Available online at www.sciencedirect.com
ScienceDirect
Digital Signal Processing 18 (2008) 568–580
Digital Signal Processing
www.elsevier.com/locate/dsp

KPCA denoising and the pre-image problem revisited

A.R. Teixeira [a], A.M. Tomé [a,*], K. Stadlthanner [b], E.W. Lang [b]

[a] *DETI/IEETA, Universidade de Aveiro, 3810-193 Aveiro, Portugal*
[b] *Institute of Biophysics, University of Regensburg, D-93040 Regensburg, Germany*

Available online 15 August 2007

Abstract

Kernel principal component analysis (KPCA) is widely used in classification, feature extraction and denoising applications. In the latter it is unavoidable to deal with the pre-image problem which constitutes the most complex step in the whole processing chain. One of the methods to tackle this problem is an iterative solution based on a fixed-point algorithm. An alternative strategy considers an algebraic approach that relies on the solution of an under-determined system of equations. In this work we present a method that uses this algebraic approach to estimate a good starting point to the fixed-point iteration. We will demonstrate that this hybrid solution for the pre-image shows better performance than the other two methods. Further we extend the applicability of KPCA to one-dimensional signals which occur in many signal processing applications. We show that artefact removal from such data can be treated on the same footing as denoising. We finally apply the algorithm to denoise the famous USPS data set and to extract EOG interferences from single channel EEG recordings.

Keywords: Kernel principal component analysis (KPCA); Pre-image; Time series analysis; Denoising

1. Introduction

The objective of noise reduction techniques is to improve noisy signals. Projective subspace techniques can be used favorably to get rid of most of the noise contributions to *multidimensional* signals [1]. Whereas signals are considered to live in a sub-manifold only, noise is assumed to fill the multidimensional space evenly. The goal of subspace methods thus is to project the noisy signal onto the signal-plus-noise, or simply, signal subspace. This way part of the noise on the signal can be removed. Hence an estimate of the clean multidimensional signal can be made by removing or nullifying the components of the signal in the noise subspace, retaining only the components in the signal subspace. The decomposition of the space into (two) subspaces can be done using either singular value decomposition (SVD) or principal component analysis (PCA). Both techniques estimate those directions, corresponding to the L largest eigenvalues of the data covariance/scatter matrix or singular values of the data matrix, which can be associated with the eigenvectors spanning the signal subspace. The remaining orthogonal directions then can be associated with the noise subspace. Reconstructing the multidimensional signal using only those L dominant components can result in a substantial noise reduction of the recorded signals. Note that this approach is most appropriate if the underlying signal represents the main contribution to the recorded signal. More recently those algorithms are applied in feature

* Corresponding author.
E-mail address: ana@ieeta.pt (A.M. Tomé).

1051-2004/$ – see front matter
doi:10.1016/j.dsp.2007.08.001

space created by a nonlinear mapping of the data. These *generically nonlinear* signal processing techniques like kernel principal component analysis (KPCA) are often used for denoising in image applications [2,3]. KPCA also represents a projective subspace technique. It transforms a signal nonlinearly into feature space and applies a linear principal component analysis to the transformed signals. But to recover the noise-reduced signal in input space after denoising in feature space, the nonlinear mapping must be reverted, i.e. the pre-image in input space must be estimated.

In this work the KPCA methodology is shortly reviewed, following a matrix manipulation approach which is especially convenient when signal reconstruction and pre-image estimation are considered. Two methods of computing the pre-image [4,5], discussed in the literature, are summarized before a particularly suitable way of finding the starting point of the fixed-point iterative [5] method is suggested. This modification renders the algorithm much more efficient and fast.

Furthermore, we adapt KPCA methods to denoise *one-dimensional* signals as well as to extract artefact-related features interfering with the signal of interest. Various signal processing applications rely on *one-dimensional* signals like biomedical signals, speech recordings, climatic and meteorological time series, just to mention a few. Clearly projective subspace techniques are only available for multidimensional signals. Hence, time series analysis techniques often rely on embedding one dimensional sensor signals in the space of their time-delayed coordinates [6–8], also called embedding space. Then resulting multidimensional signal can now be analyzed using KPCA. We apply the method to extract prominent artifacts like electro-oculograms (EOG) in electro-encephalograms (EEG). Note that in this example, the artifact-related contributions to the recorded EEG signals are considered "the signal" and the actual EEG signal is considered a "sort of a broadband noise." Consequently, we can use the projective subspace techniques referred to above to separate the dominating artifacts from the "pure" EEG signals. Again the influence of the pre-image estimation on the performance of the algorithm is discussed.

2. KPCA denoising

In practice, the goal of projective subspace techniques is to describe the data with reduced dimensionality by extracting meaningful components while still retaining the structure of the raw data. KPCA relies on a nonlinear mapping of given data to a higher dimensional space, called feature space. Then KPCA can simultaneously retain the nonlinear structure of the data while denoising is achieved with better performance because the projections are accomplished in the higher-dimensional feature space.

2.1. Kernel-PCA

In KPCA a set of multidimensional signals $\mathbf{x}_k,\ k = 1, \ldots, K$, is envisaged to be mapped through a nonlinear function $\phi(\mathbf{x}_k)$ into a feature space yielding the mapped data set $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_K)]$. In feature space then a linear PCA is performed estimating the eigenvectors and eigenvalues of a matrix of *outer* products, called a *scatter matrix* which for zero mean data is given by $\mathbf{C} = \mathbf{\Phi}\mathbf{\Phi}^T$. It can be shown that these eigenvectors and eigenvalues are related to those of a matrix of *inner* products, called a *kernel matrix* $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$. Using the kernel trick [2], the centered kernel matrix can be expressed as

$$\mathbf{K}_c = \left(\mathbf{I} - \frac{1}{K}\mathbf{j}_K\mathbf{j}_K^T\right)\mathbf{\Phi}^T\mathbf{\Phi}\left(\mathbf{I} - \frac{1}{K}\mathbf{j}_K\mathbf{j}_K^T\right) = \left(\mathbf{I} - \frac{1}{K}\mathbf{j}_K\mathbf{j}_K^T\right)\mathbf{K}\left(\mathbf{I} - \frac{1}{K}\mathbf{j}_K\mathbf{j}_K^T\right), \tag{1}$$

where $\mathbf{j}_K = [1, 1, \ldots, 1]^T$ is a vector with dimension $K \times 1$, and $\mathbf{I}$ is a $K \times K$ identity matrix. Notice that each element $k(i, j) \equiv k(\mathbf{x}_i, \mathbf{x}_j)$ of the kernel matrix depends on the inner product $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$ which can be computed using only the data $\mathbf{x}_k$ in input space. For instance, if a radial basis function (RBF) kernel is used, $k(i, j)$ is calculated according to

$$k(i, j) \equiv k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \tag{2}$$

where σ^2 is a free parameter related to the width of the kernel. It can be chosen according to any suitable data spread criterion.

