A Support Vector Machine Formulation to PCA Analysis and its Kernel Version

J.A.K. Suykens, **T. Van Gestel**, **J. Vandewalle**, **B. De Moor**

Katholieke Universiteit Leuven
Department of Electrical Engineering, ESAT-SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Tel: 32/16/32 18 02   Fax: 32/16/32 19 70
Email: johan.suykens@esat.kuleuven.ac.be

(Corresponding Author: Johan Suykens)

ESAT-SCD-SISTA Technical Report 2002-68
submitted to IEEE-TNN

Abstract

In this letter we present a simple and straightforward support vector machine formulation to the problem of PCA analysis in dual variables. By considering a mapping to a high dimensional feature space and application of the kernel trick (Mercer theorem) kernel PCA is obtained as introduced by Schölkopf *et al.* While least squares support vector machine classifiers have a natural link with kernel Fisher discriminant analysis (minimizing the within class scatter around targets +1 and -1), for PCA analysis one can take the interpretation of a one-class modelling problem with zero target value around which one maximizes the variance. The score variables are interpreted as error variables within the problem formulation. In this way primal-dual constrained optimization problem interpretations to linear and kernel PCA analysis are obtained in a similar style as for (LS)-SVM classifiers.

Keywords: PCA analysis, kernel PCA, support vector machines, kernel methods

1 Introduction

Support Vector Machines (SVMs) as originally introduced by Vapnik within the area of statistical learning theory and structural risk minimization [20] have proven to work successfully on many applications of nonlinear classification and function estimation. The problems are formulated as convex optimization problems, usually quadratic programs, for which the dual problem is solved. Within the models and the formulation one makes use of the kernel trick which is based on the Mercer theorem related to positive definite kernels. One can plug in any positive definite kernel for a support vector machine classifier or regressor with as typical choices linear, polynomial and RBF kernels. The work on SVMs has also stimulated the research on kernel based learning methods in general in recent years [15]. The conceptual idea of generalizing an existing linear technique to a nonlinear version by applying the kernel trick has become an area of active research. One important result in this direction is the extension of linear Principal Component Analysis (PCA) [8] to kernel PCA, shown by Schölkopf *et al.* [13, 14].

The aim of this paper is to present a new simple and straightforward formulation to PCA analysis and its kernel version. The formulation is in the style of SVMs, in the sense that one starts from a constrained optimization problem in primal weight space with incorporation of a regularization term and one solves the dual problem after application of the kernel trick. The nonlinear version of the formulation yields a solution which is equivalent to kernel PCA.

The formulation is made in a similar fashion as least squares support vector machine classifiers (LS-SVMs) [16]. For classification there is a close connection between LS-SVMs and kernel Fisher discriminant analysis [1, 9, 15, 19] as the *within class scatter* is minimized around targets +1 and -1. The PCA analysis problem is interpreted as a one-class modelling problem with target value zero around which one maximizes the variance. This results into a sum squared error cost function with regularization. The score variables are taken as additional error variables. As a result this paper shows an extension of LS-SVM formulations to the area of unsupervised learning. The LS-SVM approach is closely related to regularization networks, Gaussian processes, kernel ridge regression and reproducing kernel Hilbert spaces (RKHS) [3, 11, 12, 21, 22]. On the other hand, the formulations are more closely related to standard SVMs with explicit primal-dual interpretations from the viewpoint of optimization theory. Extensions of LS-SVMs have been given also to recurrent networks and control [18]. Issues of robustness and sparseness have been discussed in [17].

This paper is organized as follows. In Section 2 we briefly state the classical and well-known problem of linear PCA analysis. In Section 3 we present the new support vector machine formulation to linear PCA in dual variables. We also discuss the issue of taking into account an additional bias term and the link with centering. Finally, in Section 4 the nonlinear version is given which leads to kernel PCA.

