and ignore $\mathbf{\Phi}$ entirely.

Now the kernel trick can be applied. Suppose a function $k(\cdot,\cdot) : R^p \times R^p \rightarrow R$ operating on the input space can be found such that the feature space inner products are computed directly through *k* as in (4). Then explicit use of $\mathbf{\Phi}$ has been avoided, and yet results can be obtained as if $\mathbf{\Phi}$ were used. This direct

computation of feature space inner products without explicitly manipulating the feature space vectors themselves is known as the kernel trick.

The existence of the transformation $\mathbf{\Phi} : R^p \to H$ such that

$$\mathbf{\Phi}^T(\mathbf{X}_i)\mathbf{\Phi}(\mathbf{Y}_i) = k(\mathbf{X}_i, \mathbf{Y}_i)$$

is guaranteed by the following theorem.

Theorem 3.1 (Moore–Aronszajn) (Aronszajn, 1950). Let $k : R^p \times R^p \to R$ be a bivariate symmetric continuous real-valued function and H_k be a reproducing kernel Hilbert space (RKHS). Then there exists a transformation $\mathbf{\Phi} : R^p \to H_k$ such that

$$k(\mathbf{X}_i, \mathbf{Y}_i) = \mathbf{\Phi}^T(\mathbf{X}_i)\mathbf{\Phi}(\mathbf{Y}_i)$$

if and only if the matrix $K = (k_{ij})$ is nonnegative definite, where $k_{ij} = k(X_i, \mathbf{X}_j)$, $i, j = 1,\ldots,N$.

The matrix **K** is known as the kernel matrix. For a given bivariate function *k*, verifying the above conditions might not be easy. In practice, there exist many functions that have been shown to be valid kernels, and fortunately many of them deliver good performance on real-world data.

A short annotated list is presented in Table 1.

Table 1. Kernel functions

Kernel	$k(\mathbf{x}, \mathbf{y})$
Homogeneous polynomial kernel	$(\mathbf{x}^T \mathbf{y})^d$, d is an integer
Inhomogeneous polynomial kernel	$(\mathbf{x}^T \mathbf{y} + c)^d$, $c > 0$
Gaussian radial basis function	$\exp(-c \left\|\mathbf{x} - \mathbf{y}\right\|^2)$, $c > 0$
Laplacian	$\exp(-c\left\|\mathbf{x} - \mathbf{y}\right\|)$, $c > 0$

In order to carry out linear PCA in the feature space so that it mimics the standard treatment of PCA (as carried out in the input space), we have to find eigenvalues $\gamma \geq 0$ and nonzero eigenvectors $\mathbf{u} \in H$ of the estimated covariance matrix

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}\Phi(\mathbf{X}_i)\,\Phi^T(\mathbf{X}_i) \tag{3.2}$$

of the centred and nonlinearly transformed input vectors. The eigenequation $\mathbf{Cu} = \gamma\mathbf{u}$, where $\mathbf{u}$ is the eigenvector corresponding to the eigenvalue $\gamma \geq 0$ of $\mathbf{C}$, can be written in the equivalent form

$$\mathbf{\Phi}^T(\mathbf{X}_i)\,\mathbf{Cu} = \gamma\,\mathbf{\Phi}^T(\mathbf{X}_i)\mathbf{u}\ ,\ i = 1, 2,\ldots, N. \tag{3.3}$$

Because

$$\mathbf{Cu} = \frac{1}{N-1}\sum_{i=1}^{N}\Phi(\mathbf{X}_i)\,\Phi^T(\mathbf{X}_i)\,\mathbf{u}$$

all solutions $\mathbf{u}$ with nonzero eigenvalue γ are contained in the span of $\mathbf{\Phi}(\mathbf{X}_1),\ldots, \mathbf{\Phi}(\mathbf{X}_N)$. Hence there exist coefficients α_k, k = 1, 2,…, N, such that

$$\mathbf{u} = \sum_{k=1}^{N}\alpha_k\Phi(\mathbf{X}_k)\,. \tag{3.4}$$

Substituting (3.4) for $\mathbf{u}$ in (3.3), we obtain that

$$\frac{1}{N-1}\sum_{j=1}^{N}\alpha_j\Phi^T(\mathbf{X}_i)\sum_{k=1}^{N}\Phi(\mathbf{X}_k)\Phi^T(\mathbf{X}_k)\Phi(\mathbf{X}_j) = \gamma\sum_{k=1}^{N}\alpha_k\Phi^T(\mathbf{X}_i)\Phi(\mathbf{X}_k)\,, \tag{3.5}$$

for all i = 1, 2,…, N.

