Nonlinear Regularized Wiener Filtering With Kernels: Application in Denoising MEG Data Corrupted by ECG

Ibtissam Constantin, Cédric Richard, *Member, IEEE*, Régis Lengellé, and Laurent Soufflet

***Abstract*—Magnetoencephalographic and electroencephalographic recordings are often contaminated by artifacts such as eye movements, blinks, and cardiac or muscle activity. These artifacts, whose amplitude may exceed that of brain signals, may severely interfere with the detection and analysis of events of interest. In this paper, we consider a nonlinear approach for cardiac artifacts removal from magnetoencephalographic data, based on Wiener filtering. In recent works, nonlinear Wiener filtering based on reproducing kernel Hilbert spaces and the *kernel trick* has been proposed. However, the filter parameters are determined by the resolution of a linear system which may be ill conditioned. To deal with this problem, we introduce three kernel methods that provide powerful tools for solving ill-conditioned problems, namely, kernel principal component analysis, kernel partial least squares, and kernel ridge regression. A common feature of these methods is that they regularize the solution by assuming an appropriate prior on the class of possible solutions. We avoid the use of QRS-synchronous averaging techniques, which may induce distortions in brain signals if artifacts are not well detected. Moreover, our approach shows the nonlinear relation between magnetoencephalographic and electrocardiographic signals.**

***Index Terms*—Cardiac artifacts extraction, nonlinear Wiener filtering, regularization, reproducing kernel Hilbert spaces.**

I. Introduction

Electroencephalography (EEG) and magnetoencephalography (MEG) are noninvasive techniques able to provide direct information about the neural brain activity with high temporal resolution. They record, respectively, the electrical fields and magnetic fields generated from neural currents inside the brain. Since magnetic fields are not distorted while passing through the skull and the scalp, MEG may have better spatial resolution for source localization. EEG and MEG measurements are often corrupted by artifacts, such as cardiac artifacts, generated by heart activity [1]. These artifacts can be several times stronger in magnitude than the signals of interest and may severely impede the extraction of relevant information.

A common strategy used for artifact rejection consists in discarding portions of brain signals that exceed a preselected criterion threshold. However, this technique may lead to a significant loss of data, particularly when artifacts occur too frequently. Other approaches based on linear models have been applied extensively. The most popular one is regression in the time [2], [3], or frequency domain [4], [5]. This approach depends on one or more recorded artifact channels to estimate the denoised brain signals. Further linear methods rely on spatial filtering such as signal space projection (SSP) [1], [6], [7], principal component analysis (PCA) [8]–[11], common spatial subspace decomposition (CSSD) [12], and independent component analysis (ICA) [11], [13]–[19]. SSP subtracts from brain measurements the noise components oriented along specified spatial vectors. It requires a good model of the artifactual source or a considerable amount of data where the artifact amplitude is much higher than brain signals. PCA and CSSD decompose an epoch of MEG or EEG channels into several orthogonal components and identify and remove the artifactual components. They differ in the way these components are determined. PCA determines an orthogonal basis in which the variance of the data is large. This is achieved by the diagonalization of the covariance data matrix. The orthogonal components are found by projecting the data onto the basis vectors. In contrast, CSSD looks for orthogonal components by joint diagonalization of the data matrix and the pure artifact matrix. Both PCA and CSSD perform well if the signal and noise components are orthogonal to each other. Otherwise, they will mix the signals of interest and artifacts. ICA can be viewed as an extension of PCA and CSSD that has been developed in the context of blind source separation problems. ICA aims to decompose the data into statistically independent components by optimizing a suitable contrast function. It includes JADE [11], [20], Infomax [11], [21], FastICA [11], [22], and TDSEP [11], [23], [24]. Compared to PCA and CSSD, ICA removes the orthogonality constraint and forces components to be approximately independent. In general, there is no reason to assume that artifact signals are orthogonal to the signals of interest, and therefore ICA is usually more efficient than PCA and CSSD. Nonlinear techniques based on neural networks have also been considered. In [25], a neural network has been set up in order to estimate the ECG artifacts that corrupt MEG signals. The filter output is a sum of nonlinear functions of past samples, and the algorithm requires a triggered version of the ECG reference channel. Estimation of neural network parameters has been achieved by a backpropagation-like algorithm. Selection of the network architecture, which controls generalization and remains an open problem, has not been detailed. The authors have shown that the filter outperforms standard linear filters.

Manuscript received July 14, 2005; accepted March 1, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jan C. de Munck.

I. Constantin and L. Soufflet are with FORENAP FRP, 68250 Rouffach, France (e-mail: ibtissam.constantin@forenap.com; l.soufflet@forenap.com).

C. Richard and R. Lengellé are with the Laboratory ISTIT-M2S (FRE CNRS 2732), Troyes University of Technology, Troyes, France (e-mail: cedric.richard@utt.fr; regis.lengelle@utt.fr).

Digital Object Identifier 10.1109/TSP.2006.882115

1053-587X/$20.00 © 2006 IEEE

Recently, several nonlinear kernel-based algorithms have been proposed within the context of the theory of reproducing kernel Hilbert spaces (RKHSs). They have been applied successfully to a wide range of applications. The main ingredient of kernel-based methods is the so-called *kernel trick,* which provides the technical basis of learning in arbitrarily high-dimensional spaces, by means of kernel functions defined on pairs of input patterns. In recent work, a nonlinear Wiener filter based on RKHS and the kernel trick has been proposed. As for a linear Wiener filter, the filter parameters are determined by solving a linear system of equations. The filter design is simple, and its generalization ability is governed by a few tunable variables. Unlike neural networks, it does not require the choice of a particular architecture before training. Various filter structures can be selected by choosing different kernels from a wide class of functions verifying a certain condition. In this paper, we apply nonlinear kernel-based Wiener filtering to the problem of cardiac artifacts extraction from MEG data. We introduce three kernel methods that provide powerful tools for solving ill-conditioned problems, namely, kernel principal component analysis (KPCA) [26], [27], kernel partial least squares (KPLS) [28], and kernel ridge regression (KRR) [29], [30]. A common feature of these methods is that they regularize the solution by assuming an appropriate prior on the class of possible solutions.

This paper is organized as follows. In the next section, we present some elements on nonlinear Wiener filtering and its connections with RKHS. In Section III, some necessary prerequisites on the theory of RKHS are given. We derive the Wiener filter in RKHS in Section IV. In Section V, we describe KPCA, KPLS, and KRR methods, and we derive the regularized kernel-based Wiener filter. The filter is experimented with on simulated data for nonlinear system identification. Experimental results in MEG signals nonlinear denoising are presented in Section VI. Some concluding remarks are given in Section VII.

II. Nonlinear Wiener Filtering

We consider a *filter* every material or programmed structure applied to a quantity of interest in order to extract significant information in the sense of a given criterion. This quantity may, for example, be generated from noisy sensing devices or be issued from communication channels subject to perturbations. The canonical form of the filtering problem is proposed in Fig. 1. It comprehends an input x_n and a desired output d_n, supposed to be centered and real-valued without loss of generality. We note $e_n = d_n - y_n$ as the committed error. The objective consists in the selection of a model w and the implementation of an operational technique for determining its parameters. Linear Wiener theory is applied to jointly wide-sense stationary processes and consists in finding the parameters w_i [31]

$$y_n = \sum_{i=0}^{N-1} w_i x_{n-i} \tag{1}$$

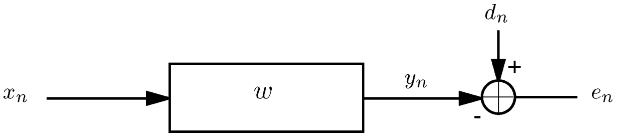

Fig. 1. Block diagram and notations.

that minimize the variance of the error e_n. By introducing the following notations:

$$\boldsymbol{x}_n = [x_n, \ldots, x_{n-N+1}]^t$$
$$\boldsymbol{w} = [w_0, \ldots, w_{N-1}]^t$$

the solution of the problem is obtained by solving Wiener–Hopf equation

$$E\left\{\boldsymbol{x}_n(d_n - \boldsymbol{w}^t\boldsymbol{x}_n)\right\} = 0 \tag{2}$$

where E denotes the mathematical expectation. The previous expression can be written in compact matrix form

$$\boldsymbol{R_x w} = \boldsymbol{R_{x}}_d \tag{3}$$

where $\boldsymbol{R_x}$ and $\boldsymbol{R_x}_d$ denote $E\{\boldsymbol{x}_n\boldsymbol{x}_n^t\}$ and $E\{d_n\boldsymbol{x}_n\}$, respectively. Note that the linearity of the filter facilitates its design, to the detriment of its capacity to give satisfactory solution to every problem. To overcome such a limit, while keeping a comparable structure to (1), we can map the observation $\boldsymbol{x}_n$ into a high-dimensional feature space by means of a nonlinear application ϕ and then consider the filter $y_n = \boldsymbol{w}^t\phi(\boldsymbol{x}_n)$. As previously, $\boldsymbol{w}$ can be determined by solving the Wiener–Hopf equation

$$E\left\{\phi(\boldsymbol{x}_n)\left(d_n - \boldsymbol{w}^t\phi(\boldsymbol{x}_n)\right)\right\} = 0. \tag{4}$$

In the literature, many nonlinear filter structures have been elaborated on the basis of this principle, e.g., polynomial filters [32]–[34]. However, a major drawback of this approach is the computational burden associated with the large dimension of the new space. The design of a second-order polynomial filter, for example, requires the use of a nonlinear map $\phi(\boldsymbol{x}_n)$, based on the components of $\boldsymbol{x}_n$ and their second-order products. The number of parameters to be estimated equals $N(N+3)/2$ and is already prohibitive with regard to the relative simplicity of the filter. This can be the source of practical difficulties when such a strategy is adopted. In the field of pattern recognition, several results on RKHS have made analogous practices possible. By authorizing the synthesis of generalized linear structures without explicitly evaluating the map $\phi(\boldsymbol{x}_n)$, RKHSs have given new perspectives within the framework of kernel methods, particularly with support vector machines [26], [29], [35], kernel Fisher discriminant [36], [37], and kernel second-order discriminant [38], [39], for solving classification and regression problems.