2 Classical principal component analysis formulation

Consider a given set of data $\{x_k\}_{k=1}^N$ with $x_k \in \mathbb{R}^n$ and N given data points for which one aims at finding projected variables $w^T x_k$ with maximal variance [7, 8, 10]. This means

$$\begin{aligned} \max_w \mathrm{Var}(w^T x) &= \mathrm{Cov}(w^T x, w^T x) = \sum_{k=1}^N (w^T x_k)^2 \\ &= w^T C\, w \end{aligned} \tag{1}$$

where $C = \sum_{k=1}^{N} x_k x_k^T$ by definition. One optimizes this objective function under the constraint that $w^T w = 1$. This gives the constrained optimization

$$\mathcal{L}(w;\lambda) = \frac{1}{2} w^T C w - \lambda (w^T w - 1) \tag{2}$$

with Lagrange multiplier λ where the solution follows from $\partial \mathcal{L}/\partial w = 0$, $\partial \mathcal{L}/\partial \lambda = 0$ and is given by the eigenvalue problem

$$C w = \lambda w. \tag{3}$$

The matrix C is symmetric and positive semidefinite. The eigenvector w corresponding to the largest eigenvalue determines the projected variable with maximal variance. Efficient and reliable numerical methods are discussed e.g. in [5].

3 A support vector machine formulation to linear principal component analysis

3.1 PCA analysis as a one-class modelling problem

Let us now reformulate the PCA problem as follows:

$$\max_{w} \sum_{k=1}^{N} (0 - w^T x_k)^2 \tag{4}$$

where 0 is considered as a single target value. While for Fisher discriminant analysis one considers two target values +1 and -1 that represent the two classes, in the PCA analysis case a zero target value is considered. Hence, one has in fact a one class modelling problem, but with a different objective function in mind. For Fisher discriminant analysis one aims at minimizing the within scatter around the targets, while for PCA analysis one is interested in finding the direction(s) for which the variance is maximal (Fig.1).

This interpretation of the problem leads to the following primal optimization problem

$$\begin{aligned} \boxed{\mathrm{P}}: \quad \max_{w,e} J_{\mathrm{P}}(w,e) = & \quad \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \\ \text{such that} & \quad e_k = w^T x_k, \ k = 1, ..., N. \end{aligned} \tag{5}$$

This formulation states that one considers the difference between $w^T x_k$ (the projected data points to the target space) and the value 0 as error variables. The projected variables correspond to what one calls the *score* variables. These error variables are maximized for the given N data points while keeping the norm of w small by the regularization term. The value γ is a positive real constant. The Lagrangian becomes

$$\mathcal{L}(w, e; \alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T x_k \right) \tag{6}$$

with conditions for optimality given by

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k x_k \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 \quad \rightarrow \quad \alpha_k = \gamma e_k, \qquad\qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \quad \rightarrow \quad e_k - w^T x_k = 0, \quad k = 1, ..., N. \end{cases} \tag{7}$$

By elimination of the variables e, w one obtains

$$\frac{1}{\gamma} \alpha_k - \sum_{l=1}^{N} \alpha_l x_l^T x_k = 0 \ , \quad k = 1, ..., N. \tag{8}$$

By defining $\lambda = 1/\gamma$ one has the following dual symmetric eigenvalue problem

$$\boxed{\text{D}} : \quad \text{solve in } \alpha :$$

$$\begin{bmatrix} x_1^T x_1 & \ldots & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & \ldots & x_N^T x_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} \tag{9}$$

which is the dual interpretation of (3). The vector of dual variables $\alpha = [\alpha_1; ...; \alpha_N]$ is an eigenvector of the Gram matrix and λ is the corresponding eigenvalue. In order to obtain the maximal variance one selects the eigenvector corresponding to the largest eigenvalue.

The score variables become

$$z(x) = w^T x = \sum_{l=1}^{N} \alpha_l x_l^T x \tag{10}$$

where α is the eigenvector corresponding to the largest eigenvalue for the first score variable. Note that all eigenvalues are positive and real because the matrix is symmetric and positive definite. One has in fact N local minima as solution to the problem for which one selects the solution of interest. The optimal solution is the eigenvector corresponding to the largest eigenvalue because in that case

$$\sum_{k=1}^{N} (w^T x_k)^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{1}{\gamma^2} \alpha_k^2 = \lambda_{max}^2, \tag{11}$$

where $\sum_{k=1}^{N} \alpha_k^2 = 1$ for the normalized eigenvector. For the different score variables one selects the eigenvectors corresponding to the different eigenvalues. The score variables are decorrelated from each other due to the fact that the α eigenvectors are orthonormal. According to [8], one can also additionally stress within the constraints of the formulation that the w vectors related to subsequent scores are orthogonal to each other.

