# Empirical Bayes approach to improve wavelet thresholding for image noise reduction

*Maarten Jansen*
*Adhemar Bultheel*

*Report TW 296, October 1999*

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

# Empirical Bayes approach to improve wavelet thresholding for image noise reduction

*Maarten Jansen*
*Adhemar Bultheel*

*Report TW 296, October 1999*

Department of Computer Science, K.U.Leuven

**Abstract**

Wavelet threshold algorithms replace wavelet coefficients with small magnitude by zero and keep or shrink the other coefficients. This is basically a local procedure, since wavelet coefficients characterize the local regularity of a function. Although a wavelet transform has decorrelating properties, structures in images, like edges, are never decorrelated completely, and these structures appear in the wavelet coefficients. We therefore introduce a geometrical prior model for configurations of large wavelet coefficients and combine this with the local characterization of a classical threshold procedure into a Bayesian framework. The threshold procedure selects the large coefficients in the actual image. This observed configuration enters the prior model, which, by itself, only describes configurations, not coefficient values. In this way, we can compute for each coefficient the probability of being "sufficiently clean". The parameters of the prior model are estimated on an empirical basis.

**Keywords :** Noise reduction, wavelet, image, threshold, Bayes, Gibbs distribution, Markov random field, pseudo likelihood
**AMS(MOS) Classification :** 41A30, 60G60, 68U10, 65D10

# Empirical Bayes approach to improve wavelet thresholding for image noise reduction

*Maarten Jansen and Adhemar Bultheel*

Department of Computer Science
K.U.Leuven
Celestijnenlaan 200A, B3001 Heverlee - Belgium
`phone: ++32 16 32 7080 fax: ++32 16 32 7996`
`e-mail: maarten.jansen@cs.kuleuven.ac.be`

Technical Report TW296, October 1999

**Abstract**

Wavelet threshold algorithms replace wavelet coefficients with small magnitude by zero and keep or shrink the other coefficients. This is basically a local procedure, since wavelet coefficients characterize the local regularity of a function. Although a wavelet transform has decorrelating properties, structures in images, like edges, are never decorrelated completely, and these structures appear in the wavelet coefficients. We therefore introduce a geometrical prior model for configurations of large wavelet coefficients and combine this with the local characterization of a classical threshold procedure into a Bayesian framework. The threshold procedure selects the large coefficients in the actual image. This observed configuration enters the prior model, which, by itself, only describes configurations, not coefficient values. In this way, we can compute for each coefficient the probability of being "sufficiently clean". The parameters of the prior model are estimated on an empirical basis.

## 1 Introduction

Wavelet thresholding [14, 15] is a popular method for the reduction of noise in images. It assumes that the original, non-corrupted image can be represented in a sparse way by a small number of large wavelet coefficients. In the case of an orthogonal transform, i.i.d. noise is spread out equally over all coefficients. Selecting the coefficients with the largest magnitude therefore removes most of the noise, while preserving the essential image information. Some of these threshold or more general shrinking procedures are based on a Bayesian model [1, 9, 31, 33, 10].

The sparsity of a wavelet representation follows from the decorrelating properties of this transform. This decorrelating is however not complete: a wavelet transform is also a multiscale data representation and the coefficients at subsequent resolution levels tend to be correlated: image features like edges cause high coefficients at several resolutions. More sophisticated methods exploit this multiscale structure [37, 12, 3]. In the case of correlated noise [22] or a non-orthogonal transform, the multiresolution structure of a wavelet transform justifies the application of scale-dependent thresholds.

This paper concentrates on a second type of correlations, which follows from the two-dimensional nature of the input: images have edges and these edges cause two-dimensional clusters of large coefficients. These correlations also appear in the original pixel representation of the image, and they do not follow in the first place from the characteristics of the wavelet transform. So, they are inherent to the very nature of the image, and therefore we introduce a prior model for configurations of large coefficients. This leads to a classical threshold procedure, incorporated into a geometrical Bayesian approach [26, 25, 24, 18]. We also pay attention to the choice of the hyperparameters.

We start with a short section on wavelet nomenclature and wavelet thresholding. Next, we explain the specific problem of a two-dimensional wavelet transform from an approximation theoretic point of view. This section is the elaboration of an idea that E. Candes presented in a discussion with one of us. We then introduce a Bayesian approach to remedy this problem. Section 5 fixes and motivates the choice of prior and conditional model. The next section composes the actual algorithm. Section 7 deals with the hyperparameters in the model and proposes how to find appropriate values for these parameters. The following section discusses some results and we end with a summary and conclusions. This text was inspired by previous work by Malfait et al. [26]. Nevertheless, as we explain in Section 8.3, our prior and condtional model (Sections 4 and 5) is original and based on a more theoretical ground. Moreover, we combine this with an empirical parameter choice (Section 7) and we motivate the Bayesian approach by a comparison with more heuristical ideas in Section 4.4.

# 2 Wavelets and wavelet thresholding

## 2.1 Wavelet terminology

This section fixes notation with respect to wavelet transforms and wavelet functions. For an introduction to wavelet theory, we refer to the extensive literature. As examples, we mention [27, 34, 21]. A discrete wavelet transform is a decomposition of an input vector into coefficients at different resolution levels. Each resolution represents image details of a specific scale. In two dimensions, we use a tensor-product like extension of a one-dimensional transform, the so called square wavelet-transform: at each level the coefficients belong to one of three *subbands* or *components*, corresponding to three *orientations* in the image: the coefficients carry vertical, horizontal, or diagonal information. This discrete transform has a continuous *interpretation*: it decomposes an expansion of a function in a scaling basis at resolution level $J$

$$f(x,y) = \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} s_{J,k,l}\varphi_{J,k}(x)\varphi_{J,l}(y)$$

into an expansion in a basis of wavelet functions $\psi$ at different scales, at different locations for the three orientations:

$$\begin{aligned} f(x,y) &= \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} s_{L,k,l}\varphi_{L,k}(x)\varphi_{L,l}(y) \\ &+ \sum_{j=L}^{J-1}\sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} w^{\mathrm{diag}}_{j,k,l}\psi_{j,k}(x)\psi_{j,l}(y) \\ &\qquad\qquad + w^{\mathrm{vert}}_{j,k,l}\psi_{j,k}(x)\varphi_{j,l}(y) \\ &\qquad\qquad + w^{\mathrm{hor}}_{j,k,l}\varphi_{j,k}(x)\psi_{j,l}(y) \end{aligned}$$

Instead of the classical fast wavelet transform, we use the so called non-decimated, redundant, or stationary wavelet transform [29, 30, 26]. While the fast transform has linear complexity, this alternative is of $\mathcal{O}(N \log N)$ complexity, both in computations and in memory requirements. On the other hand, the stationary transform is translation invariant and generates the same number of coefficients in each subband at each scale. This facilitates interscale coefficient comparisons. Moreover, the reconstruction procedure [29] from this redundant data representation is of course not unique. For manipulated data, a linear combination of all possible reconstruction schemes causes an additional smoothing of the result [29].

## 2.2 Wavelet thresholding

A wavelet threshold algorithm typically consists of three steps: first, the observational data are transformed into empirical wavelet coefficients. The next step is a manipulation of the coefficients and finally, an inverse transform of the modified coefficients yields the result.

Of course, the second step is the crucial one. Apart from the choice of the wavelet basis, all strategies and options are concentrated in this step. The manipulation is based on a *classification* of the coefficients: the most common threshold procedures use a binary classification: a coefficient is either dominated by noise *or* sufficiently important. The magnitude of the coefficient is the *criterion* of classification: we consider the absolute value of each coefficient as a measure of significance. This is motivated by the *ansatz* that a wavelet transform is a sparse data representation: a few large

Figure 1: An image (Left) with artificial, additive correlated noise (Right), SNR $= 4.97$dB. The noise is the result of a convolution of white noise with a FIR-highpass-filter.

coefficients carry nearly all information, while most coefficients are small and dominated by noise.

So, a simple threshold algorithm classifies the coefficients by their magnitudes. Coefficients below a certain *threshold* $\lambda$ are replaced by zero. Coefficients with absolute value above the threshold are kept in the *hard-threshold* approach or shrunk with a value $\lambda$ in the *soft-threshold* approach. Apart from these most popular rules, there exist a whole variety of shrinking rules, some of which result from a Bayesian model, as mentioned in the introduction.