The eigenequation (3.5) can be written as $\mathbf{K}^2\,\boldsymbol{\alpha} = N\,\gamma\,\mathbf{K}\,\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1,\ldots, \alpha_N)^T$, or as

$$\mathbf{K}^2\,\boldsymbol{\alpha} = \tilde{\gamma}\,\mathbf{K}\,\boldsymbol{\alpha}, \tag{3.6}$$

where $\tilde{\gamma} = (N-1)\,\gamma$, $\mathbf{K} = (k_{ij})$ and $k_{ij} = k(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{\Phi}^T(\mathbf{X}_i)\,\mathbf{\Phi}(\mathbf{X}_j)$, i, j = 1, 2,…, N.

To find solutions of (3.6), we solve the eigenvalue problem

$$\mathbf{K}\boldsymbol{\alpha} = \tilde{\gamma}\boldsymbol{\alpha} \tag{3.7}$$

for nonzero eigenvalues. Clearly, all solutions of (3.7) also satisfy (3.6). Moreover, it can be shown that any additional solutions of (3.7) do not make a difference in the expansion (3.4) and thus are not of interest to us.

We assumed earlier that the vectors $\{\mathbf{\Phi}(\mathbf{X}_i)\}$, i = 1, 2,…, N, are centred. In the general case we cannot centre the vectors $\{\mathbf{\Phi}(\mathbf{X}_i)\}$, because we do not know the form of the function $\mathbf{\Phi}$. Let

$$\widetilde{\Phi}(\mathbf{X}_i) = \Phi(\mathbf{X}_i) - \frac{1}{N}\sum_{k=1}^{N}\Phi(\mathbf{X}_k)$$

and

$$\widetilde{\mathbf{K}} = (\widetilde{k}(\mathbf{X}_i, \mathbf{X}_j)) = (<\widetilde{\Phi}(\mathbf{X}_i), \widetilde{\Phi}(\mathbf{X}_j)>)$$

i, j = 1, 2,…, N. We cannot compute the matrix $\widetilde{\mathbf{K}}$ directly, but we can express it in terms of the matrix $\mathbf{K}$: $\widetilde{\mathbf{K}} = \mathbf{PKP}$, where $\mathbf{P} = \mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N^T$ is the centring matrix.

Hence in the general case the construction of kernel principal components must be based on the matrix $\widetilde{\mathbf{K}}$. Because the kernel matrix $\mathbf{K}$ is nonnegative definite, $\widetilde{\mathbf{K}}$ is nonnegative define also. This results from the fact (Seber, 1984, p. 521) that if $\mathbf{A} \geq 0$, then $\mathbf{CAC}^T \geq 0$. In our case $\mathbf{C} = \mathbf{P}$, where $\mathbf{P}$ is a symmetric matrix, i.e. $\mathbf{P} = \mathbf{P}^T$. Hence, all eigenvalues of $\widetilde{\mathbf{K}}$ are nonnegative.

The coordinates of the projection of the data matrix $\begin{bmatrix} \mathbf{\Phi}^T(\mathbf{X}_1) \\ \vdots \\ \mathbf{\Phi}^T(\mathbf{X}_N) \end{bmatrix}$

on the jth kernel principal components are equal to $\widetilde{\mathbf{K}}\hat{\boldsymbol{\alpha}}_j$, j = 1, 2,…, N.

4. Classical discriminant coordinates

Consider a set of *N* p-dimensional observations $\mathbf{X}_1$, $\mathbf{X}_2$,…, $\mathbf{X}_N$, where the vectors $\mathbf{X}_i \in R^p$ are grouped in *c* disjoint classes, and every $\mathbf{X}_i$ belongs to one and only one class.

Let V = {1, 2,…, N} be the set of indices of the set of observations $\mathbf{X}_1$, $\mathbf{X}_2$,…, $\mathbf{X}_N$, and let V be partitioned into *c* disjoint subsets V_i such that

$V_i \cap V_j = \varnothing$ for $i \neq j$, $\bigcup_{j=1}^{c} V_j = V$, and each V_j contains n_j elements such that $\sum_{j=1}^{c} n_j = N$.

