Chapter 7

LS-SVM for Unsupervised Learning

In the previous Chapters we discussed the use of LS-SVMs for problems of supervised learning, including classification and function estimation. In this Chapter we show how support vector machine alike formulations can also be given to the well-known method of principal component analysis (PCA), which is a frequently used technique in unsupervised learning. The new formulation can be extended in a straightforward way to the nonlinear case by applying the kernel trick, and leads to what is presently called kernel PCA. In a similar fashion, new formulations and kernel versions to canonical correlation analysis (CCA) are obtained.

7.1 Support Vector Machines and linear PCA analysis

In the first Chapter we already explained the basic idea of linear and nonlinear PCA methods. The scope now is to first formulate linear PCA analysis within the LS-SVM classifier context and, second, to extend this formulation to a high dimensional feature space with application of the kernel trick. The kernel PCA method was originally introduced by Schölkopf *et al.* in [203]. A difference is that we now formulate an optimization problem with primal-dual interpretations where the dual problem can be related to kernel PCA. This derivation is in the style of LS-SVM primal-dual formulations and interpretations.

7.1.1 *Classical principal component analysis formulation*

In general, there exist different formulations in order to characterize PCA problems [128]. One formulation is to consider a given set of *zero mean* data $\{x_k\}_{k=1}^N$ (only input space data) as a cloud of points for which one tries to find projected variables $w^T x$ with maximal variance. This means

$$\begin{aligned} \max_w \mathrm{Var}(w^T x) &= \mathrm{Cov}(w^T x, w^T x) \simeq \frac{1}{N} \sum_{k=1}^{N} (w^T x_k)^2 \\ &= w^T C w \end{aligned} \tag{7.1}$$

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

$$\mathcal{L}(w; \lambda) = \frac{1}{2} w^T C w - \lambda (w^T w - 1) \tag{7.2}$$

with Lagrange multiplier λ. The solution follows from $\partial \mathcal{L}/\partial w = 0$, $\partial \mathcal{L}/\partial \lambda = 0$ and gives the eigenvalue problem

$$C w = \lambda w. \tag{7.3}$$

The matrix C is symmetric and positive semidefinite. The eigenvector w corresponding to the largest eigenvalue determines the projected variable having maximal variance. Efficient and reliable numerical methods are discussed e.g. in [98].

7.1.2 *Support vector machine formulation to linear PCA*

In order to establish now the link with LS-SVM methods, let us reformulate the problem as follows

$$\max_w \sum_{k=1}^{N} [0 - w^T x_k]^2 \tag{7.4}$$

where 0 is considered as a single target value. The interpretation can be made in a similar fashion as for the links between LS-SVM classifiers and Fisher discriminant analysis. The projected variable to a target space here is $z = w^T x$. 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.

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

$$\left[\begin{array}{lll} \boxed{\mathrm{P}}: & \max\limits_{w,e} J_{\mathrm{P}}(w,e) = & \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \\ & \text{such that} & e_k = w^T x_k, \; k = 1, ..., N. \end{array} \right] \tag{7.5}$$

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

$$\mathcal{L}(w,e;\alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T x_k \right) \tag{7.6}$$

with conditions for optimality given by

$$\left\{ \begin{array}{lll} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow & w = \sum_{k=1}^{N} \alpha_k x_k \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow & \alpha_k = \gamma e_k, \qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow & e_k - w^T x_k = 0, \quad k = 1, ..., N. \end{array} \right. \tag{7.7}$$

By elimination of the variables e, w one obtains

$$\frac{1}{\gamma} \alpha_k - \sum_{l=1}^{N} \alpha_l x_l^T x_k = 0 \;, \quad k = 1, ..., N. \tag{7.8}$$

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

$$\left[\begin{array}{l} \boxed{\text{D}} : \text{ solve in } \alpha : \\ \\ \begin{bmatrix} x_1^T x_1 & \dots & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & \dots & x_N^T x_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} \end{array}\right] \tag{7.9}$$

which is the dual interpretation of (7.3). When one writes this as

$$\Omega\alpha = \lambda\alpha \tag{7.10}$$

where

$$\Omega_{kl} = x_k^T x_l \,, \quad k, l = 1, ..., N. \tag{7.11}$$

One easily sees that this matrix is the Gram matrix for a linear kernel $K(x_k, x_l) = x_k^T x_l$. The vector of dual variables $\alpha = [\alpha_1; ...; \alpha_N]$ is an eigenvector of the problem 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

$$z(x) = w^T x = \sum_{l=1}^{N} \alpha_l x_l^T x \tag{7.12}$$

where α is the eigenvector corresponding to the largest eigenvalue. Note that all eigenvalues are positive and real because the matrix is symmetric and positive semidefinite. 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

$$\sum_{k=1}^{N} (w^T x_k)^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{1}{\gamma^2} \alpha_k^2 = \lambda_{max}^2, \tag{7.13}$$

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 [128], 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

$$\max_{w} \sum_{k=1}^{N} [w^T(x_k - \hat{\mu}_x)]^2 \tag{7.14}$$

where $\hat{\mu}_x = (1/N) \sum_{k=1}^{N} x_k$. The same derivations can be made and one finally obtains a centered Gram matrix as a result in the eigenvalue problem. 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. This point was also addressed in the context of (LS)-SVM classifiers. The approach of taking the eigenvalue decomposition of the centered Gram matrix is also done in *principal coordinate analysis* [99; 128].

