How to Apply Nonlinear Subspace Techniques to Univariate Biomedical Time Series

A. R. Teixeira, A. M. Tomé, *Member, IEEE*, M. Böhm, Carlos G. Puntonet, and Elmar W. Lang

***Abstract*—In this paper, we propose an embedding technique for univariate single-channel biomedical signals to apply projective subspace techniques. Biomedical signals are often recorded as 1-D time series; hence, they need to be transformed to multidimensional signal vectors for subspace techniques to be applicable. The transformation can be achieved by embedding an observed signal in its delayed coordinates. We propose the application of two nonlinear subspace techniques to embedded multidimensional signals and discuss their relation. The techniques consist of modified versions of singular-spectrum analysis (SSA) and kernel principal component analysis (KPCA). For illustrative purposes, both nonlinear subspace projection techniques are applied to an electroencephalogram (EEG) signal recorded in the frontal channel to extract its dominant electrooculogram (EOG) interference. Furthermore, to evaluate the performance of the algorithms, an experimental study with artificially mixed signals is presented and discussed.**

***Index Terms*—Electroencephalogram (EEG), electrooculogram (EOG), kernel principal component analysis (KPCA), local singular spectrum analysis (SSA), removing artifacts, subspace techniques.**

I. Introduction

In MANY biomedical signal processing applications, a sensor signal is contaminated with noise and artifact signals of substantial amplitude. The latter can sometimes be the most prominent signal component registered. Noise signals are often modeled as being additive, normally distributed, and uncorrelated with the signals of interest. Often, the signal-to-noise ratios (SNRs) are quite low. Hence, to recover the signals of interest, the task is to remove both the artifact-related components and the superimposed noise contributions.

With multidimensional signals, projective subspace techniques can then be favorably used to get rid of most of the noise contributions to the signals. However, many biomedical signals represent 1-D time series. Clearly, projective subspace techniques are not available for 1-D time series; hence, time series analysis techniques often rely on embedding a 1-D sensor signal in a high-dimensional space of time-delayed coordinates [1]–[3]. Correlations in these multidimensional signal vectors together with second-order techniques can be used to decompose the signal into uncorrelated components. The multidimensional signal is then projected to the most significant directions computed using singular value decomposition (SVD) of the data matrix $\mathbf{X}$ or principal component analysis (PCA) of the covariance matrix $\mathbf{C}$ or its related scatter matrix $\mathbf{S}$ [4].

Singular spectrum analysis (SSA) [5] used in climatic, meteorologic, and geophysics data analysis is the most widely used technique that follows this strategy. The general purpose of SSA is to decompose the embedded signal vectors into additive components. This decomposition can be used to separate noise contributions from a recorded signal by estimating those eigenvectors that span the signal subspace. These directions can be associated with the L largest eigenvalues of the eigendecomposition. As noise signals spread in all directions, the remaining orthogonal directions then only represent noise contributions. Reconstructing the signal using only those L dominant components can then result in a substantial noise reduction of the recorded signals.

The time embedding of the sensor signals transforms the 1-D time series into multidimensional signal vectors. This is a necessary step if subspace projection techniques are to be applied. However, this step often introduces nonlinearity into the signal analysis process. Of course, there also exist generically nonlinear signal processing techniques like kernel PCA (KPCA) [6], which is often used for denoising. Therefore, it will be of interest to explore these techniques in their ability to remove dominant artifacts and/or suppress noise. The kernel techniques are based on the mapping of the input data by a nonlinear function. Then, in feature space, a linear PCA is performed by estimating the eigenvectors and eigenvalues of a matrix of dot products (kernel matrix).

In this paper, we will present the concept of Local SSA, which means that after the time embedding, we cluster the resulting multidimensional signal vectors and apply the linear signal decomposition technique, i.e., SSA, only locally in each cluster [7]. However, as embedding can be regarded as a nonlinear signal manipulation, a nonlinear technique like KPCA should be even more appropriate [8]. To reduce the computational complexity, we present a variant of KPCA whose parameters are computed using the eigendecomposition of a low-rank approximation of the kernel matrix.

Manuscript received January 22, 2008; revised June 11, 2008. First published May 15, 2009; current version published July 17, 2009. The work of A. R. Teixeira was supported by a Ph.D. Scholarship (SFRH/BD/28404/2006) from the Portuguese Foundation for Science and Technology (FCT). The Associate Editor coordinating the review process for this paper was Dr. Jesús Ureña.

A. R. Teixeira and A. M. Tomé are with the Electronics, Telecommunications and Informatics Department (DETI), Instituto de Engenharia Electrónica e Telemática de Aveiro, Universidade de Aveiro, Aveiro 3810-193, Portugal (e-mail: ana@ieeta.pt).

M. Böhm and E. W. Lang are with CIMLG/Biophysics, University of Regensburg, 93040 Regensburg, Germany (e-mail: elmar.lang@biologie.uni-regensburg.de).

C. G. Puntonet is with ESTII, Departamento de Arquitectura y Tecnología de Computadores, University of Granada, 18071 Granada, Spain.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2009.2016385

The availability of digital electroencephalogram (EEG) recordings allows the study of procedures that try to remove the artifact contributions from the recorded brain signals. The primary goal will be to remove artifacts without distorting the underlying brain signals. Most of the works (as an example, see [9]) present solutions based on an analysis of multichannel recordings. In this paper, a single-channel approach is considered, and projective subspace techniques for denoising are applied. Hereby, artifact-related contributions to the recorded EEG signals will be identified as "the signal," and the actual EEG signal is considered a "sort of a broadband noise" to be separated. The philosophy behind is that artifact signals like electrooculograms (EOGs) are mostly the dominant signal contributions, much like real signals contaminated with noise. Consequently, we can use the projective subspace techniques referred to earlier to separate such artifacts from "pure" EEG signals.

II. Projective Subspace Techniques

Time-series analysis techniques often rely on embedding 1-D sensor signals in the space of their time-delayed coordinates. Embedding can be regarded as a mapping that transforms a 1-D time series $x = (x[0], x[1], \ldots, x[N-1])$ into a multidimensional sequence of $K = N - M + 1$ lagged vectors

$$\mathbf{x}_k = [x[k-1+M-1], \ldots, x[k-1]]^T, \quad k = 1, \ldots, K. \tag{1}$$

The lagged vectors $\mathbf{x}_k$ lie in a space of dimension M and constitute the columns of the *trajectory matrix* $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_K]$, $(N > M)$, i.e.,

$$\mathbf{X} = \begin{bmatrix} x[M-1] & x[M] & \cdots & x[N-1] \\ x[M-2] & x[M-1] & \cdots & x[N-2] \\ x[M-3] & x[M-2] & \cdots & x[N-3] \\ \vdots & \vdots & & \vdots \\ x[1] & x[2] & \cdots & x[N-M+1] \\ x[0] & x[1] & \cdots & x[N-M] \end{bmatrix}. \tag{2}$$

Note that the matrix has identical entries along its diagonals.

Any multidimensional signal $\mathbf{x}_k$ is projected onto the directions (eigenvectors) related to the largest eigenvalues of the covariance matrix or the related scatter matrix. The matrix can be computed in the input space (SSA or Local SSA) or after transforming the data by a nonlinear function (KPCA). The reconstruction (and the reversion of the nonlinearity for KPCA) using the same group of eigenvectors leads to $\hat{\mathbf{X}}$. Notice that, in general, the elements along each descending diagonal of $\hat{\mathbf{X}}$ will not be identical, like in case of the original trajectory matrix $\mathbf{X}$. This can be cured, however, by replacing the entries in each diagonal by their average, obtaining again a Toeplitz matrix $\mathbf{X}_r$. This procedure assures that the Frobenius norm of the difference $(\mathbf{X}_r - \hat{\mathbf{X}})$ attains its minimum value among all the possible solutions to get a matrix with all the diagonals equal [1].

The 1-D signal $\hat{x}[n]$ is then obtained by reverting the embedding, i.e., by forming the signal with the mean of the values along each descendent diagonal of $\hat{\mathbf{X}}$ [7]. Note that in the example considered later, if $\hat{x}[n]$ corresponds to the extracted EOG, then the corrected EEG is computed as $y[n] = x[n] - \hat{x}[n]$.

