ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 203-208

Kernel PLS variants for regression

L. Hoegaerts, J.A.K. Suykens, J. Vandewalle, B. De Moor
KU Leuven, Dept. of Electrical Engineering ESAT-SCD-SISTA,
Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium.
Email: {luc.hoegaerts, johan.suykens}@esat.kuleuven.ac.be

Abstract. We focus on covariance criteria for finding a suitable subspace for regression in a reproducing kernel Hilbert space: kernel principal component analysis, kernel partial least squares and kernel canonical correlation analysis, and we demonstrate how this fits within a more general context of subspace regression. For the kernel partial least squares case some variants are considered and the methods are illustrated and compared on a number of examples.

1 Introduction

Over the last years one can see certain learning algorithms being transferred to a kernel representation [10, 12]. The benefit lies in the fact that nonlinearity can be allowed, while avoiding to solve a nonlinear optimization problem. In this paper we focus on least squares regression models in the kernel context. By means of a nonlinear map into a Reproducing Kernel Hilbert Space (RKHS) [13] the data are projected to a high-dimensional space.

Estimation of regression coefficients in the RKHS will then be performed in a new basis constructed by optimization of (co)variance criteria. Next to PCA and CCA, we explore some intermediate PLS variants to see whether any gain can be involved. By reducing the new basis and projecting data to a subspace, the number of regression parameters is controlled. This scheme can reduce multicollinearity between the new variables by reducing variance on the least squares estimators. A better generalization can be obtained in this way for regression models.

LH is a PhD student supported by the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT). JS is a postdoctoral researcher with the National Fund for Scientific Research FWO-Flanders and a professor at KUL. BDM and JVDW are full professors at KUL. This research work was carried out at the ESAT laboratory of the KUL and supported by grants from several funding agencies and sources: Research Council KUL: Concerted Research Action GOA-Mefisto 666, IDO; Flemish Government: FWO-Flanders (G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02), ICCoS, ANMMM, AWI, IWT (Soft4s), STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE; Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001), IUAP V-10-29 (2002-2006), Prog. Sust. Dev. PODO-II (CP/40); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS.

This short paper is organized as follows. In section 2 we give minimal background on RKHS and introduce regression in a feature subspace. In section 3 we overview six pls variants for finding an optimal subspace. In section 4 we give their kernel versions. In section 5 we compare the methods on sinc data.

2 Subspace regression in RKHS

Assume data $\{\{(\mathbf{x_i}, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}\}$ have been given. A kernel k provides a similarity measure between pairs of data points $k : X \times X \rightarrow \mathbb{R} : (\mathbf{x_1}, \mathbf{x_2}) \mapsto k(\mathbf{x_1}, \mathbf{x_2})$. Once a kernel is chosen, one can associate to each $x \in X$ a mapping $\varphi : X \rightarrow H_k : \mathbf{x} \mapsto k(\mathbf{x}, \cdot)$, which can be evaluated at $\mathbf{x}'$ to give $\varphi(\mathbf{x})(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$. One obtains a RKHS H_k on a compact X under the condition that the kernel is positive definite [1]. The representer theorem [6] says that the solution to a regularized cost function in a RKHS is constrained to the subspace spanned by the mapped data points. The Mercer-Hilbert-Schmidt theorem states that for each positive definite kernel there exists an orthonormal set $\{\phi_i\}_{i=1}^d$ with non-negative λ_i such that we have following spectral decomposition: $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$ where $d \leq \infty$, and λ_i and ϕ_i are the eigenvalues and eigenvectors of the kernel. This allows to express $\varphi(\mathbf{x})$ in this ϕ basis, so that the φ mapping can be identified with a $d \times 1$ *feature* vector $\varphi(\mathbf{x}) = [\sqrt{\lambda_1}\phi_1(\mathbf{x}) \quad \sqrt{\lambda_2}\phi_2(\mathbf{x}) \quad ... \quad \sqrt{\lambda_d}\phi_d(\mathbf{x})]^T$. One can then construct an $n \times d$ feature matrix in RKHS: $\Phi = [\varphi^T(\mathbf{x_1}) \; \varphi^T(\mathbf{x_2}) \ldots \varphi^T(\mathbf{x_n})]^T$.