7.1.3 *Including a bias term*

While in PCA analysis one usually centers the data, the LS-SVM interpretation to PCA analysis also offers the opportunity to analyse the use of a bias term in a straightforward way. The score variables are

$$z(x) = w^T x + b \tag{7.15}$$

and one aims at optimizing the following objective

$$\max_{w,b} \sum_{k=1}^{N} [0 - (w^T x_k + b)]^2. \tag{7.16}$$

Therefore, one formulates the primal optimization problem

$$\left[\begin{array}{lll} \boxed{\mathrm{P}}: & \max_{w,b,e} J_{\mathrm{P}}(w,e) = & \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \\ & \text{such that} & e_k = w^T x_k + b, \quad k = 1, ..., N \end{array} \right] \tag{7.17}$$

with Lagrangian

$$\mathcal{L}(w,b,e;\alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T x_k - b\right) \tag{7.18}$$

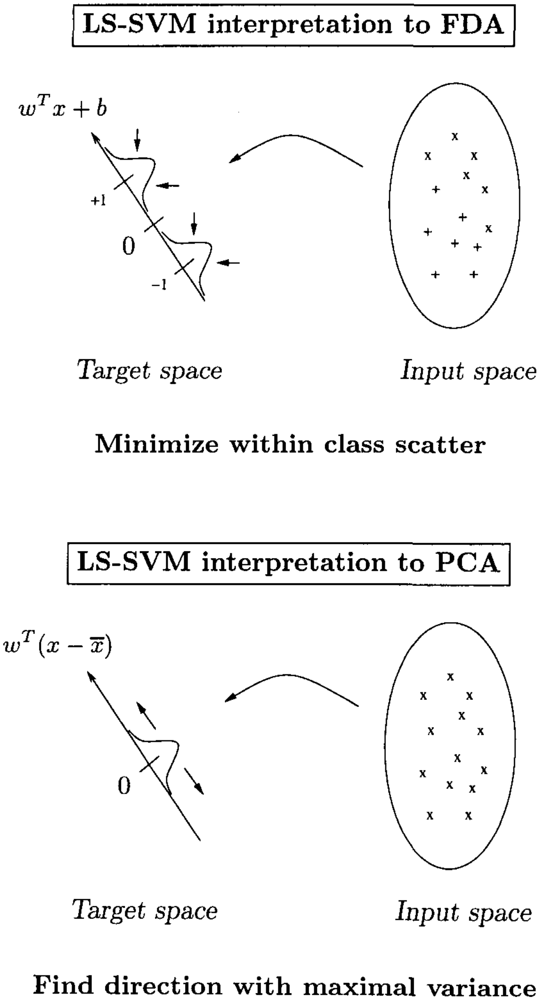
LS-SVM interpretation to FDA
$w^T x + b$
+1
0
-1
Target space
Input space
Minimize within class scatter
LS-SVM interpretation to PCA
$w^T(x - \overline{x})$
0
Target space
Input space
Find direction with maximal variance

Fig. 7.1 *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.*

giving as conditions for optimality

$$\left\{\begin{array}{llll} \frac{\partial \mathcal{L}}{\partial w}=0 & \rightarrow & w=\sum_{k=1}^{N}\alpha_k x_k & \\ \frac{\partial \mathcal{L}}{\partial e_k}=0 & \rightarrow & \alpha_k=\gamma e_k, & k=1,...,N \\ \frac{\partial \mathcal{L}}{\partial b}=0 & \rightarrow & \sum_{k=1}^{N}\alpha_k=0 & \\ \frac{\partial \mathcal{L}}{\partial \alpha_k}=0 & \rightarrow & e_k-w^T x_k-b=0, & k=1,...,N. \end{array}\right. \tag{7.19}$$

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

$$b=-\frac{1}{N}\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_l x_l^T x_k. \tag{7.20}$$

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

$\boxed{\text{D}}$: solve in α :

$$\left[\begin{array}{ccc} (x_1-\hat{\mu}_x)^T(x_1-\hat{\mu}_x) & \dots & (x_1-\hat{\mu}_x)^T(x_N-\hat{\mu}_x) \\ \vdots & & \vdots \\ (x_N-\hat{\mu}_x)^T(x_1-\hat{\mu}_x) & \dots & (x_N-\hat{\mu}_x)^T(x_N-\hat{\mu}_x) \end{array}\right]\left[\begin{array}{c}\alpha_1\\ \vdots \\ \alpha_N\end{array}\right]=\lambda\left[\begin{array}{c}\alpha_1\\ \vdots \\ \alpha_N\end{array}\right] \tag{7.21}$$

which is an eigenvalue decomposition of the centered Gram matrix

$$\Omega_c\alpha=\lambda\alpha \tag{7.22}$$

with $\Omega_c=M_c\Omega M_c$ where $M_c=I-1_v 1_v^T/N$ 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

$$z(x) = w^T x + b = \sum_{l=1}^{N} \alpha_l x_l^T x + b \tag{7.23}$$

where α is the eigenvector corresponding to the largest eigenvalue and

$$\sum_{k=1}^{N} (w^T x_k + b)^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} \frac{1}{\gamma^2} \alpha_k^2 = \lambda_{max}^2. \tag{7.24}$$

7.1.4 *The reconstruction problem*

It was already mentioned in the first Chapter that PCA analysis can be related to a reconstruction problem with information bottleneck. One aims at minimizing the reconstruction error

$$\min \sum_{k=1}^{N} \|x_k - \tilde{x}_k\|_2^2 \tag{7.25}$$

where $\tilde{x}_k$ are variables reconstructed from the score variables (Fig. 7.2). Let us denote 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 case one can take

$$\tilde{x} = Vz + \delta \tag{7.26}$$

and minimize

$$\min_{V,\delta} \sum_{k=1}^{N} \|x_k - (Vz_k + \delta)\|_2^2. \tag{7.27}$$

In matrix form this leads to a least squares solution

$$[V \ \delta] = X \begin{bmatrix} Z \\ 1_v^T \end{bmatrix}^T \left(\begin{bmatrix} Z \\ 1_v^T \end{bmatrix} \begin{bmatrix} Z \\ 1_v^T \end{bmatrix}^T \right)^{-1} \tag{7.28}$$

