

Wavelet Thresholding for Non (Necessarily) Gaussian Noise

Dissertation zur Erlangung des Doktorgrades
der Mathematischen Fakultät
der Albert-Ludwigs-Universität Freiburg i. Br.

vorgelegt von
Roland Averkamp

Dekan: Prof. Dr. W. Soergel

Referenten: Prof. Dr. L. Rüschemdorf
Prof. C. Houdré, Atlanta

Datum der Promotion: 27.10.99

Institut für Mathematische Stochastik
Universität Freiburg
Eckerstraße 1
D-79104 Freiburg i. Br.

Abstract

Soon after the discovery of orthonormal wavelets, in particular the compactly supported ones, these wavelets have been used for non-parametric function estimation. At first, only linear estimators were considered, but then non-linear shrinkage estimators of wavelet coefficients have also been studied.

In the literature, two main models are present. In the first model, the target functions are members of some smoothness class and for this class the minimax properties of estimators are investigated. This approach is facilitated by the fact that some smoothness spaces can naturally be described by norms of sequences of wavelet coefficients.

In the second approach the risk of an estimator is compared to the risk of an “ideal” estimator. This “estimator” is “ideal” because it has some knowledge of the wavelet coefficients of the function to estimate, so it is not really an estimator. The quality of estimation is then measured by the size of the ratio of the estimator's risk and the risk of the ideal estimator.

First both models have been mainly studied for Gaussian noise. Later the first model was investigated by others for other types of noises. The second model was investigated by Gao for non-Gaussian noise.

In this thesis I will consider both types of approaches for non-Gaussian noise. The content of this thesis is as follows: In the first two chapters I give a short introduction to wavelets and their use in non-parametric function estimation. The third chapter is about the ideal estimator approach for non-Gaussian noise. The fourth chapter deals with the function space approach: an addition to known results is obtained and the performance of wavelet thresholding for median filtered data is investigated. The subject of chapter 5 is an extension of Stein's unbiased risk estimation for general classes of infinitely divisible noise in the location model. Stein's unbiased risk estimate is the basis for a very adaptive thresholding estimator. The last chapter presents a comparison of the thresholds in the two approaches and a connection to kernel estimators.

Contents

| | | |
|----------|--|-----------|
| 1 | A short introduction to wavelets | 1 |
| 2 | Wavelets and non-parametric estimation | 9 |
| 3 | Donoho and Johnstone's oracle | 16 |
| 3.1 | The basics | 16 |
| 3.2 | Distributions with compact support | 34 |
| 3.3 | Very smooth densities | 37 |
| 3.4 | Conclusions | 45 |
| 4 | The function space approach | 54 |
| 4.1 | The moment conditions | 54 |
| 4.2 | Very heavy tails | 64 |
| 4.3 | Very thin tails | 71 |
| 5 | Unbiased risk estimation | 74 |
| 6 | Some last thoughts | 82 |
| 6.1 | Comparison of thresholds in the two approaches | 82 |
| 6.2 | Block thresholding and kernel estimators | 83 |

1 A short introduction to wavelets

In this chapter I want to give a short introduction to wavelets. A crude but plain definition is as follows: A wavelet is a function $\psi \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ with $\int_{\mathbb{R}} \psi(x)dx = 0$ such that the functions

$$\psi_{j,k}(x) := \sqrt{2^j} \psi(2^j x - k), \quad j, k \in \mathbb{Z},$$

form an orthonormal basis of $L^2(\mathbb{R})$. (In the sequel: if ψ is a function, then $\psi_{j,k}$ denotes the function $\sqrt{2^j} \psi(2^j \cdot - k)$.) Like the discrete Fourier transform, wavelets are a mean for decomposing functions in elementary building blocks. The Fourier transformation decomposes a periodic function into the building blocks $(\exp(ikx))_{k \in \mathbb{Z}}$, these building blocks are only localized in the frequency domain, i.e. the Fourier transforms of the building blocks are the measures $\delta_k(\cdot)$, which are concentrated around k . But for all c , $\int_{t-c}^{t+c} |\exp(ikx)|^2 dx = 2c$ for all k , i.e. the energy of the building block is uniformly distributed over the interval $[-\pi, \pi)$.

Whereas for wavelets the building blocks $\psi_{j,k}$ are localized in the frequency domain and in the time domain. The energy of $\psi_{j,k}$ is concentrated around $2^{-j}k$, i.e. $\int_{2^{-j}(k-c)}^{2^{-j}(k+c)} |\psi_{j,k}(x)|^2 dx \approx 1$ if c is “big” enough. A word on notation: $a \approx b$ means “about the same”, a/b is approximately 1 and $a_n \sim b_n$ means a_n/b_n tends to 1. For most wavelets $\widehat{\psi_{j,k}}$ (the Fourier transform of $\psi_{j,k}$) is concentrated around $2^j c_\psi$ and $-2^j c_\psi$, where c_ψ only depends on ψ . In fact many wavelets are first constructed in the frequency domain first. Figure 1 shows a Meyer wavelet and its Fourier transform. For this wavelet its Fourier transform is concentrated around 0.7 and -0.7 , i.e. $c_\psi = 0.7$.

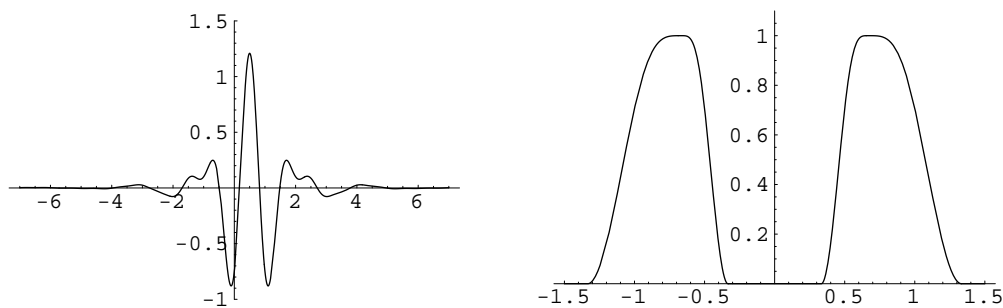


Figure 1: A Meyer wavelet and the modulus of its Fourier transform

Wavelets remind us of the windowed Fourier transform, which is discretized in time and frequency, i.e. the building blocks are $K(x - k) \exp(ijx)$ where $K(\cdot)$ is the window function. But for wavelets the size of the window changes with the frequency.

Thus one interpretation of $|\langle f, \psi_{j,k} \rangle|^2$ is “the amount of the L^2 norm of f contributed by the frequencies around $c_\psi 2^j$ at the time or place $k/2^j$ ”. Wavelet bases became practical and popular with the development of the multiresolution analysis by Mallat and Meyer ([34], [36]) and the construction of wavelets with compact support by Daubechies ([8]). Here we will mostly use compactly supported wavelets, since they have certain advantages over other kind of wavelets or wavelet-like decompositions which make them suited for statistical applications.

A multiresolution analysis is a increasing sequence of closed subspaces of $L^2(\mathbb{R})$

$$\dots V_{j-1} \subset V_j \subset V_{j+1} \dots \subset L^2(\mathbb{R})$$

such that there exists a function ϕ with $V_j = \overline{\text{sp}\{\phi(2^j x - k), k \in \mathbb{Z}\}}$, ϕ is called the scaling function or the father wavelet of the multiresolution analysis. The spaces $(V_j)_{j \in \mathbb{Z}}$ and the function ϕ have the following properties:

- $(\phi(\cdot - k))_{k \in \mathbb{Z}}$ is an orthonormal basis of V_0 , this implies that $(\sqrt{2^j} \phi(2^j x - k))_{k \in \mathbb{Z}}$ is an orthonormal basis of V_j .
- $\cap V_j = \emptyset$ and $\overline{\cup V_j} = L^2(\mathbb{R})$.

These conditions imply that the spaces V_j are dilations of each other, i.e. $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$. Further each V_j is invariant under translations of integer multiples of $1/2^j$. Now let $W_j = V_{j+1} \ominus V_j$. W_j can be viewed as the increment in information when going from an approximation in V_j to one in V_{j+1} , i.e. from coarser to finer approximation. From the properties of the (V_j) it easily follows that $L^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j$. Given these conditions on (V_j) and (W_j) , there exists a function $\psi \in L^2(\mathbb{R})$ such that $(\psi(\cdot - k))_{k \in \mathbb{Z}}$ is an orthonormal basis of W_0 . Since $L^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j$ the functions $\psi_{j,k}$ constitute an orthonormal basis for $L^2(\mathbb{R})$. Sometimes we will refer to the functions in W_j or V_j as the details of size 2^{-j} . If we talk about coarse levels, then we mean the wavelet and scaling coefficients corresponding to the spaces V_j or W_j where j is small, i.e. coefficients which describe the large details of a function. Analogously we mean by fine levels the coefficients for the spaces V_j and W_j with j large, i.e. the coefficients which describe the small details. Figure 2 might help to visualize this terminology.

Since $V_0 \subset V_1$, there are $(g_k)_{k \in \mathbb{Z}} \in \ell^2$ and $(h_k)_{k \in \mathbb{Z}} \in \ell^2$ such that

$$\phi(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} h_k \phi(2x - k)$$

and

$$\psi(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} g_k \phi(2x - k),$$

and moreover $g_k = (-1)^k h_{1-k}$. These relations are the scaling identities. The scaling function and the wavelet have the properties that $\int_{\mathbb{R}} \phi(x) dx = 1$ and that $\int_{\mathbb{R}} \psi(x) dx = 0$.

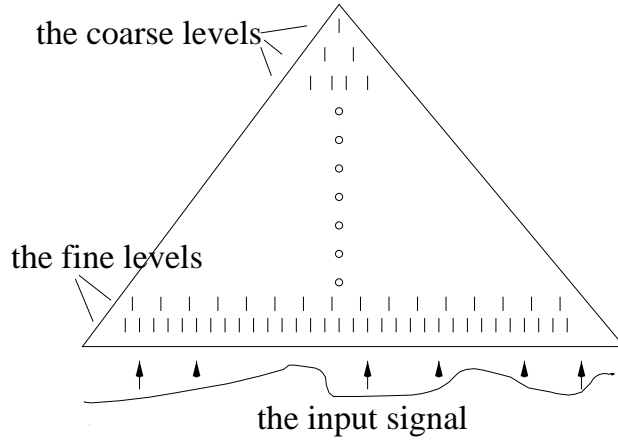


Figure 2: The discrete wavelet transform

With the multiresolution approach it is possible to construct smooth wavelets with exponential decay or wavelets which are piecewise polynomials of order k and in C^{k-1} (thus functions in V_0 are splines). Further there are efficient methods for computing the wavelet coefficients $\langle f, \psi_{j,k} \rangle$ for a given function f . The best known example of a wavelet basis based on a multiresolution analysis is the Haar basis. The scaling function is $\phi(x) = \mathbf{1}_{[0,1)}(x)$ and the wavelet is $\psi(x) = 1$ for $x \in [0, 1/2)$, $\psi(x) = -1$ for $x \in [1/2, 1)$, $\psi(x) = 0$ elsewhere. V_0 is the space of functions which are constant between the integers. Until recently this was the only known orthonormal wavelet basis with a compactly supported wavelet. In 1988, Daubechies published a method for constructing wavelets with compact support, built from a multiresolution analysis ([8]). There is a large family of wavelets with compact support, but there is 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 the base for 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 $[0, 1]$, we 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 $d_{j,k}$. We assume now that we are given the coefficients $c_{j_0,k}$ for some $j_0 > 0$. So 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 $d_{j_0-1,k}$:

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

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. We can repeat this scheme and at the level j only $2^j + N \approx n/2^{j_0-j}$ coefficients are non-zero. We will do this until we reach the level 0. To compute each coefficient we need only $O(N)$ time. Thus to compute all the coefficients $c_{j,k}$ and $d_{j,k}$ $0 \leq j \leq j_0 - 1$, $-N+2 \leq k \leq 2^j - 1$ we need $O(2^{j_0}) = O(n)$ time. This transformation is called the fast wavelet transform, see Mallat [34], 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 $W_{j_0-1}, \dots, W_{j_1}$ 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 $d_{j,k}$, $j \leq j_1$.

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

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

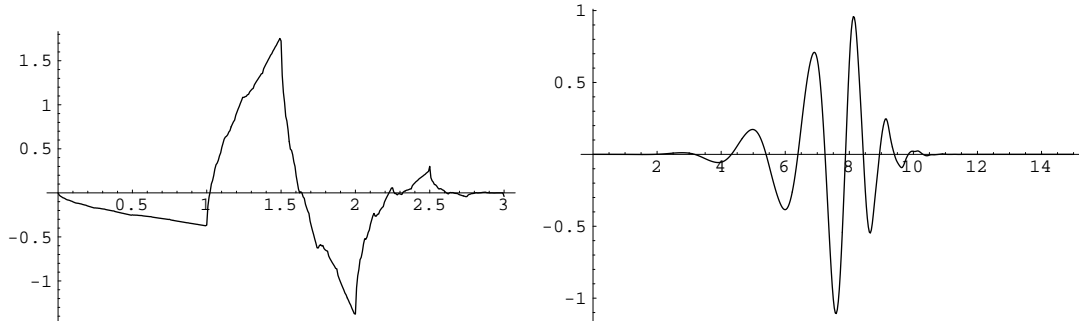


Figure 3: Two Daubechies wavelets

and

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

where u_{\cdot} and v_{\cdot} only depend on the filter (h). This claim about the length of the filters (u_{j_0-j}) and (v_{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.

Figure 3 shows two Daubechies wavelets, corresponding to discrete filters of respective length 4 and 16, thus their support is respectively of length 3 and 15. The first wavelet is Hölder continuous of order at least 0.5 and the second derivative of the second wavelet is Hölder continuous of order at least 0.4. For more information on the regularity of these wavelets see [9, ch.7]. The tradeoff between size of support and regularity is clearly visible.