A. Local SSA

Local SSA basically introduces a clustering step into the SSA technique [7] and operates in input space. A normal SSA is obtained by skipping the clustering step, i.e., choosing $q = 1$. With Local SSA, after embedding, the column vectors $\mathbf{x}_k$, $k = 1, \ldots, K$, of the trajectory matrix are clustered using any clustering algorithm (like k-means [10]). After clustering, the set of indices of the columns of $\mathbf{X}$ is subdivided into q disjoint subsets $c_1, c_2, \ldots, c_q$. Thus, the subtrajectory matrix $\mathbf{X}^{(c_i)}$ is formed with N_{c_i} columns of the matrix $\mathbf{X}$, which belong to the subset c_i of indices. Note that the model parameter q is naturally bounded from above by the number of data available. However, any reliable estimate needs a sufficient number of data points in each cluster, limiting the number of clusters to be much less than the number of available data. The following steps 1)–4) need to be repeated for every $i = 1, \ldots, q$.

1) A covariance matrix is computed in each cluster using zero-mean data obtained via

$$\mathbf{X}_c = \mathbf{X}^{(c_i)} \left(\mathbf{I} - \frac{1}{N_{c_i}} \mathbf{j}_{c_i} \mathbf{j}_{c_i}^T \right) \tag{3}$$

where $\mathbf{j}_{c_i} = [1, 1, \ldots, 1]^T$ is a vector with dimension $N_{c_i} \times 1$, and $\mathbf{I}$ is a $N_{c_i} \times N_{c_i}$ identity matrix.

2) Next, the eigenvalue decomposition of the covariance matrix is computed, i.e.,

$$\mathbf{C}^{(c_i)} = \frac{1}{N_{c_i}} \mathbf{X}_c \mathbf{X}_c^T = \frac{1}{N_{c_i}} \mathbf{S}_c = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathbf{T}}. \tag{4}$$

Afterward, denoising can be achieved by projecting the multidimensional signal into the subspace spanned by the eigenvectors corresponding to the $L_{c_i} < M$ largest eigenvalues.

3) The number of significant directions can be found by using a maximum-likelihood estimation of the parameter vector of the covariance matrix $\mathbf{C}^{(c_i)}$ of each cluster. This parameter vector $\boldsymbol{\theta}$ comprises the most significant eigenvalues and corresponding eigenvectors and the variance of the noise, which is estimated by the average over the discarded eigenvalues. The number of relevant directions $k = L_{c_i}$ can be estimated using a minimum description length criterion. It results from the value that minimizes the following expression [11]:

$$MDL(k) = -L(\widehat{\boldsymbol{\theta}}) + \frac{1}{2} P \ln N, \quad k = 0, \ldots, M-1 \tag{5}$$

where $N = N_{c_i}$ is the number of observations available to estimate the covariance matrix, and $f(\mathbf{X}^{(c_i)}|\widehat{\boldsymbol{\theta}})$ denotes the conditional probability density parameterized by $\widehat{\boldsymbol{\theta}}$. This log-likelihood function $L(\widehat{\boldsymbol{\theta}}) = \ln f(\mathbf{X}^{(c_i)}|\widehat{\boldsymbol{\theta}})$ represents the accuracy of representation of the data

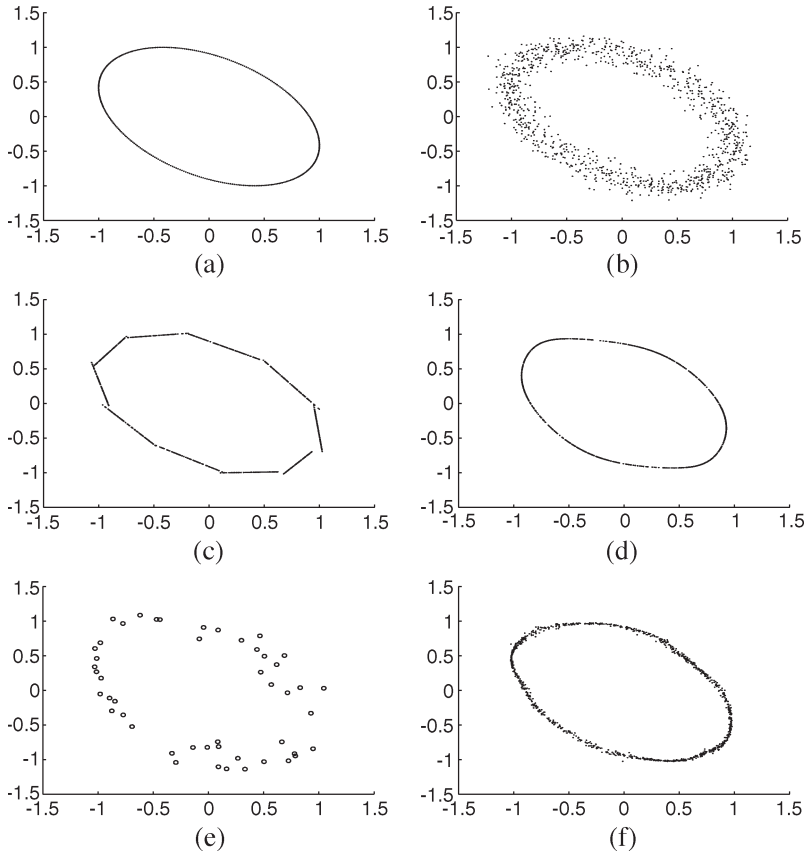

Fig. 1. Trajectory of (a) a sinusoid and (b) its noisy counterpart embedded in time-delayed coordinates $M = 2$. Graph (c) represents the Local SSA result, and graph (d) shows the corresponding trajectory obtained with KPCA. In graph (e), a subset of data points of the noisy sinusoid [graph (b)] is shown, which is used in Greedy KPCA [graph (f)]. (a) Sinusoid. (b) Sinusoid + noise. (c) Local SSA. (d) KPCA. (e) Subset R. (f) Greedy KPCA.

with the parameter vector and depends on the discarded eigenvalues, e.g.,

$$L(\widehat{\boldsymbol{\theta}}) = N(M-k)\ln\left[\frac{\prod_{i=k+1}^{M}\lambda_i^{1/(M-k)}}{\frac{1}{M-k}\sum_{i=k+1}^{M}\lambda_i}\right]. \tag{6}$$

The negative log-likelihood $-L(\widehat{\boldsymbol{\theta}})$ is recognized to be a standard measure of training error. However, it has been reported that the simple maximization of this term tends to result in the phenomenon of overfitting. Thus, the second term in (5) was added as a regularization term to penalize complexity. The value of P is related to the number of parameters in $\boldsymbol{\theta}$ and the complexity of its estimation. Considering real-valued signals, the value of P is computed according to

$$\begin{aligned} P &= k + 1 + Mk - k^2/2 - k/2 \\ &= -k^2/2 + k(M + 1/2) + 1. \end{aligned} \tag{7}$$

A simple alternative to this elaborate model-order selection is to fix the number of relevant directions instead. In some applications, even a single direction $L_{c_i} = 1$ suffices.

4) The eigenvectors related to the largest eigenvalues are used in the reconstruction process. Considering the matrix $\mathbf{U}$ with L_{c_i} eigenvectors in its columns, the reconstructed vectors in each cluster are obtained as

$$\hat{\mathbf{X}}^{(\mathbf{c_i})} = \mathbf{U}\mathbf{U}^T\mathbf{X}_c + \frac{1}{N_{c_i}}\mathbf{X}^{(c_i)}\mathbf{j}_{c_i}\mathbf{j}_{c_i}^T. \tag{8}$$

This reconstruction has to be separately done for each cluster.

The clustering is reverted by forming an estimate $\hat{\mathbf{X}}$ of the reconstructed noise-free trajectory matrix using the columns of the extracted subtrajectory matrices $\hat{\mathbf{X}}^{(c_i)}$, $i = 1, \ldots, q$, according to the contents of subsets c_i.

B. Illustrative Example

Fig. 1 illustrates the application of Local SSA to decompose a noisy sinusoid [Fig. 1(b)] into two components. The sinusoid, embedded with $M = 2$, has an elliptic trajectory in 2-D space, as shown in Fig. 1(a). Applying Local SSA using $q = 10$ clusters and projecting the data in each cluster onto the direction related to the largest eigenvector, after reconstruction, the 2-D trajectory of $\hat{\mathbf{X}}$ represents a piece-wise approximation of the original trajectory [see Fig. 1(c)]. Note that the projective subspace denoising that was globally applied, i.e., applying normal SSA, will result in a straight line that corresponds to the direction of maximum variance of the data, which would correspond to the long axis of the ellipsoid.