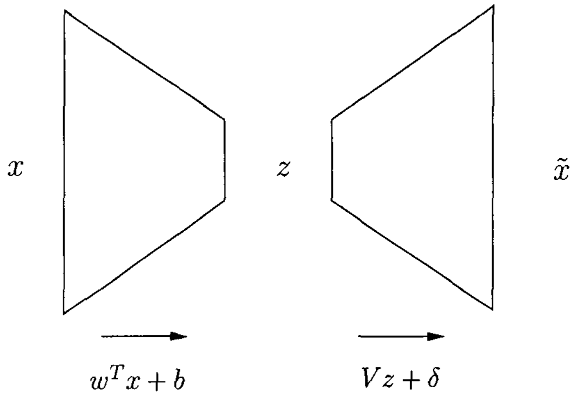

Fig. 7.2 *Reconstruction problem interpretation of linear PCA analysis related to the LS-SVM formulation with bias term. The inputs x are mapped to the score variables z to a lower dimension (dimensionality reduction step) and reconstructed into $\tilde{x}$.*

for the overdetermined problem

$$[V \ \delta] \begin{bmatrix} Z \\ 1_v^T \end{bmatrix} = X. \tag{7.29}$$

In Fig. 7.3 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

$$[V_i \ \delta_i] = X \begin{bmatrix} Z^{(i)} \\ 1_v^T \end{bmatrix}^T \left(\begin{bmatrix} Z^{(i)} \\ 1_v^T \end{bmatrix} \begin{bmatrix} Z^{(i)} \\ 1_v^T \end{bmatrix}^T \right)^{-1} \tag{7.30}$$

with $Z^{(i)} \in \mathbb{R}^{1 \times N}$ containing the scores related to the first and second largest eigenvalues, 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 [128] the use of cross-validation for PCA analysis has been discussed.

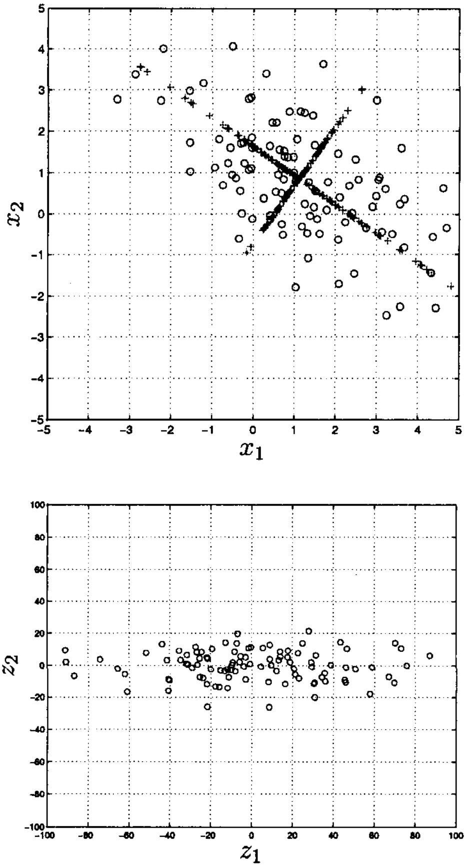

Fig. 7.3 *Illustration of PCA analysis with bias term in the problem formulation for given data points depicted as 'o': (Top) reconstructed variables $\tilde{x}_k^{(1)}$ and $\tilde{x}_k^{(2)}$ based upon the scores $z_k^{(1)}$ and $z_k^{(2)}$ depicted as '+'; (Bottom) the score variables $z_k^{(1)}$ and $z_k^{(2)}$ which are decorrelated.*

7.2 An LS-SVM approach to kernel PCA

Let us now extend the LS-SVM interpretation of linear PCA analysis to a high dimensional feature space and apply the kernel trick, as illustrated in Fig. 7.4.

Our objective is the following

$$\max_{w} \sum_{k=1}^{N} [0 - w^T(\varphi(x_k) - \hat{\mu}_{\varphi})]^2 \tag{7.31}$$

with notation $\hat{\mu}_{\varphi} = (1/N)\sum_{k=1}^{N} \varphi(x_k)$ and $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ the mapping to a high dimensional feature space which might be infinite dimensional. 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

$$\left[\begin{array}{lll} \boxed{\mathrm{P}} : & \max\limits_{w,e} J_{\mathrm{P}}(w,e) = & \gamma \frac{1}{2} \sum\limits_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \\ & \text{such that} & e_k = w^T(\varphi(x_k) - \hat{\mu}_{\varphi}), \ k = 1, ..., N. \end{array} \right] \tag{7.32}$$

This gives the Lagrangian

$$\mathcal{L}(w,e;\alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T(\varphi(x_k) - \hat{\mu}_{\varphi}) \right) \tag{7.33}$$

with conditions for optimality

$$\left\{ \begin{array}{llll} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow & w = \sum\limits_{k=1}^{N} \alpha_k (\varphi(x_k) - \hat{\mu}_{\varphi}) & \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow & \alpha_k = \gamma e_k, & k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow & e_k - w^T(\varphi(x_k) - \hat{\mu}_{\varphi}) = 0, & k = 1, ..., N. \end{array} \right. \tag{7.34}$$

By elimination of the variables e, w one obtains

$$\frac{1}{\gamma} \alpha_k - \sum_{l=1}^{N} \alpha_l (\varphi(x_l) - \hat{\mu}_{\varphi})^T (\varphi(x_k) - \hat{\mu}_{\varphi}) = 0 \ , \ k = 1, ..., N. \tag{7.35}$$