The goal is to obtain a regression estimate of the underlying function, given the data. Here, we consider the standard linear multivariate regression model in feature space $Y = \Phi W + E$, where Y represents a $n \times q$ matrix of observations of the dependent variable, W is the unknown $d \times q$ matrix of regression coefficients and E is a $n \times q$ matrix of errors with zero-mean Gaussian i.i.d. values of equal variance σ^2 (unknown). We assume all mapped data variables have been mean-centered.

A first difficulty in this setup, is that usually $d \gg n$, and potentially even $d = +\infty$. In order to restrict the infinite number of regression coefficients, one can introduce a projection of the data into a subspace of finite dimension $m \ll d$. Hence the name subspace regression. If we gather the basis vectors $\{\mathbf{v_i}\}_{i=1}^m$ of the subspace in the columns of a $d \times m$ transformation matrix V, we can express the $\varphi(\mathbf{x_i})$ in the new coordinates, so that $Z = \Phi V$ and the confined regression model becomes $Y = ZW + E$. The particular choice of the basis vectors will be discussed in section 3. To estimate the unknown true model parameters, the elements of W, we choose here to minimize the error E in least squares sense: $\min_W \|Y - ZW\|_2^2$.

A second difficulty is that in general the elements of matrix Φ are unknown because the explicit expression for φ is not available. An elegant and optimal approach is to make use of the so-called kernel trick [10], which makes use of the decomposition $V = \Phi^T A$. This allows to write $Z = \Phi V = \Phi \Phi^T A = KA$ and our central model equation becomes

$$Y = (KA)W + E. \tag{1}$$

Since we assumed that the data are centered, we need to adjust the expression $\varphi(\mathbf{x})$ with $\varphi(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n}\varphi(\mathbf{x_i})$ everywhere. For the kernel matrices this has the consequence of replacing K by [9] $K := (I_n - \frac{1}{n}1_n1_n^T)K(I_n - \frac{1}{n}1_n1_n^T)$ where I_n is the identity matrix and 1_n a vector of ones.

3 Subspace construction

Basically, we need to find directions in the variable space that form a new basis, upon which we can project. Several criteria can be chosen.

A first criterion is minimization of within-space correlation or multicollinearity, which produces uncertainty in regression coefficient estimates. Principal Component Analysis (PCA) [5] reduces this:

$$\max_{\mathbf{v}} \text{var } (\mathbf{v^T x}) = \mathbf{v^T} C_{xx}\mathbf{v} \text{ s.t. } \|\mathbf{v}\| = 1 \tag{2}$$

with $C_{xx} = X^TX$ the sample covariance matrix and $V^TV = I_p$. This involves a diagonalization procedure which requires solving an eigenvalue problem $C_{xx}\mathbf{v} = \lambda\mathbf{v}$. In dimension reduction one describes PCA as the search for a best fitting subspace in a least squares sense by minimizing $J_{\text{PCA}}(\mathbf{v}) = \sum_{i=1}^{n}\|\mathbf{x_i} - \mathbf{v}\mathbf{v^T}\mathbf{x_i}\|^2$.

A second criterion is maximization of between-space correlation. For the purpose of prediction one wishes strong correlation with the target vectors. Canonical Correlation Analysis (CCA) [4] implements this:

$$\max_{\mathbf{v},\mathbf{w}} \text{ corr}(\mathbf{v^T x}, \mathbf{w^T y}) = \frac{\mathbf{v^T C_{xy} w}}{\sqrt{\mathbf{v^T C_{xx} v}}\ \sqrt{\mathbf{w^T C_{yy} w}}} \text{ s.t. } \|\mathbf{v^T x}\| = 1 = \|\mathbf{w^T y}\|, \tag{3}$$

with $C_{xy} = X^TY$. Essentially this requires the solution of the system $\mathbf{C_{xy}w} = \lambda\ \mathbf{C_{xx}v}, \mathbf{C_{yx}v} = \lambda\ \mathbf{C_{yy}w}$. The new basises in both spaces are chosen such that the vector components (projections) of all data maximally coincide. And in least squares sense: $J_{\text{CCA}}(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{n}\|\mathbf{v^T x_i} - \mathbf{w^T y_i}\|^2$, minimizing difference between the cosine of angles of lines in both spaces.