The advantage of wavelet bases over other kinds of bases, like the Fourier basis or orthogonal polynomials, is that they are unconditional bases for a variety of function spaces. Often the norm of the function space is equivalent to a sequence norm of the wavelet coefficients (see [36]). First, by definition there is the space $L^2(\mathbb{R})$ itself as well as the other $L^p(\mathbb{R})$, $p > 1$ spaces. It is possible to decide $f \in L^p$ by only looking at the wavelet coefficients $d_{j,k}$ only. Although $L^1(0,1)$ does not have an unconditional basis, special adaptations of wavelet bases to the interval $[0,1]$ form a Schauder basis (see [9, p.304]). The Sobolev spaces

$$H_s = \{f \in L^2(\mathbb{R}) : \|\widehat{f}(\cdot) \sqrt{(1+\cdot^2)^s}\|_2 < \infty\}$$

$$(\text{=} \{f \in L^2(\mathbb{R}) : f \text{ } s \text{ times weakly differentiable, } \|f^{(s)}\|_2 < \infty\} \text{ if } s \in \mathbb{N})$$

can be characterized by

$$H_s = \left\{ f \in L^2(\mathbb{R}) : \sum_{j,k} |\langle f, \psi_{j,k} \rangle|^2 (1 + 2^{2js}) < \infty \right\}$$

if ψ and ϕ are in H_s . This is not too surprising if we remember that $\widehat{\psi_{j,k}}$ is concentrated around $c_\psi 2^j$ and take a look at Figure 1. A large class of function spaces are the Besov spaces. This class includes the Sobolev spaces. The definition of these spaces without wavelets is rather complicated. I will only give the elegant characterization of these spaces in terms of wavelet coefficients. A function f is in the Besov space $B_{p,q}^m$, $m > 0$, $1 \leq p, q \leq \infty$ if and only if

$$\left(\sum_k |\langle \phi_{0,k}, f \rangle|^p \right)^{1/p} + \left(\sum_{j \geq 0} \left(2^{j(m+1/2-1/p)} \left(\sum_k |\langle \psi_{j,k}, f \rangle|^p \right)^{1/p} \right)^q \right)^{1/q} < \infty \quad (3)$$

where ψ, ϕ are in C^r , $r > m$ with their derivatives up to order r rapidly decreasing, i.e. faster than any inverse polynomial. For our purposes the parameter q is not important. The parameter m characterizes mainly the decay of the ℓ^2 norm of the levels, while p characterizes the distribution of this ℓ^2 norm inside the levels. For example if $p = 1$, $a \in \mathbb{R}^n$ and $\|a\|_1 \leq 1$ then $\|a\|_2 \geq 1$ only if $a_k = 1$ for one k and $a_i = 0$ elsewhere, i.e. the ℓ^2 -norm is concentrated in one coefficient. This is different for $p > 2$, if $\|a\|_p = 1$ then $\|a\|_2 \geq n^{1/2-1/p}$ only if $a = (1/\sqrt[p]{n}, \dots, 1/\sqrt[p]{n})$, i.e. the ℓ^2 norm is uniformly distributed over all coefficients. It is clear that $B_{2,2}^m = H_m$. A path of Brownian motion is in $B_{p,\infty}^{1/2}$ almost surely for all $1 \leq p < \infty$, this is claimed in [23, p.105] (no proof there), it can be proved via the wavelet characterization of Besov spaces and some tedious but simple computations. The space of functions with bounded variation is a subset of $B_{1,\infty}^1$ and a superset of $B_{1,1}^1$, see [13]. Another example is the Bump Algebra, let $g_{t,s}(x) := \exp((x-t)^2/(2s^2))$, the set of functions with

$$f(x) = \sum_{i=1}^{\infty} a_i g_{t_i, s_i}(x), \quad \|a\|_1 < \infty,$$

is the space $B_{1,1}^1$ (see [36]). Note that for the spaces $B_{p,q}^m$ with $m < 1/p$, point values $f(t)$ are not well defined. For example if $m < 1/p$ and $a_{j,0} = 2^{-j/2}$ and $a_{j,k} = 0$ elsewhere, then $\sum_{j,k} (|a_{j,k}| 2^{j(m+1/2-1/p)})^q < \infty$ but $\sum_{j,k} a_{j,k} \psi_{j,k}$ does not converge at 0 if $\psi(0) \neq 0$. On the other hand, if $m > 1/p$ then $B_{p,q}^m \subset C^{m-1/p}$, if $m - 1/p$ is not an integer, this is a consequence of (3) and the results of chapter 9.2 in [9]. If $m - 1/p$ is an integer, then $B_{p,q}^m$ is a subset of a space slightly larger than $C^{m-1/p}$.

Wavelets are useful to study the local regularity of functions, see [9, p.300] for the characterization of local Hölder continuity. Further with wavelets one can

even characterize spaces of piecewise polynomials. If the length of the support of the wavelets is N , then at any level, only N coefficients are affected by a single node! So f is piecewise polynomial of order less than k and ψ is a wavelet whose first k moments vanish, then at any level at most N times the number of nodes wavelet coefficients are non-zero. The characterization of function spaces via a sequence norm of the wavelet coefficients, have in common, that the coefficients for the finer levels are more heavily weighted. This is another reason why the wavelet coefficients of smooth functions tend to zero quite fast.

The notion of wavelet we considered is rather strict, namely that for a function ψ the functions $(\psi_{j,k})$ form an orthonormal basis of $L^2(\mathbb{R})$. There are no compactly supported wavelets which are symmetric, this is considered to be a drawback in image analysis. A remedy for this deficiency are biorthogonal wavelet bases for $L^2(\mathbb{R})$, i.e. functions $\psi, \tilde{\psi} \in L^2(\mathbb{R})$ such that $(\psi_{j,k})$ and $(\tilde{\psi}_{j,k})$ are biorthogonal bases of $L^2(\mathbb{R})$ i.e.

$$f = \sum_{j,k} \langle f, \psi_{j,k} \rangle \tilde{\psi}_{j,k} = \sum_{j,k} \langle f, \tilde{\psi}_{j,k} \rangle \psi_{j,k}, \quad \forall f \in L^2(\mathbb{R}).$$

Further $\psi_{j,k}$ as well as the $\tilde{\psi}_{j,k}$ are unconditional bases of $L^2(\mathbb{R})$. It is possible to construct compactly supported biorthogonal wavelet bases from a multiresolution analysis. In this case there is a scheme similar to the fast wavelet transform, but in general it is not an orthogonal transformation. Even more general than biorthogonal wavelets are frames. The function $\psi \in L^2(\mathbb{R})$ induces a wavelet frame if for some $A, B > 0$,

$$A \sum_{j,k} \langle f, \psi_{j,k} \rangle^2 \leq \|f\|^2 \leq B \sum_{j,k} \langle f, \psi_{j,k} \rangle^2,$$

for all $f \in L^2(\mathbb{R})$. Then there is a $\tilde{\psi} \in L^2(\mathbb{R})$ with $f = \sum_{j,k} \langle f, \psi_{j,k} \rangle \tilde{\psi}_{j,k}$. ($\tilde{\psi}_{j,k}$) is called the dual frame of $(\psi_{j,k})$. The difference with biorthogonal wavelet bases, is that frames are not necessarily bases. It is quite easy to construct general wavelet frames, but the reconstruction of a function from its wavelet coefficients is more complicated.

Dropping the orthonormality and then the basis property gives greater freedom in the choice of the wavelet. One can tailor the wavelet to a specific application. But the advantage of orthonormal wavelet bases is that the orthogonality facilitates norm computations and it implies that the discrete wavelet transform is itself an orthonormal mapping.

The multiresolution analysis and wavelets are tools to construct orthonormal bases of $L^2(\mathbb{R})$, but in most cases one is only interested in functions on a bounded interval $[a, b]$ and the functions are not necessarily 0 at the interval boundaries. The simplest solution is to consider the function $f\mathbf{1}_{[a,b]}$ instead. But this leads to Gibbs phenomena at the interval boundaries, unwanted high frequencies are introduced into the signal.

In practice one extends f to the left of a by $f(a)$ and to the right of b by $f(b)$ or mirrors f at a and b . Another option is to use wavelets which are adapted to an interval ([9, p.333]). There are variants of the fast wavelet transform which are based on wavelet bases for compact intervals. These variants are orthonormal transformations from $\mathbb{R}^{2^{j_0}}$ to $\mathbb{R}^{2^{j_0}}$. An example is the discrete wavelet transform which is based on periodized wavelets. I will explain them for the interval $[0, 1]$. For $j > 0$ and $0 \leq k < 2^j$, $\psi_{j,k}$ and $\phi_{j,k}$ are replaced by

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

and the $\phi_{0,0}^{per}(\equiv 1)$ and $\psi_{j,k}^{per}$ are an orthonormal base for $L^2([0, 1])$. It is easy to see that the multiresolution properties are preserved. Unfortunately these periodized wavelets are adapted to periodic functions. If the function is not periodic, then the wavelet coefficients at the borders are bigger due to these discontinuities at the borders.

Another option are the wavelets on the interval described in [7]. These are better adapted to the boundaries of the intervals, but they are computationally more complicated. There are variants of the fast wavelet transform for these wavelet bases. These variants are orthonormal transformations from $\mathbb{R}^{2^{j_0}}$ to $\mathbb{R}^{2^{j_0}}$. The restriction on the size of the input to be a power of two is clearly a drawback. Note that we are not so much interested in the wavelet bases themselves but in the induced fast wavelet transform, i.e. a transformation for \mathbb{R}^n ; but this will become clearer in the next section.

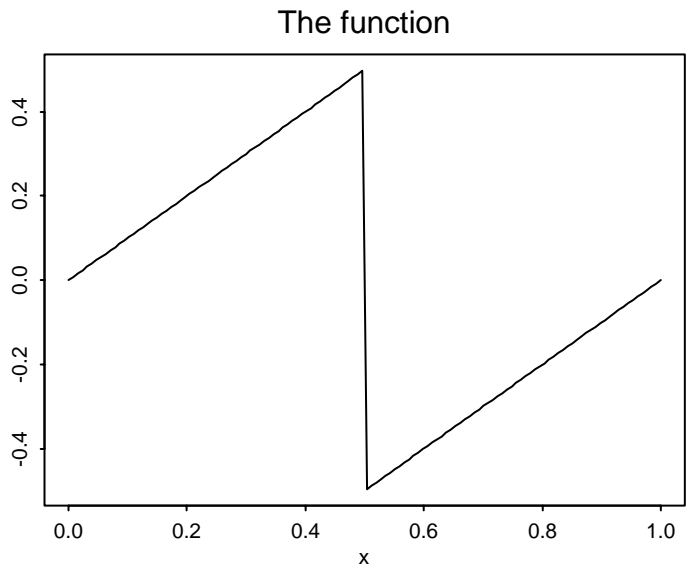
2 Wavelets and non-parametric estimation

An underlying idea in non-parametric curve/surface/signal estimation is that the function to estimate has some redundancy; 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 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 $\text{sign}(x)$. 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 consist 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 by various authors (see [25], [26], [27]).

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 scheme to Fourier coefficients (see [20], 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 $\psi_{j,k}$. Assuming that the support of ψ is $[0, N]$, there are only N affected coefficients at each level. So if we consider a wavelet basis adapted to an interval only about $\log_2(n)N$ of the first n coefficients are affected by a single discontinuity. Figure 4 shows this clearly.

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 inputted into the fast wavelet transform, i.e. noisy versions (denoted by $\tilde{c}_{j_0,k}$) 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 wavelet coefficients $d_{j,k}$ (denoted by $\tilde{d}_{j,k}$ and called the empirical wavelet coeffi-



Fourier and wavelet coefficients, ordered by size

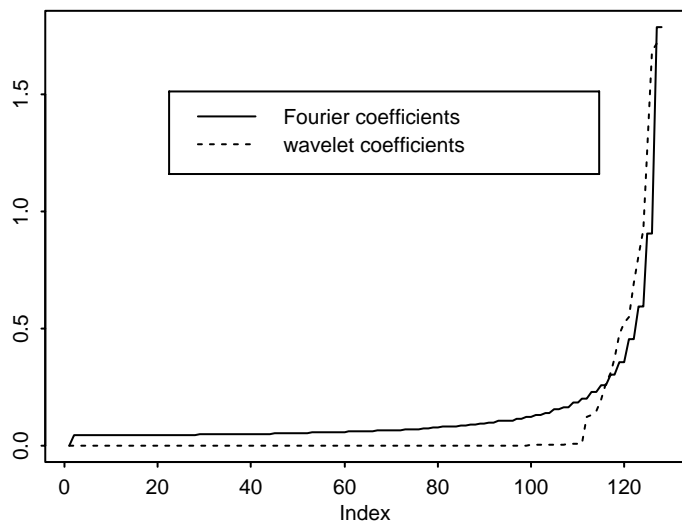


Figure 4: Comparison of the size of the absolute values of Fourier and wavelet coefficients

icients). 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 we do we 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, \dots, X_n , and has to estimate the Lebesgue density f of X_1 . The $\widetilde{c}_{j,k}$ respectively the empirical wavelet coefficients are

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

Note that $E\phi_{j,k}(X_i) = \int f(x)\phi_{j,k}(x)dx = \langle f, \phi_{j,k} \rangle = c_{j,k}$, thus $\widetilde{c}_{j,k} = c_{j,k} + e_{j,k}$ with $Ee_{j,k} = 0$ and $Var(e_{j,k}) = O(1/n)$. Remember that because of the fast wavelet transform we only have to calculate 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 a approach developed by Deylon and Juditsky for this case, see [11]. This approach has the drawback that the joint distribution of the $\widetilde{c}_{j_0,k}$ is complicated, making it hard to compute 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 . But this situation is very complicated. 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 $\widetilde{c}_{j_0,k}$ are given and of the form $\widetilde{c}_{j_0,k} = c_{j_0,k} + e_k$, where the e_k are iid random variables (thus assuming independent noise!). Although this might seem 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 $x_k = 2^{-j_0}k$, $k = 1, \dots, n = 2^{j_0}$, we have

$$c_{j_0,k} = 2^{j_0/2} \int_{\mathbb{R}} \phi(2^{j_0}x - k)f(x)dx \approx f(2^{-j_0}k)2^{-j_0/2} = f(x_k)2^{-j_0/2}, \quad (4)$$

since $\int_{\mathbb{R}} \phi(x)dx = 1$, 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 ([9, p.258]).