III. RKHS and Mercer's Condition

Let $\mathcal{H}$ be a reproducing kernel Hilbert space consisting of mappings ψ from a compact $\mathcal{X} \subset \mathrm{I\!R}^N$ to $\mathrm{I\!R}$ and let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denote the dot product defined on $\mathcal{H}$. As the Riesz representation

TABLE I
SOME EXAMPLES OF TYPICAL MERCER KERNELS

Dot-product kernels		Translation invariant kernels	
Sigmoid	$\tanh(\alpha+\beta{<}\boldsymbol{x},\boldsymbol{y}{>})$	Gaussian	$\exp(\|\boldsymbol{x}-\boldsymbol{y}\|^2/2\sigma^2)$
Homogeneous polynomial	$({<}\boldsymbol{x}\,,\boldsymbol{y}{>})^q$, $q\in\mathbb{N}^*$	Laplace	$\exp(\|\boldsymbol{x}-\boldsymbol{y}\|^2/\sigma)$
Inhomogeneous polynomial	$(\alpha+{<}\boldsymbol{x}\,,\boldsymbol{y}{>})^q$, $q\in\mathbb{N}^*$	Inverse multiquadric	$\frac{1}{\sqrt{\alpha^2+\|\boldsymbol{x}-\boldsymbol{y}\|^2}}$

theorem states, there is a unique function $\kappa(\cdot,\boldsymbol{y})$ of $\mathcal{H}$ which verifies the following reproducing property:

$$\psi(\boldsymbol{y})=\langle\psi,\kappa(\cdot,\boldsymbol{y})\rangle_{\mathcal{H}}\,,\quad\forall\psi\in\mathcal{H}\tag{5}$$

for every $\boldsymbol{y}\in\mathcal{X}$. A proof of this may be found in [40]. Here $\kappa(\cdot,\boldsymbol{y})$ is the representer of evaluation at $\boldsymbol{y}$ and κ is the reproducing kernel associated with $\mathcal{H}$. In particular, $\{\kappa(\cdot,\boldsymbol{y}):\boldsymbol{y}\in\mathcal{X}\}$ spans $\mathcal{H}$ and the dot product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ has just to be defined on it. Denoting the function $\kappa(\cdot,\boldsymbol{y})$ by $\phi(\boldsymbol{y})$, (5) implies

$$\kappa(\boldsymbol{x},\boldsymbol{y})=\langle\phi(\boldsymbol{x}),\phi(\boldsymbol{y})\rangle_{\mathcal{H}}\tag{6}$$

for all $\boldsymbol{x}$, $\boldsymbol{y}\in\mathcal{X}$. The kernel κ then evaluates the dot product of every pair of elements of $\mathcal{X}$ mapped into $\mathcal{H}$, without any explicit knowledge of either ϕ or $\mathcal{H}$. The key idea of the kernel technique used in this paper, commonly known as the kernel trick, is to choose the kernel κ rather than the mapping ϕ. Of course, not every function κ can serve as a kernel. According to the Hilbert–Schmidt theory [41], any continuous symmetric function κ can be expanded as follows:

$$\kappa(\boldsymbol{x},\boldsymbol{y})=\sum_{i=0}^{\infty}\lambda_i\psi_i(\boldsymbol{x})\psi_i(\boldsymbol{y})\tag{7}$$

where λ_i and ψ_i are eigenvalues and eigenfunctions that satisfy

$$\int\kappa(\boldsymbol{x},\boldsymbol{y})\psi_i(\boldsymbol{x})d\boldsymbol{x}=\lambda_i\psi_i(\boldsymbol{y}).\tag{8}$$

A sufficient condition to ensure that κ is a dot product in some Hilbert space $\mathcal{H}$ is that all the λ_is in (7) are positive. According to Mercer's theorem [42], this condition is achieved if and only if

$$\iint\kappa(\boldsymbol{x},\boldsymbol{y})g(\boldsymbol{x})g(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y}\geq 0\tag{9}$$

for all g fulfilling $\int g(\boldsymbol{x})^2d\boldsymbol{x}<\infty$. From (7), it is straightforward to construct a map ϕ into a potentially infinite-dimensional space which satisfies (6). For example, we might use $\phi(\boldsymbol{x})=[\sqrt{\lambda_0}\psi_0(\boldsymbol{x}),\sqrt{\lambda_1}\psi_1(\boldsymbol{x}),\ldots]^t$. In [43], it has been reported that (9) corresponds to the statement that κ is a positive definite kernel.

A typical example of kernels is the polynomial kernel $\kappa(\boldsymbol{x},\boldsymbol{y})=(\alpha+\langle\boldsymbol{x},\boldsymbol{y}\rangle)^q$, $q\in\mathbb{N}^*$, of homogeneous ($\alpha=0$) or inhomogeneous type ($\alpha>0$). It follows from [44] that polynomial kernels satisfy Mercer's condition. Radial kernels are also Mercer kernels that have received significant attention in statistical and machine learning communities. They depend on $\|\boldsymbol{x}-\boldsymbol{y}\|$. We count among them the Gaussian kernel defined by $\kappa(\boldsymbol{x},\boldsymbol{y})=\exp(\|\boldsymbol{x}-\boldsymbol{y}\|^2/2\sigma^2)$, where σ is the kernel bandwidth. This kernel is characterized by a continuum of eigenvalues which means that the components of $\phi(\boldsymbol{x})$ are not in limited number as for the polynomial kernels. Empirical findings show that Gaussian kernel ensures generally good performance under general smoothness assumptions and should be considered if no additional knowledge of the data is available. The above-mentioned examples fall in two main classes of kernels: dot-product kernels involving $\langle\boldsymbol{x},\boldsymbol{y}\rangle$ and translation invariant kernels depending on $\boldsymbol{x}-\boldsymbol{y}$. Other examples may be found in Table I. See also [45] and [46]. Moreover, there exist simple rules for designing valid kernels on the basis of given Mercer kernels, e.g., the sum and the product of two kernels are also Mercer kernels [46].

In the context of function estimation, an important theorem in RKHS [47], which provides a framework for solving a wide range of optimization problems, states that any function f minimizing a criterion of the form

$$c\left((\boldsymbol{x}_0,d_0,f(\boldsymbol{x}_0)),\ldots,(\boldsymbol{x}_{M-1},d_{M-1},f(\boldsymbol{x}_{M-1}))\right)+o\left(\|f\|_{\mathcal{H}}\right)$$

with o a monotonic increasing function on $[0,\infty[$, admits a representation of the form

$$f(\boldsymbol{x})=\sum_{i=0}^{M-1}\alpha_i\kappa(\boldsymbol{x},\boldsymbol{x}_i).\tag{10}$$

This theorem is referred to as the *representer theorem*. It shows that the optimal solution in $\mathcal{H}$ is constrained to lie in the space spanned by a set of basis functions $\kappa(.,\boldsymbol{x}_i)$, $i=0,\ldots,M-1$. The $\{\boldsymbol{x}_i\}_{i=0}^{M-1}$ are known as training samples. The essence of the proof of the representer theorem is that every function f in the Hilbert space $\mathcal{H}$ can be decomposed into two components, a component in the subspace $\mathcal{H}_M$ spanned by $\kappa(.,\boldsymbol{x}_i)$, $i=0,\ldots,M-1$, and a component orthogonal to it, i.e., $f(.)=\sum_{i=0}^{M-1}\alpha_i\kappa(.,\boldsymbol{x}_i)+f_\perp(.)$, where $\langle f_\perp,\kappa(.,\boldsymbol{x}_i)\rangle_{\mathcal{H}}=0$, for all $i=0,\ldots,M-1$. By (5), it follows that $f(\boldsymbol{x}_j)=\sum_{i=0}^{M-1}\alpha_i\kappa(\boldsymbol{x}_j,\boldsymbol{x}_i)$, which means that the values of f at the data points $\boldsymbol{x}_i$, $i=0,\ldots,M-1$ are not affected by $f_\perp$. They only depend on the coefficients α_i.