As the trajectory is inherently nonlinear, we will next consider a generically nonlinear projective subspace method, i.e., KPCA, which is a nonlinear extension of PCA. Note that unlike linear PCA, KPCA allows extracting a number of principal components that exceeds the dimensionality of the input data as the data are first mapped into a higher dimensional space. Notice that having $K \geq M$ examples of data with dimension M, working in input space, the maximum number of nonzero eigenvalues will also be M, as can be seen by computing either the covariance matrix or the matrix of dot products. In KPCA instead, the kernel matrix in the feature space will have a size of $K \times K$, and the number of nonzero eigenvalues can often be higher than M.

C. Subspace Projections and Kernel Matrices

In subspace methods, denoising is achieved by projecting the data onto basis vectors, as expressed in (8). Without loss of generality, let us consider that the datum $\mathbf{X}$ only forms one cluster $q = 1$, and that it is centered. Then, (8) simplifies to $\hat{\mathbf{X}} = \mathbf{U}\mathbf{U}^T\mathbf{X} = \mathbf{U}\mathbf{Z}$. The projections are then obtained as

$$\mathbf{Z} = \mathbf{U}^T\mathbf{X}. \tag{9}$$

As explained in the last section, the matrix of basis vectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_L]$ is formed with L eigenvectors of the covariance matrix or of a scatter matrix, which correspond to the L largest eigenvalues. However, the values of the projections can also be computed using the matrix of dot products, which is called the kernel matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. It has the same nonzero eigenvalues as the scatter matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$. Alternatively, considering an SVD of the data set and using $R > L$ nonzero singular values, we can write

$$\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^T \tag{10}$$

where $\mathbf{D}$ is a diagonal matrix with ordered eigenvalues $(\lambda_1 > \lambda_2 > \cdots > \lambda_L > \cdots > \lambda_R)$ of the kernel matrix $\mathbf{K}$ or of the scatter matrix $\mathbf{S}$, and $\mathbf{V}$ and $\mathbf{U}$ represent the R eigenvectors of the kernel and scatter matrices, respectively. Considering the SVD approximation using only the L most significant singular values and substituting them into (9), the L projections are obtained as

$$\mathbf{Z} = \mathbf{D}^{1/2}\mathbf{V}^T. \tag{11}$$

Then, the projections are related to the eigenvectors of the kernel matrix. Furthermore, the combination of (9) and (11) leads to $\mathbf{U}^T\mathbf{X} = \mathbf{D}^{1/2}\mathbf{V}^T$. Multiplying both sides of this equality by $\mathbf{V}\mathbf{D}^{-1/2}$, and considering that the columns of the eigenvector matrices are orthogonal, the basis vector matrix reads

$$\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1/2} = \mathbf{X}\mathbf{A}. \tag{12}$$

This relation shows that each eigenvector $\mathbf{u}_j$ of $\mathbf{U}$ can be represented as a linear combination of the data vectors $\mathbf{x}_k$. The coefficients of this linear combination form the components of the column vectors $\mathbf{a}_j$ of $\mathbf{A}$.

Note now that the projections used in projective subspace techniques can be expressed via dot products of data vectors. To see this, simply substitute (12) into (9), which yields

$$\mathbf{Z} = \mathbf{U}^T\mathbf{X} = \mathbf{A}^T\mathbf{X}^T\mathbf{X}. \tag{13}$$

The matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ is just the kernel matrix previously mentioned. Thus, whenever a problem can totally be phrased in terms of dot products, kernel methods can be applied.

D. Kernel Subspace Techniques

Kernel subspace techniques are projective methods in feature space created by a nonlinear transformation of the data. The data are mapped into a high (and possible infinite) dimensional space defined by a nonlinear function. However, the mapping into feature space is avoided by using kernel functions, which implicitly define a dot product in feature space computed using data in input space [6]. Then, every data manipulation (or every algorithm) can efficiently be computed as long as it can be translated into a sequence of dot products.

Consider again (13) and assume that the data have been mapped into a high-dimensional space by the mapping $\phi : \mathbf{x} \rightarrow \phi(\mathbf{x})$. The projections of the mapped data set $\mathbf{\Phi}$ are then obtained as

$$\mathbf{Z} = \mathbf{U}^T\mathbf{\Phi}. \tag{14}$$

Note that now the columns of the matrix $\mathbf{U}$ form a basis of feature space. This basis can as well be written as a linear combination of the mapped input data, i.e.,

$$\mathbf{U} = \mathbf{\Phi}_B\mathbf{A}. \tag{15}$$

With KPCA, the relation $\mathbf{\Phi}_B = \mathbf{\Phi}$ holds, whereas with Greedy KPCA, only a subset of the mapped data is considered, which yields $\mathbf{\Phi}_B = \mathbf{\Phi}_R$. Note that the column vectors $\mathbf{z}_j$ of $\mathbf{Z}$ depend on the dot products $\mathbf{\Phi}_B^T\phi(\mathbf{x}_j)$. However, to avoid an explicit mapping into feature space, all the data manipulations are achieved by dot products [6], and the kernel trick is applied. For instance, using an radial basis function (RBF) kernel, the dot product between a vector $\phi(\mathbf{x}_i)$, which belongs to subset B, and $\phi(\mathbf{x}_j)$ is computed using a kernel function that only depends on the input data, i.e.,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \tag{16}$$

To recover the noise-reduced signal, after denoising in feature space, the nonlinear mapping must be reverted (i.e., the pre-image in input space must be estimated).

Denoising using kernel methods, thus, comprises the following two steps after the computation of the projections in feature space:

1) the reconstruction in feature space;
2) the estimation of the preimage of the reconstructed point $\hat{\phi}(\mathbf{x}_j) = \mathbf{U}\mathbf{z}_j$, where $\mathbf{z}_j$ represents the projections of a noisy point $\mathbf{x}_j$.

These two steps can be joined together by minimizing the Euclidean distance of the image $\phi(\mathbf{p})$ of a yet unknown point $\mathbf{p}$ from $\hat{\phi}(\mathbf{x}_j)$, i.e.,

$$\tilde{d}^{(2)} = \left\| \phi(\mathbf{p}) - \hat{\phi}(\mathbf{x}_j) \right\|^2 = \left(\phi(\mathbf{p}) - \hat{\phi}(\mathbf{x}_j) \right)^T \left(\phi(\mathbf{p}) - \hat{\phi}(\mathbf{x}_j) \right). \quad (17)$$

The central idea of the fixed-point method [6] consists of computing the unknown preimage of a reconstructed point in the projected feature subspace by finding $\mathbf{p}$, which minimizes $\tilde{d}^{(2)}$. If an RBF kernel is considered, then the iterative procedure is described by the following equation [12]:

$$\mathbf{p}_{t+1} = \frac{\mathbf{X}_B(\mathbf{g} \diamondsuit \mathbf{k}_{p_t})}{\mathbf{g}^T \mathbf{k}_{\mathbf{p_t}}} \quad (18)$$

where $\diamondsuit$ represents a Hadamard product, and $\mathbf{g} = \mathbf{A}\mathbf{z}_j$. The components of the vector $\mathbf{k}_{\mathbf{p_t}} = \mathbf{k}(\mathbf{X}_B, \mathbf{p}_t)$ are given by the dot products between $\phi(\mathbf{p}_t)$ and the images $\mathbf{\Phi}_B$ of the training subset $\mathbf{X}_B$. The algorithm must be initialized, and $\mathbf{p}_0 \equiv \mathbf{x}_i$ is a valid choice [13], [14]. The points $\mathbf{p}_k$ then form the columns of $\hat{\mathbf{X}}$, i.e., the noise-free multidimensional signal in input space. The application of the method is illustrated in Fig. 1(d) and (f), where we can see that the denoised trajectory is smoother than the trajectory obtained with Local SSA. In the feature space, the data were projected (and reconstructed) using $L = 4$ or $L = 7$ directions, respectively, using KPCA or Greedy KPCA. The latter corresponds to using a low-rank approximation of the full kernel matrix only and will be discussed next.