Because the eigenvalues of the scatter matrix $\mathbf{C}$ coincide with the eigenvalues of the kernel matrix $\mathbf{K}$, the eigendecomposition of $\mathbf{K}_c$ provides the necessary information to compute the projection of a vector of the input space $\mathbf{y}_j$ in the feature space. Considering the matrix $\mathbf{V}$, the columns of which represent the L eigenvectors of the kernel matrix,

and $\mathbf{D}$, a diagonal matrix with the corresponding $L \leqslant K$ eigenvalues of both matrices, the image $\phi(\mathbf{y}_j)$ of a point in input space, can be projected onto the L directions spanned by the eigenvectors of the scatter matrix via

$$\mathbf{z}_j = \mathbf{D}^{-1/2}\mathbf{V}^T\left(\mathbf{I} - \frac{1}{K}\mathbf{j}_K\mathbf{j}_K^T\right)\mathbf{\Phi}^T\phi(\mathbf{y}_j), \tag{3}$$

where $\mathbf{\Phi}^T\phi(\mathbf{y}_j)$ represents a vector the components of which can be computed using the kernel trick by

$$\mathbf{k}_{y_j} = \left[k(\mathbf{x}_1, \mathbf{y}_j), k(\mathbf{x}_2, \mathbf{y}_j), \ldots, k(\mathbf{x}_K, \mathbf{y}_j)\right]^T. \tag{4}$$

There are many applications (for instance classification) where the projections provide necessary and sufficient information to characterize the problem. However, in denoising applications, for example, it is needed to reconstruct any data point in feature space from its noisy version employing the L principal components. Finally, the position of the corresponding point in input space is of interest, hence the pre-image of the denoised data sample in feature space needs to be estimated [4,5].

2.2. Reconstruction and pre-image

Consider the reconstructed point in feature space:

$$\hat{\phi}(\mathbf{y}_j) = \mathbf{\Phi}\left(\mathbf{I} - \frac{1}{K}\mathbf{j}_K\mathbf{j}_K^T\right)\mathbf{V}\mathbf{D}^{-1/2}\mathbf{z}_j = \mathbf{\Phi}\mathbf{g}. \tag{5}$$

In order to avoid to work with the mapped data set $\mathbf{\Phi}$, pre-image estimation methods described in literature combine the reconstruction in feature space and the estimation of its pre-image in input space in one step. This goal is achieved by using the Euclidian or $L2$-norm. The square of the Euclidian distance can be written using dot products which in turn can be substituted by kernel evaluations. Considering a point $\mathbf{p}$ in input space, the distance of its image $\phi(\mathbf{p})$ in feature space to the reconstructed point $\hat{\phi}(\mathbf{y}_j)$ is defined by

$$\tilde{d}^{(2)} = \left\|\phi(\mathbf{p}) - \hat{\phi}(\mathbf{y}_j)\right\|^2 = \left(\phi(\mathbf{p}) - \hat{\phi}(\mathbf{y}_j)\right)^T\left(\phi(\mathbf{p}) - \hat{\phi}(\mathbf{y}_j)\right). \tag{6}$$

Substituting the expression given in Eq. (5) to compute the reconstructed point $\hat{\phi}(\mathbf{y}_j)$, the dot product can be replaced by kernel values

$$\tilde{d}^{(2)} = k(\mathbf{p}, \mathbf{p}) - 2\mathbf{g}^T\mathbf{k}_p + \mathbf{g}^T\mathbf{K}\mathbf{g}, \tag{7}$$

where $\mathbf{k}_p$ represents a vector whose entries are computed as the dot product of $\phi(\mathbf{p})$ with images $\mathbf{\Phi}$ of the set of training data $\{\mathbf{x}_k\}$ according to Eq. (4) identifying $\mathbf{y}_j \equiv \mathbf{p}$. Both pre-image estimation methods use the definition of the Euclidian distance within different strategies, and consequently the input space point $\mathbf{p}$ must be chosen accordingly. In the following we will discuss these different strategies.

2.2.1. Distance method

Recent work [4] to estimate the pre-image of a given point in feature space is based on the fact that it is possible to compute the coordinates of a new point if we know its distances to a set of known points [9]. Hence, the distances of the reconstructed point $\hat{\phi}(\mathbf{y}_j)$ to every point in the set Φ of images of the training data $\mathbf{x}_k$ are computed. So in Eq. (7), the point $\mathbf{p}$ is chosen as an element of the training set, i.e., $\mathbf{p} \equiv \mathbf{x}_k$. Then by computing the distances to all mapped points of the training set $\mathbf{x}_k, k = 1, \ldots, K$, the following distance vector is obtained

$$\tilde{\mathbf{d}}^{(2)} = \operatorname{diag}(\mathbf{K}) - 2\mathbf{g}^T\mathbf{K} + \mathbf{g}^T\mathbf{K}\mathbf{g}. \tag{8}$$

With certain kernels, especially isotropic kernels and polynomial kernels with odd powers [4], it is possible to evaluate the distance in input space knowing the corresponding distance in feature space. If, for example, an RBF kernel is considered, there is a distinct relation between an input space distance $\mathbf{d}^{(2)}$ and the corresponding feature space distance. Once the vector of distances in feature space can be computed as

$$\tilde{\mathbf{d}}^{(2)} = 2\left(\mathbf{j}_K - \exp\left(-\frac{\mathbf{d}^{(2)}}{2\sigma^2}\right)\right) \tag{9}$$

the corresponding vector of distances in input space is then given by

$$\mathbf{d}^{(2)} = -2\sigma^2 \ln\bigl(\mathbf{j}_K - 0.5\tilde{\mathbf{d}}^{(2)}\bigr). \tag{10}$$