IV. WIENER FILTERING IN RKHS

Let $\mathcal{H}$ be an RKHS defined by a kernel κ. Let ϕ denote the mapping function from $\mathcal{X}$ to $\mathcal{H}$. Referring to the previous section, the output of the Wiener filter $y_n=\boldsymbol{w}^t\phi(\boldsymbol{x}_n)$, with w satisfying (4), may be written as

$$y_n=\sum_{i=0}^{M-1}\alpha_i\kappa(\boldsymbol{x}_n,\boldsymbol{x}_i)\tag{11}$$

where the parameters α_i are to be determined in order to minimize the variance of the error e_n. As with a linear Wiener filter, we make the assumption that the mapped function $\phi(\boldsymbol{x}_n)$ as well as the desired output d_n are centered. We shall return to this point later. Notice that the filter output is not explicitly based on the analytical expression of ϕ. It is implicitly defined by the choice of a reproducing kernel κ. For example, the use of the polynomial kernel $\kappa(\boldsymbol{x},\boldsymbol{y}) = (1 + \langle \boldsymbol{x},\boldsymbol{y}\rangle)^q$ leads to a Volterra filter of degree q, whose dual formulation is provided by $y_n = \sum_{i=0}^{M-1} \alpha_i (1 + \langle \boldsymbol{x}_n, \boldsymbol{x}_i\rangle)^q$. The latter does not require the evaluation of $\phi(\boldsymbol{x}_n)$. Remember that the components of $\phi(\boldsymbol{x}_n)$ are products of degree less than or equal to q composed of the components of $\boldsymbol{x}_n$. Finally, the solution of the problem is obtained by solving an analogous system to (3) in which $\boldsymbol{k}_n = [\kappa(\boldsymbol{x}_n, \boldsymbol{x}_0), \ldots, \kappa(\boldsymbol{x}_n, \boldsymbol{x}_{M-1})]^t$ is substituted to $\boldsymbol{x}_n$. It translates to

$$\boldsymbol{R}_{\boldsymbol{k}}\boldsymbol{\alpha} = \boldsymbol{R}_{\boldsymbol{k}d}. \tag{12}$$

As can be seen, the number of parameters to be estimated is independent of the complexity of the filter, which is characterized by the polynomial degree in the particular case of Volterra filter. The advantage of linear-in-the-parameters estimation remains. In practice, the correlation matrices $\boldsymbol{R}_{\boldsymbol{k}}$ and $\boldsymbol{R}_{\boldsymbol{k}d}$ are unknown. They must be replaced by their estimates, which can be computed easily from the vectors $\{\boldsymbol{k}_i\}_{i=0}^{M-1}$ and the corresponding desired responses $\{d_i\}_{i=0}^{M-1}$ by $\widehat{\boldsymbol{R}}_{\boldsymbol{k}} = (1/M)\sum_{i=0}^{M-1} \boldsymbol{k}_i \boldsymbol{k}_i^t = (1/M)\boldsymbol{K}^t\boldsymbol{K}$ and $\widehat{\boldsymbol{R}}_{\boldsymbol{k}d} = (1/M)\sum_{i=0}^{M-1} d_i \boldsymbol{k}_i = (1/M)\boldsymbol{K}^t\boldsymbol{d}$, where $\boldsymbol{K} = [\boldsymbol{k}_0, \ldots, \boldsymbol{k}_{M-1}]^t$ and $\boldsymbol{d} = [d_0, \ldots, d_{M-1}]^t$. Consequently, we have

$$\boldsymbol{\alpha} = (\boldsymbol{K}^t\boldsymbol{K})^{-1}\boldsymbol{K}^t\boldsymbol{d} = \boldsymbol{K}^{-1}\boldsymbol{d}. \tag{13}$$

The last equality results from the fact that $\boldsymbol{K}$ is symmetric.

As pointed out previously, we assume that the data are centered. Given $\phi(\boldsymbol{x}_i)$, $i = 0, \ldots, M-1$, this may be achieved by replacing the element K_{ij} of $\boldsymbol{K}$ with [27]

$$\begin{aligned}\overline{K}_{ij} &= \left(\phi(\boldsymbol{x}_i) - \frac{1}{M}\sum_{l=0}^{M-1}\phi(\boldsymbol{x}_l)\right)^t \left(\phi(\boldsymbol{x}_j) - \frac{1}{M}\sum_{k=0}^{M-1}\phi(\boldsymbol{x}_k)\right)\\ &= K_{ij} - \frac{1}{M}\sum_{k=0}^{M-1}(K_{ik} - K_{jk}) + \frac{1}{M^2}\sum_{l=0}^{M-1}\sum_{k=0}^{M-1} K_{lk}.\end{aligned} \tag{14}$$

In a compact matrix form, the previous expression can be written

$$\overline{\boldsymbol{K}} = \left(\boldsymbol{I} - \frac{1}{M}\mathbf{1}\right)\boldsymbol{K}\left(\boldsymbol{I} - \frac{1}{M}\mathbf{1}\right) \tag{15}$$

where $\boldsymbol{I}$ denotes the $(M \times M)$ identity matrix and $\mathbf{1}$ the $(M \times M)$ matrix whose elements all equal one. Consider now a set of new observations $\{\boldsymbol{x}_i^{test}, d_i^{test}\}_{i=0}^{L-1}$ for which we wish to evaluate the output of the filter. The corresponding filter output is given by

$$\boldsymbol{y}^{test} = \boldsymbol{K}^{test}\boldsymbol{\alpha} \tag{16}$$

where $\boldsymbol{K}^{test}$ is the $(L \times M)$ matrix whose entries are $K_{ij}^{test} = \kappa(\boldsymbol{x}_i^{test}, \boldsymbol{x}_j)$. Similarly to (14), $\boldsymbol{K}^{test}$ must be modified as

$$\overline{\boldsymbol{K}}^{test} = \left(\boldsymbol{K}^{test} - \frac{1}{M}\mathbf{1}'\boldsymbol{K}\right)\left(\boldsymbol{I} - \frac{1}{M}\mathbf{1}\right) \tag{17}$$

where $\mathbf{1}'$ is the $(L \times M)$ matrix with all its elements equal to one.

V. Regularized Wiener Filter

The resolution of (13) requires the inversion of a matrix whose dimensions are equal to the number of training samples. The problem may be ill conditioned and the solution unstable. In [48], KPCA [26], [27] has been used to obtain a regularized solution to this problem. In this section, we compare KPCA with two other kernel-based regularization techniques: KPLS [28] and KRR [29], [30]. These methods are nonlinear extensions of PCA [8]–[11], partial least squares (PLS) [49], [50], and ridge regression (RR) [51]. They have been applied successfully in various fields of signal processing—for example, within the context of object recognition, text categorization, and data analysis. In this section, they are used in a filtering context, to design a regularized Wiener filter. Throughout this section, we assume that the data are centered as explained previously.

A. KPCA

A way to regularize the solution of an ill-conditioned problem is to project the input data in a lower dimensional space and then derive the solution in the reduced space. This is a form of regularization since it restricts the class of reachable solutions of the filter design problem. KPCA is an efficient method for dimensionality reduction in RKHS. It looks for orthogonal directions in $\mathcal{H}$ so as to maximize

$$\operatorname{var}\left(\langle \phi(\boldsymbol{x}_n), \boldsymbol{v}\rangle_{\mathcal{H}}\right) \tag{18}$$

under the constraint $\langle \boldsymbol{v}, \boldsymbol{v}\rangle_{\mathcal{H}} = 1$. It can be shown that the solution is provided by the diagonalization of the correlation matrix in $\mathcal{H}$, estimated by

$$\widehat{\boldsymbol{R}}_{\phi} = \frac{1}{M}\sum_{i=0}^{M-1}\phi(\boldsymbol{x}_i)\phi^t(\boldsymbol{x}_i) = \frac{1}{M}\boldsymbol{\Phi}^t\boldsymbol{\Phi} \tag{19}$$

where $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_0), \ldots, \phi(\boldsymbol{x}_{M-1})]^t$. Therefore, the problem consists of finding the eigenvalues $\lambda > 0$ and the eigenvectors $\boldsymbol{v} \in \mathcal{H}$ of $\widehat{\boldsymbol{R}}_{\phi}$ satisfying

$$\widehat{\boldsymbol{R}}_{\phi}\boldsymbol{v} = \lambda\boldsymbol{v}. \tag{20}$$

Replacing $\widehat{\boldsymbol{R}}_{\phi}$ in (20) with its expression in (19) leads to

$$\frac{1}{M}\sum_{i=0}^{M-1}\langle \phi(\boldsymbol{x}_i), \boldsymbol{v}\rangle_{\mathcal{H}}\,\phi(\boldsymbol{x}_i) = \lambda\boldsymbol{v}. \tag{21}$$

As can be seen from (21), every eigenvector with nonzero eigenvalue lies in the span of $\{\phi(\boldsymbol{x}_i)\}_{i=0}^{M-1}$. This can be written as

$$\boldsymbol{v} = \sum_{i=0}^{M-1} a_i\phi(\boldsymbol{x}_i) = \boldsymbol{\Phi}^t\boldsymbol{a}. \tag{22}$$

Using this definition of $\boldsymbol{v}$, (20) translates to

$$\frac{1}{M}\boldsymbol{\Phi}^t\boldsymbol{\Phi}\boldsymbol{\Phi}^t\boldsymbol{a} = \lambda\boldsymbol{\Phi}^t\boldsymbol{a}. \tag{23}$$

Premultiplying with $(\boldsymbol{\Phi}\boldsymbol{\Phi}^t)^{-1}\boldsymbol{\Phi}$, (23) becomes

$$\boldsymbol{K}\boldsymbol{a} = M\lambda\boldsymbol{a}. \tag{24}$$

Let $\boldsymbol{v}_i$ denote the ith eigenvector of $\widehat{\boldsymbol{R}}_{\phi}$ corresponding to the nonzero eigenvalue λ_i and $\boldsymbol{a}_i$ the associated eigenvector of $\boldsymbol{K}$.

To ensure that the eigenvectors $\boldsymbol{v}_i$ have unit norm in the feature space, $\boldsymbol{a}_i$ should be divided by $\sqrt{M\lambda_i}$.

