On Kernel Principal Component Regression with Covariance Inflation Criterion for Model Selection

Roman Rosipal
Mark Girolami
Leonard J. Trejo

TECHNICAL REPORT

University of Paisley
School of Information and Communication Technologies
Paisley, PA1 2BE
Scotland, UK

March 2001

On Kernel Principal Component Regression with Covariance Inflation Criterion for Model Selection

Roman Rosipal
Applied Computational Intelligence Research Unit
School of Information and Communication Technologies
University of Paisley
Paisley PA1 2BE, Scotland
rosi-ci0@paisley.ac.uk

Mark Girolami
Applied Computational Intelligence Research Unit
School of Information and Communication Technologies
University of Paisley
Paisley PA1 2BE, Scotland
giro-ci0@paisley.ac.uk

Leonard J. Trejo
Computational Sciences Division
NASA Ames Research Center
Moffett Field, CA
ltrejo@mail.arc.nasa.gov

Abstract

This paper introduces Kernel Principal Component Regression (PCR) with the Covariance Inflation Criterion (CIC) for model order selection. The relation to Kernel Ridge Regression (RR) and other 'kernel' regression techniques is given and two benchmark problems demonstrate the comparable performance of CIC to cross-validation techniques. In all reported experiments CIC provides the models with equal performance in comparison to Kernel RR. Moreover, on a significant real world application, Kernel PCR with CIC resulted in smaller model compared to Kernel PCR with the cross-validation technique employed for the selection of principal components.

1 Introduction

The main problem with existing regression techniques is their poor generalization properties. In the case of ill-posed problems overfitting is the outcome of selecting an inappropriate model structure based on a finite number of examples $\{\mathbf{x}_i, y_i\}_{i=1}^n$. In practice the ill-posed problem can be viewed as a training data set which possesses a small amount of information about the desired solution. To overcome this problem a regularized formulation of regression can be considered as a variational problem

$$\min_{f \in \mathcal{H}} R_{reg}(f) = \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)) + \xi \|f\|_{\mathcal{H}}^2 \tag{1}$$

leading to a general solution of the form [18]

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + \sum_{j=1}^{l} b_j \upsilon_j(\mathbf{x}) \tag{2}$$

where the functions $\{\upsilon_j(.)\}_{j=1}^l$ span the null space of a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ [1] and the coefficients $\{c_i\}_{i=1}^n$, $\{b_j\}_{j=1}^l$ are given by the data. For the purposes of this paper we will only assume the case when $l = 1$ and $\upsilon_1(\mathbf{x}) = const \quad \forall \mathbf{x}$. $\|f\|_{\mathcal{H}}^2$ is a norm in RKHS defined by the Mercer kernel $K(\mathbf{x}, \mathbf{y})$; i.e. a positive definite function of the form $K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}).\Phi(\mathbf{y}) \rangle$ [1, 18]. $K(\mathbf{x}, \mathbf{y})$ also corresponds to a canonical dot product in a possibly high dimensional space $\mathcal{F}$ where the input data are mapped by $\Phi : R^N \to \mathcal{F}$ (see e.g. [12]). This correspondence also gave rise to the unification of Regularization Networks, SVR, Gaussian processes and

spline methods [17, 4, 19, 18, 2]. In this paper we focus our attention on the case of the quadratic loss function $V = (y_i - f(\mathbf{x}_i))^2$ and will assume non-linear regression models of the form $f(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) + b$; i.e. linear models in a feature space $\mathcal{F}$ where input data $\{\mathbf{x}_i\}_{i=1}^n$ are mapped through a non-linear function $\Phi(.)$. If y is measured with additive Gaussian noise the most appropriate loss function is quadratic.

The multicollinearity or near-linear dependence of regressors is a serious problem which can dramatically influence the effectiveness of a regression model. Multicollinearity results in large variances and covariances for the least-squares estimators of the regression coefficients. Multicollinearity can also produce estimates of the regression coefficients that are too large in absolute value. Thus the values and signs of estimated regression coefficients may change considerably given different data samples. This effect can lead to a regression model which fits the training data reasonably well, but in general bad generalization of the model can occur. This fact is in a very close relation to the argument stressed in [14], where the authors have shown that choosing the *flattest* function[1] in a feature space can, based on the smoothing properties of the selected kernel function, lead to a smooth function in the input space. We discuss two methods which deal with multicollinearity – ridge regression and principal component (PC) regression.