1) Low-Rank Approximation of a Kernel Matrix: Applying kernel methods, an eigendecomposition of the related kernel matrix, and particularly the most significant eigenvalues and corresponding eigenvectors is often required. For large training data sets, the corresponding kernel matrix $\mathbf{K}$ becomes prohibitively large. Consequently, its eigendecomposition is often impractical in real data applications. In such cases, an appropriate dimension reduction must be achieved. Few papers [15], [16] discuss the application of the Nyström extension method to compute a low-rank approximation of the kernel matrix $\tilde{\mathbf{K}} = \mathbf{V}\mathbf{D}\mathbf{V}^T$, where only the R largest eigenvalues and corresponding eigenvectors are computed. The method is based on the fact that the kernel matrix can be written in the following block notation [16], [15]:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_r & \mathbf{K}_{rs} \\ \mathbf{K}_{rs}^T & \mathbf{K}_s \end{bmatrix}. \quad (19)$$

Considering that the full matrix has a dimension of $K \times K$, the upper-left block matrix $\mathbf{K}_r$ has a dimension of $R \times R$, the upper-right block matrix $\mathbf{K}_{rs}$ has a dimension of $R \times S$, and the lower-right block matrix $\mathbf{K}_s$ has a dimension of $S \times S$, where $S = K - R$. This notation implicates that the mapped training data set of dimension K is divided into two subsets of size R and $S = K - R$, respectively. The matrix $\mathbf{K}_r$ represents the kernel matrix within subset $\mathbf{\Phi}_R$ (with R vectors), $\mathbf{K}_{rs}$ is the kernel matrix comprising subsets $\mathbf{\Phi}_R$ and $\mathbf{\Phi}_S$, and $\mathbf{K}_s$ is the kernel matrix of the subset $\mathbf{\Phi}_S$.

The low-rank approximation is written using the block matrices $\mathbf{K}_r$ and $\mathbf{K}_{rs}$ according to [16], [15]

$$\tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{K}_r \\ \mathbf{K}_{rs}^T \end{bmatrix} \mathbf{K}_r^{-1} [\mathbf{K}_r \quad \mathbf{K}_{rs}]. \quad (20)$$

It can be shown that the lower block is approximated by $\mathbf{K}_s \approx \mathbf{K}_{rs}^T \mathbf{K}_r^{-1} \mathbf{K}_{rs}$. The Nyström extensions for the R eigenvectors $\mathbf{V}$ corresponding to the R largest eigenvalues are obtained as

$$\mathbf{V}^T = \mathbf{H}^T [\mathbf{K}_r \quad \mathbf{K}_{rs}]. \quad (21)$$

The matrix $\mathbf{H}$ is computed using eigendecompositions of $R \times R$ matrices, where R is the size of subset $\mathbf{\Phi}_R$. Different approaches were considered to form the $R \times R$ matrices. In [16], only the block $\mathbf{K}_r$ is considered, whereas in [15], a matrix related to both of the upper blocks of the kernel matrix is computed in addition. The main difference between both approaches is that the eigenvectors are either nonorthogonal [16] or orthogonal [15].

2) Computing a Reduced Set of Eigenvectors: The two kernel-based approaches, i.e., KPCA and Greedy KPCA, respectively, arise from two distinct strategies to deal with the eigendecomposition of the kernel matrix $(\mathbf{K})$ of the data set. In KPCA, the matrix $\mathbf{A}$ of mixing coefficients is computed using the largest eigenvalues $(\mathbf{D})$ and corresponding eigenvectors $(\mathbf{V})$ of $\mathbf{K}$ [17]. This results in a matrix of eigenvectors

$$\mathbf{U} = \mathbf{\Phi}\mathbf{V}\mathbf{D}^{-1/2} \quad (22)$$

which form the basis for a global representation of the data vectors.

In Greedy KPCA instead, a low-rank approximation of the kernel matrix is considered. This leads to an eigendecomposition with eigenvector matrices of reduced size. In this paper, we are interested in solutions that lead to orthogonal eigenvectors $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. In [15], a solution that uses as starting point the eigendecomposition of the block matrix $\mathbf{K}_r$ is proposed. The latter is formed by randomly selecting either elements of the training set or rows/columns of $\mathbf{K}$. This result is used to transform the data and compute a new $R \times R$ matrix, whose eigendecomposition will also contribute to the eigenvector matrix. Here, we instead use the proposal in [18], which is based on the incomplete Cholesky decomposition using a symmetric pivoting scheme. The incomplete Cholesky decomposition leads to

$$\mathbf{C} = [\mathbf{L} \quad \mathbf{L}^{-T}\mathbf{K}_{rs}]. \quad (23)$$

The matrix $\mathbf{L}$ represents a triangular matrix that corresponds to the complete Cholesky decomposition of $\mathbf{K}_r = \mathbf{L}^T\mathbf{L}$. Notice that the identification of the matrix $\mathbf{L}$ naturally arises with the pivoting scheme and does not need to be known in advance. Therefore, the pivoting index of the incomplete Cholesky decomposition [18] leads to the selection of $\mathbf{\Phi}_R$ from the training set.

Considering that the kernel matrix can be approximated by the incomplete Cholesky $\tilde{\mathbf{K}} = \mathbf{C}^T\mathbf{C}$, its low-rank approximation

can also be derived from an $R \times R$ matrix defined by

$$\mathbf{Q} = \mathbf{C}\mathbf{C}^T = \mathbf{V}_q \mathbf{D} \mathbf{V}_q^T. \tag{24}$$

Note that the matrix $\mathbf{C}$ can be centered before performing the eigendecomposition, dealing that way with an approximation of the centered kernel matrix. The result of this eigendecomposition as well as the decomposition of $\mathbf{K}_r$ leads to

$$\mathbf{H} = \mathbf{L}^{-1} \mathbf{V}_q \mathbf{D}^{-1/2}. \tag{25}$$

Substituting this result into the eigenvector equation [see (21)] yields

$$\mathbf{V} = \begin{bmatrix} \mathbf{K}_r \\ \mathbf{K}_{rs}^T \end{bmatrix} \mathbf{L}^{-1} \mathbf{V}_q \mathbf{D}^{-1/2}. \tag{26}$$

It can easily be shown that the Nyström extension to the eigenvector matrix $\mathbf{V}$ has R orthogonal eigenvectors.

Again, the mapped data set can be approximated by applying an SVD decomposition, where only the R most significant singular values and the corresponding eigenvectors are considered. This leads to the following representation of the projections in feature space:

$$\begin{aligned} \mathbf{Z} = \mathbf{D}^{1/2} \mathbf{V}^T &= \mathbf{V}_q^T \mathbf{L}^{-T} [\mathbf{K}_r \quad \mathbf{K}_{rs}] \\ &= \mathbf{V}_q^T \mathbf{L}^{-T} \mathbf{\Phi}_R^T [\mathbf{\Phi}_R \quad \mathbf{\Phi}_S]. \end{aligned} \tag{27}$$

Comparing the previous result with (14), the basis vector matrix can be written as

$$\mathbf{U} = \mathbf{\Phi}_R \mathbf{L}^{-1} \mathbf{V}_q. \tag{28}$$

Note that the R vectors form an orthonormal basis in feature space, i.e., $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. The eigenvectors in the matrix $\mathbf{V}_q$ should be placed according to their corresponding eigenvalues. The first column should have the eigenvector corresponding to the largest eigenvalue and so on. Furthermore, the matrix can have $L < R$ columns to enable projections of the data onto the directions related to the L largest eigenvalues.

3) Implementation of Greedy KPCA: A very efficient implementation for the incomplete Cholesky decomposition algorithm exists (accessible in [19]), having as input the training data set $\mathbf{X}$, σ of the RBF kernel, and a threshold to control the approximation error of the decomposition. As described in [18], the matrix $\mathbf{C}$ is iteratively formed, i.e., starting with one row up to R when the error is less than the threshold. The error ϵ is approximated as $\epsilon \approx \mathrm{tr}(\mathbf{K}_s - \mathbf{K}_{rs}^T \mathbf{K}_r^{-1} \mathbf{K}_{rs})$ [20]–[22]. Note that using an RBF function, the trace is obtained as $\mathrm{tr}(\mathbf{K}) = K$, where K denotes the size of the data set. The outputs of the algorithm are the index of the pivoting scheme and the matrix $\mathbf{C}$. The former allows identifying the subset $\mathbf{\Phi_R}$ that will contribute to form R orthogonal basis vectors [see (28)]. However, there are approaches [16] where the low-rank approximation of the kernel matrix is obtained by randomly selecting a training data set. The parameters of the model depend on the eigendecomposition of the kernel matrix of the training set, which are also used to form the basis vectors as it is used in the KPCA approach [see (22)]. In the simulations to be discussed, the Cholesky decomposition was applied with training sets formed with the complete data set (with J vectors) and randomly selecting $K < J$ vectors. In the latter case, the remaining data can be considered a test set ($J - K$ vectors) as it does not contribute to the parameters of the model. Furthermore, in [23], instead of using an error threshold to stop the Cholesky decomposition, a maximum number of pivots was used.