In between these, Partial Least Squares (PLS) [14] can be positioned:

$$\max_{\mathbf{v},\mathbf{w}} \text{cov } (\mathbf{v^T x}, \mathbf{w^T y}) = \mathbf{v^T} C_{xy}\mathbf{w} \text{ s.t. } \|\mathbf{v}\| = 1 = \|\mathbf{w}\|. \tag{4}$$

Solutions can be obtained by using Lagrange multipliers, which leads to solving the PLS-SVD system $C_{xy}\mathbf{w} = \lambda\mathbf{v}, C_{yx}\mathbf{v} = \lambda\mathbf{w}$. As a least squares cost function, $J_{\text{PLS}}(\mathbf{v},\mathbf{w}) = J_{\text{PCA}}(\mathbf{v}) + J_{\text{CCA}}(\mathbf{v},\mathbf{w}) + J_{\text{PCA}}(\mathbf{w})$.

Imposing other constraints, results in other PLS variants. The original version of Wold, PLS-WA, computes consecutively the first left and right singular vectors of $(X^{(r)})^TY^{(r)}$, after which each time the data matrices are deflated (projected into the complement of the space spanned by the previous found new variables, or scores):

$$\begin{array}{rcll} X^{(r+1)} & = & X^{(r)} - \mathbf{b_r}(\mathbf{b_r^T b_r})^{\mathbf{-1}}\mathbf{b_r^T}X^{(r)} & \text{with} \quad \mathbf{b_r} = X^{(r)}\mathbf{v_r} \\ Y^{(r+1)} & = & Y^{(r)} - \mathbf{a_r}(\mathbf{a_r^T a_r})^{\mathbf{-1}}\mathbf{a_r^T}Y^{(r)} & \text{with} \quad \mathbf{a_r} = Y^{(r)}\mathbf{w_r}. \end{array} \tag{5}$$

As such the orthogonality of the scores is guaranteed in both spaces. Its directly adapted most used variant has resulted in PLS2 or PLS1, where y-space is being deflated with the x-space score.

We included PLS-U, not with an orthogonality constraint, but more strongly, uncorrelatedness with the previously found coefficients: $V_r^T C_{xx} V_r = I_p$ and $W_r^T C_{yy} W_r = I_q$. By Lagrange multipliers one arrives after some manipulation again at an eigenproblem with deflations:

$$\begin{array}{rcl} X^{(r+1)} & = & X^{(r)} - (C_{xx}V_r)((C_{xx}V_r)^T(C_{xx}V_r))^{-1}(C_{xx}V_r)^T X^{(r)} \\ Y^{(r+1)} & = & Y^{(r)} - (C_{yy}W_r)((C_{yy}W_r)^T(C_{yy}W_r))^{-1}(C_{yy}W_r)^T Y^{(r)}. \end{array} \quad (6)$$

The PLS least squares interpretation, allows to add immediately two more variants. Leaving out the compensating PCA term in x space, we obtain PLSx from CCA, with $C_{xx} = I_p$. By symmetry, also a PLSy version can be obtained.

4 Introducing the kernel function

Starting from the PLS covariance criterion, we can proceed likewise in feature space, but now with $\mathbf{v}$ and $\mathbf{w}$ as $d \times 1$ feature space vectors. To arrive at calculation with kernels instead of feature vectors one typically expands the new basis vectors as follows: $\mathbf{v} = \sum_{i=1}^n a_i \varphi(\mathbf{x_i}) = \Phi_1^T \mathbf{a}$ and $\mathbf{w} = \sum_{i=1}^n b_i \varphi(\mathbf{y_i}) = \Phi_2^T \mathbf{b}$. This will result in $\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^T K_{xx} K_{yy} \mathbf{b} / \sqrt{\mathbf{a}^T K_{xx} \mathbf{a}} \sqrt{\mathbf{b}^T K_{yy} \mathbf{b}}$, where $[K_{xx}]_{ij} = \varphi(\mathbf{x_i})^{\mathbf{T}} \varphi(\mathbf{x_j})$ and $[K_{yy}]_{ij} = \varphi(\mathbf{y_i})^{\mathbf{T}} \varphi(\mathbf{y_j})$ are the kernel Gram matrices and $\mathbf{v}$ and $\mathbf{w}$ were divided by their norm. PLS-SVD will then be determined by:

$$\begin{array}{rcl} K_{yy}\mathbf{b} & = & \lambda \mathbf{a} \\ K_{xx}\mathbf{a} & = & \lambda \mathbf{b}. \end{array} \quad (7)$$

In the same approach KPCA [9] was derived as the eigenproblem $K_{xx}\mathbf{a} = \lambda \mathbf{a}$.