Let $\overline{\mathbf{X}} = N^{-1}\sum_{i=1}^{N}\mathbf{X}_i$ be the sample mean, and let $\overline{\mathbf{X}}_j = n_j^{-1}\sum_{i\in V_j}\mathbf{X}_i$ be the mean of the jth class, for j =1, 2,…, c.

We denote by $\mathbf{S}_t$ the total scatter matrix, and by $\mathbf{S}_b$ the between-class scatter matrix. We have

$$\mathbf{S}_t = \sum_{i=1}^{N}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T,$$

$$\mathbf{S}_b = \sum_{j=1}^{c} n_j(\overline{\mathbf{X}}_j - \overline{\mathbf{X}})(\overline{\mathbf{X}}_j - \overline{\mathbf{X}})^T.$$

We seek a set of vectors $\mathbf{a}_i \in R^p$ which maximize the measure of separation of the *c* classes

$$J(\mathbf{a}) = \frac{\mathbf{a}^T\mathbf{S}_b\mathbf{a}}{\mathbf{a}^T\mathbf{S}_t\mathbf{a}},$$

subject to the additional restriction $\mathbf{a}_i^T\mathbf{S}_t\mathbf{a}_j = \delta_{ij}$ (Kronecker delta), which means that the variables $u_i = \mathbf{a}_i^T\mathbf{X}$, called discriminant coordinates, are uncorrelated. The $\mathbf{a}_i$ are called directional vectors. Finding the directional vectors reduces to solving the following generalized eigenvalue problem:

$$\mathbf{S}_b\mathbf{a}_i = \lambda_i\mathbf{S}_t\mathbf{a}_i, \; i = 1, 2, \ldots, s, \tag{4.1}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s \geq \lambda_{s+1} = 0$ and $s = \text{rank}(\mathbf{S}_b) \leq \min\{p, c-1\}$.

Because $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$, where $\mathbf{S}_w$ is the pooled within-class scatter matrix, finding classical discriminant coordinates is equivalent to finding a solution to the generalized eigenvalue problem in the form $\mathbf{S}_b\mathbf{a} = \lambda/(1 - \lambda)\mathbf{S}_w\mathbf{a}$.

Using matrix notation, the generalized problem (4.1) can be given in the form

$$\mathbf{S}_b\mathbf{A} = \mathbf{S}_t\mathbf{A}\mathbf{\Lambda}, \tag{4.2}$$

where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_s)$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_s)$.

If $\mathbf{S}_t$ is a non-singular matrix, we obtain $\mathbf{S}_t^{-1}\mathbf{S}_b\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}$, which is an ordinary eigenvalue problem.

Hence the pair $(\lambda_i, \mathbf{a}_i)$, where λ_i is an eigenvalue of the matrix $\mathbf{S}_t^{-1}\mathbf{S}_b$ and $\mathbf{a}_i$ is the corresponding eigenvector, can be used to construct discriminant coordinates. Because $\text{rank}(\mathbf{S}_b) \leq c-1$, we obtain a discriminant coordinate space with dimension at least c–1.

5. Kernel discriminant coordinates

Kernel discriminant coordinates were introduced independently by Mika et al. (1999) and by Baudat and Anouar (2000). The method is described in a book by Shawe-Taylor and Cristianini (2004).

The original space $\mathbf{R}^p$ is transformed non-linearly into a feature space H_k

$$\boldsymbol{\Phi} : \mathbf{R}^p \rightarrow H_k,$$

where $\boldsymbol{\Phi}$ is a vector-valued function, and H_k is a reproducing kernel Hilbert space (RKHS).