Consider next a subset $\mathcal{S}$ of neighbors of the reconstructed point $\hat{\phi}(\mathbf{y}_j)$, i.e., choose from the training set those S points with smallest distance $\tilde{\mathbf{d}}^{(2)}$, and select the corresponding points $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_S]$ in input space. The coordinates of the subset of points may be taken as the columns of the eigenvector matrix $\mathbf{E}$ of their covariance matrix. After centering the set of neighbors by

$$\mathbf{Q}_c = \mathbf{Q}\left(\mathbf{I} - \frac{1}{S}\mathbf{j}_S^T\mathbf{j}_S\right) \tag{11}$$

the columns of $\mathbf{W} = \mathbf{E}^T\mathbf{Q}_c$ represent the new coordinates of the points $\mathbf{Q}_c$. Their distance to the origin is obtained as $\mathbf{d}_0^{(2)} = [\|\mathbf{w}_1\|^2, \|\mathbf{w}_2\|^2, \ldots, \|\mathbf{w}_S\|^2]$. Then, the coordinates of the point $\tilde{\mathbf{p}}$ are given by

$$\mathbf{W}^T\tilde{\mathbf{p}} = -\frac{1}{2}\bigl(\mathbf{d}^{(2)} - \mathbf{d}_0^{(2)}\bigr). \tag{12}$$

The pre-image $\mathbf{p}$ of the reconstructed point $\hat{\phi}(\mathbf{y}_j)$ is finally obtained as

$$\mathbf{p} = \mathbf{E}\tilde{\mathbf{p}} + \frac{1}{S}\mathbf{Q}\mathbf{j}_S = \mathbf{E}\tilde{\mathbf{p}} + \mathbf{p}_0 \tag{13}$$

where $\mathbf{p}_0$ represents the mean of the selected neighbors. This method is usually applied considering that the number S of neighbors is less than the dimension M of the points in input space [4]. In that case M–S components of the point $\tilde{\mathbf{p}}$ vanish. Hence, the covariance matrix of the set of points $\mathbf{Q}$ can have at most S nonzero eigenvalues. Also note that the SVD of Eq. (12) represents the minimum norm solution. In that case, the second term of Eq. (13) representing the mean of the neighbors, might constitute the dominant term in the solution to the estimation of the pre-image of the reconstructed point $\hat{\phi}(\mathbf{y}_j)$.

2.2.2. Fixed point method

The central idea of this method [5] consists in computing the unknown pre-image $\mathbf{p}$ which minimizes the Euclidian distance in feature space by setting to zero the gradient of Eq. (6)

$$\frac{\partial\tilde{d}^{(2)}}{\partial\mathbf{p}} = \frac{\partial k(\mathbf{p},\mathbf{p})}{\partial\mathbf{p}} - 2\frac{\partial\mathbf{g}^T\mathbf{k}_p}{\partial\mathbf{p}} = 0. \tag{14}$$

Substituting the RBF kernel, the first term of the previous equation is zero as $k(\mathbf{p},\mathbf{p}) = 1$. Hence the zeros of the gradient are obtained by

$$\sum_{i=1}^{K} g_i(\mathbf{x}_i - \mathbf{p})\exp\left(\frac{\|\mathbf{x}_i - \mathbf{p}\|^2}{\sigma^2}\right) = \mathbf{X}(\mathbf{g}\diamondsuit\mathbf{k}_p) - \mathbf{p}\mathbf{g}^T\mathbf{k}_p = 0, \tag{15}$$

where $\diamondsuit$ represents the Hadamard product. The zeroes can thus be computed iteratively by the fixed point algorithm

$$\mathbf{p}_{t+1} = \frac{\mathbf{X}(\mathbf{g}\diamondsuit\mathbf{k}_{\mathbf{p}_t})}{\mathbf{g}^T\mathbf{k}_{\mathbf{p}_t}}. \tag{16}$$

The iterative procedure stops when $|\mathbf{p}_{t+1} - \mathbf{p}_t|$ is less than a threshold and/or a maximum number of iterations t is achieved. An equivalent iteration scheme results starting with polynomial or sigmoidal kernels. However it has been reported that the convergence of the resulting procedure could not be achieved with polynomial kernels [4]. Because of this the following discussion is restricted to RBF kernels.

2.2.3. Modified fixed point method

The fixed-point iteration can be started with any $\mathbf{p}_0$ chosen randomly, but in that case it often results in a slow convergence. Alternatively, it can be started with the given noisy point in the input space. However this is an option only when the signal-to-noise ratio is high [10]. Hence, in the following we propose a modified iteration scheme which yields superior results:

Motivation. Note that the denominator of Eq. (16) is formed by the dot product between the reconstructed point $\hat{\phi}(\mathbf{y}_j)$ and $\phi(\mathbf{p}_t)$. Thus an appropriate starting point in input space can be chosen using the nearest neighborhood

strategy borrowed from the distance method. This will avoid the numerical instability of having a very small or negative denominator.

Implementation. Note that with an RBF kernel the identification of the closest neighbors can be achieved with the dot products avoiding thus the computation of Euclidian distances in feature space. Computing the vector $\mathbf{r}$ of dot products of $\hat{\phi}(\mathbf{y}_j)$ with the training set $\mathbf{\Phi}$ yields

$$\mathbf{r} = \mathbf{g}^T \mathbf{K}. \tag{17}$$

As the dot product of every mapped data point with itself is normalized to one, the closest neighbors are obtained by identifying the set S of maximal dot products $(\hat{\phi}^T(\mathbf{y}_j)\phi(\mathbf{x}_i))$, $i = 1, \ldots, K$. The S closest neighbors, i.e., the ones that exhibit the largest dot products, are chosen. Selecting the corresponding points $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_S]$ in input space, the fixed-point iteration should start with

$$\mathbf{p}_0 = (1/S)\mathbf{Q}\mathbf{j}_S. \tag{18}$$

This strategy is more efficient than starting with the value computed by the distance method (see Eq. (13)) as suggested in [11]. The modified fixed point (mFP) algorithm will comprise the following initialization steps: the identification of S neighbors using Eq. (17) and the computation of the starting point of the fixed-point iteration (Eq. (16) with Eq. (18)).