III. Results

Two projective subspace techniques are evaluated using artificially mixed real data. Artificial data are used to quantify the performance of the algorithms and, in particular, the influence of the embedding dimension M. The algorithms were applied to every signal of the artificial data set, and the performance measures were taken between the corrected EEG and the original. For illustration purposes, the algorithms are also applied to a frontal EEG data set with a high-amplitude EOG artifact. In this paper, the complexity of the approaches will also be discussed.

A. Artificial Data Set

In this paper, a strategy, which was proposed in [24], was pursued, where each signal is obtained by linearly adding two signals: an EEG and an EOG. The data set was organized so that it can also be easily characterized by visual inspection. The artificial mixed data set was obtained by using the following.

1) Twelve EEG segments of 10-s duration, sampled at 250 Hz and with reference to scalp electrode Cz, were selected by three specialists. The segments are grouped into four types according to the visibility/dominance of one of the characteristic events: Type A—delta activity (0.5–4 Hz); Type B—theta activity (4–8 Hz); Type C—alpha band (8–13 Hz); and Type D—beta activity (13 Hz–25 Hz).
2) EOG artifacts were extracted from EEG frontal channels not belonging to the same recordings (or subjects) used to select EEG. The SSA $(q = 1)$ algorithm with $L = 1$ was used to extract the artifact. By visual inspection, the signals were recognized as clearly defined artifacts not showing any EEG-relevant information. In total, ten segments (of 10-s duration) with ocular artifacts were selected. The segments have a variable number of eye blinks (1–7), and in some examples, ocular movements are also present.

Fig. 2 shows an example for each type of EEG segment and the corresponding artificially mixed signals.

B. Parameters of Evaluation

The artificial mixtures are used to study the influence of the parameters on the performance of both algorithms. The comparisons will be made both in time and in frequency by using the correlation coefficient and the coherence function, respectively. The comparison will be made between the corrected EEG $(y[n])$ and the original EEG $(o[n])$.

The correlation coefficient evaluates the similarity of the two signals, and its value is independent of scaling or mean (m)

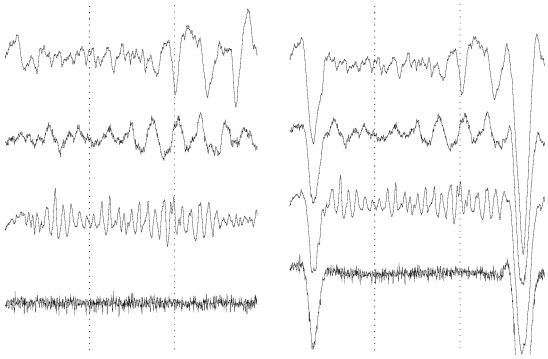

Fig. 2. Subsegments (3 s) of (left) the original EEG and (right, last row was clipped for visual proposes) the artificial mixture. (Top to bottom) Segment type: (first) Type A-Delta, (second) Type B-Theta, (third) Type C-Alpha, and (fourth) Type D-Beta.

differences. The absolute value ranges from 0 to 1 and is defined according to

$$cc_{oy} = \frac{\sum_{n=0}^{N-1} (o[n] - m_o)(y[n] - m_y)}{\sigma_o \sigma_y} \tag{29}$$

where σ represents the standard deviation of the N amplitude values of the signal. The coherence function has values in the range of 0–1 and is computed using the periodograms. Given the discrete Fourier transform (DFT) of the ith subsegment of each signal O_i and Y_i, the coherence of the mth bin in frequency is defined as

$$cf_{oy}(m) = \frac{\left|\sum_{i=1}^{I} Y_i^*(m) O_i(m)\right|^2}{\sum_{i=1}^{I} |O_i(m)|^2 \sum_{i=1}^{I} |Y_i(m)|^2}. \tag{30}$$

In the experimental results to be discussed, the segments are divided into overlapping subsegments (50% of overlap), and the DFT is computed with a resolution of 1 Hz. Furthermore, the coherence values are presented for each of the four characteristic EEG bands by averaging the bins within the frequency range of the band.

C. Evaluation of Performance

The subspace techniques discussed so far are applied to multidimensional signals resulting from an embedding of the recorded time series $x[n]$ into their delayed coordinates. Then, the embedding dimension M is a choice to be made before the application of both subspace techniques. In SSA-related literature [1], [7], it is referred that M should be higher than a threshold computed according to f_s/f_o, where f_o is related to the frequency of the artifact, and f_s is the sampling rate. A similar criterion was used in [25] to find the embedding dimension for an algorithm based on independent component analysis. After embedding, each segment of the data set is represented by a multidimensional data set $\mathbf{x}_k$, $k = 1, \ldots, J$, and will be the input of both algorithms, and the output is the extracted artifact $\hat{x}[n]$, which will be subtracted from the mixed signal to obtain the corrected EEG $(y[n])$.

1) Local SSA: In what concerns the Local SSA, to have a correct estimate of the covariance matrix in each cluster, we should have enough data in each cluster. This constitutes a practical upper bound to the number of clusters. In each cluster, the MDL criterion was used to select the subspace dimension to reconstruct the EOG signal. Note that the MDL criterion works best if enough data are available in each cluster [11]. Then, a heuristic was considered to assign the number of clusters: the clustering step is repeated until a reliable decomposition is achieved in each cluster. Starting with a maximal number of clusters, $q_{\max} = 10$, checking afterward if all the clusters end up with a cardinality higher than M, in which case the signal subspace dimension in each cluster is chosen as $L_{c_i} < (M/2)$. If both criteria are not met, then the number q of clusters is decreased, and the process is repeated. Using this strategy, the only parameter to be assigned by the user is M. Fig. 3 shows the mean correlation coefficient of the algorithm changing M from 6 to 96. The level of performance also depends on the dominant frequency range of the original EEG: from 0.9 (Type D) to 0.4 (Type A), being more reliable for segments with dominant frequencies ranging far from the frequency contents of the artifacts. However, it is possible to find a unique M for all types of segment. The number of clusters automatically assigned for the data set varied between 2 and 9 and does not depend on the EEG segment used to generate the artificial mixture. Rather, it is related to the artifact. If the segment only has one or two blinks and no baseline drifts, then the number of clusters is 2. However, for an increasing number of blinks and baseline drifts or ocular movements, the number of clusters also increases.

2) Greedy KPCA: An RBF kernel with $\sigma = \max_i(\|\mathbf{x}_i - \mathbf{x}_{\text{mean}}\|)$, $i = 1, \ldots, J$, is applied, where $\mathbf{x}_{\text{mean}}$ denotes the mean of the data set. The multidimensional data resulting from the embedding of each segment are used as training set $(K = J)$, and the Greedy KPCA was applied using an implementation adapted from [19], where the threshold to stop the algorithm was $\epsilon \leq 0.01J$. Note that the number of data vectors depends on the embedding dimension. After the eigendecomposition of matrix $\mathbf{Q}$, the number of directions L was chosen to maintain 0.95% of the variance of the data in the feature space.