Figure 1 shows an image with artificial, additive, homoscedastic correlated noise. This noise was the result of a convolution of white noise with a FIR-highpass-filter (A FIR or finite impulse response filter has a finite number of taps). The signal-to-noise ratio is 4.97 dB, where we define signal-to-noise ratio as:

$$SNR = 10 \log_{10} \left( \frac{\frac{1}{N} \sum_{i=1}^{N} (f_i - \overline{f})^2}{\sigma^2} \right),$$

with $\overline{f} = \frac{1}{N} \sum_{i=1}^{N} f_i$ and $f_i$ is the uncorrupted gray intensity of pixel $i$. We apply a redundant wavelet transform. As wavelet filter, we use the variation on the CDF-(spline)-filters "with less dissimilar lengths" [11, 2]. We choose a basis with four primal and four dual vanishing moments. These wavelets are rather popular in image processing: they are smooth and have compact support.

Since we know the uncorrupted image, we can compute at each level the threshold that minimizes the mean square error (MSE) of the wavelet coefficients. Since we use a bi-orthogonal (i.e. not an orthogonal) transform, this is not exactly the same as a minimization of the MSE of the pixel values, although a *stable* basis (with Riesz constants close to each other [13]) guarantees a quasi-equivalence. Moreover, the eventual objective is visual quality, and there seems no reason why minimization in terms of original pixels is better than minimization in terms of wavelet coefficients. Of course, the user views images in the pixel domain, but we do not interpret an image pixel by pixel. Although a discussion about the human visual system is far beyond the scope of this text, a multiresolution analysis is said to be closer to the way we view an image. Note that MSE minimization is equivalent to SNR maximization.

In practical situations, the minimum MSE threshold cannot be computed exactly. Therefore we estimate this threshold using a generalized cross validation (GCV) procedure [35, 20]. This method is as fast as the wavelet transform, requires no estimation for the noise deviation $\sigma$ and it is asymptotically optimal, i.e. if the number of data is getting large, the estimated threshold approaches the minimum MSE threshold. These properties are preserved — on a level-dependent basis — if the noise is correlated in the wavelet representation [19].

The minimum MSE threshold is based on a global compromise between noise and data: this is not the best thing we can do: instead of applying one threshold for all coefficients at a given level, we would like to decide for each coefficient separately what is best: keeping or killing. If we know the noise deviation $\sigma$ and if an *oracle* would tell us whether the noise-free magnitude $V_s$ is above or below $\sigma$, then the expected MSE is minimized by replacing $W_s$ with zero if $|V_s| < \sigma$ and keeping it otherwise [14]. This is an ideal (clairvoyant) *diagonal projection*: instead of applying the minimum MSE threshold (or its estimate) to the noise-free coefficients, we choose a threshold on these uncorrupted coefficients equal to the noise deviation. The resulting label image is shown in Figure 8 (Right picture). Since neither $V_s$ nor $\sigma$ are known in practice, this is of course an ideal case.

Figure 2 compares the result from soft-thresholding with level- and subband-dependent GCV-threshold with the result from the oracle selection. Only the three finest scales were processed. Signal-to-noise ratio is respectively 18.02 dB and 21.05 dB.

## 3 An approximation theoretic point of view

Image processing is not merely a two-dimensional translation of traditional signal processing techniques. The two-dimensional character has some important consequences, such as the existence of line singularities, manifesting as edges. The observations explained in this section also provide the basis for the development of new types of basis functions, such as ridgelets [7].

(a) (b)

Figure 2: Left: the result from thresholding with level- and subband-dependent GCV-threshold. Only the three finest scales were processed. SNR = 18.64 dB. Right: result from the optimal (clairvoyant) diagonal projection. SNR = 21.05 dB.

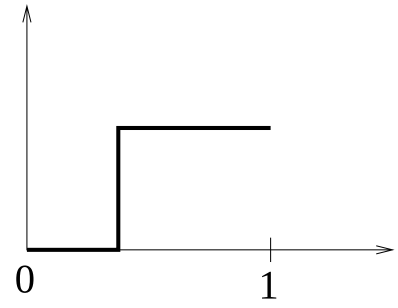



Figure 3: Step function.

### 3.1 Step function approximation in one dimension

Suppose we want to approximate the step function of Figure 3. A periodic extension of this function can be decomposed into a Fourier series:

$$f(x) = \sum_{k \in \mathbb{Z}} a_k \mathrm{e}^{-i2\pi kx}.$$

The equation sign indicates convergence, both pointwise as in $L_2[0,1]$-norm. The coefficients $a_k$ depend of course on the precise position of the singularity, but they behave like:

$$a_k = \mathcal{O}\left(\frac{1}{|k|}\right).$$

We use this Fourier expansion to *approximate* the step function by taking the $2n + 1$ harmonics with index $k = -n, \ldots, n$:

$$f_n(x) = \sum_{k=-n}^{n} a_k \mathrm{e}^{-i2\pi kx}.$$

Since the coefficients decrease for $|k| \to \infty$, this *linear* approach happens to coincide with taking the largest contributions. This Fourier basis is orthonormal, so the approximation error satisfies

$$\begin{aligned}\|\epsilon_{2n+1}\|^2_{L_2[0,1]} &= \|f - f_{2n+1}\|^2_{L_2[0,1]} = \sum_{|k|\geq n+1} |a_k|^2 \\ &= \mathcal{O}\left(2 \sum_{k=n+1}^{\infty} \frac{1}{k^2}\right) = \mathcal{O}(n^{-1}).\end{aligned}$$

We may conclude that a one dimensional Fourier decomposition performs as:

$$\|\epsilon_n\| = \mathcal{O}(n^{-1/2}).$$

This bad approximation of piecewise smooth signals is a well known drawback of the Fourier decomposition. The reason is of course that all basis functions cover the entire interval, and so all of them get in touch with the singularity, all of them have a contribution to it.

This is not the case in a wavelet decomposition, where at each scale only a constant number of coefficients are non-zero. For the Haar transform, there is only one function, say $\psi_{j,k_j}$ with a non zero coefficient $w_{j,k_j}$. Orthogonality says that:

$$\|\epsilon_J\|^2_{L_2[0,1]} = \|f - f_J\|^2_{L_2[0,1]} = \sum_{j=J+1}^{\infty} w^2_{j,k_j}.$$

If $|f(x)| \leq 1$, we have

$$|w_{j,k_j}| \leq \int_0^1 |\psi_{j,k_j}(x)|\, dx = 2^{-j}\, 2^{j/2}.$$

Hence

$$\|\epsilon_J\|^2_{L_2[0,1]} = \mathcal{O}\left(\sum_{j=J+1}^{\infty} 2^{-j}\right) = \mathcal{O}(2^{-J}).$$

This expresses that a wavelet basis indeed captures isolated singularities much more efficiently than does a Fourier basis. We do not know in advance which coefficients are going to be non-zero: this depends on the input signal and more precisely on the exact position of the jump. Therefore, keeping the non-zero wavelet coefficients is a *non-linear* approximation.

If we have a superposition of this step function $f$ and a $C^\alpha$ smooth function $g$, none of the wavelet coefficients of $h = f + g$ is exactly zero.

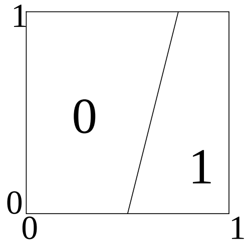



Figure 4: Two dimensional step function.

The smooth part $g$ is best approximated with a linear approach: take all coefficients up to scale $J$. If the number of vanishing moments $p \geq \alpha$, the approximation error satisfies [27]:

$$\|\epsilon_{g,J}\| = \|g - g_J\| = \mathcal{O}(2^{-J\alpha}).$$

This approximation uses $2^{J+1}$ coefficients. If we want the same order of precision for the non-smooth component, we add $\lceil J\alpha \rceil$ coefficients of $f$ in the non-linear way, and the overall approximation error $\epsilon_h(x) = \epsilon_f(x) + \epsilon_g(x)$ is then bounded by:

$$\|\epsilon_h\| \leq \|\epsilon_f\| + \|\epsilon_g\| = \mathcal{O}(2^{-J\alpha}).$$

This approximation uses $2^{J+1} + \lceil J\alpha \rceil = \mathcal{O}(2^J)$ coefficients. If we call this number $n$, we conclude that the error of a one-dimensional wavelet approximation behaves as

$$\|\epsilon_n\| = \mathcal{O}(n^{-\alpha}),$$

for smooth as well as for piecewise smooth functions. Isolated singularities have no influence on the asymptotic approximation error.