3. KPCA of one-dimensional signals

The projective subspace techniques discussed so far are clearly not available for one-dimensional time series to suppress noise contributions. But many signal processing applications rely on *one-dimensional* signals, like the biomedical signals we are going to discuss.

The transformation of one-dimensional signals to multidimensional signals can be effected by a technique called *embedding*. It is used, for example, in singular spectrum analysis (SSA) methods [7,12]. After embedding, projective techniques can be applied resulting in a noise-reduced multidimensional signal and, by reverting the embedding process, finally in a one-dimensional signal. However, other goals like feature extraction and classification can be accomplished also with this type of analysis.

Following this strategy, one-dimensional sensor signals are embedded in the space of their time-delayed coordinates [8] to form a trajectory matrix, whose column vectors span the so called embedding space. To revert the embedding process, the last step is to transform the "denoised trajectory matrix," whose columns are formed by the pre-images of the reconstructed (= denoised) vectors, into a one-dimensional time series with N samples in the time domain.

3.1. Embedding

The embedding transformation can thus be regarded as a mapping which transforms a one-dimensional time series $x = (x[0], x[1], \ldots, x[N-1])$ to a multidimensional sequence of $K = N - M + 1$ lagged vectors

$$\mathbf{x}_k = \big[x[k-1+M-1], \ldots, x[k-1]\big]^T, \quad k = 1, \ldots, K \tag{19}$$

with $M < N$ being the corresponding embedding dimension. The lagged vectors then constitute the columns of the trajectory matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]$, which represents a Toeplitz matrix. The further processing of this data matrix $\mathbf{X}$ can be performed by KPCA considering each column a point in a space of dimension M (input space). Note that the latter space is often also called feature space but it is not to be confused with the space resulting from the nonlinear transformation of the input data effected by KPCA. In the following we will use the term feature space only for this latter space.

3.2. Diagonal averaging

After applying the described steps of KPCA to each column of the trajectory matrix (X), a new matrix of denoised data is obtained $\hat{\mathbf{X}}$. So, each $\phi(\mathbf{x}_k), k = 1, \ldots, K$, is projected in feature space onto L principal directions; it is then reconstructed using these projections, and finally its pre-image $\mathbf{p}_k$ in input space is estimated following one of the described methods. The denoised points $\mathbf{p}_k$ then form the columns of $\hat{\mathbf{X}}$, the "denoised trajectory matrix." But in general this matrix does not posses the characteristic Toeplitz structure anymore, i.e., 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 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 possible solutions to get a matrix with all diagonals equal [12]. The noise-reduced one-dimensional signal, $\hat{x}[n]$, is then obtained by reverting the embedding of matrix $\mathbf{X}_r$, i.e., by forming the signal with an entry of each descendent diagonal.

4. Numerical simulations

The algorithms were implemented in MATLAB using the toolbox provided by Franc [13], where basic pattern recognition tools and kernel methods can be found. In the following the methods discussed above will first be applied to the USPS data set to illustrate the denoising performance of the modified KPCA concerning the estimation of pre-images. The second example refers to the extraction of EOG artifacts from single channel EEG recordings.

4.1. Denoising a USPS data set

In this section the USPS data set[1] consisting of 16×16 handwritten digits is used. Thus the input data vector, $\mathbf{x}_k$ has dimension 256 and is formed by row concatenation of the original image after adding white Gaussian noise $N(0, \sigma_r^2)$ to it. Notice that adding noise to each digit with fixed variance, the signal-to-noise ratio (SNR) differs for each digit. Table 1, first column, shows the SNRs for the digits represented in the second row of Fig. 1 after adding noise with variance $\sigma_r^2 = 0.25$ to the original digits represented in the first row of Fig. 1. Each type of digit is denoised separately, i.e., a kernel matrix was computed for 300 randomly chosen examples of the digit.

Each time the kernel matrix was created using an RBF kernel with $\sigma = \max_i(\|\mathbf{x}_i - \mathbf{x}_c\|),\ i = 1, \ldots, K$, with $K = 300$ and $\mathbf{x}_c$ the mean of the data set. Then, denoising was achieved by projecting each mapped digit $\boldsymbol{\Phi}(\mathbf{x}_k)$

Table 1
SNR for different pre-image estimation methods $\sigma_r^2 = 0.25$

Digit	SNR				
	Image	mFP	Mean	Distance	L
0	6.202	8.910	8.686	8.766	8
1	−0.194	2.670	4.294	3.787	4
2	2.473	4.000	4.305	4.203	8
3	3.603	7.166	6.083	6.086	16
4	1.925	4.575	4.421	3.596	26
5	3.943	7.330	5.890	5.752	16
6	3.690	7.353	7.084	6.184	16
7	1.482	4.535	4.054	3.365	16
8	4.098	7.575	6.604	6.729	8
9	4.068	7.369	5.809	5.334	16

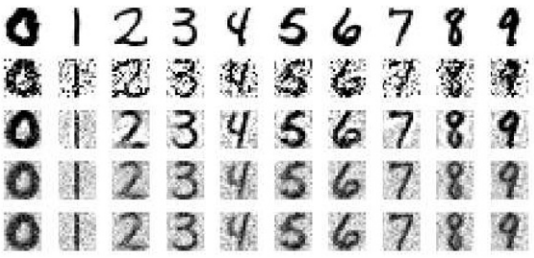

Fig. 1. Examples of each original, noisy and denoised digits using the different pre-image methods using $S = 10$ neighbors. *First line*—original digit, *second line*—noisy digit; the *following lines* the denoised digits using respectively the modified fixed point (mFP), mean and distance algorithms.

[1] http://www.kernel-machines.org.