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

$$\left[\begin{array}{l} \boxed{\mathrm{D}} : \ \text{solve in } \alpha : \\ \qquad\qquad \Omega_c \alpha = \lambda \alpha \end{array} \right] \tag{7.36}$$

with

$$\Omega_c = \begin{bmatrix} (\varphi(x_1) - \hat{\mu}_\varphi)^T(\varphi(x_1) - \hat{\mu}_\varphi) & \dots & (\varphi(x_1) - \hat{\mu}_\varphi)^T(\varphi(x_N) - \hat{\mu}_\varphi) \\ \vdots & & \vdots \\ (\varphi(x_N) - \hat{\mu}_\varphi)^T(\varphi(x_1) - \hat{\mu}_\varphi) & \dots & (\varphi(x_N) - \hat{\mu}_\varphi)^T(\varphi(x_N) - \hat{\mu}_\varphi) \end{bmatrix}. \tag{7.37}$$

One has the following elements for the centered kernel matrix

$$\Omega_{c,kl} = (\varphi(x_k) - \hat{\mu}_\varphi)^T(\varphi(x_l) - \hat{\mu}_\varphi)\,, \quad k,l = 1,...,N. \tag{7.38}$$

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

$$\begin{aligned} &(\varphi(x_k) - \hat{\mu}_\varphi)^T\ (\varphi(x_l) - \hat{\mu}_\varphi) \\ &= \left(\varphi(x_k) - \frac{1}{N}\sum_{r=1}^N \varphi(x_r)\right)^T \left(\varphi(x_l) - \frac{1}{N}\sum_{r=1}^N \varphi(x_r)\right) \\ &= \varphi(x_k)^T\varphi(x_l) - \varphi(x_k)^T\frac{1}{N}\sum_{r=1}^N \varphi(x_r) - \varphi(x_l)^T\frac{1}{N}\sum_{r=1}^N \varphi(x_r) + \\ &\quad \frac{1}{N^2}\sum_{r=1}^N\sum_{s=1}^N \varphi(x_r)^T\varphi(x_s) \\ &= K(x_k,x_l) - \frac{1}{N}\sum_{r=1}^N K(x_k,x_r) - \frac{1}{N}\sum_{r=1}^N K(x_l,x_r) + \frac{1}{N^2}\sum_{r=1}^N\sum_{s=1}^N K(x_r,x_s). \end{aligned} \tag{7.39}$$

This solution is equivalent with kernel PCA as proposed by Schölkopf *et al.* in [203]. The centered kernel matrix can be computed as $\Omega_c = M_c \Omega M_c$ with $\Omega_{kl} = K(x_k, x_l)$ with M_c the centering matrix. This issue of centering is also of importance in methods of principal co-ordinate analysis [128].

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

$$\begin{aligned} z(x) &= w^T \left(\varphi(x) - \hat{\mu}_\varphi\right) \\ &= \sum_{l=1}^{N} \alpha_l \left(\varphi(x_l) - \hat{\mu}_\varphi\right)^T \left(\varphi(x) - \hat{\mu}_\varphi\right) \\ &= \sum_{l=1}^{N} \alpha_l \left(K(x_l, x) - \tfrac{1}{N}\sum_{r=1}^{N} K(x_r, x) - \tfrac{1}{N}\sum_{r=1}^{N} K(x_r, x_l) + \right. \\ & \quad \left. \frac{1}{N^2} \sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r, x_s) \right) . \end{aligned} \tag{7.40}$$

One may choose here any positive definite kernel satisfying the Mercer condition, with the RBF kernel as a typical choice.

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 minimize the reconstruction error (Fig. 7.5). In this form of nonlinear PCA the mappings are nonlinear. The mapping from the score variables to the reconstructed input variables is done as

$$\tilde{x} = h(z) \tag{7.41}$$

such that one minimizes the reconstruction error

$$\min \sum_{k=1}^{N} \|x_k - h(z_k)\|_2^2 . \tag{7.42}$$

This form of nonlinear PCA analysis is common in the area of neural networks [23]. A different reconstruction method has been discussed by Schölkopf *et al.* in [204]. In Fig. 7.6 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 $h(\cdot)$ an MLP with one hidden layer has been taken which was trained by Bayesian learning. The eigenvalues of the kernel matrix are shown in Fig. 7.7.

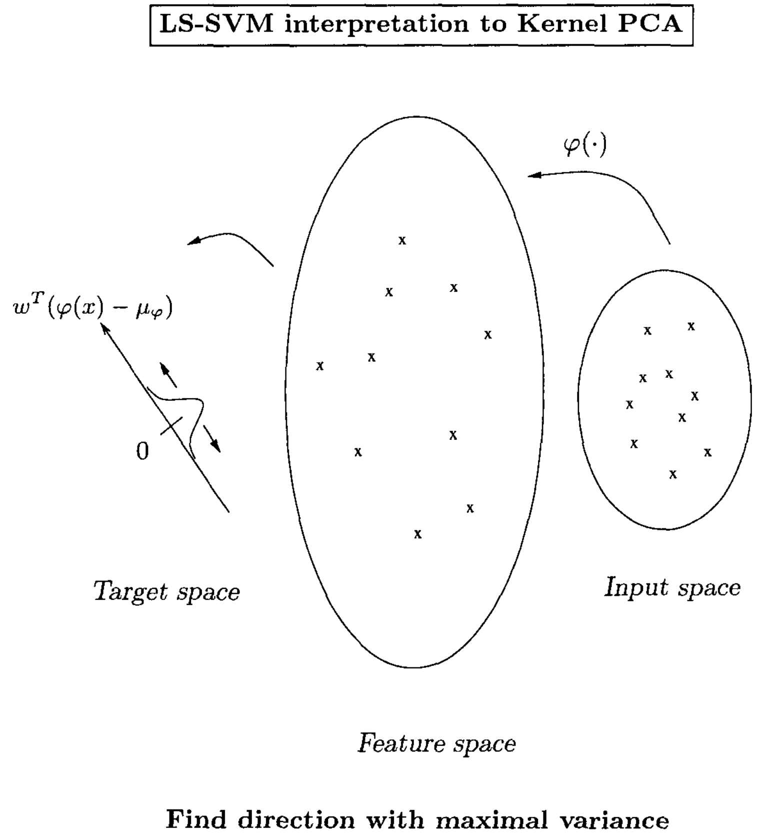
LS-SVM interpretation to Kernel PCA
$\varphi(\cdot)$
$w^T(\varphi(x) - \mu_\varphi)$
0
Target space
Input space
Feature space
Find direction with maximal variance