In [7, 8] we proposed the Kernel PCR technique based on an orthogonal projection of the original regressors in feature space $\mathcal{F}$ onto PC's found by Kernel Principal Component Analysis (PCA) [13]. We will show that a final solution of Kernel PCR leads to the form (1) and will also highlight the relation to the ridge regression technique in feature space $\mathcal{F}$.

In the current study we have used CIC, recently proposed by Tibshirani and Knight [15], for model selection in the case of Kernel PCR. For model selection in orthogonal linear regression CIC provided superior performance to the well know Bayesian and Akaike's information criteria [15]. CIC adjusts the *in-sample* training error by applying the model selection rule to permuted versions of the data set and evaluates the covariance of the predictions and true targets.

[1] The *flatness* is defined in the sense of penalizing high values of the regression coefficients estimate.

2 Kernel Principal Component Regression

Consider the standard regression model in feature space $\mathcal{F}$

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{y}$ is a vector of n observations of the dependent variable, $\mathbf{Z}$ is an $(n \times M)$ matrix of regressors whose i-th row is the vector $\Phi(\mathbf{x}_i)$ of the mapped $\mathbf{x}_i$ observation into $M \leq \infty$ dimensional feature space $\mathcal{F}$, $\boldsymbol{\gamma}$ is a vector of regression coefficients and $\boldsymbol{\epsilon}$ is the vector of error terms whose elements have equal variance σ^2, and are independent of each other. In fact we should assume the more general functional linear models, but in this paper we will work with vectors rather than a functional representation of the data. We also assume a 'centered' form of the model which can be easily achieved by centering the mapped data in $\mathcal{F}$ by (11) and (12) [13]. Thus $\mathbf{Z}^T\mathbf{Z}$ is proportional to the sample covariance matrix and Kernel PCA can be performed to extract M eigenvalues $\{\lambda_j\}_{j=1}^{M}$ and corresponding eigenvectors $\{\mathbf{V}^j\}_{j=1}^{M}$[2]. The k-th nonlinear PC of $\mathbf{x}$ is given as the projection of $\Phi(\mathbf{x})$ onto the eigenvector $\mathbf{V}^k$

$$\zeta(\mathbf{x})_k := \langle \mathbf{V}^k . \Phi(\mathbf{x}) \rangle = \sum_{i=1}^{n} \alpha_i^k K(\mathbf{x}_i, \mathbf{x}). \tag{4}$$

By the Kernel PCA projection of all original regressors onto the PC's we can rewrite (3) as

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{X} = \mathbf{Z}\mathbf{V}$ is now an $(n \times M)$ matrix of transformed regressors and $\mathbf{V}$ is a $(M \times M)$ matrix whose k-th column is the eigenvector $\mathbf{V}^k$. The columns of the matrix $\mathbf{X}$ are now orthogonal and the least squares estimate of the coefficients $\mathbf{w}$ becomes

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \boldsymbol{\Lambda}^{-1}\mathbf{X}^T\mathbf{y}, \tag{6}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$. The results obtained using all principal components in (5) is equivalent to that obtained by least squares using the

[2]We are theoretically assuming that $n > M$. Otherwise we have to deal with the singular case ($n \leq M$) allowing us to extract only up to n eigenvectors corresponding to non-zero eigenvalues.

original regressors. In fact we can express the estimate $\hat{\boldsymbol{\gamma}}$ of the original model (3) as

$$\hat{\boldsymbol{\gamma}} = \mathbf{V}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \sum_{i=1}^{M} \lambda_i^{-1}\mathbf{V}^i(\mathbf{V}^i)^T\mathbf{Z}^T\mathbf{y}$$

and its corresponding variance-covariance matrix [6] as

$$cov(\hat{\boldsymbol{\gamma}}) = \sigma^2\mathbf{V}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{V}^T = \sigma^2\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T = \sigma^2\sum_{i=1}^{M} \lambda_i^{-1}\mathbf{V}^i(\mathbf{V}^i)^T \ . \qquad (7)$$