PCA analysis is usually applied to centered data. Therefore one better considers the problem

$$\max_{w} \sum_{k=1}^{N} [w^T (x_k - \mu_x)]^2 \tag{12}$$

where $\mu_x = (1/N) \sum_{k=1}^{N} x_k$. The same derivations can be made and one finally obtains a centered Gram matrix as a result. One also sees that solving the problem in w is typically advantageous for large data sets, while for fewer given data in huge dimensional input spaces one better solves the dual problem. The approach of taking the eigenvalue decomposition of the centered Gram matrix is also done in principal co-ordinate analysis [6, 8].

3.2 Including a bias term

While in PCA analysis one usually centers the data, the new interpretation to PCA analysis also offers a way to analyse the use of a bias term.

The score variables are then

$$z(x) = w^T x + b \tag{13}$$

and one aims at optimizing the following objective

$$\max_{w,b} \sum_{k=1}^{N} [0 - (w^T x + b)]^2. \tag{14}$$

Therefore, one formulates the primal optimization problem

$$\boxed{\mathrm{P}}: \quad \begin{aligned} \max_{w,e} J_{\mathrm{P}}(w,e) = & \quad \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \\ \text{such that} & \qquad e_k = w^T x_k + b, \quad k = 1, ..., N \end{aligned} \tag{15}$$

with Lagrangian

$$\mathcal{L}(w, e; \alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T x_k - b \right) \tag{16}$$

giving the conditions for optimality

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k x_k \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow \quad \alpha_k = \gamma e_k, \qquad\qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow \quad \sum_{k=1}^{N} \alpha_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow \quad e_k - w^T x_k - b = 0, \quad k = 1, ..., N. \end{cases} \tag{17}$$

Applying $\sum_{k=1}^{N} \alpha_k = 0$ the last condition delivers an expression for the bias term

$$b = -\frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l x_l^T x_k. \tag{18}$$

By defining $\lambda = 1/\gamma$ one obtains the dual problem

$\boxed{\mathrm{D}}$: solve in α :

$$\begin{bmatrix} (x_1 - \mu_x)^T (x_1 - \mu_x) & \dots & (x_1 - \mu_x)^T (x_N - \mu_x) \\ \vdots & & \vdots \\ (x_N - \mu_x)^T (x_1 - \mu_x) & \dots & (x_N - \mu_x)^T (x_N - \mu_x) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} \tag{19}$$

which is an eigenvalue decomposition of the centered Gram matrix

$$\Omega_c \alpha = \lambda \alpha \tag{20}$$

with $\Omega_c = M_c \Omega M_c$ where $M_c = I - 1_v 1_v^T / N$, $1_v = [1; 1; ...; 1]$ and $\Omega_{kl} = x_k^T x_l$ for $k, l = 1, ..., N$. This eigenvalue problem follows from

$$\frac{1}{\gamma} \alpha_k - \sum_{l=1}^{N} \alpha_l x_l^T x_k + \frac{1}{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l x_l^T x_k = 0$$

by taking into account the fact that $\sum_{k=1}^{N} \alpha_k = 0$. One also sees that considering a bias term in the problem formulation automatically leads to a centering of the matrix.

The score variables equal

$$z(x) = w^T x + b = \sum_{l=1}^{N} \alpha_l x_l^T x + b \tag{21}$$

where α is the eigenvector corresponding to the largest eigenvalue and

$$\sum_{k=1}^{N} (w^T x_k + b)^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{1}{\gamma^2} \alpha_k^2 = \lambda_{max}^2. \tag{22}$$

3.3 The reconstruction problem

Another interpretation of PCA analysis can be made in terms of the reconstruction error [2, 8, 10]

$$\min \sum_{k=1}^{N} \|x_k - \tilde{x}_k\|_2^2 \tag{23}$$

where $\tilde{x}_k$ are variables reconstructed from the score variables. Let us denote now the data matrix and the matrix with selected score variables as $X = [x_1 x_2 ... x_N] \in \mathbb{R}^{n \times N}$ and $Z = [z_1 z_2 ... z_N] \in \mathbb{R}^{n_s \times N}$, respectively, where n_s denotes the number of selected variables which determines the dimensionality reduction.