The vector $\boldsymbol{\Phi}(\mathbf{X}_i) = \widetilde{\mathbf{X}}_i$ is called the feature vector corresponding to the observation $\mathbf{X}_i \in \mathbf{R}^p$, i = 1, 2,…, N. The non-linear transformation $\boldsymbol{\Phi}$ is in general not known; however we select a known form of the non-negative definite kernel function

$$k(\mathbf{X}, \mathbf{Y}) = \Phi^T(\mathbf{X})\Phi(\mathbf{Y}) = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}.$$

Let $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_t$ denote respectively the between-class and total scatter matrices in the feature space. We have

$$\widetilde{\mathbf{S}}_t = \sum_{i=1}^{N}(\widetilde{\mathbf{X}}_i - \overline{\widetilde{\mathbf{X}}})(\widetilde{\mathbf{X}}_i - \overline{\widetilde{\mathbf{X}}})^T,$$

$$\widetilde{\mathbf{S}}_b = \sum_{j=1}^{c} n_j(\overline{\widetilde{\mathbf{X}}}_j - \overline{\widetilde{\mathbf{X}}})(\overline{\widetilde{\mathbf{X}}}_j - \overline{\widetilde{\mathbf{X}}})^T,$$

where $\widetilde{\overline{\mathbf{X}}}_k$ and $\widetilde{\overline{\mathbf{X}}}$ are respectively the mean of the kth class and total mean in the feature space. Finding discriminant coordinates in the feature space reduces to solving the following optimization problem:

$$\mathbf{b}_i = \arg\max \frac{\mathbf{b}^T \widetilde{\mathbf{S}}_b \mathbf{b}}{\mathbf{b}^T \widetilde{\mathbf{S}}_t \mathbf{b}}, \quad \mathbf{b}_i^T \widetilde{\mathbf{S}}_b \mathbf{b}_j = \delta_{ij}. \tag{5.3}$$

We know that there exist coefficients b_{ij} such that $\mathbf{b}_i = \sum_{j=1}^{n} b_{ij} \widetilde{\mathbf{X}}_j$. Hence the optimization problem (5.3) is equivalent to the problem

$$\mathbf{b}_i = \arg\max \frac{\mathbf{b}^T \widetilde{\mathbf{K}} \mathbf{D} \widetilde{\mathbf{K}} \mathbf{b}}{\mathbf{b}^T \widetilde{\mathbf{K}} \widetilde{\mathbf{K}} \mathbf{b}}, \quad \mathbf{b}_i^T \widetilde{\mathbf{K}} \widetilde{\mathbf{K}} \mathbf{b}_j = \delta_{ij}$$

and the optimum vectors $\mathbf{b}_1, \ldots, \mathbf{b}_s$ are equal to the eigenvectors corresponding to the maximum eigenvalues in the generalized problem

$$\widetilde{\mathbf{K}} \mathbf{D} \widetilde{\mathbf{K}} \mathbf{b} = \lambda \widetilde{\mathbf{K}} \widetilde{\mathbf{K}} \mathbf{b}, \quad \widetilde{\mathbf{K}} = \mathbf{P} \mathbf{K} \mathbf{P}, \tag{5.4}$$

where $\mathbf{K} = (k_{ij})$ ($k_{ij} = k(\mathbf{X}_i, \mathbf{X}_j)$) is the kernel matrix, $\mathbf{P} = \mathbf{I}_N - N^{-1} \mathbf{1}_N \mathbf{1}_N^T$ is the centring matrix, and the matrix $\mathbf{D}$ is defined by

$$D_{ij} = \begin{cases} \dfrac{1}{n_k}, & \text{if } \mathbf{X}_i \text{ and } \mathbf{X}_j \text{ belong to the kth class,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_s]$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_s)$. In matrix notation the problem of (5.4) has the form

$$\widetilde{\mathbf{K}} \mathbf{D} \widetilde{\mathbf{K}} \mathbf{B} = \widetilde{\mathbf{K}} \widetilde{\mathbf{K}} \mathbf{B} \mathbf{\Lambda}. \tag{5.5}$$

Solving the generalized eigenvalue problem presents certain difficulties, because both matrices are non-negative definite.

