Kernel PLS Smoothing for Nonparametric Regression Curve Fitting: an Application to Event Related Potentials

Roman Rosipal[1,2], **Leonard J Trejo**[1], **Kevin Wheeler**[1]
[1]NASA Ames Research Center
Computational Sciences Division
Moffett Field, CA 94035
[2]Department of Theoretical Methods
Slovak Academy of Sciences
Bratislava 842 19, Slovak Republic

August, 2003

Abstract

We present a novel smoothing approach to nonparametric regression curve fitting. This is based on kernel partial least squares (PLS) regression in reproducing kernel Hilbert space. It is our interest to apply the methodology for smoothing experimental data, such as brain event related potentials, where some level of knowledge about areas of different degrees of smoothness, local inhomogeneities or points where the desired function changes its curvature is known or can be derived based on the observed noisy data. With this aim we propose locally-based kernel PLS regression and locally-based smoothing splines methodologies incorporating this knowledge. We illustrate the usefulness of kernel PLS and locally-based kernel PLS smoothing by comparing the methods with smoothing splines, locally-based smoothing splines and wavelet shrinkage techniques on two generated data sets. In terms of higher accuracy of the recovered signal of interest from its noisy observation we demonstrate comparable or better performance of the locally-based kernel PLS method in comparison to other methods on both data sets.

1 Introduction

The problem of smoothing, de-noising, or estimating signals has produced a wide range of methods, such as wavelet de-noising, signal averaging, or smoothing splines (Chui, 1992; Donoho & Johnstone, 1995; Başar, 1980; Wahba, 1990). One area in which noise continues to be a problem is estimation of brain event related potentials (ERPs). This is an important problem because noise currently limits the utility of ERPs for understanding brain-behavior relationships. Here we consider a new method of de-noising which may offer improved estimation of ERPs, using nonlinear regression, such as kernel partial least squares (PLS).

There has been significant advancement in developing nonparametric regression techniques during the last several decades with the aim of smoothing observed data corrupted by some level of noise. A subset of these techniques is based on defining an appropriate dictionary of basis functions from which the final regression model is constructed. The model is usually defined to be a linear combination of functions selected from the dictionary. The widely used methods like smoothing splines and wavelet shrinkage belong to this category. These smoothing techniques have also been successfully applied to problems of signal de-noising which involves a wide area of research in the signal processing community.

In this setting it is usually assumed that the signal of interest is a linear combination of the

selected basis functions $\psi_i(x) \in \mathcal{D}$

$$g(x) = \sum_{i=1}^{p} w_i \psi_i(x)$$

where $\mathcal{D}$ represents a dictionary (family) of functions and $\{w_i\}_{i=1}^{p}$ are weighting coefficients. The main problem associated with this approach is the appropriate definition of $\mathcal{D}$ and the selection of a subset of basis functions used for the final model. Using a fixed dictionary of several basis functions, for example, all polynomials up to a pre-defined order or several trigonometric functions, may provide an easier selection among basis functions, but in general may not guarantee the possibility to closely approximate the desired signal of interest. On the other side defining our solution in a "rich" functional space may guarantee exact functional approximation of the signal of interest, however in a noisy scenario we may have a bigger problem of finding an adequate final estimate of the signal of interest. Smoothing splines and closely related support vector machines (SVM) are examples of this second approach (Wahba, 1990; Vapnik, 1998; Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002). The solution is defined to lay in a rich functional space and as such it can be expressed in the basis of the space (possibly infinite). However to avoid exact fitting of a measured noisy signal we need to incorporate some a priori assumptions about the smoothness of the desired signal of interest which is usually achieved through different forms of regularization. The appropriate functional space and regularization form selection for different types of noise distribution and types of signals are the main issues associated with these methods.

In this paper we propose a novel approach which tries to combine both of these strategies. We consider our solution to be in a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Saitoh, 1988). A straightforward connection between a RKHS and the corresponding feature space representation allows us to define the desired solution in a form of penalized linear regression model. More specifically, we consider a kernel PLS regression in a feature space $\mathcal{F}$ (Rosipal & Trejo, 2001). The basis functions $\psi_i(x)$ are taken to be score vectors obtain by kernel PLS, which may be seen as the estimates of an orthogonal basis in $\mathcal{F}$ defined by the measured signal and the kernel function used. These estimates are sequentially obtained using the existing correlations between nonlinearly mapped input data into $\mathcal{F}$ and the measured noisy signal, that is, extracted basis functions closely reflect existing input-output dependencies. Thus, the construction of the basis functions is given by the measured signal itself and is not a priori pre-defined without respect to the measured signal. The second stage of the proposed approach reduces to the problem of setting the number of basis functions p. The sequential structure of the extracted basis functions with respect to increasing description of the overall variance of the measured signal motivates our use of the Vapnik-Chervonnekis (VC) based model selection criterion (Cherkassky et al., 1999; Cherkassky & Shao, 2001). A low computation cost of this in-sample model selection criterion makes it a good candidate for the considered task of smoothing observed noisy signals.

Next, we extend the methodology of kernel PLS smoothing by assuming a set of locally-based kernel PLS models. The concept of locally weighted regression was originally proposed by Cleveland (1979) and its extension to linear PLS was discussed in Centner and Massart (1998). Our concept of locally weighted kernel PLS regression differs from these previous approaches. We try to model nonlinear relations between original input and output data by a proper selection of a kernel function and a considered level of regularization. However, in practice, it might by difficult to find such a kernel PLS regression model that guarantee a perfect fit over all parts of the signal. This may occur especially in the case where the desired signal of interest consists of several different parts where degree of smoothness may not be stationary. Thus, with the aim to construct a regression model with a higher flexibility to fit the desired signal of interest we consider a locally weighted linear PLS model in $\mathcal{F}$. Note that term weighted regression is sometimes also associated with a robust regression estimate where the a set of weights is determined by the residuals between predicted and observed output values (Cleveland, 1979; Silverman, 1985). In this paper we consider a noise element of the observed signal to be homogeneous and stationary over different parts of the signal and our concept of weighting is based on spatial distances of the neighboring input points only. To stress this fact, in the next, we use the term locally-based regression instead of locally weighted regression.

It is our intention to construct an appropriate weighting scheme based on a priori information about an approximate location of points of change of the signal curvature, areas of different degrees of smoothness or other local inhomogeneities occurring in the signal. In this paper we focus on a problem of smoothing event-related potentials signals corrupted by high levels of noise where this kind of information is known from many psychological observations. The spatial localization of individual kernel PLS models is achieved by incorporating weight functions reflecting the local areas of interest. Depending on weight function selection this allows us to construct soft or hard thresholding regions where kernel PLS regression models are constructed. Final regression estimate consists of the weighted summation of individual local kernel PLS regression models. Finally, the same form of information is included into smoothing splines and locally-based smoothing spline model is proposed.

We compare our methodology of kernel PLS and locally-based kernel PLS smoothing with the wavelet based signal de-noising, smoothing splines and locally-based smoothing splines on heavisine function and generated human ERPs distributed over individual scalp areas. We investigate the situations with different levels of additive uncorrelated or spatio-temporal correlated noise added to these generated signals. The wavelet shrinkage method is inherently designed to deal with local inhomogeneities in the signal, and may serve as a reference for detecting inhomogeneities using other methods. We use heavisine function as an illustrative example on which the used methods are also compared in terms of the recovery of local inhomogeneity.

2 Methods

2.1 Basic Definitions

Consider the regression model

$$y_i = g(x_i) + \epsilon_i \tag{1}$$

where $\{y_i\}_{i=1}^n$ represent observations at equidistant design points $\{x_i\}_{i=1}^n$, $a < x_1 < x_2 < \ldots < x_n < b$ in $[a, b]$ and $\{\epsilon_i\}_{i=1}^n$ are errors not restricted to be uncorrelated or to be drawn from a pre-specified probability distribution. In this paper we consider nonparametric estimation of the function $g(.)$. We assume that $g(.)$ is a *smooth* function in a functional space $\mathcal{H}$. To restrict our estimate of $g(.)$ in $\mathcal{H}$ to be a function with the desired smoothness property we consider an estimate $\hat{g}$ to be obtained as

$$\hat{g}(.) = \arg\min_{g \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \xi \Omega(g) \right] \tag{2}$$

In this formulation ξ is a positive number (regularization coefficient, term) to control the trade-off between approximating properties and the smoothness of $g(.)$ imposed by the penalty functional $\Omega(g)$. Further we will assume that $\mathcal{H}$ is a RKHS which provides a finite dimensional solution of (2) in spite of the fact that (2) is defined over an infinite-dimensional space. Kimeldorf and Wahba (1971) have shown that the solution of (2) leads to a general finite dimensional form known as the *Representer theorem*:

$$\hat{g}(x) = \sum_{i=1}^{n} d_i K(x_i, x) + \sum_{j=1}^{l} e_j \upsilon_j(x) \tag{3}$$

where the functions $\{\upsilon_j(.)\}_{j=1}^l$ span the null space of $\mathcal{H}$ and the coefficients $\{d_i\}_{i=1}^n$, $\{e_j\}_{j=1}^l$ are given by the data. $K(x, y)$ is a positive definite kernel function; that is, a symmetric function of two variables satisfying the Mercer theorem conditions (Mercer, 1909; Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002).[1] Using the eigen-expansion any kernel function can be written

[1] We consider one-dimensional input space, however, the following theoretical results are equally valid for a higher dimensional scenario.

as

$$K(x,y) = \sum_{i=1}^{S} \lambda_i \phi_i(x)\phi_i(y) = \langle \Phi(x), \Phi(y) \rangle = \Phi(x)^T \Phi(y) \tag{4}$$

where $\{\Phi(.)\}_{i=1}^{S}$ and $\{\lambda_i\}_{i=1}^{S}$ are corresponding eigenfunctions and eigenvalues, respectively. It becomes clear that any kernel $K(x,y)$ also corresponds to a canonical (Euclidean) dot product in a possibly high-dimensional space $\mathcal{F}$ where the input data are mapped by

$$\begin{aligned} \Phi: \quad & \mathcal{X} \rightarrow \mathcal{F} \\ & x \rightarrow (\sqrt{\lambda_1}\phi_1(x), \sqrt{\lambda_2}\phi_2(x), \ldots, \sqrt{\lambda_S}\phi_S(x)) \end{aligned}$$

The space $\mathcal{F}$ is usually denoted as a *feature space* and $\{\{\sqrt{\lambda_i}\phi_i(x)\}_{i=1}^{S}, x \in \mathcal{X}\}$ as *feature mappings*.