Fig. 7.4 *LS-SVM approach to kernel principal component 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.*

7.3 Links with density estimation

A link between kernel PCA and orthogonal series density estimation has been established by Girolami [94]. A probability density function that is square integrable can be represented by a convergent orthogonal series

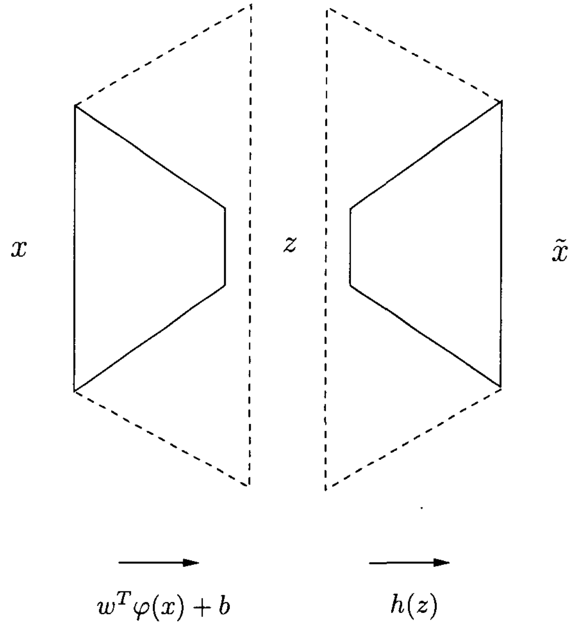

Fig. 7.5 *Reconstruction problem interpretation of kernel PCA analysis related to the LS-SVM formulation. The inputs x are mapped to the score variables z. This number can be larger than the dimension of the input space, but one selects as few as possible and minimizes the reconstruction error. In this form of nonlinear PCA the mappings are nonlinear.*

expansion such that

$$p(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x) \tag{7.43}$$

with $x \in \mathbb{R}^n$ and $\{\varphi_i(x)\}_{i=1}^{\infty}$ an orthonormal set of functions [120]. For an orthonormal series expansion in a Hilbert space with density function $p(x)$ the expansion coefficients equal [136]

$$c_i = \int \varphi_i(x) p(x) dx. \tag{7.44}$$

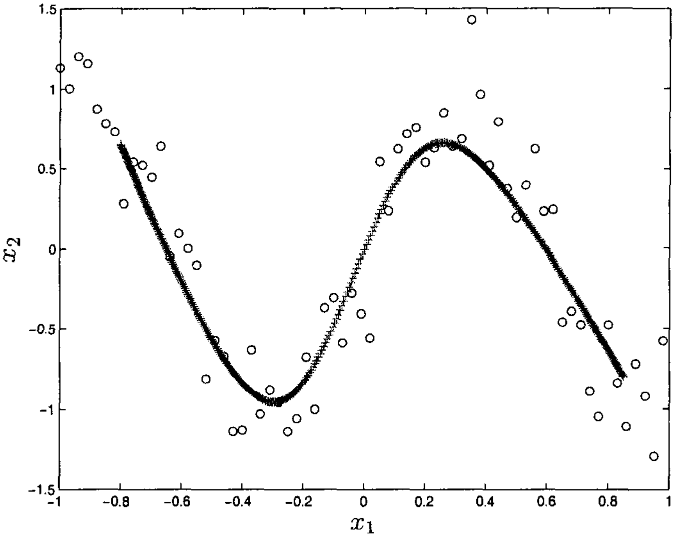

Fig. 7.6 *Illustration of kernel PCA to noisy sine function data depicted by 'o' in a two-dimensional input space. The reconstructed variables $\tilde{x}_k$ are shown as '+' and are reconstructed based upon one single score variable. This shows that the method is capable of discovering the intrinsic dimensionality equal to one of the noisy sine function line in the input space and denoise the noisy sine function.*

Towards illustrating the link with kernel PCA it is needed to consider the truncated series

$$\hat{p}_M(x) = \sum_{i=1}^{M} \hat{c}_i \varphi_i(x) \tag{7.45}$$

consisting of M terms with estimated coefficients

$$\hat{c}_i = \frac{1}{N} \sum_{k=1}^{N} \varphi_i(x_k) \tag{7.46}$$

which gives

$$\hat{p}_M(x) = \frac{1}{N} \sum_{i=1}^{M} \sum_{k=1}^{N} \varphi_i(x_k) \varphi_i(x). \tag{7.47}$$

Although it can be shown that this method is asymptotically unbiased it may produce negative point values which is an important drawback of the

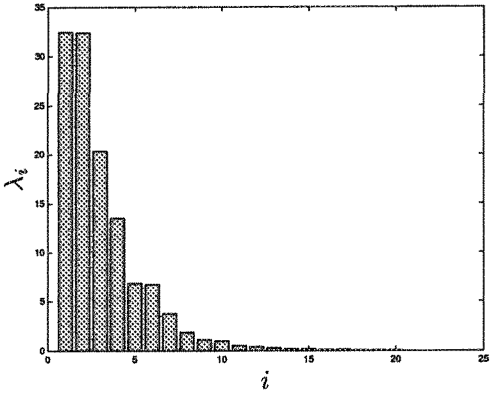

Fig. 7.7 *Eigenvalues of the centered kernel matrix (scree graph) related to the previous Figure of the noisy sine function.*

method.