Let $\boldsymbol{b}_n$ be the projection of $\phi(\boldsymbol{x}_n)$ on a subset $\{\boldsymbol{v}_i\}_{i=0}^{S-1}$ associated to the S largest eigenvalues. The ith element of $\boldsymbol{b}_n$ is called the ith principal component (PC) of $\phi(\boldsymbol{x}_n)$ [11]. We have

$$\boldsymbol{b}_n = \boldsymbol{V}^t\phi(\boldsymbol{x_n}) = \boldsymbol{A}^t\boldsymbol{k}_n \tag{25}$$

where $\boldsymbol{V} = [\boldsymbol{v}_0,\ldots,\boldsymbol{v}_{S-1}]$ and $\boldsymbol{A} = [\boldsymbol{a}_0,\ldots,\boldsymbol{a}_{S-1}]$. The last equation results from (22). We now have to determine the Wiener filter operating on $\boldsymbol{b}_n$. The output of the filter is defined by

$$y_n = \boldsymbol{w}^t\boldsymbol{b}_n \tag{26}$$

where $\boldsymbol{w}$ is the unknown parameter vector. According to (13), $\boldsymbol{w}$ is given by

$$\boldsymbol{w} = (\boldsymbol{B}^t\boldsymbol{B})^{-1}\boldsymbol{B}^t\boldsymbol{d} \tag{27}$$

where $\boldsymbol{B} = \boldsymbol{\Phi}\boldsymbol{V} = [\boldsymbol{b}_0,\ldots,\boldsymbol{b}_{S-1}]^t$. Combining (25) and (26) leads to

$$y_n = \boldsymbol{w}^t\boldsymbol{A}^t\boldsymbol{k}_n = \boldsymbol{\alpha}^t\boldsymbol{k}_n. \tag{28}$$

Note that $\boldsymbol{B}^t\boldsymbol{B}$ is diagonal: $\mathrm{diag}(M\lambda_0,\ldots,M\lambda_{S-1})$. In the case $S = M$, the mapped functions $\{\phi(\boldsymbol{x}_i)\}_{i=0}^{M-1}$ do not undergo any dimensionality reduction and the solution is identical to ordinary kernel-based Wiener filtering. However, to avoid numerical problems, the smallest eigenvalues $M\lambda_i$ of $\boldsymbol{K}$ should be discarded. Procedures for the selection of S can be found in [9] and [52]–[55]. Holdout cross-validation will be used in this paper.

B. KPLS

Like KPCA, KPLS projects the data $\phi(\boldsymbol{x}_n)$ onto a lower dimensional space to reduce the set of possible solutions. There is, however, one essential difference with KPCA: in KPCA, the projection is determined without reference to the desired response, whereas in KPLS, the observed response plays an important role. The KPLS method has been developed in [28] for multiple output variables. In this paper, we derive KPLS for one single output variable. The KPLS problem can be stated as

$$\max_{\boldsymbol{v}} \mathrm{cov}\left(\langle\phi(\boldsymbol{x}_n),\boldsymbol{v}\rangle_{\mathcal{H}}, d_n\right), \quad \text{such that } \langle\boldsymbol{v},\boldsymbol{v}\rangle_{\mathcal{H}} = 1 \tag{29}$$

which yields

$$\min_{\boldsymbol{v}}(-\boldsymbol{v}^t\boldsymbol{\Phi}^t\boldsymbol{d}), \quad \text{such that } \langle\boldsymbol{v},\boldsymbol{v}\rangle_{\mathcal{H}} = 1. \tag{30}$$

This constrained optimization problem can be solved by searching the saddle point of the Lagrangian function

$$h(\boldsymbol{v},\gamma) = -\boldsymbol{v}^{\boldsymbol{t}}\boldsymbol{\Phi}^t\boldsymbol{d} + \gamma\left(\langle\boldsymbol{v},\boldsymbol{v}\rangle_{\mathcal{H}} - 1\right) \tag{31}$$

where γ denotes the Lagrange multiplier. Differentiating with respect to $\boldsymbol{v}$ and setting the result to zero leads to

$$\boldsymbol{\Phi}^t\boldsymbol{d} = 2\gamma\boldsymbol{v}. \tag{32}$$

Considering the constraint $\langle\boldsymbol{v},\boldsymbol{v}\rangle_{\mathcal{H}} = 1$, the solution is given by

$$\boldsymbol{v} = \frac{\boldsymbol{\Phi}^t\boldsymbol{d}}{\sqrt{\boldsymbol{d}^t\boldsymbol{\Phi}\boldsymbol{\Phi}^t\boldsymbol{d}}} = \frac{\boldsymbol{\Phi}^t\boldsymbol{d}}{\sqrt{\boldsymbol{d}^t\boldsymbol{K}\boldsymbol{d}}}. \tag{33}$$

The numerator in (33) depends explicitly on $\boldsymbol{\Phi}$. However, as will be shown, the determination of the filter parameters may be achieved without explicit knowledge of $\boldsymbol{v}$. The quantity required for this purpose is

$$\boldsymbol{t} = \boldsymbol{\Phi}\boldsymbol{v} = \frac{\boldsymbol{K}\boldsymbol{d}}{\sqrt{\boldsymbol{d}^t\boldsymbol{K}\boldsymbol{d}}} \tag{34}$$

which is the first PC of $\boldsymbol{\Phi}$. The remaining PCs may be retrieved in a recursive manner as follows: we seek directions in $\mathcal{H}$ so as to provide a low-rank approximation of $\boldsymbol{\Phi}$. Let $\boldsymbol{\Phi}_0 = \boldsymbol{\Phi}$, $\boldsymbol{d}_0 = \boldsymbol{d}$, $\boldsymbol{K}_0 = \boldsymbol{K}$, and $\boldsymbol{v}_0 = \boldsymbol{v}$. In a least square sense, the best rank-one approximation of $\boldsymbol{\Phi}_0$ is given by

$$\hat{\boldsymbol{\Phi}}_0 = \boldsymbol{t}_0\boldsymbol{p}_0^t \tag{35}$$

where $\boldsymbol{p}_0$ is determined so as to minimize $\|\boldsymbol{\Phi}_0 - \boldsymbol{t}_0\boldsymbol{p}_0^t\|^2$ with respect to $\boldsymbol{p}_0$ and therefore

$$\boldsymbol{p}_0 = \frac{\boldsymbol{\Phi}_0^t\boldsymbol{t}_0}{\boldsymbol{t}_0^t\boldsymbol{t}_0}. \tag{36}$$

Likewise, the desired output vector $\boldsymbol{d}_0$ is approximated by

$$\hat{\boldsymbol{d}}_0 = \boldsymbol{t}_0 c_0 \tag{37}$$

where c_0 solves $\min_{c_0}\|\boldsymbol{d}_0 - \boldsymbol{t}_0 c_0\|^2$. It is given by

$$c_0 = \frac{\boldsymbol{d}_0^t\boldsymbol{t}_0}{\boldsymbol{t}_0^t\boldsymbol{t}_0}. \tag{38}$$

To determine the next vector $\boldsymbol{t}_0$, we deflate the matrices $\boldsymbol{\Phi}_0$ and $\boldsymbol{d}_0$ according to

$$\begin{aligned}\boldsymbol{\Phi}_1 &= \boldsymbol{\Phi}_0 - \boldsymbol{t}_0\boldsymbol{p}_0^t\\ \boldsymbol{d}_1 &= \boldsymbol{d}_0 - \boldsymbol{t}_0 c_0\end{aligned}$$

and then apply (34). The deflation of $\boldsymbol{\Phi}_0$ affects $\boldsymbol{K}_0$ as follows:

$$\begin{aligned}\boldsymbol{K}_1 &= \left(\boldsymbol{\Phi}_0 - \boldsymbol{t}_0\boldsymbol{p}_0^t\right)\left(\boldsymbol{\Phi}_0 - \boldsymbol{t}_0\boldsymbol{p}_0^t\right)^t &(39)\\ &= \left(\boldsymbol{I} - \frac{\boldsymbol{t}_0\boldsymbol{t}_0^t}{\boldsymbol{t}_0^t\boldsymbol{t}_0}\right)\boldsymbol{K}_0\left(\boldsymbol{I} - \frac{\boldsymbol{t}_0\boldsymbol{t}_0^t}{\boldsymbol{t}_0^t\boldsymbol{t}_0}\right). &(40)\end{aligned}$$

The last equality is obtained by substituting $\boldsymbol{p}_0$ with its expression in (36). After extracting S PCs, $\boldsymbol{\Phi}$ and $\boldsymbol{d}$ are approximated by

$$\hat{\boldsymbol{\Phi}} = \sum_{i=0}^{S-1}\boldsymbol{t}_i\boldsymbol{p}_i^t = \boldsymbol{T}\boldsymbol{P}^t \tag{41}$$

$$\hat{\boldsymbol{d}} = \sum_{i=0}^{S-1}\boldsymbol{t}_i c_i = \boldsymbol{T}\boldsymbol{c}. \tag{42}$$

For selecting the appropriate number of PCs, one may refer to [52]–[55]. In this paper, as for KPCA, holdout cross-validation will be used. The next step is to determine the Wiener filter in the reduced space spanned by the vectors $\boldsymbol{V}^t\phi(\boldsymbol{x}_n)$, where

$\boldsymbol{V} = [\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{S-1}]$. For notation convenience, we drop the hat from $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{d}}$. The solution is given by

$$\boldsymbol{w} = (\boldsymbol{V}^t\boldsymbol{\Phi}^t\boldsymbol{\Phi}\boldsymbol{V})^{-1}\boldsymbol{V}^t\boldsymbol{\Phi}^t\boldsymbol{d}. \tag{43}$$

Combining (41) and (42), (43) translates to

$$\boldsymbol{w} = (\boldsymbol{P}^t\boldsymbol{V})^{-1}\boldsymbol{c}. \tag{44}$$

The matrix $(\boldsymbol{P}^t\boldsymbol{V})^{-1}$ is an upper triangular matrix and is always invertible [49]. However, $\boldsymbol{P}$ and $\boldsymbol{V}$ are directly related to $\boldsymbol{\Phi}$ and cannot be computed [see (33) and (36)]. To overcome this problem, we may write [56]

$$\boldsymbol{P} = \boldsymbol{\Phi}^t\boldsymbol{T}(\boldsymbol{T}^t\boldsymbol{T})^{-1} \tag{45}$$

$$\boldsymbol{c} = (\boldsymbol{T}^t\boldsymbol{T})^{-1}\boldsymbol{T}^t\boldsymbol{d} \tag{46}$$

$$\boldsymbol{V} = \boldsymbol{\Phi}^t\boldsymbol{D}\boldsymbol{F} \tag{47}$$

where $\boldsymbol{D} = [\boldsymbol{d}_0, \ldots, \boldsymbol{d}_{S-1}]$ and $\boldsymbol{F}$ is the diagonal matrix with element $F_{ii} = (1/\sqrt{\boldsymbol{d}_i^t\boldsymbol{K}_i\boldsymbol{d}_i})$, for $i = 0, \ldots, S-1$. We then deduce

$$\boldsymbol{w} = (\boldsymbol{T}^t\boldsymbol{\Phi}\boldsymbol{\Phi}^t\boldsymbol{D}\boldsymbol{F})^{-1}\boldsymbol{T}^t\boldsymbol{d} \tag{48}$$

and therefore the regularized Wiener filter is defined as

$$y_n = \boldsymbol{w}^t\boldsymbol{V}^t\phi(\boldsymbol{x}_n) = \boldsymbol{\alpha}^t\boldsymbol{k}_n \tag{49}$$

where $\boldsymbol{\alpha} = \boldsymbol{D}(\boldsymbol{T}^t\boldsymbol{K}\boldsymbol{D})^{-1}\boldsymbol{T}^t\boldsymbol{d}$. The last equality results from (47). It appears that the filter output does not explicitly depend on $\boldsymbol{\Phi}$. The KPLS algorithm proceeds as follows.

