Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 2381 – 2398
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Multivariate denoising using wavelets and principal component analysis

Mina Aminghafari[a, b], Nathalie Cheze[a, c], Jean-Michel Poggi[a, d, *]

[a] *Laboratoire de Mathématique - U.M.R. C 8628, "Probabilités, Statistique et Modélisation", Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France*
[b] *Amirkabir University of Technology, Tehran, Iran*
[c] *Université Paris 10-Nanterre, Modal'X, France*
[d] *Université Paris 5, France*

Available online 18 January 2005

Abstract

A multivariate extension of the well known wavelet denoising procedure widely examined for scalar valued signals, is proposed. It combines a straightforward multivariate generalization of a classical one and principal component analysis. This new procedure exhibits promising behavior on classical bench signals and the associated estimator is found to be near minimax in the one-dimensional sense, for Besov balls. The method is finally illustrated by an application to multichannel neural recordings.

Keywords: Multivariate denoising; Wavelets; PCA

1. Introduction

On one hand, denoising algorithms based on wavelet decompositions are a popular method for one-dimensional statistical signal extraction and filtering. On the other hand, principal component analysis (PCA) is among the most notorious data-analysis tools designed to simplify multidimensional data by tracking new factors supposed to capture the main features.

* Corresponding author. Laboratoire de Mathématique - U.M.R. C 8628, "Probabilités, Statistique et Modélisation", Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France.
E-mail addresses: mina.aminghafari@math.u-psud.fr (M. Aminghafari), nathalie.cheze@u-paris10.fr (N. Cheze), jean-michel.poggi@math.u-psud.fr (J.-M. Poggi).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.12.010

This paper proposes a multivariate extension of wavelet denoising procedures, combining a straightforward multivariate generalization of the classical one for scalar valued signals and principal component analysis. This proposal takes place among the various recent approaches combining wavelet strategies and data analytic tools to cope with the problem of feature extraction in regression models. Numerous applied situations strongly motivate this interest. Let us mention some of them together with some references focusing on a wavelet-data analysis approach: spectral calibration problems (Vannucci et al., 2003), multivariate statistical process control (see Bakshi, 1998; Teppola and Minkkinen, 2000), blind source separation (Roberts et al., 2004), functional magnetic resonance imaging (fMRI) analysis (Meyer and Chinrungrueng, 2003), spike detection and sorting (Oweiss and Anderson, 2001).

This paper focuses on multivariate wavelet denoising and deals with regression models of the form $X(t) = F(t) + \varepsilon(t)$, where the observation X is p-dimensional, F is deterministic and is the signal to be recovered and ε is a spatially correlated noise. This kind of model is well suited for situations for which such an additive spatially correlated noise is realistic. For example, a longitudinal study on p subjects, the analysis of a part a fMRI region (involving p voxels) or the noise reduction in multichannel neural recordings (using p channels). Let us be a little bit more precise on this last example which is chosen to illustrate the behavior of the proposed procedure on a real world data set, at the end of the paper. Following Oweiss and Anderson (2001), extra-cellular neural recordings can be modeled as an invariant deterministic signal and an additive noise which obscures neural discharges from cells of interest. This noise contains a component exhibiting spatial correlation coming from background activity caused by neural cells.

To close this introduction, let us recall some facts about classical univariate wavelet denoising dealing both with signal processing and functional estimation in statistics and which is of interest in various applied fields. Valuable references are the books (Mallat, 1998; Percival and Walden, 2000; Vidakovic, 1999) and the survey paper (Antoniadis, 1997). For basics on wavelets, we refer the reader to Mallat (1998) or Misiti et al. (2003) for example.

The simplest considered model is of the following form:

$$X(t) = f(t) + \varepsilon(t), \quad t = 1, \ldots, n, \tag{1}$$

where $(X(t))_{1 \leqslant t \leqslant n}$ is observed, $(\varepsilon(t))_{1 \leqslant t \leqslant n}$ is a centered Gaussian white noise of unknown variance σ^2 and f is an unknown function to be recovered through the observations.

For a given orthogonal wavelet basis denoted by $\left((\phi_{J,k})_{k \in Z}, (\psi_{j,k})_{1 \leqslant j \leqslant J, k \in Z} \right)$ where ψ is a wavelet, ϕ the associated scaling function, J a suitably chosen decomposition level and where $g_{j,k}(x) = 2^{-j/2} g\left(2^{-j}x - k\right)$, wavelet denoising proceeds in three steps:

- *Step* 1: Compute the wavelet decomposition of the observed signal up to level J;
- *Step* 2: Threshold conveniently the wavelet detail coefficients;
- *Step* 3: Reconstruct a denoised version of the original signal, from the thresholded detail coefficients and the approximation coefficients, using the inverse wavelet transform.

Various strategies are available (see the survey paper Antoniadis et al., 2001) to perform this task and the asymptotic performance of the associated estimators is the minimax one up to

a logarithmic factor, for large classes of functions simultaneously (let us mention that block thresholding, see Hall et al., 1999, could be used to remove the logarithmic factor). For simplicity and since only relative performance between the proposed multivariate procedures are of interest, we restrict our attention to the so-called universal threshold (introduced by Donoho and Johnstone, 1994) which is of the form $\hat{\sigma}\sqrt{2\log(n)}$, where $\hat{\sigma}$ is an estimator of σ based on the detail coefficients at level 1 (the finest one). Such methods are effective because functions f belonging to various general classes are such that they admit a sparse wavelet representation (Kerkyacharian and Picard, 2000). So the energy of f is mainly concentrated in a few large wavelet coefficients which are adaptively selected by this procedure since the coefficients below the threshold are attributable to the additive noise.

This paper dedicated to a multivariate denoising procedure that takes into account the correlation structure of the noise, is organized in two main sections. Section 2 proposes a first denoising procedure which is a direct generalization of the one-dimensional strategy. The method is based on a change of basis followed by a classical one-dimensional soft-thresholding. This new procedure exhibits promising behavior on classical test signals and the associated estimator is found to be near minimax in the one-dimensional sense, for Besov balls. The change of basis is obtained from the diagonalization of a robust estimate of the noise covariance matrix given by the minimum covariance determinant estimator based on the matrix of finest details.

Section 3 first recalls the multiscale PCA proposed by Bakshi (1998) for statistical process control purposes. This scheme is discussed and then a second denoising procedure combining wavelets and PCA is proposed. The introduction of a PCA step try to take advantage of the deterministic relationships between the signals, leading to an additional denoising effect. It is then illustrated by some simulation examples and by an application to multichannel neural recordings.