To study the dependence of the performance on the subspace dimension M, the correlation coefficient cc_{oy} between the original and corrected EEGs was also considered. Fig. 3 shows the result, and we can verify that to achieve a stable behavior, the embedding dimension M can be smaller than the Local SSA. However, the level of performance is worse than that achieved with Local SSA in spite of having a similar tendency. The number of pivots needed to fulfill the error criterion changes for the different segments. Table I shows the range of values for each type of segment when $M = 46$. The number of pivots is related to the type of artifact, not to the EEG segment used in the mixture. Note that the size of the training data set for each segment is $J = 2456$, and the R median value is < 100 in all cases. The maximum values only occur for a segment that has simultaneously ocular artifacts and baselines drifts. Furthermore, notice that the maximal values of L reveal that less than 1/4 of the computed eigenvectors of $\mathbf{Q}$ are used. This

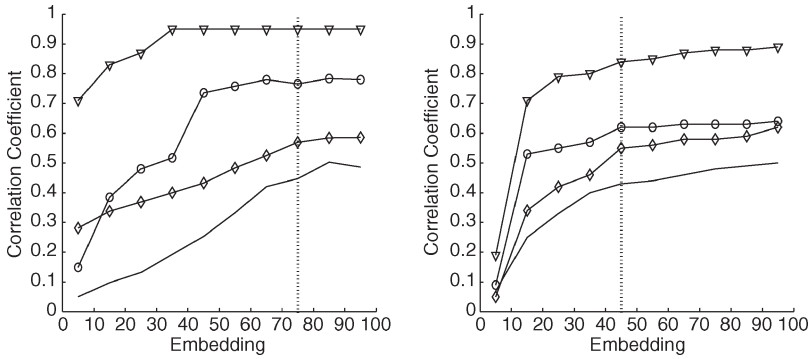

Fig. 3. Mean correlation coefficient (cc_{oy}) versus embedding dimension (M): (left) Local SSA and (right) Greedy KPCA. Segment type: ($\square$) type A, ($\diamond$) type B, ($\circ$) type C, and (∇) type D.

TABLE I
GREEDY KPCA (MIN—MINIMUM; MED—MEDIAN; MAX—MAXIMUM)

	Pivots (R)			Directions (L)		
	Min	Med	Max	Min	Med	Max
Type A	33	79	207	14	24	46
Type B	24	74	177	10	20	39
Type C	34	88	390	13	25	77
Type D	33	90	645	9	18	48

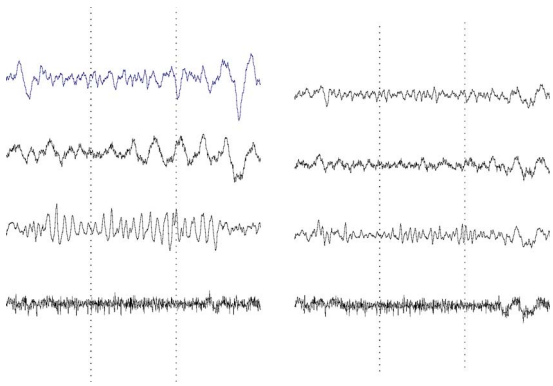

Fig. 4. Corrected versions resulting from the application of (left) Local SSA and (right) Greedy KPCA. (Top to bottom) Segment type: (first) Type A-Delta, (second) Type B-Theta, (third) Type C-Alpha, and (fourth) Type D-Beta.

fact may indicate that the Cholesky decomposition could have had an earlier stop without affecting the performance.

3) Greedy KPCA Versus Local SSA: To compare distortions in time and frequency of both algorithms, the outputs of the Local SSA with $M=76$ and the Greedy KPCA with $M=46$ were used. Fig. 4 shows the output of the algorithms for the segments illustrated in Fig. 2, and we can verify the difference on performance of the algorithms, namely, for Type C segments, where the alpha bursts almost disappear from the Greedy KPCA outcome. For Local SSA, the only visible distortion is on Type A segments, where some slow waves are not visible on the outcome. Furthermore, the algorithm SSA ($q=1$ with $L=1$) varying M from 6 to 96 in steps of 5 was also applied to each segment, and for comparison purposes, the corrected EEG version with the highest correlation with the original was selected.

Fig. 5 shows a comparison of the correlation coefficients between the output of the algorithms and the original for the segments in the data set. The level of distortion is related to the segment type. The correlation coefficient decreases as the EEG frequency content is closer to the frequency of artifact. However, in all the cases, Local SSA performs better than Greedy KPCA and SSA. The analysis in the frequency range confirms these results. Note that whatever is the segment, the beta band is always the least distorted, i.e., the coherence function cf_{oy} always has a value close to 1, and the standard deviation (across segments) is very small (see vertical line in Fig. 6). The alpha band also has values of around 0.9 in the three cases for Local SSA, whereas for the other algorithms, the values have a broader range. In particular, the Greedy KPCA has 0.4 for segments Type C, whereas the other algorithms have 0.9. The values of coherence for segment Types A and B of the Local SSA vary between 0.4 and 0.6, whereas the Greedy KPCA is 0.1–0.4.

In most of the cases, the SSA algorithm shows a performance similar to Local SSA. However, notice that the embedding dimension M was not fixed in this paper; rather, it was kept variable, and the output was chosen according to the correlation between the corrected and original signal. This way, this implementation is not useful in any practical application, as the artifact-free original signal is not available.

D. Analyzing a Frontal Channel

In this paper, the two suggested projective subspace techniques will be illustrated using a data segment from a real signal of 12-s duration and a sampling rate of 128 Hz, which was recorded from a frontal EEG channel (Fp1-Cz) contaminated with a very prominent EOG artifact. This segment was used in previous works [23], [17] to illustrate the use of different KPCA subspace algorithms. The computational complexity of KPCA depends not only on the size of the kernel matrix but also on the procedure to estimate the preimage [14] of the data points obtained after denoising the mapped data in feature space. To cope with the problems arising from the size of the kernel matrix, the data window can be subdivided into segments of proper size. However, the size of the segment has to be adapted to the structure of the recorded signal so that each segment contains the characteristics of the artifact signal to be separated. This leads to the study of greedy approaches, and in [23], the two alternatives for the Nyström approach are combined to decrease the computational load.

In this section, we applied the subspace techniques using the same strategy described for the artificial data set. Fig. 7

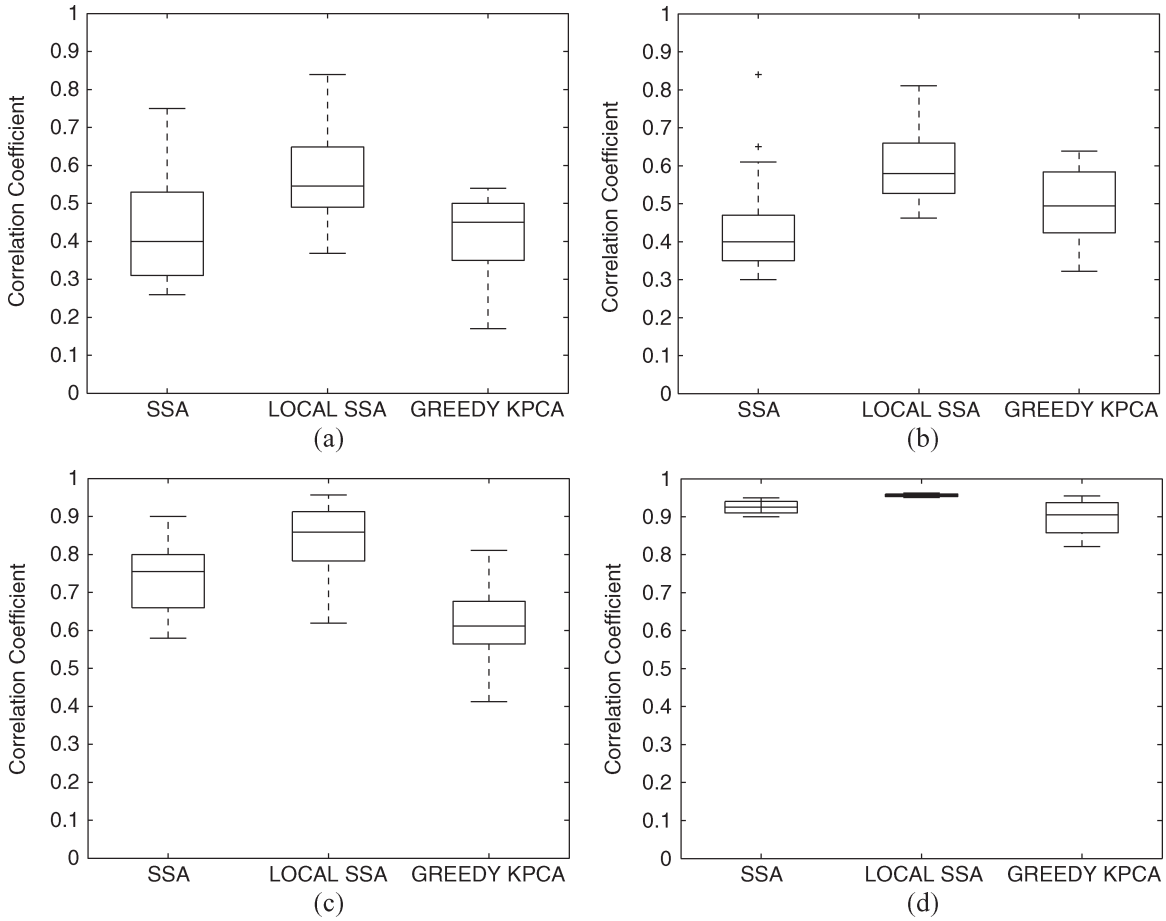

Fig. 5. Boxplots of correlation coefficients (cc_{oy}) for SSA, Local SSA, and Greedy KPCA algorithms. (a) Type A. (b) Type B. (c) Type C. (d) Type D.