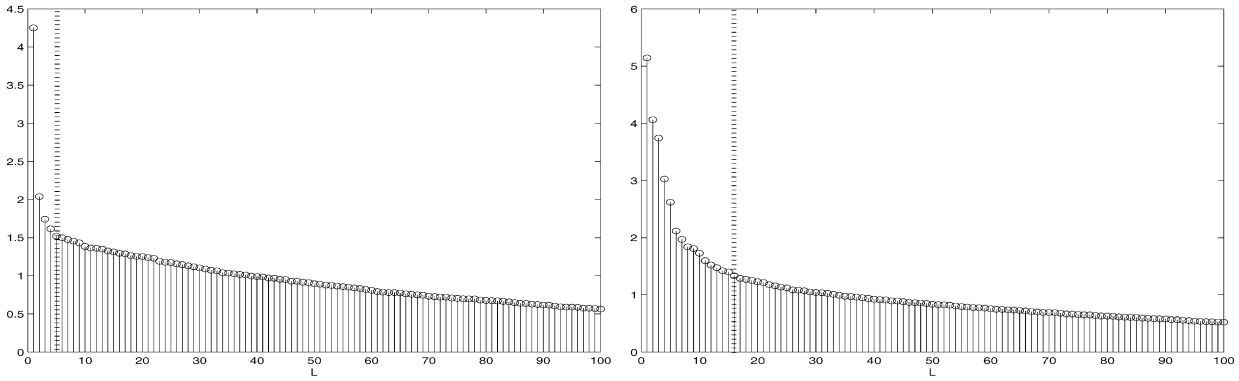

Fig. 2. Eigenspectra of kernel matrices for: *left*—digit 1, *right*—digit 5.

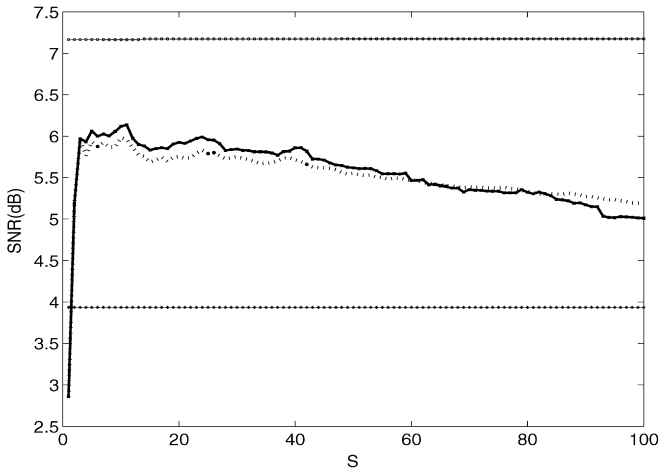

Fig. 3. SNR versus number of neighbors in the pre-image methods. *Top* trace—modified fixed point (mFP) algorithm; *bottom* trace—SNR of the noisy digit; *dotted* trace—mean of neighbors; *full* trace—distance algorithm.

onto the leading L eigenvectors corresponding to the largest eigenvalues. The eigenspectra of the kernel matrices are different. The number L of leading eigenvalues in each case was chosen according to the leveling off of the latter (see Fig. 2). The last column of Table 1 shows the number L of directions obtained for each digit. Finally, to yield a denoised version $\hat{\mathbf{x}}_k$ of the noisy digit $\mathbf{x}_k$, the pre-image $\hat{\mathbf{x}}_k$ of the reconstructed image of a digit $\hat{\mathbf{\Phi}}(\mathbf{x}_k)$ was estimated employing one of the following methods

- *Distance*: the distance method as proposed in [4], see Eq. (13).
- *Mean*: the mean of the nearest neighbors only within the distance method which corresponds to considering $\mathbf{p}_0$ as the pre-image.
- *mFP*: the fixed point method initialized with $\mathbf{p}_0$.

Figure 1 provides an illustration of the results of KPCA denoising for different digits using the different pre-image estimation methods discussed. Table 1 shows the SNRs between the original digit and the denoised version using $S = 10$ nearest neighbors. These values do not show any consistent tendency that could indicate a clear preference to any of the pre-image methods. But it is obvious that the modified fixed point (mFP) method yields better results in most cases and both the mean and the distance methods always yield very similar results.

Considering the dependence on S, it turns out that the distance method, proposed in [4], is the most sensitive concerning the number of nearest neighbors selected. Figure 3 illustrates in a more detailed manner this dependence of the SNR obtained with the mFP, the mean and the distance methods on the number of nearest neighbors.

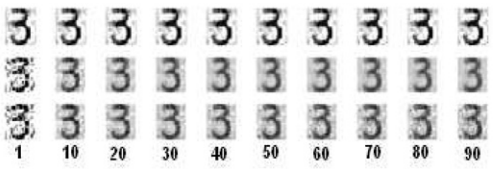

Fig. 4. Denoised digit using different values for S: *top*—modified fixed point (mFP) algorithm; *middle*—mean of neighbors; *bottom*—distance method.

Table 2
SNR as function of the number of nearest neighbors S for noise variance $\sigma^2 = 0.25$ and the pre-image estimation methods discussed

	S	SNR	
		Mean	Variance
Noisy image		2.322	3.026
mFP	1	4.5164	2.4323
	10	4.5287	2.3575
	100	4.4144	2.4711
Mean	1	2.1092	2.4853
	10	4.7978	1.8510
	100	4.2323	1.7979
Distance	1	2.1092	2.4853
	10	4.7942	1.8258
	100	3.3670	1.8589

The results obtained with the method of mean nearest neighbors are always close to the ones resulting from the distance method. And both methods show a decline of the SNR with increasing S. On the other hand, the modified fixed point is not dependent on the starting point (number of neighbors). Figure 4 illustrates the influence of the number of neighbors on the SNRs obtained. Naturally when the number of neighbors increases the solution provided by the mean of neighbors encompasses a smoothing of the background thus leading to a higher noise level.

The methods are also evaluated using the whole set of digits. The mean SNRs and their respective variances using different S are collected in Table 2. The results obtained with the modified fixed point method of course do not depend on S while in the distance method there is a difference of $\approx$1.5 dB between $S = 10$ and 100. But considering $S = 1$, both the mean and the distance methods correspond to choosing, within the given data set, the digit whose image is closest to $\hat{\phi}(\mathbf{y}_j)$. Then the SNR is similar to the one of the original noisy digit. However, considering $S = 1$ in the modified fixed point method, a reliable solution is achieved with the algorithm converging faster than when it is initialized randomly.