To avoid the problem of multicollinearity PCR uses only some of the PC's. It is clear from (7) that the influence of small eigenvalues can significantly increase the overall variance of the estimate. PCR simply deletes the PC's corresponding to small values of the eigenvalues λ_i, i.e. the PC's where multicollinearity may appear. The penalty we have to pay for the decrease in variance of the regression coefficient estimate is bias in the final estimate. However, if multicollinearity is a serious problem, the introduced bias can have a less significant effect in comparison to a high variance estimate. If the elements of $\mathbf{w}$ corresponding to deleted regressors are zero, an unbiased estimate is achieved [6].

Using the first N-nonlinear PC's to create orthogonal regressors (4) for our Kernel PCR model (5) we can formulate the solution as

$$f(\mathbf{x}, \mathbf{c}) = \sum_{k=1}^{N} w_k\zeta(\mathbf{x})_k + b = \sum_{k=1}^{N} w_k \sum_{i=1}^{n} \alpha_i^k K(\mathbf{x}_i, \mathbf{x}) + b = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + b, \qquad (8)$$

where $\{c_i = \sum_{k=1}^{N} w_k\alpha_i^k\}_{i=1}^{n}$ and b represents a bias term.

We have shown that by removing the PC's whose variances are very small we can eliminate large variances of the estimate due to multicollinearities. However, if the orthogonal regressors corresponding to those PC's have large correlation with the dependent variable y such deletion is undesirable (experimentally demonstrated in Section 4). There are several different strategies for selecting the appropriate orthogonal regressors for the final model (see [6] and ref. therein). We now consider the recently proposed CIC for model selection in KPCR as a novel alternative to methods such as cross-validation.

2.1 Covariance Inflation Criterion

We provide a brief overview of the CIC (for more detailed description see [15]). For clarity we use the same notations as in [15]. Define the (average)

optimism

$$\mathrm{op}(\beta) = E\{\mathrm{Err}(\beta) - \overline{\mathrm{err}}(\beta)\}$$

where $\overline{\mathrm{err}}(\beta)$ is a training error of the best model M_β and $\mathrm{Err}(\beta)$ its test set prediction error. The CIC represents the estimate of the optimism $\mathrm{op}(\beta)$ plus $\overline{\mathrm{err}}(\beta)$ and is defined as

$$\mathrm{cic}(\beta) = \overline{\mathrm{err}}(\beta) + \frac{2}{n}\frac{\hat{\sigma}^2}{\sigma_y^2}\sum \mathrm{cov}^0\{y_i^*, \eta_{\mathbf{z}^*}(\mathbf{x}_i, M_\beta^*)\} + \frac{2}{n}\hat{\sigma}^2, \tag{9}$$

where cov^0 represents covariance under the permutation distribution of $\mathbf{x}_i$ and y, i.e. $(\mathbf{x}_i, y_i^*)$ with $y_1^*, y_2^*, \ldots, y_n^*$ a sample drawn without replacement from $y_1, y_2, \ldots, y_n$ and the $\mathbf{x}_i$ fixed. M_β^* is the model for β estimate from the permuted data, $\hat{\sigma}^2$ is an estimate of the noise variance σ^2 and $\sigma_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$. In [15] authors proved that $\mathrm{cic}(\beta)$ is an unbiased estimate of true optimism $\mathrm{op}(\beta)$ for linear fitting and proposed the following algorithm to estimate the CIC. Fix the regressors $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and generate B random permutations of the targets $\mathbf{y}^{*b} = \{y_1^{*b}, y_2^{*b}, \ldots, y_n^{*b}\}_{b=1}^B$. Estimate the prediction model M_β based on the data sets $\mathbf{z}^{*b} = (\mathbf{x}, \mathbf{y}^{*b})$ and compute the predictions $\eta_i^{*b} = \eta_{\mathbf{z}^{*b}}(\mathbf{x}_i, M_\beta^*)$ of the model. Estimate the expression $\sum \mathrm{cov}^0\{y_i^*, \eta_{\mathbf{z}^*}(\mathbf{x}_i, M_\beta^*)\}$ by