2. Multivariate wavelet denoising

2.1. Procedure

Let us consider the following p-dimensional model:

$$X(t) = f(t) + \varepsilon(t), \quad t = 1, \ldots, n, \tag{2}$$

where $X(t)$, $f(t)$, $\varepsilon(t)$ are of size $1 \times p$ and $(\varepsilon(t))_{1 \leqslant t \leqslant n}$ is a centered Gaussian white noise with unknown covariance matrix $E\left(\varepsilon(t)^{\mathrm{T}}\varepsilon(t)\right) = \Sigma_\varepsilon$. Each component of $X(t)$ is of the previous form (1), for $1 \leqslant i \leqslant p$:

$$X^i(t) = f^i(t) + \varepsilon^i(t), \quad t = 1, \ldots, n,$$

where f^i belongs to some functional space (typically L^2 or Besov spaces).

The covariance matrix Σ_ε, supposed to be positive definite, captures the stochastic link between the components of $X(t)$ and models the spatial correlation.

Of course, it is possible to denoise each component of $X(t)$, ignoring the spatial correlation structure of the noise components. This method, called the direct one in the sequel, is taken as a starting point.

The first proposal for multivariate denoising is described as follows, starting from an $n \times p$ matrix X containing p signals (the columns of X) of dyadic length n supposed to be such that $n \gg p$.

- *Step* 1: Perform the wavelet decomposition at level J of each column of X. This step produces $J+1$ matrices $D_1, \ldots, D_J$ containing the detail coefficients at level 1 to J of the p signals and the approximation coefficients A_J of the p signals. Matrices D_j and A_J are, respectively, of size $n2^{-j} \times p$ and $n2^{-J} \times p$;
- *Step* 2: Define $\hat{\Sigma}_\varepsilon$ an estimator of the noise covariance matrix and then compute the SVD of $\hat{\Sigma}_\varepsilon$ providing an orthogonal matrix V such that $\hat{\Sigma}_\varepsilon = V \Lambda V^{\mathrm{T}}$ where $\Lambda = diag(\lambda_i, 1 \leqslant i \leqslant p)$. Apply to each detail after change of basis using V (namely $D_j V$, $1 \leqslant j \leqslant J$), the p univariate thresholding strategies using the threshold $t_i = \sqrt{2\lambda_i \ \log(n)}$ for the ith column of $D_j V$;
- *Step* 3: Reconstruct a denoised matrix $\check{X}$ (or equivalently an estimator $\hat{f}$ of f), from the simplified detail and approximation matrices, by changing of basis using V^{T} and inverting the wavelet transform.

Since the wavelet transform (in time) and the change of basis (in space) commute, it appears that this straightforward generalization of the univariate denoising is equivalent to: first, change of basis to decorrelate the p components of the noise and second, to apply p univariate wavelet denoising. Using the orthogonal matrix V defined in step 2, let us change of basis the model (2) by defining $\tilde{X}(t) = X(t)V$ and similarly $\tilde{f}(t)$ and $\tilde{\varepsilon}(t)$, leading to, for $t = 1, \ldots, n$

$$\tilde{X}(t) = \tilde{f}(t) + \tilde{\varepsilon}(t), \tag{3}$$

which involves a noise with p near uncorrelated components as soon as $\hat{\Sigma}_\varepsilon$ is a consistent estimator of Σ_ε. Then it suffices to denoise each component and to change of basis using V^{T} to get back to the original one.

Surprisingly this scheme performs better than the direct one performing the denoising of each component of (2) ignoring correlation structure, this will be illustrated in Section 2.4. Indeed, since $trace\left(F\hat{\Sigma}_\varepsilon^{-1}F^{\mathrm{T}}\right) = trace\left(\tilde{F}\Lambda^{-1}\tilde{F}^{\mathrm{T}}\right)$, where F and $\tilde{F}$ are defined as in (7) and (8), it turns out that the p signal-to-noise ratios after change of basis are not necessarily better than initial ones. It seems that the crucial point is that the change of basis merges all the original functions in each $\tilde{f}^i$ introducing a kind of redundancy leading to better estimate the main features (such as change points or singularities) of the original signals f^i to be recovered.

2.2. Theoretical results

In this section devoted to theoretical results, we consider the estimator, denoted by $\hat{f}^*$, introduced by the procedure in Section 2.1 but where the noise covariance matrix is supposed to be known and where soft thresholding is performed with the function

$$d_t(x) = \mathrm{sgn}(x)(|x| - t)_+, \tag{4}$$

where t is the positive real threshold.

First, we compute the quadratic risk of the associated estimator.

Second, we adopt the minimax viewpoint in order to examine the estimation procedure performance when the functions f^i belong to Besov balls. We prove that our estimator is near minimax in the one-dimensional sense but inherits the worst regularity rate among the p functions.

The univariate loss function, is defined by

$$R\left(f^i, \hat{f}^i\right) = \sum_{m=1}^{n} E\left(f^i(m) - \hat{f}^i(m)\right)^2. \tag{5}$$

A straightforward extension to the multivariate case is the sum over the p functions

$$R_M\left(f, \hat{f}\right) = \sum_{i=1}^{p} R\left(f^i, \hat{f}^i\right). \tag{6}$$

2.2.1. Risk computation

Let X be the $n \times p$ matrix of the n observations from the p-dimensional vector $X(t)$, then X can be written as

$$X = F + \varepsilon, \tag{7}$$

where F and ε are $n \times p$ matrices, associated with $f(t)$ and $\varepsilon(t)$, respectively.

Let V be the orthogonal matrix such that $V^{\mathrm{T}}\Sigma_\varepsilon V$ is diagonal. Then, we have

$$XV = \tilde{X} = \tilde{F} + \tilde{\varepsilon}. \tag{8}$$

For a given orthogonal compactly supported wavelet, the discrete wavelet transform (DWT) of the matrix $\tilde{X}$ can be written as follows:

$$W\tilde{X} = W\tilde{F} + W\tilde{\varepsilon}, \tag{9}$$

where W is an $n \times n$ orthogonal matrix defined using the wavelet and scaling filters impulse responses (see Percival and Walden, 2000, p. 57). Let us denote by Y^i and Y_m the ith column and the mth row of matrix Y, respectively. Then, $W_m\tilde{X}^i$ and $W_m\tilde{F}^i$ are the mth discrete wavelet transform coefficients of $\tilde{X}^i$ and $\tilde{F}^i$, respectively.