One can then take the scores resulting from kernel PCA as basis function for a density estimator. Therefore, one considers the eigenvalue decomposition of the centered kernel matrix

$$\Omega_c U = U\tilde{\Lambda} \tag{7.48}$$

where $\tilde{\Lambda} = \text{diag}([\tilde{\lambda}_1; ...; \tilde{\lambda}_N])$ contains the eigenvalues and $U = [u_1 ... u_N] \in \mathbb{R}^{N \times N}$ the corresponding eigenvectors. This can be used in order to estimate the eigenfunctions $\phi_i(x)$ and eigenvalues λ_i for the integral equation (Karhunen-Loeve expansion)

$$\int K(x, x')\phi_i(x)p(x)dx = \lambda_i \phi_i(x') \tag{7.49}$$

with the following estimates

$$\begin{aligned} \hat{\lambda}_i &= \tfrac{1}{N}\tilde{\lambda}_i \\ \hat{\phi}_i(x_k) &= \sqrt{N} u_{ki} \\ \hat{\phi}_i(x') &= \tfrac{\sqrt{N}}{\tilde{\lambda}_i}\sum_{k=1}^{N} u_{ki} K(x_k, x') \end{aligned} \tag{7.50}$$

where u_{ki} denotes the ki-th entry of the matrix U. This approach is a form of Nyström method [63; 294]. Using the eigenvectors as finite sample estimates of the corresponding eigenfunctions, the truncated estimate of the probability density function at point x' is given by

$$\begin{aligned} \hat{p}_M(x') &= \frac{1}{N} 1_v^T \sum_{i=1}^{M} \sqrt{\tilde{\lambda}_i} u_i \sum_{k=1}^{N} \frac{1}{\sqrt{\tilde{\lambda}_i}} u_{ki} K(x_k, x') \\ &= \frac{1}{N} 1_v^T U_M U_M^T \theta(x') \end{aligned} \tag{7.51}$$

where $\theta(x') = [K(x', x_1); K(x', x_2); ...; K(x', x_N)]$, $1_v = [1; 1; ...; 1]$ and $U_M \in \mathbb{R}^{N \times M}$ is the matrix with eigenvectors of Ω_c consisting of the eigenvectors corresponding to the M largest eigenvalues. For the case of $M = N$ this reduces to the well-known Parzen window density estimator $p(x') = \frac{1}{N} 1_v^T \theta(x')$. Note that in order to obtain a density the integral of the function over the entire space should be equal to one. It is assumed here that normalizations are done for the kernel K to achieve this.

Concerning the number of terms to be taken in the expansion many criteria have been investigated in literature [67]. In [94] the following has been investigated with cutoff value for determination of the value of M:

$$(\frac{1}{N} 1_v^T u_i)^2 > \frac{2N}{1+N}. \tag{7.52}$$

An estimate for the overall integrated square truncation error $\sum_{i=M+1}^{\infty} c_i^2$ is given by

$$c_i^2 \simeq \tilde{\lambda}_i (\frac{1}{N} 1_v^T u_i)^2. \tag{7.53}$$

This can also be related to the quadratic Renyi entropy

$$H_R = -\log \int p(x)^2 dx \tag{7.54}$$

which forms a measure of distribution compactness [87] and has recently been used within the context of information theoretic learning [187]. In [94] Girolami shows that

$$\int \hat{p}(x)^2 dx = \sum_{i=1}^{N} \tilde{\lambda}_i (\frac{1}{N} 1_v^T u_i)^2. \tag{7.55}$$

Large contributions to the entropy come from components that have small values of $\tilde{\lambda}_i (\frac{1}{N} 1_v^T u_i)^2$ and are related to elements with little or no structure, caused by observation noise or diffuse regions in the data. Large values of $\tilde{\lambda}_i(\frac{1}{N}1_v^T u_i)^2$ on the other hand indicate regions of high density or compactness. A more general result according to [94] is

$$\int \hat{p}(x)^2 dx = \frac{1}{N^2} 1_v^T \Omega 1_v \tag{7.56}$$

in the sense that this result is not restricted to RBF kernels.

7.4 Kernel CCA

7.4.1 *Classical canonical correlation analysis formulation*

The problem of canonical correlation analysis (CCA) is related to the PCA analysis problem [128]. In CCA analysis (originally studied by Hotelling in 1936 [116]) one is interested in finding maximal correlation between projected variables $z_x = w^T x$ and $z_y = v^T y$ where $x \in \mathbb{R}^{n_x}, y \in \mathbb{R}^{n_y}$ denote given random vectors with zero mean. Linear CCA analysis has been applied in subspace algorithms for system identification [275], with links to system theory, information theory and signal processing [61].

The objective function is to maximize the correlation coefficient

$$\begin{aligned} \max_{w,v} \rho &= \frac{\mathcal{E}[z_x z_y]}{\sqrt{\mathcal{E}[z_x z_x]}\sqrt{\mathcal{E}[z_y z_y]}} \\ &= \frac{w^T C_{\mathrm{xy}} v}{\sqrt{w^T C_{\mathrm{xx}} w}\sqrt{v^T C_{\mathrm{yy}} v}} \end{aligned} \tag{7.57}$$

with $C_{\mathrm{xx}} = \mathcal{E}[xx^T]$, $C_{\mathrm{yy}} = \mathcal{E}[yy^T]$, $C_{\mathrm{xy}} = \mathcal{E}[xy^T]$. This is usually formulated as the constrained optimization problem

$$\begin{array}{ll} \displaystyle\max_{w,v} & w^T C_{\mathrm{xy}} v \\ \text{such that} & w^T C_{\mathrm{xx}} w = 1 \\ & v^T C_{\mathrm{yy}} v = 1 \end{array} \tag{7.58}$$

which leads to a generalized eigenvalue problem. The solution follows from the Lagrangian

$$\mathcal{L}(w, v; \eta, \nu) = w^T C_{\mathrm{xy}} v - \eta \frac{1}{2}(w^T C_{\mathrm{xx}} w - 1) - \nu \frac{1}{2}(v^T C_{\mathrm{yy}} v - 1) \tag{7.59}$$

with Lagrange multipliers η, ν, which gives

$$\left\{ \begin{array}{lcl} C_{\mathrm{xy}} v & = & \eta\, C_{\mathrm{xx}} w \\ C_{\mathrm{yx}} w & = & \nu\, C_{\mathrm{yy}} v. \end{array} \right. \tag{7.60}$$