In the context of linear PCA analysis one considers a linear mapping from the scores to the reconstructed variables. In order to be able to handle also the bias term formulation one can take

$$\tilde{x} = V z + \delta \tag{24}$$

and minimize

$$\min_{V,\delta} \sum_{k=1}^{N} \|x_k - (V z_k + \delta)\|_2^2. \tag{25}$$

In matrix form this leads to a least squares solution

$$[V\ \delta] = X \begin{bmatrix} Z \\ 1_v^T \end{bmatrix} \left(\begin{bmatrix} Z \\ 1_v^T \end{bmatrix} \begin{bmatrix} Z \\ 1_v^T \end{bmatrix}^T \right)^{-1} \tag{26}$$

for the overdetermined problem

$$[V\ \delta] \begin{bmatrix} Z \\ 1_v^T \end{bmatrix} = X. \tag{27}$$

In Fig.2 an illustrative example is given of linear PCA analysis with bias term in the problem formulation. Shown are the score variables and reconstructed variables. The two components are reconstructed by $\tilde{x}^{(i)} = V_i z^{(i)} + \delta_i$ for $i \in \{1, 2\}$ where $z^{(i)} \in \mathbb{R}$ are one-dimensional variables and

$$[V_i\ \delta_i] = X \begin{bmatrix} Z^{(i)} \\ 1_v^T \end{bmatrix} \left(\begin{bmatrix} Z^{(i)} \\ 1_v^T \end{bmatrix} \begin{bmatrix} Z^{(i)} \\ 1_v^T \end{bmatrix}^T \right)^{-1} \tag{28}$$

with $Z^{(i)} \in \mathbb{R}^{1\times N}$ containing the scores related to the first and second largest eigenvalue, respectively.

In this cost function one usually considers the error on the given (training) data set. Of course issues of generalization are also relevant at this point. In [8] e.g. the use of cross-validation for PCA analysis has been discussed.

4 An LS-SVM approach to kernel PCA

We now follow the usual SVM methodology of mapping the data from the input space to a high dimensional feature space and applying the kernel trick.

Our objective is the following

$$\max_{w} \sum_{k=1}^{N} [0 - w^T(\varphi(x_k) - \mu_\varphi)]^2 \tag{29}$$

with notation $\mu_\varphi = (1/N)\sum_{k=1}^N \varphi(x_k)$ and $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n_h}$ the mapping to a high dimensional feature space which might be infinite dimensional (Fig.3). We take here the centering approach instead of using a bias term in the formulation. The following optimization problem is formulated now in the primal weight space

$$\begin{array}{rl} \boxed{\mathrm{P}} : \; \max\limits_{w,e} J_{\mathrm{P}}(w,e) = & \gamma \frac{1}{2} \sum\limits_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \\ \text{such that} \qquad & e_k = w^T(\varphi(x_k) - \mu_\varphi), \; k = 1, ..., N. \end{array} \tag{30}$$

This gives the Lagrangian

$$\mathcal{L}(w,e;\alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T(\varphi(x_k) - \mu_\varphi) \right) \tag{31}$$

with conditions for optimality

$$\left\{ \begin{array}{lll} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow & w = \sum\limits_{k=1}^{N} \alpha_k (\varphi(x_k) - \mu_\varphi) \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow & \alpha_k = \gamma e_k, \qquad\qquad\qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow & e_k - w^T(\varphi(x_k) - \mu_\varphi) = 0, \quad k = 1, ..., N. \end{array} \right. \tag{32}$$

By elimination of the variables e, w one obtains

$$\frac{1}{\gamma}\alpha_k - \sum_{l=1}^{N} \alpha_l (\varphi(x_l) - \mu_\varphi)^T (\varphi(x_k) - \mu_\varphi) = 0 \;, \; k = 1, ..., N. \tag{33}$$

Defining $\lambda = 1/\gamma$ one obtains the following dual problem

$$\begin{array}{c} \boxed{\mathrm{D}} : \; \text{solve in } \alpha : \\ \\ \Omega_c \alpha = \lambda \alpha \end{array} \tag{34}$$

with

$$\Omega_c = \left[\begin{array}{ccc} (\varphi(x_1) - \mu_\varphi)^T(\varphi(x_1) - \mu_\varphi) & \ldots & (\varphi(x_1) - \mu_\varphi)^T(\varphi(x_N) - \mu_\varphi) \\ \vdots & & \vdots \\ (\varphi(x_N) - \mu_\varphi)^T(\varphi(x_1) - \mu_\varphi) & \ldots & (\varphi(x_N) - \mu_\varphi)^T(\varphi(x_N) - \mu_\varphi) \end{array} \right]. \tag{35}$$