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., 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 “real” estimation errors. This scheme works well in practice and most other schemes for equidistant designs are not really needed.

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

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

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:

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

For the transformed data, we write $w_{j,k}$ rather than $d_{j,k}$, 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 $d_{j,k}$ s by the $c_{j_{top},k}$ s).

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 Gaussian 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 $\|W_n x\|_2^2 = \|x\|_2^2$, for $x \in \mathbb{R}^n$. Thus if $\hat{w}_{j,k}$ is an estimator for $w_{j,k}$ and \hat{f} the equivalent estimator for f then

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

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 $\tilde{w}_{j,k}$ 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; and 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. These two policies were studied by Donoho and Johnstone [14], [15], [16]. They correspond to respectively applying the operators

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

and

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

to the noisy wavelet coefficients (here and in the sequel $\text{sgn}(x) := x/|x|$ for $x \neq 0$ and $\text{sgn}(0) = 0$.) In both cases λ is called the threshold and it usually depends on the index (j, k) . Other shrinking schemes have been proposed (see [2], [6], [10]) but, all in all, the asymptotic performances of the different shrinking estimators do not vary much.

It can be shown that in the case of normal iid noise, soft and hard thresholding are not admissible (see [2] for a discussion of other shrinking schemes). This is not a surprise, if one considers the variety of possible shrinking estimators.

The Figures 5, 6, 7 show the soft, hard thresholding estimator and

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

with their respective risk functions for Gaussian noise. T_λ^M is a special case of the semisoft thresholding estimator of Bruce and Gao [3]. The parameter λ for each estimator was chosen such that for each estimator the risk $\int (T_\lambda(x+a) - a)^2\Phi(dx)$ at $a = 0$ is about 0.0004 where Φ is the standard Gaussian 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. Donoho and Johnstone were the first to apply nonlinear shrinking to wavelet coefficients, there is a variety of papers by them and others, especially on minimax estimation, see [6], [10], [14], [15], [25], [26] and [27].

Soft and hard thresholding were known before, although they were called differently for example in Bickel [4] and originate in a paper of Efron and Morris [19].

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.

Obviously 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 [40]). 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,

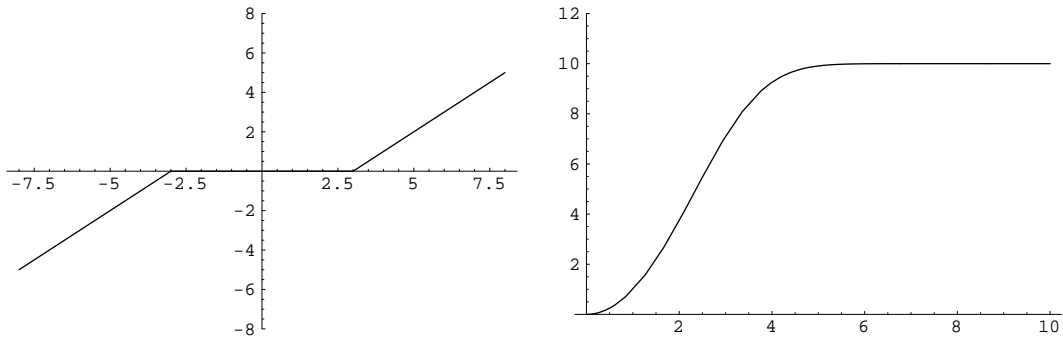


Figure 5: T_λ^S and its risk function for $\lambda = 3$

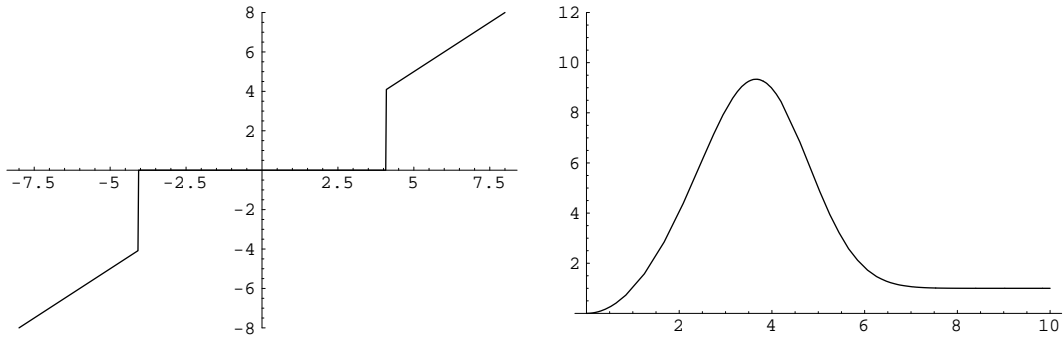


Figure 6: T_λ^H and its risk function for $\lambda = 4.1$

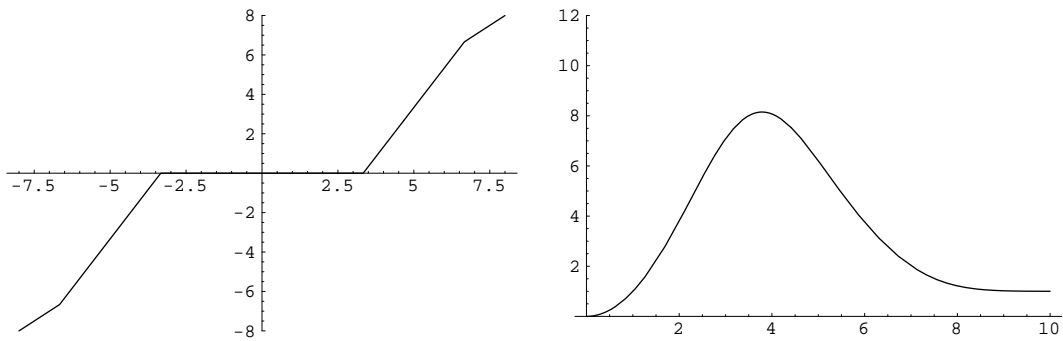


Figure 7: T_λ^M and its risk function for $\lambda = 3.3$

(see the dissertation of Gao [21]). 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 [21], [28], [29]).

3 Donoho and Johnstone's oracle

3.1 The basics

In the usual denoising setting by wavelet thresholding, one assumes that the target functions are members of a fixed subset, e.g. a ball of a smoothness class such as a Besov space which can be characterized by the sequence of the wavelet coefficients. For this target set of functions one then computes a minimax threshold and tries to prove that the risk for this threshold is close to the minimax risk in the class of all estimators ([6], [10], [14], [15], [17], [42]).

Another approach to denoising introduced by Donoho and Johnstone ([17]) 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. Of course, the mean square error of this estimator is the minimum of σ^2 and the square of the coefficient. Under the assumption of independent 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. In contrast to the smoothness class approach, this scheme does not try to minimize the maximal mean square error for a fixed class of functions, but 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.

We assume the regular design situation, i.e. we have the observations

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

To this data we apply a discrete wavelet transform $W_n : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 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 $Y_i = \theta_i + z_i$, $i = 1, \dots, n$ 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 $T_\lambda^S(x) = (|x| - \lambda)_+ \text{sgn}(x)$ when applied to Y) will be compared to the benchmark

$$B_n(\theta, \sigma^2) := \sigma^2 + \sum_{i=1}^n \min(\theta_i^2, \sigma^2), \quad (5)$$

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 W^n be a wavelet transform based on the wavelet ψ , then it follows from identity (4) that

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

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. The following result from multivariate decision theory is theirs and is crucial for their results.

Theorem 3.1 *Let $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, $n \geq 3$, where the θ_i are parameters of interest and the z_i are iid normal random variables with mean zero and variance σ^2 . Let $\lambda_n = \sigma\sqrt{2 \log n}$, then*

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

In his thesis, Gao [21] proves a similar result for iid random variables with exponential tails. In fact, a stronger result of this type holds for a wider class of distributions. But we still require the z_i to be identically distributed random variables.

Theorem 3.2 Let $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, $n \geq 4$, where the θ_i are parameters of interest and the z_i are identically distributed random variables. Their distribution μ is symmetric about 0 and $Ez_1^2 = \sigma^2$. Then, the equation

$$\left(2 \int_{\lambda}^{\infty} (x - \lambda)^2 \mu(dx)\right) (n + 1) = \lambda^2 + \sigma^2, \quad (6)$$

has a unique positive solution λ_n . Moreover,

$$\Lambda_n = \sup_{\theta \in \mathbb{R}^n} \frac{E\|T_{\lambda_n}^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} = \inf_{\lambda} \sup_{\theta \in \mathbb{R}^n} \frac{E\|T_{\lambda}^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)},$$

where $\Lambda_n = (\lambda_n^2 + \sigma^2)/(\sigma^2(1 + 1/n))$.

Proof: The proof is similar to the proof of Theorem 3.1 in [15]. For $\lambda, a \in \mathbb{R}$, set

$$\begin{aligned} p(\lambda, a) &:= E|T_{\lambda}^S(z_1 + a) - a|^2 \\ &= \int (\operatorname{sgn}(x + a)(|x + a| - \lambda)_+ - a)^2 \mu(dx). \end{aligned}$$

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

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

The function $p(\lambda, \infty) := \lim_{a \rightarrow \infty} p(\lambda, a) = \sigma^2 + \lambda^2$ is continuous and increasing on $[0, \infty)$, whereas $p(\lambda, 0) = 2 \int_{\lambda}^{\infty} (x - \lambda)^2 \mu(dx)$ is continuous and decreasing on the positive part of the support of μ , it is zero outside the support. 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 $\sup_{a \in \{0, \infty\}} \frac{p(\lambda, a)}{\sigma^2/n + \min(a^2, \sigma^2)}$ and $\Lambda_n = p(\lambda_n, \infty)/(\sigma^2(1 + 1/n))$ is equal to the minimum of this term.

We now claim that

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

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

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

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

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

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

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

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

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

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

$$\frac{p(\lambda_n, a)}{\sigma^2/n + a^2} \leq \frac{a^2 + p(\lambda_n, 0)}{\sigma^2/n + a^2}$$

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

Therefore we conclude,

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

It remains to show that if $n \geq 4$, then $p(\lambda_n, 0) \geq \sigma^2/n$.

Remember $p(\lambda_n, 0)/(\sigma^2/n) = (\sigma^2 + \lambda_n^2)(1 + 1/n)\sigma^2$, this implies, $p(\lambda_n, 0) \geq \sigma^2/n$ 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 the distribution $1/2(\delta_{-\sigma} + \delta_{+\sigma})$. Hence $p(\lambda, 0) \geq p_Y(\lambda, 0)$. But this implies that λ_n is greater or equal than the solution $\tilde{\lambda}_n$ of $p_Y(\lambda, 0)(n+1) = \lambda^2 + \sigma^2$. Thus it is enough to prove that for $n \geq 4$, $\tilde{\lambda}_n^2 \geq \sigma^2/n$. For $n = 4$, $\tilde{\lambda}_n$ is the solution of

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

It easily follows that $\tilde{\lambda}_4 = \sigma/2$. Thus $\tilde{\lambda}_4^2 = \sigma^2/4$ and since $\tilde{\lambda}_n$ is increasing $\lambda_n \geq \sigma^2/n$ for $n \geq 4$.

We now show that, λ_n is the optimal threshold, this is easy, since for $\lambda > 0$ we have

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

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

If the z_i have a distribution with compact support $[-a, a]$, then $\lim_{\lambda \rightarrow a} p(\lambda, 0) = 0$. Thus $\lambda_n \rightarrow a$ and $\Lambda_n \rightarrow a^2/\sigma^2 + 1$. In this case soft thresholding is almost as efficient as the ideal estimator.

Example: let z_1 have the distribution $1/2(\delta_{-\sigma} + \delta_{+\sigma})$, then $\lambda_n \rightarrow \sigma$ and $\Lambda_n \rightarrow \sigma^2/\sigma^2 + 1 = 2$. Thus soft thresholding is only twice as worse as the ideal estimator. For a uniform distribution the ratio is $5/2$. The ratio becomes worse if the size of the support is large in comparison to the variance.

Remark 3.3 *Of course Theorem 3.2 holds if μ is not symmetric. It was just notational convenience not to include this case. One just has to replace in formula (6)*

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

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 $(\lambda_{left}$ and $\lambda_{right})$ is the solution of

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

The proof is basically the same.

Remark 3.4 *The previous method can be applied to other loss functions, not just quadratic loss, we can replace it by any symmetric loss function h , which is increasing on the positive axis and 0 at 0. In the benchmark we have to replace σ^2 by $m_h := Eh(z_1)$ and $\sum_i \min(\theta_i^2, \sigma^2)$ by $\sum_i \min(h(\theta_i), m_h)$. The computations of the optimal thresholds do not change much, but the condition $n \geq 4$ has to be replaced, it depends not only on h , 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 not identical distributed random variables, the following result gives a good suggestion for a threshold and an upper bound for the ratio of risks, namely compute λ_n for each z_i separately, and choose as threshold the largest of these λ_n . Note that the benchmark $B_n(\theta, \sigma)$ is replaced by $\sum_{i=1}^n (\min(\theta_i^2, \sigma_i^2) + \sigma_i^2/n)$.