1) Define $\boldsymbol{K}_0 = \boldsymbol{K}$ and $\boldsymbol{d}_0 = \boldsymbol{d}$.
2) For $i = 0$ to $S-1$, compute
   - $\boldsymbol{t}_i = (\boldsymbol{K}_i\boldsymbol{d}_i/\sqrt{\boldsymbol{d}_i^t\boldsymbol{K}_i\boldsymbol{d}_i})$.
   - $c_i = (\boldsymbol{d}_i^t\boldsymbol{t}_i/\boldsymbol{t}_i^t\boldsymbol{t}_i)$.
   - $\boldsymbol{K}_{i+1} = (\boldsymbol{I} - (\boldsymbol{t}_i\boldsymbol{t}_i^t/\boldsymbol{t}_i^t\boldsymbol{t}_i))\boldsymbol{K}_i(\boldsymbol{I} - (\boldsymbol{t}_i\boldsymbol{t}_i^t/\boldsymbol{t}_i^t\boldsymbol{t}_i))$, $\boldsymbol{d}_{i+1} = \boldsymbol{d}_i - \boldsymbol{t}_ic_i$.
3) compute $\boldsymbol{\alpha} = \boldsymbol{D}(\boldsymbol{T}^t\boldsymbol{K}\boldsymbol{D})^{-1}\boldsymbol{T}^t\boldsymbol{d}$.

C. KRR

The main idea behind KRR is that the filter is determined by minimizing a cost function that includes the sum-square error function and a quadratic penalty term. The first term enforces closeness to the data, while the second ensures smoothness of the solution. A regularization parameter controls the tradeoff between these two antagonistic terms. RR in the feature space can be stated as follows:

$$\min_{\psi}\left(\sum_{i=0}^{M-1} e_i^2 + \delta\|\psi\|^2\right), \quad \text{such that } d_i - \langle\psi, \phi(\boldsymbol{x}_i)\rangle_{\mathcal{H}} = e_i, \quad i = 0, \ldots, M-1 \tag{50}$$

where δ is the regularization parameter $(\delta > 0)$. The Lagrangian function is provided by

$$h(\psi, \boldsymbol{e}, \boldsymbol{\alpha}) = \sum_{i=0}^{M-1} e_i^2 + \delta\|\psi\|^2 + \sum_{i=0}^{M-1} \alpha_i\left(d_i - \langle\psi, \phi(\boldsymbol{x}_i)\rangle_{\mathcal{H}} - e_i\right) \tag{51}$$

where α_i are the Lagrange multipliers. It follows from the saddle point condition that the partial derivatives of h with respect to the primal variables ψ and e_i have to vanish for optimality. This yields [29], [30]

$$\psi = \frac{1}{2\delta}\sum_{i=0}^{M-1}\alpha_i\phi(x_i) \tag{52}$$

$$e_i = \frac{\alpha_i}{2}. \tag{53}$$

Resubstituting these equations in (51) and rewriting this equation in matrix form leads to the following dual problem:

$$\max_{\boldsymbol{\alpha}} h(\boldsymbol{\alpha}) = \left(\boldsymbol{d}^t\boldsymbol{\alpha} - \frac{1}{4\delta}\boldsymbol{\alpha}^t\boldsymbol{K}\boldsymbol{\alpha} - \frac{1}{4}\boldsymbol{\alpha}^t\boldsymbol{\alpha}\right). \tag{54}$$

Differentiating with respect to $\boldsymbol{\alpha}$ and setting the result to zero, we obtain

$$\boldsymbol{\alpha} = 2\delta(\boldsymbol{K} + \delta\boldsymbol{I})^{-1}\boldsymbol{d} \tag{55}$$

and therefore

$$y(\boldsymbol{x}_n) = \langle\psi, \phi(\boldsymbol{x}_n)\rangle_{\mathcal{H}} = \boldsymbol{d}^t(\boldsymbol{K} + \delta\boldsymbol{I})^{-1}\boldsymbol{k}_n. \tag{56}$$

D. Experiments on Simulated Data

To assess the performance of the proposed approach, we consider the nonlinear difference equation proposed in [57]

$$d_n = 0.5d_{n-1} + 0.3d_{n-1}x_{n-1} + 0.2x_{n-1} + 0.05d_{n-1}^2 + 0.6x_{n-1}^2 \tag{57}$$

where x_n and d_n are, respectively, the input and the desired output of the system. We assume that d_n is corrupted by an additive zero-mean white Gaussian noise with standard deviation equal to 0.06. The input x_n is generated according to a normal distribution with mean 0.2 and standard deviation 0.1. Three implicit data transformations were considered by selecting the Gaussian, the inhomogeneous polynomial, and the linear kernels, with $\alpha = 1$. The criterion used to measure the performance of our approach was

$$\text{NMSE} = \frac{\sum_{i=0}^{M-1}(d_i - y_i)^2}{\sum_{i=0}^{M-1} d_i^2} \tag{58}$$

which represents the normalized mean-square error. A low value of NMSE clearly indicates a good performance of the denoising process. A 3000-sample set was generated by iterating from the initial desired output $d_0 = 0.1$. It was split into three subsets, of 1000 samples each, for training, holdout cross-validation, and testing, respectively. The filter was designed using the data in the training set. The optimum filter parameters, i.e., the dimension of the input vector N, the kernel parameters σ and q, the number of PCs S, and the regularization parameter δ were determined so as to minimize NMSE on the cross-validation set. The filter performance was evaluated on the testing set.

Table II reports the results obtained with KPCA, KPLS, and KRR, respectively. It indicates the optimum filter parameters as well as the values of NMSE on the training, cross-validation, and testing sets. We observe that the three methods considered here exhibit similar good performances on the testing set for the Gaussian and polynomial kernels. However, in terms of number

TABLE II
COMPARISON RESULTS FOR KPCA, KPLS, AND KRR ON SIMULATED DATA

	Gaussian kernel			Polynomial kernel			Linear kernel		
	KPCA	KPLS	KRR	KPCA	KPLS	KRR	KPCA	KPLS	KRR
Parameters	$N=5$ $\sigma=195$ $S=22$	$N=5$ $\sigma=177$ $S=14$	$N=5$ $\sigma=187$ $\delta=9\times10^{-13}$	$N=5$ $q=2$ $S=15$	$N=5$ $q=2$ $S=4$	$N=5$ $q=2$ $\delta=8\times10^{-12}$	$N=8$ $S=8$	$N=6$ $S=3$	$N=6$ $\delta=5\times10^{-8}$
NMSE	KPCA	KPLS	KRR	KPCA	KPLS	KRR	KPCA	KPLS	KRR
Training set	0.480	0.479	0.483	0.491	0.484	0.482	0.598	0.609	0.609
Cross-validation set	0.516	0.514	0.515	0.512	0.513	0.516	0.621	0.624	0.624
Testing set	0.527	0.527	0.529	0.523	0.529	0.529	0.640	0.632	0.632