Proposition 1. *Let* $I(i) = \left\{1 \leqslant m \leqslant n, |W_m\tilde{X}^i| > t_i\right\}$. *Then, the risk of estimator* $\hat{f}^*$ *is given by*

$$R_M\left(f, \hat{f}^*\right) = \sum_{i=1}^{p}\left(card(I(i))\left(\lambda_i + t_i^2\right) + \sum_{m \notin I(i)}\left(W_m\tilde{F}^i\right)^2\right). \tag{10}$$

Proof. The estimator $\hat{f}^*$ is defined using the p estimators $\hat{\tilde{f}}^i$ after the change of basis using V. Then using (8) and (9), $\hat{f}^*$ leads to $\hat{F}$. Matrices $\hat{F}$ and $\hat{\tilde{F}}$ are related by

$$\hat{F} = \hat{\tilde{F}}V^{\mathrm{T}}, \tag{11}$$

where the estimator $\hat{\tilde{F}}$ of $\tilde{F}$ is such that $\hat{\tilde{F}} = W^{\mathrm{T}}U$. The components of the n-dimensional column vector U^i are $d_{t_i}\left(W_m\tilde{X}^i\right)$ where d_{t_i} is the soft thresholding function defined in (4) with threshold t_i.

The risks in the new basis and in the initial basis are the same

$$R_M\left(f, \hat{f}\right) = R_M\left(\tilde{f}, \hat{\tilde{f}}\right). \tag{12}$$

Indeed,

$$\begin{aligned}
R_M\left(f, \hat{f}\right) &= trace\left(E\left(\left(F-\hat{F}\right)\left(F-\hat{F}\right)^{\mathrm{T}}\right)\right)\\
&= trace\left(E\left(\left(F-\hat{F}\right)VV^{\mathrm{T}}\left(F-\hat{F}\right)^{\mathrm{T}}\right)\right)\\
&= trace\left(E\left(\left(\tilde{F}-\hat{\tilde{F}}\right)\left(\tilde{F}-\hat{\tilde{F}}\right)^{\mathrm{T}}\right)\right)\\
&= trace\left(E\left(\left(\tilde{F}-\hat{\tilde{F}}\right)^{\mathrm{T}}\left(\tilde{F}-\hat{\tilde{F}}\right)\right)\right)\\
&= \sum_{i=1}^{p}\left(E\left(\left(\tilde{F}^i-\hat{\tilde{F}}^i\right)^{\mathrm{T}}\left(\tilde{F}^i-\hat{\tilde{F}}^i\right)\right)\right) = \sum_{i=1}^{p} R\left(\tilde{f}^i, \hat{\tilde{f}}^i\right).
\end{aligned}$$

Since W is an orthogonal matrix, we can derive

$$\begin{aligned}
R\left(\tilde{f}^i, \hat{\tilde{f}}^i\right) &= E\left(\left(\tilde{F}^i-\hat{\tilde{F}}^i\right)^{\mathrm{T}}\left(\tilde{F}^i-\hat{\tilde{F}}^i\right)\right)\\
&= E\left(\left(\tilde{F}^i-\hat{\tilde{F}}^i\right)^{\mathrm{T}}W^{\mathrm{T}}W\left(\tilde{F}^i-\hat{\tilde{F}}^i\right)\right)\\
&= E\left(\left(W\tilde{F}^i-U^i\right)^{\mathrm{T}}\left(W\tilde{F}^i-U^i\right)\right)\\
&= \sum_{m=1}^{n} E\left(W_m\tilde{F}^i - d_{t_i}\left(W_m\tilde{X}^i\right)\right)^2,
\end{aligned}$$

where $\left(W_m\tilde{F}^i - d_{t_i}\left(W_m\tilde{X}^i\right)\right)^2 = \begin{cases} \left(W_m\tilde{\varepsilon}^i + t_i\ \mathrm{sgn}\left(W_m\tilde{X}^i\right)\right)^2 & \text{if } \left|W_m\tilde{X}^i\right| > t_i \\ \left(W_m\tilde{F}^i\right)^2 & \text{if } \left|W_m\tilde{X}^i\right| \leqslant t_i \end{cases}$ and $E\left(W_m\tilde{\varepsilon}^i\right)^2 = \lambda_i$ since $E\left(\tilde{\varepsilon}^{\mathrm{T}}\tilde{\varepsilon}\right) = \Lambda$. Then we obtain (10). □

Remark 1. If we define the risk as $R_M^{\Sigma}\left(f, \hat{f}\right) = trace\left(E\left(\left(F-\hat{F}\right)\Sigma^{-1}\left(F-\hat{F}\right)^{\mathrm{T}}\right)\right)$ including explicitly the covariance structure, instead of (6), then result (10) becomes

$$R_M^{\Sigma}\left(f, \hat{f}^*\right) = \sum_{i=1}^{p}\frac{1}{\lambda_i}\left(card(I(i))\left(\lambda_i + t_i^2\right) + \sum_{m\notin I(i)}\left(W_m\tilde{F}^i\right)^2\right).$$

This comes from

$$R_M^{\Sigma}\left(\tilde{f},\hat{\tilde{f}}\right)=trace\left(E\left(\left(\tilde{F}-\hat{\tilde{F}}\right)\Lambda^{-1}\left(\tilde{F}-\hat{\tilde{F}}\right)^{\mathrm{T}}\right)\right)=\sum_{i=1}^{p}\frac{1}{\lambda_i}R\left(\tilde{f}^i,\hat{\tilde{f}}^i\right).$$

2.2.2. Near minimaxity

The risk (10) depends on f. Then, it is traditional in nonparametric estimation to calculate its maximum over a smoothness class and to compare it to the risk of minimax estimators. We consider here functions belonging to balls of Besov spaces denoted $B_{r,q}^{\alpha}$ with norm $||.||_{r,q}^{\alpha}$.

The following result is based on Theorem 1.2 of Donoho (1995) assuming that the considered orthogonal wavelet ψ belongs to the class C^k and has d vanishing moments, for some positive integers k and d.

Proposition 2. *Suppose that ε is a centered Gaussian white noise with covariance matrix Σ_ε. Let $f_i\in\mathscr{F}_C$, $i=1,\ldots,p$ where $\mathscr{F}_C$ denotes the Besov ball of scalar valued functions $\left\{g,\|g\|_{r,q}^{\alpha}\leqslant C\right\}$ with $1/r<\alpha<\min(k,d)$ and $C>0$. Then, there is a positive constant denoted by c_1 depending on $\mathscr{F}_C$ and ψ, such that*

$$\sup_{f\in(\mathscr{F}_C)^p}R_M\left(f,\hat{f}^*\right)\leqslant c_1 p\,\log(n)\inf_{\hat{g}}\sup_{g\in\mathscr{F}_{C\sqrt{p}}}R\left(g,\hat{g}\right),\tag{13}$$

where $\hat{f}^*$ is the wavelet based estimator described previously using universal soft thresholding defined in step 2 (*see Section* 2.1).

Proof. Since $\tilde{f}^i=\sum_{j=1}^{p}V_j^i f^j$, $f^j\in B_{r,q}^{\alpha}$ implies that $\tilde{f}^i\in B_{r,q}^{\alpha}$ (see DeVore and Lorentz, 1993, pp. 54–56). More precisely, $f^j\in\mathscr{F}_C$, $j=1,\ldots,p$, implies that $\tilde{f}^i\in\mathscr{F}_{C\sqrt{p}}$. Indeed,

$$\left|\left|\tilde{f}^i\right|\right|_{r,q}^{\alpha}\leqslant\sum_{j=1}^{p}\left|V_j^i\right|\left|\left|f^j\right|\right|_{r,q}^{\alpha}\leqslant C\sqrt{p\sum_{j=1}^{p}\left(V_j^i\right)^2}=C\sqrt{p}.$$

For each i, let $\hat{\tilde{f}}^{*i}$ be the estimator of $\tilde{f}^i$ induced by the definition of $\hat{f}^{*i}$. Then, $\hat{\tilde{f}}^{*i}$ is the estimator considered in the near minimaxity theorem of Donoho (1995) with threshold $t_i=\sqrt{2\lambda_i\,\log(n)}$.

Furthermore, using (12), the risk can be written as follows:

$$\begin{aligned}\sup_{f\in(\mathscr{F}_C)^p}R_M\left(f,\hat{f}^*\right)&=\sup_{\left(f^j\in\mathscr{F}_C\right)_{1\leqslant j\leqslant p}}\sum_{i=1}^{p}R\left(f^i,\hat{f}^{*i}\right)\\&=\sup_{\left(f^j\in\mathscr{F}_C\right)_{1\leqslant j\leqslant p}}\sum_{i=1}^{p}R\left(\tilde{f}^i,\hat{\tilde{f}}^{*i}\right),\end{aligned}$$

then, since $\tilde{f}^i\in\mathscr{F}_{C\sqrt{p}}$, we have

$$\sup_{f\in(\mathscr{F}_C)^p}R_M\left(f,\hat{f}^*\right)\leqslant\sup_{\left(\tilde{f}^j\in\mathscr{F}_{C\sqrt{p}}\right)_{1\leqslant j\leqslant p}}\sum_{i=1}^{p}R\left(\tilde{f}^i,\hat{\tilde{f}}^{*i}\right).$$

So by the near minimaxity theorem, we get

$$\sup_{f\in(\mathcal{F}_C)^p} R_M\left(f,\hat{f}^*\right) \leqslant \sum_{i=1}^{p} \sup_{\tilde{f}^i\in\mathcal{F}_{C\sqrt{p}}} R\left(\tilde{f}^i,\hat{\tilde{f}}^{*i}\right)$$

$$\leqslant pc_1\log(n)\inf_{\hat{\tilde{f}}^i}\sup_{\tilde{f}^i\in\mathcal{F}_{C\sqrt{p}}} R\left(\tilde{f}^i,\hat{\tilde{f}}^i\right). \qquad \square$$

So the multivariate estimator $\hat{f}^*$ is near minimax in the one-dimensional sense. This is expected because it estimates p scalar-valued functions.

Remark 2. Let us mention that if $f^i \in B^{\alpha_i}_{r,q}$, then $f^i \in B^{\alpha}_{r,q}$ with $\alpha = \min_{1\leqslant i\leqslant p}\alpha_i$ so the estimator inherits the rate of convergence of the less regular function. It follows that this estimator is not adaptive in the multidimensional sense.

2.3. Noise covariance estimation

Now, let us turn back to the practical issue of the noise covariance estimation. In the one-dimensional case, a convenient estimation of the noise variance is needed in order to fine tune the threshold. Since typically (under very general assumptions) the finer detail coefficients are essentially a Gaussian noise possibly contaminated by a few large coefficients due to the signal, a robust estimator based on the median absolute deviation is used.

Similarly for a natural extension to the multivariate case, a robust estimator of the covariance matrix must be defined. This issue is of course crucial for factor analysis as PCA and highly robust covariance estimators with respect to outliers have been developed in this context. For the same reasons as in the one-dimensional case, D_1 the matrix of details at level 1, is essentially a p-dimensional white noise of covariance matrix Σ_ε corrupted by some coefficients due to the signals to be recovered. So, it is again convenient to use a robust estimator applied to D_1. For this purpose, we use the minimum covariance determinant (MCD) estimator proposed by Rousseeuw (1984) and we apply it to D_1.

The MCD looks for that h-subset (typically $h = 0.75n$) of the data with the smallest determinant of its covariance matrix. More recently, Rousseeuw and Van Driessen (1999) proposed an algorithm allowing quick computation of the MCD. We use in this paper the Matlab clear and efficient code named *fastmcd* available from ftp://win-ftp.uia.ac.be/pub/software/statis/newfiles/fastmcdm.gz.

2.4. Simulated examples

In this section, we present some selected and typical examples among numerous numerical trials. We apply the proposed procedure on simulated examples built from the well known signals introduced by Donoho and Johnstone (1994) and widely considered as bench

examples. Two methods are compared:

- p direct univariate denoising in the initial basis which is considered as the reference method;
- its multivariate alternative: p univariate denoising in the new basis defined by diagonalizing the robust estimate of the noise covariance matrix.

Let us introduce some preliminary material, useful for the simulation examples all along the paper. We consider a four-dimensional model of the form (2) with $n = 1024$. We consider noisy versions of the well known functions called "Blocks", "Bumps", "HeaviSine", "Doppler" (see Donoho and Johnstone, 1994 for the definitions) suitably normalized in order to avoid huge scale effect. The actual model is then defined by f^1= Blocks, f^2= Bumps, f^3= HeaviSine and $f^4 = 10\times$ Doppler and the covariance matrix given by

$$\Sigma_\varepsilon = \Sigma_1 = \begin{pmatrix} 1 & 0.8 & 0.6 & 0.7 \\ 0.8 & 1 & 0.5 & 0.6 \\ 0.6 & 0.5 & 1 & 0.7 \\ 0.7 & 0.6 & 0.7 & 1 \end{pmatrix}. \tag{14}$$

The noise covariance matrix Σ_ε is estimated using the MCD estimator applied to D_1: $\hat{\Sigma}_\varepsilon = MCD\,(D_1)$. Let us mention that the estimations carried out for all simulated examples are accurate.

For each model, we simulate 50 realizations of $\left(X^1(t), \ldots, X^4(t)\right)_{t=1,\ldots,1024}$ and then obtain 50 realizations of each denoised signal $\left(\check{X}^i_k\right)_{k=1,\ldots,50}$.

To evaluate the performance of each method, we introduce the following error criterion, which is an unbiased estimate of a normalized version of the individual risk (5), calculated for each $\left(\check{X}^i\right)$

$$R\left(\check{X}^i, f^i\right) = \frac{1}{50}\sum_{k=1}^{50}\frac{1}{1024}\sum_{t=1}^{1024}\left(\check{X}^i_k(t) - f^i(t)\right)^2. \tag{15}$$

The smaller R, the better the denoising. In addition, we compute the global error performance indicator:

$$R_{\text{glob}}\left(\check{X}\right) = \sum_{i=1}^{4} R\left(\check{X}^i, f^i\right).$$

We perform the wavelet decomposition at level $J = 5$ (a usual choice when $n = 2^{10}$) and at level $J = 3$ (this last value is provided by an automatic choice discussed in Section 3.2) using the near symmetric wavelet of order 4 (with four vanishing moments) due to Daubechies, 1992.

The performance of the two methods (and for the two values of J) are given in Table 1. The relative performance is defined by the ratio of the risk of the reference method and the risk of our method.

For $J = 5$, as it can be seen the global performance is improved after change of basis. The same occurs for each individual function except for the more regular HeaviSine. The reason

Table 1
Multivariate denoising of noisy versions of Blocks, Bumps, HeaviSine and 10× Doppler with $\Sigma_\varepsilon = \Sigma_1$: performance of our method compared to the reference one

Level	Method	$R\left(., f^1\right)$	$R\left(., f^2\right)$	$R\left(., f^3\right)$	$R\left(., f^4\right)$	R_{glob}
$J = 5$	Reference	0.635	0.585	0.239	0.589	2.048
	Our method	0.528	0.494	0.327	0.549	1.898
	Relative performance	120%	118%	73%	107%	108%
$J = 3$	Reference	0.521	0.407	0.360	0.512	1.800
	Our method	0.485	0.399	0.378	0.493	1.755
	Relative performance	107%	102%	95%	104%	103%

Table 2
Multivariate denoising of noisy versions of Blocks, Bumps, HeaviSine and 10× Doppler with $\Sigma_\varepsilon = \Sigma_2$: performance of our method compared to the reference one

Level	Method	$R\left(., f^1\right)$	$R\left(., f^2\right)$	$R\left(., f^3\right)$	$R\left(., f^4\right)$	R_{glob}
$J = 5$	Reference	0.633	0.587	0.238	0.587	2.045
	Our method	0.557	0.520	0.329	0.518	1.924
	Relative performance	113%	110%	72%	113%	106%
$J = 3$	Reference	0.519	0.406	0.364	0.510	1.799
	Our method	0.497	0.407	0.380	0.488	1.772
	Relative performance	104%	100%	74%	104%	101%

is that the first method slightly oversmoothes while the second slightly undersmoothes and captures in a better way the discontinuities.

When $J = 3$, the comments are qualitatively the same but the performance is better and the difference between the two methods is smaller.

Now let us consider the same model but with the following covariance structure allowing positive and negative covariance terms:

$$\Sigma_\varepsilon = \Sigma_2 = \begin{pmatrix} 1 & -0.8 & 0.6 & -0.7 \\ -0.8 & 1 & -0.5 & 0.6 \\ 0.6 & -0.5 & 1 & -0.7 \\ -0.7 & 0.6 & -0.7 & 1 \end{pmatrix}. \tag{16}$$

The results are given in Table 2.

The conclusions are similar, except that the differences between the methods become smaller.

3. Multivariate wavelet denoising using PCA

The multivariate procedure previously examined can be generalized by looking at the deterministic relationships between the p signals. The idea is to use principal component analysis, not to discover new variables which could be of interest, but to kill unsignificant principal components to obtain an additional denoising effect.

In this section, we first recall the multiscale PCA proposed by Bakshi (1998) in another context and we discuss it from the denoising perspective. Next, a second denoising procedure combining wavelets and PCA is proposed. Finally it is illustrated by some simulated examples and applied to reduce noise in multichannel neural recordings.

3.1. Multiscale principal component analysis

A multiscale version of principal component analysis (called MSPCA) has been proposed by Bakshi in order to fine tune the limits in statistical process control (SPC). Then dimensionality reduction issues are of interest as well as selection of relevant factors across scales. The MSPCA scheme is as follows:

- *Step* 1: Perform the wavelet transform at level J of each column of X;
- *Step* 2: For $1 \leqslant j \leqslant J$, perform the PCA of the matrix D_j and select an appropriate number p_j of useful principal components or suppress the detail D_j;
- *Step* 3: Similarly, perform the PCA of the matrix A_J and select p_{J+1} principal components;
- *Step* 4: From the simplified detail and approximation matrices reconstruct a new matrix $\check{X}$ containing the main features of the original matrix X, by inverting the wavelet transform;
- *Step* 5: Finally perform the PCA of the matrix $\check{X}$ and build a convenient statistic for SPC.

Let us remark that to reduce dimensionality, the final PCA step is necessary since, in general, $rank(\check{X}) = rank(X)$. Indeed, a PCA is performed at each level leading to an optimal basis depending on the level and then no dimensionality reduction is reached in the initial common basis.

Nevertheless, steps 2 and 3 provide a compact representation of the p original signals of length n under the form of $n\left(\sum_{j=1}^{J} p_j 2^{-j} + p_{J+1}2^{-J}\right)$ wavelet coefficients.

This procedure is attractive from the SPC objective. But let us remark that, since wavelets are known to be effective for sparse signals, such level by level PCA may destroy sparseness and then make difficult its use for denoising.

Nevertheless, let us interpret the MSPCA scheme in terms of model (2) from the denoising perspective, to suggest another way to take advantage of the deterministic relationships between the p signals.

Step 2 for $j = 1$ reduces to compute a matrix diagonalizing D_1. This is similar to our procedure taking the matrix $D_1^{\mathrm{T}} D_1$ as an estimator of Σ_ε instead of $MCD\,(D_1)$.

If $p_j = 0$, step 2 performs a crude denoising killing all the coefficients, even those attributable to signals. In fact, when j is small, step 2 focuses on noise components and when j is large, it focuses on the dimensionality reduction of the deterministic function f.

In this last case, suppressing the components associated to lower eigenvalues leads to capture the significant ones as well as to denoise signals.

For intermediate scales, the MSPCA scheme is difficult to interpret.

3.2. General denoising procedure

Our procedure for extending denoising to multivariate signals, sketched in Section 2, does not take into account the deterministic relations between the p signals f^i. The previous section suggests a natural way to handle it:

- to apply the previously considered thresholding strategy involving change of basis using V for the details;
- to perform a PCA and select the convenient number of components for the approximations.

More precisely, the general procedure for multivariate denoising is as follows:

- *Step* 1: Perform the wavelet transform at level J of each column of X;
- *Step* 2: Define $\hat{\Sigma}_\varepsilon$ the estimator of the noise covariance matrix as $\hat{\Sigma}_\varepsilon = MCD\,(D_1)$ and then compute V such that $\hat{\Sigma}_\varepsilon = V \Lambda V^{\mathrm{T}}$ where $\Lambda = diag\,(\lambda_i, 1 \leqslant i \leqslant p)$. Apply to each detail after change of basis (namely $D_j V$, $1 \leqslant j \leqslant J$), the p univariate thresholding strategies using the threshold $t_i = \sqrt{2\lambda_i \log(n)}$ for the ith column of $D_j V$;
- *Step* 3: Perform the PCA of the matrix A_J and select the appropriate number p_{J+1} of useful principal components;
- *Step* 4: Reconstruct a denoised matrix $\check{X}$, from the simplified detail and approximation matrices, by changing of basis using V^{T} and inverting the wavelet transform;

Let us remark that a useful alternative scheme could be to cancel step 3 (which is natural when compression purposes are also of interest) by selecting $p_{J+1} = p$ and to introduce step 5:

- *Step* 5: Perform a final PCA of the matrix $\check{X}$ obtained at step 4 and select $\check{p}$ principal components.

The proposed procedure both generalizes univariate denoising and captures the attractive property of dimensionality reduction of the MSPCA scheme leading to an additional denoising effect.

Remark 3. The problem of selecting the appropriate number of principal components has received a lot of attention and many approaches are available, including ones based on extensive crossvalidation, out of the scope of this paper.

For steps 3 and 5, one can use the well known and widely used Kaiser criterion selecting the components corresponding to eigenvalues greater than the mean of all the eigenvalues (see also Karlis et al., 2003). Nevertheless, such choices lead to retain a too small number of components in the toy examples examined in the next section. The following rule of thumb

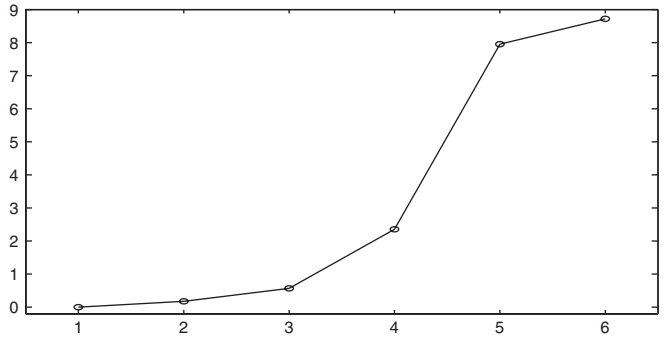

Fig. 1. Plot of $\log(1 + r_j)$ for $1 \leqslant j \leqslant J_{\max} = 6$, defined in Remark 4.

is used: retain the components associated to eigenvalues greater than 5% of the sum of all the eigenvalues.

Remark 4. Another practical issue is the choice of the decomposition level J, which can be considered as a low frequency cutoff. Of course model selection type rule can be used to choose it, but here the comparison between the MSPCA scheme and our procedure suggests the entirely different following rule: select J, $1 \leqslant J \leqslant J_{\max}$, corresponding to the first significant jump (see Fig. 1) of $r_j = trace\left(\Lambda^{(j)} - \Lambda^{(1)}\right)\left(\Lambda^{(j)} - \Lambda^{(1)}\right)^{\mathrm{T}}$ where $\Lambda^{(j)}$ is the diagonal matrix obtained from the PCA of D_j. If such a jump does not exist (for example when the signal-to-noise ratio is too small) select $J = J_{\max}$. The underlying idea of this heuristic is the following. For high frequency detail D_1 the matrix $V\Lambda^{(1)}V^{\mathrm{T}}$ is close to $\hat{\Sigma}_\varepsilon = MCD\,(D_1)$ since the coefficients due to functions f are generally rare and of little influence. When j increases, D_j captures more and more components due to f and $\Lambda^{(j)} - \Lambda^{(1)}$ becomes larger and larger.

3.3. Simulated examples

Let us consider an example which addresses multivariate denoising including dimensionality reduction tools. We start with two previously considered initial functions: "Blocks" the more irregular one and "HeaviSine" the more regular, and we consider the sum and the difference of these two functions, leading to the actual model defined by $f^1 =$ Blocks, $f^2 =$ HeaviSine, $f^3 = f^1 + f^2$ and $f^4 = f^1 - f^2$ and the covariance matrix $\Sigma_\varepsilon = \Sigma_2$ defined in (16).

Three methods are compared:

- the first multivariate procedure which is considered here as the reference method: p univariate denoising in the new basis diagonalizing covariance matrix, corresponding to steps 1, 2 and 4 of the general procedure;
- two variants corresponding to steps 1–4 for the first one and steps 1, 2, 4 and 5 for the second one.

Table 3
Multivariate denoising of noisy versions of Blocks, HeaviSine and their sum and difference with $\Sigma_\varepsilon = \Sigma_2$: performance of the three methods

	$R\left(., f^1\right)$	$R\left(., f^2\right)$	$R\left(., f^3\right)$	$R\left(., f^4\right)$	R_{glob}
Steps 1,2,4	0.469	0.369	0.457	0.427	1.722
Steps 1,2,3,4 $p_{J+1} = 2$	0.343	0.213	0.390	0.310	1.256
Relative performance	136%	173%	117%	138%	138%
Steps 1,2,4,5 $\check{p} = 2$	0.301	0.200	0.391	0.325	1.217
Relative performance	156%	185%	117%	131%	141%

The decomposition level J, the number of principal components p_{J+1} or $\check{p}$ are selected according to the rules introduced in the previous section. They lead to select $J = 3$ and p_{J+1} or $\check{p} = 2$ for all but a few realizations. The heuristic rule to select J is based on the detection of the first breakdown of the slope of the graph of Fig. 1.

The performance of the three methods are given in Table 3.

Let us focus on the two last methods including PCA. First, they outperform tremendously the first method. Second, they perform well, even on the hard function "HeaviSine" (see the first experiment in Section 2.4). Third, they exhibit close global performance and seem to be equally interesting. Fourth, the last one performing PCA after the inverse wavelet transformation improves the performance on "HeaviSine". This suggests that final PCA has an additional denoising effect with respect to the PCA of the approximation coefficients leading to slightly oversmooth irregular functions (like "Blocks") and to better denoise more regular ones (like "HeaviSine").

Let us mention that, on these examples, direct univariate denoising is slightly improved by a final PCA but not sufficiently to outperform the two last methods.

The results obtained from the same model with the covariance matrix $\Sigma_\varepsilon = \Sigma_1$ are qualitatively similar but the results are globally less impressive (the global relative performance is 114% instead of 141%).

To synthesize, let us turn back to the effect of the choice of the decomposition level J. We give an example of the obtained results using the last considered method and the univariate method considered as the reference method in Section 2.4, for $J = 3$ and 5, for one realization focusing on the functions of interest f^1 and f^2. The results are given in Fig. 2. The figure is composed of four rows of plots. The first one contains the original signals and the second the noised versions. The two last rows correspond to $J = 3$ and 5, respectively. The denoised signals using steps 1, 2, 4, 5 with $\check{p} = 2$ can be found in the four plots labelled A,B,C and D. Those obtained using univariate denoising are drawn in the plots labelled a,b,c and d.

A visual comparison leads to the superiority of the procedure using $J = 3$: the two signals of interest are better recovered, especially around the discontinuities. The price to pay is to locally slightly undersmooth the signal where the underlying function is regular. In addition this figure illustrates the satisfactory behavior of our procedure.

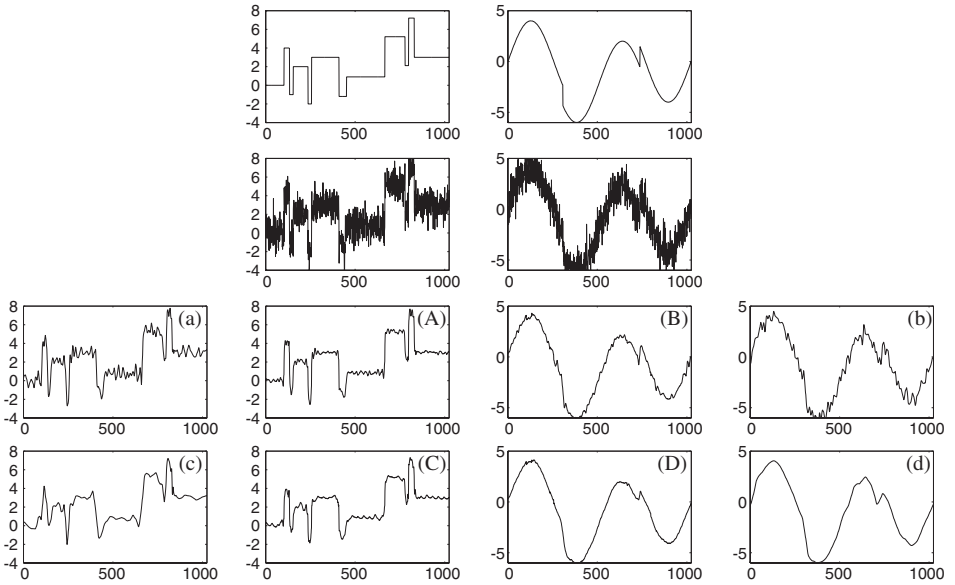

Fig. 2. Multivariate denoising of noisy versions of Blocks, HeaviSine and their sum and difference with $\Sigma_\varepsilon = \Sigma_2$: denoising signals obtained from the two methods. First row: original signals, second row: noisy signals. The two last rows correspond to $J = 3$ and 5, respectively. Denoised signals using steps 1, 2, 4, 5 with $\check{p} = 2$: plots A,B,C,D and univariate denoising: plots a,b,c,d.

3.4. Application to multichannel neural recordings

Real multichannel data exhibiting spatial correlation are examined and fully described in Bierer and Anderson (1999) and Oweiss and Anderson (2001). The data are collected using a 16-channel microprobe and are recorded from dorsal cochlear nucleus (DCN) of guinea pigs. The processed data, recorded at a rate of 20 kHz/channel, are for 100 sweeps of 1950 time samples each. Neural noise is highly correlated among closely spaced electrode sites so it is natural to examine the recordings corresponding to four groups: channels 1–4, 5–8, 9–12, 13–16, because of the geometry of the sensor electrodes. Denoising takes place to identify isolated spikes before spike sorting. This initial task, called spike detection, aims at improving the signal-to-noise ratio (SNR) by removing noise taking into account the spatial correlation.

The following results illustrate on these data the application of the final procedure proposed in this paper. Parameters are slightly adapted to the specific nature of the problem: the Daubechies wavelet of order 2 is chosen and hard thresholding is used, since only irregular features (spikes) are of interest. The decomposition level is equal to $J = 6$ (this value is also chosen in Oweiss and Anderson, 2001). We analyze performance using the SNR. This quantity is essentially based on the "size" of the spikes and the noise "size" given by the standard deviation of the signal without spikes. We use here the following definition of the SNR in dB of a signal x: $SNR(x) = 20 \log_{10}(\max |x|/s(x))$, where $s(x)$ is a robust estimator of the standard deviation of x.

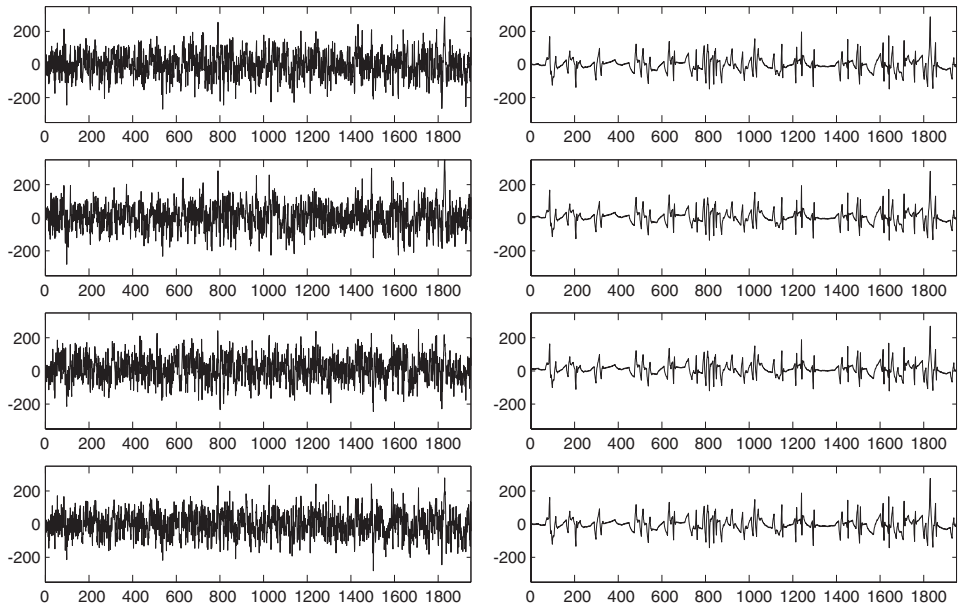

Fig. 3. Original and denoised signals for channels 1–4.

Table 4
Relative performance for channels 1–4

Channel	1	2	3	4
$SNR_{\text{fin}}/SNR_{\text{orig}}$	176 %	135%	190%	172%

Let us examine a typical situation by focusing on the first sweep. Our method selects automatically only one principal component, leading to denoised signals which are the same up to a multiplicative constant, since they are reconstructed from a single one.

Fig. 3 shows on the left the four original signals for the channels 1–4 and on the right the denoised versions. The original signals are of similar magnitude and shape, and spikes are not clearly apparent.