$$\sum_{i=1}^n \sum_{b=1}^B (y_i^{*b} - \bar{y}_i)\eta_i^{*b}/B,$$

where the $\bar{y}_i$ is the true mean of the y_i^{*b} under permutation sampling. Calculate for the tuning parameter range, β (the model order parameter). In our experiments we found $B = 15$ to be satisfactory. A brief review of Kernel RR is now given.

3 Kernel Ridge Regression

Kernel RR is another technique to deal with multicollinearity by assuming the linear regression model (3) whose solution is now achieved by minimizing

$$R_{rr}(\boldsymbol{\gamma}) = \sum_{i=1}^n [y_i - f(\mathbf{x}_i, \boldsymbol{\gamma})]^2 + \xi\|\boldsymbol{\gamma}\|^2, \tag{10}$$

where $f(\mathbf{x}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}^T\Phi(\mathbf{x}) + \mathbf{b}$ and ξ is a regularization term. The least-squares estimate of $\boldsymbol{\gamma}$ is biased but the variance is decreased. Similar to the

Kernel PCR case we can express the variance-covariance matrix of the $\boldsymbol{\gamma}$ estimate [6] as

$$cov(\hat{\boldsymbol{\gamma}}) = \sigma^2 \sum_{i=1}^{M} \lambda_i (\lambda_i + \xi)^{-2} \mathbf{V}^i (\mathbf{V}^i)^T.$$

We can see, that in contrast to Kernel PCR, the variance reduction in Kernel RR is achieved by giving less weight to small eigenvalue PC's via the factor ξ.

In practice we usually do not know the explicit mapping $\Phi(.)$. However if the transformation is known, computation in the high-dimensional feature space $\mathcal{F}$ may be numerically intractable. We can derive the desired solution using the 'kernel' trick, i.e. to use the fact that $K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}).\Phi(\mathbf{y}) \rangle$ and express the solution in this dot product form [11, 2]

$$f(\mathbf{x}) = \mathbf{y}^T (\mathbf{K} + \xi I)^{-1} \mathbf{k} \, ,$$

where $\mathbf{K}$ is the Gram matrix consisting of dot products of the mapped input data $\mathrm{K}_{ij} = \langle \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j) \rangle$ $i, j = 1, \ldots, n$ and $\mathbf{k}$ is the vector of dot products of a new mapped input example $\Phi(\mathbf{x})$ and the vectors of the training set; $\mathrm{k}_i = \langle \Phi(\mathbf{x}_i).\Phi(\mathbf{x}) \rangle$.[3]

In this paper we assume centralized Kernel RR [9]; i.e. we assume the sample mean of the mapped data $\Phi(\mathbf{x}_i)$ and targets y_i to be zero. The centralization of the individual mapped data points is accomplished by the "centralization" of $\mathbf{K}$ and $\mathbf{K}_t$ matrices given by the

$$\mathbf{K} = (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \mathbf{K} (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \tag{11}$$

$$\mathbf{K}_t = (\mathbf{K}_t - \frac{1}{n} \mathbf{1}_{n_t} \mathbf{1}_n^T \mathbf{K}) (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \tag{12}$$

where $\mathbf{I}$ is an n dimensional identity matrix and $\mathbf{1}_n$, $\mathbf{1}_{n_t}$ represent the vectors whose elements are all ones, with length n and n_t, respectively. $\mathbf{K}_t$ represents the $(n_t \times n)$ kernel matrix for all n_t testing data points.

[3] The same form can be derived in the case of the dual representation of the Regularization Network minimizing (1) using the loss function $V(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$ [5, 3] or through the techniques derived from Gaussian processes [2].