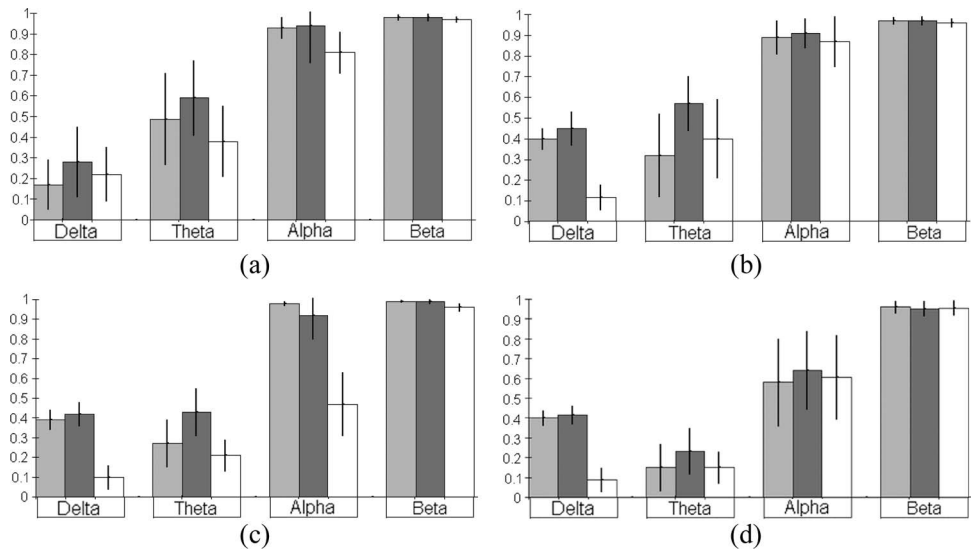

Fig. 6. Coherence values in the different frequency bands for (gray bar) SSA, (black bar) Local SSA, and (white bar) Greedy KPCA. (a) Type A. (b) Type B. (c) Type C. (d) Type D.

shows the output of the Local SSA for which the number of clusters was automatically assigned to $q = 4$. For this signal taking $q = 6$, the power line interference (50 Hz) is extracted with the EOG artifact, as shown in [23]. The Greedy KPCA was applied using both strategies described in Section II-D3. Using the complete data set and an error threshold to stop the incomplete Cholesky decomposition leads to the selection of $R = 53$ pivots. Fig. 8 (third trace) illustrates this choice by

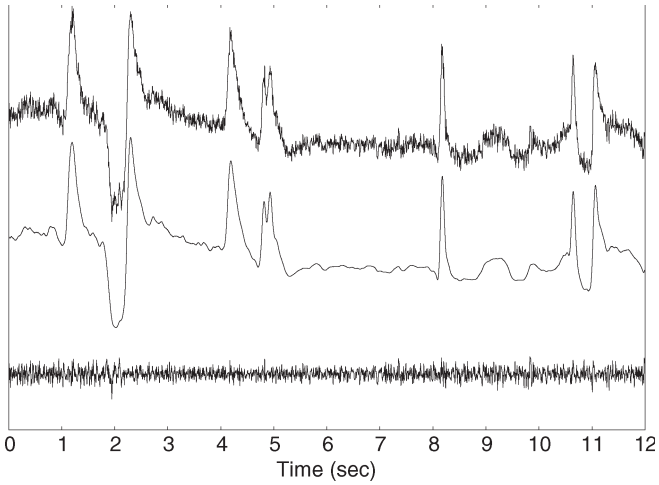

Fig. 7. Local SSA ($M = 41$, $q = 4$). (Top to bottom) Original, artifact, and corrected EEGS.

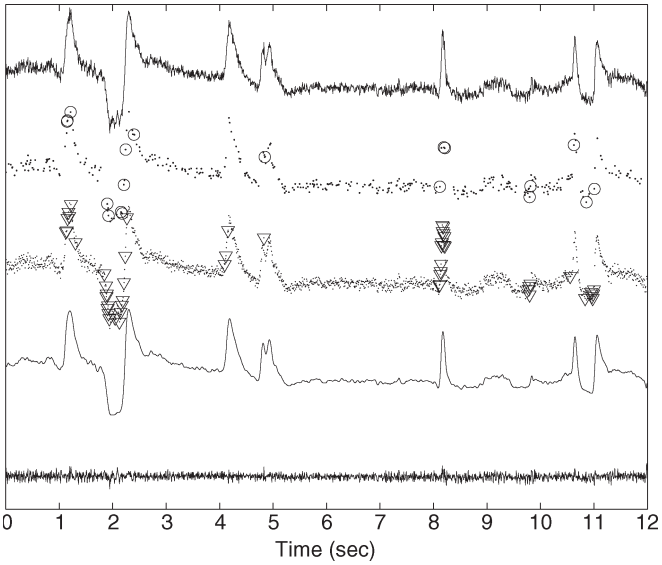

Fig. 8. Greedy KPCA ($M = 25$). (Top to bottom) Original, pivot selection ($R = 20$), pivot selection ($R = 53$), artifact, and corrected EEG.

plotting the first row of the data matrix ordered according to their time argument. For the other alternative, the training set is formed with 25% of the data, and the Cholesky decomposition is performed up to a maximum of $R = 20$ pivots. Fig. 8 (second trace) indicates the localization of the pivots in the latter case, and it can be verified that they match the same regions of the signal. The corrected signal and the EOG signal are very similar to the signals obtained with KPCA applied to subsegments of 4 s, as described in [23]. The correlation coefficient between the corrected EEGs of the different algorithms is in the range of 0.91–0.95, and between the extracted artifacts, it is 0.99 for every possible combination of signals.

Fig. 9 shows the frequency contents of the corrected EEG by the discussed algorithms in comparison to the original. The output of the Local SSA algorithm coincides with the original for frequencies higher than 10 Hz. Moreover, the Greedy KPCA with $R = 20$ pivots is the best versions, as verified in the figure.

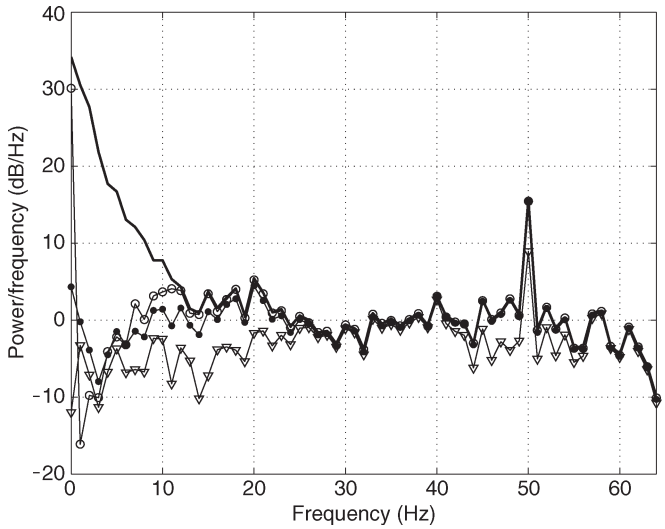

Fig. 9. Power spectral density (in decibels per hertz) versus frequency (in hertz) of the original EEG (-) and corrected EEG by Local SSA ($\circ$), Greedy KPCA with 20 Pivots ($\bullet$), and Greedy KPCA (∇).