For KPLS-WA and KPLS-U we can substitute X by Φ_1 and Y by Φ_2. But because of the unknown elements of Φ, we cannot directly obtain the SVD of $(\Phi_1^r)^T \Phi_2^r$. The NIPALS-PLS algorithm [14] allows to circumvent this issue and delivers the $\mathbf{a}$ and $\mathbf{b}$ as first singular vectors of $K_{xx}^{(r)} K_{yy}^{(r)}$ and $K_{yy}^{(r)} K_{xx}^{(r)}$. Then the deflation expressions become for PLS-U at step r:

$$\begin{array}{rcl} K_{xx}^{(r+1)} & = & K_{xx} - K_{xx}^2 A_r (A_r^T K_{xx}^3 A_r)^{-1} A_r^T K_{xx}^2 \\ K_{yy}^{(r+1)} & = & K_{yy} - K_{yy}^2 B_r (B_r^T K_{yy}^3 B_r)^{-1} B_r^T K_{yy}^2. \end{array} \quad (8)$$

And for PLS-WA (which resembles the kernel version of PLS1 [8]) we have:

$$\begin{array}{rcl} K_{xx}^{(r+1)} & = & K_{xx}^{(r)} - \mathbf{a_r}\mathbf{a_r^T} K_{xx}^{(r)} - K_{xx}^{(r)} \mathbf{a_r}\mathbf{a_r^T} + \mathbf{a_r}\mathbf{a_r^T} K_{xx}^{(r)} \mathbf{a_r}\mathbf{a_r^T} \\ K_{yy}^{(r+1)} & = & K_{yy}^{(r)} - \mathbf{b_r}\mathbf{b_r^T} K_{yy}^{(r)} - K_{yy}^{(r)} \mathbf{b_r}\mathbf{b_r^T} + \mathbf{b_r}\mathbf{b_r^T} K_{yy}^{(r)} \mathbf{b_r}\mathbf{b_r^T}. \end{array} \quad (9)$$

For KPLSx and KPLSy we point out that its solutions can be considered in the optimization context as special cases of the 'regularized' KCCA variant which was proposed in the context of primal-dual LS-SVM formulations [11, 12]:

$$\begin{array}{rcl} K_{yy}\mathbf{b} & = & \lambda(\nu_1 K_{xx} + I)\mathbf{a}, \\ K_{xx}\mathbf{a} & = & \lambda(\nu_2 K_{yy} + I)\mathbf{b}, \end{array} \quad (10)$$

if one fixes the positive parameters (ν_1, ν_2) respectively as $(1, 0)$ and $(0, 1)$. Even KPLS-SVD is a subcase, with $(\nu_1, \nu_2) = (0, 0)$. The 'regular' KCCA was originally reported by [7] and [2], in an independent component analysis (ICA) context. So all these closely related methods deliver us useful subspaces spanned by the columnspace of A (eigenvectors $\{a_i\}_{i=1}^m$, ordered corresponding to nondecreasing values of the eigenvalues λ_i) in (1).

5 Experiments

To demonstrate some characteristics of these subspace regression methods, we applied them to the sinc function. We considered a domain dataset of 100 equally-spaced points in the interval [-10,10]. The corresponding output values were centralized. We used a Gaussian kernel $\exp(-\|\mathbf{x_i} - \mathbf{x_j}\|_2^2/h^2)$ with $h = 1$. In Fig.1(left) a typical picture of the first three components qualitatively show a good correlation with the targets. Non-equally-spaced sampling causes the components to be more irregular and oscillatory, while prediction will be less performant in undersampled regions and near boundaries. In Fig.1(right) we show an interpolation example on the same data, but with added Gaussian noise with standard deviation $\sigma = 0.2$. The other methods give similar component profiles and prediction results.