One has the following elements for the centered kernel matrix

$$\Omega_{c,kl} = (\varphi(x_k) - \mu_\varphi)^T (\varphi(x_l) - \mu_\varphi)\,, \qquad k, l = 1, ..., N. \tag{36}$$

For the centered kernel matrix one can apply the kernel trick as follows for given points x_k, x_l:

$$\begin{aligned}
&(\varphi(x_k) - \mu_\varphi)^T \ (\varphi(x_l) - \mu_\varphi) \\
&= \left(\varphi(x_k) - \frac{1}{N}\sum_{r=1}^{N} \varphi(x_r)\right)^T \left(\varphi(x_l) - \frac{1}{N}\sum_{r=1}^{N} \varphi(x_r)\right) \\
&= \varphi(x_k)^T \varphi(x_l) - \varphi(x_k)^T \frac{1}{N}\sum_{r=1}^{N} \varphi(x_r) - \varphi(x_l)^T \frac{1}{N}\sum_{r=1}^{N} \varphi(x_r) + \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} \varphi(x_r)^T \varphi(x_s) \\
&= K(x_k, x_l) - \frac{1}{N}\sum_{r=1}^{N} K(x_k, x_r) - \frac{1}{N}\sum_{r=1}^{N} K(x_l, x_r) + \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r, x_s)
\end{aligned} \tag{37}$$

with application of the kernel trick $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$ based on the Mercer Theorem. A typical choice is the RBF kernel $K(x_k, x_l) = \exp(-\|x_k - x_l\|_2^2/\sigma^2)$. This solution is equivalent with the kernel PCA solution as proposed by Schölkopf *et al.* in [13]. The centered kernel matrix can be computed as $\Omega_c = M_c \Omega M_c$ with $\Omega_{kl} = K(x_k, x_l)$. This issue of centering is also of importance in methods of principal co-ordinate analysis [8].

The optimal solution to the formulated problem is obtained by selecting the eigenvector corresponding to the largest eigenvalue. The projected variables become

$$\begin{aligned}
z(x) &= w^T \left(\varphi(x) - \mu_\varphi\right) \\
&= \sum_{l=1}^{N} \alpha_l \left(\varphi(x_l) - \mu_\varphi\right)^T \ \left(\varphi(x) - \mu_\varphi\right) \\
&= \sum_{l=1}^{N} \alpha_l \Bigg(K(x_l, x) - \tfrac{1}{N}\sum_{r=1}^{N} K(x_r, x) - \tfrac{1}{N}\sum_{r=1}^{N} K(x_r, x_l) + \\
&\qquad \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r, x_s) \Bigg) .
\end{aligned} \tag{38}$$

For the nonlinear PCA case the number of score variables n_s can be larger than the dimension of the input space n. One selects then as few score variables as possible and minimizes the reconstruction error. In this form of nonlinear PCA the mappings are nonlinear. The mapping from the score variables to the reconstructed input variables is done as

$$\tilde{x} = h(z) \tag{39}$$

such that one minimizes the reconstruction error

$$\min \sum_{k=1}^{N} \|x_k - h(z_k)\|_2^2. \tag{40}$$

This form of nonlinear PCA analysis is common in the area of neural networks [2]. A different reconstruction method has been discussed by Schölkopf *et al.* in [14] and illustrated on several examples including denoising. Furthermore, the link between kernel PCA and density estimation has been recently discussed in [4]. In Fig.4 an illustrative example is given of kernel PCA with RBF kernel applied to a noisy sine function problem. The intrinsic dimensionality of the problem is 1. Based upon the second score variable a good reconstruction with *denoising* of the given data can be made. For the nonlinear mapping $g(\cdot)$ a MLP with one hidden layer has been taken which was trained by Bayesian learning. The eigenvalues of the centered kernel matrix are shown in Fig.5.

5 Conclusion

A new formulation has been given to PCA analysis as a support vector machine. The use of a mapping to a high dimensional feature space leads to the kernel PCA version of Schölkopf *et al.* The formulation considers the problem as a one-class modelling problem with zero target value around which one maximizes the variance. A straightforward comparison can be made with the problem of Fisher discriminant analysis where the within class scatter is minimized around target values +1 and -1. Natural links exist with LS-SVM classifiers. This result also further extends (LS)-SVM methods towards unsupervised learning.