IV. Concluding Remark

Artifact reduction in EEG recordings [26] is a very important problem that needs to be addressed in a systematic way. Such artifacts can often be found only in certain channel recordings, like EOG artifacts in frontal channel recordings. Hence, it is of practical interest to be able to efficiently remove such artifacts from single-channel recordings without the need to do a full multichannel analysis. Thus, we have been studying the feasibility of projective subspace techniques to address the problem using a single-channel approach. We consider such projections either in input space or in a high-dimensional feature space, after a nonlinear mapping of the data from the input space to the feature space. We have proposed the Local SSA to adapt a linear technique to nonlinear problems in input space and also to adapt a generically nonlinear technique, namely KPCA, to deal with very-high-dimensional problems of prohibitive computational complexity. This leads us to propose Greedy KPCA based on a Nyström extension of low-rank approximations to the kernel matrix. Numerical simulations using artificially mixed data reveal that both techniques are feasible to remove the high-amplitude ocular movements and blinks. Local SSA shows a better performance as the corrected EEG exhibits less distortion in all the frequency bands. However, it was shown that both approaches have a similar performance in what concerns frequency distortions in the frequency range of beta and alpha bands. The frontal EEG signal analysis confirms these results. Concerning computational complexity, Local SSA suffers from the burden imposed by the recursive procedure to choose the number of clusters (q), whereas in Greedy KPCA, it is the size of the training set that renders the choice of the pivots a slow process. The hybrid approach applied to the real example might then be an alternative that must further be studied as well as the stopping criterion of the Cholesky decomposition.

Both algorithms (Local SSA and Greedy KPCA) are incorporated in the EEGLAB [27] environment. This open-software tool based on MATLAB offers visualization facilities that will allow accomplishing a clinical evaluation task. However, the

new facilities (plugins) need to be improved to cope with the long segments of the signal. Our goal is to use this facility in the visualization of critical segments of signals from a database of epileptic patients recorded in long-term monitoring sessions and study the impact of the application of the algorithms. In the described scenario, the algorithm can be applied in parallel to channels that suffer from high-amplitude artifacts. This could be useful to detect the onset of a focal seizure.

ACKNOWLEDGMENT

The authors would like to thank A. Martins da Silva and Hospital Geral de Santo António for providing the signals and the helpful discussions.

REFERENCES

[1] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, *Analysis of Time Series Structure: SSA and Related Techniques*. London, U.K.: Chapman & Hall, 2001.

[2] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[3] C. H. You, S. N. Koh, and S. Rahardja, "Signal subspace speech enhancement for audible noise reduction," in *Proc. IEEE ICASSP*, Philadelphia, PA, 2005, vol. I, pp. 145–148.

[4] R. Duda, P. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ: Wiley, 2001.

[5] M. Ghil, M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, "Advanced spectral methods for climatic time series," *Rev. Geophys.*, vol. 40, no. 1, pp. 3.1–3.41, 2002.

[6] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[7] A. R. Teixeira, A. M. Tomé, E. Lang, P. Gruber, and A. Martins da Silva, "Automatic removal of high-amplitude artifacts from single-channel electroencephalograms," *Comput. Methods Programs Biomed.*, vol. 83, no. 2, pp. 125–138, Aug. 2006.

[8] A. R. Teixeira, A. M. Tomé, E. Lang, R. Schachtner, and K. Stadlthanner, "On the use of KPCA to extract artifacts in one-dimensional biomedical signals," in *Proc. IEEE MLSP*, S. McLoone, J. Larsen, M. V. Hulle, A. Rogers, and S. C. Douglas, Eds., Dublin, Ireland, 2006, pp. 385–390.

[9] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, Mar. 2000.

[10] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[11] A. P. Liavas and P. A. Regalia, "On the behavior of information theoretic criteria for model order selection," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1689–1695, Aug. 2001.

[12] B. Schölkopf, S. Mika, C. J. Barges, P. Knirsch, K.-R. Müller, G. Ratsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1000–1017, Sep. 1999.

[13] T. Takahashi and T. Kurita, "Robust de-noising by kernel PCA," in *Proc. ICANN*, J. Dorronsoro, Ed. Madrid, Spain: Springer-Verlag, 2002, vol. 2415, pp. 739–744.

[14] A. Teixeira, A. Tomé, K. Stadlthanner, and E. Lang, "KPCA denoising and the pre-image problem revisited," *Dig. Signal Process.*, vol. 18, no. 4, pp. 568–580, Jul. 2008.

[15] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, Feb. 2004.

[16] C. K. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, pp. 682–688.

[17] A. R. Teixeira, A. M. Tomé, E. W. Lang, and K. Stadlthanner, "Nonlinear projective techniques to extract artifacts in biomedical signals," in *Proc. EUSIPCO*, Florence, Italy, 2006.

[18] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 1–48, 2002.

[19] F. R. Bach, *Kernel Independent Component Analysis*, 2003. [Online]. Available: http://www.di.ens.fr/~fbach/kernel-ica/index.htm

[20] V. Franc and V. Hlavá, "Greedy algorithm for a training set reduction in the kernel methods," in *Proc. 10th Int. Conf. Comput. Anal. Images Patterns*. Groningen, Holland: Springer-Verlag, 2003, pp. 426–433.

[21] G. C. Cawley and N. L. C. Talbot, "Efficient formation of a basis in a kernel induced feature space," in *Proc. Eur. Symp. Artif. Neural Netw.*, M. Verleysen, Ed., Bruges, Belgium, 2002, pp. 1–6.

[22] G. Baudat and F. Anouar, "Feature vector selection and projection using kernels," *Neurocomputing*, vol. 55, no. 1/2, pp. 21–38, Sep. 2003.

[23] A. R. Teixeira, N. Alves, A. M. Tomé, M. Böhm, E. W. Lang, and C. G. Puntonet, "Single-channel electroencephalogram analysis using non-linear subspace techniques," in *Proc. IEEE Int. Symp. Intell. Signal Process. (WISP)*, Madrid, Spain, 2007, pp. 865–870.

[24] C. W. Anderson, J. N. Knight, T. O'Connor, M. J. Kirby, and A. Sokolov, "Geometric subspace methods and time-delay embedding for EEG artifact removal and classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 142–146, Jun. 2006.

[25] C. J. James and D. Lowe, "Extracting multisource brain activity from a single electromagnetic channel," *Artif. Intell. Med.*, vol. 28, no. 1, pp. 89–104, May 2003.

[26] R. J. Croft and R. J. Barry, "Removal of ocular artifact from the EEG: A review," *Neurophysiol. Clin.*, vol. 30, no. 1, pp. 5–19, Feb. 2000.

[27] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics," *J. Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

A. R. Teixeira received the B.S. degree in mathematics applied to technology in 2003 from the University of Porto, Porto, Portugal, and the M.Sc. degree in electrical engineering in 2006 from the University of Aveiro, Aveiro, Portugal, where she is currently working toward the Ph.D. degree also in electrical engineering with the Signal Processing Group, Instituto de Engenharia Electrónica e Telemática de Aveiro.

Her research interests include biomedical digital signal processing and principal and independent component analysis.

A. M. Tomé (S'86–M'90) received the Ph.D. degree in electrical engineering from University of Aveiro, Aveiro, Portugal, in 1990.

She is currently an Associate Professor of electrical engineering with the Electronics, Telecommunications and Informatics Department (DETI), Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA), University of Aveiro. Her research interests include digital and statistical signal processing, independent component analysis, and blind source separation, as well as classification and pattern recognition applications.

M. Böhm received the Diploma degree from University of Regensburg, Regensburg, Germany, in 2004. He is currently working toward the Doctoral degree with the Institute of Biophysics, Computational Intelligence and Machine Learning Group, University of Regensburg.

His scientific interests are in the fields of blind source separation, bio-inspired optimization, and brain modeling.

Carlos G. Puntonet received the Ph.D. degree in electronics from University of Granada, Granada, Spain, in 1994.

He is currently an Associate Professor with the "Departamento de Arquitectura y Tecnología de Computadores," University of Granada. His research interests lie in the fields of signal processing, linear and nonlinear independent component analysis, artificial neural networks, and optimization methods.

Elmar W. Lang received the Ph.D. degree in physics from University of Regensburg, Regensburg, Germany, in 1980.

He is currently an Adjunct Professor of biophysics with the University of Regensburg, where he is heading the Computational Intelligence and Machine Learning Group (CIMLG). His current research interests include biomedical signal and image processing, independent component analysis and blind source separation, neural networks for classification, and stochastic process limits in queueing applications.