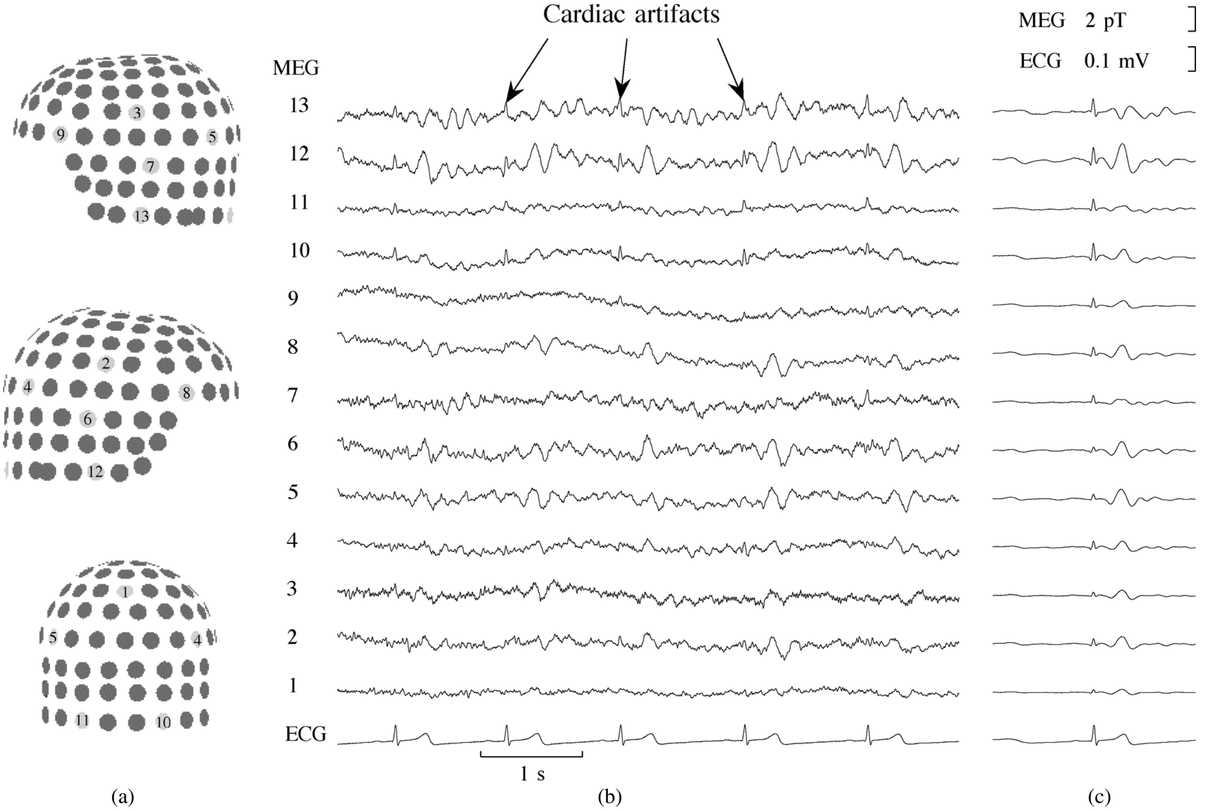

Fig. 2. (a) Subset of sensors locations. (b) MEG signals. The simultaneous recorded ECG is shown at the bottom. (c) The averaged signals in synchrony with the R wave of the ECG.

of selected PCs, we notice that KPLS requires fewer PCs than KPCA. This could be expected since KPLS takes into consideration the desired output to determine the projection in the reduced-space. Comparing the three kernels, we can see that both Gaussian and polynomial kernels are equally efficient and produce an NMSE on the testing set in the range 52.3–52.9%. They clearly outperform the linear kernel.

VI. APPLICATION TO MEG SIGNALS DENOISING

The magnetic field generated by the brain neuronal activity is extremely weak, varying from 0.01 to a few picotesla (pT). The cardiac magnetic field that reaches a few hundred picotesla is much stronger and can severely alter MEG measurements. In [1], it has been argued that cardiac artifacts observed in MEG recordings are mainly produced by the electrical heart activity. The contribution of blood pulsations is insignificant. Cardiac artifacts can be identified by recording the ECG signal. In Fig. 2(b), a subset of MEG channels measured from different locations of the scalp (a) is shown with the simultaneously recorded ECG. The sharp R wave of the ECG is clearly visible on MEG data. Fig. 2(c) shows the averaged signals in synchrony with the R wave of the ECG. The most contaminated channels are located in the temporal area.

A. Data Acquisition

MEG measurements were performed in a magnetically shielded room, using a whole-head MEG system (BTi Magnes 2500 WH) with 148 sensors. ECG was simultaneously recorded as an external channel. Each MEG sensor consists of a magnetometer coupled to a superconducting quantum interference device (SQUID). SQUIDs are immersed in liquid helium at a temperature of −269 °C. They convert the magnetic flux to voltage. Magnetometers are more sensitive than gradiometers and allow the measurement of signals generated by deep sources in the brain. Since they are very sensitive to noise, a set of 11 reference sensors (5 gradiometers and 6 magnetometers) located far enough from the scalp is used to detect the environmental noise. The subjects were recorded eyes closed in different vigilance states (awake and sleep), in a lying position, the head centered touching the inner back of the MEG helmet.

TABLE III
COMPARISON RESULTS FOR KPCA, KPLS, AND KRR ON MEG DATA

	Gaussian kernel			Polynomial kernel			Linear kernel		
	KPCA	KPLS	KRR	KPCA	KPLS	KRR	KPCA	KPLS	KRR
Parameters	$N = 12$ $\sigma = 0.05$ $S = 14$	$N = 11$ $\sigma = 0.048$ $S = 5$	$N = 7$ $\sigma = 0.11$ $\delta = 0.086$	$N = 7$ $q = 60$ $S = 21$	$N = 6$ $q = 61$ $S = 14$	$N = 6$ $q = 64$ $\delta = 0.016$	$N = 8$ $S = 5$	$N = 8$ $S = 4$	$N = 8$ $\delta = 0.001$
C_p	KPCA	KPLS	KRR	KPCA	KPLS	KRR	KPCA	KPLS	KRR
Training set	0.799	0.792	0.810	0.813	0.811	0.811	0.859	0.859	0.858
Cross-validation set	0.848	0.852	0.860	0.861	0.860	0.862	0.905	0.905	0.905
Testing set	0.866	0.867	0.884	0.887	0.885	0.890	0.915	0.915	0.915

MEG and ECG signals were digitized at a 254.31 Hz sampling frequency and bandpass filtered [0.1; 50 Hz].

B. *Kernel-Based Wiener Filtering*

We applied our method on MEG data highly corrupted by ECG, recorded from the sensor at position 13 in Fig. 2. The recorded ECG was used as the reference signal (input of the kernel regularized Wiener filter). The MEG signal was used as the desired output, so the residue is the denoised MEG signal. As a measure of performance, we used the normalized mean-square error (58). Three databases of 3000 samples were constituted for training, holdout cross-validation, and testing, respectively. The optimum filter parameters were determined by minimizing NMSE on the cross-validation set. Table III shows the results obtained for the three kernels. We notice that the Gaussian kernel achieves the best performance whereas the linear kernel yields the lowest performance. This result is observed for all methods KPCA, KPLS, and KRR. The minimum value of NMSE on the testing set is reached with KPCA. In order to verify whether the observed differences in NMSEs for nonlinear and linear kernels are statistically significant, we performed the one-sided tests

$$\mathrm{T}_1 \quad \mathrm{H}_1 : \mathrm{NMSE}_{\mathrm{gaus}} = \mathrm{NMSE}_{\mathrm{lin}}$$
$$\mathrm{H}_2 : \mathrm{NMSE}_{\mathrm{gaus}} < \mathrm{NMSE}_{\mathrm{lin}}$$

and

$$\mathrm{T}_2 \quad \mathrm{H}_1 : \mathrm{NMSE}_{\mathrm{poly}} = \mathrm{NMSE}_{\mathrm{lin}}$$
$$\mathrm{H}_2 : \mathrm{NMSE}_{\mathrm{poly}} < \mathrm{NMSE}_{\mathrm{lin}}$$

of comparison of NMSEs which, considering the definition of NMSE, are equivalent to tests of comparison of the estimated variances of the residues. Under some technical conditions (independence and gaussianity of the residue), and under the null hypothesis, it can be shown that the probability density function of the statistic F, defined as the ratio between two NMSEs, is central Fisher with $M - N$ degrees of freedom for the numerator and the denominator. If the level of significance is chosen to be equal to 0.1, then the hypothesis H_2 is accepted if the observed F ratio is smaller than the critical value $F_{0.1}(2988, 2992) = 0.954$. For the Gaussian kernel, the corresponding value of F when KPCA is used is $F_{obs} = 0.866/0.915 = 0.946$. Since F_{obs} falls in the region of rejection of H_1 for a 0.1-level test, the hypothesis H_1 should be discarded. This confirms that the Gaussian kernel yields better performance than the linear kernel. Regarding T_2, no significant differences between polynomial and linear kernels were found. Likewise, we resorted to F statistic in order to determine whether NMSEs are significantly different across KPCA, KPLS, and KRR. The results indicated that there is no significant differences between the performance of the three methods. Concerning the number of PCs, and in agreement with what was stated previously, we notice that KPLS entails an NMSE almost equal to that obtained with KPCA, with significantly fewer PCs. Figs. 3 and 4 present KPLS results for the Gaussian and linear kernels. The highest two curves show

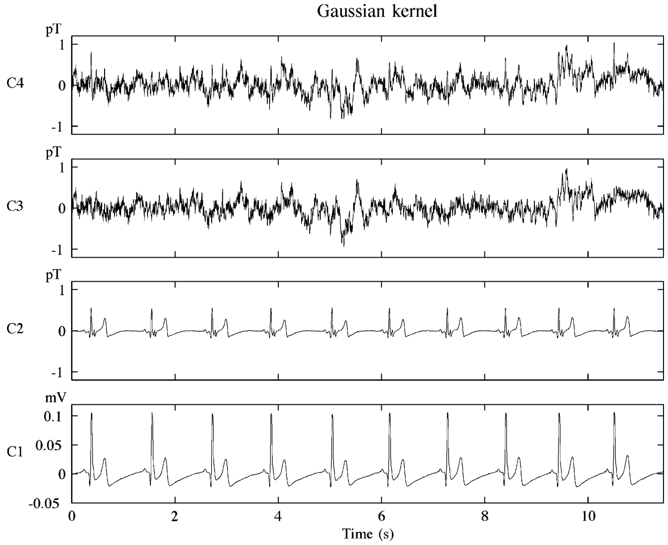

Fig. 3. KPLS results with Gaussian kernel. C1: ECG reference signal. C2: estimated ECG contribution in MEG signal. C3: denoised MEG. C4: corrupted MEG.