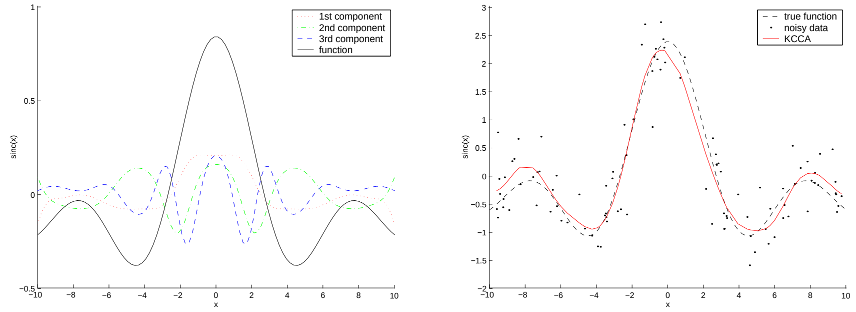

Figure 1: (Left) Visualisation of the first three components extracted by KCCA; (Right) Example of prediction on a noisy sinc ($\sigma = 0.2$, $m = 4$, $h = 1$, $\nu_1 = 1 = \nu_2$).

We compared for 100 randomizations on sinc data sets (noise added with $\sigma = 0.2$) with results shown in Table 1. The number of components $m \in \{1, 2, .., 30\}$ was determined by 10-fold cross-validation (CV), $h = 0.15$ was fixed. On average all methods perform equally well in terms of mean square error (MSE) on an independent test data set, with comparable variance. KPLS1 performs best in terms of lowest number of components, followed by KCCA and KPLSy. When instead the deflation of y-space on the x scores of KPLS1 is plugged into PLS-U and PLS-WA, one improves to $m = 5 \pm 3$ and 5 ± 4, respectively. As for the parameters ν_1 and ν_2 we may conclude that the regression result is fairly insensitive to ν_2, but that large values of ν_1 cause overfitting. The use of other kernels, like the polynomial or the sigmoidal kernel, did not produce better results. The results are similar to LS-SVMs for regression (kernel ridge regression or regularization network [3] with additional bias term) using the LS-SVMlab software http://www.esat.kuleuven.ac.be/sista/lssvmlab/. We also tested the methods on the larger, real-world dataset benchmark of Boston Housing data and observed again comparable performance.

method	comps.	Mean Square Error	method	comps.	Mean Square Error
KPCA	25 ± 4	0.3875 ± 0.0914	KPLSx	19 ± 6	0.3864 ± 0.0972
KPLS-SVD	19 ± 6	0.3888 ± 0.1014	KPLSy	15 ± 7	0.3908 ± 0.0861
KPLS-WA	20 ± 5	0.3801 ± 0.0863	KPLS1	3 ± 1	0.3825 ± 0.0879
KPLS-U	24 ± 5	0.3778 ± 0.0812	KCCA	11 ± 7	0.3903 ± 0.0445

Table 1: Comparison of KPLS variants and KCCA, which correspond to forms of subspace regression in RKHS with different choices of matrix A in (1). Shown are the number of selected components and test set performance for 100 randomizations.

6 Conclusions

We presented three related methods -PCA, PLS and CCA- in a unified view from the viewpoint of RKHS regression, where they deliver a suitable subspace in order to reduce the potentially high number of regression parameters in high-dimensional feature space. We considered some additional kernel variants of the PLS case and they all compare to KPCA and KCCA in performance, where a proper deflation scheme keeps their number of needed components smaller.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 686:337–404, 1950.

[2] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

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

[4] R. Gittins. *Canonical analysis.* Springer-Verlag, 1985.

[5] I.T. Jolliffe. *Principal Component Analysis.* Springer Verlag, 1986.

[6] G.S. Kimeldorf and G. Wahba. Tchebycheffian spline functions. *J. Math. Ana. Applic.*, (33):82–95, 1971.

[7] P.L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems, submitted for publication.*

[8] R. Rosipal and L.J. Trejo. Kernel partial least squares regression in rkhs. *Journal of Machine Learning Research*, 2:97–123, Dec 2001.

[9] B. Schölkopf, A.J. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[10] B. Schölkopf and A. Smola. *Learning with kernels.* MIT Press, 2002.

[11] J.A.K. Suykens, T. Van Gestel, J. Vandewalle, and B. De Moor. A Support Vector Machine Formulation to PCA Analysis and its Kernel Version. *IEEE Transactions on Neural Networks*, in press.

[12] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines.* World Scientific, Singapore, 2002.

[13] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* SIAM, 1990.

[14] H. Wold. Estimation of principal components and related models by iterative least squares. In Krishnaiah, editor, *Multiv. An.*, pages 391–420. N.Y., 1966.