One of the ways of obtaining an approximate solution to (5.5), also used in ridge regression, is to regularize the matrix $\widetilde{\mathbf{K}} \widetilde{\mathbf{K}}$, that is to replace $\widetilde{\mathbf{K}} \widetilde{\mathbf{K}}$ with a new non-singular matrix $\widetilde{\mathbf{K}} \widetilde{\mathbf{K}} + \varepsilon \mathbf{I}$ (Friedman, 1989; Mika et al., 1999). We then solve the new generalized eigenvalue problem

$$\widetilde{\mathbf{K}} \mathbf{D} \widetilde{\mathbf{K}} \mathbf{B} = (\widetilde{\mathbf{K}} \widetilde{\mathbf{K}} + \varepsilon \mathbf{I}) \mathbf{B} \mathbf{\Lambda}.$$

Because the matrix $\widetilde{\mathbf{K}}\widetilde{\mathbf{K}} + \varepsilon\mathbf{I}$ is non-singular, the above problem can be reduced to the classical eigenvalue problem

$$(\widetilde{\mathbf{K}}\widetilde{\mathbf{K}} + \varepsilon\mathbf{I})^{-1}\widetilde{\mathbf{K}}\mathbf{D}\widetilde{\mathbf{K}}\mathbf{B} = \mathbf{B}\boldsymbol{\Lambda},$$

whose solution has the form

$$\mathbf{B} = \mathrm{eig}[(\widetilde{\mathbf{K}}\widetilde{\mathbf{K}} + \varepsilon\mathbf{I})^{-1}\widetilde{\mathbf{K}}\mathbf{D}\widetilde{\mathbf{K}}].$$

Hence the pair $(\lambda_i, \mathbf{b}_i)$, where λ_i is an eigenvalue of the matrix $(\widetilde{\mathbf{K}}\widetilde{\mathbf{K}} + \varepsilon\mathbf{I})^{-1}\widetilde{\mathbf{K}}\mathbf{D}\widetilde{\mathbf{K}}$ and $\mathbf{b}_i$ is the corresponding eigenvector, can be used to construct kernel discriminant coordinates. Because rank $(\widetilde{\mathbf{K}}\mathbf{D}\widetilde{\mathbf{K}}) \leq$ c–1, we obtain a discriminant coordinate space with dimension at least c–1. The coordinates of the projection of the data matrix

$$\begin{bmatrix} \widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}(\mathbf{X}_1) \\ \vdots \\ \widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}(\mathbf{X}_N) \end{bmatrix}$$

on the jth kernel principal components are equal to $\widetilde{\mathbf{K}}\hat{\boldsymbol{\alpha}}_j$, j = 1, 2,…, c–1.

6. A new discriminant space

Consider the (c–1)-dimensional space of the first kernel discriminant variables. The training sample $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ transformed into this space will be denoted by $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N\}$, where $\mathbf{Y}_i \in R^{c-1}$. We then further consider the (c–1)-dimensional space of the first kernel principal components. The training sample transformed into this space will be denoted by $\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N\}$, where $\mathbf{Z}_i \in R^{c-1}$.

We will create a new 2(c–1)-dimensional space determined by the first (c–1) kernel discriminant variables and the first (c–1) kernel principal components. The directional vectors determining this new space are normed so that their length is equal to 1. The new space combines the advantages of the space of kernel discriminant coordinates and the space of kernel principal components. In this space it is possible to apply a variety of classification algorithms, obtaining an improvement in classification quality.

7. Example

Three types of discriminant spaces were considered: the (c−1)-dimensional space of kernel principal components described in section 3, the (c−1)-dimensional space of kernel discriminant coordinates described in section 5, and the combined 2(c−1)-dimensional space of kernel discriminant coordinates and kernel principal components.

The kernel spaces were built using the Gaussian kernel

$$k(\mathbf{x},\mathbf{y}) = \exp(-d\,\|\mathbf{x}-\mathbf{y}\|^2),$$

where d is a positive constant. We consider the lower triangle in the table of squared distances between elements of the training sample $\{\mathbf{X}_1, \mathbf{X}_2,\ldots, \mathbf{X}_N\}$. The value of d was taken to be the reciprocal of the arithmetic mean of the elements of the lower triangle. The matrix $\widetilde{\mathbf{K}}\widetilde{\mathbf{K}} + \varepsilon\mathbf{I}$ was taken with $\varepsilon = 10^{-5}$.

To check the usefulness of the three aforementioned discriminant spaces in the process of classification, each space was used to classify objects belonging to 25 different training samples from the University of California database (Bache and Lichman, 2013).