4.2. EEG data

Biomedical signals are often contaminated with artifacts which severely distort the signals to be investigated. As an example we will study the removal of prominent EOG artifacts from EEG recordings. A segment of the signal of 12 s of duration containing high-amplitude EOG artifacts was considered and divided into 4 sub-segments with $N = 384$ samples. KPCA was applied separately to each sub-segment. The one-dimensional signal was embedded in $M = 11$ dimensions, and the number $L = 6$ principal components for reconstruction was the same for all subsegments. The pre-image estimated in input space then obviously corresponds to the embedded, multidimensional version of the one-dimensional EOG contaminating the original EEG recording. This one-dimensional EOG can be obtained after reverting the embedding process. Finally the extracted EOG signal is subtracted from the original EEG recording to yield a corrected version of the EEG. The first experiment illustrates the impact of the method to estimate the pre-image upon the results and the second experiment illustrates the performance of the KPCA method when compared with local SSA.

4.2.1. Estimation of the pre-image

Visual inspection of the extracted signals confirmed that the results strongly depend on the method to estimate the pre-image, corroborating results obtained with the USPS data set. Further experiments showed that the performance of the distance method is strongly dependent on the number S of neighbors, yielding in some cases unreliable solutions (see Fig. 8, for example). The results of the modified fixed point algorithm are illustrated in Fig. 5 and are independent of the number of neighbors used to estimate the starting point. The rate of convergence can be improved considerably when using $\mathbf{p}_0$ as the starting point instead of using a random initialization (see Fig. 6).

To provide a global overview of the performance of the methods, correlation coefficients between a reference signal and signals resulting from changing either the method of estimating the pre-image or varying the number of neighbors S are calculated. Figure 7 shows the results considering as reference the signal obtained with the modified fixed point method initialized with the best match in the training set ($S = 1$). The correlation coefficient for mFP equals $cc(\text{mFP}) = 1$ whatever is the number of neighbors. Note that with $S = \{3, 6, 7, 12\}$ neighbors, the distance method does not yield a reliable solution. Figure 8 illustrates this for one such solution where we can confirm that the extracted signal does not correspond to the EOG component. Furthermore, note that if $S > M$, the correlation coefficients are small. If $S < 20$, the result of the mean method is very close to the result of the modified fixed point

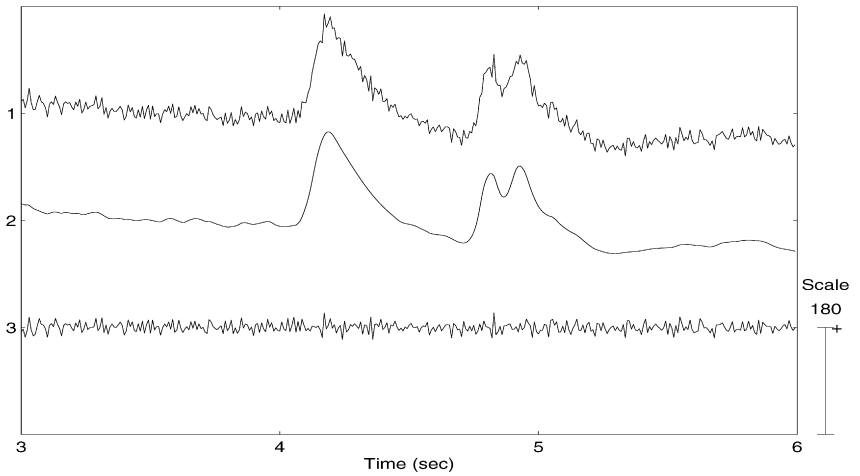

Fig. 5. A segment of the EEG signal processed with KPCA (reconstructed with $L = 6$ principal components and using mFP to estimate the pre-image (*top*—the original EEG, *middle*—the extracted EOG artifact, *bottom*—the corrected EEG).

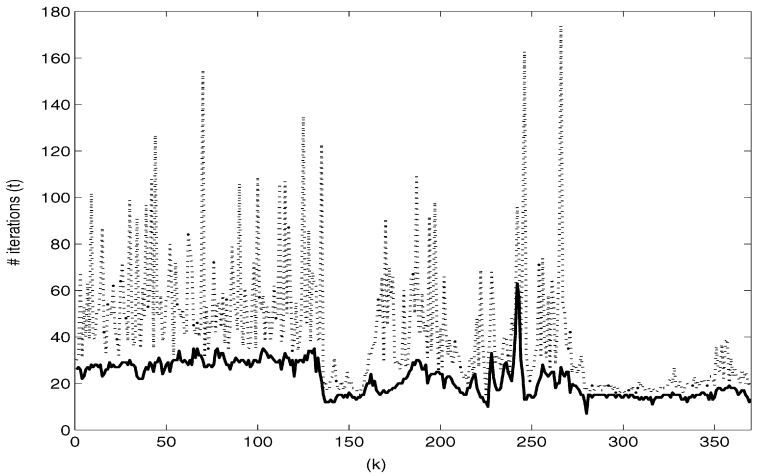

Fig. 6. The number of iterations needed for denoising data vector $\mathbf{x}_k$, $k = 1, \ldots, K = 374$ using the modified fixed point (full line) or the fixed point with random initialization (dotted line).

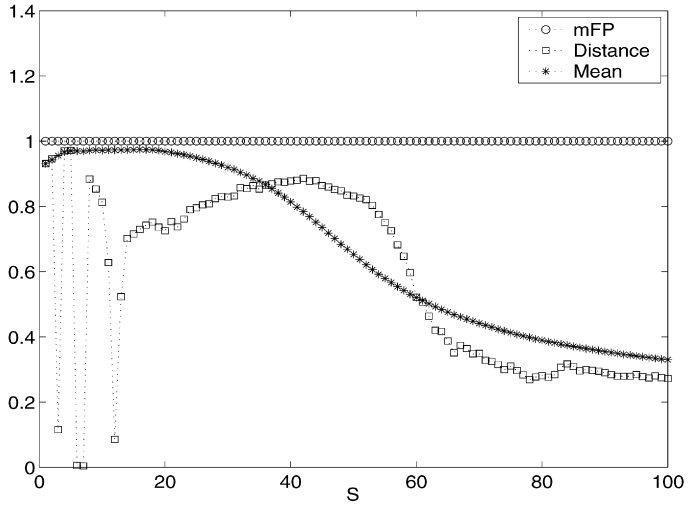

Fig. 7. Correlation coefficients between the signal obtained with mFP ($S = 1$) and all the signals obtained changing the pre-image method and/or varying S.