4 Experiments

4.1 Benchmarks

First we tested our method on the well known Friedman#1 and Boston housing datasets. In both cases we used a 2-nd order polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}.\mathbf{y} \rangle + 1)^2$ which leads to the projection of the data sets into 66 and 91 dimensional feature spaces, respectively. For the Friedman data we randomly generated 100 training sets of size 200 and validation sets of size 40 and one 1000 example test set. In the case of the Boston housing data set we randomly split the data into 100 – 401/81/25 train/validation/test – partitions. The validation sets were used for cross-validation model selection in the case of Kernel PCR and for setting the regularization term ξ in the Kernel RR model. Because the CIC for model selection is based only on the *in-sample error* for comparative reasons we did not use a validation set for building the final models. In Figure 1 the number of selected regressors (model order) as a function of the total number of regressors used is depicted. We compared the cross-validation (CV) technique and CIC. Based on the eigen spectrum regressors corresponding to the first 20 most significant (largest eigenvalues) PC's (Friedman data set) were used in all models. On the Boston data set the first 5 most significant PC's were used. First regressors entering the model were selected by t-statistics representing their importance in the model. Although the best performing (test error) final models are on average created by 60% of all the available regressors, it is observed that some of the regressors with smaller variance entered the final models. In addition the final models selected by CIC gave on average a 2.5% lower test-set mean-square error (MSE) on both data sets in comparison to CV.

On three samples of the Boston housing data we observed a significant increase in the error using the Kernel RR method and we decided to create 3 new train/test partitions. In fact, the high prediction error variance using the 'kernel' regression methods on this data set was also observed in [11]. On both sets of data we did not observe significant differences in MSE on test data.

4.2 Human Signal Detection Performance Monitoring

We have used Event Related Potentials (ERPs) and performance data from an earlier study [16]. Eight male Navy technicians experienced in the operation of display systems performed a signal detection task. In this study

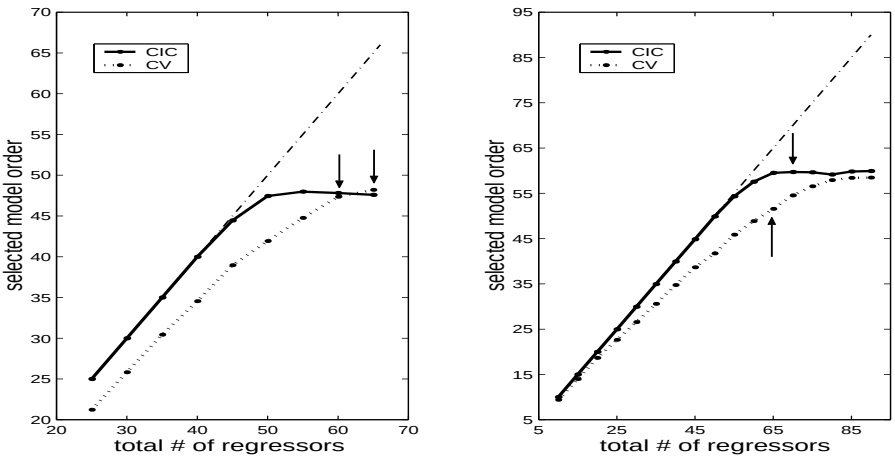

Figure 1: Comparison of the CIC (solid line) and cross-validation (CV) (dotted line) techniques for model selection on Friedman#1 (left) and Boston housing (right) data sets. The arrows indicate the models on which the best performance on test set was achieved. The results are averaged over 100 simulations.

we randomly selected two data sets. Performance of the operators was measured as a linear composite of speed, accuracy, and confidence. A single measure, PF1, was derived using factor analysis of the performance data for all subjects, and validated within subjects. The computational formula for PF1 was

$$\text{PF1} = 0.33*\text{Accuracy} + 0.53*\text{Confidence} - 0.51*\text{Reaction Time}$$

using standard scores for accuracy, confidence, and reaction time based on the mean and variance of their distributions across all subjects. PF1 varied continuously, being high for fast, accurate, and confident responses and low for slow, inaccurate, and unconfident responses. ERPs were recorded from midline frontal, central, and parietal electrodes (Fz, Cz, and Pz). The full description of the experimental setting can be found in [16].