Theorem 3.5 *Suppose we have the observations $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, with $n \geq 4$, where θ_i are the parameters of interest and the z_i are random variables with $Ez_i = 0$ and $Ez_i^2 = \sigma_i^2$. The distribution μ_i of z_i is symmetric about 0. Let $\bar{\sigma}^2 = (\sum_{i=1}^n \sigma_i^2)/n$. As we already know, the equations*

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

have unique solutions $\lambda_{n,i}$. 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

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

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

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

where $p_i(\lambda, a) = \int |T_\lambda^S(x+a) - x|^2 \mu_i(dx)$. As was shown in the last proof, the left part is smaller than the minimum of

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

the first term is less or equal to 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 $\lambda_{n,i}$ derived in the proof of Theorem 3.2. \square

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. 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 3.5 we just have to find an upper bound for all the optimal thresholds. For given symmetric distributions μ_1 and μ_2 with $\mu_1([a, \infty)) < \mu_2([a, \infty))$, let $\lambda_i, i = 1, 2$, be the solutions of

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

Clearly λ_2 is larger than λ_1 . If the e_i are iid with support $[-a, a]$ 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 [41, p.855]) we have now

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

Thus an upper bound for the optimal thresholds of the empirical wavelet coefficients is the solution of equation (6) for the symmetric distribution μ_2 defined

by $\mu_2([t, \infty)) := \exp(-\frac{t^2}{2a^2})$. Later we will see that this solution of the equations for μ_2 is asymptotically like $\sqrt{2a^2 \log n}$.

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 3.6 *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 $X_i, i = 1, \dots, n$, be iid random variables with distribution μ . Assume g is a convex function on \mathbb{R}^+ , then for any $a_1, \dots, a_n \in \mathbb{R}^+$, $b_1, \dots, b_n \in \mathbb{R}^+$, respectively written in decreasing order with $\sum_{i=1}^k a_i^2 \leq \sum_{i=1}^k b_i^2$, $k = 1, \dots, n$, 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 $a \rightarrow g(ax^2)$ is convex for all $x \in \mathbb{R}$. It is easy to see, that for $c_1, \dots, c_n \in \mathbb{R}$ the distribution of $\sum_{i=1}^n c_i X_i$ is again a variance mixture of normal random variables with mixing measure $\nu_{c_1^2} \star \dots \star \nu_{c_n^2}$, 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

$$\begin{aligned} R(a_1^2, \dots, a_n^2) &:= Eg \left(\left(\sum_{i=1}^n a_i X_i \right)^2 \right) \\ &= \int_{-\infty}^{+\infty} g(x^2) \left(\int_{(0, \infty)} \dots \int_{(0, \infty)} \varphi_{\sum u_i a_i^2}(x) \nu(du_1) \dots \nu(du_n) \right) dx \\ &= \int_{(0, \infty)} \dots \int_{(0, \infty)} \left(\int_{-\infty}^{+\infty} g(x^2) \varphi_{\sum u_i a_i^2}(x) dx \right) \nu(du_1) \dots \nu(du_n) \\ &= ER^N \left(\sum_{i=1}^n Y_i a_i^2 \right), \end{aligned}$$

where Y_i are iid random variables with distribution ν .

Since R^N is convex it follows from Marshall and Olkin[35, ch.12, beginning of Sect.G] that R is Schur-convex, i.e. if $\alpha_1, \dots, \alpha_n$ and β_1, \dots, β_n in decreasing order and $\sum_{i=1}^k \alpha_i \leq \sum_{i=1}^k \beta_i$, $k = 1, \dots, n$, $\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i$, then:

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

Thus the assertion follows. \square

This theorem gives us the needed bounds for $p(\lambda, 0)$, indeed note that $(|x| - \lambda)_+^2 = (\sqrt{x^2} - \lambda)_+^2 = g(x^2)$ with $g(x) = (\sqrt{x} - \lambda)_+^2$ and g is convex.

What kind of distributions are variance mixtures of normal random variables? This family is closed under mixtures and convolutions. If we parameterize 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

$$f(\sqrt{|x|}) = \int_{(0, \infty)} \frac{1}{\sqrt{2\pi a}} \exp(-|x|/(2a)) \nu(da).$$

Thus f is a variance mixture of normal densities if and only if f is symmetric and $f(\sqrt{x})$, $x \in [0, \infty)$ is a Laplace transform. In turn, it is well known ([18, p.415]) that a function g is a Laplace transform if and only if it is completely monotone, i.e.

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

For $f(x) = \exp(-x^c)$, $0 < c \leq 1$ 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 hence the density

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

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

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

are variance mixtures as well. In Feller ([18, 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 Gaussian 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, but there exists a random variable v whose distribution is a normal mixture with $P(|e_1| > x) \leq KP(|v| > x)$, $K \geq 1$, 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 ([32, Lemma 4.6]). Let λ_n be the positive solution of $E(|Kv| - \lambda)_+^2(n+1) = \lambda^2 + \sigma^2$ then

$$\sup_{\theta \in \mathbb{R}} \frac{E(T_{\lambda_n}^S(\sum_i a_i e_i + \theta) - \theta)^2}{\sigma^2/n + \min(a, \sigma^2)} \leq \Lambda_n, \quad (7)$$

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 [32, Lemma 4.6]),

$$ET_\lambda^S \left(\sum_i a_i e_i \right)^2 \leq \frac{1}{2} ET_\lambda^S \left(2Kt_0 \sum_i a_i v_i \right)^2 + \frac{1}{2} ET_\lambda^S \left(2K \sum_i a_i u_i \right)^2 =: h(\lambda), \quad (8)$$

where the u_i are iid random variables with distribution $(\delta_1 + \delta_{-1})/2$. So if λ_n is the solution of $h(\lambda)(n+1) = \lambda^2 + \sigma^2$, then again relation (7) holds with $\Lambda_n = (\lambda_n^2 + \sigma^2)/((1 + 1/n)\sigma^2)$.

It would be desirable to generalize Theorem 3.6 to a wider class of distributions, or to obtain a version of Theorem 3.6 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 ([1]).

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

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

where the infimum is for all estimators and λ_n and Λ_n are now computed for the normal distribution ([15]). 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 of λ_n is given for distributions with exponentially decaying density.

Theorem 3.7 *Let μ be a distribution on \mathbb{R} with variance σ^2 and density $\exp(-h(x))$, where h is symmetric, continuous and increasing on $[0, \infty)$. Further let*

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

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

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

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

Then

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

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 strictly increasing. Let $1 > \delta > 0$ then

$$q(\lambda)f((1 + \delta)\lambda) \xrightarrow{\lambda \rightarrow \infty} 0 \text{ and } q(\lambda)f((1 - \delta)\lambda) \xrightarrow{\lambda \rightarrow \infty} \infty.$$

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

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

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.

Lets turn to the second asymptotic:

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

The second summand on the right 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 the assumptions, 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

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

□

Remark 3.8 *If the density f is asymptotically like an inverse fractional polynomial, i.e. there is a $c > 3$ with*

$$\liminf_{x \rightarrow \infty} f(x)x^c = \limsup_{x \rightarrow \infty} f(x)x^c \in (0, 1),$$

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

Remark 3.9 *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 3.10 *Suppose we have the observations $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, where θ_i are the parameters of interest and the z_i are iid random variables with $Ez_1 = 0$ and $Ez_1^2 = \sigma^2$.*

The Lebesgue density of the distribution μ of z_1 is of the form $\exp(-h(x))$, where h is symmetric, continuous and increasing on $[0, \infty)$.

Further

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

then

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

where the infimum is for all estimators.

Proof: We will prove this bound by computing Bayes risks. The proof is quite similar to the one in [15]. 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 $\varepsilon_n \rightarrow 0$ and $a_n \rightarrow \infty$.

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 on the first respectively 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 $\beta f(a_n + c) \geq \frac{\varepsilon_n}{1-\varepsilon_n} f(0)$, $\varepsilon_n n \rightarrow \infty$ and set

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

We will now prove $p_n = Q_n(N_n - n\varepsilon_n > m_n) = o(\varepsilon_n)$ by using Bennett's inequality ([41, 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 $\psi(0) = 1$.

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

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 $a_n \geq d_n \geq 0$.

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

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 $\varepsilon_n n \rightarrow \infty$ and $\beta f(a_n + c) \geq \frac{\varepsilon_n}{1-\varepsilon_n} f(0)$. The second condition is equivalent to $f(a_n + c) \geq \frac{\varepsilon_n}{\beta(1-\varepsilon_n)} f(0)$.

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. \square

Remark 3.11 *The condition*

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

in Theorem 3.10 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 3.12 *What happens if the additional σ^2 in equation (5) 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 side by c_n . The \sim remains valid if $c_n/(\varepsilon_n n) \rightarrow 0$. 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 3.13 *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 symmetric, continuous and increasing on $[0, \infty)$. Further*

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

We apply a periodic wavelet transform W_n to these observations, for every n we take the same type of wavelet. Let $Y = W_n(X)$ and $\theta = W_n(f)$.

Then

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

where the infimum is for all estimators for θ .

Remark 3.14 *In contrast to Theorem 3.10, 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 $z_{j,k}$, $0 \leq k \leq 2^j - 1$, $0 \leq j \leq m - 1$, will denote the k^{th} wavelet coefficient at level j of $W_n(X)$, $W_n(f)$ respectively $W_n(e)$. Let $\hat{\theta}_{j,k}$ be an estimator for $\theta_{j,k}$, $\hat{\theta}_{j,k}$ may depend on all X_i . We will compute an asymptotic lower bound for

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

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 $\varepsilon_n \rightarrow 0$ and $a_n \rightarrow \infty$. 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, \dots, 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

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

and

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

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 $f_{2i}, \dots, f_{2i+N-1}$. As one easily

checks, the projection of Q_n^i on $f_{2i}, \dots, f_{2i+N-1}$ is $\varepsilon_n \delta_{(d_0, \dots, d_{N-1})a_n} + (1 - \varepsilon_n) \delta_{(0, \dots, 0)}$ (δ 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 $X_{2i}, \dots, X_{2i+N-1}$ if $(f_{2i}, \dots, f_{2i+N-1}) = a_n(d_0, \dots, d_{N-1})$ respectively if $(f_{2i}, \dots, f_{2i+N-1}) = (0, \dots, 0)$, 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 $\hat{\theta}_i$ be an estimator for θ_i , then

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

We proceed to calculate 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

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

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 $\varepsilon_n n \rightarrow \infty$. Set $m_n := (\frac{n}{2} \varepsilon_n)^{2/3}$. 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 $0 \leq \widehat{\theta}_{m-1,k} \leq a_n$. Now,

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

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

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

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

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

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 $a_n \sim \text{constant } h^{-1}(\log n)$ the theorem is proved. \square

3.2 Distributions with compact support

In the previous section we considered noise whose tail is heavier than the tail of the normal distribution and computed lower asymptotic bounds for thresholding. We made use of the fact that one half of the coefficients is in the finest level. The distribution of the noise in this level is “close” to the distribution of the original noise. Now we want to consider noise whose support is compact. We want to show that the asymptotic performance of thresholding in this case is not better than in the Gaussian case. Here we will use the fact that the distribution of the noise in the upper level coefficients is near the normal distribution.

Theorem 3.15 *Let the observations $X_i = f_i + e_i$, $i = 1, \dots, n = 2^m$ be given, where the f_i are the parameters of interest and the e_i are iid random variables whose distribution has compact support $[-a, a]$. We apply a periodic wavelet transform W_n to these observations, for every n we take the same type of wavelet. The wavelet base has a Hölder regularity of $\beta > 0$. Let $Y = W_n(X)$, $\theta = W_n(f)$ and $z = W_n(e)$. Then with*

$$\Lambda_n := \inf_{(\lambda_{j,k}) \in \mathbb{R}^n} \sup_{\theta \in \mathbb{R}^n} \frac{\sum_{j,k} E|T_{\lambda_{j,k}}^S(Y_{j,k}) - \theta_{j,k}|^2}{\sigma^2 + B_n(\theta, \sigma^2)} \quad (9)$$

$$\liminf_{n \rightarrow \infty} \Lambda_n / (2 \log n) \geq 1.$$

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

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

We will now show that $\max_{j \leq h, k, i} |w_{j,k,i}^n| 2^{(m-j)/2} \leq C$ 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

$$\sup_{i,k} |2^{(m-j)/2} \langle \psi_{j,k}, \phi_{m,i} \rangle - \psi(2^{j-m}i - k)| \leq C 2^{\beta(j-m)} \quad (10)$$

for a constant C and $m - j > j_0$ for some j_0 ([9, p.205]). But W^n is based on periodized wavelets on $[0, 1]$, i.e. for $j \geq 0$ and $0 \leq k < 2^j$, $\psi_{j,k}$ and $\phi_{j,k}$ are replaced by

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

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 $\psi_{j,k}^{per}$ only at most N wavelets are involved, the same bounds as in (10) hold, but with C replaced by a constant C_2 depending on N and C . Hence

$$\sup_{k,i} |w_{j,k,i}^n| \leq 2^{(j-m)/2} N \|\psi\|_\infty + C_2 2^{(j-m)(\beta+1/2)} \leq C_3 2^{(j-m)/2}, \quad (11)$$

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

$$\begin{aligned} \Lambda_n &= \inf_{(\lambda_{j,k}) \in \mathbb{R}^n} \sup_{\theta \in \mathbb{R}^n} \frac{\sum_{j,k} E |T_{\lambda_{j,k}}^S(Y_{j,k}) - \theta_{j,k}|^2}{\sigma^2 + B_n(\sigma^2, \theta)} \\ &\geq \inf_{(\lambda_{j,k}) \in \mathbb{R}^n} \max \left(\frac{\sum_{j \leq h, k} E T_{\lambda_{j,k}}^S(z_{j,k})^2}{\sigma^2}, \frac{\sum_{j \leq h, k} (\lambda_{j,k}^2 + \sigma^2)}{(2^{h+1} + 1)\sigma^2} \right), \end{aligned} \quad (12)$$

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

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

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 $j \leq h$. To bound the first term in the maximum in (12) we will use the following lemma (see [32, Lemma 8.1]):

Lemma 3.16 *Let (X_i) be a finite sequence of independent mean zero real random variables such that $\|X_i\|_\infty \leq c$ 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 $t \geq K(\gamma)b$ and $tc \leq \varepsilon(\gamma)b^2$ where $b = (\sum_i EX_i^2)^{1/2}$,*

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

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

$$\sup_{j \leq h, k, i} \|w_{j,k,i}^n e_i\|_\infty \leq C_3 a 2^{(h-m)/2}$$

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

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

Further note that $ET_{\lambda_{j,k}}^S(z_{j,k})^2 \geq P(|z_{j,k}| \geq \lambda_{j,k} + 1)$. Also, because of the uniform convergence in the CLT, (see [38] p. 149, 5.4) there exists an $n_0 = n_0(K(\gamma))$ such that for all $n \geq n_0$, $P(|z_{j,k}| \geq K(\gamma) + 1) \geq (1 - \Phi(K(\gamma) + 1))/2$ where Φ is the normal distribution function.