7.4.2 *Support vector machine formulation to linear CCA*

In a similar fashion as for PCA analysis one can develop a support vector machine type formulation. This is done by considering the primal problem

$$\left[\begin{array}{lll} \boxed{\mathrm{P}}: & \max\limits_{w,v,e,r} \; \gamma \sum\limits_{k=1}^{N} e_k r_k - \nu_1 \frac{1}{2} \sum\limits_{k=1}^{N} e_k^2 - \nu_2 \frac{1}{2} \sum\limits_{k=1}^{N} r_k^2 - \frac{1}{2} w^T w - \frac{1}{2} v^T v & \\ & \text{such that } e_k = w^T x_k, & k = 1, ..., N \\ & \qquad\qquad r_k = v^T y_k & k = 1, ..., N \end{array} \right] \tag{7.61}$$

with Lagrangian

$$\begin{aligned} \mathcal{L}(w, v, e, r; \alpha, \beta) = \gamma \sum_{k=1}^{N} e_k r_k - \nu_1 \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \nu_2 \frac{1}{2} \sum_{k=1}^{N} r_k^2 - \frac{1}{2} w^T w - \\ \frac{1}{2} v^T v - \sum_{k=1}^{N} \alpha_k [e_k - w^T x_k] - \sum_{k=1}^{N} \beta_k [r_k - v^T y_k] \end{aligned} \tag{7.62}$$

where α_k, β_k are Lagrange multipliers. The objective function in the primal problem does not have the same expression as the correlation coefficient but takes the contribution of the numerator with a plus sign and the contributions of the denominator with a minus sign. However, the cost function can be considered as a generalization of the objective $\min_{w,v} \sum_k \|w^T x_k - v^T y_k\|_2^2$ which is known in the area of CCA analysis [96]. Additional unknown error variables e, r are taken which are equal to the score variables z_x, z_y. By considering the expressions of the score variables as constraints one is able to find a meaningful dual problem for which one is able to create a kernel version.

The conditions for optimality are

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k x_k \\ \frac{\partial \mathcal{L}}{\partial v} = 0 & \rightarrow \quad v = \sum_{k=1}^{N} \beta_k y_k \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 & \rightarrow \quad \gamma v^T y_k = \nu_1 w^T x_k + \alpha_k, \qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial r_k} = 0 & \rightarrow \quad \gamma w^T x_k = \nu_2 v^T y_k + \beta_k, \qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow \quad e_k = w^T x_k, \qquad k = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \beta_k} = 0 & \rightarrow \quad r_k = v^T y_k, \qquad k = 1, ..., N \end{cases} \tag{7.63}$$

which results in the following dual problem after defining $\lambda = 1/\gamma$

$\boxed{\text{D}}$: solve in α, β :

$$\left[\begin{array}{ccc|ccc} & & & y_1^T y_1 & \dots & y_1^T y_N \\ & 0 & & \vdots & & \vdots \\ & & & y_N^T y_1 & \dots & y_N^T y_N \\ \hline x_1^T x_1 & \dots & x_1^T x_N & & & \\ \vdots & & \vdots & & 0 & \\ x_N^T x_1 & \dots & x_N^T x_N & & & \end{array}\right] \left[\begin{array}{c} \alpha_1 \\ \vdots \\ \alpha_N \\ \hline \beta_1 \\ \vdots \\ \beta_N \end{array}\right]$$

$$= \lambda \left[\begin{array}{ccc|ccc} \nu_1 x_1^T x_1 + 1 & \dots & \nu_1 x_1^T x_N & & & \\ \vdots & & \vdots & & 0 & \\ \nu_1 x_N^T x_1 & \dots & \nu_1 x_N^T x_N + 1 & & & \\ \hline & & & \nu_2 y_1^T y_1 + 1 & \dots & \nu_2 y_1^T y_N \\ & 0 & & \vdots & & \vdots \\ & & & \nu_2 y_N^T y_1 & \dots & \nu_2 y_N^T y_N + 1 \end{array}\right] \left[\begin{array}{c} \alpha_1 \\ \vdots \\ \alpha_N \\ \hline \beta_1 \\ \vdots \\ \beta_N \end{array}\right] . \tag{7.64}$$

This is a generalized eigenvalue problem to be solved with eigenvectors $[\alpha; \beta]$ and corresponding eigenvalues λ. One can then select the value of λ and the corresponding eigenvectors such that the correlation coefficient is maximized

$$\max \rho(\lambda). \tag{7.65}$$

The resulting score variables are

$$\begin{aligned} z_{x_k} &= e_k = \sum_{l=1}^{N} \alpha_l x_l^T x_k \\ z_{y_k} &= r_k = \sum_{l=1}^{N} \beta_l y_l^T y_k. \end{aligned} \tag{7.66}$$

The eigenvalues λ will be both positive and negative for the CCA problem. Also note that one has $\rho \in [-1, 1]$.

7.4.3 *Extension to kernel CCA*

Let us now extend this formulation for linear CCA to a nonlinear version by mapping the input space to a high dimensional feature space. The score variables are now

$$\begin{aligned} z_x &= w^T(\varphi_1(x) - \hat{\mu}_{\varphi_1}) \\ z_y &= v^T(\varphi_2(y) - \hat{\mu}_{\varphi_2}) \end{aligned} \tag{7.67}$$

where $\varphi_1(\cdot) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_{h_x}}$ and $\varphi_2(\cdot) : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_{h_y}}$ are mappings (which can be chosen to be different) to high dimensional feature spaces and $\hat{\mu}_{\varphi_1} = (1/N)\sum_{k=1}^N \varphi_1(x_k)$, $\hat{\mu}_{\varphi_2} = (1/N)\sum_{k=1}^N \varphi_2(y_k)$. It is important to take centering into account for the nonlinear CCA problem.