The desired output PF1 was linearly normalized to have a range of 0 to 1. For each subject we split the data into 10 different 55% and 45% training and testing partitions. Eleven-fold CV to estimate desired parameters was applied on each training partition. After CV a final model was tested on an independent testing partition. For detail experimental setting and the principal component selection strategy used in the case of CV we refer the

reader to [10]. In the case of CIC the first 50% of the regressors entered all models and 2.5% of the regressors corresponding to the smallest eigenvalues were discarded from the models. As in the former case t-statistics were used for adaptive selection of the regressors. Described results, for each setting of the parameters, are an average of 10 runs each on a different partition of training, validation and testing data. The validity of the models was measured in terms of normalized MSE (NMSE) and in terms of test proportion correct (TPC), defined as the proportion of data for which PF1 was correctly predicted with 10% tolerance, i.e. ± 0.1 in our case. In this study we used a Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = e^{-(\frac{\|\mathbf{x}-\mathbf{y}\|^2}{L})}$, with the L values on which individual methods achieved the best results in our former study [7]. In Figure 2 we depicted an example of cic(β) as a function of model order β. In Table 1 we summarize the results achieved on two subjects A (592 ERPs) and B (776 ERPs). We can see that all methods achieved similar prediction results. The CIC criterion picked up the models which on average uses only 66% of the overall regressors which is a significant reduction. Moreover, the number of selected principal components was less than 90% in comparison to Kernel PCR with CV criterion employed.

Method	NMSE		TPC		# of regressors	
	A	B	A	B	A	B
KPCR + CIC	**0.119**	**0.182**	**90.6**	**84.7**	**200.4**	**258.5**
	(0.03)	(0.03)	(0.02)	(0.02)	(19.7)	(16.9)
KPCR + CV	**0.118**	**0.175**	**90.3**	**84.6**	**225.3**	**289.6**
	(0.03)	(0.02)	(0.02)	(0.02)	(32.0)	(54.5)
Kernel RR	**0.117**	**0.173**	**91.2**	**84.8**	–	–
	(0.03)	(0.02)	(0.02)	(0.02)	–	–

Table 1: The comparison of the NMSE and TPC prediction errors for subjects A and B. The values represent an average of 10 simulations and corresponding standard deviation is presented in parentheses. The last column represents a model order selected by CIC and CV, respectively.

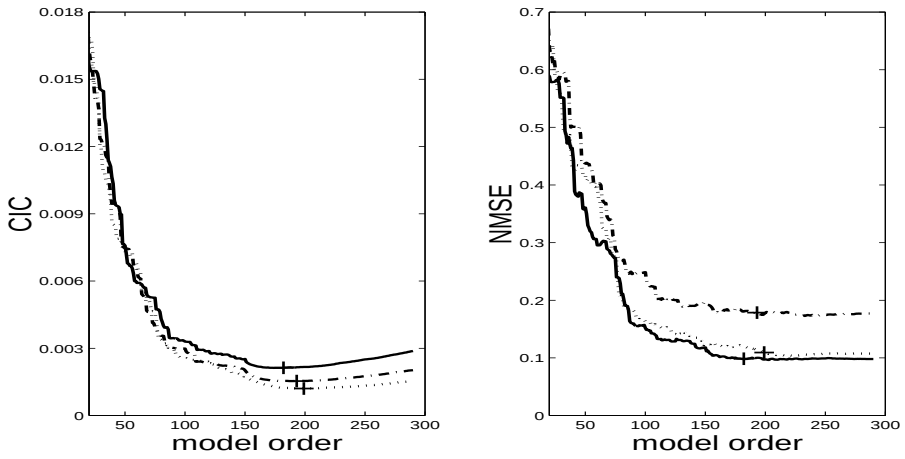

Figure 2: Three examples of CIC (left) and NMSE (right) as the function of model order (number of selected regressors) achieved on subject A. On the rest of runs and on subject B the similar behavior was observed. The '+' mark indicates selected model (min. of CIC).