Now let $(\lambda_{j,k,n})$ be one set of optimal thresholds for the right side in (9) and $0 < \alpha < 1$, then one of the following conditions holds (we always assume $j \leq h!$)

$$\left| \left\{ \lambda_{j,k,n} \in \left(K(\gamma)\sigma, \frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 a} \right) \right\} \right| \geq (1 - \alpha)2^{h+1}, \quad (13)$$

$$|\{\lambda_{j,k,n} \leq K(\gamma)\sigma\}| \geq \alpha 2^h \quad (14)$$

or

$$\left| \left\{ \lambda_{j,k,n} \geq \frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 a} \right\} \right| \geq \alpha 2^h \quad (15)$$

Thus for $n \geq n_0$ we get a lower bound in the first case (13) 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 lemma above 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 $\lambda_{j,k,n}$ 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 3.7 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 (12), 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 (12) 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. $\liminf \Lambda_n / (2 \log n) \geq 1$. \square

3.3 Very smooth densities

In this section the conditions 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 densities. The reason is that one cannot beat thresholding for very smooth densities with compact support.

Theorem 3.17 *Let the observations $X_i = s_i + e_i$, $i = 1, \dots, n = 2^m$ 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 . Their distribution μ has unimodal density f with support $[a, b]$. Further $\sqrt{f} \in C^2([a, b])$, and $f \in C^k([a, b])$ $k \geq 2$, with $f^{(k)}(a) \neq 0$ and $f^{(k)}(b) \neq 0$, also for $\ell < k$ $f^{(\ell)}(a) = f^{(\ell)}(b) = 0$. We apply a periodic wavelet transform W_n to these observations, for every n we take the same type of wavelet. Let $Y = W_n(X)$ and $\theta = W_n(s)$.*

Then

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

where the infimum is for 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

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

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

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

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 $\frac{|f'(x)|^{5/2}}{f(x)^{3/2}} = O((x-a)^{k-5/2})$ 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}

$$\left| \sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} \right| = O(t_n/\sqrt{n}) \quad (16)$$

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 because of (16) $E|h_{n,i}|^p = O((t_n/\sqrt{n})^p)$. 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

$$\int_a^b \left(\frac{\sqrt{f(x - t_{n,i}/\sqrt{n})}}{\sqrt{f(x)}} - 1 - \frac{t_{n,i}}{2\sqrt{n}} g(x) \right)^2 f(x) dx$$

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

Like in Strasser ([39, p.379]), a simple Taylor expansion yields

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

where $r(0) = 1$ and $|r(x_1) - r(x_2)| \leq C|x_1 - x_2|$ 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 depends only on t_n and n .

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

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

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

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 $\sum_{i=1}^n E h_{n,i}^2$ are stochastically equivalent with respect to P_0^n . This is equivalent to $\sum_{i=1}^n X_{n,i} \rightarrow 0$ in probability where $X_{n,i} = h_{n,i}^2 - E h_{n,i}^2$. Note that

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

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

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

Since $E X_{n,i} = 0$,

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

Now

$$P_0^n \left(\left| \sum_{i=1}^n X_{n,i} \right| \geq \varepsilon \right) \leq \sum_{i=1}^n P_0^n(|X_{n,i}| > 1) + P_0^n \left(\left| \sum_{i=1}^n \tilde{X}_{n,i} \right| \geq \varepsilon \right),$$

since

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

and

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

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 have now

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

Again, like in Strasser ([39, p.381 top]), it holds

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

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

Hence

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

Like in Lemma 74.3. in [39] it follows from

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

that

$$\begin{aligned}
&\sum_{i=1}^n \left(E_{P_0} \left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) \right)^2 - E_{P_0} h_{n,i}(x_i)^2 \right) \\
&= \sum_{i=1}^n E_{P_0} \left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) - h_{n,i}(x_i) \right) \left(\frac{t_{n,i}}{\sqrt{n}} g(x_i) + h_{n,i}(x_i) \right)
\end{aligned}$$

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

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

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

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 have now an asymptotic expansion for the likelihood ratios which is

necessary to calculate 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 $\varepsilon_n \rightarrow 0$ and $a_n \rightarrow \infty$, 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 $\tilde{\theta}_{j,k}$ 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

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

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

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

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

Simple computations yield now that

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

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 the Theorems 3.10 and 3.13 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 than the Bayes risk is larger than $(1 - \varepsilon_n)^2 \varepsilon_n a_n^2 C_1 p_1$.

$$\begin{aligned} & P_{a_n} \left(\left(\frac{dP_0}{(1 - \varepsilon_n)dP_0 + \varepsilon_n dP_{a_n}} \right)^2 \geq C_1 \right) \\ & = P_0 \left(\left(\frac{dP_{-a_n}}{(1 - \varepsilon_n)dP_{-a_n} + \varepsilon_n dP_0} \right)^2 \geq C_1 \right) \end{aligned}$$

$$\begin{aligned}
&= P_0 \left(\left(1 - \varepsilon_n + \frac{\varepsilon_n dP_0}{dP_{-a_n}} \right)^2 \leq 1/C_1 \right) \\
&\geq P_0 \left(\frac{dP_{-a_n}}{dP_0} \geq C_2 \varepsilon_n \right),
\end{aligned} \tag{17}$$

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 we will investigate now instead 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 (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 3.15. 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 $t_{r,i} = a_n c_i \sqrt{r}$):

$$\frac{dP_{a_n}}{dP_0} = \exp \left(a_n U_n - \frac{a_n^2}{2} \gamma^2 + o_{P_0}(1) \right) \tag{18}$$

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 an $N(0, \gamma^2)$ 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 [38] 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 (17) and (18),

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

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

$$\frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n \gamma^2}{2} + \gamma = \Phi^{-1} \left(\frac{1 - p_1}{2} \right) \gamma, \tag{19}$$

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 calculations yield now that the solution of equation (19) is

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

Now we can apply the machinery from the proofs of the Theorems 3.10 and 3.13 and this gives

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

□

3.4 Conclusions

In the previous sections, we examined the ideal estimator approach for soft thresholding; and computing the optimal thresholds was easy in this context. The optimal threshold for soft thresholding is the solution of a simple equality involving only the distribution of the noise. If a closed form of the distribution is not at hand, then it is more difficult to compute the optimal threshold.

Variations of the Theorems 3.2, 3.5 and 3.7 also hold for related type of estimators, for example hard thresholding and the estimator T_{λ}^M of the second section. In each case, the size of the threshold parameter and the performance of the estimator is 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 3.7, 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 $\sim \text{constant} \log^2 n$. In practice the higher levels are not thresholded. Note that this does not change the asymptotic performance of our scheme. After all only a smaller and smaller fraction of the wavelet coefficients is not thresholded. But there is another method, where the threshold for the n^{th} coefficient is always the same, the thresholds depend on the coefficient. This is presented next.

Theorem 3.18 *Let $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, where the θ_i are parameters of interest and 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 3.2). Let $\tilde{\lambda}_i$ be such that $p(\tilde{\lambda}_i, 0) = 2\sigma^2/i$ and let*

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

Then

$$\lim_{n \rightarrow \infty} \frac{\Lambda_n}{\tilde{\Lambda}_n} = 1.$$

Proof: It is clear that if $p(\tilde{\lambda}_i, 0)$ is decreasing in i , then $\tilde{\lambda}_i$ is increasing. Our first step is to prove $\tilde{\lambda}_n/\sqrt{2 \log n \sigma^2} \rightarrow 1$. 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 [41, 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. $\lim_{x \rightarrow 0} q^{-1}(x)/\sqrt{2 \log(1/x) \sigma^2} = 1$. Thus our claim follows: $\lim_{n \rightarrow \infty} \tilde{\lambda}_n/\sqrt{2 \sigma^2 \log n} = 1$.

Now, let $\theta \in \mathbb{R}^n$, if $|\theta_j| > \sigma$, $1 \leq j \leq n$, then, since $p(\tilde{\lambda}_j, \theta_j) \leq p(\tilde{\lambda}_j, \infty)$ (see proof of Theorem 3.2)

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

If $|\theta_j| < \sigma$ and $(\sum_{i=1}^n p(\tilde{\lambda}_i, \theta_i))/B_n(\theta, \sigma^2) \geq 1$ then, since $p(\tilde{\lambda}, \theta_j) \leq p(\tilde{\lambda}, 0) + \theta_j^2$ (again see proof of Theorem 3.2):

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

Thus if $\sup_{\theta \in \mathbb{R}^n} \frac{\sum_{i=1}^n p(\tilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} \geq 1$ then it follows from the preceding calculations that

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

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

The above result could be transferred to other distributions, where the density has an exponential decay like in Theorem 3.7. The main point is to choose λ_i such that $\sum_{i=1}^n p(\tilde{\lambda}_i, 0) \approx \Lambda_n \sigma^2$. 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 to the thresholds in Theorem 3.18. The thresholds are quite close to the thresholds in the next section, where the choice of thresholds is based on another minimax approach.

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 if in the case of noise with bounded support one wants to be better than soft thresholding one has to exploit special properties of the noise distribution, maybe something like a running median. 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. Also there are other continuous shrinkers that tend to the identity function for large values. In our minimax

| $n \rightarrow$ | 32 | 128 | 512 | 2048 | 65536 | 2^{24} | 2^{32} |
|----------------------------|------|------|------|------|-------|----------|----------|
| ϕ | 1.28 | 1.67 | 2.04 | 2.40 | 3.22 | 4.35 | 5.31 |
| $\exp(- x)$ | 1.58 | 2.19 | 2.85 | 3.55 | 5.43 | 8.70 | 12.15 |
| $\exp(-\sqrt{ x })$ | 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 8: The optimal thresholds for some densities

framework it does not pay off to use such an estimator. Further this chapter showed us that in our context one cannot do asymptotically better than soft thresholding. Another reason for negligence of the performance of our estimator is that we do not measure the risk directly but compare it to our benchmark.

The results in this chapter are chiefly of 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 (see [15]), namely the solution of equation (6):

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

The same holds 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. I performed some calculations with Mathematica to compute the optimal thresholds, the results can be seen in Figure 8. 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 I added the optimal thresholds for distributions with polynomial decay, the uniform distribution and some mixtures of them. All distributions are scaled so that their variance is 1.

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. Example: $\exp(-x^2)$ is a functional part of the density of a normal distribution. The rationale 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). The Figures 9 and 10 show the ratio of the asymptotic threshold term and the optimal threshold for the normal distribution and the Laplace distribution. The values

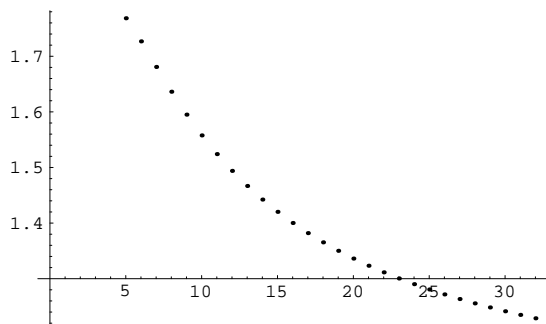


Figure 9: Ratio of asymptotic threshold and optimal threshold for the Normal distribution

on the horizontal axis are the dual logarithms of n . As one sees the asymptotic works very slowly.

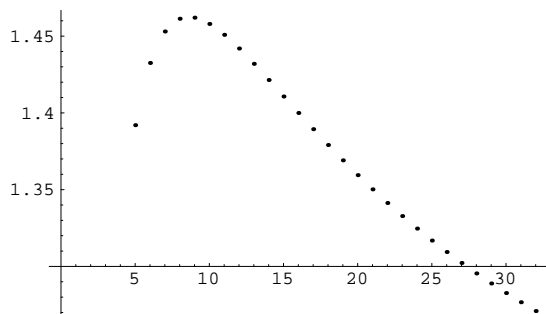


Figure 10: Ratio of asymptotic threshold and optimal threshold for the Laplace distribution

For non-Gaussian noise it is complicated to compute the distribution of the noise in the upper levels of the transformed signal. The consequence of the theorems 3.5 and 3.6 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 our minimax approach it does not pay off to waste thoughts on this. In the next section we will see, that depending on the noise, from some level on upward the thresholds can be chosen as for Gaussian noise. 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 $\hat{\lambda}_n$ such that $(1+n)p_X(\lambda, 0) = \lambda^2 + 1$. Let Y be an independent copy of X and $\tilde{\lambda}_n$ 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 $\tilde{\lambda}_n = 1.81$ and for $n = 2^{16}$ $\lambda_n = 4.1$ and $\tilde{\lambda}_n = 3.55$. 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, computing the optimal threshold (in the sense of Theorem 3.2) is very difficult for convolutions. After all we have to solve the following equation

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

When using an 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, the variance has to be estimated. The oracle method is highly sensitive to the estimated variance. If the thresholds are chosen based on a variance estimation which is too 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.