Let us denote by SNR_{orig} the SNR of the original signal and SNR_{fin} the SNR of the denoised signal. In Table 4, one can find for the channels 1–4, the relative performance defined by $SNR_{\text{fin}}/SNR_{\text{orig}}$, for each channel.

Starting from signals for which SNR_{orig} lies between 10 and 14 dB, the procedure performs well since the ratio reaches 190%. Let us remark that this value is close to the relative performance mentioned in Oweiss and Anderson (2001) and obtained using a method especially designed for spike detection.

In addition let us mention that, following experiments not reported here, the PCA step contributes differently according to the channel but, it never deteriorates significantly the

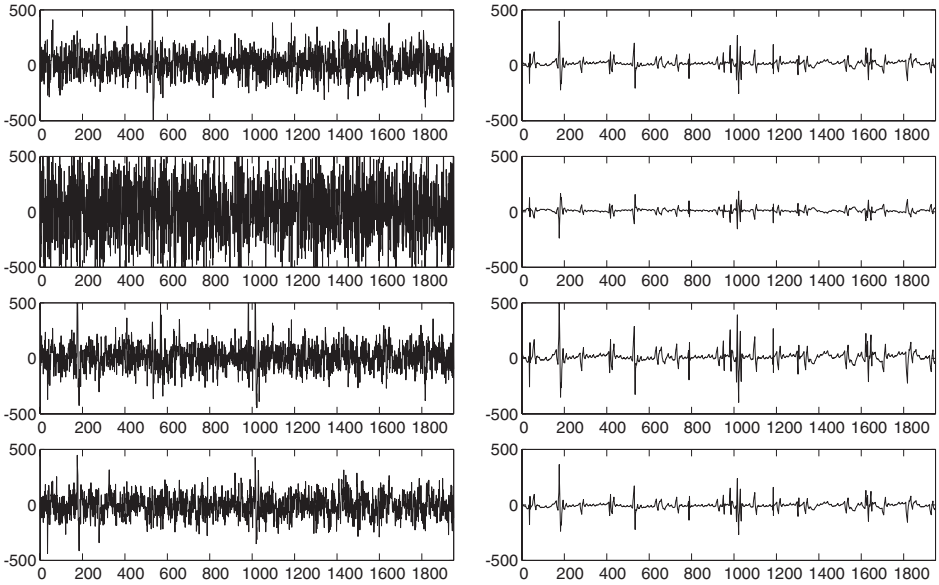

Fig. 4. Original and denoised signals for channels 9–12.

Table 5
Relative performance for channels 9–12

Channel	9	10	11	12
$SNR_{\text{fin}}/SNR_{\text{orig}}$	141%	181%	164%	194%

performance and can strongly improve it, for example for channel 3 the associated benefit is about 30%.

Of course, when the spikes are more visible in the original signals, the denoised signals magnify more clearly the spikes. This is the case for the channels 9–12, see Fig. 4.

Nevertheless, Table 5 shows that the method performs in a similar way, since the relative performance has the same order of magnitude and varies similarly over the channels.

4. Conclusion

We have proposed a multivariate denoising procedure combining wavelets and PCA, that takes into account the correlation structure of the noise. This new procedure exhibits promising behavior on classical bench signals and seems to perform well when it is applied to multichannel neural recordings, the real world example which illustrates the method.

This work could be extended in various directions, let us mention some of them for future work. First, the way to select the parameters of the method, for example the maximum level

of wavelet decomposition and the appropriate number of principal components to retain, could be studied more deeply including the examination of theoretical implications of the associated choices. Second, it would be interesting to extend this work to the use of libraries of basis functions such as wavelet packets, in order to allow to select an accurate signal representation with respect to a given entropy-like function. Third, the extension of the method to more general noise models could be of interest, both from theoretical and practical viewpoints, for example to handle non white noise.

Acknowledgements

The authors thank Anestis Antoniadis for valuable discussions, Karim Oweiss and Yasir Suhail for making available to us the multichannel neural recordings that we used to illustrate our method here and the three anonymous referees for helpful comments and suggestions.

References

Antoniadis, A., 1997. Wavelet in statistics: a review. J. Ital. Statist. Soc. 6, 97–144.
Antoniadis, A., Bigot, J., Sapatinas, T., 2001. Wavelet estimators in nonparametric regression: a comparative simulation study. J. Statist. Software 6, 1–83.
Bakshi, B., 1998. Multiscale PCA with application to MSPC monitoring. AIChE J. 44, 1596–1610.
Bierer, S., Anderson, D., 1999. Multi-channel spike detection and sorting using an array processing technique. Neurocomputing 26–27, 947–956.
Daubechies, I., 1992. Ten Lectures on Wavelets. SIAM, Philadelphia, PA.
DeVore, R., Lorentz, G., 1993. Constructive Approximation. Springer, Berlin.
Donoho, D., 1995. De-noising by soft-thresholding. IEEE Trans. Inform. Theory 41 (3), 613–627.
Donoho, D., Johnstone, I., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81 (3), 425–455.
Hall, P., Kerkyacharian, G., Picard, D., 1999. On the minimax optimality of block thresholded wavelet estimators. Statist. Sinica 9 (1), 33–49.
Karlis, D., Saporta, G., Spinakis, A., 2003. A simple rule for the selection of principal components. Comm. Statist. Theory Methods 32, 643–666.
Kerkyacharian, G., Picard, D., 2000. Thresholding algorithms, maxisets and well-concentrated bases. Test 9 (2), 183–244.
Mallat, S., 1998. A Wavelet Tour of Signal Processing. Academic Press, New York.
Meyer, F., Chinrungrueng, J., 2003. Analysis of event-related fMRI data using local clustering bases. IEEE Trans. Med. Imaging 933–939.
Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.-M., 2003. Les ondelettes et leurs applications. Hermes,
Oweiss, K., Anderson, D., 2001. Noise reduction in multichannel neural recordings using a new array wavelet denoising algorithm. Neurocomputing 38–40, 1687–1693.
Percival, D., Walden, A., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge.
Roberts, S., Roussos, R., Choudrey, E., 2004. Hierarchy, priors and wavelets: structure and signal modelling using ICA. Signal Process. 84, 283–297.
Rousseeuw, P., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871–880.
Rousseeuw, P., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Teppola, P., Minkkinen, P., 2000. Wavelet-PLS regression models for both exploratory data analysis and process monitoring. J. Chemometrics 14, 383–399.
Vannucci, M., Brown, P., Fearn, T., 2003. A decision theoretical approach to wavelet regression on curves with a high number of regressors. J. Statist. Plann. Inference 112 (1–2), 195–212.
Vidakovic, B., 1999. Statistical Modeling by Wavelets. Wiley, New York.