### 3.2 Approximations in two dimensions

Now suppose we are given a two dimensional function $f(x,y) \in L_2[0,1]^2$, which is 0 in one part of the square and 1 in the other part. The boundary between these two parts is a simple line, as in Figure 4. The coefficients of the Fourier expansion of this function

$$f(x,y) = \sum_{k \in \mathbb{Z}} \sum_{l \in \mathbb{Z}} a_{k,l} \, \mathrm{e}^{i2\pi kx} e^{i2\pi ly}$$

can be found as:

$$\begin{aligned} a_{k,l} &= \int_0^1 \int_0^1 f(x,y) \, \mathrm{e}^{-i2\pi kx} e^{-i2\pi ly} \, dx dy \\ &= \mathcal{O}\left(\frac{1}{|kl|}\right). \end{aligned}$$

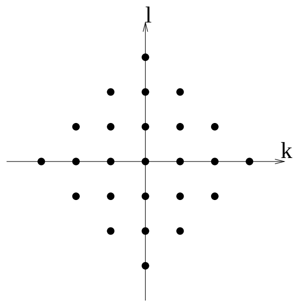



Figure 5: Position of indices in $\mathbb{Z}^2$ corresponding to coefficients in a linear Fourier approximation.

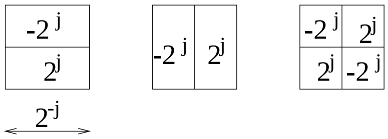



Figure 6: Two-dimensional Haar basis functions.

A linear approximation keeps all coefficients with indices $k$ and $l$ such that $|k| + |l| \leq m$, for a given $m$. Figure 5 shows the position of these indices in $\mathbb{Z}^2$ for $m = 3$. The approximation error satisfies:

$$\|\epsilon_n\|^2 = \sum_{k \in \mathbb{Z}} \sum_{l:|k|+|l| \geq m+1} |a_{k,l}|^2 = \mathcal{O}(m^{-1})$$

Since $n = (2m+1)^2$, we have

$$\|\epsilon_n\| = \mathcal{O}(n^{-1/4}).$$

This says that for an equal order of magnitude of the error as in the one-dimensional case, we need the square of the number of coefficients.

We now proceed to a (Haar) wavelet expansion. The basis contains three types of functions: horizontally oriented, vertically oriented and diagonally oriented functions as depicted in Figure 6. At each scale we have $K2^j$ non-zero coefficients with all three types of basis functions. The exact value of the constant $K \geq 1$ depends on the length of the singularity, i.e. its orientation in the image. If $|f(x,y)| \leq 1$, the coefficients satisfy:

$$|w_{j,k,l}^{\text{diag}}| \leq \int_0^1 \int_0^1 |\psi_{j,k,l}^{\text{diag}}(x,y)|\, dx dy = 2^j\ 2^{-j} \times 2^{-j},$$

and similarly for horizontal and diagonal subbands. The approximation error becomes:

$$\|f - f_J\|^2 \leq 3 \sum_{j=J+1}^{\infty} K 2^j 2^{-j2} = \mathcal{O}(2^{-J}).$$

This is the same order of approximation as in the 1D-case, we now need

$$n = \sum_{j=0}^{J} K 2^j = \mathcal{O}(2^J)$$

coefficients. So the order of approximation is now:

$$\|f - f_n\| = \mathcal{O}(n^{-1/2}),$$

while in the one-dimensional case we had $n = J$ and

$$\|f - f_n\| = \mathcal{O}(2^{-n}).$$

This is not just squaring the number of coefficients to obtain a comparable error in two dimensions. This dramatic change comes from the difference between a point singularity in one dimensional signals and a line singularity in 2 dimensions. A point has no dimension, at each scale it only interferes with a fixed number of basis functions. A line has however a certain length. Consequently, the number of basis functions meeting this line increases for finer scales.

For a piecewise smooth function of the form $h = g + f$, with $g \in C^\alpha$, we have $\|g - g_J\| = \mathcal{O}(2^{-J\alpha})$, provided that the wavelet basis has $p \geq \alpha$ vanishing moments. This linear approximation uses $n = \mathcal{O}(2^{2J})$ coefficients, so the order of approximation is $\mathcal{O}(n^{-\alpha/2})$. We need coefficients at $\lceil 2J\alpha \rceil$ resolution levels to represent $f$ with the same accuracy. This means $\mathcal{O}(2^{2J\alpha})$ additional non-zero coefficients. The total number of coefficients to achieve $\|h - h_n\| = \mathcal{O}(2^{-J\alpha/2})$ is then $n = \mathcal{O}(2^{2J\alpha} + 2^J) = \mathcal{O}(2^{2J\alpha})$. A wavelet approximation for a piecewise smooth function in two dimensions thus has an accuracy of $\mathcal{O}(n^{-1/2})$. Unlike in the one-dimensional case, the line singularity does have an important impact on the quality of the wavelet approximation: all the benefits from using more than one vanishing moment seem to be lost.

## 3.3 Other basis functions

From the previous analysis, we conclude that wavelets may not provide the ultimate representation for images. It is of course true that wavelets *are* successful, but looking for better generalizations of the wavelet idea for more-dimensional applications is an interesting — though difficult — challenge. The previous argument comes from approximation theory, but it has consequences in statistical estimation: a basis that performs well in approximation of piecewise smooth functions is appropriate for noise reduction. This is a motivation for the development of new types of basis functions, like ridgelets [7].

This text takes a different approach: it applies the classical two-dimensional wavelets and concentrates on the coefficients in the basis for a description of edges. This description is based on a random prior model for these coefficients and leads to a Bayesian algorithm.

# 4 The Bayesian approach

## 4.1 Motivation and objectives

In one dimension, as in two dimensions, wavelet basis functions are localized in space and scale (frequency). As a consequence, manipulating a coefficient has a local effect, both in space and frequency. This is an important advantage of wavelet based methods.

On the other hand, usual *classification* rules are local too, and do not take into account all the correlations that exist among neighboring coefficients. Although a wavelet transform has decorrelating properties, this decorrelation is not complete. We distinguish two types of correlations:

1. Important image features correspond to large coefficients at different scales: these coefficients are of course correlated. This type of correlation is inherent to all wavelet decompositions: it reflects the multiscale nature of it. There exist some *deterministic* algorithms [37, 3] that take this multiresolution character into account. Other algorithms start from different variations of a stochastic 'tree' model for uncorrupted wavelet coefficients in a multiscale structure [8, 12, 23].

2. The second type of correlation is within one scale and is specific for two-dimensional inputs, like images: important coefficients tend to be clustered on the location of edges.

We assume that classical thresholding, possibly extended to deal with interscale correlations, performs sufficiently well for the first type of inter-coefficient dependencies. This paper concentrates on the second type of correlations. For the stochastic description of these structures, we need *geometrical* prior models. This leads to a multiscale version of *Markov Random Field Models* (MRF). Similar approaches are in [4, 6, 26, 25, 24].

The basis for our approach remains the thresholding philosophy. The optimal selection based on the 'oracle' information, as explained in Section 2.2, is an ideal benchmark, and we hope that the incorporation of a prior model helps in mimicking this oracle. The Bayesian approach aims at two objectives at once: by taking into account the geometrical structures in the coefficients, we want to come closer to the ideal coefficient selection. As becomes clear from the subsequent sections, both objectives are reflected by the Bayesian model that we use.

## 4.2 Plugging the threshold procedure into a fully random model

Whereas typical threshold algorithms are based on this heuristical approach that the largest coefficients capture the essential image features, *Bayesian* methods start from a full model for wavelet coefficients of the following type:

$$\boldsymbol{W} = \boldsymbol{V} + \boldsymbol{N}$$

This is an additive model for wavelet coefficients where $\boldsymbol{N}$ is the noise vector, $\boldsymbol{V}$ is the uncorrupted wavelet coefficient vector, and $\boldsymbol{W}$ is the input (empirical) wavelet coefficient vector. Both the noise and the noise-free data are viewed as realizations of a probability distribution. We now describe how threshold procedures fit into this model.

A wavelet threshold algorithm consists of three steps: first, the observational data are transformed into empirical wavelet coefficients. The next step is a manipulation of the coefficients and finally, an inverse transform of the modified coefficients yields the result. When extending this thresholding with a Bayesian approach, we leave the three steps intact, but we build in more uncertainties in the second step. As explained in Section 2.2, the selection criterion used in the second step is based on a measure of *regularity.* This measure of significance $\boldsymbol{M}$ is a function of the observation:

$$\boldsymbol{M} = m(\boldsymbol{W}),$$

Wavelet coefficients with a significance below a threshold $\lambda$, are classified as noisy. With each wavelet coefficient $W_s$, the algorithm associates a 'label' or 'mask' variable $X_s$ such that:

$$X_s = \begin{cases} 0, & \text{if } W_s \text{ is noisy according} \\ & \text{to the criterion, i.e. if } M_s < \lambda \\ 1, & \text{then } W_s \text{ is sufficiently clean, i.e. if} \\ & M_s \geq \lambda \end{cases} \tag{1}$$

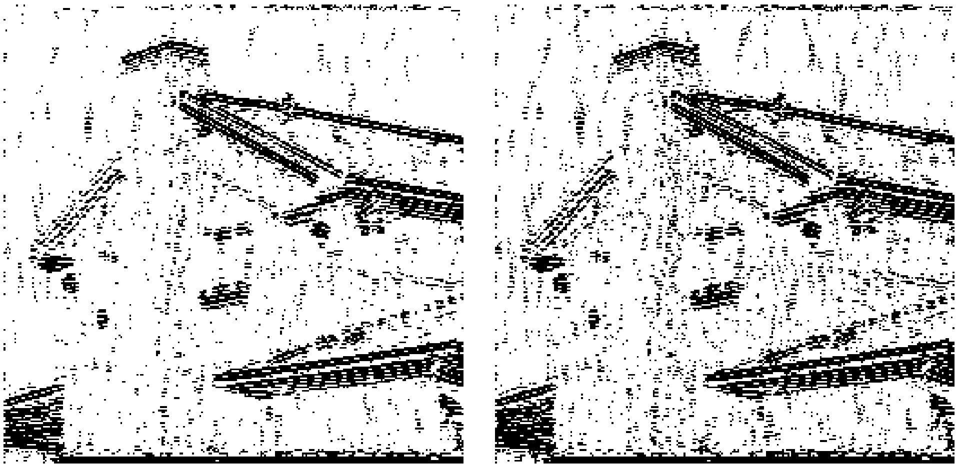
In these and following equations, $s$ represents the 'multidimensional' index of a wavelet coefficient on a given resolution level $j$ and for a given component $m$ (vertical, horizontal, or diagonal): $s = (k, l)$. To avoid overloaded notations we omit the resolution level $j$ and the component $m$ in our equations, and use the simple index $s$. So, if no confusion is possible, we write $W_s$ instead of $W^m_{j;s}$ or $W^m_{j;k,l}$. This classification is followed by the modification step: If $W_{\lambda s}$ denotes the modified coefficient, with subscript $\lambda$ referring to the threshold value, we write:

$$W_{\lambda s} = h(W_s, M_s, X_s),$$

for some *action* $h(W_s, M_s, X_s)$. The classic *hard-threshold* procedure corresponds to

$$h(W_s, M_s, X_s) = W_s X_s.$$

It keeps the 'uncorrupted' coefficients and replaces the noisy ones by zero.



(a) (b)

Figure 7: Mask or label images, corresponding to the horizontal component of the one but finest scale. Black pixels represent coefficients with magnitude above the threshold. Left: using the minimum MSE threshold. Right: using a GCV estimate of this threshold.

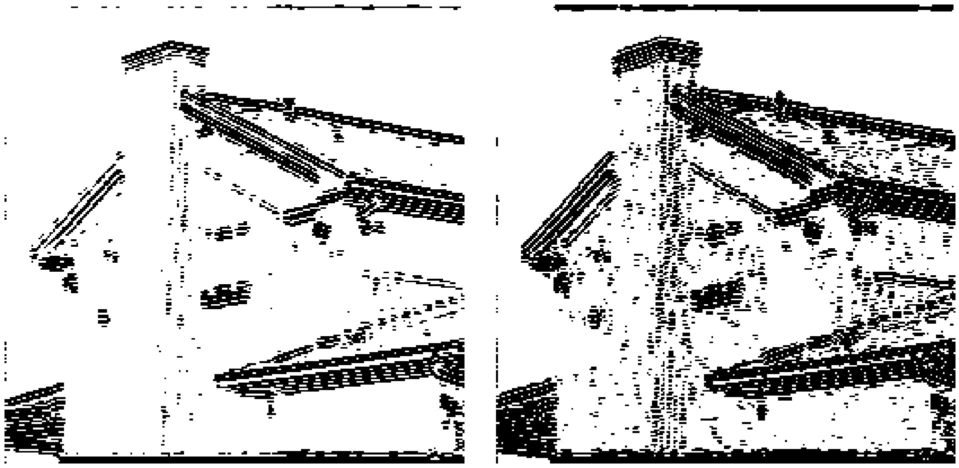
### 4.3 Threshold mask images

Figure 7 visualizes the binary label image $\boldsymbol{X}$, i.e. it shows in black the position of the selected wavelet coefficients for the horizontal subband at the one but finest resolution level from the non-decimated wavelet transform of the noisy image in Figure 1. The one but finest scale in the wavelet transform is two scales below the original image resolution, and, as before, we use the variation on the CDF-(spline)-filters "with less dissimilar lengths" [11, 2]. Primal and dual wavelets have four vanishing moments. The mask on the left-hand side was obtained by soft-thresholding using the minimum MSE threshold. The mask on the right-hand side is obtained by a generalized cross validation threshold. Applying soft-thresholding using this mask (and its analogues for other components and scales) leads to the output in Figure 2(a).

If we apply the same threshold to the noise-free coefficients, we get the left picture in Figure 8. We see that many of the isolated pixels have disappeared: they were due to noise. Applying the optimal selection, inspired by the 'oracle' information leads to an even more structured image in Figure 8(b).

### 4.4 Binary image enhancement methods

A comparison of the different label images clearly illustrates that thresholding each coefficient separately does not take into account the image structures. An obvious way to recover the optimal mask of Figure 8 (b), is trying



(a) (b)

Figure 8: Same mask images as in Figure 7, here based on noise-free coefficients. Left: black pixels indicate noise-free coefficients with magnitude above the previous threshold. Right: black pixels indicate noise-free coefficients with magnitude larger than noise deviation. This corresponds to the ideal wavelet selection: if an "oracle" [14] tells us whether or not a coefficient is dominated by noise, this is the best thing we can do.

classic enhancement methods. Figure 9 (a) shows the label image after applying a median filter to the approximate minimum MSE labels in Figure 7. Another possibility is the application of so called erosion-dilation methods: these methods proceed in two steps: in the first step, black pixels with less than, for instance, two black neighbors are removed. This erosion can be repeated several times, before going to the dilation. This second step tries to restore the eroded objects by turning white background pixels into black object pixels, if there is already an object in the neighborhood. (This neighborhood is typically a $3 \times 3$ box containing the actual pixel in its center.) Figure 9 (b) contains the result of this operation. It is hard to preserve the fine edge structures, while removing the noisy pixels. These operations act on the label images only and forget about the background behind them: these pixels come from a wavelet coefficient classification. We would prefer a method that can deal with the geometry *and* the local coefficient values at the *same* time. Bayes' rule tells us how we can do so.

### 4.5 Bayesian classification

The classification (1) in a threshold algorithm is a deterministic function of the empirical coefficients: thresholding on magnitudes corresponds to a simple step function, as illustrated in Figure 10. Recall that, in this text, the measure of significance is the coefficient magnitude: $\boldsymbol{M} = |\boldsymbol{W}|$. However, it

(a) (b)

Figure 9: Result of elementary binary image enhancement methods on the approximate minimum MSE label image in Figure 7. Left: a median filter; Right: an erosion-dilation procedure.

would be interesting to examine measures based on interscale-correlations: e.g. $M_{j;s}^{m} = \prod_i W_{j+i;s}^{m}$, where $j$ is the current resolution level and $m$ the orientation ($m = \text{HOR}, \text{VER}, \text{DIAG}$).

Because we want to take the spatial configurations into account, we give up this tight relation between a coefficient value and its classification. We introduce a *prior model* for coefficient classification configurations $\boldsymbol{X}$. This prior should express the belief that clusters of noise-free coefficients have a higher probability than configurations with many isolated labels. In particular, edge-shaped clusters should be promoted. The prior model rests on the *classification* of the coefficients, not on the uncorrupted coefficients themselves. A similar idea is the use of Hidden Markov Models [8, 12, 23, 4, 6].

Next, the *conditional model* (likelihood function) states that if the classification label for a coefficient equals one, this coefficient is *probably* above the threshold. A zero label means that the corresponding coefficient is probably small. The classical, deterministic approach can be seen as an extreme case of this probability model, where, for example, a label $X = 0$ tells that the coefficient is *certainly* in the range $[-\lambda, \lambda]$. This appears in Figure 10(b).

If we have a prior distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$ and a conditional model $f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})$, then Bayes' rule allows to compute the *posterior* probability:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}) = \frac{\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})}{f_{\mathbf{M}}(\boldsymbol{m})}$$

In a given experiment, where $\boldsymbol{m}$ is fixed, the denominator is a constant. As we explain in Section 6.2, it is sufficient to know the *relative* probabilities

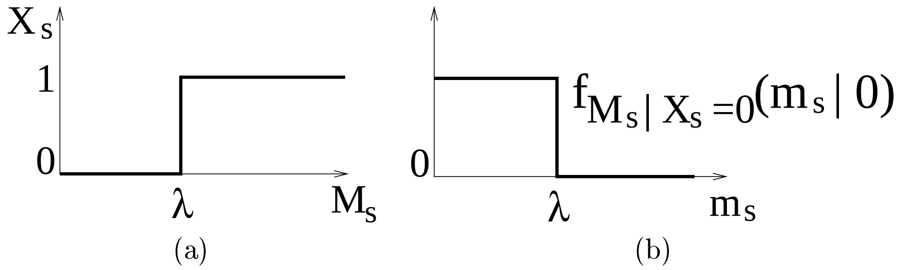



Figure 10: Left: The deterministic classification function for coefficient magnitude thresholding: if a coefficient magnitude $M$ is below the threshold value $\lambda$, it is classified as noisy ($X = 0$), otherwise it is called sufficiently clean ($X = 1$). Right: this deterministic approach is a special case of the Bayesian model, where the conditional density is zero for coefficient magnitudes above the threshold if $X = 0$ and beneath the threshold if $X = 1$.

of configurations, and therefore we write:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{M} = \boldsymbol{m}) = C \cdot \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})$$

# 5 Prior and conditional model

## 5.1 The prior model

As explained above, we are looking for a multivariate model for a binary image $\boldsymbol{X}$. Expressing a probability function for all $2^{NM}$ possible values in a $N$ by $M$ label image may be cumbersome. We therefore construct the model starting from local descriptions of clustering. The prior probability function can always be written as:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z} \exp[-H(\boldsymbol{x})], \tag{2}$$

with the *partition function* $Z$:

$$Z = \sum_{\boldsymbol{x}} \exp[-H(\boldsymbol{x})].$$

$H(\boldsymbol{x})$ is the *energy function* of a configuration $\boldsymbol{x}$: the lower the energy, the higher the prior probability. To express that this energy comes from *local* interactions only, we first define for each pixel index $s$ in the lattice $S$ its set of neighbors, i.e. the set of indices $\partial s \subset S$ that interact with $s$. We assume that $s \notin \partial s$ and $s \in \partial t \Leftrightarrow t \in \partial s$. A set of indices that are all neighbors to each other is called a *clique*. The set of all cliques is

$$\mathcal{C} = \{C \in 2^S \mid \forall s, t \in C : t \in \partial s\}$$

If the total energy of a configuration equals the sum of its clique potential functions:

$$H(\boldsymbol{x}) = \tau \sum_{C \in \mathcal{C}} U_C(\boldsymbol{x}_C),$$

the probability function $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$ is called a *Gibbs distribution* with respect to the given neighborhood system $\{S, \partial\}$. The hyperparameter $\tau$ measures the *rigidity* of the configuration. The higher $\tau$, the less likely are status changes due to noise. As the equation indicates, the clique potential $U_C$ only depends on the values of $\boldsymbol{x}$ in the sites that belong to $C$.

Whereas Gibbs distributions are based on local energies, Markov Random Fields (MRF) are based on local statistical dependencies. A Markov Random Field, relative to a neighborhood system $\{S, \partial\}$ is a probability function $P$ with the two-dimensional Markov property:

$$\mathrm{P}(X_s = x_s | \boldsymbol{X}_{S \setminus \{s\}} = \boldsymbol{x}_{S \setminus \{s\}}) = \mathrm{P}(X_s = x_s | \boldsymbol{X}_{\partial s} = \boldsymbol{x}_{\partial s}).$$

This definition does not use the notion of *clique*.

The Hammersly-Clifford theorem states that MRF's and Gibbs distributions are the same:

**Theorem 1 (Hammersly-Clifford)** *A probability function is a Markov Random Field with respect to a neighborhood system if and only if it is a Gibbs distribution with respect to same neighborhood system.*

Proofs are in [5, 36]. This theorem facilitates the computation of conditional probabilities in the lattice of a Gibbs distribution: this computation only uses local information. The computation of marginal probabilities of a MRF is well served by the Gibbs distribution property. Especially in expressions like

$$\frac{\mathrm{P}_{\mathbf{X}}(\boldsymbol{u})}{\mathrm{P}_{\mathbf{X}}(\boldsymbol{v})}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ differ in a couple of lattice points $s$ only, it is easy to use the theorem and limit the calculations to the potentials of cliques that contain one of these points $s$:

$$\frac{\mathrm{P}_{\mathbf{X}}(\boldsymbol{u})}{\mathrm{P}_{\mathbf{X}}(\boldsymbol{v})} = \exp\left(-\tau \sum_{s} \sum_{C \in \mathcal{C}, s \in C} U_C(\boldsymbol{u}_C) - U_C(\boldsymbol{v}_C)\right).$$

Actually, a Gibbs distribution is mostly the only practically possible specification of a Markov Random Field: it is hard to check whether a collection of local conditional probabilities form a coherent set for Markov Random Field [16].

The application of this MRF's and Gibbs distributions in image analysis and image processing is still growing and our list [16, 17, 36] is nothing but a snapshot. An often used Gibbs distribution is the Ising model. The neighbors of a pixel with index $s$ is a $3 \times 3$ submatrix, excluding $s$ in its center. The model only considers pairs of neighbors. Other cliques in the system have no energy. The total energy is:

$$H(\boldsymbol{x}) = \sum_{\{s,t\} \in \mathcal{C}} x_s x_t.$$

For our experiments, we use a slightly different model, in a $5 \times 5$-neighborhood system. We only consider $3 \times 3$ cliques, i.e. the largest possible type in a 5 by 5 neighborhood system. The potentials of all other types of cliques are set to zero. For a 3 by 3 clique $C$ we set:

$$U_C(\boldsymbol{x}_C) = \frac{\sum_{s \in C} x_s \sum_{t \in C \cap \partial s} (1 - 2x_s x_t)}{\sum_{s \in C} x_s},$$

i.e. for each $x_s = 1$, we subtract the number of neighbors within the clique with value one from the number of neighbors with zero value. The sum of these results is divided by the number of labels with value $x_s = 1$, to obtain a mean value. The idea behind this potential function is to compute a kind of "average degree of isolation" within the clique for the pixels with value one. The background pixels are considered to be neutral. Unlike the classical Ising model, this function is not symmetric for interchanges of ones and zeros.

### 5.2 The conditional model

We also need a conditional density $f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})$. Whereas the prior describes the clustering of significant wavelet coefficients, this conditional model deals with the *local* significance measure. Therefore we write

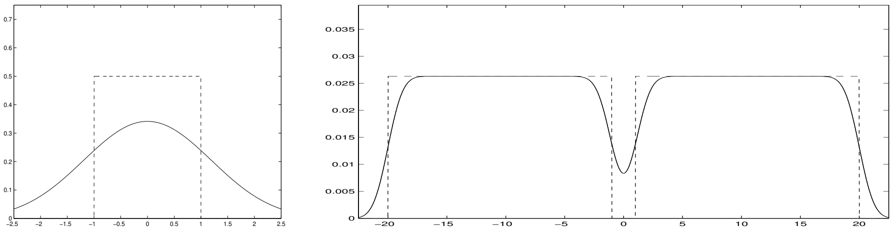
$$f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x}) = \prod_{s \in S} f_{M_s|X_s}(m_s|x_s).$$

This density expresses that if the label $X_s = 1$, i.e. if the corresponding wavelet coefficient is sufficiently noise-free, a large value of $M_s$ is probable. Referring to the ideal selection procedure, we now impose the idea that selected coefficients should have an untouched value above the noise deviation $\sigma$. This means that if $X_s = 1$, $V_s$ is for instance uniformly distributed on $[-\mu, -\sigma] \cup [\sigma, \mu]$, $\mu$ being the maximum coefficient magnitude, which is a parameter of the model that has to be determined. If the noise $N_s \sim n(0, \sigma)$ has a Gaussian density, it is easy to verify that

$$f_{W_s|X_s}(w|1) = \frac{1}{2(\mu - \sigma)} \left[ \Phi(w + \mu) - \Phi(w + \sigma) + \Phi(w - \sigma) - \Phi(w - \mu) \right],$$

where $\Phi(z)$ is the cumulative Gaussian distribution. A similar argument leads to the following conditional model for coefficients dominated by noise:

$$f_{W_s|X_s}(w|0) = \frac{1}{2\sigma} \left[ \Phi(w + \sigma) - \Phi(w - \sigma) \right].$$




(a): $f_{W_s|X_s}(w|0)$ (b): $f_{W_s|X_s}(w|1)$

Figure 11: Conditional probability densities in Bayesian model. The model expresses that if a label $X_s = 0$, the corresponding coefficient *probably* has a small magnitude, but magnitude is no longer a strict selection criterion: a small coefficient might be important and a large coefficient might be replaced by zero.

Figure 11 shows these density functions for $\sigma = 1$ and $\mu = 20$. This model expresses that if a label $X_s = 0$, the corresponding coefficient *probably* has a small magnitude, but magnitude is no longer a strict selection criterion: a small coefficient might be important and a large coefficient might be replaced by zero.

Other, perhaps more realistic models follow from the assumption that the important, noise-free coefficients are exponentially distributed:

$$f_{V_s|X_s}(v|1) = \frac{\rho \mathrm{e}^{\rho\sigma}}{2} \mathrm{e}^{-\rho|v|}.$$

This leads to the following expression for the coefficients corrupted by noise:

$$\begin{aligned} f_{W_s|X_s}(w|1) &= \frac{\rho\, \mathrm{e}^{\sigma\rho + \sigma^2\rho^2/2}}{2} \\ &\quad \left[\mathrm{e}^{-\rho w}\Phi(w - \sigma - \sigma^2\rho) + \mathrm{e}^{\rho w}(1 - \Phi(w + \sigma + \sigma^2\rho))\right]. \end{aligned}$$

Even more general models for noise-free wavelet coefficients are Laplacian distributions [32]:

$$f_V(v) = K\, \mathrm{e}^{-|kv|^\kappa}.$$

Typical values for $\kappa$ range between 0.5 and 0.8.

# 6 The Bayesian algorithm

## 6.1 Posterior probabilities

From Bayes' rule, we can compute the posterior probabilities

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}) = \frac{\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}) f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})}{f_{\mathbf{M}}(\boldsymbol{m})}$$

With these probabilities, a *Bayesian decision rule* leads to an estimation of the optimal label. The Maximum A Posteriori (MAP) procedure chooses the mask $\boldsymbol{x}$ with the highest posterior probability. The Maximal Marginal Posterior (MMP) rule is a more local approach: it computes in each site $s$ the marginal probabilities:

$$\mathrm{P}(X_s = 1|\boldsymbol{M} = \boldsymbol{m}) = \sum_{\boldsymbol{x}} x_s \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}),$$

and if this probability is more than 0.5, the pixel gets value 1. Both decision rules have a binary outcome: each coefficient is classified as noisy ($X = 0$) or relatively uncorrupted ($X = 1$). We would like to exploit the entire posterior probability: the posterior mean value

$$E(X_s|\boldsymbol{M} = \boldsymbol{m}) = \mathrm{P}(X_s = 1|\boldsymbol{M} = \boldsymbol{m})$$

preserves all information. It is a minimum least squares estimator. This classification leads to a posterior 'expected action':

$$\begin{aligned}\mathrm{E}(W_{\lambda s}|\boldsymbol{M}) = {} & h(W_s, m_s(\boldsymbol{W}), 1)\mathrm{P}(X_s = 1|\boldsymbol{M}) \\ & + h(W_s, m_s(\boldsymbol{W}), 0)\mathrm{P}(X_s = 0|\boldsymbol{M}).\end{aligned}$$

If $h(W_s, M_s, X_s) = X_s W_s$, this is:

$$\mathrm{E}(W_{\lambda s}|\boldsymbol{M}) = W_s \mathrm{E}(X_s|\boldsymbol{M} = \boldsymbol{m}) = W_s \mathrm{P}(X_s = 1|\boldsymbol{M}).$$

Unlike most thresholding methods, this is not a binary procedure: using the posterior probability leads to a more continuous approach.

## 6.2 Stochastic sampling

The computation of $\mathrm{P}(X_s = 1|\boldsymbol{W})$ involves the probability of all possible configurations $\boldsymbol{x}$. Because of the enormous number of configurations, this is an intractable task. The sum we have to compute is of the following form:

$$\mu_s = \sum_{\boldsymbol{x}} f_s(\boldsymbol{x}) \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m})$$

where in this case $f_s(\boldsymbol{x}) = x_s$ and $\mu_s = \mathrm{P}(X_s = 1|\boldsymbol{M} = \boldsymbol{m})$.

To estimate this type of sum (or integral for random variables on a continuous line), one typically uses stochastic samplers. These methods generate subsequent samples $\boldsymbol{X}^{(i)}$, not selected uniformly, but in proportion to their probability. This allows to approximate the matrix of required marginal probabilities by the mean value of the generated masks:

$$\hat{\mu}_s = \sum_i X_s^{(i)}.$$

Mostly, the samples are generated, not independently of each other, but in a chain, hence the name Markov Chain Monte Carlo (MCMC) estimation. The next sample is generated, starting from the previous one. One advantage of this procedure is that knowledge of the *relative* probabilities of the candidates is sufficient. The probability ratio of two subsequent samples:

$$r^{(i)} = \frac{\mathrm{P}(\boldsymbol{X}^{(i+1)}|\boldsymbol{M})}{\mathrm{P}(\boldsymbol{X}^{(i)}|\boldsymbol{M})}$$

is the only quantity needed by the algorithm, and if

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M}) = \frac{1}{Z f_{\mathbf{M}}(\mathbf{m})} \exp[-H(\boldsymbol{x})] f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x}),$$

there is no need for the enormous computation of the partition function $Z f_{\mathbf{M}}(\mathbf{m})$.

We use the classic Metropolis MCMC sampler [28]. The chain of states is started from an initial state $\boldsymbol{X}^{(0)}$. The successive samples $\boldsymbol{X}^{(i)}$ are then produced as follows: a candidate intermediate state is generated by a local random perturbation of the actual state. Then the probability ratio $r$ of the actual state and its perturbation is computed. Since the Gibbs distribution is based on local potential functions, only positions $s$ whose mask labels are switched by the perturbation or which have a switched label in their neighborhood $\partial s$ are involved in the computation. If the candidate has a higher probability than the actual state, i.e. if the probability ratio is larger than one, then the new state is accepted, otherwise it is accepted with probability equal to $r$. To generate a completely new sample, we repeat this local switching for all locations in the grid.

# 7 Parameter estimation

## 7.1 Parameters of the conditional model

The conditional model $f_{V_s|X_s}(v|1)$ or $f_{W_s|X_s}(w|1)$ is for instance uniform or exponential. This model contains a hyperparameter. It is not so hard to fill in this parameter using the observed, noisy wavelet coefficients. In our approach, we mostly use the uniform model on $[\sigma, \mu]$ for which it is easy to prove that the expected highest magnitude $\mathrm{E}|V|_{\max}$ equals:

$$\mathrm{E}|V|_{\max} = \frac{N\mu + \sigma}{N+1}.$$

A good measure for the noise variance is the average energy removed by the minimum MSE-threshold:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} (W_{\lambda i} - W_i)^2.$$

Since the influence of the noise on the largest coefficients is relatively small, we take:

$$\hat{\mu} = \frac{(N+1)|W|_{\max} - \hat{\sigma}}{N}.$$

## 7.2 Full Bayes or empirical Bayes

The prior energy model contains a parameter $\tau$:

$$H(\boldsymbol{x}) = \tau \sum_{C \in \mathcal{C}} U_C(\boldsymbol{x}_C)$$

It determines the local rigidity of the prior. The higher its value, the larger the energy difference between the two states of a given pixel in the label image. The choice $\tau = 0$, for instance, disregards spatial structures.

To find a good value for this parameter, there exists at least two approaches. The *fully* Bayesian approach considers this parameter as an instance of still another density and assigns a prior distribution $f_\tau$ to $\tau$. The posterior density for this parameter is then:

$$f_{\tau|\mathbf{X}}(\tau|\boldsymbol{x}) \propto f_\tau(\tau)\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\tau).$$

The posterior probability of the full set of unknowns $\boldsymbol{X}$, and $\tau$, given the observation $\boldsymbol{M} = \boldsymbol{m}$ then satisfies:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}, \tau|\boldsymbol{M} = \boldsymbol{m}) \propto f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}|\tau)f_\tau(\tau).$$

From these expressions, we can — for instance — find values for $\boldsymbol{M}$ and $\tau$ with maximum posterior probability.

The *empirical* Bayes approach maximizes the likelihood of the actual label image $\boldsymbol{x}$:

$$\mathrm{L}(\tau) = \mathrm{P}(\boldsymbol{X} = \boldsymbol{x}),$$

where the probability function on the right hand side depends on the rigidity $\tau$ through the energy function $H(\boldsymbol{x})$. This method has two practical problems. First, the computation of the likelihood function is extremely hard, due to the intractable partition function in (2). Therefore, and since the rigidity parameter controls the local behavior of the label image, we use a pseudo-likelihood method: we maximize the product of "local" likelihood functions:

$$\mathrm{PL}(\tau) = \prod_{s \in S} P(X_s = x_s|\boldsymbol{X}_{\partial s} = \boldsymbol{x}_{\partial s}, \tau).$$

This leaves us with the second problem: we have no real instance of the probability function $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$, because we only have *noisy* measurements $\boldsymbol{M}$. The probability function of $\boldsymbol{X}$ supposes that $\boldsymbol{X}$ is the optimal selection of wavelet coefficients. This selection is based on the uncorrupted values $\boldsymbol{V}$ being above or below $\sigma$, which we do not know. Nevertheless, we assume that the *local* behavior of the mask obtained by thresholding the *noisy* coefficients approaches the rigidity of the optimal selection. The choice of the threshold is of course crucial in this approximation: we cannot take $\lambda = \sigma$, pretending $\boldsymbol{W} \approx \boldsymbol{V}$, since this would generate highly noisy masks, with little structure from the optimal selection. A mask generated by the minimum MSE or GCV is generally still too noisy, as becomes clear from a comparison of the labels in Figure 7 with the ideal one in Figure 8(b). This can be helped by applying a median filter to the minimum GCV labels, as in Figure 9. As mentioned before, this median filter does not take into account the background of the individual labels, like the conditional density in a Bayesian approach. Therefore it is less appropriate for the actual correction of the label images, but it may do a good job in estimating the rigidity factor $\tau$ of the optimal selection mask on a local basis. Another possibility is the universal threshold: this threshold eliminates all noise with high probability, at the risk of loosing parts of the underlying structure.

# 8 The algorithm and its results

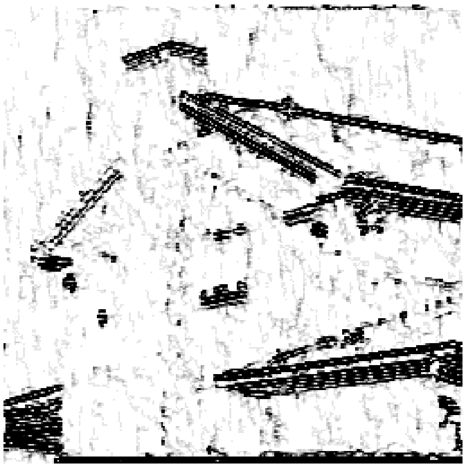
## 8.1 Algorithm overview

This is a schematic overview of the subsequent steps of the algorithm:

1. Compute the stationary wavelet transform $\boldsymbol{W}$ of the input.
2. At each level and for each component, select the appropriate threshold. This threshold generates an initial label image $\boldsymbol{X}^{(0)}$.
3. Apply a median filter to $\boldsymbol{X}^{(0)}$ and estimate the prior parameter $\alpha$ from the result, using a maximum pseudo-likelihood estimator.
4. Run a stochastic sampler to estimate for each coefficient at the given resolution level the probability $P(X_s|\boldsymbol{W})$. Use $\boldsymbol{X}^{(0)}$ from the previous step as the starting sample. A Markov Chain Monte Carlo algorithm produces the sequence of samples.
5. $\hat{W}_{\lambda s} \leftarrow W_s P(X_s = 1|\boldsymbol{W})$.
6. Inverse wavelet transform yields the result.

## 8.2 Results and discussion

We now apply the procedure to the image with artificial noise in Figure 1. Figure 12(a) shows the mask image after ten MCMC-iterations. To be more correct: this image represents for each coefficient the posterior probability $P(X_s = 1|\boldsymbol{W})$ of its label being one. More iterations (up to 100) did not improve the output quality. This output appears in Figure 12(b). Signal-to-noise ratio is 18.50 dB. Looking at the posterior probabilities, and comparing



(a)

(b)

Figure 12: Left: label image for the wavelet coefficients of the image in Figure 1 after ten MCMC iterations. Consequently, this image has 11 grey values. A pixel value is an estimate of the marginal posterior probability $P(X_s = 1|\boldsymbol{M})$. Right: the algorithm output. Three resolution levels were processed. Signal-to-noise ratio is 18.50 dB.

this with the objective mask in Figure 8(b), we see that most spurious labels from the threshold procedure indeed have a low posterior probability. The important structures that are present in the label image corresponding to the MSE-threshold, are preserved: the coefficients belonging to these structures have high probabilities. However, it seems to be hard to recover clusters of small coefficients, even if these structures appear in the optimal selection.

We also illustrate the method with the 'realistic" MRI-image of a knee in Figure 13(a). This image has 128 by 128 pixels. Figure 13(b) has the output of the Bayesian algorithm, applied to the first and second resolution level. Figure 13(c) shows the label image for the vertical subband at the second resolution level, to be compared with the selection of a minimum GCV-threshold, depicted in Figure 13(d). The latter selection is based on local regularity (magnitude) and shows far less geometrical structure.

### 8.3 Related methods

Our prior model was designed to describe geometrical correlations among coefficients within a given subband (scale and component). This type of correlation typically appears in two-dimensional wavelet transforms, especially in image analysis. Interscale correlations, present in all dimensions, are not captured by our prior model, although this is possible, as in [12].

Another difference is the meaning of the label values $X_s$, and, conse-

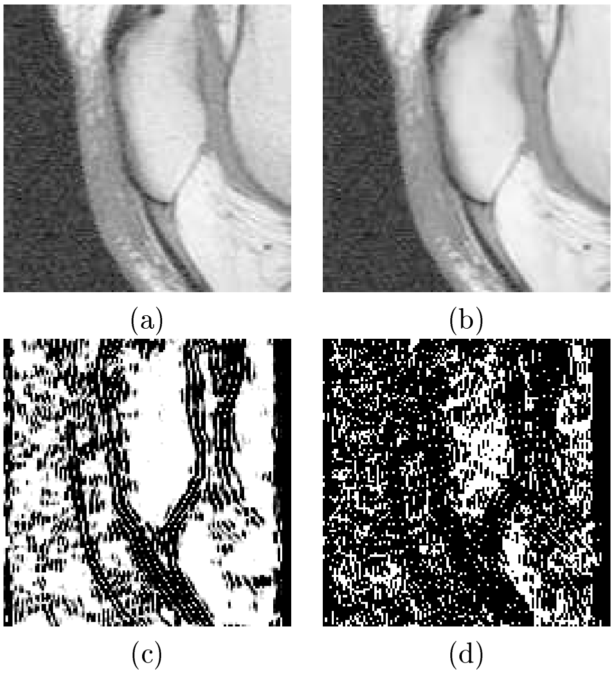


Figure 13: An example wit 'realistic' (no artificial) noise: (a) The input image, an MRI image ($128 \times 128$ pixels) with noise. (b) Output of the Bayesian algorithm, applied to the first and second resolution level of the image in (a). (c) and (d): Selection masks for vertical subband at the one but finest resolution level. The image in (c) has eleven grey levels, it represents for each coefficient an MCMC-estimate of the posterior probabilities of being important. The MCMC procedure used ten iterations, hence eleven grey levels, from zero to one. The last image is binary: black pixels correspond to coefficients that are preserved by a minimum GCV-threshold. This selection is based on local regularity (magnitude) and shows far less geometrical structure.

quently the design of the conditional model. Unlike the labels in [9, 12], a label one in our algorithm means that the corresponding noise-free coefficient is certainly larger than $\sigma$. The conditional model is explicitly inspired by the idea of finding the optimal diagonal projection of [14]. We do not compute a posterior mean $\mathrm{E}(V_s|\boldsymbol{W})$, but rather a posterior expected action: $\mathrm{E}(W_{\lambda s}|\boldsymbol{W})$.

This algorithm was inspired by previous work by Malfait et al. [26, 25], although their algorithm is based on Hölder regularity, and therefore looks at the evolution of coefficients through scales. Our algorithm uses coefficient magnitudes at one scale only, because this leads to more stable computations. Second, unlike the work by Malfait et al. the algorithm described in this text aims at the optimal coefficient selection, and the conditional model has been designed with this objective in mind. Third, all model parameters in our algorithm are determined automatically, in an empirical or heuristical way: there is no need for learning, the algorithm adapts itself to a given image.

## 9 Summary and conclusions

This paper has investigated the possibilities of a Bayesian procedure to improve the results of a wavelet thresholding procedure. This procedure was designed for application in image noise reduction and it combines two objectives:

1. We want to capture the correlations in wavelet coefficients due to edge singularities. This type of singularities is specific for more-dimensional data, like images. The *prior* model in our procedure takes these line singularities into account: the model is based on *geometrical* properties: it favors clusters of important coefficients.

2. With the aid of this geometrical prior, we aim at mimicking the optimal coefficient selection. This is reflected in the *conditional* model.

The algorithm succeeds in finding more structure in the coefficient selection, which results in an output with better preserved edges. It would be interesting to quantify this gain in contrast. A more sophisticated conditional model, based on Laplacian distributions for uncorrupted wavelet coefficients, as well as the never ending search for good prior models are other topics for further research.

## Acknowledgement

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's

Office for Science, Technology and Culture. The scientific responsibility rests with its authors. The first author is financed by a grant from the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT).

## References

[1] F. Abramovich, F. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 58, 1997.

[2] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using the wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220, 1992.

[3] R. Baraniuk. Optimal tree approximation with wavelets. In M. A. Unser, A. Aldroubi, and Laine A. F., editors, *Wavelet Applications in Signal and Image Processing VII*, volume 3813 of *SPIE Proceedings*, pages 206–214, July 1999.

[4] M. G. Bello. A combined Markov Random Field and wave-packet transform-based approach for image segmentation. *IEEE Transactions on Image Processing*, 3(6):834–846, 1994.

[5] J. E. Besag. Spatial interaction and the spatial analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[6] C. A. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, 1994.

[7] E. Candes. *Ridgelets: theory and applications.* PhD thesis, Department of Statistics, Stanford University, August 1998.

[8] P. Charbonnier, L. Blanc-Féraud, and M. Barlaud. Noisy image restoration using multiresolution Markov Random Fields. *Journal of Visual Communication and Image Representation*, 3(4):338–346, 1992.

[9] H. Chipman, E. Kolaczyk, and R. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.*, 92:1413–1421, 1997.

[10] M. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrica*, 85:391–401, 1998.

[11] A. Cohen, I. Daubechies, and J. Feauveau. Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560, 1992.

[12] M. S. Crouse, R.D. Nowak, and R. G. Baraniuk. Wavelet-based signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 46, Special Issue on Wavelets and Filterbanks:886–902, 1998.

[13] I. Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Regional Conf. Series in Appl. Math., Vol. 61. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[14] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[15] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

[16] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[17] J. Heikkinen and H. Högmander. Fully Bayesian approach to image restoration with an application in biogeography. *Applied Statistics*, 43(4):569–582, 1994.

[18] M. Jansen and A. Bultheel. Geometrical priors for noisefree wavelet coefficient configurations in image de-noising. In P. Müller and B. Vidakovic, editors, *Bayesian inference in wavelet based models*, pages 223–242. Springer-Verlag, 1999.

[19] M. Jansen and A. Bultheel. Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise. *IEEE Transactions on Image Processing*, 8(7):947–953, July 1999.

[20] M. Jansen, M. Malfait, and A. Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44, January 1997.

[21] B. Jawerth and W. Sweldens. An overview of wavelet based multiresolution analyses. *SIAM Review*, 36(3):377–412, 1994.

[22] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, 59:319–351, 1997.

[23] M. R. Luettgen, W. C. Karl, A. S. Willsky, and R. R. Tenney. Multiscale representations of Markov Random Fields. *IEEE Transactions on Signal Processing*, 41(12):3377–3395, December 1993.

[24] M. Malfait. *Stochastic Sampling and Wavelets for Bayesian Image Analysis.* PhD thesis, Department of Computer Science, K.U.Leuven, Belgium, 1995.

[25] M. Malfait. Using wavelets to suppress noise in biomedical images. In A. Aldroubi and M. Unser, editors, *Wavelets in Medicine and Biology,* Chapter 8, pages 191–208. CRC Press, 1995.

[26] M. Malfait and D. Roose. Wavelet based image denoising using a markov random field a priori model. *IEEE Transactions on Image Processing*, 6(4):549–565, 1997.

[27] S. G. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 525 B Street, Suite 1900, San Diego, CA, 92101-4495, USA, 1998.

[28] N. Metropolis, M. Rosenbluth, et al. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[29] G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, Lecture Notes in Statistics, pages 281–299, 1995.

[30] J.-C. Pesquet, H. Krim, and H. Carfantan. Time invariant orthonormal wavelet representations. *IEEE Transactions on Signal Processing*, 44(8):1964–1970, 1996.

[31] F. Ruggeri and B. Vidakovic. A Bayesian decision theoretic approach to wavelet thresholding. Preprint 95-35, Duke University, Durham, NC, 1995.

[32] E. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In M. A. Unser, A. Aldroubi, and Laine A. F., editors, *Wavelet Applications in Signal and Image Processing VII*, volume 3813 of *SPIE Proceedings*, pages 206–214, July 1999.

[33] E. P. Simoncelli and E.H. Adelson. Noise removal via Bayesian wavelet coring. In *proceedings 3rd International Conference on Image Processing*, September 1996.

[34] G. Strang and T. Nguyen. *Wavelets and filter banks.* Wellesley-Cambridge Press, Box 812060, Wellesley MA 02181, fax 617-253-4358, 1996.

[35] N. Weyrich and G. T. Warhola. De-noising using wavelets and cross validation. In S.P. Singh, editor, *Approximation Theory, Wavelets and Applications*, volume 454 of *NATO ASI Series C*, pages 523–532, 1995.

[36] G. Winkler. *Image analysis, random fields and dynamic Monte Carlo methods.* Applications of Mathematics. Springer, 1995.

[37] Y. Xu, J. B. Weaver, D. M. Healy, and J. Lu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3(6):747–758, 1994.