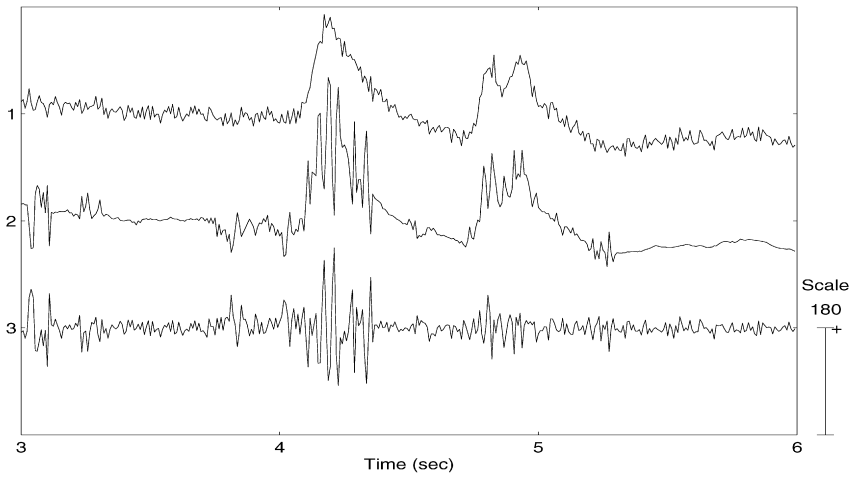

Fig. 8. A segment of the EEG signal processed with KPCA (reconstructed with $L = 6$ principal components) using the distance method ($S = 12$) to estimate the pre-image (*top*—the original EEG, *middle*—the extracted EOG artifact, *bottom*—the corrected EEG).

algorithm (mFP). This is corroborated by estimates of the power spectral densities of both the corrected EEG signals and the extracted EOG as obtained with the mean and modified fixed point algorithms, respectively.

4.2.2. KPCA compared to local SSA

In previous work [14] a modification of the well known singular spectrum analysis (SSA) was proposed for denoising and feature extraction from one-dimensional signals. The method is called local singular spectrum analysis (local SSA). The method is briefly summarized in appendix but consists in a local PCA decomposition performed in input space while KPCA performs a PCA after transforming the input data nonlinearly to a feature space. The algorithm is applied to the same segments of the EEG recordings, contaminated by large EOG artifacts, to which also KPCA has been applied. Using an embedding dimension $M = 41$ and an optimal number of clusters $q = 6$, local SSA achieves an artifact separation (not shown here) which is upon visual inspection indistinguishable from the results of the KPCA analysis shown in Fig. 8a. Comparing the power spectral densities resulting from both the local SSA and the modified KPCA, Fig. 9 shows that the low frequency content of the corrected EEG is affected differently by both methods. The modified KPCA seems to preserve more spectral information in the very low frequency regime ($f \leqslant 3$ Hz) but yields

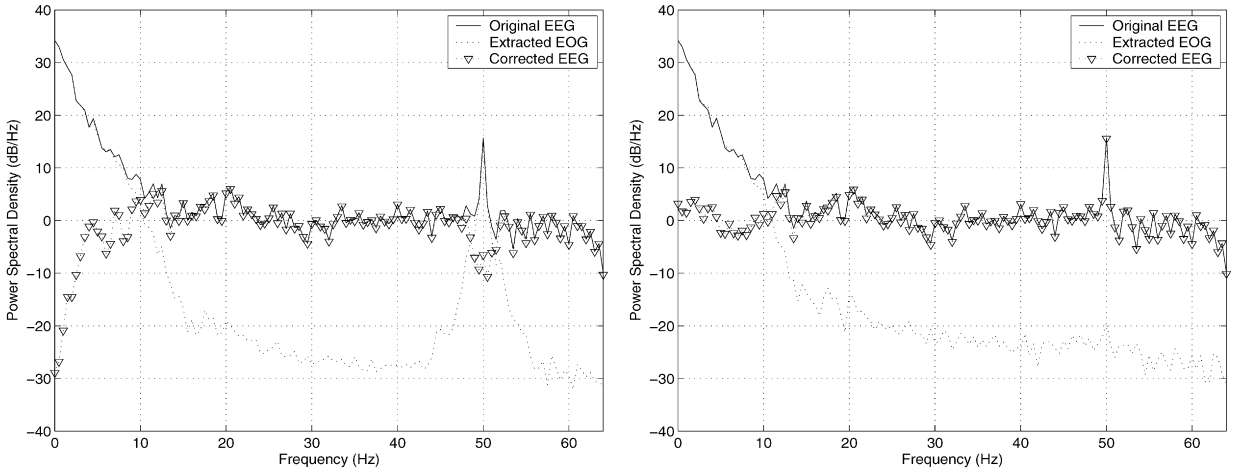

Fig. 9. Power spectral densities (psd) resulting from: *left*—local SSA ($q = 6$), *right*—kernel-PCA.

similar results at higher frequencies: the correlation coefficient (cc) between the extracted EOGs resulting from both the KPCA and the local SSA is $cc = 0.999$ and the one between the corrected EEGs is $cc = 0.833$. The latter value results from the fact that local SSA also extracts the 50 Hz line interference while the KPCA does not.

5. Conclusions

In this work we considered the estimation of the pre-image within KPCA. This becomes important in applications like denoising. We compared the two methods of pre-image estimation discussed in the literature and suggested two simple modifications yielding a hybrid approach which proofed very effective in practical applications with RBF kernels: (a) the mean of neighbors; (b) the mean of nearest neighbors as starting point to fixed point (mFP). The comparison on a denoising task revealed that the solution obtained by the distance method strongly depends on the number S of nearest neighbors chosen and sometimes does not yield reliable results. If S is smaller than the dimension of the data space, the solution can often be closely approximated by simply choosing the mean of the nearest neighbors which speeds up computation considerably. Concerning the fixed point algorithm a random initialization as suggested in the literature often results in very slow convergence. Initializing the algorithm with the mean of the nearest neighbors considerably speeds up convergence and yields a very robust algorithm. The methodology used for denoising can as well be applied to the problem of extracting prominent artifacts from one-dimensional signals like single channel EEG recordings. To this end, we adapt kernel principal component analysis to deal with one-dimensional signals by embedding them into the space of their delayed coordinates. We demonstrate the performance of the proposed KPCA algorithms to remove high-amplitude EOG interferences from EEG signals. The result of the artifact removal strongly depends on the way the pre-image is estimated. We show that again the fixed point algorithm initialized with the mean of the nearest neighbors (mFP) of each mapped data point is the most robust and reliably extracts the EOG artifacts. In a previous work [14,15] local SSA was proposed for denoising and feature extraction from one-dimensional signals. We present a short comparison to this method and show that the low frequency content of the corrected EEG is affected differently by both methods.