Of course the oracle is not always the optimal estimator, so let X be a mean zero random variable with variance 1 and consider the linear shrinker $x \rightarrow \alpha x$, $\alpha \in (0, 1)$, 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 oracle. But it is easy to see that this ratio is bounded away from 0, let $p(\cdot, \cdot)$ the risk function for soft thresholding for some distribution μ with variance 1 and mean zero, for simplicity we assume μ is also 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 Gaussian distribution c_μ is larger than 0.7. Let Λ_n and λ_n now have their usual meaning in the sense of Theorem 3.2 with respect to the distribution μ . If $\tilde{\lambda}_i$ $i = 1, \dots, n$ is a set of thresholds, then

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

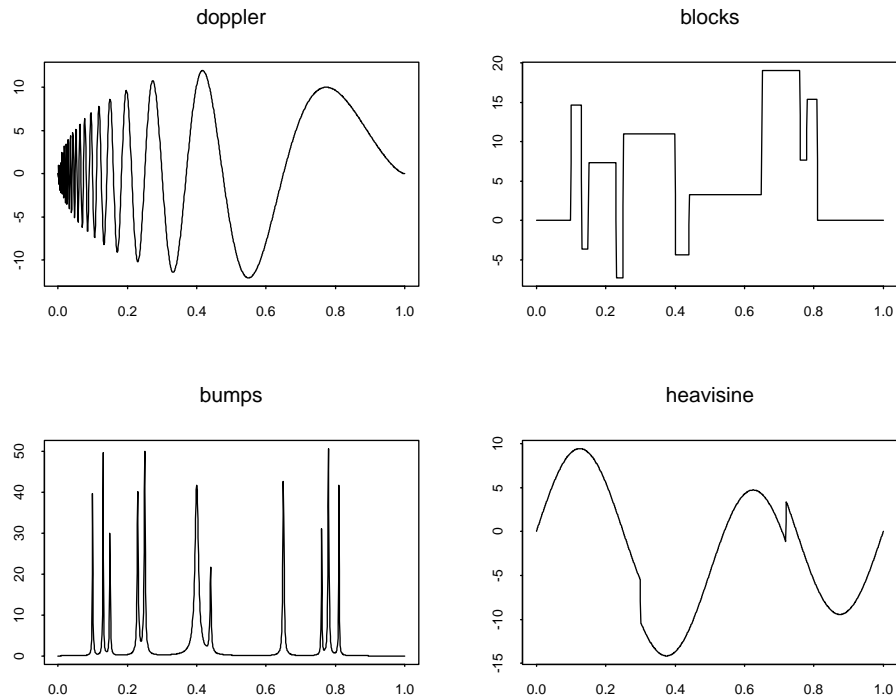
Λ_n/c_μ seems to be big, but if for example for $0 < \varepsilon_1 < 1$, n^{ε_1} of the a_i are larger than 1 then for any soft thresholding estimator the risk will be of size at least $O(n^{\varepsilon_1})$. If the thresholds are chosen too small or too large, then it can easily

happen that the risk is of size $O(n^{\varepsilon_2})$ with $\varepsilon_2 > \varepsilon_1$. So if one wants to play safe then thresholds based on the ideal estimator method are a good choice.

As we already mentioned, the oracle 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.

Interesting is a comparison of the performance of the different thresholds, and the performance of soft thresholding for non-Gaussian noise. For this I undertook a small Monte-Carlo study. The target signals are depicted in Figure 11, they were introduced by Donoho and Johnstone ([15]). Note that the functions blocks

Figure 11: The 4 Signals of Donoho and Johnstone



and heavisine are not continuous and hence not in a Besov space $B_{p,q}^m$ where $m > 1/p$. The simulation was performed 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 [9]). As lengths of the signal were chosen 512 and 8192. The noise is Gaussian noise and random variables with a Student distribution with four degrees of freedom, scaled to have variance 1. The density of this distribution decays like $1/x^5$, so its tail is quite heavy. The thresholds used are the optimal thresholds of Theorem

3.2 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

| signal | Gaussian noise | | | Student noise | | |
|-----------|----------------|------|------|---------------|------|------|
| | 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

| signal | Gaussian noise | | | Student noise | | |
|-----------|----------------|-------|-------|---------------|-------|-------|
| | 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 I tried the estimator of Theorem 3.18, 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 $\tilde{\lambda}_{2^j+1}$ of Theorem 3.18.

Thresholds of Theorem 3.18

| signal | Gaussian noise | | Student noise | |
|-----------|----------------|-------|---------------|-------|
| | 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 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 performance of the estimator which does not threshold the first five levels and the estimator of Theorem 3.18 are comparable. Surprising is the good performance of the James-Stein estimator, it is also robusiter if the noise is not Gaussian noise.

Some estimators were suggested, that apply a linear estimator to the wavelet coefficients, where some constants are chosen depending on the data. Because of the following result of Donoho and Johnstone ([16, Theorem 5]) the estimators are not that useful.

Let X_1, \dots, X_n be iid mean zero Gaussian 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 with iid Gaussian 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.

4 The function space approach

In the previous chapter we compared the performance of wavelet estimators to a benchmark and tried to obtain minimax results for this ratio. The other well known approach is to assume the function is a member of a bounded subset of a smoothness class and to compute minimax bounds for estimation in this context. Donoho and Johnstone had early results for this type of estimation problem see [15],[17] or [13], later Deylon and Juditsky [12] and Neumann and Spokoiny [37], extended this to non-Gaussian noise. In this chapter will give an extension of the results of Deylon and Juditsky and Neumann and Spokoiny. Deylon and Juditsky showed that for rather general noises, one can have the same minimax rate as for Gaussian noise. Neumann and Spokoiny showed that under somewhat stronger conditions the ratio of the minimax risk for Gaussian noise and other types of noises tends to 1. I will show that under lesser conditions on the noise, for soft thresholding the ratio of the minimax risk for Gaussian noise and other types of noises tends to 1.

In the second part I will show that if the noise is heavy tailed then by median filtering the data and then applying wavelet thresholding it is still possible to have the same minimax rate as for Gaussian noise. The chapter concludes with two examples for when there are estimators which are better than wavelet thresholding.

4.1 The moment conditions

The model we consider is quite similar to the one in the previous chapters. We are given data $X_i = f_i + n^{-1/2}z_i$, $i = 1, \dots, n$, $n = 2^h$, $h \in \mathbb{N}$, where (f_i) is the signal in our data which we want to estimate and the (z_i) are iid random variables which represent the noise. Actually we think of $f_i = f_{n,i} = f(i/n)/\sqrt{n}$, so we assume our data is sampled from some real signal with rate $1/n$ and multiplied by $1/\sqrt{n}$. The mean of z_1 is 0 and the variance is 1. We will apply a wavelet transform W_n to the data, so we have then noisy wavelet coefficients.

$$w_k = a_k + e_k; k = 1, \dots, 2^{j_0} \text{ and } w_{j,k} = a_{j,k} + e_{j,k}; j \leq j_0, k = 1, \dots, 2^j$$

where j_0 is a fixed constant, i.e. we assume we stop the wavelet transformation at the level j_0 . (Of course the coefficients depend also on n). Again we use a wavelet transform adapted to an interval, either by periodization or boundary corrections. This model is reasonable if we assume that $f(i/n)/\sqrt{n}$ is a good approximation for the scaling coefficient with the index (h, i) , but this has already been discussed in the introduction. The difference to the previous model in chapter 3 is that we do not compare the performance of estimators with a benchmark, but we restrict the possible values of $(a_{\cdot,\cdot})$. We will assume,

$$\left(\sum_k |a_k|^p \right)^{1/p} + \left(\left(\sum_{j \geq j_0} 2^{js} \left(\sum_k |a_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q} \leq A,$$

for some constant A , where $s := m + 1/2 - 1/p$ and $m > 1/p$. It was already explained in the first chapter that if the $(a_{\cdot, \cdot})$ are the wavelet coefficients of f and the wavelet basis is sufficiently smooth then $\|f\|_{B_{p,q}^m} \leq C_2 A$, where $C_2 = C_2(m, p, q)$ is a constant. As was mentioned earlier, the condition $m > 1/p$ is necessary to have well defined point values of functions in $B_{p,q}^m$. This implies that really “nice” functions like $\text{sign}(x)$ do not belong to any of the Besov spaces we consider.

If the z_i are Gaussian iid random variables, then the minimax rate in this model is $n^{-2m/(2m+1)}$, i.e.

$$\inf_{\hat{a}} \sup_{a, \|a\|_{B_{p,q}^m} \leq C} E \|\hat{a} - a\|_2^2 \sim cn^{-\frac{2m}{2m+1}}, \quad (20)$$

where the infimum is for all estimators and c is a positive constant (see [37]).

Further estimators based on soft thresholding achieve this rate. Later it was shown by Neumann and Spokoiny ([37]) that for noise with all moments, soft thresholding achieves the same rate as in the Gaussian case, the actual performance, not just the rate in both cases is the same, i.e. the c in relation (20) is the same. If the noise satisfies certain regularity conditions, then even the minimax rate is the same. Deylon and Juditsky ([12]) extended these results by showing that even for more general distributions, soft thresholding can achieve the same rate as in the Gaussian case. I want to complement this result, by showing that if the noise fulfills some moment conditions, soft thresholding actually achieves the same asymptotic performance as soft thresholding in the Gaussian case. (Well, we need fairly high moments.) We use an idea already employed by Deylon and Juditsky, namely it is possible to assume the noise is bounded.

Theorem 4.1 *Assume the situation at the beginning of the chapter holds, and $p, q \geq 1$ and $m > 1/p$ are constants. The (z_i) have finite moments of order L where L satisfies the respectively the conditions*

$$L > \frac{6}{2m/(2m+1)} \text{ if } p \geq 2 \quad (21)$$

and

$$L > \frac{6(m + 1/2 - 1/p)(2m + 1)}{(m + 1/2 - 1/p)(2m + 1) - m} \text{ if } 1 \leq p \leq 2. \quad (22)$$

Further their distribution is symmetric and continuous. Then

$$\liminf_{n \rightarrow \infty} \frac{\inf_{(\lambda)} \sup_{a, \|a\|_{B_{p,q}^m} \leq A} E_{\Phi} \sum_{j,k} |T_{\lambda,j,k}^S(w_{j,k} - a_{j,k})|^2}{\inf_{(\tilde{\lambda})} \sup_{a, \|a\|_{B_{p,q}^m} \leq A} E \sum_{j,k} |T_{\tilde{\lambda},j,k}^S(w_{j,k} - a_{j,k})|^2} \geq 1, \quad (23)$$

where E_{Φ} stands for the mean when the z_i have a normal distribution.

The requirement for symmetric random distributions is for technical reasons, because we will need that $E z_i \mathbf{1}_{\{|z_i| > \lambda\}} = 0$ for all λ . The requirement for continuous

distributions is made to simplify partial integration. These requirements can be circumvented by more technical efforts in the proof.

To prove this theorem we will need some lemmas.

Lemma 4.2 *Let $s := m + 1/2 - 1/p$ and and assume*

$$\left(\sum_k |a_k|^p \right)^{1/p} + \left(\left(\sum_{j \geq j_0} 2^{js} \left(\sum_k |a_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q} \leq A,$$

for some $A > 0$. Then for all $l \geq j_0$

$$\sum_{j \geq l} \|a_{j,\cdot}\|_2^2 \leq \begin{cases} A^2(2^{-2m})^l / (1 - 2^{-2m}) = O(n^{-\alpha 2m}) & : p \geq 2 \\ A^2(2^{-2s})^l / (1 - 2^{-2s}) = O(n^{-\alpha 2s}) & : 1 \leq p < 2 \end{cases}$$

where $n = 2^h$ and for all α with $2^l \geq n^\alpha = 2^{\alpha h}$.

Proof: As usual, in the level j there are 2^j coefficients $a_{j,\cdot}$. It is clear that for $j \geq j_0$, $2^{js} \|a_{j,\cdot}\|_p \leq A$. If $p \geq 2$ then:

$$\begin{aligned} \|a_{j,\cdot}\|_2^2 &\leq ((2^j)^{1/2-1/p} \|a_{j,\cdot}\|_p)^2 \\ &\leq 2^{2j(1/2-1/p)} A^2 2^{-2js} \\ &= A^2 2^{-2j(s-1/2+1/p)} = A^2 2^{-2jm}. \end{aligned}$$

If $1 \leq p < 2$ then

$$\|a_{j,\cdot}\|_2^2 \leq \|a_{j,\cdot}\|_p^2 \leq A^2 2^{-2js}.$$

Thus

$$\sum_{j \geq l} \|a_{j,\cdot}\|_2^2 \leq \begin{cases} A^2(2^{-2m})^l / (1 - 2^{-2m}) & : p \geq 2 \\ A^2(2^{-2s})^l / (1 - 2^{-2s}) & : 1 \leq p < 2 \end{cases}.$$

So if $2^l \geq 2^{\alpha h}$ then

$$\sum_{j \geq l} \|a_{j,\cdot}\|_2^2 \leq \begin{cases} A^2 n^{-2m\alpha} / (1 - 2^{-2m}) & : p \geq 2 \\ A^2 n^{-2s\alpha} / (1 - 2^{-2s}) & : 1 \leq p < 2 \end{cases}.$$

□

This Lemma implies that if we want to achieve the same minimax rate as for Gaussian noise, we do not have to care about coefficients in the levels $l = \alpha h$ and below as long as $\alpha > 1/(2m + 1)$ if $p \geq 2$ and $\alpha > 2m/((2m + 1)s)$ if $1 \leq p \leq 2$. The reason is the L^2 -norm of these coefficients is of size $o(n^{-2m/(2m+1)})$. For $p \geq 2$, let l be such that $2n^{1/(2m+1)} \geq 2^l > n^{1/(2m+1)}$, then the simple estimator which discards the coefficients in the level l and below achieves the minimax rate since

$$\sum_{j \geq l, k} a_{j,k}^2 = O\left(n^{-\frac{2m}{2m+1}}\right) \text{ and } \sum_{j < l, k} E z_{j,k}^2 = O\left(n^{-\frac{2m}{2m+1}}\right).$$

We will want to use the following result of Kolmogorov (see [41, p.855]).