One starts from the primal problem

$$\left[\begin{array}{lll} \boxed{\mathrm{P}} : \max\limits_{w,v,e,r} & \gamma \sum_{k=1}^{N} e_k r_k - \nu_1 \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \nu_2 \frac{1}{2} \sum_{k=1}^{N} r_k^2 - \frac{1}{2} w^T w - \frac{1}{2} v^T v & \\ \text{such that} & e_k = w^T(\varphi_1(x_k) - \hat{\mu}_{\varphi_1}), & k = 1, ..., N \\ & r_k = v^T(\varphi_2(y_k) - \hat{\mu}_{\varphi_2}), & k = 1, ..., N \end{array} \right] \tag{7.68}$$

with Lagrangian

$$\mathcal{L}(w,v,e,r;\alpha,\beta) = \gamma\sum_{k=1}^{N} e_k r_k - \nu_1\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \nu_2\frac{1}{2}\sum_{k=1}^{N} r_k^2 - \frac{1}{2}w^T w - \frac{1}{2}v^T v - \sum_{k=1}^{N} \alpha_k[e_k - w^T(\varphi_1(x_k) - \hat{\mu}_{\varphi_1})] - \sum_{k=1}^{N} \beta_k[r_k - v^T(\varphi_2(y_k) - \hat{\mu}_{\varphi_2})] \tag{7.69}$$

where α_k, β_k are Lagrange multipliers. Note that w and v might be infinite dimensional now. For this reason it is necessary to derive the dual problem.

The conditions for optimality are

$$\left\{ \begin{array}{lll} \frac{\partial\mathcal{L}}{\partial w} = 0 & \rightarrow & w = \sum_{k=1}^{N} \alpha_k(\varphi_1(x_k) - \hat{\mu}_{\varphi_1}) \\ \frac{\partial\mathcal{L}}{\partial v} = 0 & \rightarrow & v = \sum_{k=1}^{N} \beta_k(\varphi_2(y_k) - \hat{\mu}_{\varphi_2}) \\ \frac{\partial\mathcal{L}}{\partial e_k} = 0 & \rightarrow & \gamma v^T(\varphi_2(y_k) - \hat{\mu}_{\varphi_2}) = \nu_1 w^T(\varphi_1(x_k) - \hat{\mu}_{\varphi_1}) + \alpha_k \\ & & \qquad\qquad k = 1, ..., N \\ \frac{\partial\mathcal{L}}{\partial r_k} = 0 & \rightarrow & \gamma w^T(\varphi_1(x_k) - \hat{\mu}_{\varphi_1}) = \nu_2 v^T(\varphi_2(y_k) - \hat{\mu}_{\varphi_2}) + \beta_k \\ & & \qquad\qquad k = 1, ..., N \\ \frac{\partial\mathcal{L}}{\partial \alpha_k} = 0 & \rightarrow & e_k = w^T(\varphi_1(x_k) - \hat{\mu}_{\varphi_1}) \\ & & \qquad\qquad k = 1, ..., N \\ \frac{\partial\mathcal{L}}{\partial \beta_k} = 0 & \rightarrow & r_k = v^T(\varphi_2(y_k) - \hat{\mu}_{\varphi_2}) \\ & & \qquad\qquad k = 1, ..., N \end{array} \right. \tag{7.70}$$

which results in the following dual problem after defining $\lambda = 1/\gamma$

$$\left[\begin{array}{l} \boxed{\mathrm{D}}: \text{ solve in } \alpha, \beta : \\ \\ \left[\begin{array}{cc} 0 & \Omega_{c,2} \\ \Omega_{c,1} & 0 \end{array} \right] \left[\begin{array}{c} \alpha \\ \beta \end{array} \right] \\ \\ = \lambda \left[\begin{array}{cc} \nu_1\Omega_{c,1} + I & 0 \\ 0 & \nu_2\Omega_{c,2} + I \end{array} \right] \left[\begin{array}{c} \alpha \\ \beta \end{array} \right] \end{array} \right] \tag{7.71}$$

where

$$\begin{aligned} \Omega_{c,1_{kl}} &= (\varphi_1(x_k) - \hat{\mu}_{\varphi_1})^T(\varphi_1(x_l) - \hat{\mu}_{\varphi_1}) \\ \Omega_{c,2_{kl}} &= (\varphi_2(y_k) - \hat{\mu}_{\varphi_2})^T(\varphi_2(y_l) - \hat{\mu}_{\varphi_2}) \end{aligned} \tag{7.72}$$

are the elements of the centered Gram matrices for $k, l = 1, ..., N$. In practice these matrices can be computed by $\Omega_{c,1} = M_c\Omega_1 M_c$, $\Omega_{c,2} = M_c\Omega_2 M_c$ with centering matrix M_c. The eigenvalues and eigenvectors that give an optimal correlation coefficient value are selected. The resulting score variables can be computed by applying the kernel trick with kernels $K_1(x_k, x_l) = \varphi_1(x_k)^T\varphi_1(x_l)$, $K_2(y_k, y_l) = \varphi_2(y_k)^T\varphi_2(y_l)$.

Using the CCA method one is in search of finding interesting relations between variables. One may apply this for example for input selection. At this point it is important to make a good choice of the tuning parameters and of the kernels and their tuning parameters. One may use an additional validation set to ensure meaningful generalization of the method. Further extensions towards independent component analysis (ICA) in relation to multiway parafac-candecomp models may be studied in view of rank-one approximation to high order tensors and a generalized Rayleigh quotient as defined in [302]. The CCA formulation can also be further related to partial least squares (PLS), for which kernel versions have been studied in [195].