2.2 Kernel Partial Least Squares Regression

Assume a nonlinear transformation of $x \in \mathcal{R}$ into a feature space $\mathcal{F}$. Using the straightforward connection between a RKHS and $\mathcal{F}$ Rosipal and Trejo (2001) have extended the linear PLS regression model into its nonlinear kernel form. Effectively this extension represents the construction of linear PLS model in $\mathcal{F}$. Denote $\mathbf{\Phi}$ the $(n \times S)$ matrix of zero mean mapped input data $\Phi(x)$ into an S-dimensional feature space $\mathcal{F}$ and denote $\mathbf{Y}$ the $(n \times M)$ matrix of zero mean outputs. The kernel PLS method finds weight vectors $\mathbf{a}, \mathbf{b}$ such that

$$max_{|\tilde{\mathbf{a}}|=|\tilde{\mathbf{b}}|=1}[cov(\mathbf{\Phi}\tilde{\mathbf{a}}, \mathbf{Y}\tilde{\mathbf{b}})]^2 = [cov(\mathbf{\Phi}\mathbf{a}, \mathbf{Y}\mathbf{b})]^2 = [cov(\mathbf{t}, \mathbf{u})]^2$$

where $cov(\mathbf{t}, \mathbf{u}) = \mathbf{t}^T\mathbf{u}/n$ denotes the sample covariance between the $\mathcal{F}$-space score vectors (components, latent variables) $\mathbf{t}$ and Y-space score vectors $\mathbf{u}$. It can be shown that we can estimate the score vector $\mathbf{t}$ as first eigenvector of the following eigenvalue problem (Höskuldsson, 1988; Rosipal, 2003)

$$\mathbf{K}\mathbf{Y}\mathbf{Y}^T\mathbf{t} = \lambda\mathbf{t} \tag{5}$$

where $\mathbf{K}$ represents the $(n \times n)$ *Gram matrix* of the cross dot products between all input data points $\{\Phi(x)\}_{i=1}^{n}$, that is, $\mathrm{K}_{ij} = K(x_i, x_j)$ where $K(.,.)$ is a selected kernel function. The Y-scores $\mathbf{u}$ are then estimated as

$$\mathbf{u} = \mathbf{Y}\mathbf{Y}^T\mathbf{t} \tag{6}$$

After the extraction of new scores vectors $\mathbf{t}, \mathbf{u}$ the matrices $\mathbf{K}$ and $\mathbf{Y}$ are deflated. The deflation of these matrices takes the form

$$\mathbf{K} \leftarrow (\mathbf{I}_n - \mathbf{t}\mathbf{t}^T)\mathbf{K}(\mathbf{I}_n - \mathbf{t}\mathbf{t}^T) \quad , \quad \mathbf{Y} \leftarrow (\mathbf{I}_n - \mathbf{t}\mathbf{t}^T)\mathbf{Y} \tag{7}$$

where $\mathbf{I}_n$ is an n-dimensional identity matrix.

Finally, taking into account normalized scores $\mathbf{t}$ we define the estimate of the PLS regression model in $\mathcal{F}$ as

$$\hat{\mathbf{Y}} = \mathbf{K}\mathbf{U}(\mathbf{T}^T\mathbf{K}\mathbf{U})^{-1}\mathbf{T}^T\mathbf{Y} = \mathbf{T}\mathbf{T}^T\mathbf{Y} \tag{8}$$

Denote $\mathbf{d}^m = \mathbf{U}(\mathbf{T}^T\mathbf{K}\mathbf{U})^{-1}\mathbf{T}^T\mathbf{y}^m$, $m = 1, \ldots, M$ where the $(n \times 1)$ vector $\mathbf{y}^m$ represents the mth output variable. Then we can rewrite the solution of the kernel PLS regression (8) for the mth output variable as

$$\hat{g}^m(x, \mathbf{d}^m) = \sum_{i=1}^{n} d_i^m K(x, x_i)$$

which agrees with the solution of the regularized formulation of regression (2) given by the Representer theorem (3). Using equation (8) we may also interpret the kernel PLS model as a linear regression model of the form

$$\hat{g}^m(x, \mathbf{c}^m) = c_1^m t_1(x) + c_2^m t_2(x) + \ldots + c_p^m t_p(x) = \sum_{i=1}^{p} c_i^m t_i(x) \tag{9}$$

where $\{t_i(x)\}_{i=1}^{p}$ are p score vectors evaluated at the data point x and $\mathbf{c}^m = \mathbf{T}^T\mathbf{y}^m$ is the vector of weights for the mth regression model.

At the beginning of the section we assumed a zero mean regression model. To center the mapped data in a feature space $\mathcal{F}$ we can simply apply the following procedure (Schölkopf et al., 1998; Rosipal & Trejo, 2001)

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

where $\mathbf{1}_n$ represent a $(n \times 1)$ vector with elements equal to one.

It is worth noting that the score vectors $\{\mathbf{t}_i\}_{i=1}^{p}$ may be represented as functions of the original input data x. Then, the proposed kernel PLS regression technique can be seen as a method of sequential construction of a basis of orthogonal functions $\{t_i(x)\}_{i=1}^{p}$ which are evaluated at the discretized locations $\{x_i\}_{i=1}^{n}$. It is also important to note that the score vectors are extracted such that they increasingly describe overall variance in the input data space and more interestingly also describe the overall variance of the observed output data samples. In Fig.1 (top plot) we demonstrate this fact on example taken from Wahba (1990). We compute the function $g(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$ at $n = 101$ equally spaced data points x_i in the interval $[0, 3.25]$ and add independently, identically distributed noise samples ϵ_i generated according to $\mathcal{N}(0, 0.04)$.

This example suggest that to more precisely model the first negative part of the generated $g(.)$ function we need to use score vectors which generally reflect higher frequency parts of the investigated signal (see also Fig. 2). On contrary this may potentially lead to lowering the accuracy of our estimate over "smoother" parts of the signal. In our case second positive part of the $g(.)$ function. However, in many practical situations we may observe data segments where the observed function changes dramatically its curvature or even more we may have a priori information about the approximate locations of these segments. In the next subsection we describe the construction of locally-based kernel PLS regression models which incorporate this information.

2.2.1 Locally-Based Kernel Partial Least Squares Regression

To start this section we first demonstrate how the locally-based approach to kernel PLS described below may improve our estimate on the previous example. Consider that to model the first part of the generated function $g(.)$ we use the data in the range $[0, 1.5]$ and remaining data will be used to model second part of the function. To avoid discontinuities on the common edge of these two models we will introduce a form of soft clustering. This simply means that to construct both models we use all available data points, however, we introduce different weighting of individual data points. We consider weighting functions to be uniform over the individual parts of the signal and to be exponentially decreasing at the edges (the used weighting functions are depicted at the top of the bottom plot in Fig. 1). From Fig. 1 (bottom plot) we may observe a smoother estimate over the second part of the generated g(.) function using this locally-based PLS approach in comparison to the results obtained with the global kernel PLS regression. Both estimates over the first part of the function provides comparable results. However, in the case of locally-based kernel PLS regression we used only four score vectors extracted in each locally-based kernel PLS model in comparison to eight score vectors (selected based on minimum mean square error on clean signal) used in the global kernel PLS model. The overall mean squared error in the case of locally-based kernel PLS regression decreased by a factor of two, as compared to the global kernel PLS regression model.

Now, we provide a more rigorous description of the method. First, we consider the soft or hard clustering of the input data and their associated outputs. We introduce a weighting function $r(x)$ which reflects importance of the point x in a kernel PLS model. The points having very small or zero values of the function $r(.)$ will effectively be excluded from the construction of orthogonal PLS basis (PLS score vectors extraction) for a regression model and vice-versa. The weighting functions are defined is such a way that the overall kernel PLS model is decomposed into several

local kernel PLS sub-models.[2] The final model is then based on a composition of the individual locally-based kernel PLS models.

Let the $(n \times 1)$ vector $\mathbf{r}$ represent the values of the weighting function $r(.)$ at the training data points $\{x_i\}_{i=1}^n$. The weighted centering of $\mathbf{\Phi}$ given by $\mathbf{r}$ is

$$\mathbf{\Phi}_r = \mathbf{R}_d(\mathbf{\Phi} - \mathbf{1}_n \frac{\mathbf{r}^T\mathbf{\Phi}}{r_s}) = (\mathbf{R}_d - \frac{\mathbf{r}\mathbf{r}^T}{r_s})\mathbf{\Phi}$$

where $r_s = \sum_{i=1}^n r_i$ and $\mathbf{R}_d$ is the $(n \times n)$ diagonal matrix with elements on diagonal equal to r_i. Consequently, the centered Gram matrix will have the form

$$\mathbf{K}_r = \mathbf{\Phi}_r\mathbf{\Phi}_r^T = (\mathbf{R}_d - \frac{\mathbf{r}\mathbf{r}^T}{r_s})\mathbf{K}(\mathbf{R}_d - \frac{\mathbf{r}\mathbf{r}^T}{r_s})$$

Similarly, we do weighted centering of the output data

$$\mathbf{Y}_r = \mathbf{R}_d(\mathbf{Y} - \mathbf{1}_n \frac{\mathbf{r}^T\mathbf{Y}}{r_s}) = (\mathbf{R}_d - \frac{\mathbf{r}\mathbf{r}^T}{r_s})\mathbf{Y}$$

Consider that we define Z clusters based on which Z locally-based kernel PLS models are constructed. Define centered Gram matrix $\mathbf{K}_r^z$ and the outputs matrix $\mathbf{Y}_r^z$ constructed using the weight function $r^z(.)$ associated with the zth cluster. Using (5) and (6) where $\mathbf{K}$ is replaced by $\mathbf{K}_r^z$ and $\mathbf{Y}$ is replaced by $\mathbf{Y}_r^z$, respectively, we obtain the scores $\mathbf{t}^z, \mathbf{u}^z$ of the zth kernel PLS model. After each step we deflate $\mathbf{K}_r^z$ and $\mathbf{Y}_r^z$ matrices in the same way as described in the previous section (eqs. (7)).

Denoting by $\mathbf{T}^z$ and $\mathbf{U}^z$ the matrices with columns consisting from the extracted $\mathbf{t}^z, \mathbf{u}^z$ scores the kernel PLS regression estimate for the zth cluster is given as

$$\hat{\mathbf{Y}}_r^z = \mathbf{T}^z(\mathbf{T}^z)^T\mathbf{Y}_r^z$$

To express this estimate in the original not centered variables we can write

$$\hat{\mathbf{Y}}^z = \mathbf{R}_d^{-1}\hat{\mathbf{Y}}_r^z + \mathbf{1}_n \frac{(\mathbf{r}^z)^T\mathbf{Y}}{r_s^z}$$

where r_s^z is the sum of the elements of the weighting vector $\mathbf{r}^z$ defined for the zth cluster by the weight function $r^z(.)$. To be consistent with our previous notation we denote by

$$\hat{g}^m(x_i)^z \stackrel{\text{def}}{=} (\hat{y}_i^z)^m \ , \ i = 1, \ldots, n \ ; \ m = 1, \ldots, M \ ; \ z = 1, \ldots, Z$$

the locally-based kernel PLS estimate for the zth cluster for the $m$the output variable at the data point x_i. The final locally-based kernel PLS regression model consists of the weighted summation of Z individual local kernel PLS regression estimates. This estimate for the input point x_i is given as

$$\hat{g}^m(x_i) = \sum_{z=1}^{Z} r_i^z \hat{g}^m(x_i)^z / \sum_{z=1}^{Z} r_i^z \ , \ i = 1, \ldots, n \ ; \ m = 1, \ldots, M$$

where $\{r_i^z\}_{i=1}^n$ are the elements of the weighting vector $\mathbf{r}^z$.