Acknowledgements

Our research is supported by grants from several funding agencies and sources: Research Council KUL: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), G.0407.02 (support vector machines), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP-TR-18: Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS; BDM and JVDW are a full professor at K.U.Leuven Belgium, JS is a professor at K.U.Leuven Belgium and a postdoctoral researcher with FWO Flanders. This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the K.U.Leuven.

References

[1] Baudat G., Anouar F., "Generalized discriminant analysis using a kernel approach," *Neural Computation,* **12**, 2385–2404, 2000.

[2] Bishop C.M., *Neural networks for pattern recognition*, Oxford University Press, 1995.

[3] Evgeniou T., Pontil M., Poggio T., "Regularization networks and support vector machines," *Advances in Computational Mathematics,* **13**(1), 1–50, 2000.

[4] Girolami M. (2002) "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, **14**(3), 669–688.

[5] Golub G.H., Van Loan C.F., *Matrix Computations*, Baltimore MD: Johns Hopkins University Press, 1989.

[6] Gower J.C., "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, **53**, 325–338, 1966.

[7] Hotelling H., "Relations between two sets of variates," *Biometrica*, **28**, 321–377, 1936.

[8] Jolliffe I.T., *Principal Component Analysis*, Springer Series in Statistics, Springer-Verlag, 1986.

[9] Mika S., Rätsch G., Weston J., Schölkopf B., Müller K.-R., "Fisher discriminant analysis with kernels," In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, 41-48. IEEE, 1999.

[10] Pearson K., "On lines and planes of closest fit to systems of points in space," *Phil. Mag. (6)*, **2**, 559–572, 1901.

[11] Poggio T., Girosi F., "Networks for approximation and learning," *Proceedings of the IEEE*, **78**(9), 1481–1497, 1990.

[12] Saunders C., Gammerman A., Vovk V., "Ridge Regression Learning Algorithm in Dual Variables," *Proc. of the 15th Int. Conf. on Machine Learning ICML-98*, Madison-Wisconsin, 515–521, 1998.

[13] Schölkopf B., Smola A., Müller K.-R., "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation,* **10**, 1299–1319, 1998.

[14] Schölkopf B., Mika S., Burges C., Knirsch P., Müller K.-R., Rätsch G., Smola A., "Input space vs. feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, **10**(5), 1000–1017, 1999.

[15] Schölkopf B., Smola A., *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[16] Suykens J.A.K., Vandewalle J., "Least squares support vector machine classifiers," *Neural Processing Letters*, **9**(3), 293–300, 1999.

[17] Suykens J.A.K., De Brabanter J., Lukas L., Vandewalle J., "Weighted least squares support vector machines : robustness and sparse approximation," *Neurocomputing*, in press.

[18] Suykens J.A.K., Vandewalle J., De Moor B., "Optimal control by least squares support vector machines," *Neural Networks*, **14**(1), 23–35, 2001.

[19] Van Gestel T., Suykens J.A.K., Lanckriet G., Lambrechts A., De Moor B., Vandewalle J., "A Bayesian Framework for Least Squares Support Vector Machine Classifiers, Gaussian Processes and Kernel Fisher Discriminant Analysis," *Neural Computation*, to appear issue May 2002.

[20] Vapnik V., *The nature of statistical learning theory*, Springer-Verlag, New-York, 1995.

[21] Wahba G., *Spline Models for Observational Data*, Series in Applied Mathematics, **59**, SIAM, Philadelphia, 1990.

[22] Williams C.K.I., Rasmussen C.E., "Gaussian processes for regression," In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo (Eds), *Advances in Neural Information Processing Systems* **8**, 514–520. MIT Press, 1996.

Captions of Figures

Figure 1: Both Fisher discriminant analysis (FDA) (supervised learning) and PCA analysis (unsupervised learning) can be derived from the viewpoint of LS-SVMs as a constrained optimization problem formulated in the primal space and solved in the dual space of Lagrange multipliers. In FDA the within class scatter is minimized around targets +1 and -1. PCA analysis can be interpreted as maximizing the variance around target 0, i.e. as a one-class target zero modelling problem.

Figure 2: Illustration of PCA analysis with bias term in the problem formulation for given data points depicted as 'o': (Top) reconstructed variables $\tilde{x}_k^{(1)}$ and $\tilde{x}_k^{(2)}$ based upon the scores $z_k^{(1)}$ and $z_k^{(2)}$ depicted as '+'; (Bottom) the score variables $z_k^{(1)}$ and $z_k^{(2)}$ which are decorrelated.