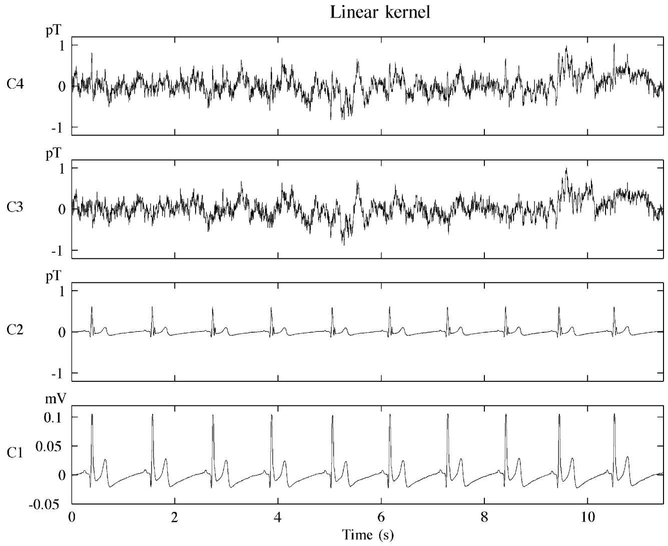

Fig. 4. KPLS results with linear kernel. C1: ECG reference signal. C2: estimated ECG contribution in MEG signal. C3: denoised MEG. C4: corrupted MEG.

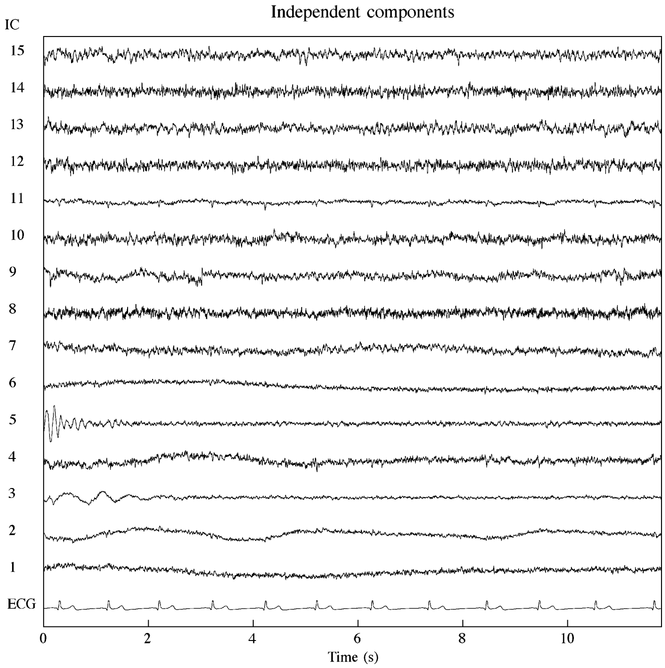

Fig. 5. The first 15 independent components determined by TDSEP. The ECG reference signal is shown at the bottom. IC 11 can be readily attributed to ECG source.

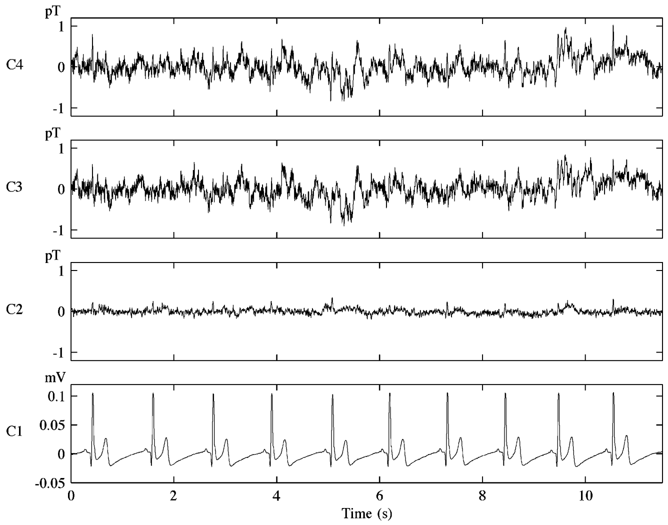

Fig. 6. TDSEP output on the testing set. C1: ECG reference signal. C2: estimated ECG contribution in MEG signal. C3: denoised MEG. C4: corrupted MEG.

the contaminated and the denoised MEG signals. Curves C1 and C2 show the ECG reference channel and the estimated artifact signal.

C. Comparison With ICA

In order to illustrate the competitiveness of our approach, it was compared to state-of-the-art ICA method, which is commonly used in filtering brain signals. ICA is an unsupervised technique that decomposes the measured data into statistically independent components. Among various algorithms devised for solving the ICA problem, we used TDSEP [11], [23], [24], which is based on the joint diagonalization of several time-lagged correlation matrices, defined as

$$\boldsymbol{R_z}(\tau) = E\left\{\boldsymbol{z}_n \boldsymbol{z}_{n-\tau}^t\right\} \tag{59}$$

where $\boldsymbol{z}_n$ denotes the observed data vector at time n and τ is some lag constant. In [23], the authors have shown that the use of multiple time delays improves TDSEP performance. The simultaneous diagonalization is achieved by first whitening the data using PCA and then performing a number of orthogonal transformations or Jacobi rotations. TDSEP is unable to separate signals that have identical spectra. It has been pointed out in [58] that the time delay operation corresponds to filtering with a sinusoidal comb filter whose comb finger distance is inversely proportional to the time delay. This implies that using larger time delays yields higher frequency resolution. In this application, we used the same experimental conditions as previously. In order to mitigate the overlearning effect which is particularly severe in MEG applications [14], [59], a dimensionality reduction was applied to the high-dimensional MEG data during the whitening stage by discarding the smallest eigenvalues of $\boldsymbol{R_z}(0)$. The best results were obtained with $\tau = [0, \ldots, 600]$ and 21 PCs. They are reported in Figs. 5 and 6. Fig. 5 displays the first 15 independent components. Component IC 11 reflects the cardiac contamination present in MEG recordings. Fig. 6 depicts the output of TDSEP on the testing set. We notice that the resulting performances of TDSEP are quite unsatisfactory when compared to the results obtained with Wiener filtering. This is confirmed by the value of NMSE on the testing set, which is equal to 0.958. This result can be explained by the blind nature of ICA and the lack of information about ECG source. We conclude that, when an ECG recorded signal is available, Wiener filtering techniques are more convenient for filtering MEG data than ICA. Otherwise, ICA should be applied. Within this context, a nonlinear kernel version of TDSEP [60] could be used to remove cardiac artifacts from MEG signals.

VII. Conclusion

Kernel-based methods have become a standard tool in data modelling. The purpose of using kernels is to avoid explicit mapping in a high-dimensional feature space for solving nonlinear problems. In this paper, we presented an efficient computational approach to nonlinear Wiener filtering problem based on kernels. We showed that the Wiener–Hopf equation is solved by the resolution of a linear system which may suffer from ill conditioning. To overcome this problem, we proposed three kernel-based regularization methods: KPCA, KPLS, and KRR. We applied our method to the problem of cardiac artifacts reduction from MEG data. A nonlinear model constructed with Gaussian kernel outperformed the linear model. Finally, a comparison with a current state-of-the-art method, ICA, was provided. ICA adopts a different approach for solving the filtering problem, based on the assumption that the underlying components of the measured data are statistically independent. We showed that Wiener filtering induces significantly better performance than ICA and should be used if an ECG reference signal is available.

This paper leaves much room for further improvements. We considered in the experiments three kernels, which have shown outstanding performance in many problems: the Gaussian, polynomial, and linear kernels. The choice of a particular kernel was determined by minimizing NMSE on the validation set. Other kernel selection methods have been proposed in the literature [61]–[65], which can be used in many kernel algorithms. This subject deserves further in-depth studies in order to fully exploit the potential of kernel methods. On the other hand, the kernel-based Wiener filter discussed in this paper requires stationarity of the data to be processed. In the nonstationary case, adaptive algorithms should be used in order to update the parameters of the filter when a new observation is available. In the literature, there exist a wide variety of linear adaptive filtering schemes such as the least mean square and the recursive least squares algorithms. Nevertheless, the extension of these algorithms to their kernel counterparts is a very challenging task since, as is suggested by the representer theorem, the number of kernel functions in the filter output grows linearly with the number of observations. Recently, attempts have been made to defeat the problem [66]–[68]. Another important direction of research would be to investigate this problem in depth in order to extend kernel-based Wiener filtering to the nonlinear adaptive case.

REFERENCES

[1] V. Jousmäki and R. Hari, "Cardiac artifacts in magnetoencephalogram," *J. Clin. Neurophysiol.*, vol. 13, no. 2, pp. 172–176, Mar. 1996.

[2] S. A. Hillyard and R. Galambos, "Eye-movement artifact in the CNV," *Electroenceph. Clin. Neurophysiol.*, vol. 28, no. 2, pp. 173–182, 1970.

[3] R. Verleger, T. Gasser, and J. Möcks, "Correction of EOG artifacts in event-related potentials of EEG: Aspects of reliability and validity," *Psychophysiology*, vol. 19, no. 4, pp. 472–480, 1982.

[4] J. L. Whitton, F. Lue, and H. Moldofsky, "A spectral method for removing eye-movement artifacts from the EEG," *Electroenceph. Clin. Neurophysiol.*, vol. 44, pp. 735–741, 1978.

[5] J. C. Woestenburg, M. N. Verbaten, and J. L. Slangen, "The removal of eye-movement artifact from the EEG by regression analysis in the frequency domain," *Biol. Psychol.*, vol. 16, pp. 127–147, 1983.

[6] M. Huotilainen, R. J. Ilmoniemi, H. Tiitinen, J. Lavaikainen, K. Alho, M. Kajola, and R. Näätänen, "The projection method in removing eye-blink artefacts from multichannel MEG measurements," in *Biomagnetism: Fundamental Research and Clinical Applications*, C. Baumgartner, L. Deecke, G. Stroink, and S. J. Williamson, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 363–367.

[7] M. A. Uusitalo and R. J. Ilmoniemi, "The signal-space projection (SSP) method for separating MEG or EEG into components," *Med. Biol. Eng. Comput.*, vol. 35, pp. 135–140, 1997.

[8] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, pp. 559–572, 1901.

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

[10] P. K. Sadasivan and D. N. Dutt, "SVD based technique for noise reduction in electroencephalographic signals," *Signal Process.*, vol. 55, pp. 179–189, 1996.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Interscience, 2001.