Lemma 4.3 Let $X_i, i = 1, \dots, n$ be independent random variables with mean zero. Let $s_n := \sqrt{\sum_{i=1}^n EX_i^2}$ and $|X_i| < K, i = 1, \dots, n$ and $S_n = \sum_{i=1}^n X_i$. Then

$$P(S_n/s_n > \lambda) < \begin{cases} \exp\left(\frac{\lambda^2}{2}\left(1 - \frac{\lambda K}{2s_n}\right)\right) & : \lambda \leq s_n/K \\ \exp\left(\frac{-\lambda s_n}{4K}\right) & : \lambda \geq s_n/K \end{cases}.$$

The next Lemma is a simple application of the previous one.

Lemma 4.4 Let X_1, \dots, X_k be independent mean zero random variables and let $\varepsilon > 0$. Assume $\|X_i\|_\infty \leq K_n = O(n^{-\varepsilon}), i = 1, \dots, k$ and $\sum_{i=1}^k EX_i^2 = 1$. Let F be the distribution function of $\sum_{i=1}^k X_i$. Let $\lambda_n = \log n$ and $c_n := (1 - K_n/(2\lambda_n))$. Then for a with $0 < a < \lambda_n$

$$\int_a^\infty x^2 F(dx) \leq (a^2 + 2)/c_n \exp(-c_n a^2/2) + o(\exp(1/K_n)).$$

Proof: Using Lemma 4.3 we obtain

$$\begin{aligned} \int_a^\infty x^2 dF(x) &= a^2(1 - F(a)) + 2 \int_a^\infty x(1 - F(x))dx \\ &\leq a^2 \exp(-c_n a^2/2) + 2 \int_a^\infty x \exp(-c_n x^2/2)dx \\ &\quad + 2 \int_{1/K_n}^\infty x \exp(-x/(4K_n))dx \\ &= \dots + 2/c_n \exp(-c_n a^2/2) - 8K_n x \exp(-x/(4K_n))|_{1/K_n}^\infty \\ &\quad + 8K_n \int_{1/K_n}^\infty \exp(-x/(4K_n))dx \\ &= \dots + \dots + 8 \exp(-1/(4K_n^2)) + 32K_n^2 \exp(-1/(4K_n^2)) \\ &\leq (a^2 + 2)/c_n \exp(-c_n a^2/2) + o(\exp(1/K_n)). \end{aligned}$$

Further we need the following result about large deviations which is a simple extension of Lemma 5.8 in [38], the difference to this Lemma is that the requirement of identical distributions of the random variables is dropped. In the sequel Φ will denote the distribution of a $N(0, 1)$ random variable.

Lemma 4.5 Let $C > 0$ and $\varepsilon \in (0, 1)$, then there exists an increasing sequence (β_n) converging to 1 such that for independent mean zero random variables X_1, \dots, X_k with $\sum_{i=1}^k EX_i^2 = 1$ and $\max_{i=1, \dots, k} E|X_i|^3 \leq Cn^{-3/2}$ and all x with $|x| \leq \varepsilon\sqrt{2\log n}$

$$\beta_n \leq \frac{P(\sum_{i=1}^k X_i \leq x)}{\Phi((-\infty, x])} \leq 1/\beta_n \text{ and } \beta_n \leq \frac{P(\sum_{i=1}^k X_i > x)}{\Phi((x, \infty))} \leq 1/\beta_n. \quad (24)$$

Proof: The proof with the help of Esseen's inequality ([38, Theorem 5.4]) is essentially the same as for Lemma 5.8 in [38].

Lemma 4.6 Let $X_i, i = 1, \dots, n$ be iid random variables with $m_4 := EX_i^4 < \infty$, $EX_i = 0$, $EX_i^2 = 1$. Let $Y := \sum_{i=1}^n a_i X_i$ where $\sum_{i=1}^n a_i^2 = 1$. Then $3 \leq EY^4 \leq m_4$ if $m_4 \geq 3$ and $3 \geq EY^4 \geq m_4$ if $m_4 \leq 3$.

Note that 3 is the fourth moment of a $N(0, 1)$ random variable.

Proof:

$$\begin{aligned}
E \left(\sum_i a_i X_i \right)^4 &= \sum_i a_i^4 m_4 + 3 \sum_{i,j, i \neq j} a_i^2 a_j^2 \\
&= \sum_i a_i^4 m_4 + 3 \left(\sum_i a_i^2 \left(\sum_{j, j \neq i} a_j^2 \right) \right) \\
&= m_4 \sum_i a_i^4 + 3 \left(\sum_i a_i^2 (1 - a_i^2) \right) \\
&= m_4 \sum_i a_i^4 + 3 \left(\sum_i a_i^2 - \sum_i a_i^4 \right) \\
&= (m_4 - 3) \sum_i a_i^4 + 3 \\
&= m_4 + (m_4 - 3) \left(\sum_{i=1}^4 a_i^4 - 1 \right)
\end{aligned}$$

The assertion follows now for both cases since $\sum_i a_i^4 \leq 1$. \square

Proof of Theorem 4.1: We will show that we can discard the coefficients from a certain level downward because the ℓ^2 -norm of these is small compared to the minimax risk. For the other levels we will show that with the right thresholds we can achieve the same minimax performance as for Gaussian noise.

So first let us consider the following situation, let $\alpha \in (0, 1)$ be fixed and l such that $2^l \leq n^\alpha < 2^{l+1}$ and (c_n) is a sequence which tends to infinity. Assume $T_{j,k}$ are functions with $\sup_x |T_{j,k}(x) - x| < d_n$, with $d_n = O(\log n / \sqrt{n})$. Define $A_n := \{\max_i |z_i| < c_n\}$, $\varepsilon_n := 1 - P(A_n)$, and let $\tilde{z}_i := \min(\max(z_i, -c_n), c_n)$ and $\tilde{e}_{j,k} := (W(\tilde{z}/\sqrt{n}))_{j,k}$.

$$\begin{aligned}
&E \sum_{j \leq l, k} |T_{j,k}(a_{j,k} + e_{j,k}) - a_{j,k}|^2 \\
&= E \sum_{j \leq l, k} |T_{j,k}(a_{j,k} + e_{j,k}) - a_{j,k}|^2 \mathbf{1}_{A_n} + \sum_{j \leq l, k} |T_{j,k}(a_{j,k} + e_{j,k}) - a_{j,k}|^2 \mathbf{1}_{A_n^C} \\
&\leq E \sum_{j \leq l, k} |T_{j,k}(a_{j,k} + \tilde{e}_{j,k}) - a_{j,k}|^2 + \sum_{j \leq l, k} |T_{j,k}(a_{j,k} + e_{j,k}) - a_{j,k}|^2 \mathbf{1}_{A_n^C} \\
&\leq E \dots + \sum_{j \leq l, k} \sqrt{E |T_{j,k}(a_{j,k} + e_{j,k}) - a_{j,k}|^4} \sqrt{P(A_n^C)} \\
&\leq E \dots + 2n^\alpha \sqrt{8(M + d_n^4)} \sqrt{\varepsilon_n},
\end{aligned}$$

where $M := \max(3, Ez_1^4)/n^2 \geq Ee_{j,k}^2$ because of Lemma 4.6.

If $\varepsilon_n = O(1/n^2)$ then the term $2n^\alpha \sqrt{8(M + d_n^4)} \sqrt{\varepsilon_n}$ is of size $o(1/n)$. Thus if we discard the coefficients from the level l downward and use thresholds that are smaller than $\log n / \sqrt{n}$, then w.l.o.g. we can assume the z_i are bounded by c_n , since $o(1/n)$ is small compared to the minimax risk. Consider the coefficients in the level l (with $2^l \leq n^\alpha < 2^{l+1}$) and above. Let w be the noise part in one of these coefficients, then with $n = 2^h$

$$w = \sum_{i=1}^n v_i \frac{z_i}{\sqrt{n}} \text{ and } \max_i |v_i| \leq C_1 2^{-(h-l)/2} \leq C_1 \sqrt{n^{\alpha-1}}, \quad (25)$$

where C_1 only depends on the type of the wavelet transform. To assume that the z_i are bounded and thus to be able to use the Lemmas 4.3 and 4.4, we will need that $P(\max_{1 \leq i \leq n} |z_i 2^{-(h-l)/2}| \geq n^{-\varepsilon}) = O(1/n^2)$ for some $\varepsilon > 0$. The z_i have moments of order L , hence

$$\begin{aligned} P(\max_{1 \leq i \leq n} |z_i 2^{-(h-l)/2}| \geq n^{-\varepsilon}) &\leq nP(|z_1| 2^{-(h-l)/2} \geq n^{-\varepsilon}) \\ &\leq nE|z_1|^L n^{-L(1-\alpha)/2} n^{\varepsilon L} \\ &\leq E|z_1|^L n^{1-L((1-\alpha)/2-\varepsilon)}. \end{aligned}$$

This implies that $P(\max_{1 \leq i \leq n} |z_i 2^{-(h-l)/2}| \geq n^{-\varepsilon}) = O(1/n^2)$ if $1 - L((1-\alpha)/2 - \varepsilon) \leq -2$ and this is given if

$$L \geq \frac{6}{(1-\alpha) - 2\varepsilon}.$$

Now let α such that $\sup_{a, \|a\|_{B_{p,q}^m} \leq A} \sum_{j \geq \alpha h} |a_{j,k}|^2 = o(n^{-2m/(2m+1)})$, if $p \geq 2$ then Lemma 4.2 implies $1 - \alpha < 2m/(2m+1)$ and hence

$$L > \frac{6}{2m/(2m+1) - 2\varepsilon}. \quad (26)$$

In the same way for $1 \leq p < 2$ we need $\alpha > \frac{m}{(m+1/2-1/p)(2m+1)}$ and thus

$$L > \frac{6}{\frac{(m+1/2-1/p)(2m+1)-m}{(m+1/2-1/p)(2m+1)} - 2\varepsilon}. \quad (27)$$

Since ε can be arbitrarily small, these are the same conditions as in the theorem. Here are some examples for the moment conditions for $1 \leq p < 2$:

$$\begin{aligned} m = 1/2, p = 2 &\Rightarrow L > 12, \\ m = 1, p = 1 &\Rightarrow L > 18, \\ m = 1, p = 3/2 &\Rightarrow L > 10, \\ m = 2, p = 3/2 &\Rightarrow L > 7.7. \end{aligned}$$

As we already know the numerator in equation (23) is $\sim \text{constant } n^{-2m/(2m+1)}$. From now on we assume $\alpha \in (0, 1)$ with $2^l \leq n^\alpha < 2^{l+1}$ such that

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} \sum_{j>l, k} a_{j,k}^2 = o(1/n^{2m/(2m+1)}) \text{ and } L \geq 6/((1-\alpha) - 2\varepsilon),$$

for some $\varepsilon > 0$. This is possible because of the conditions imposed on L . Thus to prove the theorem we can choose $\tilde{\lambda}_{j,k} = \infty$ for $j > l$.

In view of the previous computations, w.l.o.g. we can assume $\|z_i 2^{(h-l)/2}\|_\infty \leq n^{-\varepsilon}$, $i = 1, \dots, n$, provided we use thresholds smaller than $(\log n)/\sqrt{n}$. This condition implies that the noise terms in the wavelet coefficients are sums of independent random variables which are smaller than $n^{-\varepsilon}/\sqrt{n}$. In the sequel, for a distribution μ , $\lambda \geq 0$ and $a \in \mathbb{R}$, let $p_\mu(\lambda, a) := \int (T_\lambda^S(x+a) - a)^2 \mu(dx)$. If the e_i are iid $N(0, 1/n)$ random variables, then distribution of the noise in each coefficient is $\Phi_n := N(0, 1/n)$. Let $(\lambda_{j,k})$ be the optimal minimax thresholds for this situation. Let $\mu_{j,k}$ denote the distribution of the random variable $e_{j,k}$, i.e. the distribution of the noise in the coefficient with index (j, k) . We remember that $Ee_{j,k}^2 = 1/n$. Of course l , $\mu_{j,k}$ and $\lambda_{j,k}$ depend on n , but for simplicity we choose not to indicate this with notation.

If with suitable thresholds $\tilde{\lambda}_{j,k}$,

$$\liminf_{n \rightarrow \infty} \inf_{j \leq l, k} \inf_a \frac{p_{\Phi_n}(\lambda_{j,k}, a) + 1/n^2}{p_{\mu_{j,k}}(\tilde{\lambda}_{j,k}, a)} \geq 1 \quad (28)$$

holds, then soft thresholding achieves the same minimax performance as in the Gaussian case. The reason is that

$$\sum_{j \leq l, k} \frac{1}{n^2} \leq \frac{1}{n} = o\left(n^{-\frac{2m}{2m+1}}\right)$$

and $n^{-2m/(2m+1)}$ is the minimax rate for Gaussian noise. To spare us some distinction of cases, for the remainder of the proof we assume that $a > 0$.

Let λ_n be the threshold such that $p_{\Phi_n}(\lambda_n, 0) = 1/n^2$. If $\lambda > \lambda_n$ then

$$(T_{\lambda_n}^S(x+a) - a)^2 < (T_\lambda^S(x+a) - a)^2 \text{ for } x \in (-\lambda_n - a, \lambda_n).$$

Further $\int_{\lambda_n}^\infty (x - \lambda_n)^2 \Phi_n(dx) = p_{\Phi_n}(\lambda_n, 0)/2$ and

$$\int_{-\infty}^{-\lambda_n - a} ((x + \lambda_n + a) - a)^2 \Phi_n(dx) \leq \int_{-\infty}^{-\lambda_n} (x + \lambda_n)^2 \Phi_n(dx) = p_{\Phi_n}(\lambda_n, 0)/2.$$

Thus

$$\begin{aligned} p_{\Phi_n}(\lambda, a) &\leq \int_{-\lambda_n - a}^{\lambda_n} (T_\lambda^S(x+a) - a)^2 \Phi_n(dx) + \int_{\lambda_n}^\infty (T_{\lambda_n}^S(x+a) - a)^2 \Phi_n(dx) \\ &\quad + \int_{-\infty}^{-\lambda_n - a} (T_{\lambda_n}^S(x+a) - a)^2 \Phi_n(dx) \\ &\leq p_{\Phi_n}(\lambda, a) + 1/n^2. \end{aligned}$$

Hence without loss of generality we will assume $\sup_{j \leq l, k} \lambda_{j,k} \leq \lambda_n \sim \sqrt{2 \log n} / \sqrt{n}$ and proving that for suitable thresholds $\tilde{\lambda}_{j,k}$

$$\liminf_{n \rightarrow \infty} \inf_{j \leq l, k} \inf_a \frac{p_{\Phi_n}(\lambda_{j,k}, a)}{p_{e_{j,k}}(\tilde{\lambda}_{j,k}, a)} \geq 1 \quad (29)$$

holds is enough to complete the proof. In the following we assume μ_n is the distribution of $e_{j,k}$ and $p_n := p_{\mu_n} = p_{e_{j,k}}$.