Classification was performed by the method of linear discriminant functions. Table 2 shows that the largest percentage of correct classifications is achieved in the combined space of kernel discriminant coordinates and kernel principal components. This is the case for 23 of the training samples. For one training sample (Ring) we have the same percentage of correct classifications in the kernel discriminant coordinate space as in the combined space of kernel discriminant coordinates and kernel principal components. For one training sample (Hayes-Roth) the kernel discriminant coordinate space gives better results than the combined space. The results imply that the combined space should be chosen as the best discriminant space.

Table 2. Comparison of percentages of correct classification

Data	N	c	p	Percentage of correct classification
Balance	625	3	4	87.36 (546/625) 89.12 (557/625) 90.88 (568/625)
Car	1728	4	6	69.56 (1202/1728) 83.33 (1440/1728) 86.46 (1494/1728)
Choice	1473	3	9	45.76 (674/1473) 45.96 (677/1473) 54.11 (797/1473)
Danesym	300	3	2	67.33 (202/300) 68.33 (205/300) 71.00 (213/300)
Danford	45	4	3	42.22 (19/45) 44.44 (20/45) 53.33 (24/45)
Dermatology	358	6	34	75.14 (269/358) 83.80 (300/358) 95.25 (341/358)
EnVowel	990	11	10	75.96 (752/990) 93.84 (929/990) 95.86 (949/990)
Glass	214	6	9	62.15 (133/214) 64.49 (138/214) 73.83 (158/214)
HayesRoth	132	3	3	56.82 (75/132) 85.61 (113/132) 80.30 (106/132)
Image	2310	7	19	76.84 (1637/2310) 78.40 (1811/2310) 87.06 (2011/2310)
Lenses	24	3	4	70.83 (17/24) 62.50 (15/24) 75.00 (18/24)
Oil	56	3	5	76.79 (40/56) 76.79 (40/56) 80.36 (45/56)
PlVowel	1130	6	5	62.74 (709/1130) 63.98 (723/1130) 73.45 (830/1130)
Ring	1000	2	2	65,90 (659/1000) 71,70 (717/1000) 71,70 (717/1000)
Risk	87	3	2	97.70 (85/87) 97.70 (85/87) 100.00 (87/87)
Smith	45	4	3	55.56 (25/45) 60.00 (27/45) 66.67 (30/45)
Soybean	47	4	35	85.11(40/47) 87.23(41/47) 95.74(45/47)
Tae	151	3	5	41.72 (63/151) 45.70 (69/151) 46.36 (70/151)
Three-class	300	3	2	90.00 (270/300) 91.33 (274/300) 92.67 (278/300)
Thyroid	215	3	5	86.51 (186/215) 85.58 (184/215) 93.49 (201/215)
Wale-form	900	3	21	85.11 (766/900) 88.33 (795/900) 93.67 (843/900)
Wine	178	3	13	70.22 (125/178) 71.34 (127/178) 71.91 (128/178)
WineQuality	1599	6	11	49.22 (787/1599) 49.53 (792/1599) 51.34 (821/1599)
Vehicle	846	4	18	43.38 (367/846) 47.28 (400/846) 47.75 (404/846)
Zoo	101	7	16	96.04 (97/101) 96.04 (97/101) 99.01 (100/101)

REFERENCES

Aronszajn N. (1950): Theory of reproducing kernels. Transactions of the American Mathematical Society 68: 337–404.

Badat G., Anouar F. (2000): Generalized discriminant analysis using a kernel approach. Neural Computation 12: 2385–2404.

Bache K., Lichman M. (2013): UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Fisher R.A. (1936): The use of multiple measurements in taxonomic problem. Annals of Eugenics 7: 179–188.

Friedman J.H. (1989): Regularized discriminant analysis. Journal of the American Statistical Association 84: 165–175.

Hotelling H. (1933): Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24: 417–441, 498–520.

Mika S., Rätsch G., Weston J., Schölkopf B., Müller K.R. (1999): Fisher discriminant analysis with kernels. In Y.H. Hu, J. Larsen, E. Wilson, and S. Douglas (eds.), Neural Networks for Signal Processing IV: 41–48.

Schölkopf B., Smola A., Müller K.B. (1998): Nonlinear component analysis as a kernel eigenvalues problem. Neural Computation 10: 1299–1319.

Seber G.A.F. (1984): Multivariate Observations. Wiley, New York.

Shawe-Taylor J., Cristianini N. (2004): Kernel methods for pattern analysis. Cambridge University Press, Cambridge, UK.