Finally, let us make several general comments on the locally-based and also global kernel PLS methodology:

a) First we have to stress that we defined the weighting function $r(.)$ based on existing general knowledge about the signal of interest, visual inspection of the noisy data, detection of the segments with significant change of the curvature, degrees of smoothness, etc. The obvious question which may occur is how the segmentation (clustering) of the input data will be "transformed" into the

[2]This is analogous to the strategy used to construct the mixture of probabilistic principal component analyzers (Tipping & Bishop, 1999) where the function $r(.)$ represents a posterior *responsibilities* of individual data points x.

clustering of the data in a feature space $\mathcal{F}$. This is an important issue due to the fact that we consider local PLS in $\mathcal{F}$, not in the original input space. We may believe that we can invoke good clustering in a feature space $\mathcal{F}$ if the nonlinear mapping to $\mathcal{F}$ will be smooth and will preserve topological relations of the original input data. In all our experiments with the kernel PLS model or its locally-based version we have used the Gaussian kernel function $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{l}\right)$ where l is the parameter controlling the width of the Gaussian function. By using a wider Gaussian kernel function we can not successfully localize very small segments of input data. Data corresponding to small distances in input space will be "flocked" together in $\mathcal{F}$. In opposite for a smaller width l of the Gaussian kernel function the input data with greater distances will be too "spread" in $\mathcal{F}$ and intended localization may be lost. The inspection of the graphs reflecting dependence of feature space distances on input space distance of the equidistant points generated in the interval $[-1, 1]$ suggest to use l parameter in the range of $[0.1, 1.4]$. The theoretical value of two times the variance of uniformly distributed data in $[-1, 1]$ equals $0.\bar{6}$ and in all our experiments with the Gaussian kernel function this motivates our choice of the width l to be equal 0.6.

b) As we might already observe the kernel PLS method iteratively extracts orthogonal score vectors describing overall variance of the input as well as output data. This effectively creates a sequence of orthogonal functions with increasing "complexity". Here we define the complexity in the sense that the first kernel PLS score vectors will pick up the trend of the output function and will represent rather smooth slowly-varying functions. In contrast higher score vectors will represent higher frequency components of the output signal or noise. We note that this hypothesis in general will depend on the selected kernel function. However, in our setting we have used the Gaussian kernel function which has a nice smoothing property based on the fact that higher frequency components are suppressed (Girosi et al., 1995; Schölkopf & Smola, 2002). We may hypothesize that the above argument will be true for such a class of "smooth" kernel functions. In Fig. 2 we demonstrate this argument using several score vectors extracted in previous example where we have used Gaussian kernel with the width l equal to 1.8.[3]

c) Described locally-based and global kernel PLS regression is well defined for univariate as well as multivariate outputs scenario. In the case of multivariate outputs this approach opens the possibility for spatio-temporal modeling of a collection of signals of interest. This can be useful in the case that observed signals represent different time realizations or measurements at different spatial locations. Or also in the situations where both, different time and spatial measurements of the same or similar signals of interest $g(.)$ are collected. We simply arrange the output matrix $\mathbf{Y}$ to be the matrix with columns representing these different temporal or spatial signal(s) of interest. The extracted score vectors will represent common features of these different realizations. Finally, we would like to note that the "kernel trick" (4) allows us to easily extend the methodology to the case of multidimensional inputs and also for the case of not equally sampled data.

d) As we noticed in the previous example the locally-based kernel PLS approach provided (in terms of MSE) a better estimate of the generated function $g(.)$. This was achieved with smaller number of four different score vectors used in individual local kernel PLS models. On several different data sets we experimentally observed that to properly smooth the noisy data over the segments with low curvature of $g(.)$ individual locally-based kernel PLS needed less than 8-10 score vectors. In many cases less than five score vectors provided best results. This was observed on different smoothing problems described in the paper and other non-published data sets as well. However, results of Rosipal and Trejo (2001) indicate that if the regression task with smaller level of output noise is of interest this number may increase. In the current paper we provide results where the maximum number of score vectors used in individual local kernel PLS models was restricted to be the first four score vectors. The final number $p \leq 4$ of score vectors was selected using the model selection criterion described in the next section.

[3]We just remind the reader that the kernel PCA decomposition without any connection to the output function would lead to the extraction of the principal components similar to a trigonometric series expansion for a Gaussian kernel (Schölkopf & Smola, 2002; Rosipal & Trejo, 2001).

2.2.2 Model Selection

We have already noticed that kernel PLS extracts score vectors with increasing complexity in the sense of remark b) in the previous section. This construction of a sequence of functions with increasing complexity also motivates our use of the idea of Vapnik's *structural risk minimization* approach for model selection, that is, in our case this is the determination of the number of score vectors p used in each locally-based kernel PLS regression model. Vapnik (1998) has shown that for regression problems with a squared loss function the following bound on an estimate of in-sample prediction error (PE) holds with probability $1-\eta$

$$PE \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}(x_i))^2 \left(1 - c\sqrt{\frac{h(\ln(\frac{an}{h})+1) - \ln\eta}{n}}\right)_+^{-1} \tag{10}$$

where h is VC dimension of the set of approximating functions, c is a constant reflecting the "tails of the loss function distribution", a is a theoretical constant and

$$(x)_+ = \begin{array}{ll} x & if \;\; x > 0 \\ 0 & otherwise \end{array}$$

The first term on right hand side of (10) represents *empirical error* while the second term is often called *penalization factor*, which for increasing complexity of the regression model inflates empirical error.

For practical use it is difficult to compute the exact VC dimension for an arbitrary set of functions, moreover, it can be infinite for some classes of functions. However, constructing a regression function to be a linear combination of a finite (fixed) set of basis functions, ordered based on the increasing complexity, Cherkassky et al. (1999) and Cherkassky and Shao (2001) suggested to take the following heuristic penalization factor

$$\left(1 - \sqrt{v - v\ln v + \frac{\ln n}{2n}}\right)_+^{-1}$$

where $v = p/n$ with p representing VC dimension of the considered regression function (9) with p terms. To complete this replacement Cherkassky et al. (1999) set $\eta = 1/\sqrt{n}$ and they have considered parameters a, c to be equal one.[4] In comparison with other model selection criteria, it was demonstrated that the new model selection criterion motivated by the structural risk minimization theory and VC dimension may provide comparable or better results (Cherkassky et al., 1999; Cherkassky & Shao, 2001).

2.3 Univariate Smoothing Splines

We again consider model (1) and we further assume that the function $g(.)$ belongs to the Sobolev Hilbert space

$$\mathcal{W}_2^2[0,1] = \{f : f, f' \text{ absolutely continuous on } [0,1]; \int_0^1 (f''(x))^2 dx < \infty\}$$

which determines the smoothness properties of $g(.)$. Without lack of generality we consider input data to be in the interval $[0,1]$. Wahba (1990) has shown that the same general framework for the solution of regularized functional (2) in a RKHS can also be applied for smoothing splines. More precisely, the RKHS is a Sobolev Hilbert space and the penalty functional $\Omega(g)$ is the semi-norm of the space. In our case of $\mathcal{W}_2^2[0,1]$ the following regularized functional is considered

$$\hat{g}(x) = \arg\min_{g \in \mathcal{W}_2^2[0,1]} \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - g(x_i))^2 + \gamma \int_0^1 (f''(x))^2 dx\right] \tag{11}$$

[4]For more detail description of this reasoning see (Cherkassky et al., 1999; Cherkassky & Shao, 2001) .

which provides a unique solution known as a *natural cubic spline*. This solution has the following proprieties

$$\hat{g}(x) = \begin{cases} \pi^1 & \text{for } x \in [0, x_1] \\ \pi^3 & \text{for } x \in [x_j, x_{j+1}] \\ \pi^1 & \text{for } x \in [x_n, 1] \\ C^2 & x \in [0, 1] \end{cases}$$

where π^k are polynomials of degree k and C^2 represents functions of 2 continuous derivatives. The smoothness of the solution is controlled through the parameter $\gamma \geq 0$. Adapting theory of RKHS and the Representer theorem (3) we can write the solution of (11) in the form (Wahba, 1990)

$$\hat{g}(x) = e_1 \upsilon_1(x) + e_2 \upsilon_2(x) + \sum_{i=1}^{n} d_i K(x, x_i) \tag{12}$$

where $\upsilon_1(x) = 1,\ \upsilon_2(x) = x - 1/2$ and

$$\begin{aligned} K(x, x_i) = \ & \tfrac{1}{4}((x - \tfrac{1}{2})^2 - \tfrac{1}{12})((x_i - \tfrac{1}{2})^2 - \tfrac{1}{12}) \\ & -((|x - x_i| - \tfrac{1}{2})^4 - \tfrac{1}{2}(|x - x_i| - \tfrac{1}{2})^2 + \tfrac{7}{240})/24 \end{aligned} \tag{13}$$

To estimate the vectors of unknown coefficients $\mathbf{e} = (e_1, e_2)^T, \mathbf{d} = (d_1, d_2, \ldots, d_n)^T$ we need to solve the following optimization problem

$$\arg\min_{\mathbf{e},\mathbf{d}} \left[\frac{1}{n} \|\mathbf{y} - (\boldsymbol{\Upsilon}\mathbf{e} + \mathbf{K}\mathbf{d})\|^2 + \gamma \mathbf{d}^T \mathbf{K} \mathbf{d} \right] \tag{14}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T, \mathbf{K}$ is the Gram matrix with (i, j)th entry to be equal $K(x_i, x_j)$ of (13) and $\boldsymbol{\Upsilon}$ is the $(n \times 2)$ matrix with (i, j)th entry $\upsilon_j(x_i)$. It can can be further shown (Wahba, 1990; Green & Silverman, 1994) that we can express this solution in the form

$$\hat{\mathbf{g}} = \mathbf{S}_\gamma \mathbf{y}$$

where $\hat{\mathbf{g}}$ denotes $(n \times 1)$ vector of fitted values $\hat{g}(x_i)$ at the training input data points $\{x_i\}_{i=1}^n$ and $\mathbf{S}_\gamma$ is the *smoother matrix* which depends only on $\{x_i\}_{i=1}^n$ and γ.

In the case of independent, identically normally distributed errors ϵ_i the minimum of *generalized cross-validation* (GCV) function was proved to provide a good estimate for γ (Wahba, 1990)

$$\text{GCV}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} [\frac{y_i - \hat{g}(x_i)}{1 - trace(\mathbf{S}_\gamma)/n}]^2$$

Finally, we extend smoothing splines to the locally-based smoothing splines model by minimizing weighted form of (14). Using the previously defined diagonal weighting matrix $\mathbf{R}_d$ this is given by replacing the $\mathbf{K}$ and $\boldsymbol{\Upsilon}$ matrices in (14) by their weighted versions $\mathbf{R}_d\mathbf{K}\mathbf{R}_d$ and $\mathbf{R}_d\boldsymbol{\Upsilon}$.

3 Data Construction

3.1 Heavisine function

To examine the performance of kernel PLS and locally-based kernel PLS methods we selected a standard example of de-noising heavisine function taken from Donoho and Johnstone (1995)

$$g(x) = 4sin(4\pi x) - sign(x - 0.3) - sign(0.72 - x)$$

The function of a period one has two jumps at $x_1 = 0.3$ and $x_2 = 0.72$. The function scaled into interval $[-1, 1]$ is depicted in Fig. 7b. We computed heavisine function on five sets of equally spaced data points with number of samples 124, 256, 512, 1024 and 2048, respectively. We added Gaussian noise of two levels such that signal-to-noise ratio (SNR) was 1dB and 5dB, respectively.[5] We created 50 different replicates of the noisy heavisine function for each set of different lengths and noise levels.

[5] In all our experiments we define SNR to be $10 \log_{10} \frac{S}{N}$, where $S = \sum_{i=1}^{n} g(x_i)^2$ is a sum of squared amplitudes of clean signal and $N = \sum_{i=1}^{n} \epsilon_i^2$ is the sum of squared amplitudes of noise, respectively.

3.2 Event Related Potentials

We simulated brain event related potentials (ERPs) and ongoing electroencephalogram (EEG) activity using the dipole simulator program of the BESA software package.[6] In this scenario we consider EEG to represent spatially and temporally correlated noise added to the ERPs. This simulation provides a reasonable but simplified model of real-world ERPs measurements.

The sources of simulated ERPs were represented by five dipoles with different spatial locations and orientations. The placement of the dipoles with their sample activation function is presented in Fig. 3. The dipoles that contribute to the ERP produce a composite waveform with four prominent peaks: N1 (negative peak with the latencies in the range 100-140 ms), P2 (positive peak with the latencies in the range 200-240 ms), N2 (negative peak with the latencies in the range 250-300 ms) and P3 (positive peak with the latency in the range 340-400 ms) as observed on C_z electrode. These four peaks correspond to well known components of human auditory or visual ERP (Hillyard & Kutas, 1983; Naatanen & Picton, 1987; Naatanen & Picton, 1986; Parasuraman & Beatty, 1980; Picton et al., 1974). We generated 20 different realizations of ERPs on the scalp by randomly changing the latencies of peaks and amplitudes of the individual activation functions. These dipoles were used to generate a full scalp topography containing 19 data channels located using the International 10-20 System (Jasper, 1958). An average of the signal at the mastoid electrodes A1 and A2 was used as the reference signal for other electrodes. A sample of ERP for one of 20 different trials is also shown in Fig. 3. Individual data epochs were designed to be 800 ms long starting 100 ms before the event. We have used a sampling rate equal to 640Hz resulting in 512 data points per epoch. In the next step coherent noise modeling of ongoing EEG activity was added to the generated ERPs. The noise data waveforms are summed over the contributions from 200 different sources (dipoles) with random locations and orientations. The noise is generated such that there is a high correlation between signal amplitudes from close electrodes. The frequency characteristic reflects characteristics of EEG spectra ($1/\sqrt{f+1}$) with added dominant α-band frequency about 10Hz. The weighting of the α-band in comparison to the other frequencies was changed in each realization but the proportion of the α-band stays in the range of $40\% - 60\%$. Two different levels of the amplitudes of the noise signal waveforms were set. This created two sets of noisy ERPs with averaged SNR over the electrodes and trials to be equal 5dB and -1.5dB, respectively. The same sampling rate as described above was used. Finally, temporally and spatially uncorrelated Gaussian noise was added to the generated data to represent measurement noise and other non-biological sources. For the first set of the data the zero mean noise with standard deviation equal to 0.5μV was generated. In the second case standard deviation of the white Gaussian noise was increased to 1μV. The noise was generated for each channel individually and referenced to average of A_1 and A_2 electrodes. This resulted in final averaged SNR over electrodes and trials to be equal 1.3dB and -4.6dB, respectively.

4 Experiments

To evaluate our results we have used two measures of goodness of fit to the clean signal $g(.)$. First, the normalized root mean squares error (NRMSE) was defined as

$$\text{NRMSE} = \sqrt{\frac{\sum_{i=1}^{n}(g(x_i) - \hat{g}(x_i))^2}{\sum_{i=1}^{n}(g(x_i) - \bar{g}(x))^2}}\ ,\ \bar{g}(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

The second measure we have used is Spearman's rank correlation coefficient (SRC) between the de-noised signal $\hat{g}(.)$ and the original noise-free signal $g(.)$ (Heckman & Zamar, 2000). Although both measures can provide good intuition about the goodness of fit and similarity of the shapes between our estimate of the signal of interest and the signal itself, they reflect mainly the overall characteristics of the fit. Thus we have also visually inspected smoothness of the individual estimates $\hat{g}(.)$ and goodness of fit of inhomogeneities in the signal $g(.)$.

[6] `http://www.besa.de`

4.1 Heavisine function

Both the kernel PLS (KPLS) and locally-based kernel PLS (LKPLS) methods were compared with smoothing splines (SS), locally-based smoothing splines (LSS) and Wavelet Shrinkage (WS) methods.

In the case of KPLS and LKPLS equally spaced data in the interval $[-1, 1]$ were used. The theoretical value of two times the variance of uniformly distributed data in $[-1, 1]$ equals $0.\bar{6}$ and in all our experiments with the Gaussian kernel function this motivates our choice of the width l to be equal 0.6. Visual inspection of a sample of a noisy heavisine function (Fig. 7a) suggests to consider four segments (intervals) $\mathcal{S}_1$ =([-1,-0.5],[-0.5,0],[0,0.5],[0.5,1]) for LKPLS. These segments reflect four "bumps" visible on the noisy function over which individual local kernel PLS models are constructed. We set the weighting function $r(.)$ to be equal to 1 within the individual segments. The segments are overlapped with exponentially decaying values in a way that it reaches values close to zero in the middle of the adjacent segments. In the second case we assume that the approximate location of the local inhomogeneity close to the point 0.45 is known in advance. We rearrange the four segments to be $\mathcal{S}_2$ =([-1,-0.5],[-0.5,0.4],[0.4,0.5],[0.5,1]) and set the weighting function $r(.)$ to be equal to 1 within the individual segments. We used the same decay of $r(.)$ as in the first case to overlap the segments. The same segmentation and weighting functions rescaled into the interval $[0, 1]$ were used in the case of LSS.

We investigated different settings of the WS method as implemented in the Matlab Wavelet Toolbox.[7] Comparing the results in the terms of NRMSE and SRC on recovery of the original noise-free heavisine function we examined different families of wavelets, different numbers of wavelet decomposition levels and different threshold selection rules as implemented in the `wden` function of the Wavelet Toolbox. The best results were achieved and are reported using the Daubechies 4 family of wavelets, periodic extension at the edges, SURE as the threshold selection rule, threshold rescaling using a single estimation of level noise based on the first-level of coefficients and soft thresholding of detail coefficients at level $N - 1$, where N is maximal number of decomposition levels as defined by the number of data points used. These settings are in good agreements with findings of Donoho and Johnstone (1995).

First, the KPLS regression approach was compared with LKPLS using two different segmentations of the observed noisy signal. In the case of KPLS we observed that by fixing number of score vectors to be in the range $8 - 15$ we may achieve good performance of the method. We have also observed that the VC-based model selection criterion described in Section 2.2.2 has a tendency to underestimate a number of selected score vectors in the case of a large number of data points (512,1024,2048). Consequently KPLS regression resulted in lower accuracy than that found with a fixed number of score vectors. In the case of small number of data points (128,256) the VC-based model selection criterion provided comparable or better results than the case where a fixed number of score vectors was used. The number of score vectors in these cases was set to 12. In the case of LKPLS, limiting the number of scores vector to the first four and using the VC-based model selection criterion to select the final number provided good results. In Fig. 4 a boxplot with lines at the lower quartile, median, and upper quartile values and a whisker plot of the distribution of NRMSE over 50 different runs corresponding to the case of SNR=5dB is depicted (qualitatively similar results were achieved in the case of SNR=1dB). In terms of the median and upper and lower quartiles of NRMSE the LKPLS method outperformed KPLS for samples of 512,1024 and 2048 points in both SNR scenarios. In terms of SRC we observed qualitatively the same results as described for NRMSE. In the case of SNR=1dB the median of SRC for the KPLS and LKPLS methods was in the range between $0.952 - 0.997$ with smaller values starting for the case of smaller number of samples used. In the case of SNR=5dB the range of median values of SRC for KPLS and LKPLS was between $0.983 - 0.997$ corresponding to the high "agreement of shapes" between estimates $\hat{g}(.)$ and the original noise-free heavisine function $g(.)$. To statistically compare the best results achieved with individual methods the nonparametric sign test and the Wilcoxon matched-pairs test were used to test the hypothesis about the differences within the pairs of NRMSE and SRC (only the cases where significant difference in both NRMSE and SRC terms was observed are

[7] `http://www.mathworks.com/`

considered). The significance level for all tests was set to $\alpha = 0.01$. In the case of SNR=1dB and samples of 256,512, 1024 and 2048 points two-sided as well as one-side alternative of both tests provided statistically significant superiority of LKPLS-$\mathcal{S}_2$ over KPLS in both, NRMSE and SRC, terms. For the same sample sizes LKPLS-$\mathcal{S}_1$ resulted in smaller median NRMSE values in comparison to KPLS, however, no statistically significant differences were observed in this case. KPLS significantly outperformed LKPLS-$\mathcal{S}_1$ for sample size equal to 128. In the case of SNR=5dB the statistically significant superiority of LKPLS-$\mathcal{S}_2$ over KPLS was observed for all different sample size sets. Similar to the previous SNR=1dB case LKPLS-$\mathcal{S}_1$ in comparison to KPLS provided smaller median NRMSE but comparable median SRC values. Again, no statistically significant differences were observed between LKPLS-$\mathcal{S}_1$ and KPLS.

In the next step SS and LSS using two different segmentations were compared. Although SS using the GCV criterion resulted in good recovery of the noise-free heavisine function, this model selection criterion failed when the LSS models were considered. More precisely, we observed that GCV applied to individual local smoothing splines models tends to select small values of γ leading to strong overfitting. Therefore, we constrained the range of the smoothing parameter with respect to the observed test errors. The minimum value of γ considered for GCV was set to $1e^{-5}$ in the case of SNR=1dB and to $1e^{-6}$ in the case of SNR=5dB. In Fig. 5 a boxplot of NRMSE using these setting of γ is depicted. In contrast to LKPLS we can not see any obvious improvement of LSS in comparison to SS. The same observation was true for SRC. Qualitatively similar results (not shown) were obtained for SNR=1dB. In the case of SNR=1dB two-sided as well as one-side alternatives of both nonparametric tests did not show any statistically significant differences between SS and LSS-$\mathcal{S}_1$. The same was true for SS compared with LSS-$\mathcal{S}_2$ except sample size equal to 512 in which case LSS-$\mathcal{S}_2$ significantly outperformed SS. In the case of SNR=5dB the statistically significant superiority of SS over LSS-$\mathcal{S}_1$ was observed for samples of 128 and 256 points and for sample size equal to 128 in comparison to LSS-$\mathcal{S}_2$. In contrast, LSS-$\mathcal{S}_2$ outperformed SS for samples of 1024 and 2048 points. Finally, we note that the median of SRC for SS and LSS varied in the range $0.940 - 0.997$ for SNR=1dB and in the range $0.961 - 0.998$ for SNR=5dB, respectively, providing good agreement of the recovered curves with the original noise-free heavisine signal.

Next, the LKPLS-$\mathcal{S}_2$ and LSS-$\mathcal{S}_2$ approaches were compared with WS. The results in terms of NRMSE for the case of SNR=5dB are depicted in Fig. 6. In general we observed superiority of both locally-based approaches over WS. In the case of SNR=1dB two-sided as well as one-side alternative of both nonparametric tests showed statistically significant superiority of LKPLS-$\mathcal{S}_2$ over WS for all sample sizes except 128. LSS-$\mathcal{S}_2$ outperformed WS for all sample sizes. Tests also showed statistically significant superiority of LKPLS-$\mathcal{S}_2$ over LSS-$\mathcal{S}_2$ for sample size of 2048 points. In contrast, LSS-$\mathcal{S}_2$ outperformed LKPLS-$\mathcal{S}_2$ in the case of sample size equal to 128. In the case of SNR=5dB, LKPLS-$\mathcal{S}_2$ and LSS-$\mathcal{S}_2$ showed significant superiority over WS for all samples sizes except 2048. LKPLS-$\mathcal{S}_2$ preformed better than LSS-$\mathcal{S}_2$ for sample sizes 128,256 and 512 and no significant differences were observed between these two locally-based methods for sample sizes 1024 and 2048.

Using the LKPLS-$\mathcal{S}_2$, LSS-$\mathcal{S}_2$ and WS methods we visually inspected all 50 trials in the case of 2048 samples and both SNRs. In the case of SNR=5dB we could confirm the detection of local inhomogeneity close to the point 0.45 in almost all trials using the LKPLS-$\mathcal{S}_2$ (47 trials) and WS (49 trials) methods. In the case of LSS-$\mathcal{S}_2$ it was possible in 39 trials. In the case of SNR=1dB this was possible in 34 trials with LKPLS-$\mathcal{S}_2$, in 19 trials with LSS-$\mathcal{S}_2$ and in 20 trials with WS. The overall smoothness of LKPLS-$\mathcal{S}_2$ and LSS-$\mathcal{S}_2$ and their fit to the original noise-free heavisine signal visually appeared better then WS. In terms of the median NRMSE, the improvements of the LKPLS-$\mathcal{S}_2$ and LSS-$\mathcal{S}_2$ methods over WS were 16.6% and 12.4%, respectively.

In Fig. 7 we plotted examples of smoothed curves with the value of NRMSE closest to the median of NRMSE achieved by the given method over all 50 trials. We may see that using this form of comparison only LKPLS and LSS provide the results where we can visually detect the local inhomogeneity close to the point 0.45.

4.2 Event Related Potentials

In the case of LKPLS we again used the equally spaced data in the interval $[-1, 1]$ and the width l of the Gaussian kernel equal to 0.6. To set the weighting function $r(.)$ we created an average of the first five ERP trials on electrode P_z to visually set the segments over which the LKPLS and LSS regression models were constructed. We also took into account existing knowledge about the shape of real-world ERPs (Hillyard & Kutas, 1983; Naatanen & Picton, 1987; Naatanen & Picton, 1986; Parasuraman & Beatty, 1980; Picton et al., 1974) (see also Section 3.2). This motivates our choice of three segments ([-1,-0.3],[-0.3,0.1],[0.1,1]). We set $r(.)$ to be equal 1 over individual segments and then overlap the segments with exponentially decaying values of $r(.)$ reaching the values close to zero (less then $10e^{-5}$) on interval 0.4. In the case of LSS the segments and the weighting functions were rescaled into the interval $[0, 1]$.

In the case of WS we have used thresholds rescaled by a level-dependent estimation of the noise level (Johnstone & Silverman, 1997). This provided better results in this case due to the temporally and spatially correlated noise that we added to ERPs. Other have found that in the case of colored noise GCV or CV criteria fail to provide an appropriate estimate of the smoothing parameter γ in the SS approach (Diggle & Hutchinson, 1989; Wang, 1998; Opsomer et al., 2001). Although there exist several modifications of GCV for the case of colored noise usually some a priori knowledge about the covariance matrix of the correlated noise or its parametric specifications is needed (Diggle & Hutchinson, 1989; Wang, 1998; Opsomer et al., 2001). Thus, similar to the previous case we have constrained the range of the smoothing parameter γ for SS and LSS with respect to the observed test errors.

Signal averaging was also used to extract the ERPs embedded in noise (Başar, 1980). This approach has been proven to be quite useful due to the overlap of the ERP and noise spectra. Although the assumption of stationary, time-invariant ERPs justify this method, it may provide a useful smooth estimate of ERPs when there exists slight variability among individual ERP realizations. However, in this case the information about the variation in amplitudes and latencies over individual trials will be smeared out. The estimate of the ERP at each electrode was constructed to be the average of 20 different realizations of the ERP measured at the electrode. This was taken to represent the same de-noised signal for all trials and NRMSE and SRC between this estimate and individual clean ERP realizations was then computed. We have to stress that in contrast to the averaging approach the results of other smoothing methods described are single-trial oriented.

In the case of KPLS and LKPLS we also compared a univariate approach in which each electrode represents single KPLS or LKPLS regression model with the approaches where spatial, temporal and spatio-temporal setting of the multivariate outputs was used. In the spatial setting individual columns are constructed using measurements at different electrodes while in temporal they are represented by different time measurements at the same electrode. In spatio-temporal setting we combine both approaches. We have to stress that this spatio-temporal information is only used to extract common PLS score vectors while regression models are consequently built for each electrode individually. This allows us to obtained several different estimates for a particular electrode by using the KPLS or LKPLS models with different settings of the output matrix. Finally we may create a final estimate at the electrode by combining these individual estimates. We investigated different settings of local spatial distribution of the electrodes as well as more global settings where the measurements from spatially more distributed electrodes created the final multidimensional output matrix $\mathbf{Y}$. The similar modification of short-term and long-term temporal setting was investigated and mutually combined. We observed small differences among the results achieved over individual trials, however, in terms of averaged NRMSE and SRC over all trials the differences were small. Next we report the results where overall spatial information from all electrodes was used, that is, columns of the output matrix are measurements at individual electrodes over one trial. A number of selected score vectors for the regression models corresponding to individual electrodes was set based on the minimum of the VC-based model selection criterion for the particular model. We set the maximum number of score vectors equal to 12 in the case of KPLS and to four in the case of LKPLS.

The median values of NRMSE computed over 20 different trials on each of 19 electrodes for

two levels of averaged SNR are plotted in Fig. 8 and 9. From these plots we may observe that, in terms of the median NRMSE and SRC, LKPLS provides better results in comparison to KPLS and WS on almost all electrodes. We used two-sided as well as one-sided alternatives of both nonparametric tests to test differences between matched-pairs of the NRMSE and SRC values. The significance level for all tests was set to $\alpha = 0.01$. We observed statistically significant superiority of LKPLS over KPLS on four electrodes out of 19 in the case of SNR=1.3dB. In the case of SNR=-4.6dB we observed this superiority of LKPLS over KPLS on 6 electrodes. In comparison to WS the LKPLS method was significantly better on 11 electrodes for SNR=1.3dB and 16 electrodes for SNR=-4.6dB, respectively. Although the results using LKPLS are worse than applying the signal averaging technique we may also observe that the same technique of averaging applied to smoothed, de-noised estimates of LKPLS outperformed averaging applied to the raw ERPs. We may observe that for a smaller level of noise the median of SRC for averaged LKPLS estimates is higher then the median of SRC using raw noisy ERPs data averaging technique over some of the electrodes. This suggests that single trial LKPLS smoothing may better follow latency and amplitude changes among individual ERPs realizations. In Fig. 10 we plotted an example taken at electrode C_4 which partially supports this assumption. We may see that average taken over all 20 noisy ERPs trials tends to by slightly shifted in the latency of the second negative peak and lower in the amplitude of the first positive peak as compared to the clean ERP. It is important to note that the results on some of the electrodes indicate median NRMSE greater than one and also lower median correlation coefficient (SRC < 0.8). This occurs on electrodes with smaller SNR (frontal and temporal electrodes) and in these cases we cannot expect appropriate reconstruction of the clean ERP.

Furthermore, we were interested to see if a proper selection of smoothness in the case of SS and LSS may provide results which would be better in comparison to that achieved with LKPLS. With no restriction on the minimum value of the smoothing parameter γ used in GCV (minimum γ considered was $1e^{-8}$), both SS and LSS have tendency to follow the noisy part of the signal resulting in significantly worse performance in comparison to LKPLS, KPLS and WS. Thus, in the next step, we gradually constrained the minimum values of γ considered and investigated the performance of SS and LSS on the observed test set errors. We have to stress that this is only possible due to the fact that we know the clean ERP signals. For the case of averaged SNR=1.3dB, we observed that by constraining γ to $1e^{-5}$ LSS yielded the median NRMSE and SRC values very comparable to those of the LKPLS method (Fig. 11). We did not observe any statistically significant differences between LSS and LKPLS in this case. This value of the minimum γ was increased to $1e^{-4}$ in the case of SNR=-4.6dB and we have observed statistically significant superiority of LSS over LKPLS on four electrodes. These were posterior electrodes (P_3, P_7, O_1, O_2) characterized by the dominant P3 component and small or none P2 and N2 components (see Fig. 3E). By merging the second and third segments we observed an improved performance of LKPLS in terms of the median NRMSE and SRC on these electrodes. Only a small improvement in terms of the median NRMSE and SRC on three electrodes (P_3, P_7, O_2) was observed for LSS. We did not observe any statistically significant differences between LKPLS and LSS (using either two or three segments) on these four electrodes. Further increase of the minimum value of the γ parameter in GCV resulted in excessively smooth curves obtained by LSS with increasing median NRMSE and decreasing median SRC. Similar behavior was observed in the case of SS. However, in contrast to the KPLS case, we did not observed statistically significant improvement of LSS over SS for SNR=1.3dB. In the case of SNR=-4.6dB LSS significantly outperformed SS on two electrodes.

5 Discussion and Conclusion

We have described a new smoothing technique based on kernel PLS regression. On two different data sets we have shown that the proposed methodology may provide comparable results with the existing smoothing splines and wavelet shrinkage techniques. By expressing smoothing splines in RKHS we could directly see existing connections with our kernel PLS regression method. The close connections between smoothing splines and recently elaborated methods of regularization

networks and support vector machines, which also motivated our kernel PLS methodology, was already pointed out by Smola et al. (1998); Girosi (1998); Wahba (1999) and Evgeniou et al. (2000). Recent interest in the use and development of different types of kernel functions in kernel-based learning gives hope to extend the methodology of nonparametric curve fitting discussed in this paper. An example of this may be the recent theoretical work on statistical asymptotic properties of Gaussian and a periodic Gaussian kernels (Lin & Brown, 2002). In agreement with theoretical results, the authors have shown that periodic Gaussian kernel may provide better results in the case of very smooth functions. In the case of functions of moderate smoothness, comparable results with periodic cubic splines were achieved.

We also proposed a locally-based kernel PLS regression model. This method was designed to incorporate a priori knowledge about the approximate location of changes in signal curvature, areas of the different degree of smoothness, discontinuities or other inhomogeneities occurring in the signal. The motivation to incorporate the existing knowledge about the signal of interest came from the problem of smoothing ERP signals corrupted by high levels of noise. Knowledge about the approximate shape of ERPs is known from many psychological observations, however, the exact shape of ERPs varies from trial to trial. The relatively good results for locally-based kernel PLS on the ERP data set justify usefulness of the approach. However, where there is large variation in the latency of ERP score vectors, additional operations may be needed to determine the proper intervals and weighting functions for locally-based kernel PLS. This may involve the use of the methods for automatic segmentation of the input space (see Rasmussen and Ghahramani (2002); Tipping and Bishop (1999) and references therein). Our results also suggest that the locally-based PLS method tends to be equal or superior to the global kernel PLS approach on the ERP data set. The results also encourage us to considering the locally-based kernel PLS methodology on other problems of de-noising biological data.

On the heavisine function corrupted with high levels of white Gaussian noise we observed that, by including the knowledge about approximate location of a local inhomogeneity, locally-based kernel PLS resulted in comparable or better results in comparison to locally-based smoothing splines and wavelet shrinkage methods—the methods which are also design to deal with local inhomogeneities in the signal. Wavelet shrinkage is a method which has been shown to nicely localize spikes in signals composed of smooth components (Donoho & Johnstone, 1995). On heavisine data we have observed that locally-based kernel PLS may provide comparable (SNR=5dB) or better detection of local inhomogeneity (SNR=1dB) than wavelet shrinkage. When the underlaying true function is smooth the smoothing based techniques may result in a smoother recovery of the noise-free signal without unnecessary wiggles. This was also observed on our data sets.

The concept of localization used for building locally-based kernel PLS and smoothing splines may be also potentially implemented into the other kernel-based regression models, for example, support vector regression, kernel principal components regression or kernel ridge regression (Vapnik, 1998; Cristianini & Shawe-Taylor, 2000; Schölkopf & Smola, 2002; Rosipal et al., 2000; Rosipal et al., 2001; Saunders et al., 1998). Locally-based kernel PLS creates different set of score vectors (basis functions) for each segment of data which results in a opportunity of higher flexibility of the overall kernel PLS model. This is in contrast to the locally-based smoothing splines method where each local smoothing splines model uses a dictionary of piecewise-polynomials functions (in our case piecewise-cubic polynomials) to fit the output data. The flexibility of this approach is in the selection of different degree of smoothness for each local smoothing spline model. However, on both data sets we observed that this may lead to overfitting when the GCV criterion for the estimate of the degree of smoothness is used.

Our experiments confirmed that one of the main issues of existing nonparametric regression techniques is appropriate selection of regularization. On a data set with uncorrelated Gaussian noise added to the noise-free heavisine signal, we have observed good performance of smoothing splines using the GCV criterion for setting the degree of smoothness. However, in the case of colored noise this criterion without a priori knowledge or an appropriate estimate of the variance-covariance matrix generally tends to underestimate smoothing parameter γ resulting in overfitting (Diggle & Hutchinson, 1989; Wang, 1998; Opsomer et al., 2001). The same overfitting effect was also observed in the case of locally-based smoothing splines. By limiting the considered minimum

value of γ we have achieved good results with both smoothing splines approaches. This constraint of the range of smoothing parameter in the case of smoothing splines models has its analogy with limiting the maximum number of score vectors used in the kernel PLS model or in its locally-based modification. This limitation reflects our prior assumption of smoothness of the underlaying signal of interest. In the case that the signal has higher frequency components only smooth approximation of these parts of the signal can be achieved. Good demonstration of this was reported on heavisine data (Fig. 7d). However, it still remains an open question how to select score vectors reflecting higher frequency parts of the signal and not noise in the case of considered low SNRs. On the other hand, we may also achieve a higher flexibility of the kernel PLS approaches by controlling the width of the Gaussian kernel function with respect to the considered target function. However, in our preliminary studies on heavisine data we observed that leave-one-out model selection criterion used to set the width parameter of the Gaussian kernel had a tendency to overfitting. Thus, in our locally-based kernel PLS and kernel PLS models we fixed the width parameter to a predefined value and partially compensated for a lost flexibility of the model by creating a target dependent basis of score vectors. We used the VC-based model selection criterion to set the final number of score vectors. On the investigated data sets we have observed a good behavior of this model selection criterion when the maximum number of score vectors allowed to enter the model was generally smaller than 6-8 in the case of locally-based kernel PLS and smaller than 14-16 in the case of kernel PLS. Although this in-sample model selection criterion provided us satisfactory results with a low computational cost, it remains an open task to compare different existing model selection criteria on the considered low SNR experiments.

On the artificially generated ERP data set we have observed that by using different spatio-temporal arrangements of the signals from different electrodes in our multivariate outputs (locally-based) kernel PLS models we achieved very comparable results. This may be consider as both negative and positive. On the negative side, our belief that this arrangement may provide us some additional common information from measurements on different electrodes and trials in comparison to single electrode and single trial was not confirmed. More detailed inspection suggests that the differences between these two strategies of setting the outputs start to be more evident when higher numbers of PLS score vectors (describing also noisy part of the signal) are included into the final regression model. However, this is contradictory to our goal of smoothing the ERP signals where a lower number of score vectors is needed. Our observations indicate that these very first score vectors are generally very similar in the case of univariate (single trial single electrode) and also in the case of multivariate outputs based on a different spatio-temporal setting. Thus, it is not surprising that using these score vectors in the final regression model results in comparable performance. On the positive side, the possibility to extract desired score vectors using all electrodes from one or several trials occurred in our case to be computationally easier in comparison to the extraction of score vectors from each electrode and trial independently.

Acknowledgments

The authors would like to thank Dr. Peter Tiňo for helpful discussions during the study. This work was supported by funding from the NASA CICT/ITSR and IS/HCC programs.

References

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

Başar, E. (1980). *EEG-Brain dynamics. Relation between EEG and brain evoked potentials.* Amsterdam: Elsevier.

Centner, V., & Massart, D. (1998). Optimization in Locally Weighted Regression. *Analytical Chemistry*, *70*.

Cherkassky, V., & Shao, X. (2001). Signal estimation and denoising using VC-theory. *Neural Networks*, *14*, 37–52.

Cherkassky, V., Shao, X., Mulier, F., & Vapnik, V. (1999). Model Complexity Control for Regression Using VC Generalization Bounds. *IEEE Transactions on Neural Networks*, *10*, 1075–1089.

Chui, C. (1992). *Introduction to wavelets*. Academic Press.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, *74*.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

Diggle, P., & Hutchinson, M. (1989). On spline smoothing with autocorrelated errors. *The Australian Journal of Statistics*, *31*, 161–182.

Donoho, D. L., & Johnstone, I. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, *90*, 1200–1224.

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, *13*, 1–50.

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

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization Theory and Neural Network Architectures. *Neural Computation*, *7*, 219–269.

Green, P., & Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Heckman, N., & Zamar, R. (2000). Comparing the shapes of regression functions. *Biometrika*, *87*, 135–144.

Hillyard, S., & Kutas, M. (1983). Electrophysiology of cognitive processing. *Annual review of Psychology*, *34*, 33–61.

Höskuldsson, A. (1988). PLS Regression Methods. *Journal of Chemometrics*, *2*, 211–228.

Jasper, H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.

Johnstone, I., & Silverman, B. (1997). Wavelet Threshold Estimators for Data with Correlated noise. *Journal of the Royal Statistical Society, series B*, *59*, 319–351.

Kimeldorf, G., & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*, 82–95.

Lin, Y., & Brown, L. (2002). *Statistical Properties of the Method of Regularization with Periodic Gaussian Reproducing Kernel* (Technical Report no. 1062). Department of Statistics, University of Wisconsin, Madison, WI.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions Royal Society London*, *A209*, 415–446.

Naatanen, R., & Picton, T. (1986). N2 and automatic versus controlled processes. *Electroencephalography and Clinical Neurophysiology Supplement*, *38*, 169–186.

Naatanen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, *24*, 375–425.

Opsomer, J., Wang, Y., & Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, *16*, 134–153.

Parasuraman, R., & Beatty, J. (1980). Brain events underlying detection and recognition of weak sensory signals. *Science*, *210*, 80–83.

Picton, T., Hillyard, S., Kraus, H., & Galambos, R. (1974). Human auditory evoked potentials. I. Evaluation of components. *Electroencephalography and Clinical Neurophysiology*, *36*, 179–190.

Rasmussen, C. E., & Ghahramani, Z. (2002). Infinite mixtures of gaussian process experts. *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Rosipal, R. (2003). Kernel Partial Least Squares for Nonlinear Regression and Discrimination. *Neural Network World*, *13*, 291–300.

Rosipal, R., Girolami, M., & Trejo, L. (2000). Kernel PCA for Feature Extraction of Event-Related Potentials for Human Signal Detection Performance. *Proceedings of ANNIMAB-1 Conference* (pp. 321–326). Götegorg, Sweden.

Rosipal, R., Girolami, M., Trejo, L., & Cichocki, A. (2001). Kernel PCA for Feature Extraction and De-Noising in Nonlinear Regression. *Neural Computing & Applications*, *10*, 231–243.

Rosipal, R., & Trejo, L. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, *2*, 97–123.

Saitoh, S. (1988). *Theory of Reproducing Kernels and its Applications*. Harlow, England: Longman Scientific & Technical.

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge Regression Learning Algorithm in Dual Variables. *Proceedings of the 15th International Conference on Machine Learning* (pp. 515–521). Madison, Wisconsin.

Schölkopf, B., Smola, A., & Müller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, *10*, 1299–1319.

Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press.

Silverman, B. (1985). Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting. *Journal of the Royal Statistical Society, series B*, *47*, 1–52.

Smola, A., Schölkopf, B., & Müller, K. (1998). The connection between regularization operators and support vector kernels. *Neural Networks*, *11*, 637–649.

Tipping, M., & Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, series B*, *61*, 611–622.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Wahba, G. (1990). *Splines Models of Observational Data*, vol. 59 of *Series in Applied Mathematics*. Philadelphia: SIAM.

Wahba, G. (1999). Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola (Ed.), *Advances in Kernel Methods - Support Vector Learning*, 69–88. Cambridge, MA: The MIT Press.

Wang, Y. (1998). Smoothing Spline Models With Correlated Random Errors. *Journal of the American Statistical Association*, *93*, 341–348.

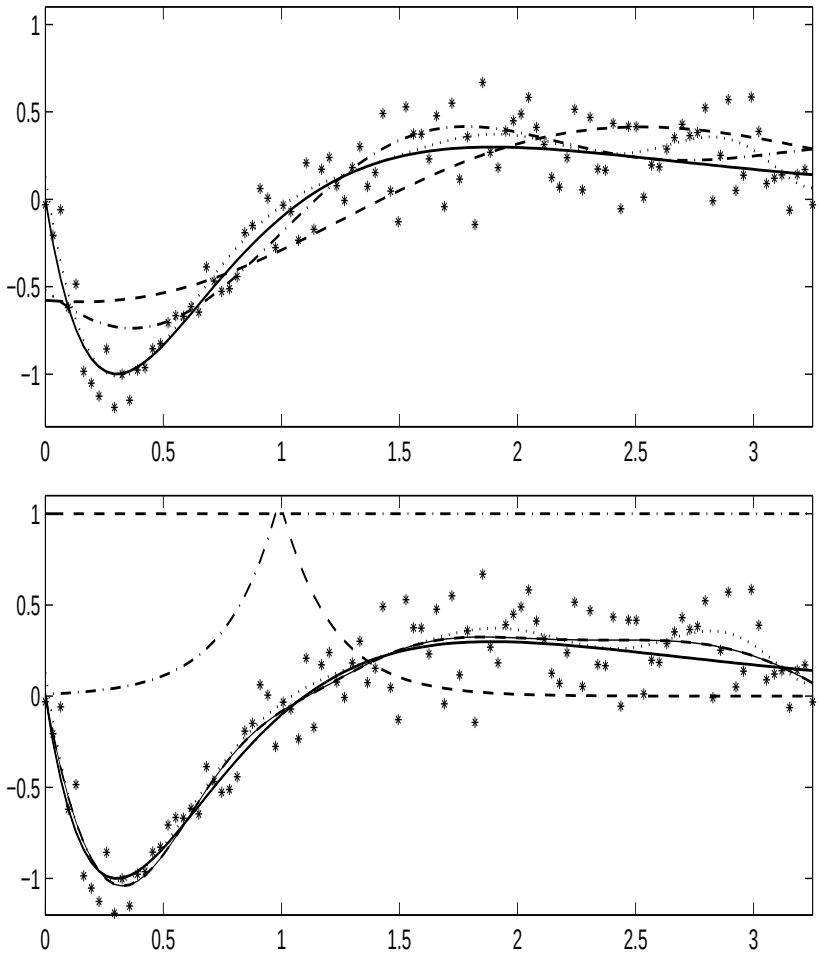

Figure 1: Smoothing of noisy $g(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$ function. The clean function is shown solid line. A noisy signal generated by adding white Gaussian noise with standard deviation equal to 0.2 is represented by asterisks. *Top*: Comparison of the kernel PLS (KPLS) regression models using different numbers of score vectors. Dashed line - KPLS with the first component (describing 64.0% of variance in input space and 66.3% variance in output space). Dash-dotted line - KPLS with the first four score vectors (describing 99.7% of variance in input space and 77.9% variance in output space). Dotted line - KPLS with the first eight score vectors (describing almost 100% of variance in input space and 86.7% variance in output space). *Bottom:* Comparison of the KPLS and locally-based KPLS (LKPLS) regression models. Dashed line - LKPLS with the first four score vectors in each model, upper dashed and dash-dotted lines represent the used weighting functions. Mean squared error (MSE) on clean function was equal to $1.9e^{-3}$. Dotted line - KPLS with the first eight score vectors. MSE on clean function was equal to $4.3e^{-3}$. Two weighting functions used in LKPLS are depicted at the top (dashed and dash-dotted lines).