[12] Y. Wang, "Reduction of cardiac artifacts in magnetoencephalogram," in *Proc. 12th Int. Conf. Biomagn.*, Espoo, Finland, 2000.

[13] S. Makeig, A. J. Bell, and T.-P. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 145–151.

[14] R. Vigário and E. Oja, "Independence: A new criterion for the analysis of the electromagnetic fields in the global brain?," *Neural Netw.*, vol. 13, pp. 891–907, 2000.

[15] A. K. Barros, R. Vigário, V. Jousmäki, and N. Ohnishi, "Extraction of event-related signals from multichannel bioelectrical measurements," *IEEE Trans. Biomed. Eng.*, vol. 47, pp. 583–588, May 2000.

[16] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, pp. 589–593, May 2000.

[17] L. Vigon, M. R. Saatchi, J. E. W. Mayhew, and R. Fernandes, "Quantitative evaluation of techniques for ocular artefact filtering of EEG waveforms," *Proc. Inst. Elect. Eng. Sci. Meas. Technol.*, vol. 147, no. 5, pp. 219–228, Sept. 2000.

[18] T. H. Sander, G. Wübbeler, A. Lueschow, G. Curio, and L. Trahms, "Cardiac artifact subspace identification and elimination in cognitive MEG data using time-delayed decorrelation," *IEEE Trans. Biomed. Eng.*, vol. 49, pp. 345–354, Apr. 2002.

[19] K.-R. Müller, R. Vigário, F. Meinecke, and A. Ziehe, "Blind source separation techniques for decomposing event-related brain signals," *Int. J. Bifurcat. Chaos*, vol. 14, no. 2, pp. 773–791, 2004.

[20] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. Inst. Elect. Eng. F*, vol. 140, no. 6, pp. 362–370, 1993.

[21] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[22] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, pp. 1483–1492, 1997.

[23] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[24] A. Ziehe and K.-R. Müller, "TDSEP—An effective algorithm for blind separation using time structure," in *Proc. Int. Conf. Artif. Neural Netw.*, Skövde, Sweden, 1998, pp. 675–680.

[25] V. Tresp, I. Leuthausser, M. Schlang, R. Neuneier, K. Abraham-Fuchs, and W. Harer, "An efficient model for systems with complex responses," in *Neural Networks Signal Process. II—Proc. 1992 IEEE Signal Process. Workshop*, Helsingoer, Denmark, 1992, pp. 493–502.

[26] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks.*, vol. 12, pp. 181–201, Mar. 2001.

[27] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[28] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, 2001.

[29] N. Cristianini and J. Shawe-Taylor*, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[30] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc. 15th Int. Conf. Machine Learning*, Madison, WI, 1998, pp. 515–521.

[31] S. Haykin*, Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[32] G. Palm and T. Poggio, "The Volterra representation and the Wiener expansion: Validity and pitfalls," *SIAM J. Appl. Math.*, vol. 33, pp. 195–216, 1977.

[33] M. Schetzen*, The Volterra and Wiener Theories of Nonlinear Systems*. New York: Wiley, 1980.

[34] W. J. Rugh*, Nonlinear System Theory*. Baltimore, MD: Johns Hopkins Univ. Press, 1981.

[35] V. N. Vapnik*, The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[36] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks Signal Processing. IX—Proc. 1999 IEEE Signal Process. Workshop*, Madison, WI, 1999, pp. 41–48.

[37] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, Oct. 2000.

[38] F. Abdallah, C. Richard, and R. Lengellé, "Kernel second-order discriminants versus support vector machines," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, China, 2003, pp. 149–152.

[39] ——, "An improved training algorithm for nonlinear kernel discriminants," *IEEE Trans. Signal Process.*, vol. 52, pp. 2798–2806, Oct. 2004.

[40] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[41] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. New York: Interscience, 1953.

[42] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Trans. Roy. Soc. London*, vol. A 209, pp. 415–446, 1909.

[43] J. Stewart, "Positive definite functions and generalizations: An historical survey," *Rocky Mountain J. Math.*, vol. 6, pp. 409–434, 1976.

[44] T. Poggio, "On optimal nonlinear associative recall," *Biolog. Cybern.*, vol. 19, pp. 201–209, 1975.

[45] M. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. Mach. Learn. Res.*, pp. 299–312, 2001.

[46] R. Herbrich, *Learning Kernel Classifiers. Theory and Algorithms, Cambridge*. Cambridge, MA: MIT Press, 2002.

[47] G. S. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Math. Anal. Applicat.*, vol. 33, pp. 82–95, 1971.

[48] I. Constantin, C. Richard, R. Lengellé, and L. Soufflet, "Regularized kernel-based Wiener filtering: Application to magnetoencephalographic signals denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 289–292.

[49] A. Höskuldsson, "PLS regression methods," *J. Chemometrics*, vol. 2, pp. 211–218, 1988.

[50] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics Intel. Lab. Syst.*, vol. 58, pp. 109–130, 2001.

[51] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 3, pp. 55–67, 1970.

[52] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components analysis," *Technometrics*, vol. 20, pp. 397–405, 1978.

[53] M. Stone and R. J. Brooks, "Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression," *J. Roy. Statist. Soc. Ser. B*, vol. 52, no. 2, pp. 237–269, 1990.

[54] Q.-S. Xu and Y. Z. Liang, "Monte Carlo cross validation," *Chemometrics Intel. Lab. Syst.*, vol. 56, pp. 1–11, 2001.

[55] L. Stordrange, F. O. Libnau, D. Malthe-Sörenssen, and O. M. Kvalheim, "Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques," *J. Chemometrics*, vol. 16, pp. 529–541, 2002.

[56] P. J. Lewi, "Pattern recognition, reflection from a chemometric point of view," *Chemometrics Intel. Lab. Syst.*, vol. 28, pp. 23–33, 1995.

[57] S. A. Billings and W. S. F. Voon, "Correlation-based model validity tests for non-linear models," *Int. J. Contr.*, vol. 44, pp. 235–244, 1986.

[58] T. H. Sander, M. Burghoff, G. Curio, and L. Trahms, "Single evoked somatosensory MEG responses extracted by time delayed decorrelation," *IEEE Trans. Signal Process.*, vol. 53, pp. 3384–3392, Sep. 2005.

[59] A. Hyvärinen, J. Särelä, and R. Vigário, "Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size," in *Proc. Int. Workshop Ind. Comp. Anal. Blind Separation Signals*, Aussois, France, 1999, pp. 425–429.

[60] S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller, "Kernel-based nonlinear blind source separation," *Neural Comput.*, vol. 15, pp. 1089–1124, 2003.

[61] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, "On kernel-target alignment," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2001, pp. 367–373.

[62] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," in *Proceedings of the International Conference on Machine Learning*. San Mateo, CA: Morgan Kaufmann, 2002, pp. 323–330.

[63] O. Bousquet and D. J. L. Herrmann, "On the complexity of learning the kernel matrix," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 399–406.

[64] K. Crammer, J. Keshet, and Y. Singer, "Kernel design using boosting," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 537–544.

[65] C. S. Ong, A. J. Smola, and R. C. Williamson, "Learning the kernel with hyperkernels," *J. Mach. Learn. Res.*, vol. 6, pp. 1043–1071, 2005.

[66] T. J. Dodd, V. Kadirkamanathan, and R. F. Harrison, "Function estimation in Hilbert space using sequential projections," in *Proc. IFAC Int. Conf. Intel. Contr. Syst. Signal Process.*, Algarve, Portugal, 2003, pp. 113–118.

[67] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Process.*, vol. 52, pp. 2165–2176, Aug. 2004.

[68] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Trans. Signal Process.*, vol. 52, pp. 2275–2285, Aug. 2004.

Ibtissam Constantin received the Dipl.-Ing. degree in electrical engineering and the M.S. degree in industrial control from the Faculty of Engineering, University of Lebanon, in 2000 and 2002, respectively. She is currently pursuing the Ph.D. degree at Troyes University of Technology, Troyes, France.

Her scientific interests are in the field of machine leaning and kernel methods.

Cédric Richard (M'01) was born in Sarrebourg, France, on January 24, 1970. He received the Dipl.-Ing. and M.S. degrees in 1994 and the Ph.D. degree in 1998 from Compiègne University of Technology, Compiègne, France, all in electrical and computer engineering.

From 1999 to 2003, he was an Associate Professor at Troyes University of Technology, Troyes, France. Since 2003, he has been a Professor at the Systems Modelling and Dependability Laboratory, Troyes University of Technology. His current research interests involve time-frequency analysis, statistical estimation and decision theories, and pattern recognition.

Régis Lengellé was born on March 30, 1958. He received the Dipl.-Ing, M.S., and Ph.D. degrees from Compiègne University of Technology, Compiègne, France, in 1980, 1981, and 1983 respectively, and the Habilitation à Diriger des Recherches from the University Henri Poincaré Nancy I, France, in 1994.

From 1985 to 1993, he was an Associate Professor at Compiègne University of Technology. Since 1994, he has been a Professor at Troyes University of Technology, Troyes, France, with research interests focused in signal processing, statistical decision theory, and pattern recognition with applications to systems monitoring.

Laurent Soufflet was born in Paris, France, in 1960. He received the M.S. and Ph.D. degrees in electronics from the University of Mulhouse, France, in 1984 and 1991, respectively.

His thesis was on the three-dimensional representation of the cerebral electrical activity. Since 1991, he has been with FORENAP, a research center dedicated to psychiatry and pharmacology, at the psychiatric hospital of Rouffach, Alsace, France. He is responsible for the Signal and Image Processing Department and is a MEG expert. His current research and development includes EEG/MEG signal processing, source localization, and three-dimensional and multimodality brain imaging techniques.