5 Conclusions

On two benchmark and real world data sets we have demonstrated the comparable performance of Kernel PCR with CIC for model selection in comparison to Kernel RR. The connection between both methods was given. The computational cost of Kernel PCR is comparable with Kernel PCA. In fact independence of the regressors allows us to compute the estimates $\hat{w}_i$ of (6) adaptively and this is a significant time reduction during model selection.

CIC is based on estimation of *in-sample error*, i.e no validation set is needed. It was pointed out in [15] that CIC is a good measure for comparing models but may be less appropriate than methods based on *extra-sample error*, e.g. cross-validation, when good performance of the model on unseen test data is required. However, the selection of the representative validation data set is needed. In our study on two benchmark data sets the similar behavior of cross-validation and CIC was observed. On two subjects selected from the data set reflecting the problem of estimating human signal detection performance from the Event Related Potentials we observed that the Kernel PCR method with CIC resulted in similar performance but with a smaller number of principal components selected. In practical situations splitting

of the available data set into training and validation sets leads not only to less accurate estimates of the components but also has the potential to the decrease their number when $n \ll M$. Thus, the possibility to select "correct" principal componets by *in-sample* model selection procedures may be fruitful here.

Acknowledgements

ERPs data were obtained under a grant from the US Navy Office of Naval Research (PE60115N), monitored by Joel Davis and Harold Hawkins. The first author is funded by a research grant for the project "Objective Measures of Depth of Anaesthesia"; University of Paisley and Glasgow Western Infirmary NHS trust, and is partially supported by Slovak Grant Agency for Science (grants No. 2/5088/00 and No. 00/5305/468).

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[3] T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

[4] F. Girosi. An equivalence between sparse approximation and Support Vector Machines. *Neural Computation*, 10(6):1455–1480, 1998.

[5] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. Technical Report A.I. Memo No. 1430, MIT, 1993.

[6] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[7] R. Rosipal, M. Girolami, and J.L. Trejo. Kernel PCA for Feature Extraction of Event-Related Potentials for Human Signal Detection Performance. In *Proceedings of ANNIMAB-1 Conference*, pages 321–326, Götegorg, Sweden, 2000. Springer.

[8] R. Rosipal, M. Girolami, L. J. Trejo, and A. Cichocki. Kernel PCA for Feature Extraction and De-Noising in Non-Linear Regression. to appear *Neural Computing & Applications* , 2001.

[9] R. Rosipal, L. J. Trejo, and A.Cichocki. Kernel Principal Component Regression with EM Approach to Nonlinear Principal Components Extraction. Technical report, University of Paisley, CIS, 2000.

[10] R. Rosipal and L.J. Trejo. Kernel Partial Least Squares Regression in RKHS. Technical report, University of Paisley, 2001.

[11] C. Saunders, A. Gammerman, and V. Vovk. Ridge Regression Learning Algorithm in Dual Variables. In *Proc. of the 15th International Conference on Machine Learning*, 1998.

[12] B. Schölkopf, S.Mika, C.J.C.Burges, P.Knirsch, K.R.Müller, G.Rätsch, and A.J.Smola. Input Space vs. Feature Space in Kernel-Based Methods. *IEEE Transactions on Neural Networks* , 10(5):1000–1017, 1999.

[13] B. Schölkopf, A. J. Smola, and K. R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

[14] A.J. Smola, B. Schölkopf, and K. R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[15] R. Tibshirani and K. Knight. The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. B*, 61(3):529–546, 1999.

[16] L. J. Trejo, A. F. Kramer, and J. A. Arnold. Event-related Potentials as Indices of Display-monitoring Performance. *Biological Psychology*, 40:33–71, 1995.

[17] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1998.

[18] G. Wahba. *Splines Models of Observational Data*, volume 59 of *Series in Applied Mathematics.* SIAM, Philadelphia, 1990.

[19] C.K.I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M.I. Jordan, editor, *Learning and Inference in Graphical Models.* Kluwer, 1998.