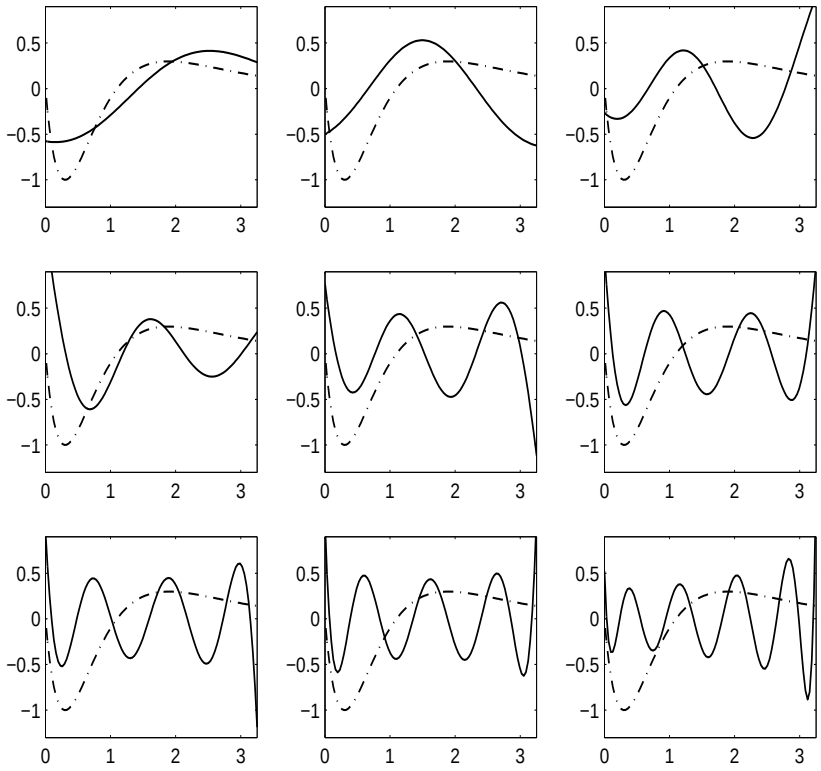

Figure 2: The first five and the ninth (from top left to bottom right) $\mathcal{F}$-space score vectors computed from noisy signal described in Fig. 1. The clean function is shown dash-dotted.

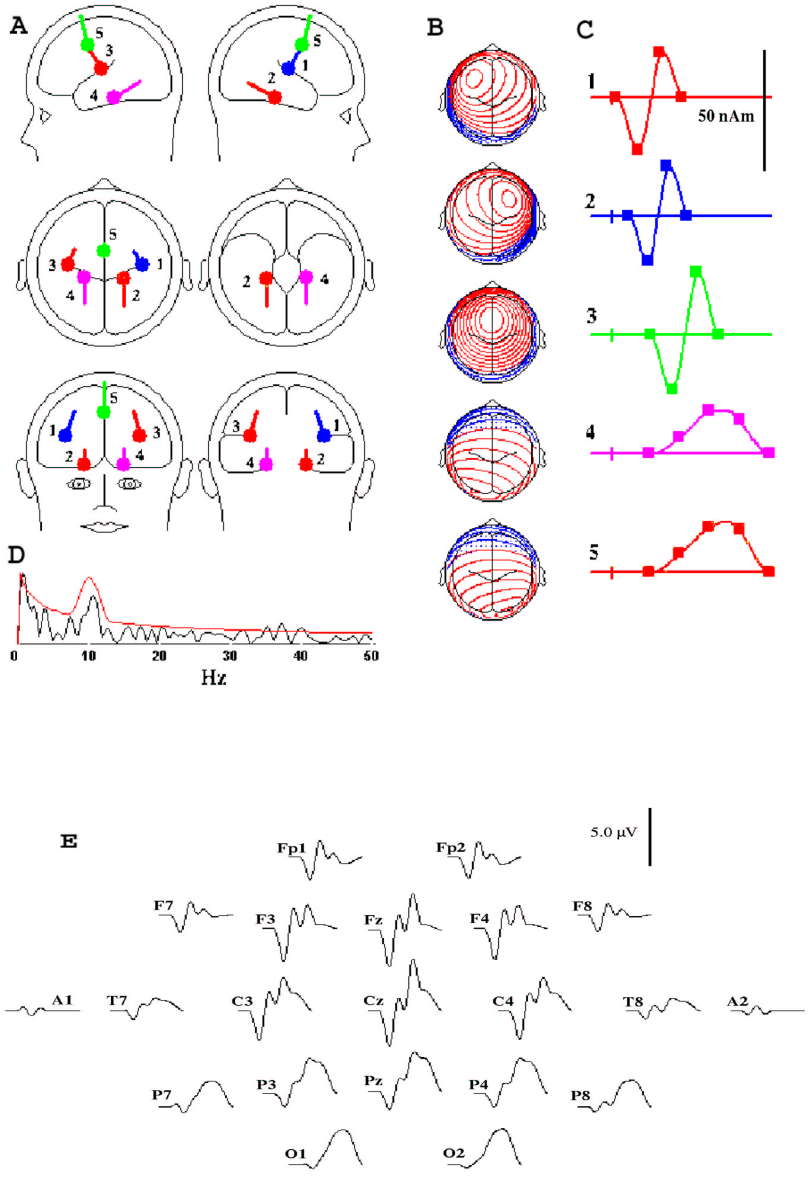

Figure 3: Simulated event related potentials (ERPs). A: Location and orientation of dipoles. B: Maps showing scalp topography of each dipole C: Activation function of each dipole. D: The averaged (over electrodes) noise amplitude spectrum and the weighting function of a sample of noise added to the ERP. The spectra simulate ongoing EEG activity with dominant α-band frequency to be 50% of the spectral power. E: Simulated ERP waveshapes at different electrodes (scalp locations).

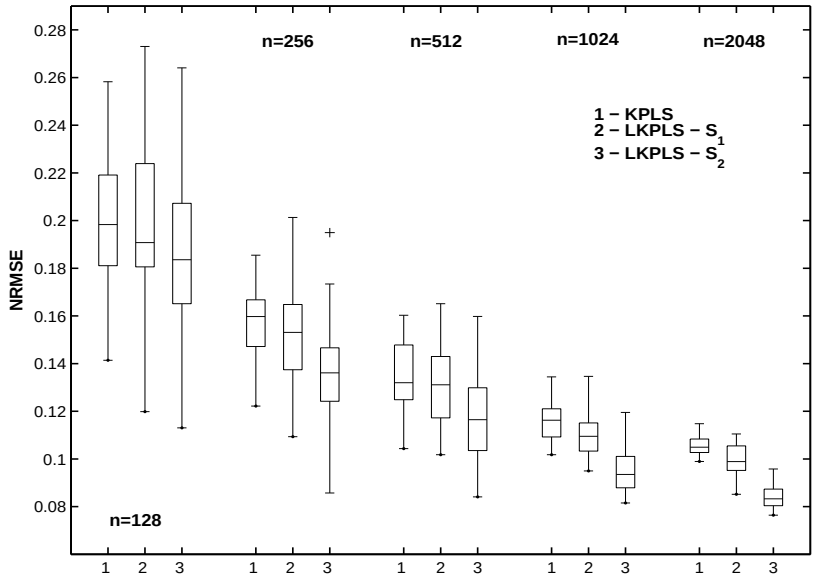

Figure 4: Results on heavisine function. Boxplots with lines at the lower quartile, median, and upper quartile values and whisker plot for three different nonparametric smoothing methods and different number (n) of samples used. The performance of kernel PLS (KPLS, left-hand boxplots in the individual triplets) is compared with locally-based kernel PLS using set of segments $\mathcal{S}_1$ (LKPLS-$\mathcal{S}_1$, middle boxplots in the triplets) and locally-based kernel PLS using set of segments $\mathcal{S}_2$ (LKPLS-$\mathcal{S}_2$, right-hand boxplots in the triplets) in terms of normalized root mean squared error (NRMSE). The boxplots are computed on results from 50 different replicates of the noisy heavisine function (SNR=5dB).

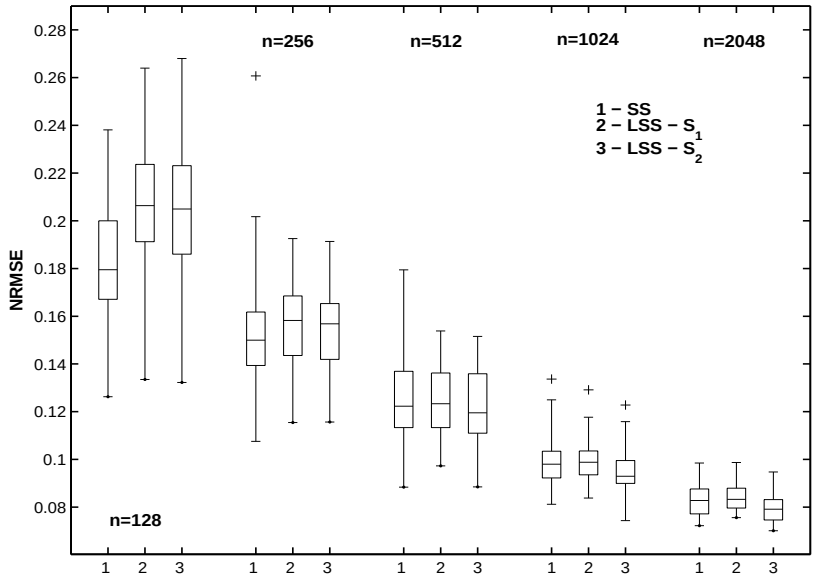

Figure 5: Results on heavisine function. Boxplots with lines at the lower quartile, median, and upper quartile values and whisker plot for three different nonparametric smoothing methods and different number (n) of samples used. The performance of smoothing splines (SS, left-hand boxplots in the individual triplets) is compared with locally-based smoothing splines using set of segments $\mathcal{S}_1$ (LSS-$\mathcal{S}_1$, middle boxplots in the triplets) and locally-based smoothing splines using set of segments $\mathcal{S}_2$ (LSS-$\mathcal{S}_2$, right-hand boxplots in the triplets) in terms of normalized root mean squared error (NRMSE). The boxplots are computed on results from 50 different replicates of the noisy heavisine function (SNR=5dB).

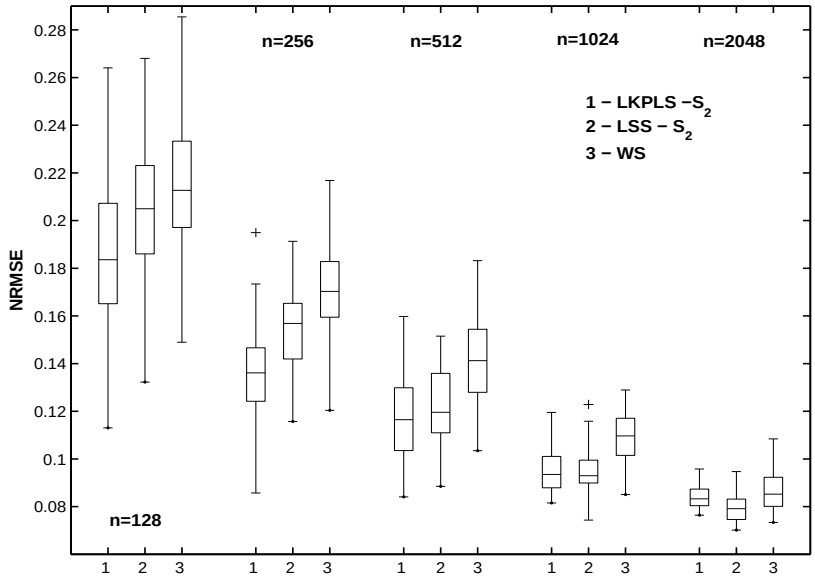

Figure 6: Results on heavisine function. Boxplots with lines at the lower quartile, median, and upper quartile values and whisker plot for three different nonparametric smoothing methods and different number (n) of samples used. The performance of locally-based kernel PLS (LKPLS-$\mathcal{S}_2$, left-hand boxplots in the individual triplets) and locally-based smoothing splines (LSS-$\mathcal{S}_2$, middle boxplots in the triplets) both using set of segments $\mathcal{S}_2$ is compared with wavelet shrinkage (WS, right-hand boxplots in the triplets) in terms of normalized root mean squared error (NRMSE). The boxplots are computed on results from 50 different replicates of the noisy heavisine function (SNR=5dB).