Figure 3: LS-SVM approach to kernel Fisher discriminant analysis: the input data are mapped to a high dimensional feature space and next to the score variables. The score variables are interpreted as error variables in a one-class modelling problem with target zero for which one aims at having maximal variance.

Figure 4: Illustration of kernel PCA to noisy sine function data depicted by 'o' in a two-dimensional input space. The reconstructed variables $\tilde{x}_k$ are shown as '+' and are reconstructed based upon one single score variable. This shows that the method is capable of discovering the intrinsic dimensionality equal to one of the noisy sine function line in the input space and denoise the noisy sine function.

Figure 5: Eigenvalues of the centered kernel matrix (scree graph) related to the previous Figure of the noisy sine function.

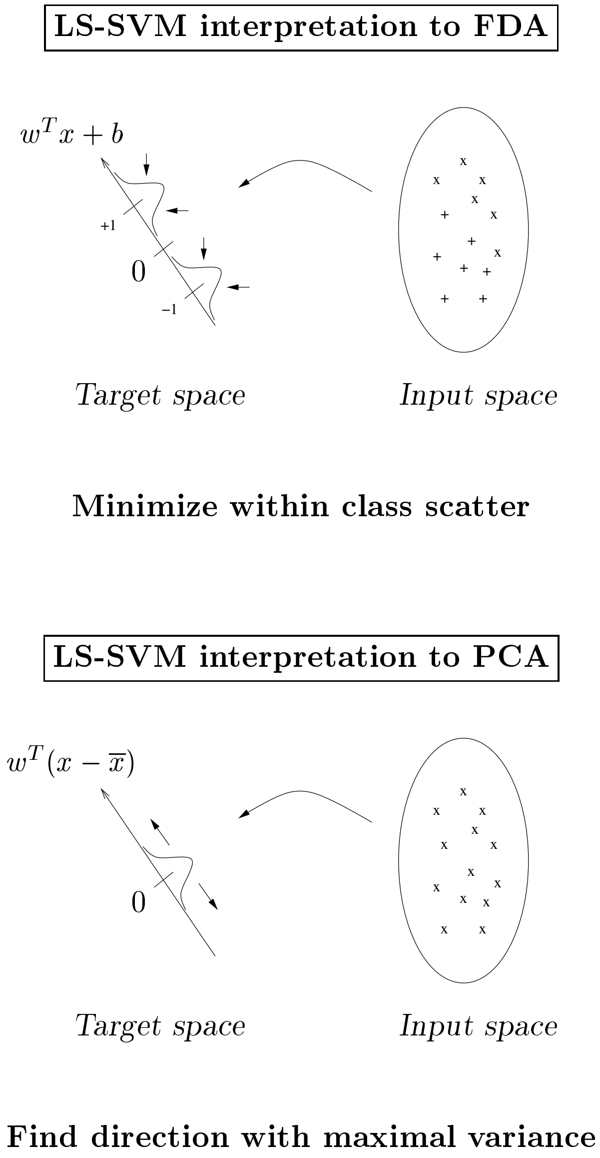

LS-SVM interpretation to FDA
$w^T x + b$
+1
0
−1
Target space
Input space
Minimize within class scatter
LS-SVM interpretation to PCA
$w^T(x - \overline{x})$
0
Target space
Input space
Find direction with maximal variance

Figure 1:

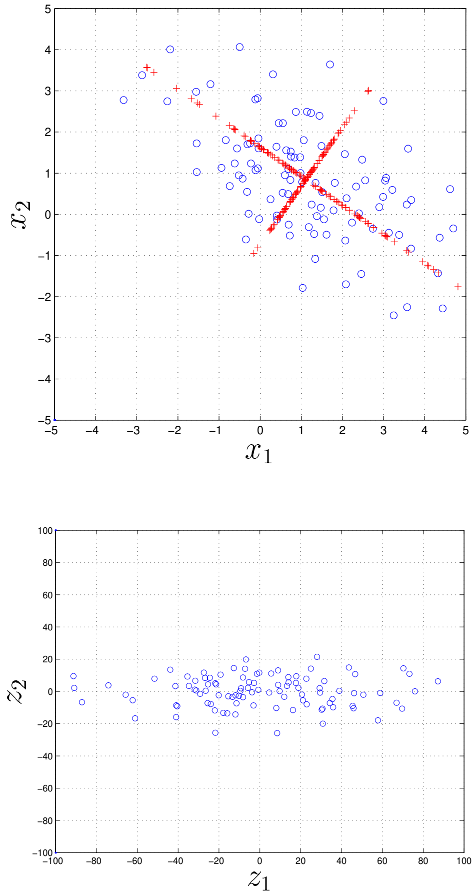
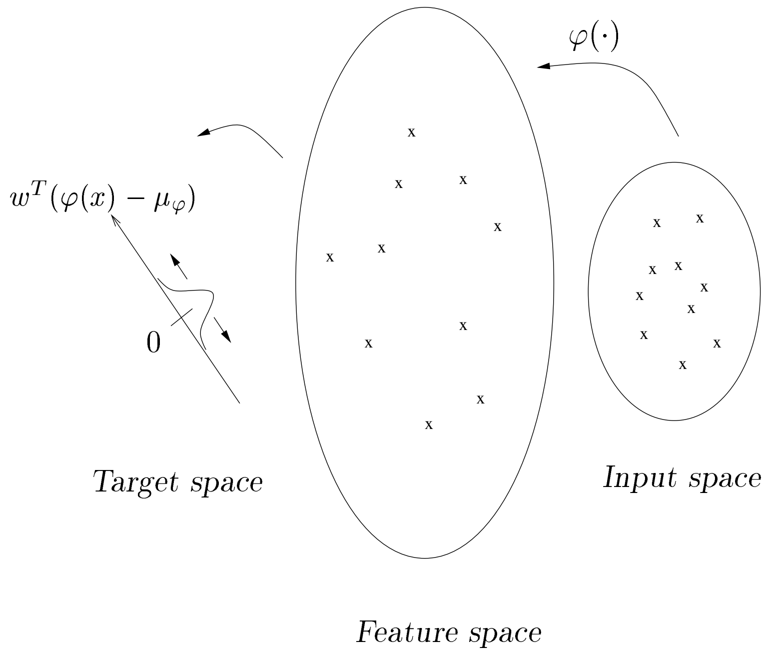


Figure 2:

$\boxed{\textbf{LS-SVM interpretation to Kernel PCA}}$



Find direction with maximal variance

Figure 3:

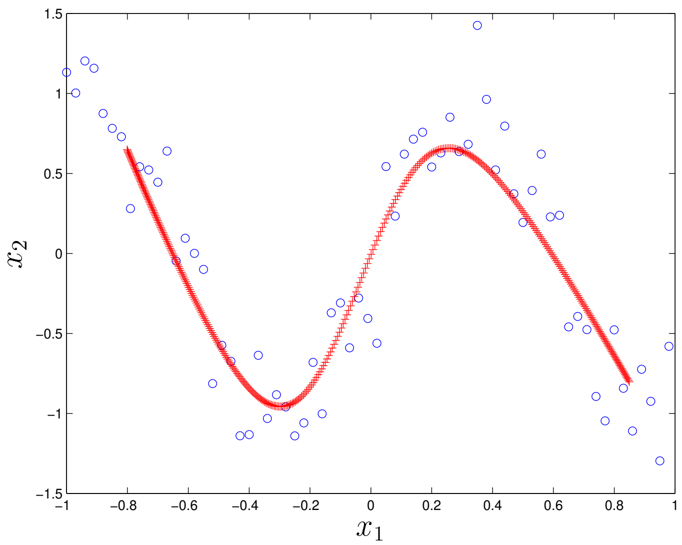


Figure 4:

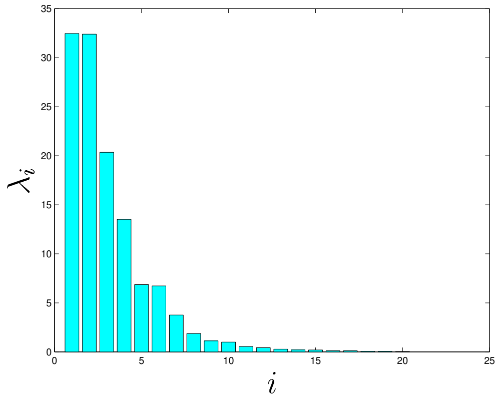


Figure 5: