WAVELET THRESHOLDING FOR NON (NECESSARILY) GAUSSIAN NOISE: IDEALISM*†

R. Averkamp‡ and C. Houdré§

December 4, 1999

Abstract

For various types of noise (exponential, normal mixture, heavy tailed, compactly supported, ...) wavelet tresholding methods are studied. Problems linked to the existence of optimal thresholds are tackled, and minimaxity properties of the methods also analyzed. A coefficient dependent method for choosing thresholds is also presented.

1 Introduction

An underlying idea in non-parametric curve/surface/signal estimation is that the function to estimate has some redundancy; and this is often reflected by assuming that it belongs to a particular class. To name but a few, the class could be a bounded subset of a Sobolev space, or it could be the piecewise constant functions with a limited number of discontinuities. In other words, the prior assumption is that the function is "nice". Being nice does not mean smooth like infinitely differentiable, but rather that there is not too much

*The results presented here are part of the first author's doctoral dissertation defended at the University of Freiburg, Germany.

†Key words: Wavelets, Thresholding, Minimax.

AMS Classification: Primary 62G07, 62C20. Secondary 60G70, 41A25.

‡Research supported in part by the NSF Grant DMS-9632032, while this author was visiting the Southeast Applied Analysis Center, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA.

§Research supported in part by the NSF Grant DMS-9803239.

information in this curve. It could be discontinuous but only at a limited number of places; and in this sense $\sin(1000x)$ is less nice than $\mathrm{sgn}(x)$ (here and in the sequel $\mathrm{sgn}(x) := x/|x|$ for $x \neq 0$ and $\mathrm{sgn}(0) = 0$). Moreover, often the function to estimate is assumed to only have one mode or to be monotone. Of course, there are many exceptions to that type of assumption, for example, sound recordings or images. An image might consists of a smooth blue sky, not so smooth clouds and rather unsmooth trees. The redundancies of images and sound recordings are not that easily translated into mathematical terms. To tackle this type of problem, wavelets can be extremely useful as shown, in recent years, by various authors [1], [25], [26], [27], [8], [12], [13], [11], [15], [16], [18], [23], [36], [42], ... (we refer the reader to the recent book of Vidakovic [41] for a much more precise and complete bibliography).

Indeed, the heuristic for the use of wavelets in non-parametric estimation is that the expansion of a "nice" function in a wavelet basis is sparse, i.e., only a few of the wavelet coefficients are big and the rest is small and thus negligible. Thus in order to estimate the function, one has to guess the important wavelet coefficients, estimate them and discard the rest. Assuming that the wavelet coefficients are zero or negligible for fine levels is like making the assumption of smoothness. A different point of view is that one is not interested in these fine levels anyway, because they cannot be used.

The Fourier basis is an orthonormal basis of $L^2([0,1])$ and it is possible to apply the above method to Fourier coefficients (see [19], where an adaptive linear estimator is used). It is well known that the n^{th} Fourier coefficient of a function in C^k is of size $O(1/n^k)$; but the Fourier basis expansion is rather sensitive to discontinuities, a discontinuity in the functions affects *all* Fourier coefficients. The wavelet expansion is more robust, discontinuities affect only those wavelet coefficients $c_{j,k}$ where the discontinuity is in the support of $\psi_{j,k}$ (throughout if ψ is a function, then $\psi_{j,k}(\cdot) = 2^{j/2}\psi(2^j \cdot -k)$, $j,k \in \mathbb{Z}$). Assuming that the support of ψ is $[0,N]$, there are only N coefficients affected at each level. So if we consider a wavelet basis adapted to an interval only about $N \log_2 n$ of the first n coefficients are affected by a single discontinuity. Figure 1 shows this clearly.

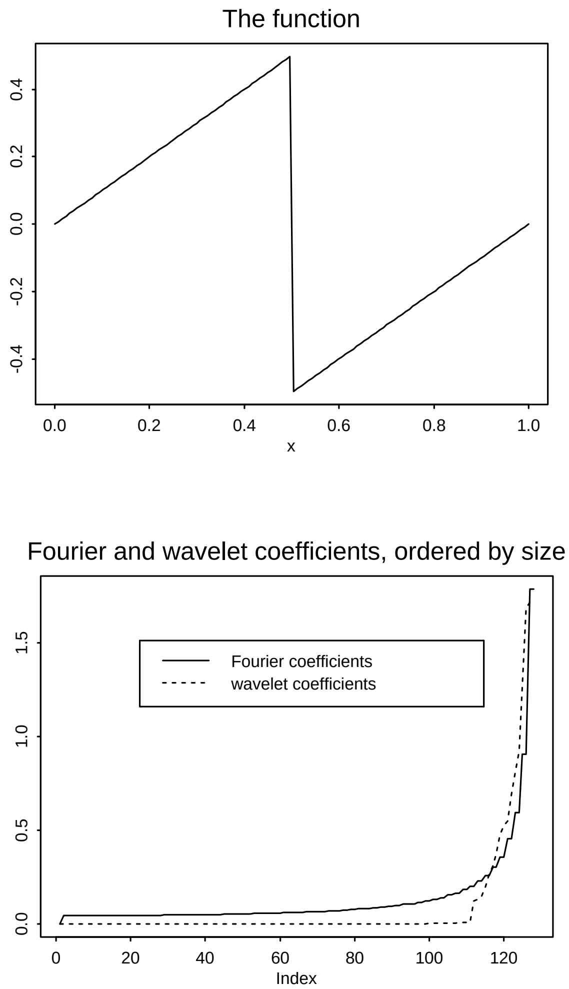


Figure 1: Comparison of the size of the absolute values of Fourier and wavelet coefficients

Let us recall a few essentials about wavelets and refer the reader to the books of Daubechies [10] and Meyer [34], [35] for a much more complete introduction. Using a multiresolution approach, there is a large family of wavelets with compact support, but there is also a trade off between the regularity, i.e. smoothness of these wavelets and the size of their support. The Haar wavelet has the smallest possible support, but it is not continuous. Since these wavelets have compact support there are only finitely many non-zero coefficients h_k in the scaling identity. Also the support of ψ and ϕ are intervals of integer lengths. These properties of compactly supported wavelets are at the root of a very efficient analog of the discrete Fourier transform. Let ϕ and ψ be the compactly supported scaling function and wavelet of a multiresolution analysis. The support of ϕ and ψ is $[0, N-1]$ and the length of the filter (h_i) is N. Then

$$\phi_{j,k} = \sum_{i=0}^{N-1} h_i \phi_{j+1,2k+i},$$

and

$$\psi_{j,k} = \sum_{i=0}^{N-1} (-1)^i h_{N-1-i} \phi_{j+1,2k+i} \ .$$

Now let f be a function with support $[0,1]$, and denote the scaling coefficient $\langle \phi_{j,k}, f\rangle$ by $c_{j,k}$ and the wavelet coefficients $\langle \psi_{j,k}, f\rangle$ by $d_{j,k}$. Assume that we are given the coefficients $c_{j_0,k}$ for some $j_0 > 0$, i.e., that we are given the projection of f onto V_{j_0}. If f is continuous one can view the $c_{j_0,k}$ as an approximation of $\sqrt{2^{-j_0}}$ times the values of f at the places $k/2^{j_0}$, since the integral of ϕ is one and the support of $\phi_{j,k}$ is $[k/2^{j_0}, (k+N-1)/2^{j_0}]$. Another way we might get the $c_{j_0,k}$ is to sample f at the places $k/2^{j_0}$ and to compute the linear combination of the $\phi_{j_0,k}$ which interpolates the samples. Starting with the level j_0 means that we are not interested in details of f with a spread in time smaller than $O(1/2^{j_0})$. Since f has support $[0,1]$, only those $c_{j_0,k}$ with $-N+2 \leq k \leq 2^{j_0} - 1$ are non-zero. Because of the scaling identities it is now easy to compute the $c_{j_0-1,k}$ and $d_{j_0-1,k}$:

$$c_{j_0-1,k} = \sum_{i=0}^{N-1} h_i c_{j_0,2k+i} \text{ and } d_{j_0-1,k} = \sum_{i=0}^{N-1} (-1)^i h_{N-1-i} c_{j_0,2k+i}.$$

Let $n = 2^{j_0} + N$ be the number of $c_{j_0,k}$ which are non-zero. The computation of the coefficients for the level $j_0 - 1$ takes $O(Nn)$ time and again only

$2^{j_0-1} + N \approx n/2$ of the $c_{j_0-1,k}$ and $d_{j_0-1,k}$ are non-zero. This method can be repeated, and at the level j only $2^j + N \approx n/2^{j_0-j}$ coefficients are non-zero. This is done until reaching the level 0. To compute each coefficient only $O(N)$ time is needed. Thus to compute all the coefficients $c_{j,k}$ and $d_{j,k}$ $0 \le j \le j_0 - 1$, $-N+2 \le k \le 2^j - 1$ we need $O(2^{j_0}) = O(n)$ time. This transformation is called the fast wavelet transform, see Mallat [32] where it is called the cascade algorithm there.

Note that we do not need to compute the coefficients for j_0 levels, we can stop earlier at some level j_1. Then we have a decomposition of f into $W_{j_0-1}, \ldots, W_{j_1}$ and V_{j_1}. Often a further decomposition of V_{j_1} is not needed. For example a simple smoothing operation is to project f onto V_{j_1}. When using wavelets for compression, most of the information of a function is in its projection on V_{j_1} and so it is not likely that it is easy to compress the coefficients $d_{j,k}$, $j \le j_1$.

The inverse transformation, i.e. computing the $c_{j_0,k}$ from the $d_{j,k}$ and $c_{0,k}$ is also simple, the problem is solved if we can compute each $c_{j,k}$ from the $d_{j-1,k}$ and $c_{j-1,k}$. Using the orthonormality properties of the wavelet and the scaling function it is easy to show that:

$$\phi_{j,k} = \sum_{\ell \text{ even}} h_\ell \phi_{j-1,(k-\ell)/2} + \sum_{\ell \text{ even}} h_{N-1-\ell} (-1)^\ell \psi_{j-1,(k-\ell)/2}, \quad k \text{ even},$$

and

$$\phi_{j,k} = \sum_{\ell \text{ odd}} h_\ell \phi_{j-1,(k-\ell)/2} + \sum_{\ell \text{ odd}} h_{N-1-\ell} (-1)^\ell \psi_{j-1,(k-\ell)/2}, \quad k \text{ odd}.$$

Thus computing the inverse of the fast wavelet transform takes again $O(n)$ time. Note that

$$\phi_{j,k} = \sum_{i=0}^{2^{j_0-j}(N-1)} u_{j_0-j,i+2^{j_0-j}k} \phi_{j_0,i}, \tag{1.1}$$

and

$$\psi_{j,k} = \sum_{i=0}^{2^{j_0-j}(N-1)} v_{j_0-j,i+2^{j_0-j}k} \phi_{j_0,i}, \tag{1.2}$$

where $u_{\cdot,\cdot}$ and $v_{\cdot,\cdot}$ only depend on the filter (h). This claim about the length of the filters (u_{j_0-j}) and (u_{j_0-j}) can be proved via a simple induction argument. Later we will see that $\max_i |u_{j_0-j,i}| = O(2^{(j_0-j)/2})$ and $\max_i |v_{j_0-j,i}| = O(2^{(j_0-j)/2})$.

Also note that the fast wavelet transform is actually a repeated discrete filtering method. This method is known in signal theory as subband coding. If the first $k+1$ moments (orders 0 to k) of a wavelet with compact support of a multiresolution analysis vanish, then there are coefficients $(a_{m,i})$, $i \in \mathbb{Z}$, $0 \leq m \leq k$, such that $x^m = \sum_i a_{m,i}\phi(x-i)$, where the convergence is understood pointwise.

From now on we use an orthonormal wavelet basis from a multiresolution analysis adapted to an interval.

Non-parametric estimation via wavelet methods is then usually divided in two parts. The first part transforms the data into something which can be inputed into the fast wavelet transform, i.e. noisy versions (denoted by $\widetilde{c}_{j_0,k}$) of the scaling coefficients $c_{j_0,k}$, with j_0 large. Then the fast wavelet transform is applied to this data, and gives noisy versions of the the wavelet coefficients $d_{j,k}$ (denoted by $\widetilde{d}_{j,k}$ and called the empirical wavelet coefficients). In the second part, estimates $\widehat{d_{j,k}}$ of the $d_{j,k}$ are computed using the $\widetilde{d}_{j,k}$ and using the heuristic that the wavelet transform of the signal is sparse and that the noise is uniformly distributed over the empirical wavelet coefficients. From these estimates, estimation of the original function can easily be obtained.

How to compute these "noisy" versions $\widetilde{c}_{j_0,k}$ of the $c_{j_0,k}$? In a density estimation framework, one has n iid random variables $X_1, \ldots, X_n$, and needs to estimate the Lebesgue density f of X_1. The $\widetilde{c}_{j,k}$, respectively the empirical wavelet coefficients are given by:

$$\widetilde{c}_{j,k} = n^{-1}\sum_i \phi_{j,k}(X_i) \text{ and } \widetilde{d}_{j,k} = n^{-1}\sum_i \psi_{j,k}(X_i).$$

Note that $E\phi_{j,k}(X_i) = \int_{\mathbb{R}} f(x)\phi_{j,k}(x)dx = \langle f, \phi_{j,k}\rangle = c_{j,k}$, thus $\widetilde{c}_{j,k} = c_{j,k} + e_{j,k}$ with $Ee_{j,k} = 0$ and $Var(e_{j,k}) = O(1/n)$. Recall also that because of the fast wavelet transform we only have to compute the $\widetilde{c}_{j_0,k}$ via these formulas.

In a regression framework, we observe data $Y_i = f(X_i) + e_i$, where the e_i are iid random variables representing the observation error and the X_i are either iid random variables, independent of the e_i, or are deterministic design points, the simplest case being when the design points are equidistant. There is an approach developed for this case in [12]; which however has the drawback that the joint distribution of the $\tilde{c}_{j_0,k}$ is complicated, making it hard to analyze properties of the empirical wavelet coefficients.

In a more complex situation, one is given n pairs of random variables (X_i, Y_i) and wants to estimate $E(Y|X)$. In particular, given $(f(x_i)+e_i, x_i)$, we might want to estimate f where the (x_i, e_i) are iid random variables with $E(e_i|x_i = \cdot) = 0$, and where the corresponding conditional probabilities $P^{e_i|x_i=t}$ might heavily depend on t. In general it is impossible to compute the distribution of the noise in the wavelet coefficients. In the above problems, a simpler framework which can be seen as a first order approximation is to assume that the $\tilde{c}_{j_0,k}$ are given and of the form $\tilde{c}_{j_0,k} = c_{j_0,k} + e_k$, where the e_k are iid random variables (thus assuming independent noise!). Although this might appear rather naive, it is close to the equidistant design case, where $Y_k = f(x_k) + e_k$ is given. Indeed, assuming for simplicity that f is defined on $[0,1]$ and that $x_k = 2^{-j_0}k$, $k = 1, \ldots, n = 2^{j_0}$, we have

$$c_{j_0,k} = 2^{j_0/2} \int_{\mathbb{R}} \phi(2^{j_0}x - k) f(x) dx \approx f(2^{-j_0}k) 2^{-j_0/2} = f(x_k) 2^{-j_0/2}, \quad (1.3)$$

since for the scaling function ϕ, we have $\int_{\mathbb{R}} \phi(x)dx = 1$, and if f is continuous at x_k. This approximation is even better if we use a wavelet basis with a scaling function ϕ with vanishing moments of order 1 to k (except order 0), i.e., using coifflets ([10, p.258]).

A word on notation: throughout the text, $a \approx b$ is used to mean "about the same", i.e., a/b is approximately 1 and $a_n \sim b_n$ means that a_n/b_n tends to 1.

Another way to view the regular design situation is to assume, since the $f(x_k)$ are close to the $c_{j_0,k}$, that the discrete wavelet transform of the $f(x_k)$ has the same properties as the true wavelet coefficients of f, i.e., that it is sparse. The continuous function f is then estimated by interpolating the estimates of the $f(x_k)$. Thus estimation and interpolation are separated. The errors incurred by these approximations and discretizations are negligible in comparison to the "true" estimation errors. This method works well in practice and most other methods for equidistant designs are not really needed.

Assume now that we obtained (by some preprocessing) noisy observations of the $c_{j_0,k}$ $(=: f_k)$

$$\tilde{c}_{j_0,k} = f_k + e_k, \quad k = 1, \ldots, 2^m = n,$$

where the e_k are iid random variables which represent the noise or observation error. Applying a fast wavelet transform W_n to the data $(\tilde{c}_{j_0,k})$ gives:

$$\widetilde{w}_{j,k} = (W_n(f))_{j,k} + (W_n(e))_{j,k} =: w_{j,k} + z_{j,k}$$

For the transformed data, we write $w_{\cdot,\cdot}$ rather than $d_{\cdot,\cdot}$ since the coefficients do not exactly correspond to the wavelet coefficients; the reason being that we have to take care of the boundary effects and also do not compute the whole "triangle" of coefficients but stop at some level (and replace the top $d_{j,k}$ by the $c_{j_{top},k}$).

The mapping W_n being orthonormal, if the e_k are uncorrelated, so are the $z_{j,k}$; and if the e_i are iid $N(0,\sigma^2)$ random variables so are the $z_{j,k}$! This fact greatly simplifies the normal iid case. Often one tries to reduce the estimation with other kinds of noise to the case of Gaussian noise by some asymptotic reasoning; or one just postulates Gaussian noise.

Another consequence of the orthonormality of W_n is the fact that norms are preserved: $\|W_n x\|_2^2 = \|x\|_2^2$, for $x \in \mathbb{R}^n$. Thus if $\widehat{w}_{j,k}$ is an estimator for $w_{j,k}$ and $\widehat{f}$ the equivalent estimator for f then

$$\|f - \widehat{f}\|_2^2 = \sum_{j,k} |\widehat{w}_{j,k} - w_{j,k}|^2.$$

Since the wavelet transform of a "nice" function is sparse, we expect only a small fraction of the wavelet coefficients to be big and the rest to be small and thus negligible. So if a $\widetilde{w}_{j,k}$ is small, it is reasonable to regard it as mostly noise and to set $w_{j,k}$ to zero; if it is big, it is reasonable to keep it. This is known as hard thresholding. Soft thresholding shrinks everything towards zero by a certain amount, thus reducing the variance of the estimation at the cost of a higher bias. Nonlinear shrinking policies were first applied to wavelet coefficients by Donoho and Johnstone [15], [16], [17], [14] and in a function space framework by DeVore and Lucier [11]. Hard and soft thresholding correspond to respectively applying the operators

$$T_\lambda^H(x) = x\mathbf{1}_{\{|x|>\lambda\}},$$

and

$$T_\lambda^S(x) = (|x| - \lambda)_+ \mathrm{sgn}(x),$$

to the noisy wavelet coefficients. In both cases λ is called the threshold and it usually depends on the index (j,k). There is by now a variety of papers on these rules, especially on minimax estimation, see [8], [25], [26], [27], [42], ... Although called differently, soft and hard thresholding had already appeared before, see for example Bickel [7]; and both procedures seem to originate in a paper of Efron and Morris [20]. Other shrinking schemes have also been proposed (see e.g., [5], [6]) but, all in all, the asymptotic performances of the different shrinking estimators do not vary much. For normal iid noise, soft and hard thresholding are not admissible (see [5] for a discussion of other shrinking schemes). This is not a surprise if one considers the variety of possible shrinking estimators.

Figures 2, 3, 4 show the soft, hard thresholding estimator and

$$T_\lambda^M(x) := x\mathbf{1}_{\{|x|\geq\lambda\}} + 2(|x|-\lambda/2)_+\mathrm{sgn(x)}\mathbf{1}_{\{|x|<\lambda\}},$$

with their respective risk functions for Gaussian noise. T_λ^M is a special case of the semisoft thresholding estimator of Bruce and Gao [6]. The parameter λ for each estimator was chosen such that for each estimator the risk, $\int_{\mathbb{R}}(T_\lambda^\cdot(x+a)-a)^2\Phi(dx)$, at $a=0$, is about 0.0004 where Φ is the standard normal distribution. The parameters are $\lambda_S=3$, $\lambda_H=4.1$ and $\lambda_M=3.3$. T_λ^M has the advantage of having a smaller bias for large values than soft thresholding and of being continuous, unlike the hard thresholding estimator.

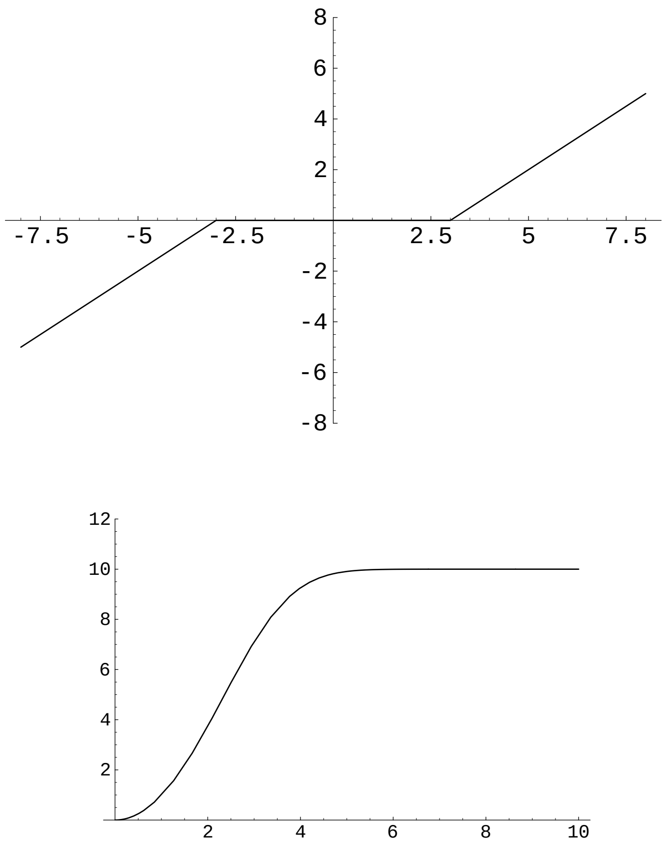


Figure 2: T_λ^S and its risk function for $\lambda = 3$

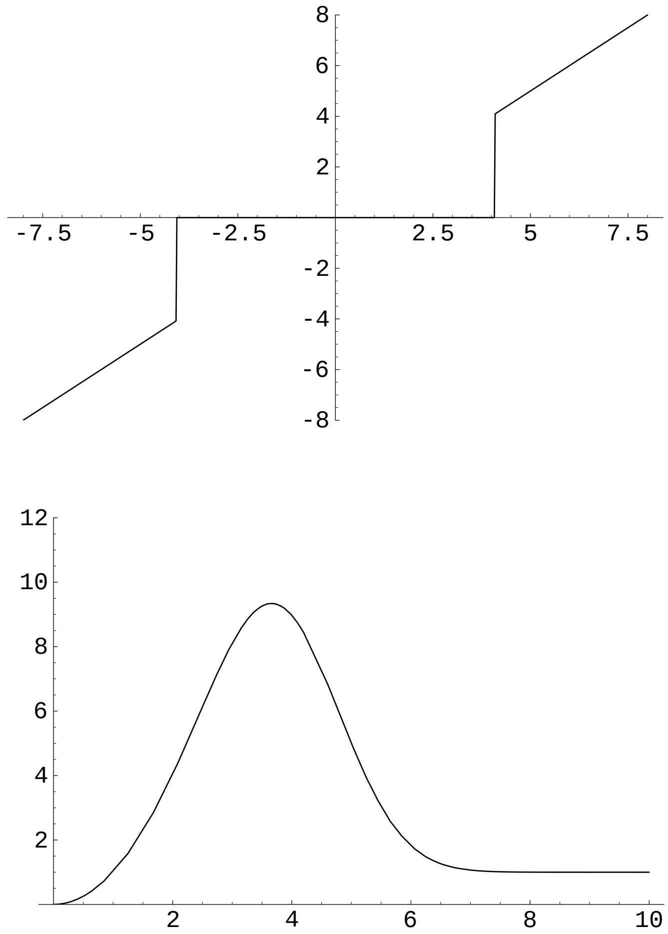


Figure 3: T_λ^H and its risk function for $\lambda = 4.1$

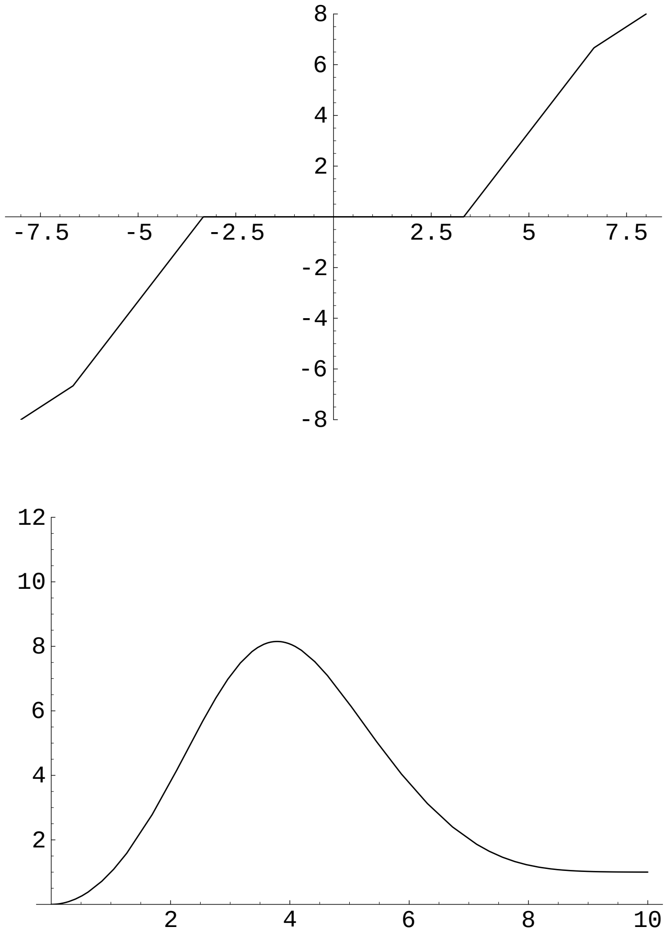


Figure 4: T_λ^M and its risk function for $\lambda = 3.3$

More important than the type of shrinking procedure is the amount of shrinking applied to each empirical wavelet coefficient. The thresholds could be the same for all the coefficients or could depend on the level. Further the thresholds depend on the expected sparseness of the wavelet coefficients. In practice the choice of thresholds should be data driven.

Clearly the thresholds should depend on the type of noise and on the variance of the noise in the initial data. For Gaussian noise we do know the exact distribution of the noise in the wavelet coefficients; it is iid Gaussian. If we use a transform that is not orthogonal, for example a fast wavelet transform based on biorthogonal wavelets, then the multivariate distribution of the noise is uniquely determined by its covariance matrix, and it is straightforward to compute this matrix. In fact, the fast wavelet transform is only a filtering scheme, so one can use other filtering schemes which correspond not only to biorthogonal wavelet bases, but also to wavelet frames (or the "lifting approach" see [39]). Of course, the noise in the coefficients is no longer uncorrelated, but for filtering schemes it is quite easy to keep track of the correlations. (See [30] for a fast method to compute the covariance matrix.) More generally, the core of wavelet thresholding is to apply a linear transformation to discrete data = signal + noise. Since we expect the signal to be concentrated in a few coefficients, it is easier to estimate the transformed signal and then to transform back. So one actually only needs invertible linear transformations which transform "nice" signals into sparse signals.

In real life the assumption of Gaussian noise is not often given. For large datasets one can rely on the central limit theorem and one gets the same asymptotic results as in the Gaussian case, but this will not do for small datasets, (see [22], for example). Matters get worse if the requirement of independence in the initial noise is dropped, often the initial data to which the wavelet transform is applied is already the result of preprocessing (when dealing with irregular spaced design, random design, density estimation), then the noise is neither identical distributed nor independent. In particular, for heavy tailed noise the thresholds are sometimes too small. So far only few results deal directly with non-Gaussian noise (see [22], [28], [29]). It is the purpose of the present work to help understand the non normal framework. To briefly describe it, in the next section we recall the "ideal" denoising method of Donoho and Jonstone and study it for certain classes of noise with tails heavier than normal. Minimax type results are obtained in Section 3. In Section 4, compactly supported noises are tackled. Section 5 deals with noise having very smooth density. In the last section, a different approach to choosing thresholds is studied (the threshold of the kth coefficient is always the same, no matter the signal length). Various simulations and computations are also presented. At times, our results complement the Gaussian framework. Many of the results presented here have been previously announced in [2] and were presented at the 1996–1997 wavelet special year in Montréal. A companion paper [3] studies the function space approach to denoising in a non (necessarily) normal framework.

2 The ideal method

An approach to denoising introduced by Donoho and Johnstone ([16]) is the following method: Given noisy wavelet coefficients, i.e. the true wavelet coefficient plus a random term which represents the noise and assuming that one has knowledge of the true wavelet coefficients, an ideal estimator is to set a noisy coefficient to zero if the variance σ^2 of the noise is greater than the square of the true wavelet coefficient; otherwise the noisy coefficient is kept. Clearly, the mean square error of this estimator is the minimum of σ^2 and the square of the coefficient. Under the assumption of iid normal noise (see also [24] for normal correlated noise), these authors show that the soft thresholding estimator achieves a risk at most $O(\log n)$ times the risk of this ideal estimator. Moreover, no estimator is asymptotically better. This method tries to minimize the quotient of the mean square error and of a functional of the function itself. This functional depends on the wavelet coefficients and measures how easily the estimation can be performed. It is clear that it is more difficult to denoise a signal whose energy is uniformly distributed over the wavelet coefficients than to denoise a signal whose energy is concentrated in a few big wavelet coefficients. This ease of estimation is measured by the mean square error of the ideal estimator.

The "ideal method" does not require any a-priori knowledge on the function to be denoised, but might not be optimal if a smoothness class information is available. For many "smooth" functions "most of" the wavelet coefficients are rather small and only a small part of the wavelet coefficients is big. This means that the risk of the ideal estimator is small and this, in turn, implies that the risk of soft thresholding is small for these functions. It is clear that this method is applicable to other sparse estimation problems, not just in the wavelet area.

Assume the regular design situation, i.e. we have the observations

$$X_i = f_i + e_i, \quad i = 1, \ldots, n = 2^m.$$

To this data we apply a discrete wavelet transform $W_n : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ which is adapted to an interval by boundary corrections or by periodizing the wavelet basis, in any case W_n is an orthogonal transformation.

Let $Y = W_n(X)$ be the empirical wavelet coefficients, let $\theta = W_n(f)$ and let $z = W_n(e)$. Thus $Y_i = \theta_i + z_i$, $i = 1, \ldots, n$, and the respective mean square errors in estimating θ and f are equal. Assuming some knowledge of the true wavelet coefficients θ, consider the following estimator for θ_i: $\check{\theta}_i = Y_i$ if $\theta_i^2 \geq \sigma^2$ and $\check{\theta}_i = 0$ if $\theta_i^2 \leq \sigma^2$. In plain words, an empirical wavelet coefficient is kept if its contribution to the energy of the function is greater than the variance of the noise, otherwise it is discarded. The performance of other estimators (in particular of the soft thresholding estimator $T_\lambda^S(x) = (|x| - \lambda)_+ \mathrm{sgn}(x)$ when applied to Y) will be compared to the benchmark

$$B_n(\theta, \sigma^2) := \sigma^2 + \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2), \tag{2.1}$$

which is the mean square error of $\check{\theta}$ plus σ^2. Since the mean square error of $\check{\theta}$ is 0 if $\theta = 0$, σ^2 is added. This is close to assuming that at least one θ_i^2 is greater than the variance σ^2. Notice that $B_n(\theta, \sigma^2)$ is small in comparison to $n\sigma^2$ (the total variance in the signal X, i.e. the sum of the variances of each component) if θ (the wavelet transform) is sparse. In fact $B_n(\theta, \sigma^2)$ is itself a measure for the sparsity of θ. The additional summand σ^2 is a measure of our expectation of the sparsity of the signal. As we will see later, in the Gaussian case the additional summand σ^2 can be replaced by $\sigma^2 \log n$ and neither the threshold nor the ratio of the optimal thresholding and the benchmark do change much, the asymptotic behavior is the same. Of course changing the additional summand to tune the estimator is contrary to our faith not to make any prior assumptions about the signal.

A note is necessary: let f be a function defined on $[0, 1]$ and let W^n be a wavelet transform based on the wavelet ψ, then it follows from (1.3) that

$$W^n(f(i/n)_{i=1,\ldots,n})_{j,k} \sim \sqrt{n}\langle f, \psi_{j,k}\rangle.$$

So if in the following the thresholds are increasing with n, it is important to remember that the true wavelet coefficients are also increasing with n, whereas the variance of the noise in the coefficients remains constant.

If the e_i are iid normal random variables, then the z_i are normal and independent since W_n is an orthogonal transformation. This is the original case investigated by Donoho and Johnstone. For $Y_i = \theta_i + z_i$, $i = 1, \ldots, n$, $n \geq 3$, where the θ_i are parameters of interest and where the z_i are zero mean iid normal random variables with variance σ^2, they show that

$$\sup_{\theta \in \mathbb{R}^n} \frac{E\|T^S_{\lambda_n}(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} \leq (1 + 2\log n),$$

where $\lambda_n = \sigma\sqrt{2\log n}$; and in [22] a similar result is proved for iid random variables with exponential tails. A stronger result is actually proved in [16]:

$$\sup_{\theta \in \mathbb{R}^n} \frac{E\|T^S_{\lambda_n^*}(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} \leq \Lambda_n^*,$$

where $\Lambda_n^* := \inf_\lambda \sup_{\theta \in \mathbb{R}} \dfrac{E\|T^S_\lambda(Y) - \theta\|^2}{n^{-1} + \min(\theta^2, \sigma^2)} \leq (1 + 2\log n)$ and where $\lambda_n^* \leq \sigma\sqrt{2\log n}$ is the largest λ attaining Λ_n^*.

In fact, a result of this type continues to hold for a wider class of distributions, but it still requires the z_i to be identically distributed random variables (throughout this work, θ_i are real valued parameters of interest).

Theorem 2.1 *Let $Y_i = \theta_i + z_i$, $i = 1, \ldots, n$, $n \geq 4$, where the z_i are identically distributed symmetric random variables with law μ and such that $Ez_1^2 = \sigma^2$. Then, the equation*

$$2(n+1)\int_\lambda^\infty (x - \lambda)^2 \mu(dx) = \lambda^2 + \sigma^2, \tag{2.2}$$

has a unique positive solution λ_n. Moreover,

$$\Lambda_n := \sup_{\theta \in \mathbb{R}^n} \frac{E\|T^S_{\lambda_n}(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} = \inf_\lambda \sup_{\theta \in \mathbb{R}^n} \frac{E\|T^S_\lambda(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} = \frac{n(\lambda_n^2 + \sigma^2)}{(n+1)\sigma^2}.$$

Proof: The proof is similar to the proof of Theorem 2 in [16]. For $\lambda, a \in \mathbb{R}$, set

$$\begin{aligned} p(\lambda, a) &:= E|T^S_\lambda(z_1 + a) - a|^2 \\ &= \int_{\mathbb{R}} (\operatorname{sgn}(x + a)(|x + a| - \lambda)_+ - a)^2 \mu(dx). \end{aligned}$$

For $\lambda > 0$, set $\Lambda := \sup_{a \in \mathbb{R}} \dfrac{p(\lambda, a)}{\sigma^2/n + \min(a^2, \sigma^2)}$. Then

$$\sum_{i=1}^{n} E|T_\lambda^S(Y_i) - \theta_i|^2 \leq \Lambda \sum_{i=1}^{n} \left(\frac{\sigma^2}{n} + \min(\theta_i^2, \sigma^2) \right) = \Lambda B_n(\theta, \sigma^2).$$

The function $p(\lambda, \infty) := \lim_{a \to \infty} p(\lambda, a) = \sigma^2 + \lambda^2$ is continuous and increasing on $[0, \infty)$, whereas $p(\lambda, 0) = 2 \int_\lambda^\infty (x - \lambda)^2 \mu(dx)$ is continuous and non–increasing on $[0, \infty)$ (decreasing on the positive part of the support of μ, and zero outside the support.) Moreover, $p(0,0) = p(0, \infty) = \sigma^2$. Hence λ_n, which is the unique solution of

$$\frac{p(\lambda, 0)}{\sigma^2/n} = \frac{p(\lambda, \infty)}{\sigma^2 + \sigma^2/n},$$

minimizes $\sup_{a \in \{0, \infty\}} \dfrac{p(\lambda, a)}{\sigma^2/n + \min(a^2, \sigma^2)}$ and $\Lambda_n = p(\lambda_n, \infty)/(\sigma^2(1 + 1/n))$ is equal to the minimum of this term. We now claim that

$$\sup_{a \in \mathbb{R}} \frac{p(\lambda_n, a)}{\sigma^2/n + \min(a^2, \sigma^2)} = \sup_{a \in \{0, \infty\}} \frac{p(\lambda_n, a)}{\sigma^2/n + \min(a^2, \sigma^2)}.$$

In the following let λ be fixed, let δ_a be the Dirac measure with unit mass at a and let

$$f_{\lambda,a}(x) := (T_\lambda^S(x) - a)^2 = \begin{cases} (x - a + \lambda)^2 & : x \in (-\infty, -\lambda] \\ a^2 & : x \in (-\lambda, \lambda) \\ (x - a - \lambda)^2 & : x \in [\lambda, \infty) \end{cases}.$$

If $\lambda > h > 0$, $a > 0$ then

$$\begin{aligned} p(\lambda, a + h) &= \int_{\mathbb{R}} f_{\lambda, a+h}(x)(\mu * \delta_{a+h})(dx) \\ &= \int_{\mathbb{R}} f_{\lambda, a+h}(x + h)(\mu * \delta_a)(dx) \\ &\geq \int_{\mathbb{R}} f_{\lambda,a}(x)(\mu * \delta_a)(dx) = p(\lambda, a), \end{aligned}$$

where the inequality holds since

$$\begin{aligned} f_{\lambda, a+h}(x + h) &= \begin{cases} (x - a + \lambda)^2 & : x \in (-\infty, -\lambda - h] \\ (a + h)^2 & : x \in (-\lambda - h, \lambda - h) \\ (x - a - \lambda)^2 & : x \in [\lambda - h, \infty) \end{cases} \\ &\geq f_{\lambda,a}(x). \end{aligned}$$

Thus $p(\lambda, a)$ is non–decreasing in a on $(0, \infty)$ and non–increasing on $(-\infty, 0)$ since μ is symmetric. Therefore

$$\sup_{|a|\geq\sigma} \frac{p(\lambda, a)}{\sigma^2/n + \min(a^2, \sigma^2)} = \frac{p(\lambda, \infty)}{\sigma^2/n + \sigma^2} \left(= \frac{p(\lambda, -\infty)}{\sigma^2/n + \sigma^2} \right).$$

We claim $a^2 + p(\lambda, 0) \geq p(\lambda, a)$. Indeed,

$$\begin{aligned} a^2 + f_{\lambda,0}(x) &= \begin{cases} a^2 + (x+\lambda)^2 & : x \in (-\infty, -\lambda] \\ a^2 & : x \in (-\lambda, \lambda) \\ a^2 + (x-\lambda)^2 & : x \in [\lambda, \infty) \end{cases} \\ &\geq \begin{cases} (x+\lambda)^2 & : x \in (-\infty, -\lambda - a] \\ a^2 & : x \in (-\lambda - a, \lambda - a) \\ (x-\lambda)^2 & : x \in [\lambda - a, \infty) \end{cases} \\ &= f_{\lambda,a}(x+a). \end{aligned}$$

Thus

$$\begin{aligned} a^2 + p(\lambda, 0) &= \int_{\mathbb{R}} (a^2 + f_{\lambda,0}(x))\mu(dx) \\ &\geq \int_{\mathbb{R}} f_{\lambda,a}(x+a)\mu(dx) \\ &= \int_{\mathbb{R}} f_{\lambda,a}(x)(\mu * \delta_a)(dx) \\ &= p(\lambda, a). \end{aligned}$$

For the moment assume that $p(\lambda_n, 0) \geq \sigma^2/n$, then for $a \in [-\sigma, \sigma]$,

$$\begin{aligned} \frac{p(\lambda_n, a)}{\sigma^2/n + \min(a^2, \sigma^2)} &= \frac{p(\lambda_n, a)}{\sigma^2/n + a^2} \\ &\leq \frac{a^2 + p(\lambda_n, 0)}{\sigma^2/n + a^2} \\ &= 1 + \frac{(p(\lambda_n, 0) - \sigma^2/n)}{\sigma^2/n + a^2} \\ &\leq 1 + \left(\frac{p(\lambda_n, 0) - \sigma^2/n}{\sigma^2/n} \right) \\ &= \frac{p(\lambda_n, 0)}{\sigma^2/n}. \end{aligned}$$

Therefore we conclude that,

$$\sup_{a\in\mathbb{R}} \frac{p(\lambda_n,a)}{\sigma^2/n+\min(a^2,\sigma^2)} = \sup_{a\in\{0,\infty\}} \frac{p(\lambda_n,a)}{\sigma^2/n+a^2} = \Lambda_n.$$

It remains to show that if $n \geq 4$, then $p(\lambda_n,0) \geq \sigma^2/n$. Recall that $p(\lambda_n,0)/(\sigma^2/n) = (\sigma^2+\lambda_n^2)/(1+1/n)\sigma^2$. Thus, the conditions $p(\lambda_n,0) \geq \sigma^2/n$ and $\lambda_n^2 \geq \sigma^2/n$ are equivalent. We have

$$p(\lambda,0) = Eg(z_1^2) \geq g(Ez_1^2) = g(\sigma^2) = Eg(Y^2) = p_Y(\lambda,0),$$

where $g(x) = (\sqrt{|x|}-\lambda)_+^2$ (g is convex !) and Y has distribution $(\delta_{-\sigma}+\delta_{+\sigma})/2$. Hence $p(\lambda,0) \geq p_Y(\lambda,0)$. But this implies that λ_n is greater or equal to the solution $\widetilde{\lambda}_n$ of $(n+1)p_Y(\lambda,0) = \lambda^2+\sigma^2$. Thus it is enough to prove that for $n \geq 4$, $\widetilde{\lambda}_n^2 \geq \sigma^2/n$. For $n=4$, $\widetilde{\lambda}_n$ is the solution (smaller or equal to σ) of

$$5(\sigma-\lambda)^2 = \lambda^2+\sigma^2,$$

, i.e., $\widetilde{\lambda}_4 = \sigma/2$. Thus $\widetilde{\lambda}_4^2 = \sigma^2/4$ and since $\widetilde{\lambda}_n$ is increasing in n, $\lambda_n \geq \sigma^2/n$ for $n \geq 4$.

To finish the proof, we show that λ_n is the optimal threshold. For $\lambda > 0$ we have (choosing $\theta = 0$ or $\theta = \infty$)

$$\begin{aligned}\sup_{\theta\in\mathbb{R}^n} \frac{E\|T_\lambda^S(Y)-\theta\|^2}{B_n(\theta,\sigma^2)} &\geq \max\left(\frac{np(\lambda,0)}{\sigma^2}, \frac{np(\lambda,\infty)}{(n+1)\sigma^2}\right)\\ &\geq \Lambda_n,\end{aligned}$$

where the second inequality holds since λ_n was chosen to minimize the second term and Λ_n is the minimum of this term. □

Remark 2.2 *If the z_i are compactly supported, say on $[-M,M]$, it follows that* $\lim_{\lambda\to M} p(\lambda,0) = 0$. *Thus $\lambda_n \to M$ and $\Lambda_n \to \dfrac{\sigma^2+M^2}{\sigma^2}$. In this case soft thresholding is almost as efficient as the ideal estimator. For example, if z_1 has law $(\delta_{-\sigma}+\delta_{+\sigma})/2$, then $\lambda_n \to \sigma$ and $\Lambda_n \to 2$. Thus soft thresholding is only twice as worse as the ideal estimator. For the uniform distribution on $[-M,M]$, the ratio is 4. The ratio becomes worse if the size of the support is large in comparison to the variance.*

Remark 2.3 *Theorem 2.1 continues to hold if μ is not symmetric. It was just notational convenience not to include this case. One just has to replace in (2.2)*

$$2\int_{\lambda}^{\infty}(|x|-\lambda)^2\mu(dx) \text{ by } \int_{|x|>\lambda}^{\infty}(|x|-\lambda)^2\mu(dx).$$

Also it might be better to use different thresholds for each side, but this is another class of estimators. Anyway, then the optimal pair of thresholds (λ_{left} and λ_{right}) is the solution of

$$\begin{aligned}(n+1)&\left(\int_{-\infty}^{\lambda_{left}}(|x|-\lambda_{left})^2\mu(dx)+\int_{\lambda_{right}}^{\infty}(|x|-\lambda_{right})^2\mu(dx)\right)\\ &=\max(\lambda_{left}^2,\lambda_{right}^2)+\sigma^2.\end{aligned}$$

The proof of this claim is basically the same as the one presented above.

Remark 2.4 *The previous method can also be applied to other loss functions. The quadratic loss can be replaced by any even loss function ℓ, which is increasing on the positive axis and 0 at 0. In the benchmark we have to replace σ^2 by $m_\ell := E\ell(z_1)$ and $\sum_i \min(\theta_i^2,\sigma^2)$ by $\sum_i \min(\ell(\theta_i), m_\ell)$. The computations of the optimal thresholds do not change much, but the condition $n \geq 4$ has to be changed, it depends not only on ℓ, but also on the distribution of z_1.*

Unfortunately the previous result is not directly applicable to the noisy wavelet coefficients, the requirement of identically distributed random variables is too strong. In the case of non identically distributed random variables, the following result gives a good suggestion for a threshold and an upper bound for the ratio of risks, namely compute $\lambda_{n,i}$ for each z_i separately, and choose for global threshold the largest of those $\lambda_n = \sup_i \lambda_{n,i}$. This will give a bound on the ratio of risks which is more handy (although bigger) than $\sup_i \Lambda_{n,i}$ (note that $\inf_i \Lambda_{n,i}$ provides a lower bound). Note also that below the benchmark $B_n(\theta,\sigma)$ has been replaced by $\sum_{i=1}^n(\min(\theta_i^2,\sigma_i^2)+\sigma_i^2/n)$.

Theorem 2.5 *Let $Y_i = \theta_i + z_i$, $i = 1,\ldots,n$, with $n \geq 4$, where the z_i are symmetric (about zero) random variables with law μ_i such that $Ez_i^2 = \sigma_i^2$. Let $\bar\sigma^2 = (\sum_{i=1}^n \sigma_i^2)/n$. As we already know, the equation*

$$2(n+1)\int_{\lambda}^{\infty}(x-\lambda)^2\mu_i(dx) = \lambda^2+\sigma_i^2, \quad \lambda > 0,$$

has a unique solution $\lambda_{n,i}$. *Let* $\lambda_n \geq \sup_i \lambda_{n,i}$ *and* $\Lambda_n := \sup_i(\lambda_n^2+\sigma_i^2)/(\sigma_i^2(1+1/n))$. *Then*

$$\sup_{\theta\in\mathbb{R}^n} \frac{E\|T^S_{\lambda_n}(Y)-\theta\|^2}{\bar\sigma^2+\sum_{i=1}^n \min(\theta_i^2,\sigma_i^2)} \leq \Lambda_n.$$

Proof: The proof is similar to the proof of the previous theorem. All we need to show is:

$$\sup_{a\in\mathbb{R}} \frac{p_i(\lambda_n,a)}{\sigma_i^2/n+\min(a^2,\sigma_i^2)} \leq \Lambda_n, \quad i=1,\ldots,n,$$

where $p_i(\lambda,a) = \int_{\mathbb{R}} |T^S_\lambda(x+a)-a|^2\mu_i(dx)$. As was shown in the previous proof, the left hand side above is smaller than the maximum of

$$\frac{\lambda_n+\sigma_i^2}{(1+1/n)\sigma_i^2} \text{ and } \sup_{a\in[-\sigma_i,\sigma_i]} \frac{p_i(\lambda_n,a)}{\sigma_i^2/n+a^2}.$$

Now, the above first term is less or equal than Λ_n and for the second term we have

$$\begin{aligned}
\sup_{a\in[-\sigma_i,\sigma_i]} \frac{p_i(\lambda_n,a)}{\sigma_i^2/n+a^2} &\leq \sup_{a\in[-\sigma_i,\sigma_i]} \frac{p_i(\lambda_n,0)+a^2}{\sigma_i^2/n+a^2} \\
&\leq \sup_{a\in[-\sigma_i,\sigma_i]} \frac{p_i(\lambda_{n,i},0)+a^2}{\sigma_i^2/n+a^2} \\
&= \frac{\lambda_{n,i}^2+\sigma_i^2}{\sigma_i^2(1+1/n)} \\
&\leq \Lambda_n,
\end{aligned}$$

where the second and third inequality hold since $\lambda_{n,i} \leq \lambda_n$, and the equality holds because of the properties of $\lambda_{n,i}$ obtained in the proof of Theorem 2.1.□

We want to apply the above theorem to the case where $Y_i = \theta_i + z_i$ are empirical wavelet coefficients. Since the z_i are linear combinations of the initial noise, $\lambda_{n,i}$ is quite complicated to compute. An alternative is to find an upper bound for the $\lambda_{n,i}$ which only depends on the initial noise (e). Since W_n is orthogonal, we know that $z_k = \sum_{i=1}^n w_{k,i}e_i$ with $\sum_{i=1}^n w_{k,i}^2 = 1$. The tail behavior of z_k is thus crucial for $\lambda_{n,k}$. First we want to compute an upper bound for the performance of soft thresholding. Because of Theorem 2.5 we just have to find an upper bound for all the optimal thresholds. For given symmetric distributions μ_1 and μ_2 with $\mu_1([a,\infty)) \leq \mu_2([a,\infty))$, $a \geq 0$, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, let λ_i, $i = 1,2$, be the solutions of

$$2(n+1)\int_\lambda^\infty (x-\lambda)^2 \mu_i(dx) = \lambda^2 + \sigma^2, \quad c > 0,\ n \in \mathbb{N},\ i = 1,2.$$

Clearly λ_2 is larger than λ_1. If the e_i are iid (or just independent or bounded martingale differences) with support $[-M, M]$ and mean zero then $z_{j,k} = \sum_i w_i e_i$ with $\sum_i w_i^2 = 1$ and by a result of Hoeffding (see [40, p.855]) we have for $t \geq 0$,

$$P(z_{j,k} \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i w_i^2 (M-(-M))^2}\right) = \exp\left(\frac{-t^2}{2M^2}\right).$$

Thus an upper bound for the optimal thresholds of the empirical wavelet coefficients is the solution of (2.2) for the symmetric distribution μ_2 defined by $\mu_2([t,\infty)) := \exp\left(\frac{-t^2}{2M^2}\right)$. Later we will see that for μ_2 this solution is asymptotically like $\sqrt{2M^2 \log n}$. Actually, estimates more precise than $\exp\left(\frac{-t^2}{2M^2}\right)$ are available but all in all, the asymptotics of the threshold do not change much with these.

If e_1 has bigger tail than the normal distribution, then the central limit theorem applies and the distribution of $\sum_i w_{k,i} e_i$ "tends" to the normal distribution, i.e. its tail is becoming smaller. This is the heuristic which leads to the following:

Theorem 2.6 *Let μ be a distribution on $\mathbb{R}$ which is a variance mixture of normal distributions, i.e. it is absolutely continuous with density*

$$\int_0^\infty \varphi_a(x)\nu(da),$$

where ν is a probability measure on $(0,\infty)$ and where φ_a is the centered normal density of variance a. Let X_i, $i = 1,\ldots,n$, be iid random variables with distribution μ. If g is a convex function on $\mathbb{R}^+$, then for any $a_1,\ldots,a_n \in \mathbb{R}$, $b_1,\ldots,b_n \in \mathbb{R}$, whose respective squares are written in non–increasing *order* $(\cdots a_i^2 \geq a_{i+1}^2 \geq \cdots, \cdots b_i^2 \geq b_{i+1}^2 \geq \cdots)$ *with* $\sum_{i=1}^k a_i^2 \leq \sum_{i=1}^k b_i^2$, $k = 1, \ldots, n$, *and* $\sum_{i=1}^n a_i^2 = \sum_{i=1}^n b_i^2$,

$$Eg\left(\left(\sum_{i=1}^n a_i X_i\right)^2\right) \leq Eg\left(\left(\sum_{i=1}^n b_i X_i\right)^2\right).$$

Proof: Let $R^N(a) := Eg(aN^2)$ where N is a $N(0,1)$ random variable. R^N is a convex function since $a \to g(ax^2)$ is convex for all $x \in \mathbb{R}$. It is easy to see, that for $c_1, \ldots, c_n \in \mathbb{R}$ the distribution of $\sum_{i=1}^n c_i X_i$ is again a variance mixture of normal random variables with mixing measure $\nu_{c_1^2} \star \ldots \star \nu_{c_n^2}$, where ν_a is an abbreviation for the measure $\nu(\cdot/a)/a$, $a > 0$ and ν_0 is the Dirac measure with mass at 0. It follows that

$$\begin{aligned} R(a_1^2, \ldots, a_n^2) &:= Eg\left(\left(\sum_{i=1}^n a_i X_i\right)^2\right) \\ &= \int_{-\infty}^{+\infty} g(x^2)\left(\int_0^\infty \cdots \int_0^\infty \varphi_{\sum u_i a_i^2}(x)\nu(du_1)\ldots\nu(du_n)\right) dx \\ &= \int_0^\infty \cdots \int_0^\infty \left(\int_{-\infty}^{+\infty} g(x^2)\varphi_{\sum u_i a_i^2}(x)dx\right)\nu(du_1)\ldots\nu(du_n) \\ &= ER^N\left(\sum_{i=1}^n Y_i a_i^2\right), \end{aligned}$$

where Y_i are iid random variables with distribution ν. Since R^N is convex it follows from Marshall and Olkin[33, ch.12, beginning of Sect.G] that R is Schur-convex, i.e. if $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ in decreasing order and $\sum_{i=1}^k \alpha_i \leq \sum_{i=1}^k \beta_i$, $k = 1, \ldots, n$, $\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i$, then:

$$R(\alpha_1, \ldots, \alpha_n) \leq R(\beta_1, \ldots, \beta_n).$$

Thus the assertion follows. □

This theorem gives the needed bounds on $p(\lambda, 0)$; indeed $(|x| - \lambda)_+^2 = (\sqrt{x^2} - \lambda)_+^2 = g(x^2)$ with $g(x) = (\sqrt{x} - \lambda)_+^2$ and g is convex for $x \geq 0$ (see Proposition 5 in [2]).

What kind of distributions are variance mixtures of normal random variables? This family is closed under mixtures and convolutions. Parametrizing by the variance, the mixture measure of a convolution is the convolution of mixture measures. Let f be a variance mixture of normal densities, then

$$f\left(\sqrt{|x|}\right) = \int_0^\infty \frac{1}{\sqrt{2\pi a}} \exp(-|x|/(2a))\nu(da).$$

Thus f is a variance mixture of normal densities if and only if f is even and $f(\sqrt{x})$, $x \in [0, \infty)$ is a Laplace transform. In turn, it is well known ([21, p.415]) that a function g is a Laplace transform if and only if it is completely monotone, i.e.

$$(-1)^n g^{(n)}(x) \geq 0, \ \forall x \in [0, \infty), \ \forall n \in \mathbb{N}.$$

If $f(x) = \exp(-x^c)$, $0 < c \leq 1$, it can be shown by induction that

$$f^{(k)}(x) = \sum_{i=1}^{n_k} (-1)^k x^{a_{k,i}} b_{k,i} \exp(-x^c),$$

with $a_{k,i} \leq 0$ and $b_{k,i} \geq 0$. Thus f is a Laplace transform and so is the density

$$h(x) = C_1 \exp(-C_2 x^c), \quad 0 < c < 2,$$

where C_1, C_2 are appropriate constants. Another example of a completely monotone function is $1/(1+x)^n$, which implies that densities of the form

$$\frac{constant}{(1+x^2)^n}, \ n \geq 1$$

are variance mixtures as well. In Feller ([21, XII.4]) two criteria are mentioned, namely if f and g are completely monotone then fg is completely monotone and if f is completely monotone and g is positive and its derivative is completely monotone then $f \circ g$ is completely monotone.

Often the class of variance mixtures of the normal distribution is not large enough for a specific application. But sometimes we are able to carry over the properties of these densities to other "nearby" densities, in fact, we just need an upper bound for $p(0, \lambda)$. If the distribution of the e_i is not a normal mixture but is only symmetric, and if there exists a random variable v whose distribution is a normal mixture such that $P(|e_1| > x) \leq KP(|v| > x)$, $K \geq 1$, then

$$ET_\lambda^S\left(\sum_i a_i e_i\right)^2 \leq ET_\lambda^S\left(K\sum_i a_i v_i\right)^2,$$

where the v_i are iid copies of v. This is a consequence of a version of the concentration principle ([31, Lemma 4.6]). Let λ_n be the positive solution of $(n+1)E(|Kv|-\lambda)_+^2 = \lambda^2+\sigma^2$ then

$$\sup_{\theta\in\mathbb{R}} \frac{E\left(T_{\lambda_n}^S\left(\sum_i a_i e_i + \theta\right) - \theta\right)^2}{\sigma^2/n + \min(a,\sigma^2)} \leq \Lambda_n, \tag{2.3}$$

where $\Lambda_n = (\lambda_n^2+\sigma^2)/((1+1/n)\sigma^2)$. If $P(|e_1| > x) \leq KP(|v| > x)$ holds only for $t \geq t_0$ then (again by [31, Lemma 4.6]),

$$ET_\lambda^S\left(\sum_i a_i e_i\right)^2 \leq \frac{1}{2}ET_\lambda^S\left(2Kt_0\sum_i a_i v_i\right)^2 + \frac{1}{2}ET_\lambda^S\left(2K\sum_i a_i u_i\right)^2 =: h(\lambda), \tag{2.4}$$

where the u_i are iid random variables with distribution $(\delta_1+\delta_{-1})/2$. So if λ_n is the solution of $(n+1)h(\lambda) = \lambda^2+\sigma^2$, then again relation (2.3) holds with $\Lambda_n = (\lambda_n^2+\sigma^2)/((1+1/n)\sigma^2)$.

It would be desirable to generalize Theorem 2.6 to a wider class of distributions, or to obtain a version of Theorem 2.6 which directly describes the behavior of the tails of the sums. There are results of this type for special classes of distributions, where the tail of the sums is bounded by the tail of the initial random variable ([4]).

3 Near minimaxity

If the noise is normal, thresholding achieves the minimax rate in the class of all estimators, i.e.

$$\lim_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)}\Lambda_n^{-1} = 1,$$

where the infimum is taken over all estimators and where Λ_n is now computed for the normal distribution ([16]). For a special class of distributions one can also show that soft thresholding is asymptotically "near" minimax, i.e. the 1 on the above right-hand side is replaced by a constant. This result is a natural consequence of the next two theorems. First an asymptotic rate for λ_n is given for distributions with exponentially decaying density.

Theorem 3.1 *Let μ be a distribution on $\mathbb{R}$ with variance σ^2 and density $\exp(-h(x))$, where h is even, continuous and increasing on $[0,\infty)$. Further let*

$$\liminf_{x\to\infty}\frac{h(cx)}{h(x)} > 1$$

for all $c > 1$. Let λ_n be the solution of

$$2(n+1)\int_\lambda^\infty (x-\lambda)^2\mu(dx) = \lambda^2+\sigma^2, \quad \lambda > 0,$$

and $\Lambda_n := (\lambda_n^2+\sigma^2)/(\sigma^2(1+1/n))$. Then

$$\lim_{n\to\infty}\frac{h^{-1}(\log(n))}{\lambda_n} = 1, \text{ and } \lim_{n\to\infty}\frac{(h^{-1}(\log(n))^2+\sigma^2)}{\sigma^2\Lambda_n} = 1.$$

Proof: First let us note a consequence of the conditions imposed on h. If λ tends to infinity $h(c\lambda) - h(x+\lambda) - \log(\lambda^2+\sigma^2)$ tends to infinity if $c > 1$, and tends to negative infinity if $0 < c < 1$, for all $x \in \mathbb{R}$.
Set

$$q(\lambda) := \frac{\lambda^2+\sigma^2}{2\int_\lambda^\infty (x-\lambda)^2 f(x)dx} - 1, \quad \lambda > 0.$$

Clearly q is increasing. Let $1 > \delta > 0$ then

$$q(\lambda)f((1+\delta)\lambda) \underset{\lambda\to\infty}{\longrightarrow} 0 \text{ and } q(\lambda)f((1-\delta)\lambda) \underset{\lambda\to\infty}{\longrightarrow} \infty.$$

We prove the first asymptotic: let $\lambda \geq 1$, then

$$\begin{aligned}
&\left(\frac{\lambda^2+\sigma^2}{2\int_\lambda^\infty (x-\lambda)^2 f(x)dx} - 1\right) f((1+\delta)\lambda) \\
\leq\ & \frac{\lambda^2+\sigma^2}{\int_\lambda^\infty (x-\lambda)^2 f(x)dx} f((1+\delta)\lambda) \\
=\ & \left(\int_0^\infty x^2 \frac{f(x+\lambda)}{f((1+\delta)\lambda)(\lambda^2+\sigma^2)}dx\right)^{-1} \underset{\lambda\to\infty}{\longrightarrow} 0,
\end{aligned}$$

since

$$\frac{f(x+\lambda)}{f((1+\delta)\lambda)(\lambda^2+\sigma^2)} = \exp(h((1+\delta)\lambda) - \log(\lambda^2+\sigma^2) - h(x+\lambda))$$

and this tends for all x to infinity by the assumptions. Let us turn to the second asymptotic:

$$q(\lambda)f((1-\delta)\cdot\lambda) \geq \left(2\int_0^\infty x^2 \frac{f(x+\lambda)}{f((1-\delta)\lambda)(\lambda^2+\sigma^2)}\,dx\right)^{-1} - f((1-\delta)\lambda) \underset{\lambda\to\infty}{\longrightarrow} \infty.$$

The second summand on the right hand side converges to 0. The integrand in the first summand is equal to

$$x^2 \exp(h((1-\delta)\lambda) - h(x+\lambda) - \log(\lambda^2+\sigma^2)),$$

the exponent tends, by assumption, to $-\infty$, for all $x \in \mathbb{R}$. Therefore the integral tends to 0 and the whole term to infinity.

Thus for x sufficiently large (depending on δ),

$$\frac{1}{f((1-\delta)x)} \leq q(x) \leq \frac{1}{f((1+\delta)x)},$$

and hence

$$\frac{1}{1-\delta} f^{-1}(1/y) \geq q^{-1}(y) \geq \frac{1}{1+\delta} f^{-1}(1/y),$$

for y large enough. Since λ_n is the solution of $q(\lambda_n) = n$, it follows that

$$\lim_{n\to\infty} \frac{\lambda_n}{f^{-1}(1/n)} = \lim_{n\to\infty} \frac{\lambda_n}{h^{-1}(\log n)} = 1.$$

$\square$

Remark 3.2 *If the density f is asymptotically like an inverse fractional polynomial, i.e., if there is a $c > 3$ such that*

$$\liminf_{x\to\infty} f(x)x^c = \limsup_{x\to\infty} f(x)x^c \in (0,\infty),$$

then easy computations yield $p(\lambda, 0) \sim \alpha/\lambda^{c-3}$ and this implies $\lambda_n \sim \sqrt[c-1]{\alpha n}$ for some constant $\alpha > 0$.

Remark 3.3 *If the additional summand σ^2 in $B_n(\theta,\sigma^2)$ is replaced by $c_n\sigma^2$, $c_n = O((\log n)^\beta)$, $\beta > 0$, then after changing q to $q(\lambda) := (\lambda^2+\sigma^2)/p(\lambda,0) - c_n$, the optimal thresholds are still the solution of $q(\lambda) = n$, and the asymptotic behavior of q^{-1} is not changed. Thus the asymptotic behavior of the thresholds does not change too.*

Theorem 3.4 *Let $Y_i = \theta_i + z_i$, $i = 1,\ldots,n$, where the z_i are iid random variables with $Ez_1 = 0$ and $Ez_1^2 = \sigma^2$. Let the Lebesgue density of the distribution μ of z_1 be of the form $\exp(-h(x))$, where h is even, continuous and increasing on $[0,\infty)$.*
Further, let

$$\liminf_{x\to\infty} \frac{h^{-1}(x)}{h^{-1}(x - 2\log x)} = 1,$$

then

$$\liminf_n \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)} \frac{\sigma^2}{(h^{-1}(\log n))^2} \geq 1,$$

where the infimum is taken over all estimators $\widehat{\theta}$ of θ.

Proof: We will prove this bound by computing Bayes risks. The proof is similar to the proof of Theorem 4 in [16]. For $0 < \varepsilon < 1$ and $a > 0$, let $F_{\varepsilon,a} := \varepsilon\delta_a + (1-\varepsilon)\delta_0$, where δ_c denotes the Dirac measure with mass at c. The a-priori measure for $\theta \in \mathbb{R}^n$ will be $Q_n := \otimes_{i=1}^n F_{\varepsilon_n,a_n}$, we will specify ε_n and a_n later. For now, it suffices to assume that $\varepsilon_n \to 0$ and $a_n \to \infty$.
First we consider the one-dimensional case. We compute the Bayes risk for estimating $\theta_1 \in \mathbb{R}$ given $Y_1 = \theta_1 + z_1$ and the a-priori measure for θ_1 is $F_{\varepsilon,a}$. Let $M := F_{\varepsilon,a} * f$, i.e.

$$M(A,B) = (1-\varepsilon)\delta_0(A)\int_B f(x)dx + \varepsilon\delta_a(A)\int_B f(x-a)dx.$$

Let Π_1 and Π_2 be the projection from $\mathbb{R}^2$ to the first, respectively the second coordinate. Then the Bayes estimator in this context for θ is

$$\begin{aligned} d_{\varepsilon,a}(x) &= E_M(\Pi_1|\Pi_2 = x) \\ &= \frac{0(1-\varepsilon)f(x) + a\varepsilon f(x-a)}{(1-\varepsilon)f(x) + \varepsilon f(x-a)} \\ &= \frac{\varepsilon f(x-a)}{\varepsilon f(x-a) + (1-\varepsilon)f(x)} a. \end{aligned}$$

Thus,

$$
\begin{aligned}
& E_{F_{\varepsilon,a}} E_{\theta_1} (d_{\varepsilon,a} - \theta_1)^2 \\
&= \varepsilon \int (d_{\varepsilon,a}(x) - a)^2 f(x-a) dx + (1-\varepsilon) \int d_{\varepsilon,a}(x)^2 f(x) dx \\
&\geq \varepsilon a^2 \int \left(1 - \frac{\varepsilon f(x-a)}{\varepsilon f(x-a) + (1-\varepsilon) f(x)} \right)^2 f(x-a) dx \\
&= \varepsilon a^2 \int \left(\frac{(1-\varepsilon) f(x)}{\varepsilon f(x-a) + (1-\varepsilon) f(x)} \right)^2 f(x-a) dx \\
&= (1-\varepsilon)^2 \varepsilon a^2 \int \frac{f(x)^2}{(\varepsilon f(x-a) + (1-\varepsilon) f(x))^2} f(x-a) dx.
\end{aligned}
$$

Now let $\alpha \in (0,1)$, then there exists a $c > 0$ satisfying,

$$\int_{-c}^{c} f(x) dx \geq \alpha.$$

Let $\beta > 0$ and assume a and ε satisfy

$$\beta f(a+c) \geq \frac{\varepsilon}{1-\varepsilon} f(0).$$

If $x \in (a-c, a+c)$, then

$$\beta f(x) \geq \frac{\varepsilon}{1-\varepsilon} f(x-a).$$

This implies

$$\frac{f(x)^2}{(\varepsilon f(x-a) + (1-\varepsilon) f(x))^2} \geq \frac{f(x)^2}{((1-\varepsilon) f(x)(1+\beta))^2}, \text{ for all } x \in (a-c, a+c).$$

Hence, by integrating only on $(a-c, a+c)$ we get

$$
\begin{aligned}
E_{F_{\varepsilon,a}} E_{\theta_1} (d_{\varepsilon,a} - \theta_1)^2 &\geq \frac{(1-\varepsilon)^2}{(1-\varepsilon)^2} \frac{\alpha}{(1+\beta)^2} \varepsilon a^2 \\
&= \frac{\alpha}{(1+\beta)^2} \varepsilon a^2.
\end{aligned}
$$

Now let α, c and β be fixed, let ε_n, a_n be sequences such that $\beta f(a_n + c) \geq \frac{\varepsilon_n}{1-\varepsilon_n} f(0)$, $\varepsilon_n n \to \infty$ and set

$$\begin{aligned} m_n &:= (n\varepsilon_n)^{2/3}, \\ N_n &:= \#\{\theta_i \neq 0, i = 1, \ldots, n\}, \\ A_n &:= \{N_n \leq n\varepsilon_n + m_n\}, \\ p_n &:= Q_n(A_n^c). \end{aligned}$$

We will now prove $p_n = Q_n(N_n - n\varepsilon_n > m_n) = o(\varepsilon_n)$ by using Benett's inequality ([40, p.851]). This inequality gives the first of the following inequalities, here ψ is a fixed continuous function which is decreasing on $[0, \infty)$ and $\psi(0) = 1$.

$$\begin{aligned} p_n &\leq \exp\left(-\frac{m_n/\sqrt{n}}{2\varepsilon_n(1-\varepsilon_n)}\psi\left(\frac{m_n/\sqrt{n}}{\varepsilon_n(1-\varepsilon_n)\sqrt{n}}\right)\right) \\ &\leq \exp\left(-\frac{(\varepsilon_n n)^{2/3}}{2\varepsilon_n\sqrt{n}}\psi\left(\frac{(\varepsilon_n n)^{2/3}}{n\varepsilon_n(1-\varepsilon_n)}\right)\right) \\ &= \exp\left(-\frac{1}{2}\varepsilon_n^{-1/3}n^{1/6}\psi\left(\frac{(\varepsilon_n n)^{-1/3}}{1-\varepsilon_n}\right)\right) \\ &\leq \exp\left(-\frac{1}{4}\varepsilon_n^{-1/3}n^{1/6}\right) \\ &\qquad \text{for } n \text{ large enough since } n\varepsilon_n \to \infty \text{ and } \psi(0) = 1 \\ &= o(\varepsilon_n). \end{aligned}$$

Now we are ready to tackle the Bayes risk of Q_n. Let d_n be the Bayes estimator for Q_n, it is clear $a_n \geq d_n \geq 0$.

$$\begin{aligned} & E_{Q_n}E_\theta \frac{\|d_n(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} \\ &\geq \frac{1}{\sigma^2(1 + n\varepsilon_n + m_n)} E_{Q_n}E_\theta \sum_{i=1}^n (d_n^i(Y) - \theta_i)^2 \mathbf{1}_{A_n} \\ &\geq \frac{1}{\sigma^2(1 + n\varepsilon_n + m_n)} \left(E_{Q_n}E_\theta \sum_{i=1}^n (d_n^i(Y) - \theta_i)^2 - np_n a_n^2\right) \\ &\qquad \text{since } E_\theta(d_n^i(Y) - \theta_i)^2 \mathbf{1}_{A_n^c} \leq p_n a_n^2 \\ &= \frac{1}{\sigma^2(1 + n\varepsilon_n + m_n)} \left(\sum_{i=1}^n E_{Q_n}E_{\theta_i}(d_n^i(Y) - \theta_i)^2 - np_n a_n^2\right) \end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\left(n\varepsilon_n a_n^2\frac{\alpha}{(1+\beta)^2}-np_n a_n^2\right)\\
&\sim \frac{\alpha}{(1+\beta)^2\sigma^2}a_n^2,
\end{aligned}
$$

since $p_n a_n^2 = o(\varepsilon_n a_n^2)$ and $m_n = o(\varepsilon_n n)$.
How can we choose ε_n and a_n, such that a_n is as big as possible? We remember that a_n and ε_n have to satisfy $\varepsilon_n n \to \infty$ and $\beta f(a_n + c) \geq \dfrac{\varepsilon_n}{1-\varepsilon_n} f(0)$. The second condition is equivalent to $f(a_n + c) \geq \dfrac{\varepsilon_n}{\beta(1-\varepsilon_n)} f(0)$. Thus, if $\varepsilon_n = \log n/n$ we can choose

$$a_n = h^{-1}(\log n - \log\log n + \log(1 - \log(n)/n) + \log\beta - \log f(0)) - c.$$

Because of the conditions imposed on h^{-1} we have $a_n \sim h^{-1}(\log n)$. Since α, c and β are arbitrary, the theorem follows. □

Remark 3.5 *The condition*

$$\liminf_{x\to\infty}\frac{h^{-1}(x)}{h^{-1}(x-2\log x)} = 1$$

in Theorem 3.4 is satisfied if h grows at least as fast as a fractional polynomial. Combining the last two theorems we see that if the noise in the data is iid and its density is asymptotically like $\exp(-h(x))$ where h is a fractional polynomial, then soft thresholding is optimal in the minimax sense, it has the asymptotically best ratio of risk and benchmark.

Remark 3.6 *What happens if the additional σ^2 in equation (2.1) is replaced by other values c_n? Important is the last step in the previous theorem, namely*

$$\frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\left(n\varepsilon_n a_n^2\frac{\alpha}{(1+\beta)^2}-np_n a_n^2\right) \sim \frac{\alpha}{(1+\beta)^2\sigma^2}a_n^2.$$

We have to replace the 1 in the denominator on the left hand side by c_n. The $\sim$ remains valid if $c_n/(\varepsilon_n n) \to 0$. To keep the same rate for a_n it is necessary that $\log\varepsilon_n \sim -\log n$. Thus, if for example $c_n \sim (\log n)^p$ then with $\varepsilon_n = (\log n)^{p+1}/n$ we get the same asymptotic rate for a_n and the Bayes risk for Q_n.

Theorem 3.7 *Let the observations $X_i = f_i + e_i$, $i = 0, \dots, n-1 = 2^m - 1$, be given, where the f_i are the parameters of interest and the e_i are iid random variables whose distribution μ has the density $g(x) := \exp(-h(x))$ where h is even, continuous and increasing on $[0, \infty)$. Further, let*

$$\limsup_{x \to \infty} \frac{h(cx)}{h(x)} < \infty, \quad \forall c > 0.$$

Apply a periodic wavelet transform W_n to these observations, taking for every n the same type of wavelet. Let $Y = W_n(X)$ and let $\theta = W_n(f)$, then

$$\liminf_{n \to \infty} \inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^n} \frac{E\|\widehat{\theta} - \theta\|^2}{B_n(\theta, \sigma^2)} \frac{1}{(h^{-1}(\log n))^2} > 0,$$

where the infimum is taken over all estimators $\widehat{\theta}$ of θ.

Remark 3.8 *In contrast to Theorem 3.4, this theorem is only about the rate of the ratios. The idea of the proof is the same in both theorems, but the proof is now complicated by the fact that we do not have a closed expression for the density of the noise in the empirical wavelet coefficients. Since we have to rely on rough bounds for these tails, we only get a statement about the rate.*

Proof: We will apply a wavelet transform W_n derived from a multiresolution analysis with a wavelet which has compact support. The fixed filter length is N, N is even. $Y_{j,k}$, $\theta_{j,k}$ and $z_{j,k}$, $0 \le k \le 2^j - 1$, $0 \le j \le m-1$, will denote the k^{th} wavelet coefficient at level j of $W_n(X)$, $W_n(f)$ respectively $W_n(e)$. Let $\widehat{\theta}_{j,k}$ be an estimator for $\theta_{j,k}$, $\widehat{\theta}_{j,k}$ may depend on all X_i. We will compute an asymptotic lower bound for

$$\sup_{\theta} \frac{\sum_{j,k} E|\widehat{\theta}_{j,k} - \theta_{ij}|^2}{B_n(\theta, \sigma^2)},$$

by computing a Bayes risk.
For $1 > \varepsilon > 0$ and $a > 0$ let $F_{\varepsilon,a}$ be defined as in the last proof. The a priori measure for $\theta_{j,k}$ is F_{ε_n, a_n} if $j = m-1$ and δ_0 otherwise. For the moment it suffices to know $\varepsilon_n \to 0$ and $a_n \to \infty$. The a priori measure Q_n for $\theta \in \mathbb{R}^n$ is the product measure of the a priori measures of the coordinates.
In the following we will omit the level coefficient $m-1$, since we are only concerned with the coefficients in this level.

Let us observe the following: For n greater than a certain bound, the filter coefficients $d_0, \ldots, d_{N-1}$ to compute the $\theta_{j,k}$ do no longer depend on n. Thus $\theta_i = \sum_{\ell=0}^{N-1} d_\ell f_{2i+\ell}$ where the indices for f are considered modulo n, we are using a periodic wavelet transform! On the other hand, given the a priori measure Q_n (i.e. assuming the wavelet coefficients at levels $j < m-1$ are 0) one computes the f_i by

$$f_i := \sum_{\ell=0,\ \ell \text{ even}}^{N-1} d_\ell \theta_{(i-\ell)/2}, \qquad \text{if } i \text{ is even,}$$

and

$$f_i := \sum_{\ell=0,\ \ell \text{ odd}}^{N-1} d_\ell \theta_{(i-\ell)/2}, \qquad \text{if } i \text{ is odd.}$$

This follows immediately from the fact that W_n is orthonormal, i.e. $W_n^T = W_n^{-1}$ in the language of matrices.
For $0 \leq i \leq n/2 - 1$, we set

$$Q_n^i(\cdot) := \frac{Q_n(\cdot \cap \{\theta_{i+j} = 0, 1 \leq |j| \leq N/2 - 1\})}{Q_n(\{\theta_{i+j} = 0, 1 \leq |j| \leq N/2 - 1\})},$$

where $0/0$ is understood as 0. If Q_n^i is the a priori measure, then θ_i and f_j are independent, if $j < 2i$ or $j > 2i + N - 1$. Hence the Bayes estimator for w_i given the a priori measure Q_n^i may depend only on $f_{2i}, \ldots, f_{2i+N-1}$. As one easily checks, the projection of Q_n^i on $f_{2i}, \ldots, f_{2i+N-1}$ is $\varepsilon_n \delta_{(d_0,\ldots,d_{N-1})a_n} + (1-\varepsilon_n)\delta_{(0,\ldots,0)}$ ($\delta_\cdot$ denotes the Dirac measure). Hence the Bayes estimator b_i for θ_i given Q_n^i is

$$b_i(x) := \frac{g_{a_n}(x)\varepsilon_n}{g_{a_n}(x)\varepsilon_n + (1-\varepsilon_n)g_0(x)} a_n, \quad x \in \mathbb{R}^N,$$

where g_{a_n} and g_0 are the densities of $X_{2i}, \ldots, X_{2i+N-1}$ if $(f_{2i}, \ldots, f_{2i+N-1}) = a_n(d_0, \ldots, d_{N-1})$ respectively if $(f_{2i}, \ldots, f_{2i+N-1}) = (0, \ldots, 0)$, i.e. $g_{a_n}(x) = \prod_{\ell=0}^{N-1} g(x_\ell - a_n d_\ell)$ and $g_0(x) = \prod_{\ell=0}^{N-1} g(x_\ell)$.

Now let $\widehat{\theta}_i$ be an estimator for θ_i, then

$$E_{Q_n} E_{\theta_i}(\widehat{\theta}_i - \theta_i)^2 \geq (1-\varepsilon_n)^{N-2} E_{Q_n^i} E_{\theta_i}(b_i - \theta_i)^2.$$

We proceed to compute the Bayes risk of b_i given the a priori measure Q_n^i:

$$\begin{aligned} E_{Q_n^i}E_{\theta_i}(b_i-\theta_i)^2 &= (1-\varepsilon_n)E_{\theta_i=0}b_i^2+\varepsilon_n E_{\theta_i=a_n}(b_i-a_n)^2 \\ &\geq \varepsilon_n\int_{\mathbb{R}^N}\left(\frac{g_{a_n}(x)\varepsilon_n}{g_0(x)(1-\varepsilon_n)+g_{a_n}(x)\varepsilon_n}a_n-a_n\right)^2 g_{a_n}(x)dx \\ &= \varepsilon_n a_n^2\int_{\mathbb{R}^N}\left(\frac{g_0(x)(1-\varepsilon_n)}{g_0(x)(1-\varepsilon_n)+g_{a_n}(x)\varepsilon_n}\right)^2 g_{a_n}(x)dx. \end{aligned}$$

Set $d_{max} := \max_{i=0,\dots,N-1}|d_i|$. Let $1>\alpha>0$, then there exists a $c>0$ such that $\int_{[-c,c]^N} g_0(x)dx \geq \alpha$. Now define $I_n := [-c,c]^N + a_n(d_0,\dots,d_{N-1})$, and assume that ε_n and a_n are such that

$$(1-\varepsilon_n)g_0(x) > 2\varepsilon_n g_{a_n}(x)$$

for all $x\in I_n$. Then

$$\begin{aligned} E_{Q_n^i}E_{\theta_i}(b_i-\theta_i)^2 &\geq \varepsilon_n a_n^2\int_{I_n}\left(\frac{g_0(x)(1-\varepsilon_n)}{g_0(x)(1-\varepsilon_n)3/2}\right)^2 g_{a_n}(x)dx \\ &\geq \frac{4}{9}\varepsilon_n a_n^2\alpha. \end{aligned}$$

which implies

$$E_{Q_n}E_w(\widehat{\theta}_i-\theta_i)^2 \geq (1-\varepsilon_n)^{N-2}\varepsilon_n a_n^2\alpha 4/9.$$

Thus for α large enough and n greater than a certain bound, the Bayes risk for estimating a single θ_i is greater or equal than $\frac{4}{10}\varepsilon_n a_n^2$.

Let $\varepsilon_n = \frac{\log n}{n}$ and a_n be sequences which satisfy $(1-\varepsilon_n)g_0(x) > 2\varepsilon_n g_{a_n}(x)$ for all $x\in I_n$ and $\varepsilon_n n\to\infty$. Set $m_n := \left(\frac{n}{2}\varepsilon_n\right)^{2/3}$. Define

$$\begin{aligned} N_n &:= \#\{\theta_i\neq 0, i=0,\dots,n/2-1\}, \\ A_n &:= \left\{N_n \leq \frac{n}{2}\cdot\varepsilon_n + m_n\right\}, \\ p_n &:= Q_n(A_n^c) = o(\varepsilon_n). \end{aligned}$$

The inequality for p_n was part of the last proof.
Let $\widehat{\theta}_{j,k}$ be an estimator for $\theta_{j,k}$, without loss of generality we assume that

$0 \le \widehat{\theta}_{m-1,k} \le a_n$. Now,

$$
\begin{aligned}
& E_{Q_n} E_\theta \frac{\sum_{j,k} |\widehat{\theta}_{j,k} - \theta_{j,k}|^2}{B_n(w,\sigma^2)} \\
\ge \quad & E_{Q_n} E_\theta \frac{\sum_k |\widehat{\theta}_k - \theta_k|^2}{\sigma^2(1+N_n)} \text{ leaving index } m-1 \text{ out} \\
\ge \quad & \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} E_{Q_n} E_\theta \sum_k |\widehat{\theta}_k - \theta_k|^2 \mathbf{1}_{A_n} \\
& \qquad \text{with } E_\theta \mathbf{1}_{A_n^c} |\widehat{\theta}_k - \theta_k|^2 \le Q_n(A_n^c) a_n^2 = p_n a_n^2 \\
\ge \quad & \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} E_{Q_n} E_\theta \sum_k |\widehat{\theta}_k - \theta_k|^2 - \frac{n}{2} p_n a_n^2 \\
\ge \quad & \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} \sum_k E_{Q_n} E_\theta |\widehat{\theta}_k - \theta_k|^2 - \frac{n}{2} p_n a_n^2 \\
\ge \quad & \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} \left(\frac{n}{2} \frac{4}{10} \varepsilon_n a_n^2 - \frac{n}{2} p_n a_n^2 \right), \qquad \text{for } n \text{ large} \\
\ge \quad & \frac{1}{3\sigma^2} a_n^2 \qquad \text{for } n \text{ large,}
\end{aligned}
$$

since $p_n = o(\varepsilon_n)$.

We take $\varepsilon_n := \frac{\log n}{n}$, assume n is so large that $\varepsilon_n < 1/2$, then we choose

$$a_n := h^{-1} \left(\frac{1}{N} (\log n - \log\log n) - \log \left(g(0) \sqrt[N]{4} \right) \right) - c.$$

Then

$$
\begin{aligned}
& h(a_n + c) = -\log(g(0) \sqrt[N]{4}) + \frac{1}{N}(\log n - \log\log n) \\
\Leftrightarrow \quad & \exp(-h(a_n + c)) = g(0) \sqrt[N]{\frac{\log n}{n}} \sqrt[N]{4} \\
\Leftrightarrow \quad & g(a_n + c) = g(0) \sqrt[N]{\varepsilon_n} \sqrt[N]{4}.
\end{aligned}
$$

Since $\varepsilon_n < 1/2$ and $|d_{max}| \le 1$,

$$g(d_{max} a_n + c) \ge g(0) \sqrt[N]{\varepsilon_n} \sqrt[N]{2/(1-\varepsilon_n)},$$

it follows

$$(1-\varepsilon_n)\prod_{\ell=0}^{N-1} g(d_\ell a_n + c) \geq 2\varepsilon_n g(0)^N,$$

and finally

$$(1-\varepsilon_n) g_0(x) \geq 2\varepsilon_n g_{a_n}(x),$$

for all $x \in I_n$. Thus our choice of a_n fulfills the required conditions. Since $a_n \sim$ constant $h^{-1}(\log n)$ the theorem is proved. □

4 Compactly supported distributions

In the previous section we considered noise whose tail was heavier than the tail of the normal distribution and computed lower asymptotic bounds for thresholding. We made use of the fact that half of the coefficients are in the finest level. The distribution of the noise in this level is "close" to the distribution of the original noise. Now we consider noise whose support is compact, and show that the asymptotic performance of thresholding in this case is not better than in the normal case. Here we will use the fact that the distribution of the noise in the upper level coefficients is near the normal distribution. From the discussion before Theorem 2.6, the result below can be complemented by the fact that for compactly supported noise e, Λ_n (as defined below) is such that $\Lambda_n = O(\log n)$.

Theorem 4.1 *Let the observations $X_i = f_i + e_i$, $i = 1, \ldots, n = 2^m$, be given, where the f_i are the parameters of interest and the e_i are iid random variables whose distribution has compact support $[-a, a]$. Apply a periodic wavelet transform W_n to these observations taking for every n the same type of wavelet such that the corresponding wavelet base has a Hölder regularity of $\beta > 0$. Let $Y = W_n(X)$, let $\theta = W_n(f)$, let $z = W_n(e)$ and let*

$$\Lambda_n := \inf_{(\lambda_{j,k})\in\mathbb{R}^n} \sup_{\theta\in\mathbb{R}^n} \frac{\sum_{j,k} E|T^S_{\lambda_{i,j}}(Y_{j,k}) - \theta_{j,k}|^2}{\sigma^2 + B_n(\theta,\sigma^2)}. \tag{4.1}$$

Then $\liminf_{n\to\infty} \dfrac{\Lambda_n}{2\log n} \geq 1.$

Proof: For a fixed $q \in (0,1)$ let $h = h(n) = [\log_2(n^q)]$. We will need a lower bound for $p(0,\lambda)$ for the noise $(z_{j,k})$ in the wavelet coefficients. As we already saw,

$$z_{j,k} = \sum_{i=1}^{n} w_{j,k,i}^n e_i \text{ with } \sum_{i=1}^{n} (w_{j,k,i}^n)^2 = 1.$$

We will now show that $\max_{j \le h,k,i} |w_{j,k,i}^n| 2^{(m-j)/2} \le C$ for some constant C. This inequality is needed for applying a result about tail behavior to $z_{j,k}$. If we use a multiresolution analysis which is not restricted to an interval as base for W^n, then

$$w_{j,k,i}^n = \langle \psi_{j,k}, \phi_{m,i} \rangle.$$

It is well known that if ϕ is Hölder continuous with exponent β then

$$\sup_{i,k} |2^{(m-j)/2} \langle \psi_{j,k}, \phi_{m,i} \rangle - \psi(2^{j-m} i - k)| \le C 2^{\beta(j-m)} \tag{4.2}$$

for a constant C and $m - j > j_0$ for some j_0 ([10, p.205]). But W^n is based on periodized wavelets on $[0,1]$, i.e. for $j \ge 0$ and $0 \le k < 2^j$, $\psi_{j,k}$ and $\phi_{j,k}$ are replaced by

$$\psi_{j,k}^{per}(x) := \sum_{i \in \mathbb{Z}} \psi_{j,k}(x+i) \text{ and } \phi_{j,k}^{per}(x) := \sum_{i \in \mathbb{Z}} \phi_{j,k}(x+i).$$

Thus we have

$$w_{j,k,i}^n = \langle \psi_{j,k}^{per}, \phi_{m,i}^{per} \rangle.$$

Since for m large enough $\phi_{m,i}^{per}(x) = \phi_{m,i}(x - [x])$ and in the construction of each $\psi_{\cdot,\cdot}^{per}$ only at most N wavelets are involved, the same bounds as in (4.2) hold, but with C replaced by a constant C_2 depending on N and C. Hence

$$\sup_{k,i} |w_{j,k,i}^n| \le 2^{(j-m)/2} N \|\psi\|_\infty + C_2 2^{(j-m)(\beta+1/2)} \le C_3 2^{(j-m)/2}, \tag{4.3}$$

where $C_3 := C_2 + N\|\psi\|_\infty$. Now

$$\begin{aligned} \Lambda_n &= \inf_{(\lambda_{j,k}) \in \mathbb{R}^n} \sup_{\theta \in \mathbb{R}^n} \frac{\sum_{j,k} E|T_{\lambda_{j,k}}^S(Y_{j,k}) - \theta_{j,k}|^2}{\sigma^2 + B_n(\sigma^2, \theta)} \\ &\ge \inf_{(\lambda_{j,k}) \in \mathbb{R}^n} \max \left(\frac{\sum_{j \le h,k} E T_{\lambda_{j,k}}^S(z_{j,k})^2}{\sigma^2}, \frac{\sum_{j \le h,k} (\lambda_{j,k}^2 + \sigma^2)}{(2^{h+1}+1)\sigma^2} \right), \end{aligned} \tag{4.4}$$

where we get the lower bound by plugging in the following values for θ:

$$\theta_{j,k} = 0 \text{ for all } j,k \text{ respectively } \theta_{j,k} = \begin{cases} \infty & : j \le h \\ 0 & : \text{else} \end{cases}.$$

From now on we will only be concerned with the levels $0, \dots, h$. If in the rest of the proof (j,k) is an index then implicitly $j \le h$. To bound the first term in the maximum in (4.4) we will use the following lemma (see [31, Lemma 8.1]):

Lemma 4.2 *Let (X_i) be a finite sequence of independent mean zero real random variables such that $\|X_i\|_\infty \le c$ for all i. Then for every $\gamma > 0$, there exist positive numbers $K(\gamma)$ (large enough) and $\varepsilon(\gamma)$ (small enough) such that for every t satisfying $t \ge K(\gamma)b$ and $tc \le \varepsilon(\gamma)b^2$ where $b = \left(\sum_i EX_i^2\right)^{1/2}$,*

$$P\left(\sum_i X_i > t\right) \ge \exp(-(1+\gamma)t^2/2b^2).$$

Let $\gamma > 0$, then $\|e_k\|_\infty \le a < \infty$ for all k. Thus

$$\sup_{j \le h,k,i} \|w^n_{j,k,i} e_i\|_\infty \le C_3 a 2^{(h-m)/2}$$

and $\sum_i E(w_{j,k,i}e_i)^2 = \sigma^2$. Thus for $t \ge K(\gamma)\sigma$ and $t \le \dfrac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 a}$ we have

$$P(|z_{j,k}| > t) \ge 2\exp(-(1+\gamma)t^2/(2\sigma^2)).$$

Further note that $ET^S_{\lambda_{j,k}}(z_{j,k})^2 \ge P(|z_{j,k}| \ge \lambda_{j,k} + 1)$. Also, because of the uniform convergence in the CLT, (see [37] p. 149, 5.4) there exists an $n_0 = n_0(K(\gamma))$ such that for all $n \ge n_0$, $P(|z_{j,k}| \ge K(\gamma)+1) \ge (1-\Phi(K(\gamma)+1))/2$ where Φ is the normal distribution function.

Now let $(\lambda_{j,k,n})$ be one set of optimal thresholds for the right side in (4.1) and $0 < \alpha < 1$, then one of the following conditions holds (we always assume $j \le h$!)

$$\left|\left\{\lambda_{j,k,n} \in (K(\gamma)\sigma, \frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 a})\right\}\right| \ge (1-\alpha)2^{h+1}, \tag{4.5}$$

$$|\{\lambda_{j,k,n} \le K(\gamma)\sigma\}| \ge \alpha 2^h, \tag{4.6}$$

or

$$\left|\left\{\lambda_{j,k,n} \geq \frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 a}\right\}\right| \geq \alpha 2^h. \tag{4.7}$$

Thus for $n \geq n_0$ we get a lower bound in the first case (4.5) for Λ_n of

$$\Lambda_n \geq (1-\alpha)\max\left(\frac{\sum_{j\leq h,k} P(|z_{j,k}| \geq \lambda_{j,k,n}+1)}{\sigma^2}, \frac{\sum_{j,k}(\lambda_{j,k,n}^2+\sigma^2)}{(2^h+1)\sigma^2}\right).$$

At this point we apply the above lemma and get

$$\Lambda_n \geq (1-\alpha)\max\left(\sum_{j\leq h,k}\frac{\exp(-(1+\gamma)(\lambda_{j,k,n}+1)^2/(2\sigma^2))}{\sigma^2}, \frac{\sum_{j\leq h,k}(\lambda_{j,k,n}^2+\sigma^2)}{(2^{h+1}+1)\sigma^2}\right).$$

It is easy to see that the minimum on the right side is achieved if all the $\lambda_{j,k,n}$ are equal to the solution λ_n of

$$2^{h+1}\exp(-(1+\gamma)(\lambda_n+1)^2/(2\sigma^2)) = \frac{2^{h+1}(\lambda_n^2+\sigma^2)}{(2^{h+1}+1)}.$$

Now simple calculations like in the proof of Theorem 3.1 yield that λ_n behaves asymptotically like

$$\sqrt{\frac{\sigma^2 2\log 2^{h+1}}{(1+\gamma)}} \sim \sqrt{\frac{\sigma^2 2q\log n}{(1+\gamma)}},$$

where $\sim$ holds since $n^q/2 \leq 2^h \leq n^q$. This gives an asymptotic lower bound for Λ_n of $(1-\alpha)2q\log n/(1+\gamma)$.

With (4.4), the next two cases yield lower bounds for Λ_n of

$$\frac{\alpha 2^h}{\sigma^2}P(|z_{j,k}| \geq K(\gamma)+1) \geq \frac{\alpha 2^h(1-\Phi(K(\gamma)+1))}{2\sigma^2} \sim C_4 2^h \geq C_4 n^q/2$$

and respectively

$$\frac{\alpha 2^{h+1}}{(2^h+1)\sigma^2}\left(\frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 a}\right)^2 \sim C_5 2^{(m-h)} \geq C_5 n^{1-q}/2,$$

where C_4 and C_5 are constants. Thus the right-hand side of (4.4) grows as least as fast as $(1-\alpha)2q\log n/(1+\gamma)$.

Since α, q and γ were arbitrary, this gives a asymptotic lower bound for Λ_n of size about $2\log n$, i.e. $\liminf \Lambda_n/(2\log n) \geq 1$. □

5 Very smooth densities

In this section the conditions that the densities must satisfy are rather strict, but the point made is that, if one wants to beat soft thresholding, one has to use estimators which exploit the special structure of the densities involved. The reason is that one cannot beat thresholding for very smooth densities with compact support.

Theorem 5.1 *Let the observations $X_i = s_i + e_i$, $i = 1, \ldots, n = 2^m$ be given, where the s_i are the parameters of interest and the e_i are mean zero iid random variables with variance σ^2. and law μ having unimodal density f with support $[a, b]$. Further let $\sqrt{f} \in C^2([a, b])$, let $f \in C^k([a, b])$ $k \geq 2$, with $f^{(k)}(a) \neq 0$ and $f^{(k)}(b) \neq 0$, also for $\ell < k$, let $f^{(\ell)}(a) = f^{(\ell)}(b) = 0$. Apply a periodic wavelet transform W_n to these observations taking for every n the same type of wavelet. Let $Y = W_n(X)$ and let $\theta = W_n(s)$, then*

$$\liminf_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)} (\log n)^{-1} > 0,$$

where the infimum is taken over all estimators for θ.

The conditions on f and $\sqrt{f}$ are fulfilled for example if $f \in C^2([a, b])$ and for $x \geq 0$, $f(x + a) = c_1 x^2 + o(x^3)$ and $f(b - x) = c_2 x^2 + o(x^3)$.

The previous computations of the minimax bounds were based on the a-priori measure $F_{\varepsilon,a}$ and its Bayes risk, where ε and a were carefully chosen. In that context we were able to compute the Bayes estimator. This estimator was based on likelihood ratios. Here we want to apply the same scheme to coarse level wavelet coefficients. Again the Bayes estimator is based on likelihood ratios. We will compute an asymptotic expansion of these likelihood ratios.

Proof: Let λ be the Lebesgue measure, $P_0 = f\lambda$ and $P_h = P_0 * \varepsilon_h = f(x - h)\lambda$. Finally let

$$g := -2\frac{\sqrt{f}'}{\sqrt{f}} = \frac{-f'}{f}, \quad \text{on } (a, b),\ g = 0 \text{ elsewhere.}$$

Let $t_{n,i}$, $i = 1, \ldots, n$ denote a triangular array of numbers with $t_n := \sup_{i=1,\ldots,n} |t_{n,i}|$ and $t_n/n^c \to 0$ for all $c > 0$. Our goal is to prove that

$$\log\left(\prod_{i=1}^{n} \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} t_{n,i} g(x_i) - \frac{\|g\|_{P_0}^2}{2n}\sum_{i=1}^{n} t_{n,i}^2 + o_{P_0^n}(1),$$

where $o_{P_0^n}(1)$ denotes a sequence of random variables that converge in probability to 0, we will show that this stochastic convergence depends only on t_n and n. For simplicity and ease of notation we will assume $t_{n,i} \geq 0$. This will spare us some simple distinction of cases. Note that

$$E_{P_0}|g(x)|^{5/2} = \int_a^b \left(\frac{|f'(x)|}{f(x)}\right)^{5/2} f(x)dx = \int_a^b \frac{|f'(x)|^{5/2}}{f(x)^{3/2}}dx.$$

The restrictions on f imply that for $x \in (a, b)$ near a

$$f(x) = c_2(x-a)^k + o((x-a)^k)$$

and

$$f'(x) = kc_2(x-a)^{k-1} + o((x-a)^{k-1}),$$

where c_2 is a constant. Thus $\dfrac{|f'(x)|^{5/2}}{f(x)^{3/2}} = O((x-a)^{k-5/2})$, near a. A similar statement holds for f near b. Together this implies $E_{P_0}|g|^{5/2} < \infty$. Now let us turn to the likelihood quotient, define

$$h_{n,i} := 2\left(\sqrt{\frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}} - 1\right).$$

Let $p \in (2, 2+2/k)$, then

$$\int_{\mathbb{R}} |h_{n,i}|^p dP_0 = 2^p \int_a^b \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)}\right)^p f(x)^{1-p/2}dx.$$

Because of the differentiability of $\sqrt{f}$

$$\left|\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)}\right| = O(t_n/\sqrt{n}) \tag{5.1}$$

uniformly. We need $\int f(x)^{1-p/2}dx < \infty$. Because of the behavior of f near a and b

$$f(x+a)^{1-p/2} = O(x^{k-kp/2}) \text{ and } f(x+b)^{1-p/2} = O(x^{k-kp/2}).$$

Since $k - kp/2 > -1$, $E|h_{n,i}|^p < \infty$ and since by (5.1) $E|h_{n,i}|^p = O((t_n/\sqrt{n})^p)$. Note that it follows in the same way that $E|h_{n,i}|^2 = O(t_n^2/n)$. The next step is to show

$$\int_{-\infty}^{\infty} \left(h_{n,i}(x) - \frac{t_{n,i}}{\sqrt{n}}g(x)\right)^2 f(x)dx = O\left(\frac{t_n^2}{n^{3/2}}\right).$$

One fourth of the left side is equal to

$$
\begin{aligned}
&\int_a^b \left(\frac{\sqrt{f(x - t_{n,i}/\sqrt{n})}}{\sqrt{f(x)}} - 1 - \frac{t_{n,i}}{2\sqrt{n}} g(x) \right)^2 f(x)dx \\
&= \int_a^b \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} - \frac{t_{n,i}}{2\sqrt{n}} g(x)\sqrt{f(x)} \right)^2 dx \\
&= \int_a^b \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} + \frac{t_{n,i}}{\sqrt{n}} \sqrt{f(x)}' \right)^2 dx \\
&= \int_a^{a+t_{n,i}/\sqrt{n}} \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} + \frac{t_{n,i}}{\sqrt{n}} \sqrt{f(x)}' \right)^2 dx \\
&\quad + \int_{a+t_{n,i}/\sqrt{n}}^b (\ldots)^2 dx \\
&\le \int_{a+t_n/\sqrt{n}}^b \left(c_1 \frac{t_{n,i}}{\sqrt{n}} \right)^4 dx + 2\int_a^{a+t_n/\sqrt{n}} \left(\frac{t_n}{\sqrt{n}} \sqrt{f}' \right)^2 + \sqrt{f(x)}^2 dx, \\
&\qquad \text{since } \sqrt{f} \in C^2[a,b],\ c_1 > 0 \text{ is a constant} \\
&\le c_2 \left(\frac{t_n^4}{n^2} + \left(\frac{t_n}{\sqrt{n}} \right)^3 + \left(\frac{t_n}{\sqrt{n}} \right)^{k+1} \right), \quad \text{where } c_2 > 0 \text{ is a constant} \\
&\qquad \text{since } f(x+a) = O(x^k) \text{ and } \|\sqrt{f}'\|_\infty < \infty \\
&= O\left(\frac{t_n^3}{n^{3/2}} \right).
\end{aligned}
$$

As in Strasser ([38, p.379]), a simple Taylor expansion yields

$$
\begin{aligned}
&\log \left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i) \right) \\
&\quad = \sum_{i=1}^n 2\log\left(\frac{1}{2} h_{n,i}(x_i) + 1 \right) \\
&\quad = \sum_{i=1}^n h_{n,i}(x_i) - \frac{1}{4}\sum_{i=1}^n h_{n,i}(x_i)^2 + \frac{1}{4}\sum_{i=1}^n (1 - r(h_{n,i}(x_i)))h_{n,i}(x_i)^2,
\end{aligned}
$$

where $r(0) = 1$ and $|r(x_1) - r(x_2)| \le C|x_1 - x_2|$ if $|x_1| < 1$, $|x_2| < 1$, for some $C > 0$. We want to show that the last summand tends to 0 in probability, the rate depending only on t_n and n.

$$
\begin{aligned}
&P_0^n\left(\sum_{i=1}^n (1-r(h_{n,i}(x_i)))h_{n,i}(x_i)^2 \geq \varepsilon\right) \\
&\leq \quad P_0^n\left(\sup_i |1-r(h_{n,i}(x_i))| \geq \frac{1}{t_n^2 \log n}\right) + P_0^n\left(\frac{1}{t_n^2 \log n}\sum_{i=1}^n h_{n,i}(x_i)^2 \geq \varepsilon\right)
\end{aligned}
$$

Since $E_{P_0}|h_{n,i}|^2 = O(t_n^2/n)$,

$$
\frac{1}{t_n^2 \log n}\sum_{i=1}^n h_{n,i}(x_i)^2
$$

converges to zero in L^1. Further

$$
\begin{aligned}
P_0^n(t_n^2 \log n \sup_i |1-r(h_{n,i}(x_i))| \geq 1) \quad &\leq \quad \sum_{i=1}^n P_0\left(|1-r(h_{n,i}(x_i))| \geq \frac{1}{\log n t_n^2}\right) \\
&\leq \quad \sum_{i=1}^n P_0\left(|h_{n,i}(x_i)| \geq \frac{1}{C \log n t_n^2}\right) \\
&\leq \quad \sum_{i=1}^n E|h_{n,i}|^p (\log n)^p t_n^{2p} C^p \\
&= \quad O\left(\frac{t_n^{p3}(\log n)^p}{n^{p/2-1}}\right).
\end{aligned}
$$

Thus the uniform stochastic convergence is proved since $p > 2$.

Now we show that the terms $\sum_{i=1}^n h_{n,i}^2$ and $\sum_{i=1}^n Eh_{n,i}^2$ are stochastically equivalent with respect to P_0^n. This is equivalent to $\sum_{i=1}^n X_{n,i} \to 0$ in probability where $X_{n,i} = h_{n,i}^2 - Eh_{n,i}^2$. Note that

$$
E|X_{n,i}|^{p/2} = O(E(h_{n,i}^2)^{p/2} + (Eh_{n,i}^2)^{p/2}) = O\left(\frac{t_n^p}{n^{p/2}}\right).
$$

Let $\widetilde{X}_{n,i} := X_{n,i}\mathbf{1}_{\{|X_{n,i}|\leq 1\}}$, then $\widetilde{X}_{n,i}^2 \leq |X_{n,i}|^{p/2}$ and

$$
\sum_{i=1}^n E\widetilde{X}_{n,i}^2 \leq \sum_{i=1}^n |X_{n,i}|^{p/2} = O\left(\frac{t_n^p}{n^{p/2-1}}\right).
$$

Since $EX_{n,i} = 0$,

$$\begin{aligned}
\sum_{i=1}^{n} |EX_{n,i}\mathbf{1}_{\{|X_{n,i}|\le 1\}}| &\le \sum_{i=1}^{n} E|X_{n,i}|\mathbf{1}_{\{|X_{n,i}|>1\}} \\
&\le \sum_{i=1}^{n}\left(1P(|X_{n,i}|\ge 1) + \int_1^\infty \frac{E|X_{n,i}|^{p/2}}{x^{p/2}}dx\right) \\
&= O\left(\frac{t_n^p}{n^{p/2-1}}\right).
\end{aligned}$$

Now

$$P_0^n\left(\left|\sum_{i=1}^{n} X_{n,i}\right| \ge \varepsilon\right) \le \sum_{i=1}^{n} P_0^n(|X_{n,i}| > 1) + P_0^n\left(\left|\sum_{i=1}^{n}\widetilde{X}_{n,i}\right| \ge \varepsilon\right),$$

since

$$\sum_{i=1}^{n} P_0^n(|X_{n,i}|\ge 1) \le \sum_{i=1}^{n} E|X_{n,i}|^{p/2} = O\left(\frac{t_n^p}{n^{p/2-1}}\right),$$

and

$$\begin{aligned}
\sum_{i=1}^{n} P\left(|\widetilde{X}_{n,i}|\ge\varepsilon\right) &\le \frac{E\left(\sum_{i=1}^n \widetilde{X}_i\right)^2}{\varepsilon^2} \\
&\le \frac{E\sum_{i=1}^n E\widetilde{X}_{n,i}^2 + \left(\sum_{i=1}^n |E\widetilde{X}_{n,i}|\right)^2}{\varepsilon^2} \\
&= O\left(\frac{t_n^p}{n^{p/2-1}}\right),
\end{aligned}$$

we obtain that $\sum_{n,i} X_{n,i}$ converges to zero uniformly, i.e. the rate depends only on t_n and n. Thus we now have

$$\begin{aligned}
\log\left(\prod_{i=1}^{n}\frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}\right) &= \sum_{i=1}^{n} h_{n,i} - \frac{1}{4}\sum_{i=1}^{n} h_{n,i}^2 + o_{P_0^n}(1) \\
&= \sum_{i=1}^{n}(h_{n,i} - E_{P_0}h_{n,i}) - \frac{1}{4}\sum_{i=1}^{n} Eh_{n,i}^2 + \sum_{i=1}^{n} E_{P_0}h_{n,i} + o_{P_0^n}(1).
\end{aligned}$$

Proceeding as in Strasser ([38, p.381 top]), gives

$$E_{P_0}h_{n,i} = -\frac{1}{4}E_{P_0}h_{n,i}^2 - P_{t_{n,i}/\sqrt{n}}(N_{n,i}),$$

where $N_{n,i} = \left\{ \frac{dP_0}{dP_{t_{n,i}/\sqrt{n}}} = 0 \right\}$. Hence

$$\log\left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right)$$
$$= \sum_{i=1}^n (h_{n,i} - E_{P_0} h_{n,i}) - \frac{1}{2}\sum_{i=1}^n E_{P_0} h_{n,i}^2 - \sum_{i=1}^n P_{t_{n,i}/\sqrt{n}}(N_{n,i}) + o_{P_0^n}(1).$$

Like in Lemma 74.3. in [38] it follows from

$$\sum_{i=1}^n E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) - h_{n,i}(x_i)\right)^2 = O\left(n\frac{t_n^3}{n^{3/2}}\right) \to 0,$$

that

$$\sum_{i=1}^n \left(E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i)\right)^2 - E_{P_0} h_{n,i}(x_i)^2\right)$$
$$= \sum_{i=1}^n E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) - h_{n,i}(x_i)\right)\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) + h_{n,i}(x_i)\right)$$
$$\leq \sum_{i=1}^n \sqrt{E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) - h_{n,i}(x_i)\right)^2}\sqrt{E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) + h_{n,i}(x_i)\right)^2}$$
$$\leq \sqrt{\sum_{i=1}^n E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) - h_{n,i}(x_i)\right)^2 \sum_{i=1}^n E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) + h_{n,i}(x_i)\right)^2}$$
$$= O\left(\sqrt{n\left(\frac{t_n^3}{n^{3/2}}\right)(nt_n^2/n)}\right) = O\left(\frac{t_n^{5/2}}{\sqrt[4]{n}}\right)$$

and since $E_{P_0} g(x_i) = 0$

$$E_{P_0^n}\left(\sum_{i=1}^n (h_{n,i}(x_i) - E_{P_0} h_{n,i}) - \frac{t_{n,i}}{\sqrt{n}} g(x_i)\right)^2$$
$$= \sum_{i=1}^n E_{P_0^n}\left(h_{n,i}(x_i) - \frac{t_{n,i}}{\sqrt{n}} g(x_i) - E_{P_0} h_{n,i}\right)^2$$
$$\leq \sum_{i=1}^n E_{P_0^n}\left(h_{n,i}(x_i) - \frac{t_{n,i}}{\sqrt{n}} g(x_i)\right)^2$$
$$= O\left(\frac{t_n^3}{\sqrt{n}}\right).$$

Hence

$$
\begin{aligned}
&\log\left(\prod_{i=1}^{n}\frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right)\\
&= \sum_{i=1}^{n}\frac{t_{n,i}}{\sqrt{n}}g(x_i)-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{t_{n,i}}{\sqrt{n}}\right)^2 E_{P_0}g(x_i)^2-\sum_{i=1}^{n}P_{t_{n,i}/\sqrt{n}}(N_{n,i})+o_{P_0^n}(1).
\end{aligned}
$$

where the stochastic convergence depends only on t_n and n. Since

$$
\begin{aligned}
P_{t_{n,i}/\sqrt{n}}(N_n) &= \int_b^{b+t_{n,i}/\sqrt{n}} f\left(x-\frac{t_{n,i}}{\sqrt{n}}\right)dx\\
&= \int_0^{t_{n,i}/\sqrt{n}}\frac{f''(b)}{2}x^2+o(x^2)dx = O\left(\frac{t_n^3}{n^{3/2}}\right),
\end{aligned}
$$

it follows

$$
\log\prod_{i=1}^{n}\frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}t_{n,i}g(x_i)-\frac{E_{P_0}g^2}{2}\sum_{i=1}^{n}t_{n,i}^2+o_{P_0^n}(1),
$$

and again the stochastic convergence of $o_{P_0^n}(1)$ to 0 depends only on n and t_n. We now have an asymptotic expansion for the likelihood ratios which is needed to compute Bayes risks. Again we will consider wavelet coefficients at the level $l=(\log_2 n)/2$, the coefficients in the middle of the pyramid of the wavelet coefficients. The Bayes measure for each $\theta_{j,k}$, $j\neq l$ is δ_0, for $j=l$ it is F_{ε_n,a_n} with $\varepsilon_n\to 0$ and $a_n\to\infty$, the exact values will be specified later. The overall Bayes measure is the product measure of the individual Bayes measures. First we compute the Bayes risk for a single coefficient $\theta_{l,k}$. Let $z=W(e)$, and $\widetilde{\theta}_{j,k}$ random variables which are distributed as described by the Bayes measure, independent of $e.$. The Bayes estimator for the coefficient with index l,h is

$$
E(\widetilde{\theta}_{l,h}|z_{j,k}+\widetilde{\theta}_{j,k}, j=0,\dots,m-1;k=1,\dots,2^j),
$$

and the Bayes risk for estimating is (the expectation is for the noise and the Bayes measure simultaneously)

$$
\begin{aligned}
&E(E(\widetilde{\theta}_{l,h}|z_{j,k}+\widetilde{\theta}_{j,k}, j=0,\dots,m-1;k=1,\dots,2^j)-\widetilde{\theta}_{l,h})^2\\
&\geq E(E(\widetilde{\theta}_{l,h}|z_{j,k}+\widetilde{\theta}_{j,k}, j=0,\dots,m-1;k=1,\dots,2^j;\ \widetilde{\theta}_{j,k},(j,k)\neq(l,h),\\
&\qquad j=0,\dots,m-1;k=1,\dots,2^j)-\widetilde{\theta}_{l,h})^2\\
&= E(E(\widetilde{\theta}_{l,h}|z_{l,h}+\widetilde{\theta}_{l,h}, z_{j,k},(j,k)\neq(l,h), j=0,\dots,m-1;k=1,\dots,2^j)-\widetilde{\theta}_{l,h})^2\\
&\geq E(E(\widetilde{\theta}_{l,h}|e.+\widetilde{\theta}_{l,h}c_{l,h,\cdot})-\widetilde{\theta}_{l,h})^2,
\end{aligned}
$$

The last inequality holds since $\sum_i (e_i + \widetilde{\theta}_{l,h} c_{l,h,i}) c_{l,h,i} = z_{l,h} + \widetilde{\theta}_{l,h}$ and for $(j,k) \neq (l,h)$, $\sum_i (e_i + \widetilde{\theta}_{l,h} c_{l,h,i}) c_{j,k,i} = z_{j,k}$, where $(c_{j,k,i})$ are the coefficients of the wavelet transform (the wavelet transform is orthogonal!).
Simple computations now yield that

$$E(\widetilde{\theta}_{l,h} | e_{\cdot} + \widetilde{\theta}_{l,h} c_{l,h,\cdot}) = a_n \frac{\varepsilon_n dP_{a_n}}{\varepsilon_n dP_{a_n} + (1-\varepsilon_n) dP_0},$$

with P_0 being the distribution for $e.$ and dP_{a_n} the distribution for $e.+a_n(c_{l,h,\cdot})$. With the background of the proofs of Theorem 3.4 and 3.7 it is easy to see that when choosing $\varepsilon_n = \log(\sqrt{n})/\sqrt{n}$, the Bayes risk for this coefficient is larger than

$$(1-\varepsilon_n)^2 \varepsilon_n a_n^2 \int_{\mathbb{R}^n} \left(\frac{dP_0}{(1-\varepsilon_n) dP_0 + \varepsilon_n dP_{a_n}} \right)^2 dP_{a_n}.$$

Let p_1, C_1 be constants between 0 and 1. If the integrand is larger than C_1 with probability p_1 then the Bayes risk is larger than $(1-\varepsilon_n)^2 \varepsilon_n a_n^2 C_1 p_1$.

$$\begin{aligned} & P_{a_n} \left(\left(\frac{dP_0}{(1-\varepsilon_n) dP_0 + \varepsilon_n dP_{a_n}} \right)^2 \geq C_1 \right) \\ = \;& P_0 \left(\left(\frac{dP_{-a_n}}{(1-\varepsilon_n) dP_{-a_n} + \varepsilon_n dP_0} \right)^2 \geq C_1 \right) \\ = \;& P_0 \left(\left(1 - \varepsilon_n + \frac{\varepsilon_n dP_0}{dP_{-a_n}} \right)^2 \leq 1/C_1 \right) \\ \geq \;& P_0 \left(\frac{dP_{-a_n}}{dP_0} \geq C_2 \varepsilon_n \right), \end{aligned} \tag{5.2}$$

where $C_2 = 1/(1/\sqrt{C_1} - 1)$. Clearly the asymptotic properties of dP_{-a_n}/dP_0 and dP_{a_n}/dP_0 are the same, so instead we investigate when $P_0(dP_{a_n}/dP_0 > C_2 \varepsilon_n) > p_1$. We have $Y_{l,h} = \sum_{i=1}^n c_{l,h,i} X_i$. Since $l = (\log_2 n)/2$, only about $r = O(\sqrt{n})$ of the $c_{l,h,i}^n$ are non-zero (see relation (1.2)), thus (with some renaming, c_i for $c_{l,h,i}$ and Z_i for X_i.) $Y_{l,h} = \sum_{i=1}^r c_i Z_i$ with $\sup_i c_i^2 \leq q_1/r$, where q_1 is some global constant, as we already know from the proof of Theorem 4.1. Thus changing the mean in the $Y_{l,h}$ by a is equivalent to changing the mean in each Z_i by $c_i a$ (the inverse wavelet transform of $w_{l,h}$

equal to a and all the other coefficients $w_{j,k}$ zero is $(ac_{l,h,i})_{i\in\mathbb{N}}$, this follows from the orthogonality of the wavelet transform). Hence it follows from our preparations (with $t_{r,i} = a_n c_i \sqrt{r}$):

$$\frac{dP_{a_n}}{dP_0} = \exp\left(a_n U_n - \frac{a_n^2}{2}\gamma^2 + o_{P_0}(1)\right), \tag{5.3}$$

with $U_n = \frac{1}{\sqrt{r}} \sum_{i=1}^r (c_i\sqrt{r}) g(x_i)$ and $\gamma^2 = E_{P_0} g(x_i)^2$. Note that $\sum_{i=1}^r (\sqrt{r} c_i)^2 = r$. Because of the central limit theorem U_n converges in distribution to a $N(0,\gamma^2)$ distribution, i.e. $U_n = V_n + R_n$, where V_n is an $N(0,\gamma^2)$ distribution and R_n converges to 0 in probability. Since $E|g(X_i)|^{5/2}$, Theorem 5.8 in [37] which is about the uniform convergence in the central limit theorem ensures that this convergence only depends on $t_r := \sup_i |t_{r,i}|$ and r. Thus combining (5.2) and (5.3),

$$P_0\left(V_n + o_{P_0}(1) \geq \frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2}\right) > p_1$$

is needed. Let $\varepsilon_n := \log\sqrt{n}/\sqrt{n}$, and let a_n be the solution of

$$\frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2} + \gamma = \Phi^{-1}\left(\frac{1-p_1}{2}\right)\gamma, \tag{5.4}$$

where Φ is the error function. Now for some n_0 and all $n \geq n_0$

$$\begin{aligned}
&P_0\left(V_n + o_{P_0}(1) \geq \frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2}\right) \\
&\geq \ P_0\left(V_n \geq \frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2} + \gamma\right) - \frac{1-p_1}{2} \\
&= \ 1 - \Phi\left(\frac{\log(\varepsilon_n C_2)}{\gamma a_n} + \frac{a_n\gamma}{2} + 1\right) - \frac{1-p_1}{2} \\
&= \ p_1.
\end{aligned}$$

Simple computations yield that the solution of the equation (5.4) is

$$a_n = \sqrt{\left(\frac{1-\Phi^{-1}(\frac{1-p_1}{2})}{\gamma}\right)^2 - \frac{2\log(\varepsilon_n C_2)}{\gamma^2}} - \frac{1-\Phi^{-1}(\frac{1-p_1}{2})}{\gamma}$$

$$
\begin{aligned}
&= \sqrt{\frac{\log n}{\gamma^2} - 2\frac{\log C_2 + \log\log\sqrt{n}}{\gamma^2} + \left(\frac{1-\Phi^{-1}(\frac{1-p_1}{2})}{\gamma}\right)^2} \\
&\quad - \frac{1-\Phi^{-1}(\frac{1-p_1}{2})}{\gamma} \\
&\sim \frac{\sqrt{\log n}}{\gamma}
\end{aligned}
$$

Finally, applying the machinery of the proofs of Theorem 3.4 and 3.7 gives:

$$
\liminf_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta} \frac{E\|\theta - \widehat{\theta}\|^2}{\sigma^2 + \sum_{i=1}^n \min(\theta_i^2, \sigma^2)} \frac{\gamma^2}{\log n} > 0.
$$

□

6 Concluding remarks

In the previous sections, we examined the ideal estimation approach, compared it to soft thresholding and computed the optimal thresholds. The optimal threshold for soft thresholding is the solution of an integral equation involving only the distribution of the noise. If a closed form expression for this distribution is not at hand, then it is difficult to compute the optimal threshold.

Variations of Theorem 2.1, 2.5 and 3.1 also hold for related type of estimators, for example hard thresholding and the estimator T_λ^M of the introductory section. In each case, the size of the threshold parameter and the performance of the estimator are governed by the tail behavior of the noise distribution. The asymptotic performance is the same as for soft thresholding, as long the density satisfies the conditions of Theorem 3.1, i.e. that it decays like $\exp(-h(x))$, where h grows faster than x^ε, $\varepsilon > 0$.

The main advantage of the ideal estimator scheme is that no a-priori knowledge is required, the only heuristic is that the wavelet representation of the function is sparse. Another heuristic is, that the wavelet transform of the signal is not only sparse, but that in the finer levels it is sparser than in the coarser levels. For example, a discontinuity affects $O(1/2^j)$ of the coefficients at the level j. Further we also saw how the wavelet coefficients decay with the levels. So if we know in which smoothness class a function belongs, then it might be better to apply different thresholds to different levels.

It is also possible to apply the ideal estimator method only level-wise, but this is no longer a minimax scheme. The thresholds are a little bit too small, the sum of the risks at 0 of the coefficients is of size $\sim$ *constant* $\log^2 n$. In practice the higher levels are not thresholded. Note that this does not change the asymptotic performance of the method since after all only a smaller and smaller fraction of the wavelet coefficients is not thresholded. Another method, where the threshold for the k^{th} coefficient is always the same (no matter what the length of the input vector is), is possible. The thresholds depend on the coefficients. This is presented next in the normal case.

Theorem 6.1 *Let $Y_i = \theta_i + z_i$, $i = 1, \ldots, n$, where the z_i are iid normal random variables with mean zero and variance σ^2, and where Λ_n and $p(\cdot,\cdot)$ have their usual meaning (as in Theorem 2.1). Let $\widetilde{\lambda}_i$ be such that $p(\widetilde{\lambda}_i, 0) = 2\sigma^2/i$ and let*

$$\widetilde{\Lambda}_n = \sup_{\theta \in \mathbb{R}^n} \frac{E \sum_{i=1}^n |T^S_{\widetilde{\lambda}_i}(Y_i) - \theta_i|^2}{\sigma^2 + B_n(\theta, \sigma^2)}.$$

Then

$$\lim_{n\to\infty} \frac{\Lambda_n}{\widetilde{\Lambda}_n} = 1.$$

Proof: It is clear that if $p(\widetilde{\lambda}_i, 0)$ is decreasing in i, then $\widetilde{\lambda}_i$ is increasing. Our first step is to prove that $\widetilde{\lambda}_n / \sqrt{2 \log n\sigma^2} \to 1$. First note that

$$p(\lambda, 0) \geq \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\lambda+1}^{\lambda+2} \exp(-x^2/(2\sigma^2))dx \geq \exp(-(\lambda+3)^2/(2\sigma^2)),$$

for λ large enough. Also for x large enough,

$$\frac{1}{\sqrt{2\pi\sigma^2}} 2x^2 \exp(-x^2/(2\sigma^2)) \leq \exp(-(x-1)^2/(2\sigma^2)),$$

and by Mill's ratio (see [40, p.850]),

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_\lambda^\infty \exp(-x^2/(2\sigma^2))dx \sim \frac{1}{\sqrt{2\pi}} \frac{\sigma}{\lambda} \exp(-\lambda^2/(2\sigma^2)).$$

With these relations it follows that for large λ

$$\exp(-(\lambda+3)^2/(2\sigma^2)) < p(\lambda, 0) < \exp(-(\lambda-2)^2/(2\sigma^2)).$$

Let $q(\lambda) := p(\lambda, 0)$, q is increasing and its inverse q^{-1} behaves near 0 like $\sqrt{2\log(1/x)\sigma^2}$, i.e. $\lim_{x\to 0} q^{-1}(x)/\sqrt{2\log(1/x)\sigma^2} = 1$. Thus our claim follows: $\lim_{n\to\infty} \widetilde{\lambda}_n/\sqrt{2\sigma^2\log n} = 1$.
Now, let $\theta \in \mathbb{R}^n$, if $|\theta_j| > \sigma$, $1 \le j \le n$, then, since $p(\widetilde{\lambda}_j, \theta_j) \le p(\widetilde{\lambda}, \infty)$ (see the proof of Theorem 2.1)

$$\frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta)}{B_n(\theta, \sigma^2)} \le \frac{\sum_{i=1, i\ne j}^n p(\widetilde{\lambda}_i, \theta_i) + p(\widetilde{\lambda}_j, \infty)}{B_n(\theta, \sigma^2)}.$$

If $|\theta_j| < \sigma$ and $(\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i))/B_n(\theta, \sigma^2) \ge 1$ then, since $p(\widetilde{\lambda}, \theta_j) \le p(\widetilde{\lambda}, 0) + \theta_j^2$ (again see the proof of Theorem 2.1):

$$\begin{aligned}\frac{\sum_{i=1, i\ne j}^n p(\widetilde{\lambda}_i, \theta_i) + p(\widetilde{\lambda}_j, \theta_j)}{\sigma^2 + \sum_{i=1, i\ne j}^n \min(\theta_i^2, \sigma^2) + \theta_j^2} &\le \frac{\sum_{i=1, i\ne j}^n p(\widetilde{\lambda}_i, \theta_i) + p(\widetilde{\lambda}_j, 0) + \theta_j^2}{\sigma^2 + \sum_{i=1, i\ne j}^n \min(\theta_i^2, \sigma^2) + \theta_j^2} \\ &\le \frac{\sum_{i=1, i\ne j}^n p(\widetilde{\lambda}_i, \theta_i) + p(\widetilde{\lambda}_j, 0)}{\sigma^2 + \sum_{i=1, i\ne j}^n \min(\theta_i^2, \sigma^2)}.\end{aligned}$$

Thus if $\sup\limits_{\theta\in\mathbb{R}^n} \dfrac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} \ge 1$, it follows from the computations above that

$$\begin{aligned}\sup_{\theta\in\mathbb{R}^n} \frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} &\le \sup_{\theta\in\{0,\infty\}^n} \frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} \\ &= \sup_{J\subset\{1,\ldots,n\}} \frac{\sum_{i\in J} p(\widetilde{\lambda}_i, 0) + \sum_{i\in J^c}(\widetilde{\lambda}_i^2 + \sigma^2)}{\sigma^2(|J^c| + 1)} \\ &\le \sup_{J\subset\{1,\ldots,n\}} \frac{(2\log n + 2)\sigma^2 + |J^c|(\widetilde{\lambda}_n^2 + \sigma^2)}{\sigma^2(|J^c| + 1)} \\ &\sim \Lambda_n,\end{aligned}$$

since $\sum_{i=1}^n 1/i \le \log n + 1$ and $(\widetilde{\lambda}_n^2 + \sigma^2)/\sigma^2 \sim \Lambda_n \sim 2\log n$. $\square$

The above result can be transfered to other distributions, where the density has an exponential decay like in Theorem 3.1. The main point is to choose $\widetilde{\lambda}_i$ such that $\sum_{i=1}^n p(\widetilde{\lambda}_i, 0) \approx \Lambda_n\sigma^2$. We could also choose the thresholds level-wise, this would lead to thresholds of size $\sim \sqrt{2m}$ for the m^{th} level, but there is no big difference with the thresholds in Theorem 6.1. These thresholds are quite close to thresholds where the choice of thresholds is based on another minimax approach (see [3]).

The minimax bounds for thresholding depend on the tail behavior of the noise in the wavelet coefficients. If only a small fraction of the noise coefficients has a heavy tail, then the minimax bound will be large. This was exploited for the noise with compact support, there we only considered the topmost $\sqrt{n}$ coefficients. For a lower bound for the heavy tail noise, we considered only the finest scale wavelet coefficients which are about one half of all wavelet coefficients. The minimax result for the noise with very smooth density is not general but shows that, in the case of noise with bounded support, if one wants to do better than soft thresholding one has to exploit special properties of the noise distribution, maybe something like a running median (see [3]). It was noted earlier that the soft thresholding estimator is not admissible. In particular, the behavior for large values is bad, it seems that hard thresholding is a better choice there. Also there are other continuous shrinkers that tend to the identity function for large values. In a minimax framework it does not pay off to use such an estimator. Further we showed that in our context one cannot do asymptotically better than soft thresholding. Another reason for neglecting the performance of our estimator is that the risk is not measured directly but compared to benchmarks.

The results presented here have mainly an asymptotic nature. For Gaussian noise the thresholds $\lambda_n = \sqrt{2\log n}$ are asymptotically optimal, but especially for small n they are larger than the actual optimal thresholds, namely the solution of the equation (2.2):

$$2(n+1)\int_{\lambda}^{\infty}(x-\lambda)^2\Phi(dx) = \lambda^2 + \sigma^2.$$

(see also [16] and the tables below). The same is also true for noise with distributions like $\exp(-h(x))$ where the asymptotically optimal thresholds are of size $h^{-1}(\log(n))$, if h is a fractional polynomial. Although there are in general no closed form formulas for the thresholds, it is possible to compute numerical approximations quite easily. The results of some computations, using Mathematica, to obtain the optimal thresholds, are presented in Figure 5. The type of noise distributions considered are the normal distribution, the Laplace distribution and the distribution with the density $c_1\exp(-c_2\sqrt{|x|})$. Additionally the optimal thresholds for distributions with polynomial decay,

$n \to$	32	128	512	2048	65536	2^{24}	2^{32}
ϕ	1.28	1.67	2.04	2.40	3.22	4.35	5.31
$\exp(-\lvert x \rvert)$	1.58	2.19	2.85	3.55	5.43	8.70	12.15
$\exp(-\sqrt{\lvert x \rvert})$	2.18	3.26	4.60	6.21	11.6	24.2	42.1
$\mathbf{1}_{-1,1}(x)$	1.04	1.26	1.42	1.53	1.66	1.72	1.74
$1/x^{10}+1$	1.11	1.40	1.68	1.99	2.96	5.53	10.3
$1/(x^4+1)$	1.99	3.28	5.30	8.46	27.0	171	1088
$1/((x+20)^4+1)$	2.78	4.67	7.64	12.3	40.0	256	1625
$1/((x+0.1)^4+1)$	2.08	3.44	5.57	8.91	28.5	181	1149
$999\phi(x)+1/(x+1)^4$	1.28	1.67	2.05	2.42	3.51	22.0	141
$99\phi(x)+1/(x+1)^4$	1.29	1.69	2.10	2.57	7.23	47.9	306
$9\phi(x)+1/(x+1)^4$	1.37	1.93	2.89	4.82	16.2	105	666

Figure 5: The optimal thresholds for some densities

the uniform distribution and some mixtures of them are also given. All distributions are scaled to have variance one.

In the table the densities are labeled only by the functional part of the densities, i.e. the actual density is the functional part scaled such that it has variance 1. For example, $\exp(-x^2)$ is a functional part of the normal density. The rational for a maximal $n = 2^{32}$ is that most of today's (1999) computers are not able to work with datasets larger than 2^{32} (32 bit address bus). Figure 6 and 7 show the ratio of the asymptotic threshold term and the optimal threshold for the normal distribution and the Laplace distribution. On the horizontal axis the base 2 logarithm of n are plotted. As one sees the asymptotic works very slowly.

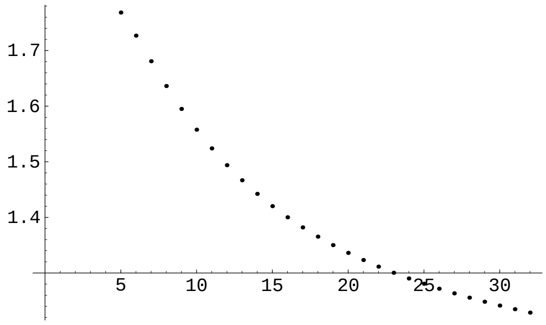


Figure 6: Ratio of asymptotic threshold and optimal threshold for the Normal distribution

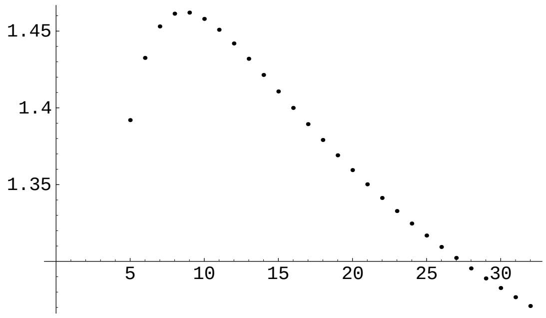


Figure 7: Ratio of asymptotic threshold and optimal threshold for the Laplace distribution

For non-Gaussian noise the distribution of the noise in the upper levels of the transformed signal is difficult to compute. A consequence of Theorem 2.5 and 2.6 was to use thresholds based on the initial noise. If this noise is heavy tailed than the thresholds for the upper levels are higher than actually necessary. In a minimax approach it does not really matter. As it will be seen in [3], and depending on the noise, from some level on upward the thresholds can be chosen as in the normal case. Also for the lower levels, the thresholds based on the initial noise is also often to high. For example, let X be a random variable with a Laplace distribution with $EX^2 = 1$ and λ_n such that $(1+n)p_X(\lambda, 0) = \lambda^2 + 1$. Let Y be an independent copy of X and $\widetilde{\lambda}_n$ be the solution of $(1+n)p_{(X+Y)/\sqrt{2}}(\lambda, 0) = \lambda^2 + 1$. Then for $n = 512$, $\lambda_n = 1.99$ and $\widetilde{\lambda}_n = 1.81$ and for $n = 2^{16}$, $\lambda_n = 4.1$ and $\widetilde{\lambda}_n = 3.55$. So it is reasonable to expect that the optimal thresholds for higher level wavelet coefficients are much smaller than the optimal thresholds for the Laplace distribution.

Unfortunately, it is very difficult to compute the optimal threshold (in the sense of Theorem 2.1) for convolutions. After all we have to solve

$$(n+1)\int_{|x|>\lambda} (|x| - \lambda)^2 \mu(dx) = \lambda^2 + \sigma^2;$$

and when using a numerical iterative method, this involves several computations of high dimensional integrals if μ is a convolution and the density of μ is not explicitly known.

In practice the variance of the data is usually not known, and so it has to be estimated. The ideal method is highly sensitive to the estimated variance. If the thresholds are chosen based on a variance estimation which is to small, then the absolute difference in risk to the thresholding based on a better variance estimation might be small for a certain signal, but the ratio of the risk and the benchmark might be much higher for the soft thresholding with a smaller threshold.

Clearly, the ideal estimator is not always the optimal estimator, indeed let X be a zero mean random variable with unit variance and let $x \to \alpha x$, $\alpha \in (0, 1)$, be a linear shrinker. Then $E(\alpha(X + a) - a)^2 = (1-\alpha)^2 a^2 + \alpha^2$. If $|a| = 1$ and $\alpha = 1/2$ then $((1-\alpha)^2 a^2 + \alpha^2)/(\min(a^2, 1)) = 1/2$. Also in some cases knowing the best threshold for soft thresholding can be better than the risk of the ideal method. However, it is easy to see that this ratio is bounded away from 0. Indeed, let $p(\cdot, \cdot)$ be the risk function for soft thresholding and some distribution μ with variance 1 and mean zero, where for simplicity we also assume that μ is symmetric about 0. Since $p(\lambda, a)$ is increasing in a:

$$\begin{aligned} c_\mu &:= \inf_{a\in\mathbb{R}} \inf_{\lambda} \frac{p(\lambda, a)}{\min(a^2, 1)} \\ &\geq \inf_{0\leq a\leq 1} \inf_{\lambda} \frac{p(\lambda, a)}{a^2} \\ &\geq \inf_{0\leq a\leq 1} \inf_{\lambda} \frac{a^2\mu(-\infty, \lambda - a) + \int_\lambda^\infty (x-\lambda)^2\mu(dx)}{a^2}. \end{aligned}$$

If $\lambda \geq 1/2$, then the last term is larger than $\mu((-\infty, -1/2))$, if $\lambda \leq 1/2$ then it is larger than $2\int_{1/2}^\infty (x - 1/2)^2\mu(dx)$. For example, for the normal distribution c_μ is larger than 0.7. Let Λ_n and λ_n now have there usual meaning in the sense of Theorem 2.1 with respect to the distribution μ. If $\widetilde{\lambda}_i$ $i = 1, \ldots, n$ is a set of thresholds, then

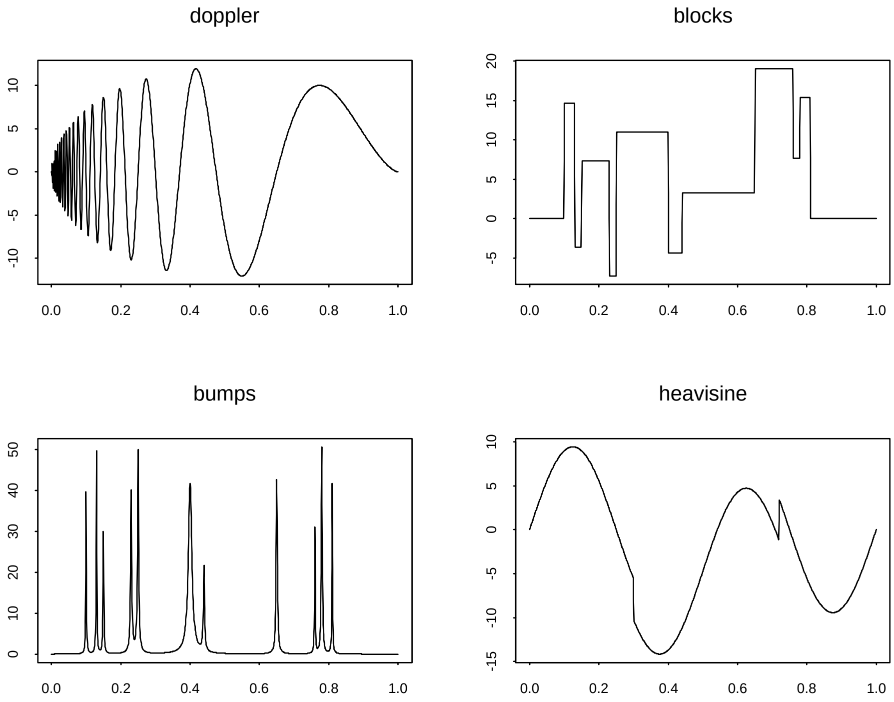
$$\sup_{a\in\mathbb{R}^n} \frac{\sum_{i=1}^n p_\mu(\lambda_n, a_i)}{1 + \sum_{i=1}^n p_\mu(\widetilde{\lambda}_i, a_i)} \leq \frac{\Lambda_n}{c_\mu}.$$

Λ_n/c_μ seems to be big, but if for example for $0 < \varepsilon_1 < 1$, n^{ε_1} of the a_i are larger than 1 then for any soft thresholding estimator the risk will be of size at least $O(n^{\varepsilon_1})$. If the thresholds are chosen too small or too large, then it can easily happen that the risk is of size $O(n^{\varepsilon_2})$ with $\varepsilon_2 > \varepsilon_1$. So if one wants to play safe there thresholds based on the ideal estimator method are a good choice.

As we already mentioned, the ideal method can also be used in other circumstances than wavelet thresholding. Any sparse estimation problem is a good candidate for this method. For example when using an experimental new estimator, all that has to be known is the variance of the noise in the coefficients. Often it is not possible to quantify the expectation of the sparseness of the signal.

A comparison of the performance of the different thresholds, and of the performance of soft thresholding for non-Gaussian noise is of interest. To do this, a small Monte-Carlo study was performed. The target signals are depicted in Figure 8, they were introduced by Donoho and Johnstone ([16]). Note that the functions blocks and heavysine are not continuous and hence not in a Besov space $B^m_{p,q}$ where $m > 1/p$. The simulation was performed

Figure 8: The 4 Signals of Donoho and Johnstone



with S+ from StatSci and the software package wavethresh for S+ from Guy Nason. As wavelet base the least asymmetric wavelets of Daubechies, with a filter length of 16 were chosen (see [10]). The lengths of the signal were chosen as 512 and 8192. The noise is respectively normal and Student with four degrees of freedom, scaled to have variance 1. The density of this Student distribution decays like $1/x^5$, so its tail is quite heavy. The thresholds used are the optimal thresholds of Theorem 2.1 for Gaussian noise, respectively for $n = 512$ and $n = 8192$. For each combination of noise, signal and signal length different thresholding methods were applied: all levels are thresholded, the first three levels are not thresholded and the first 5 levels are not thresholded, denoted by 0, 3 and 5 in the tables. The numbers in the tables are the means of the square errors for 100 runs, divided by the length of the signal. Clearly the decision which levels are not thresholded has a large influence on the performance of the estimator.

Optimal Gaussian thresholds

Signal length = 512

	Gaussian noise			Student noise		
signal	0	3	5	0	3	5
doppler	0.45	0.40	0.35	0.51	0.49	0.39
blocks	0.98	0.93	0.77	0.97	1.03	0.80
bumps	1.11	1.12	1.02	1.17	1.17	1.07
heavisine	0.24	0.18	0.15	0.34	0.22	0.21

Signal length = 8192

	Gaussian noise			Student noise		
signal	0	3	5	0	3	5
doppler	0.074	0.07	0.06	0.099	0.095	0.094
blocks	0.22	0.22	0.20	0.24	0.24	0.24
bumps	0.21	0.21	0.19	0.24	0.23	0.23
heavisine	0.046	0.043	0.034	0.074	0.064	0.068

Additionally the estimator of Theorem 6.1 was used, but the coefficients of one level were thresholded with the largest threshold for that level of the original estimator, i.e. the level j was thresholded with $\widetilde{\lambda}_{2^j+1}$ of Theorem 6.1.

Thresholds of Theorem 6.1

	Gaussian noise		Student noise	
signal	512	8192	512	8192
doppler	0.39	0.046	0.49	0.095
blocks	0.91	0.18	1.07	0.26
bumps	1.16	0.16	1.20	0.25
heavisine	0.17	0.028	0.23	0.070

Also for comparison the James-Stein estimator was applied level-wise to the wavelet coefficients. The first 5 levels were treated as one level.

$$T_L(x_1, \dots, x_n) := (x_1, \dots, x_n)\frac{(\|x\|_2^2 - \sigma^2(n+2))_+}{\|x\|^2}.$$

where σ^2 is the variance of the noise. This estimator tries to shrink the values with an estimate of the best linear shrinkage coefficient.

James-Stein estimator

	Gaussian noise		Student noise	
signal	512	8192	512	8192
doppler	0.55	0.077	0.58	0.078
blocks	0.75	0.26	0.74	0.26
bumps	0.91	0.14	0.92	0.14
heavisine	0.18	0.040	0.19	0.043

The result of this small study is that the performances of the estimator which does not threshold the first five levels and of the estimator of Theorem 6.1 are comparable. Surprisingly good is the performance of the James-Stein estimator, it is also more robust if the noise is not normal.

Other estimators, that apply a linear estimator to the wavelet coefficients (where some constants are chosen depending on the data) have been suggested. Because of the following result of Donoho and Johnstone ([17, Theorem 5]) these estimators might not be that useful.

Let $X_1, \ldots, X_n$ be iid centered normal random variables with variance σ^2, then for all $a \in \mathbb{R}^n$

$$\inf_{\alpha} E(\alpha(X+a)-a)^2 \geq E(T_L(X+a)-a)^2 - 2\sigma^2.$$

So what does this mean? Assume we are given a signal of length n, which is contaminated by iid normal noise, the noise has variance 1. Then the risk of any estimator which shrinks linearly each level of the wavelet transform of the data with a fixed coefficient is larger than the risk of the estimator which applies T_L level-wise minus $2\log_2 n$ (we have $\log_2 n$ levels). Note that this property is independent of the signal itself. The situation might change if the linear shrinkage coefficients are chosen dependent on the noisy wavelet transform, this is what T_L does.

To finish this study, various (self explanatory) simulations are presented.

References

[1] Antoniadis, A., Smoothing noisy data with tapered coiflets series. Scand. J. Statist. 23, 313-330 (1996).

[2] Averkamp R. and Houdré C., Wavelet thresholding for non (necessarily) Gaussian noise: A preliminary report. Spline functions and the theory of wavelets (Montreal, PQ 1996) CRM Proceedings and Lecture Notes **18**, AMS Providence, RI (1999), 347–354.

[3] Averkamp R. and Houdré C., Wavelet thresholding for non (necessarily) Gaussian noise: Functionality. Preprint (1999).

[4] Barikov, N.K., Extrema of the distributions of quadratic forms of Gaussian variables. Theory Probab. Appl. 34, (1989).

[5] Bruce, A.G. and Hong-Ye, G., WaveShrink with Semisoft Shrinkage. StaSci Research Report No. 39 (1995).

[6] Bruce, A. and Gao, H.Y. WaveShrink with firm shrinkage, Research Report 39, Statistical Science Division, MathSoft Inc. (1996).

[7] Bickel, P., Minimax estimation of the mean of a normal distribution subject to doing well at a point. in Recent Advances in Statistics (M. H. Rizvi, J. S. Rustagi, D. Siegmund, eds.) Academic Press Inc., New York (1983).

[8] Chambolle A., DeVore, R.A., Lee N.Y. and Lucier B.J., Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage. IEEE Transactions on Image Processing, Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis, 7, 319-335 (1998).

[9] Cohen, A., Daubechies, I. and Vial, P., Wavelets on the interval and fast wavelet transforms, Applied and Computational Harmonic Analysis 1, No.1, 54-81 (1993).

[10] Daubechies, I., Ten lectures on wavelets. No. 61 in CBMS-NSF in Applied Mathematics, Siam, Philadelphia (1992).

[11] DeVore, R.A. and Lucier, B., Fast wavelet techniques for near optimal image processing. 1992 IEEE Military Communications Conference, 2–12 (1992).

[12] Delyon, B. and Juditsky A., Estimating Wavelet Coefficients. Wavelets and Statistics, Lecture Notes in Statistics 103, Springer Verlag (1995).

[13] Delyon, B. and Juditsky A., On Minimax Wavelet Estimators. Applied Computational Harmonic Analysis 3, 215-228 (1996).

[14] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D., Density estimation via wavelet thresholding. Annals of Statistics 24 (1996).

[15] Donoho, D.L., De-noising by soft-thresholding. IEEE Trans. Inf. Theory 41, No. 3, 613-627 (1994).

[16] Donoho, D.L. and Johnstone, I.M., Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, No. 3, 425-455 (1994).

[17] Donoho, D.L. and Johnstone, I.M., Adapting to Unknown Smoothness via Wavelet Shrinkage. J. Amer. Statist. Assoc. 90, 1200-1224 (1995).

[18] Donoho, D.L. and Johnstone, I.M., Neo-classical minimax problems, thresholding and adaptive function estimation. Bernoulli 2, No. 1, 39-62 (1996).

[19] Efromovich, S.Y. and Pinsker, M.S., Estimation of a square integrable probability density of a random variable. Problems of Information Transition 18, (1982).

[20] Efron, B. and Morris, C., Limiting the risk of bayes and empirical bayes estimators – part i: The bayes case. J. Amer. Statist. Assoc. Vol. 66 (1971).

[21] Feller, W., An Introduction to Probability Theory and its Applications, Vol. II, John Wiley & Sons (1966).

[22] Gao, H.Y., Wavelet Estimation of Spectral Densities in Time Series Analysis. Ph.D. dissertation, University of California, Berkeley (1993).

[23] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A., Wavelets Approximation and Statistical Applications. Lecture Notes in Statistics 129, Springer Verlag (1998).

[24] Johnstone, I.M. and Silverman, B.W., Wavelet threshold estimators for data with correlated noise. J. R. Statist. Soc. B 59, No. 2, 319-351 (1997).

[25] Kerkyacharian, G. and Picard, D., Estimation de densite par methode de noyau et d'ondelettes: Les liens entre la geometrie du noyau et les contraintes de regularite. C. R. Acad. Sci., Paris, Ser. I 315, No. 1, 79-84 (1992).

[26] Kerkyacharian, G. and Picard, D., Density estimation by kernel and wavelets methods: Optimality of Besov spaces. Stat. Probab. Lett. 18, No. 4, 327-336 (1993).

[27] Kerkyacharian, G. and Picard, D., Density estimation in Besov spaces. Stat. Probab. Lett. 13, No.1, 15-24 (1992).

[28] Kolaczyk, E.D., Non-Parametric Estimation of Gamma-Ray Bursts Intensities Using Haar Wavelets. The Astrophysical Journal Vol. 483, 1997

[29] Kolacyk, E.D., A Method for Wavelet Shrinkage Estimation for Certain Poisson Intensity Signals Using Corrected Thresholds. Statistica Sinica 9, No. 1, 119-136 (1999).

[30] Kovac, A. and Silverman, B.W., Extending the scope of wavelet regression methods by coefficient-dependent thresholding. preprint (1999).

[31] Ledoux, M. and Talagrand, M., Probability in Banach Spaces. Springer Verlag (1991).

[32] Mallat, S., Multiresolution approximation and wavelets. Trans. Amer. Math. Soc. 315 (1989).

[33] Marshall, A.W. and Olkin I., Inequalities: Theory of Majorization and its Applications. Academic Press (1979).

[34] Meyer, Y., Ondelettes et opérateurs I: Ondelettes. Paris: Hermann, Éditeurs des Sciences et des Arts (1990).

[35] Meyer, Y., Wavelets, algorithms and applications. Translated from the French and with a foreword by Robert D. Ryan, SIAM, Philadelphia, PA (1993).

[36] Neumann, M.H. and Spokoiny, V.G., On the efficiency of wavelet estimators under arbitrary error distributions. Mathematical Methods of Statistics Vol. 4, No. 2 (1995).

[37] Petrov, V., Limit Theorems of Probability Theory. Clarendon Press Oxford (1995).

[38] Strasser, H., Mathematical Theory of Statistics. de Gruyter, New York, Berlin (1985).

[39] Sweldens, W., The lifting scheme: A construction of second generation wavelets. SIAM J. Math. Anal. 29, No.2, 511-546 (1998).

[40] Shorak, R.S. and Wellner J.A., Empirical Processes with Applications to Statistics, John Wiley & Sons (1986).

[41] Vidakovic, B., Statistical Modeling by wavelets, John Wiley & Sons (1999).

[42] Wang, Y., Function estimation via Wavelet Shrinkage for Long-Memory Data. Ann. Stat. 24, No. 2, (1996).

R. Averkamp
Institut für mathematische Stochastik
Freiburg University
Eckerstraße 1
79104 Freiburg, Germany

C. Houdré
Southeast Applied Analysis Center
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332, USA