Acknowledgments

A.R. Teixeira received a Ph.D. Scholarship (SFRH/BD/28404/2006) supported by the Portuguese Foundation for Science and Technology (FCT). This work was also supported by grants from DAAD and CRUP which is gratefully acknowledged.

Appendix A

For convenience the main steps of local SSA are reviewed [15].

- 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 [16]). The set of indices of the columns of $\mathbf{X}$ is subdivided into q disjoint subsets $c_1, c_2, \ldots, c_q$. Thus sub-trajectory matrix $\mathbf{X}^{(c_i)}$ is formed with N_{c_i} columns of the matrix $\mathbf{X}$ which belong to the subset of indices c_i.
- A covariance matrix is computed in each cluster. And the eigenvalue decomposition of each the covariance matrices is computed. The 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.
- The clustering is reverted by forming an estimate $\hat{\mathbf{X}}$ of the reconstructed, noise-free trajectory matrix using the columns of the extracted sub-trajectory matrices, $\hat{\mathbf{X}}^{c_i},\ i = 1, \ldots, q$, according to the contents of subsets c_i. Then the corresponding one dimensional signal is obtained as explained in Section 3.2.

References

[1] P. Gruber, K. Stadlthanner, M. Böhm, F.J. Theis, E.W. Lang, A.M. Tomé, A.R. Teixeira, C.G. Puntonet, J.M. Górriz Saéz, Denoising using local projective subspace methods, Neurocomputing 69 (2006) 1485–1501.

[2] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based algorithms, IEEE Trans. Neural Networks 12 (2) (2001) 181–202.

[3] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.

[4] J.T. Kwok, I.W. Tsang, The pre-image problem in kernel methods, IEEE Trans. Neural Networks 15 (6) (2004) 1517–1525.

[5] B. Schölkopf, S. Mika, C.J. Barges, P. Knirsch, K.-R. Müller, G. Ratsch, A.J. Smola, Input space versus feature space in kernel-based methods, IEEE Trans. Neural Networks 10 (5) (1999) 1000–1016.

[6] Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement, IEEE Trans. Acoust. Speech Signal Process. 3 (4) (1995) 251–266.

[7] M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, et al., Advanced spectral methods for climatic time series, Rev. Geophys. 40 (1) (2002) 3.1–3.41.

[8] C.H. You, S.N. Koh, S. Rahardja, Signal subspace speech enhancement for audible noise reduction, in: ICASPP 2005, vol. I, Philadelphia, USA, 2005, pp. 145–148.

[9] J.C. Gower, Adding a point to vector diagram in multivariate analysis, Biometrika 55 (1968) 582–585.

[10] T. Takahashi, T. Kurita, Robust de-noising by kernel PCA, in: J.R. Dorronsoro (Ed.), ICANN2002, LNCS, vol. 2415, Springer-Verlag, Madrid, 2002, pp. 739–744.

[11] K.I. Kim, M.O. Franz, B. Schölkopf, Itereative kernel component analysis for image modeling, IEEE Trans. Pattern Anal. Machine Intel. 27 (9) (2005) 1351–1365.

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

[13] V. Franc, V. Hlaváč, Stastical pattern recognition toolbox for Matlab, 2004.

[14] A.R. Teixeira, A.M. Tomé, E.W. Lang, P. Gruber, A. Martins da Silva, On the use of clustering and local singular spectrum analysis to remove ocular artifacts from electroencephalograms, in: IJCNN2005, Montréal, Canada, 2005, pp. 2514–2519.

[15] A.R. Teixeira, A.M. Tomé, E.W. Lang, P. Gruber, A. Martins da Silva, Automatic removal of high-amplitude artifacts from single-channel electroencephalograms, Comp. Methods Programs Biomed. 83 (2) (2006) 125–138.

[16] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford Univ. Press, Oxford, 1995.

A.R. Teixeira received the diploma degree in mathematics applied to technology from the University of Porto, Porto, Portugal, in 2003, and the M.Sc. degree in electronics and telecommunications at the University of Aveiro, Aveiro, Portugal, in 2005. Currently, she is doing the Ph.D. in electrical engineering in the Signal Processing Lab of IEETA/DETI at the University of Aveiro. Her research interests include biomedical digital signal processing and principal and independent component analysis.

A.M. Tomé received the Ph.D. degree in electrical engineering from the University of Aveiro, Aveiro, Portugal, in 1990. Currently, she is an Associate Professor of electrical engineering with the DETI/IEETA of the 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.

K. Stadlthanner received the diploma degree in physics from the University of Regensburg in 2003. He also received the Ph.D. degree in computer science from the University of Granada, Spain, in 2006 and the Ph.D. degree in biophysics from the University of Regensburg, Germany, in 2007. He is currently working in a research laboratory with Phillips AG, Aachen. His scientific interests are in the fields of biological signal processing and image analysis by means of machine learning techniques and neural networks.

E.W. Lang received the diploma degree in physics in 1977, the Ph.D. degree in physics in 1980, and habilitated in biophysics in 1988 from the University of Regensburg, Regensburg, Germany. He is an Adjunct Professor of biophysics at the University

of Regensburg, where he is heading the Neuro- and Bioinformatics Group. Currently, he serves as an Associate Editor of *Neurocomputing* and *Neural Information Processing—Letters and Reviews*. His current research interests include biomedical signal and image processing, independent component analysis and blind source separation, neural networks for classification and pattern recognition, and stochastic process limits in queueing applications.