

Wavelet thresholding for some classes of non-Gaussian noise

A. Antoniadis*, D. Leporini and J.-C. Pesquet

Laboratoire IMAG-LMC, Université Joseph Fourier, B.P. 53, 38041 Grenoble Cedex 09, France and Université de Marne-la-Vallée, Cité Descartes, 5, Boulevard Descartes, Champs sur Marne, 77454 Marne la Vallée Cedex 2, France

Wavelet shrinkage and thresholding methods constitute a powerful way to carry out signal denoising, especially when the underlying signal has a sparse wavelet representation. They are computationally fast, and automatically adapt to the smoothness of the signal to be estimated. Nearly minimax properties for simple threshold estimators over a large class of function spaces and for a wide range of loss functions were established in a series of papers by Donoho and Johnstone. The notion behind these wavelet methods is that the unknown function is well approximated by a function with a relatively small proportion of nonzero wavelet coefficients. In this paper, we propose a framework in which this notion of sparseness can be naturally expressed by a Bayesian model for the wavelet coefficients of the underlying signal. Our Bayesian formulation is grounded on the empirical observation that the wavelet coefficients can be summarized adequately by exponential power prior distributions and allows us to establish close connections between wavelet thresholding techniques and Maximum A Posteriori estimation for two classes of noise distributions including heavy-tailed noises. We prove that a great variety of thresholding rules are derived from these MAP criteria. Simulation examples are presented to substantiate the proposed approach.

Key Words and Phrases: wavelets, denoising, regularization, MAP, non-Gaussian noises, exponential power distributions, Cauchy distribution.

1 Introduction

There has recently been a great deal of research interest in wavelet thresholding techniques for signal and image denoising applications (*e.g.* DONOHO, 1995, SIMONCELLI and ADELSON, 1996, ABRAMOVICH *et al.*, 1998, LEPORINI, 1998, ANTONIADIS *et al.*, 1997). In a series of papers, DONOHO and JOHNSTONE (1994, 1995, 1998), and DONOHO *et al.* (1995) developed wavelet shrinkage and thresholding

*Anestis.Antoniadis@imag.fr;
Jean-Christophe.Pesquet@univ-mlv.fr

methods for reconstructing signals from noisy data, where the noise is assumed to be white and Gaussian. They have also shown that the resulting estimates of the unknown function are nearly minimax over a large class of function spaces (typically Besov and Triebel bodies) and for a wide range of loss functions. The model generally adopted for the observed process is: $y(i) = f(i) + \zeta(i)$, $i \in \{1, \dots, K = 2^J\}$ ($J \in \mathbb{N}^*$),¹ where $\{\zeta(i)\}$ is usually assumed to be a random noise vector with independent and identically distributed (i.i.d.) Gaussian components with zero mean and variance σ^2 . Note however that the assumption of Gaussianity is alleviated in the sequel. Estimation of the underlying unknown signal f is of interest. We subsequently consider a (periodic) discrete wavelet expansion of the observation signal, leading to the following additive model: $W_y^{j,k} = W_f^{j,k} + W_\zeta^{j,k}$, $k \in \{1, \dots, 2^{-j}K\}$, where $W_f^{j,k}$ and $W_\zeta^{j,k}$ denote respectively the discrete wavelet coefficients of $f(i)$ and $\zeta(i)$ at resolution level j (see MALLAT, 1989a). Under the Gaussian noise assumption, thresholding techniques successfully utilize the unitary transform property of the wavelet decomposition to distinguish statistically the signal components from those of the noise. In order to fix terminology, we recall that a thresholding rule sets to zero all coefficients $W_y^{j,k}$ with an absolute value below a certain threshold $\chi_j > 0$. The classical hard (resp. soft) threshold estimate $\widehat{W}_f^{j,k}$ of the wavelet coefficient $W_f^{j,k}$ is obtained according to:

$$\widehat{W}_f^{j,k} = \mathbb{1}_{\{|W_y^{j,k}| > \chi_j\}} W_y^{j,k} \left(\text{resp. } \widehat{W}_f^{j,k} = \text{sign}(W_y^{j,k}) \max(0, |W_y^{j,k}| - \chi_j) \right),$$

where $\mathbb{1}_A$ denotes the usual indicator function of a set A and $\text{sign}(\cdot)$ is the signum function. Of course, the crucial step of such procedures is the choice of a thresholding (or shrinkage) method and, subsequently, that of the threshold value (see NASON, 1995). For such problems a number of approaches have been proposed in the literature, including minimax (DONOHO and JOHNSTONE, 1994, 1995, 1998), cross-validation (NASON, 1995, 1996), hypotheses testing (ABRAMOVICH and BENJAMINI, 1996, OGDEN and PARZEN, 1996a, 1996b), and Bayesian methods (VIDAKOVIC, 1998, CHIPMAN *et al.*, 1997, ABRAMOVICH *et al.*, 1998 and RUGGERI and VIDAKOVIC, 1999).

Most wavelet methods based on the Bayesian approach lead to shrinkage rules instead of thresholding and involve specifying a prior distribution on the wavelet coefficients. Bayesian wavelet shrinkage rules are obtained by specifying a certain prior for both f and σ^2 . VIDAKOVIC (1998) assumed that $W_f^{j,k}$ are independent and identically t -distributed with n degrees of freedom and that σ^2 is independent of $W_f^{j,k}$ with an exponential distribution. However, his wavelet shrinkage rule, either based on the posterior mean or via a Bayesian hypotheses testing procedure, requires numerical integration. CHIPMAN *et al.* (1997) also assumed an independent prior for $W_f^{j,k}$. Since a signal is likely to have a sparse wavelet distribution with a heavy tail, they considered a mixture of two zero-mean normal components for $W_f^{j,k}$; one has a

¹The set of positive integers, reals, nonzero reals, nonnegative reals and positive reals are respectively denoted by \mathbb{N}^* , \mathbb{R} , \mathbb{R}^* , \mathbb{R}_+ and \mathbb{R}_+^* .

very small variance and the other has a large variance. Their shrinkage rule based on the posterior mean has a closed-form representation. Both CLYDE *et al.* (1998) and ABRAMOVICH *et al.* (1998) considered a mixture of a normal component and a point mass at zero for wavelet coefficients $W_f^{j,k}$. CLYDE *et al.* (1998) assumed that the prior distribution for σ^2 is inverse gamma, and is independent of $W_f^{j,k}$. They used the stochastic search variable selection algorithm (GEORGE and McCULLOCH, 1997) to search for nonzero wavelet coefficients of the signal, and used the Markov chain Monte Carlo technique to obtain the posterior mean by averaging over all models. Moreover, closed-form approximations to the posterior mean and the posterior variance were also provided. ABRAMOVICH *et al.* (1998) considered a sum of weighted absolute errors as their loss function, resulting in a *thresholding* rule that is Bayes, rather than a shrinkage rule, which is obtained from a Bayesian approach using squared error loss. Their thresholding rule based on the posterior median also has a closed-form representation, under the assumption that σ^2 is known.

In this paper, we exhibit close connections between wavelet thresholding and Maximum *A Posteriori* (MAP) estimation (or, equivalently, wavelet regularization) using exponential power prior distributions. These distributions are in particular the priors that are put on $W_f^{j,k}$. Our approach differs from those previously mentioned by using a different prior and also different loss functions. One of the main advantages of our approach is to naturally provide a thresholding rule, and consequently, a threshold value adapted to the signal/noise under study. Moreover, we will also show that the MAP estimation is also closely related to wavelet regularization of ill-posed inverse stochastic problems with appropriate penalties and loss functions, that parallel Bridge estimation techniques for nonparametric regression as introduced by FRANK and FRIEDMAN (1993) and further extended by FU (1998).

For the sake of simplicity (but see our discussion later), we assume in the sequel that the wavelet coefficients of the signal and the noise are two independent sequences of i.i.d. random variables. In most of the above mentioned Bayesian approaches to wavelet regression the assumed independence of the wavelet coefficients is defended by the strong decorrelational property of the discrete wavelet transform. As mentioned at the end of Section 3.1, these assumptions do not imply that the original processes are i.i.d. and they are, in particular, valid when the processes are radially decomposable. In order to simplify the presentation of our result further, we will drop the dependence on the resolution level j and the time index k of the quantities involved subsequently. As however illustrated by the simulation examples in Section 3, level dependent distributions can be adopted for the wavelet coefficients.

2 Connections between MAP estimation and thresholding rules

At a given resolution level, we choose to model a wavelet coefficient W_f by an Exponential Power Distribution $\mathcal{EPD}(\alpha, \beta)$:

$$p(\cdot; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|\cdot|/\alpha)^\beta}, \quad (\alpha, \beta) \in \mathbb{R}_+^{*2}.$$

Note that the $\mathcal{EPD}(\alpha, \beta)$ model was first proposed in MALLAT (1989b) for wavelet coefficients of signals and images, and subsequently applied to image coding in BUCCIGROSSI and SIMONCELLI (1997) and image denoising using *a posteriori* mean estimates in SIMONCELLI and ADELSON (1996) and SIMONCELLI (1999). Such estimates are however unable to provide thresholding rules. We denote by $-L(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ the log-likelihood function (up to an additive constant) of the noise wavelet coefficients. In this part, $-L$ will correspond to any type of noise and, in the two next sections, some particular type of noises (and hence of likelihood functions) will be studied.

By assuming that $L(0) = 0$, we proceed to determine the MAP estimate of W_f . By denoting by $\mathcal{F}_y(\omega)$ the log-likelihood function of the posterior distribution of W_f , we have:

$$\widehat{W}_f = \arg \min_{\omega} \mathcal{F}_y(\omega), \quad \text{with } \mathcal{F}_y(\omega) = L(W_y - \omega) + \frac{|\omega|^\beta}{\alpha^\beta}. \quad (1)$$

Provided that:

A1. $L(\cdot)$ is continuous at 0, differentiable on \mathbb{R}^* and its derivative $L'(\cdot)$ satisfies:

$$\forall \omega \in \mathbb{R}^*, \quad \text{sign} \{L'(\omega)\} = \text{sign}(\omega),$$

it is straightforward to see that $\mathcal{F}_y(\cdot) \geq 0$ admits a global minimum. Furthermore, any minimizer of $\mathcal{F}_y(\omega)$ necessarily belongs to the interval $[\min(0, W_y), \max(0, W_y)]$, so that MAP estimation indeed corresponds to a shrinkage method. For simplicity, we further assume that:

A2. the function $L(\cdot)$ is symmetric about 0, which implies that \widehat{W}_f is an odd function of W_y . However, it must be pointed out that the presented results can be readily extended to the nonsymmetric case.

As it is easy to see, solving the minimization problem stated in eq. (1), when the noise is assumed to be white Gaussian noise, leads to a special case of Bridge regression, introduced by FRANK and FRIEDMAN (1993), which minimizes the residual sum of squares $L(W_y - \omega)$ penalized by $\lambda|\omega|^\beta$, $\lambda > 0$. While Frank and Friedman did not solve for the estimator of Bridge regression for any given $\beta > 0$, they pointed out that if the true model includes many zero parameters, Bridge regression will perform well for small values of β . TIBSHIRANI (1996) obtained similar results by comparing the *Lasso* with the Bridge through intensive simulation studies. In fact, under the classical Gaussian noise assumption, these results actually appear as a particular case of nonsmooth regularization for local strong homogeneity recovery recently introduced in NIKOLOVA (1997), which imply that a Bridge penalty of small β values favors models with regression parameters either of many zeros or of large absolute values from a long tailed density.

When the noise distribution is further assumed to be an $\mathcal{EPD}(\bar{\alpha}, \bar{\beta})$ distribution, the componentwise minimization problem can be restated as the minimization of a functional of the form

$$\mathcal{F}_y(\omega) = |W_y - \omega|^{\bar{\beta}} + \lambda|\omega|^\beta.$$

In terms of the observed process, and using the characterization of Besov spaces in terms of (regular enough) wavelet decompositions, the above problem corresponds to the following variational problem: find a function \hat{f} that minimizes over all possible functions φ the functional

$$\|y - \varphi\|_{B_{\bar{s}, \bar{\beta}}}^{\bar{\beta}} + \tilde{\lambda}\|\varphi\|_{B_{s, \beta}}^\beta,$$

where $\bar{s} = 1/\bar{\beta} - 1/2 \geq 0$, $s = 1/\beta - 1/2 > 0$, $2 \geq \bar{\beta} > \beta > 0$ and where $B_{s, \beta}$ denotes the Besov space, containing, roughly speaking, functions with s derivatives in L_β . In this framework, the norm (or quasi-norm) involved in the definition of the risk of an estimator is a Besov space norm with indices \bar{s} and $\bar{\beta}$ and leads to variational problems of the same nature as those studied by CHAMBOLLE *et al.* (1998) in terms of the Peetre's K-functional of y between L_2 (i.e. $\bar{s} = 0$ and $\bar{\beta} = 2$) and $B_{s, \beta}$ (see also DELYON and JUDITSKY, 1996, DECHEVSKY and PENEV, 1998). Choosing a known value of s and therefore of β is linked to any additional information we may have about the unknown curve. Typically, the value of s is an upper bound for the true smoothness index corresponding to f .

We prove hereafter that the exponent parameter $\bar{\beta}$ determines the nature of the estimates, and we establish in particular close connections with wavelet thresholding techniques and MAP estimation for two classes of noise probability density functions.

2.1 EPD Distribution for the noise wavelet coefficients

A first useful result is:

LEMMA 1. *Assume A1, A2 and the following additional conditions:*

A3. *the noise log-likelihood $-L(\cdot)$ is convex on \mathbb{R}_+ ;*

A4. *there exists $\chi > 0$ such that $L(\omega) \leq |\omega|^\beta / \alpha^\beta$ iff $|\omega| \leq \chi$.*

Then, when $\beta \leq 1$, the minimizer of $\mathcal{F}_y(\cdot)$ corresponds to a hard thresholding rule with threshold value χ .

PROOF. Due to A2, we can restrict our attention to the case $W_y \geq 0$. We easily deduce from A3 that the function $\mathcal{F}_y(\cdot)$ is concave on $[0, W_y]$, so that its minimum is reached either at 0 or at W_y . Assumption A4 allows us to show that $\mathcal{F}_y(0) \leq \mathcal{F}_y(W_y)$ iff $W_y \leq \chi$, thus ensuring that a hard thresholding estimate is obtained. □

We now consider the particular case of a noise following an exponential power distribution.

PROPOSITION 1. If $W_\xi \sim \mathcal{EPD}(\bar{\alpha}, \bar{\beta})$ with $\bar{\beta} > \beta$ and $\beta \in (0, 1]$, the MAP estimation of W_f leads to thresholding rules. In particular, when $\bar{\beta} \leq 1$, hard thresholding rules are obtained. In the case $\bar{\beta} > 1$, we have, when $|W_y| \rightarrow \infty$,

$$\widehat{W}_f = W_y - \left(\frac{\beta \bar{\alpha}^{\bar{\beta}}}{\beta \alpha^\beta} \right)^{1/(\bar{\beta}-1)} |W_y|^{(\beta-1)/(\bar{\beta}-1)} \text{sign}(W_y) + o(|W_y|^{(\beta-1)/(\bar{\beta}-1)}).$$

PROOF. First note that A1 and A2 are satisfied. Furthermore, when $\bar{\beta} \leq 1$, A3 and A4 hold, leading to hard thresholding policies with threshold value $\chi = (\bar{\alpha}^{\bar{\beta}}/\alpha^\beta)^{1/(\bar{\beta}-\beta)}$ (cf. Lemma 1).

When $\bar{\beta} > 1$, we consider

$$\mathcal{F}_y(\omega) = \frac{|W_y - \omega|^{\bar{\beta}}}{\bar{\alpha}^{\bar{\beta}}} + \frac{|\omega|^\beta}{\alpha^\beta}, \tag{2}$$

and proceed to calculate its derivative $\mathcal{F}'_y(\cdot)$ when, say, $W_y \geq 0$. We easily show that $\mathcal{F}'_y(\omega) = 0$ iff

$$W_y = \left(\frac{\beta \bar{\alpha}^{\bar{\beta}}}{\beta \alpha^\beta} \right)^{1/(\bar{\beta}-1)} \omega^{(\beta-1)/(\bar{\beta}-1)} + \omega = g(\omega) \tag{3}$$

By studying the function $g(\cdot)$, it can be proved that, if $W_y \leq g(\theta)$ with

$$\theta = \left(\frac{\beta \bar{\alpha}^{\bar{\beta}}}{\beta \alpha^\beta} \right)^{1/(\bar{\beta}-\beta)} \left(\frac{1 - \beta}{\bar{\beta} - 1} \right)^{(\bar{\beta}-1)/(\bar{\beta}-\beta)},$$

$\mathcal{F}_y(\cdot)$ is an increasing function on \mathbb{R}_+^* and, consequently, $\widehat{W}_f = 0$. This clearly shows that a thresholding rule is obtained. Note however that this result only provides a lower bound $g(\theta)$ on the threshold value. More precisely, it can be deduced from Condition (3) that, if $W_y > g(\theta)$, $\mathcal{F}_y(\cdot)$ has a unique minimum on \mathbb{R}_+^* at $\omega = \widehat{\omega}_y > \theta$. Then, $\widehat{W}_f = \widehat{\omega}_y$ if $\mathcal{F}_y(\widehat{\omega}_y) < \mathcal{F}_y(0)$ and is equal to 0 otherwise. The previous inequality can be shown to hold asymptotically (when $W_y \rightarrow \infty$) as $\widehat{\omega}_y$ is solution of (3) and therefore such that

$$\widehat{\omega}_y = W_y - \left(\frac{\beta \bar{\alpha}^{\bar{\beta}}}{\beta \alpha^\beta} \right)^{1/(\bar{\beta}-1)} |W_y|^{(\beta-1)/(\bar{\beta}-1)} + o(|W_y|^{(\beta-1)/(\bar{\beta}-1)}). \quad \square$$

We point out that the condition $\bar{\beta} > \beta$ makes sense in a practical perspective as it means that the wavelet decomposition ‘‘compresses’’ the signal better than the noise.² In the context of wavelet regularization with an $L_{\bar{\beta}}$ risk function, this assumption is also made in DELYON and JUDITSKY (1996) or DECHEVSKY and PENEV (1998).

²When $\bar{\beta} > \beta$, the pdf of W_f is greater than the pdf of W_ξ around 0, for a fixed variance. So, one may consider that the wavelet decomposition provides a more parsimonious representation for the signal than for the noise.

Although the threshold value cannot be analytically derived in the general case, we apply the previous result to some cases of particular interest. In a straightforward manner, it can be checked that

COROLLARY 1. *The minimizer of (2) for a Laplacian prior ($\beta = 1$) and $\bar{\beta} > 1$ corresponds to a soft thresholding rule with threshold value*

$$\chi = (\bar{\alpha}^{\bar{\beta}}/\bar{\beta}\alpha)^{1/(\bar{\beta}-1)}.$$

Note that this property holds in the Gaussian case ($\bar{\beta} = 2$). We further investigate the Gaussian noise assumption with the following more general result.

COROLLARY 2. *The MAP estimation with $\beta \leq 1$ and $\bar{\beta} = 2$ leads to thresholding rules corresponding to the threshold value:*

$$\chi = \frac{2 - \beta}{2(1 - \beta)} \left(\frac{2\sigma^2(1 - \beta)}{\alpha^\beta} \right)^{1/(2-\beta)}, \tag{4}$$

where $\sigma^2 = \bar{\alpha}^2/2$ denotes the variance of the noise. Moreover, we have:

$$\lim_{\substack{W_y \rightarrow \pm\chi \\ |W_y| > \chi}} \widehat{W}_f = \pm \left(\frac{2\sigma^2(1 - \beta)}{\alpha^\beta} \right)^{1/(2-\beta)}.$$

PROOF. By combining the relation $\mathcal{F}'_y(\widehat{\omega}_y) = 0$ with the Gaussian assumption, it can be proved after some algebra that $\mathcal{F}_y(\widehat{\omega}_y) < \mathcal{F}_y(0)$ iff

$$\widehat{\omega}_y > \left(\frac{2\sigma^2(1 - \beta)}{\alpha^\beta} \right)^{1/(2-\beta)} \tag{5}$$

where the lower bound is greater than or equal to θ . As W_y satisfies (3) with $\omega = \widehat{\omega}_y$, and $g(\cdot)$ is an increasing function on $[\theta, \infty)$, (5) is equivalent to

$$W_y > g\left(\left(\frac{2\sigma^2(1 - \beta)}{\alpha^\beta} \right)^{1/(2-\beta)} \right) = \chi,$$

which finally provides the expected results. □

For illustration, a threshold rule associated with the considered statistical models is presented in Fig. 1. Note that the function \widehat{W}_f/W_y tends to 1 in as $|W_y| \rightarrow \infty$. When the value of $|W_y|$ is large, one can consider that the observed value of W_y is not noise and then one does not shrink the value of W_y , which would result in underestimating W_f . This improves the properties of the soft-thresholding rule which always shifts the estimate W_y by a fixed amount.

2.2 Cauchy distribution for the noise wavelet coefficients

The previous results may be of interest in dealing with heavy-tailed exponential power noise distributions. We extend this idea to other noise classes with the following lemma.

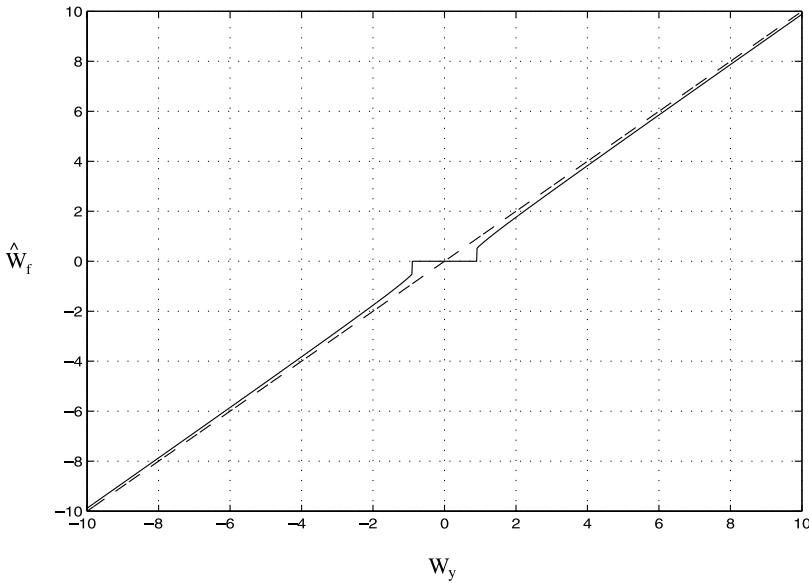


Fig. 1. Threshold rule corresponding to $\mathcal{EPD}(1,2)$ noise and $\mathcal{EPD}(1,0.6)$ prior distributions.

LEMMA 2. Assume A1, A2 and the following properties:

- A5. there exists $\bar{\epsilon} > 0$ such that $L(\cdot)$ is convex on $[0, \bar{\epsilon}]$;
- A6. $L'(\omega) = o(\omega^{\beta-1})$ when $\omega \rightarrow 0^+$;
- A7. there exists $C \in \mathbb{R}_+^*$ such that, for all local minimizer $\hat{\omega}_y > 0$ of $\mathcal{F}_y(\cdot)$,
 $W_y - \hat{\omega}_y \leq C$;
- A8. $L(\omega) = o(\omega^\beta)$ when $\omega \rightarrow \infty$.

Then, the minimizer of $\mathcal{F}_y(\cdot)$ with $\beta \leq 1$ corresponds to a double thresholding rule, i.e.

$$\exists (\chi_L, \chi_U) \in (\mathbb{R}_+^*)^2 \quad \text{such that} \quad \widehat{W}_f \neq 0 \Rightarrow \chi_L$$

PROOF. Due to Assumptions A1 and A2, we can focus on the case $W_y \geq 0$ and look for the minimum of $\mathcal{F}_y(\cdot)$ on the interval $[0, W_y]$. According to A5, $\forall (\omega, W_y) \in \mathbb{R}^2$ with $0 < \omega < W_y \leq \bar{\epsilon}$, $L'(W_y - \omega) \leq L'(W_y)$. We then have

$$\mathcal{F}'_y(\omega) = -L'(W_y - \omega) + \frac{\beta}{\alpha^\beta} \omega^{\beta-1} \geq -L'(W_y) + \frac{\beta}{\alpha^\beta} W_y^{\beta-1}.$$

Now using A6, we can assert that there exists $\eta > 0$ such that $\forall (\omega, W_y) \in \mathbb{R}^2$ with $0 < \omega < W_y < \eta$, $\mathcal{F}'_y(\omega) \geq 0$. This implies the existence of a lower positive threshold value. Furthermore, as a direct consequence of A1, $L(\cdot)$ is a nonnegative function. This fact combined with A7 guarantees that, for any local minimizer $\hat{\omega}_y > 0$ of $\mathcal{F}_y(\cdot)$, $\mathcal{F}_y(\hat{\omega}_y) \geq \hat{\omega}_y^\beta / \alpha^\beta \geq (W_y - C)^\beta / \alpha^\beta$. As $\mathcal{F}_y(0) = L(W_y)$, Assumption A8 allows us to conclude that there exists $\chi_U > 0$ such that, for all $W_y > \chi_U$, $\mathcal{F}_y(0) < \mathcal{F}_y(\hat{\omega}_y)$. This shows the existence of an upper threshold. \square

Interestingly, these estimates are closely related to the constrained minimax thresholding introduced in KRIM and SCHICK (1999) where the boundedness of the signal coefficients is however assumed. We apply the previous proposition to Cauchy noise distributions.

PROPOSITION 2. *If $W_\xi \sim \mathcal{C}(0, \bar{\alpha})$ where $\mathcal{C}(0, \bar{\alpha})$ denotes the centred Cauchy distribution with inverse scale parameter $\bar{\alpha} > 0$, the MAP estimation with $\beta \in (0, 1]$ leads to double thresholding rules.*

This result is readily proved by verifying that the assumptions of Lemma 2 are satisfied. An example of such a double threshold estimator is presented in Fig. 2. For illustration, we focus in the sequel on the Laplacian prior ($\beta = 1$). Then, the nonconvex function to be minimized is $\mathcal{F}_y : \omega \mapsto \log(1 + \bar{\alpha}^2(W_y - \omega)^2) + |\omega|/\alpha$.

COROLLARY 3. *If $\alpha\bar{\alpha} \leq -1$, the minimization of the previous expression leads to the degenerate MAP estimate: $\widehat{W}_f = 0$. If $\alpha\bar{\alpha} > 1$, the MAP estimate corresponds to the double soft thresholding rule defined by*

$$\widehat{W}_f = \begin{cases} \text{sign}(W_y)(|W_y| - \chi_L) & \text{if } \chi_L < |W_y| < \chi_U \\ 0 & \text{otherwise.} \end{cases},$$

where $\chi_L = \alpha - \bar{\alpha}^{-1}\sqrt{\alpha^2\bar{\alpha}^2 - 1}$ and χ_U is the unique solution in (χ_L, ∞) of the equation:

$$\log(1 + \bar{\alpha}^2\chi_U^2) = \log(1 + \bar{\alpha}^2\chi_L^2) + \frac{\chi_U - \chi_L}{\alpha}. \tag{6}$$

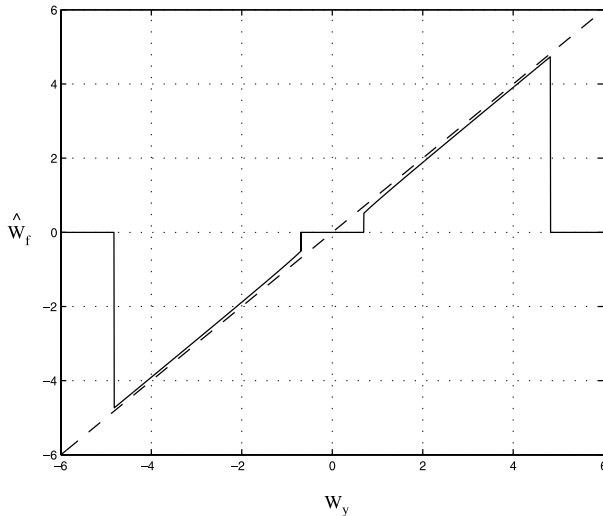


Fig. 2. Threshold rule corresponding to $\mathcal{C}(0, 2)$ noise and $\mathcal{EPD}(0.55, 0.7)$ prior distributions.

PROOF. We again consider the case where $W_y > 0$. The derivative of $\mathcal{F}_y(\cdot)$ is then given by:

$$\mathcal{F}'_y(\omega) = -\frac{2\bar{\alpha}^2(W_y - \omega)}{1 + \bar{\alpha}^2(W_y - \omega)^2} + \frac{1}{\alpha}.$$

It is straightforward to show that the existence of a local minimum at $\hat{\omega}_y \in (0, W_y)$ is equivalent to: $\alpha\bar{\alpha} > 1$ and $W_y > \chi_L$ (with χ_L as defined above). In this case, the location of the minimum is simply $\hat{\omega}_y = W_y - \chi_L$. It remains to study the sign of $\mathcal{F}_y(0) - \mathcal{F}_y(\hat{\omega}_y)$. It can be proved that $\mathcal{F}_y(0) > \mathcal{F}_y(\hat{\omega}_y)$ (and, thus, $\widehat{W}_f = \hat{\omega}_y$) iff $h(W_y) > h(\chi_L)$ where $h(\omega) = \log(1 + \bar{\alpha}^2\omega^2) - \omega/\alpha$. After some algebra, this inequality is shown to be satisfied iff $\chi_L < W_y < \chi_U$ where χ_U is the unique solution greater than $\alpha + \bar{\alpha}^{-1}\sqrt{\bar{\alpha}^2\bar{\alpha}^2 - 1} > \chi_L$ of the equation $h(\chi_U) = h(\chi_L)$. This leads to the implicit definition of χ_U in (6). \square

The intuition behind the double thresholding rule is that, for such heavy tail noise distributions, the impulsiveness of the noise induces wavelet coefficients with high amplitudes which need to be eliminated by the estimator.

At this point, one may wonder how a non-Gaussian noise distribution can lead to an independent distribution of the wavelet coefficients. A class of noise distributions which would lead to the analysis we have adopted in this paper is the class of radially decomposable distributions described in FANG *et al.* (1989). Recall that a K -dimensional random vector \mathbf{Z} is said to be radially decomposable if there exist a positive random scalar R and an independent K -dimensional random vector \mathbf{W} such that \mathbf{Z} and $R\mathbf{W}$ are identically distributed. In this decomposition R is called the radial component and \mathbf{W} the base component. If the noise vector \mathbf{Z} belongs to the class of α -symmetric distributions with $\alpha = 1$ (a subclass of the family of radially decomposable distributions), the analysis of NG and FRASER (1994) conducted for general linear models, implies that one may assume that the components of the noise on the wavelet coefficients are i.i.d standard Cauchy in making inference about the wavelet coefficients.

Another issue which has not been addressed in this paper is the minimax asymptotic behavior of the estimators derived in the above sections. For such a study, one could rely on and extend the methods derived in NEUMANN and VON SACHS (1995), and more particularly those of JUDITSKY (1997) which relate to wavelet shrinkage in non-Gaussian noise under L_p losses. However, such an approach is outside the scope of this paper and can be a subject of future research. In the next section, we apply the considered statistical models and their corresponding thresholding rules to some signal denoising examples.

3 Denoising examples

Our purpose here is to illustrate the behavior of the proposed MAP estimates for signals corrupted by non-Gaussian noises. For simplicity, we will restrict our

attention to low-complexity methods for the estimation of the hyperparameters, although more sophisticated algorithms could be envisaged, including Markov Chain Monte Carlo methods (see *e.g.* SMITH and ROBERTS, 1993, TIERNEY, 1994, LEPORINI and PESQUET, 2001). An interesting Empirical Bayes approach has also been proposed in JOHNSTONE and SILVERMAN (1998) corresponding to a marginal maximum likelihood estimation for mixture models.

3.1 Cauchy noise distribution

We first consider an example of a Doppler-like signal embedded in Cauchy noise. As the Cauchy distribution is an *alpha*-stable distribution, the wavelet coefficients $W_{\xi}^{j,k}$ of the noise $\xi(i)$ are Cauchy processes.

In a first set of simulations, a noise is synthesized with i.i.d. $\mathcal{C}(0, \bar{\alpha})$ wavelet coefficients (see the remark at the end of the previous section). The parameter $\bar{\alpha}$ is not assumed to be known and, consequently, it must be estimated during the denoising procedure.

The common way to specify a stable distribution is by its characteristic function $\phi(t)$. Recall that the log-characteristic function of a symmetric Cauchy distribution with inverse scale parameter $\bar{\alpha}$ is given by

$$\log\{\phi(t)\} = -\frac{|t|}{\bar{\alpha}}.$$

Various methods have been suggested in the literature to estimate the parameters of a stable distribution from an i.i.d. sample. Most of these methods are based on the empirical characteristic function $\hat{\phi}(\cdot)$. For example, a moment estimate for $1/\bar{\alpha}$ is obtained by solving by least-squares $|t_i|/\bar{\alpha} = -\log(|\hat{\phi}(t_i)|)$ (see NIKIAS and SHAO, 1995).

Level-dependent Laplacian priors are adopted for the wavelet coefficients of the signal. In order to estimate the model parameters, we first realize a rough estimation of the signal of interest based on median filtering. This means that an estimate of the signal is obtained as

$$\hat{f}(i) = \text{median}\{y(i-P), \dots, y(i+P)\}$$

where $P \in \mathbb{N}$ is the filter length. Estimations of the (level-dependent) dispersion parameters α_j are obtained from the wavelet coefficients of this signal estimate using a method of moments approach. The inverse scale parameter $\bar{\alpha}$ of the noise is estimated using the empirical characteristic function as described above. The MAP estimates of the signal coefficients are then computed. The original process and a noisy version where $\mathcal{C}(0, 100)$ noise has been added are presented in Fig. 3. The signal to noise ratio can be defined as $\text{SNR} = \bar{\alpha}K^{-1/2}(\sum_{i=1}^K f(i)^2)^{1/2}$ and is here equal to 29.30. As expected, the estimation procedure relying upon the MAP criterion suppresses the severe outliers generated by the heavy-tailed distribution. The indicated normalized mean square error (NMSE) and normalized mean absolute error (NMAE) were computed using 100 noise realizations. These results have been

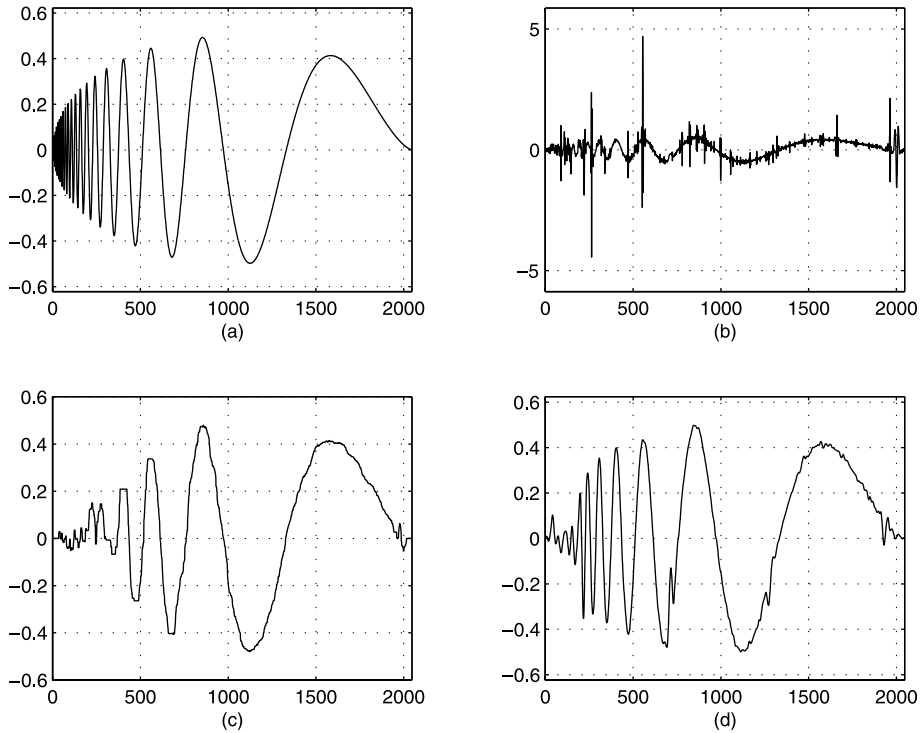


Fig. 3. (a) Original Doppler-like signal, (b) signal corrupted with noise having i.i.d. $\mathcal{C}(0, 100)$ wavelet coefficients, (c) reconstructed signal using median filtering with length 79 (NMSE = 0.1318, NMAE = 0.2090), and (d) reconstructed signal using MAP estimation (NMSE = 0.0355, NMAE = 0.0974).

obtained using a four-level wavelet decomposition implemented with eight tap symlet filters.

For comparison, the result of a median filtering of the noisy signal is also provided. It is worth noting that this estimation has been obtained with optimized window length. This optimization has been realized empirically so as to minimize the mean square estimation error (so, using the signal to be estimated). It is clear that any automatic procedure for the determination of the window length would lead to decreased estimation performances.

Another example of a Cauchy noise is given which corresponds to an autoregressive model of order 1, $\xi(i) = 0.9\xi(i - 1) + Z(i)$, driven by an i.i.d. $\mathcal{C}(0, 5000)$ Cauchy noise $Z(i)$. We then have $\text{SNR} = 146.51$. In this case, the wavelet coefficients are dependent $\mathcal{C}(0, \bar{\alpha}_j)$ random variables. They are however weakly dependent as can be checked by computing the normalized codifference of the wavelet coefficients. At resolution level j , the standardized codifference can be calculated by

$$\text{NC}_j = 2 - \frac{\text{E}\{|W_\xi^{j,k} - W_\xi^{j,k-1}|^p\}^{1/p}}{\text{E}\{|W_\xi^{j,k}|^p\}^{1/p}}$$

with $p = 0.9$. Using a three-level symlet wavelet decomposition, the corresponding values estimated from 1000 noise realizations of length 2048 are: $NC_1 = 0.1663$, $NC_2 = 0.2250$ and $NC_3 = 0.1909$.

The coefficients $\bar{\alpha}_j$ are estimated at each resolution level j as previously and the same signal denoising algorithm is applied. As shown by Fig. 4, although the assumption of independence of the wavelet coefficients is not theoretically satisfied, the proposed estimation method also leads to improved performances over median filtering with optimized length.

3.2 EPD Distribution

In the second example, we consider a 256×256 image with intensity values ranging from 0 to 255 which is corrupted by impulsive noise. For the sake of simplicity, we have adopted one-dimensional notations in the previous parts of the paper but the presented results can straightforwardly be extended to the 2D case. The main difference for images is that additional indices are required. In particular, when a 2D separable wavelet decomposition is performed, an index $d \in \{1, 2, 3\}$ must be

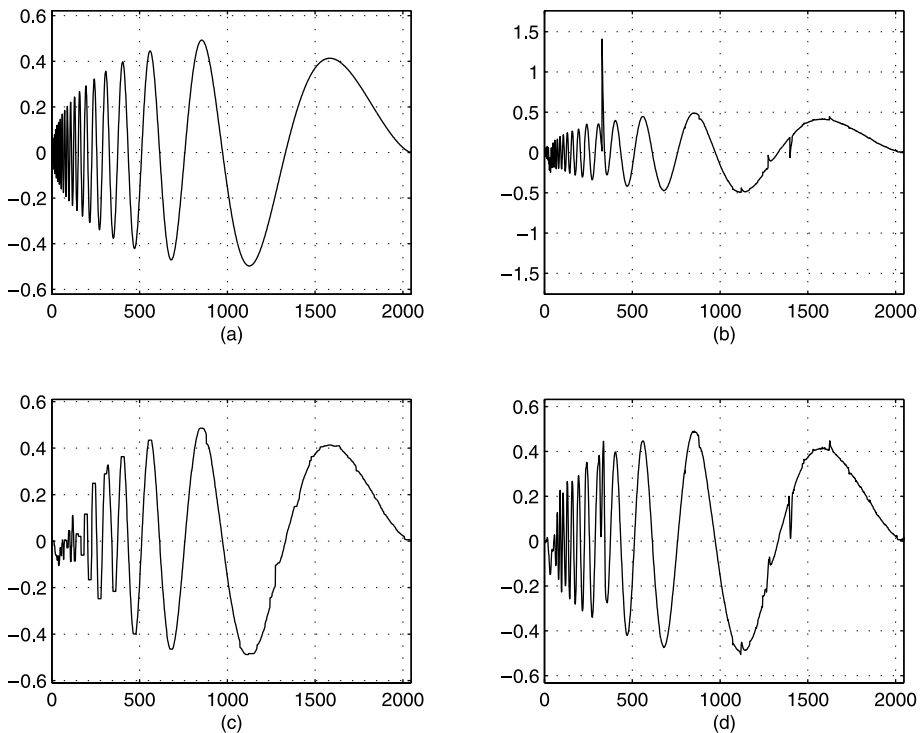


Fig. 4. (a) Original Doppler-like signal, (b) signal corrupted with Cauchy AR(1) noise, (c) reconstructed signal using median filtering with length 31 (NMSE = 0.0553, NMAE = 0.0863), and (d) reconstructed signal using MAP estimation (NMSE = 0.0347, NMAE = 0.0504).

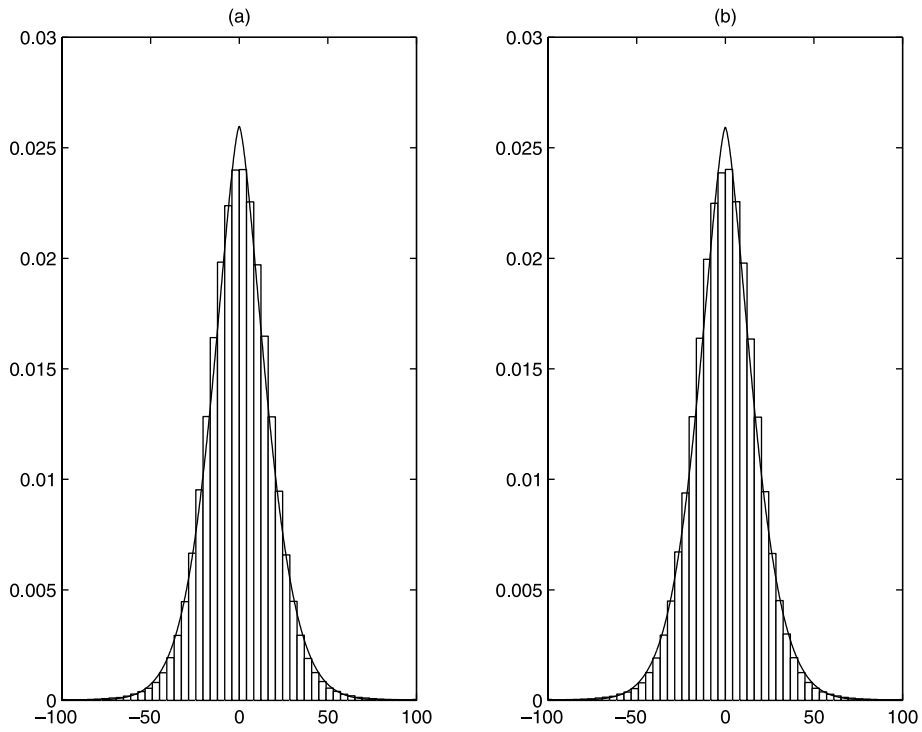


Fig. 5. Histogram and \mathcal{EPD} modelling distribution (resulting from a Maximum Likelihood estimation) of the noise wavelet coefficients at resolution level $j = 1$: (a) $d = 1, 2$, $\bar{\beta}_{1,1} = 1.495$, $\bar{\alpha}_{1,1} = 21.31$, (b) $d = 3$, $\bar{\beta}_{1,3} = 1.5$, $\bar{\alpha}_{1,3} = 21.36$.

introduced in order to distinguish the horizontal, vertical or diagonal orientations of the three analysis wavelets used.

The probability distribution of the noise is here assumed to be known and it is given by the following mixture of Gaussians:

$$0.95\mathcal{N}(0, 225) + 0.05\mathcal{N}(0, 2500).$$

This allows us to estimate the more appropriate $\mathcal{EPD}(\bar{\alpha}_{j,d}, \bar{\beta}_{j,d})$ model for each orientation and each resolution level. The parameters $\bar{\alpha}_{j,d}$ and $\bar{\beta}_{j,d}$ have been determined from noise samples using a Maximum Likelihood approach.³ Figs. 5 and 6 show the relevance of this modelling. Due to the existence of a multiresolution central limit theorem (see LEPORINI and PESQUET, 1999) for a wide class of second-order stationary processes, parameters $\bar{\beta}_{j,d}$ tend to 2 as j increases.

Estimations of the parameters $(\alpha_{j,d}, \beta_{j,d})$ of the prior distributions are subsequently derived from the second and fourth order moments of the noisy image. The resulting image estimate is presented in Fig. 9 whereas the original and noisy images are given by Figs. 7 and 8. For evaluation, we show in Fig. 10 the performances of the

³Note that, for symmetry reasons, we can impose the constraints: $\bar{\alpha}_{j,1} = \bar{\alpha}_{j,2}$ and $\bar{\beta}_{j,1} = \bar{\beta}_{j,2}$.

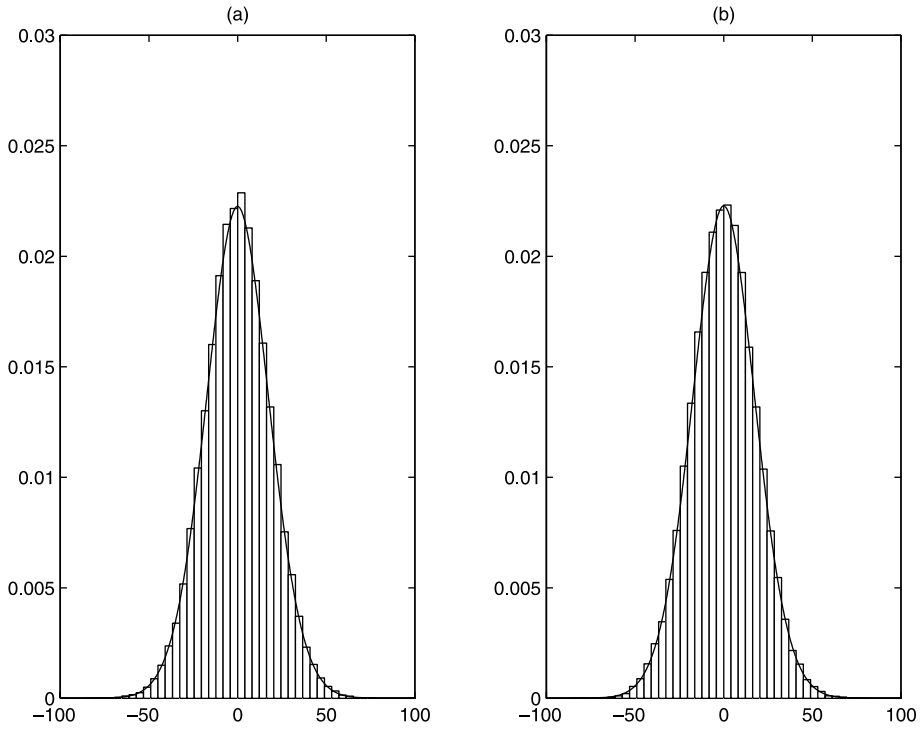


Fig. 6. \mathcal{EPD} model fitting of the noise wavelet coefficients at resolution level $j = 3$: (a) $d = 1, 2$, $\hat{\beta}_{3,1} = 1.9$, $\hat{\alpha}_{3,1} = 25.31$, (b) $d = 3$, $\hat{\beta}_{3,3} = 1.91$, $\hat{\alpha}_{3,3} = 25.27$.



Fig. 7. Original image.



Fig. 8. Noisy image (NMSE = $15.48 \cdot 10^{-3}$, NMAE = $10.26 \cdot 10^{-2}$).



Fig. 9. MAP estimate using level-dependent \mathcal{EPD} models (NMSE = $9.676 \cdot 10^{-3}$, NMAE = $7.793 \cdot 10^{-2}$).

sureshrink estimate presented in DONOHO and JOHNSTONE (1995) which assumes a level-dependent Gaussian noise. We also compare our approach with a minimax procedure which is based on the \mathcal{EPD} assumption for the noise (see Fig. 11). Then, for each value of (j, d) , a soft thresholding estimate is obtained by determining numerically the threshold value $\chi_{j,d} > 0$ which is the unique solution of the equation (cf AVERKAMP and HOUDRÉ, 1999):



Fig. 10. Sureshrink estimate with level-dependent noise variances (NMSE = $13.34 \cdot 10^{-3}$, NMAE = $9.410 \cdot 10^{-2}$).



Fig. 11. Minmax level-dependent soft thresholding (NMSE = $14.1 \cdot 10^{-3}$, NMAE = $9.243 \cdot 10^{-2}$).

$$2(2^{16-2j} + 1) \int_{\lambda_{j,d}}^{\infty} (\omega - \lambda_{j,d})^2 p(\omega; \bar{\alpha}_{j,d}, \bar{\beta}_{j,d}) d\omega = \lambda_{j,d}^2 + \bar{\alpha}_{j,d}^2 \frac{\Gamma(3/\bar{\beta}_{j,d})}{\Gamma(1/\bar{\beta}_{j,d})}.$$

The NMSE and NMAE have again been evaluated for each method and eight tap symlet filters have been used in this example to realize a three-level wavelet decomposition.



Fig. 12. MAP estimate using level-dependent \mathcal{EPD} models combined with a translation-invariant averaging method (NMSE = $6.724 \cdot 10^{-3}$, NMAE = $6.529 \cdot 10^{-2}$).

Note finally that the MAP estimation can be combined with the translation-invariant averaging technique described in NASON and SILVERMAN (1995), COIFMAN and DONOHO (1995). As demonstrated by Fig. 12, this results in a clear improvement both in terms of quantitative performances and visual quality.

Acknowledgements

This research was supported by ‘Projet AMOA, IMAG’. The editor Professor P. H. Franes, the associated editor and the two referees are greatly acknowledged for their comments and suggestions.

References

- ABRAMOVICH, F. and Y. BENJAMINI (1996), Adaptive thresholding of wavelet coefficients, *Computational Statistics and Data Analysis* **22**, 351–361.
- ABRAMOVICH, F., T. SAPATINAS and B. W. SILVERMAN (1998), Wavelet thresholding via a Bayesian approach, *Journal of the Royal Statistic Society Series B* **60**, 725–749.
- ANTONIADIS, A., I. GIJBELS and G. GRÉGOIRE (1997), Model selection using wavelet decomposition and applications, *Biometrika* **84**, 751–763.
- AVERKAMP, R. and C. HOUDRÉ (1999), Wavelet thresholding for non (necessarily) Gaussian noise: idealism. Internal report, Georgia Institute of Technology, Atlanta.
- BUCCIGROSSI, R. W. and E. P. SIMONCELLI (1997), Progressive wavelet image coding based on a conditional probability model in: *Proceedings of the IEEE Conference on Acoustics, Speech, Signal Processing*, pages IV.2957–IV.2960, Munich, Germany.

- CHAMBOLLE, A., R. A. DEVORE, N.-Y. LEE and B. J. LUCIER (1998), Nonlinear wavelet image processing: variational problems, compression and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing* **7**, 319–335.
- CHIPMAN, H. A., E. D. KOLACZYCK and R. E. McCULLOCH (1997), Adaptive Bayesian wavelet shrinkage, *Journal of the American Statistical Association* **22**, 1413–1421.
- CLYDE, M., G. PARMIGIANI and B. VIDAKOVIC (1998), Multiple shrinkage and subset selection in wavelets, *Biometrika* **85**, 391–401.
- COIFMAN, R. and D. DONOHO (1995), Translation-invariant de-noising: in ANTONIADIS, A. and G. OPPENHEIM, (eds.), *Wavelets and statistics*. Lecture Notes in Statistics, Springer Verlag.
- DECHEVSKY, L. T. and S. I. PENEV (1998), On penalized wavelet estimation, Report No. S98-15, Department of Statistics, The University of New South Wales.
- DELYON, B. and A. JUDITSKY (1996), On minimax wavelet estimators, *Applied and Computational Harmonic Analysis* **3**, 215–228.
- DONOHO, D. L. (1995), Denoising by soft-thresholding. *IEEE Transactions on Information Theory* **41**, 613–627.
- DONOHO, D. L. and I. M. JOHNSTONE (1994), Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**, 425–455.
- DONOHO, D. L. and I. M. JOHNSTONE (1995), Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association* **90**, 1200–1224.
- DONOHO, D. L. and I. M. JOHNSTONE (1998), Minimax estimation via wavelet shrinkage, *Annals of Statistics* **26**, 879–921.
- DONOHO, D. L., I. M. JOHNSTONE, G. KERKYACHARIAN and D. PICARD (1995), Wavelet shrinkage: asymptopia? (with discussion), *Journal of the Royal Statistical Society Series B* **57**, 301–369.
- FANG, K. T., S. KOTZ and K. W. NG (1989), *Symmetric multivariate and related distributions*, Chapman and Hall, London and New York.
- FRANK, I. and J. H. FRIEDMAN (1993), A statistical view of some chemometrics regression tools, *Technometrics* **35**, 109–148.
- FU, W. J. (1998), Penalized regressions: the Bridge versus the Lasso, *Journal of Computational and Graphical Statistics* **7**, 397–416.
- GEORGE, E. I. and R. McCULLOCH (1997), Approaches to Bayesian variable selection, *Statistica Sinica* **7**, 339–373.
- JOHNSTONE, I. M. and B. W. SILVERMAN (1998), Empirical Bayes approaches to mixture problems and wavelet regression, Technical Report, University of Bristol and Stanford University.
- JUDITSKY, A. (1997), Wavelet estimators: adapting to unknown smoothness, *Mathematical Methods of Statistics* **6**, 1–25.
- KRIM, H. and I.-C. SCHICK (1999), Minimax description length for signal denoising and optimized representation, *IEEE Transactions on Information Theory* **45**, 898–908.
- LEPORINI, D. (1998), Modélisation statistique et paquets d'ondelettes : application au débruitage de signaux transitoires d'acoustique sous-marine, Thèse de Doctorat de l'Université Paris XI, Orsay.
- LEPORINI D. and J.-C. PESQUET (1999), High-order wavelet packets and cumulant field analysis, *IEEE Transaction on Information Theory* **45**, 863–877.
- LEPORINI, D. and J.-C. PESQUET (2001), Bayesian wavelet denoising: Besov priors and non-Gaussian noises, *Signal Processing* **81**, 55–67.
- MALLAT, S. (1989a), Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$, *Transactions of the American Mathematical Society* **315**, 69–87.
- MALLAT, S. (1989b), A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-11**, 674–693.

- NASON, G. P. (1995), Choice of the threshold parameter in wavelet function estimation, in: A. ANTONIADIS and G. OPPENHEIM (eds.), *Wavelets and statistics*, 261–280, Lecture Notes in Statistics, Springer Verlag
- NASON, G. P. (1996), Wavelet shrinkage using cross-validation, *Journal of the Royal Statistical Society Series B* **58**, 463–479.
- NASON, G. P. and B. W. SILVERMAN (1995), The stationary wavelet transform and some statistical applications, in: A. ANTONIADIS and G. OPPENHEIM (eds.), *Wavelets and statistics*, 281–299. Lecture Notes in Statistics, Springer Verlag
- NEUMANN, M. H. and R. VON SACHS (1995), Wavelet thresholding: beyond the Gaussian i.i.d situation, in: A. ANTONIADIS and G. OPPENHEIM (eds.), *Wavelets and statistics*, Lecture Notes in Statistics, Springer Verlag, 301–330.
- NG, K. W. and D. A. S. FRASER (1994), Inference for linear models with radially decomposable error in: *Multivariate analysis and its applications*, 359–367 IMS Lecture notes, Vol 24.
- NIKIAS, C. L. and M. SHAO (1995), *Signal processing with alpha-stable distributions and applications*, Wiley-Interscience, New York.
- NIKOLOVA, M. (1997), Estimées localement fortement homogènes, *Comptes Rendus de l'Academie des Sciences Serie I* **325**, 665–670.
- OGDEN, R. T. and E. PARZEN (1996a), Change-point approach to data analytic wavelet thresholding, *Statistics and Computing*, **63**, 93–99.
- OGDEN, R. T. and E. PARZEN (1996b), Data dependent wavelet thresholding in nonparametric regression with change-point applications, *Computational Statistics* **22**, 53–70.
- RUGGERI, F. and B. VIDAKOVIC (1999), A Bayesian decision theoretic approach to wavelet thresholding, *Statistica Sinica* **9**, 183–197.
- SIMONCELLI, E. (1999), Bayesian denoising of visual images in the wavelet domain, in: P. MULLER and B. VIDAKOVIC (eds.), *Bayesian inference in wavelet-based models*, Lecture Notes in Statistics, Springer Verlag.
- SIMONCELLI, E. P. and E. H. ADELSON (1996), Noise removal via Bayesian wavelet coring, in: *Proceedings of the IEEE International Conference on Image Processing*, Lausanne, Switzerland, I.379–I.382.
- SMITH, A. and G. ROBERTS (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society Series B* **55**, 3–23.
- TIBSHIRANI, R. (1996), Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.
- TIERNEY, L. (1994), Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1762.
- VIDAKOVIC, B. (1998), Nonlinear wavelet shrinkage with Bayes rules and Bayes factors, *Journal of the American Statistical Association* **93**, 173–179.

Received: September 2000. Revised: December 2001.