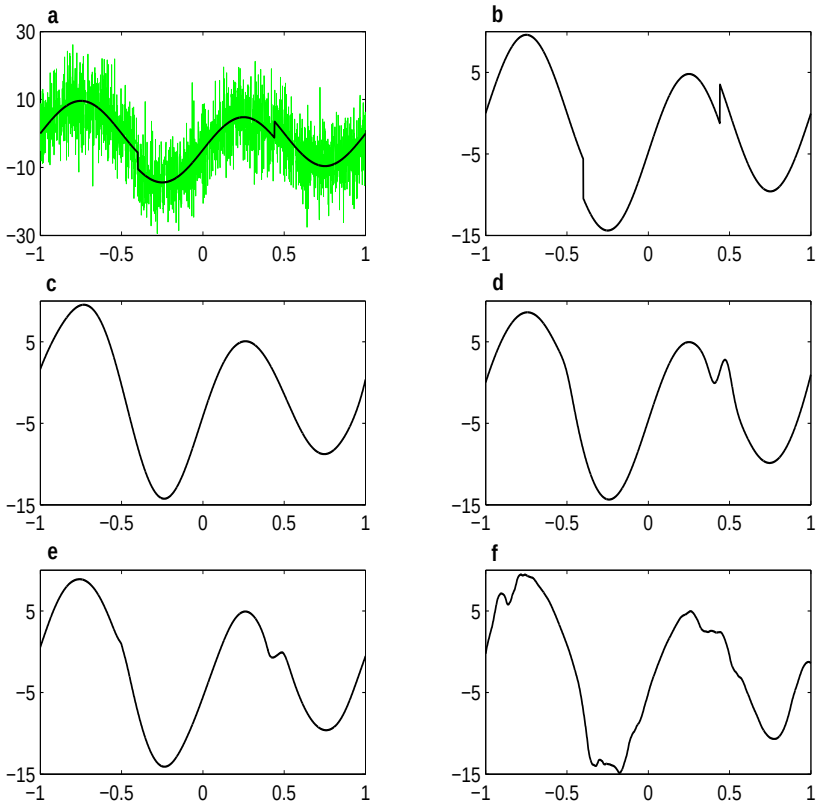

Figure 7: Results on heavisine function, SNR=1dB, number of samples $n = 2048$ a) example of noisy heavisine function b) clean heavisine function c) kernel PLS, normalized root mean squared error (NRMSE) was equal to 0.119 d) locally-based kernel PLS using set of segments $\mathcal{S}_2$, NRMSE=0.095 e) locally-based smoothing splines using set of segments $\mathcal{S}_2$, NRMSE=0.105 f) wavelet shrinkage, NRMSE=0.120.

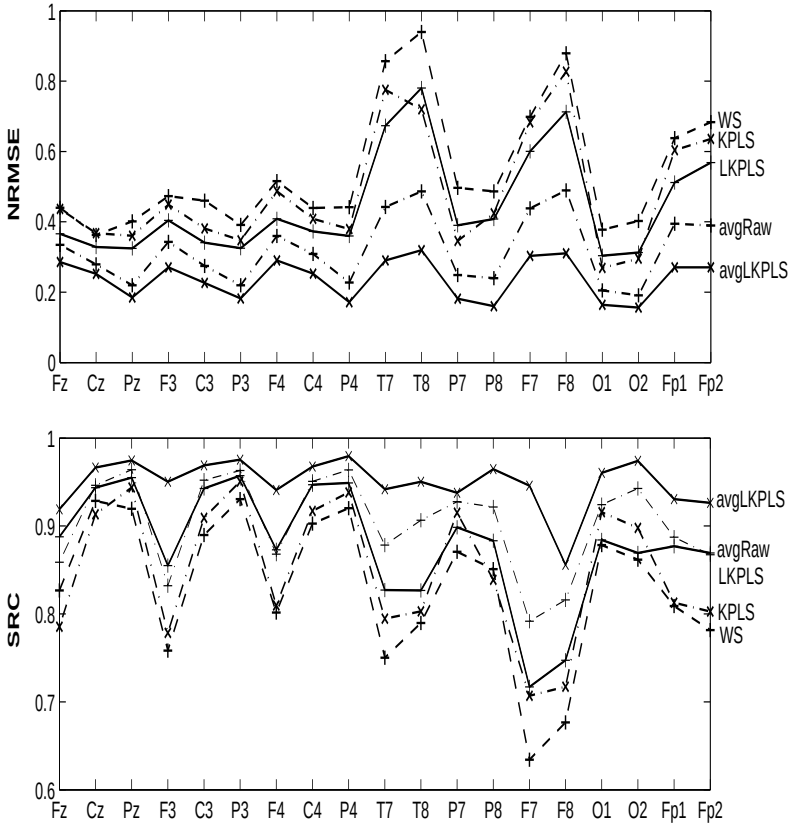

Figure 8: Results on noisy event related potentials (ERPs)–20 different trials were used. Averaged SNR over the trials and electrodes was equal to 1.3dB and 512 samples were used. Comparison of different nonparametric smoothing methods in terms of median normalized root mean squared error (NRMSE) (upper graph) and Spearman's rank correlation coefficient (SRC) (lower graph). In the upper graph the plots from the top to the bottom represent wavelet shrinkage (WS), kernel PLS (KPLS), locally-based kernel PLS (LKPLS), averaged raw ERP (avgRaw), averaged smoothed curves as result of LKPLS (avgLKPLS). The order of the plots is reversed for SRC.

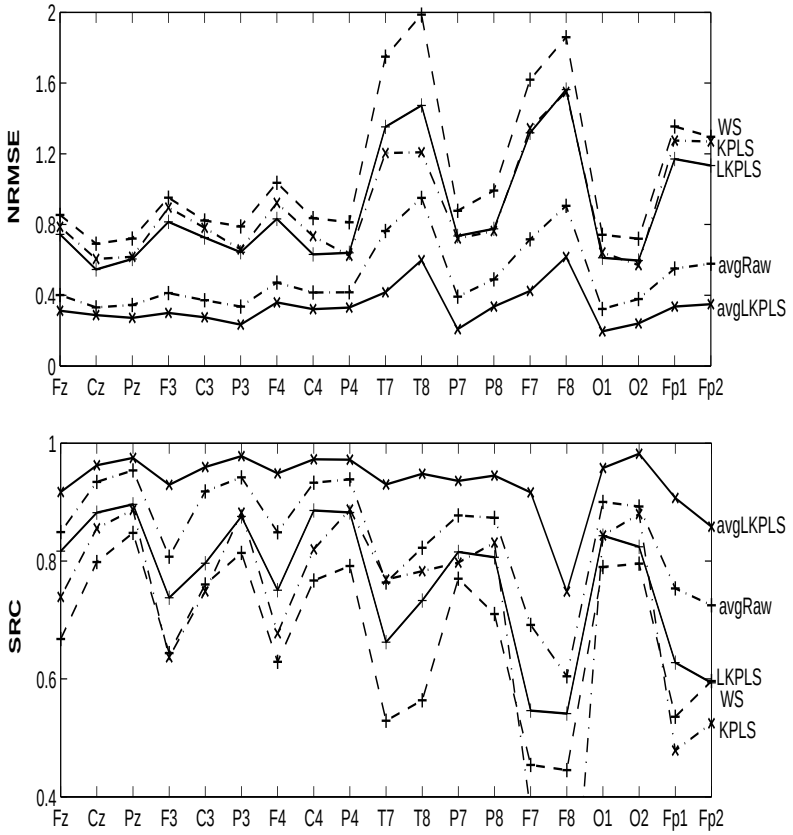

Figure 9: Results on noisy event related potentials (ERPs)–20 different trials were used. Averaged SNR over the trials and electrodes was equal to -4.6dB and 512 samples were used. Comparison of different nonparametric smoothing methods in terms of median normalized root mean squared error (NRMSE) (upper graph) and Spearman's rank correlation coefficient (SRC) (lower graph). In the upper graph the plots from the top to the bottom represent wavelet shrinkage (WS), kernel PLS (KPLS), locally-based kernel PLS (LKPLS), averaged raw ERP (avgRaw), averaged smoothed curves as result of LKPLS (avgLKPLS). The order of the plots is reversed for SRC.

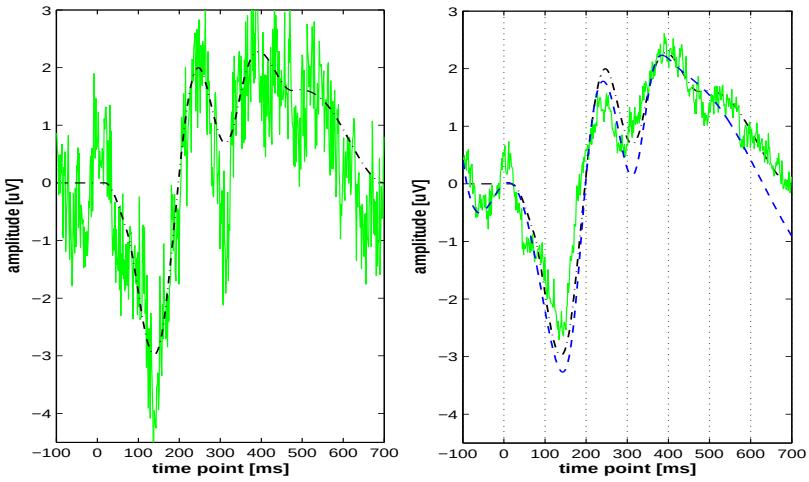

Figure 10: Example of smoothing a trial of ERP on C_4 electrode, SNR=4.5dB, number of samples $n = 512$ a) left graph: dash-dotted–clean ERP, solid line–noisy ERP b) right graph: dash-dotted line–clean ERP, dashed line–smoothed noisy ERP using locally-based kernel PLS, solid line–average of 20 different noisy ERP trials on C_4 electrode.

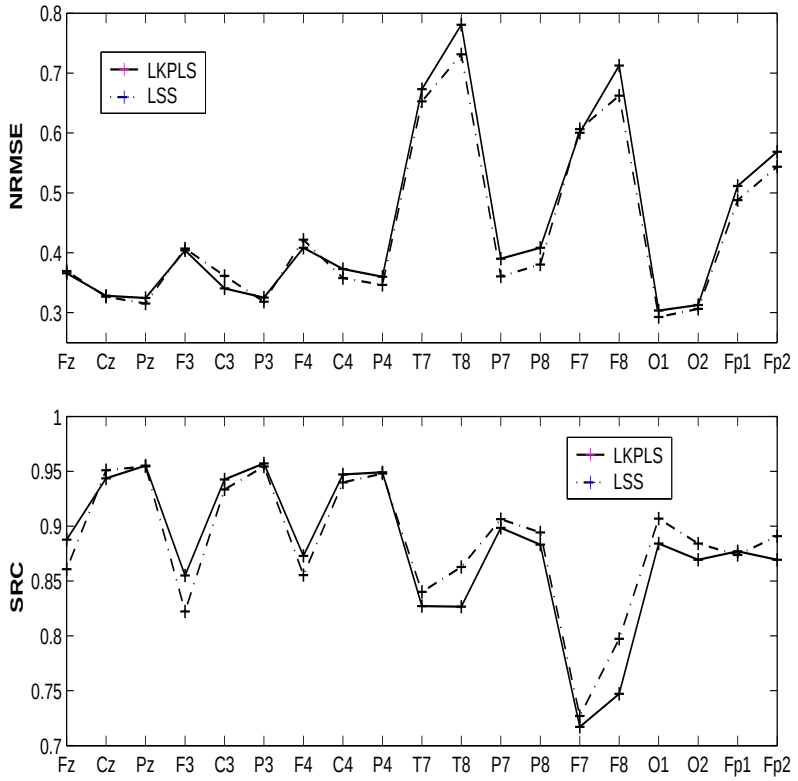

Figure 11: Results on noisy event related potentials (ERPs)–20 different trials were used. Averaged SNR over the trials and electrodes was equal to 1.3dB and 512 samples were used. Comparison of locally-based smoothing splines (dash-dotted lines) and locally-based kernel PLS (solid lines) in terms of median, upper and lower quantiles of normalized root mean squared error (NRMSE) (upper graph) and Spearman's rank correlation coefficient (SRC) (lower graph).