Note: Since doing computations with the variance $1/n$ is so nasty, we will multiply the random variables and thresholds by \sqrt{n} , and the risks by n . The size of the fraction in relation (29) is not changed by this transformation.

We will need the following inequality,

$$\begin{aligned} p_{\Phi}(\lambda, a) &= a^2 \mu(-\lambda < x + a < \lambda) + \int_{\lambda-a}^{\infty} (x - \lambda)^2 \Phi(dx) \\ &\quad + \int_{-\infty}^{-\lambda-a} (x + \lambda)^2 \Phi(dx) \\ &\geq a^2 \Phi(-\lambda - a < x < \lambda - a) + \int_{\lambda}^{\infty} (x - \lambda)^2 \Phi(dx) \\ &\quad + \int_{-\infty}^{-\lambda-a} (x + \lambda + a)^2 \Phi(dx) \\ &\geq a^2 \Phi(-\lambda - a < x < \lambda - a) + \frac{p_{\Phi}(\lambda, 0) + p_{\Phi}(\lambda + a, 0)}{2}, \end{aligned} \quad (30)$$

since Φ is symmetric. And we need another inequality, if $\lambda \geq 1$ then

$$\begin{aligned} p_{\Phi}(\lambda, 0) &\geq 2\Phi(\{x > \lambda + 1\}) \\ &\geq \frac{2}{\sqrt{2\pi}} \left(\frac{1}{\lambda + 1} - \frac{1}{(\lambda + 1)^3} \right) \exp(-(\lambda + 1)^2/2) \\ &\geq \frac{1}{\sqrt{2\pi}\lambda} \exp(-(\lambda + 1)^2/2), \end{aligned} \quad (31)$$

where the second inequality follows from an lower bound for the error function (see [41, p.850]). From μ_n we know that we can apply Lemma 4.3 and Lemma 4.4. By Lemma 4.5 there exists a sequence (β_n) which converges to 1 and an $\varepsilon_1 > 0$ (one can choose $\varepsilon_1 = (1 - \alpha)/2$ because of (25)) independent of the index of the wavelet coefficient such that for $\alpha_n := \sqrt{\varepsilon_1 2 \log n}$ and for all c with $|c| < \alpha_n$

$$\beta_n \leq \frac{\Phi((c, \infty))}{\mu_n((c, \infty))} \text{ and } \beta_n \leq \frac{\Phi((-\infty, c))}{\mu_n((-\infty, c))}. \quad (32)$$

We make now a distinction of cases to prove relation (29), the first case is $\lambda < \alpha_n/2$, the other is $\lambda \geq \alpha_n/2$.

Assume now $\lambda < \alpha_n/2$. For fixed λ define $r_a(x) := (T_\lambda^S(x+a) - a)^2$, r_a is a function with one local minimum with value 0 at λ , if $\lambda = 0$ then the minimum is at $[-\lambda, \lambda]$. Hence $r'_a(x) \geq 0$ for $x \geq \lambda$ and $r'_a(x) \leq 0$ for $x \leq \lambda$. Thus from

$$\int_{-\infty}^{\infty} r_a(x) d\Phi(x) = \int_{-\infty}^{\lambda} (-r'_a(x)) \Phi(x) dx + \int_{\lambda}^{\infty} r'_a(x) (1 - \Phi(x)) dx,$$

and

$$\int_{-\alpha_n}^{\alpha_n} r_a(x) d\mu_n(x) \leq \int_{-\alpha_n}^{\lambda} (-r'_a(x)) \mu_n((-\infty, x]) dx + \int_{\lambda}^{\alpha_n} r'_a(x) \mu_n([x, \infty)) dx,$$

and inequality (32) it follows easily that

$$\frac{\int_{-\infty}^{\infty} r_a(x) d\Phi(x)}{\int_{-\alpha_n}^{\alpha_n} r_a(x) d\mu_n(x)} \geq \beta_n. \quad (33)$$

Further because of Lemma 4.4,

$$\begin{aligned} \int_{\{|x| > \alpha_n\}} r_a(x) d\mu_n(x) &\leq \int_{\{|x| > \alpha_n\}} (\lambda + |x|)^2 d\mu_n(x) \\ &\leq \int_{\{|x| > \alpha_n\}} 4x^2 d\mu_n(x) \\ &\leq 4((\alpha_n^2 + 2)/c_n^2 \exp(-\alpha_n^2/2c_n) + o(\exp(-n^\varepsilon))) \\ &= o(p_\Phi(\lambda, 0)), \end{aligned}$$

with $c_n = 1 - n^{-\varepsilon} \log n/2$. The last identity follows from $\lambda < \alpha_n/2$ via identity (31). Since $p_\Phi(\lambda, 0) \leq p_\Phi(\lambda, a)$, the assertion (29) follows for $\lambda < \alpha_n/2$.

Now the second case:

$$\lambda \geq \alpha_n/2.$$

Choose the smallest $\tilde{\lambda}$ such that

$$p_\Phi(\lambda + 1, 0) \geq p_n(\tilde{\lambda}, 0) \text{ and } \tilde{\lambda} \geq \lambda.$$

It is a simple consequence of Lemma 4.4 and relation (31) that $\lambda/\tilde{\lambda} \rightarrow 1$ uniformly for $\lambda \geq \alpha_n/2$ (remember that we assume $\lambda \leq \lambda_n \sim \sqrt{2 \log n}$). We have to make another distinction of cases, the first case is $|a| < 1$.

Because of $p_n(\tilde{\lambda}, a) \leq a^2 + p_n(\tilde{\lambda}, 0)$ and (30) it follows

$$\begin{aligned} \inf_{|a| \leq 1} \frac{p_\Phi(\lambda, a)}{p_n(\tilde{\lambda}, a)} &\geq \inf_{|a| \leq 1} \frac{a^2 \Phi((-\lambda - 1, \lambda - 1)) + (p_\Phi(\lambda + 1, 0) + p_\Phi(\lambda, 0))/2}{a^2 + p_n(\tilde{\lambda}, 0)} \\ &\geq \inf_{|a| \leq 1} \Phi((-\alpha_n/2 - 1, \alpha_n/2 - 1)) \rightarrow 1, \end{aligned}$$

the second inequality holds since $\lambda \geq \alpha_n/2$.

The case $a > 1$ is more complicated: If $a > 1$ then since $p_\Phi(\lambda, a) \geq p_\Phi(\lambda, 1)$ clearly $p_\Phi(\lambda, a) > 1/2$ and $p(\tilde{\lambda}, a) > 1/2$ if α_n is sufficiently large, further

$$\begin{aligned} \int_{\{|x|>\alpha_n\}} (T_{\tilde{\lambda}}^S(x+a) - a)^2 \mu_n(dx) &\leq \int_{\{|x|>\alpha_n\}} 4(x^2 + \tilde{\lambda}^2) \mu_n(dx) \\ &= o(1), \end{aligned}$$

since $\tilde{\lambda} \sim \lambda \leq \lambda_n \sim \sqrt{2 \log n}$. Like in the paragraph above for equation (33) it easy to see

$$\inf_{\lambda \geq \alpha_n/2} \inf_{a \geq 1} \frac{\int_{-\infty}^{\infty} (T_{\tilde{\lambda}}^S(x+a) - a)^2 \Phi(dx)}{\int_{-\alpha_n}^{\alpha_n} (T_{\tilde{\lambda}}^S(x+a) - a)^2 \mu_n(dx)} \rightarrow 1,$$

thus

$$\liminf_{n \rightarrow \infty} \inf_{\lambda \geq \alpha_n/2} \inf_{a \geq 1} \frac{p_\Phi(\tilde{\lambda}, a)}{p_n(\tilde{\lambda}, a)} \geq 1.$$

To finish the proof we have to show now

$$\liminf_{n \rightarrow \infty} \inf_{\lambda \geq \alpha_n/2} \inf_{a > 1} \frac{p_\Phi(\lambda, a)}{p_\Phi(\tilde{\lambda}, a)} = \liminf_{n \rightarrow \infty} \inf_{\lambda \geq \alpha_n/2} \inf_{a > 1} \frac{\int_{-\infty}^{\infty} (T_\lambda^S(x+a) - a)^2 \Phi(dx)}{\int_{-\infty}^{\infty} (T_{\tilde{\lambda}}^S(x+a) - a)^2 \Phi(dx)} \geq 1. \quad (34)$$

First let us note that if $x \leq \lambda - a$ then clearly $(T_\lambda^S(x+a) - a)^2 \geq (T_{\tilde{\lambda}}^S(x+a) - a)^2$. Again there are two cases, $a < \alpha_n/4$ and $a \geq \alpha_n/4$.

Assume $a < \alpha_n/4$ then for $x > \alpha_n/4$, $x^2 \geq T_{\tilde{\lambda}}((x+a) - a)^2$ and hence $\sup_{a < \alpha_n/4} \int_{\alpha_n/4}^{\infty} (T_{\tilde{\lambda}}^S(x+a) - a)^2 \Phi(dx) = o(1)$. If $x < \alpha_n/4 < \lambda - \alpha_n/4 \leq \lambda - a$, then $(T_\lambda^S(x+a) - a)^2 \geq (T_{\tilde{\lambda}}^S(x+a) - a)^2$, thus the assertion follows for $a < \alpha_n/4$ since $p_\Phi(\lambda, a) > 1/2$.

Now the second case: if $a > \alpha_n/4$ then

$$\inf_{a > \alpha_n/4} \inf_{x \leq \tilde{\lambda} - \alpha_n/4} \frac{T_\lambda^S((x+a) - a)^2}{T_{\tilde{\lambda}}^S((x+a) - a)^2} \geq \left(\frac{\alpha_n/4 - \lambda}{\alpha_n/4 - \tilde{\lambda}} \right) \rightarrow 1.$$

Since

$$\begin{aligned} \int_{\tilde{\lambda} - \alpha_n/4}^{\infty} T_{\tilde{\lambda}}((x+a) - a)^2 \Phi(dx) &\leq \int_{\tilde{\lambda} - \alpha_n/4}^{\infty} (\alpha_n/4 + x)^2 \Phi(dx) \\ &\leq \int_{\alpha_n/4}^{\infty} 2x^2 + \alpha_n^2/8 \Phi(dx) = o(1). \end{aligned}$$

and $p_\Phi(\lambda, a) > 1/2$ the the relation (34) holds for $a > \alpha_n/4$. \square

4.2 Very heavy tails

The conditions for Theorem 4.1 are quite strong, we want to try now to apply wavelet thresholding to noises with rather heavy tails. Under these conditions we will be able to achieve at least the same minimax rate as in the Gaussian case, but the constants are larger. The idea we will employ is simple, we will apply a median filter to our data and then apply wavelet thresholding to this data.

In the paper of Deylon and Juditsky conditions are given that soft thresholding achieves the same minimax rate as in the Gaussian case. In some case a finite third moment is enough. Kovac and Silverman ([30]) used median based filtering for detecting outliers.

In the following given a_1, \dots, a_{2k+1} , $\text{med}(a_1, \dots, a_{2k+1})$ will denote the number x such that $|\{i : a_i \geq x\}| = k + 1$ and $|\{i : a_i \leq x\}| = k + 1$. We will use the abbreviation $\text{med}(a_i, 2k + 1)$ for $\text{med}(a_{i-k}, \dots, a_{i+k})$. If i is smaller than k , then $\text{med}(a_i, 2k + 1) := \text{med}(a_1, \dots, a_{2k+1})$, a similar boundary correction is performed for the largest indexes. The following Lemma, whose proof is clear, makes the advantage of the median filter clear:

Lemma 4.7 *Let X_1, \dots, X_{2k-1} be independent random variables and $c, p > 0$ such that*

$$\max_{i=1, \dots, 2k-1} P(|X_i| > x) \leq \frac{c^p}{x^p},$$

then

$$P(\text{med}(X_1, \dots, X_{2k-1}) > x) \leq \binom{2k-1}{k} \frac{c}{x^{kp}}.$$

Clearly $\text{med}(X_1, \dots, X_{2k-1})$ has moments of order $kp - \varepsilon$, for all $\varepsilon > 0$.

Thus for example the median of 7 Cauchy distributed random variables has moments of order up to $4 - \varepsilon$, $\varepsilon > 0$. The downside of this approach is that we introduce an additional bias. But let us give our main result now, the situation is as usual, we are given $X_i = f_i + n^{-1/2}z_i$, $i = 1, \dots, n$, $n = 2^h$, where f_i is the signal and the z_i are iid random variables with symmetric distribution. Again W_n is a discrete wavelet transform for \mathbb{R}^n , based on a wavelet base with compact support, $a = W_n(f)$ and $a_{j,k} = \sum_{i=1}^n c_{j,k,i} f_i$. In the following let $m > 0$, $1 \leq p, q \leq \infty$, $m > 1/p$ and $s = m + 1/2 - 1/p$.

Theorem 4.8 *Assume $P(z_1 > x) = O(1/x^c)$ for some $c > 0$ and let $A, B > 0$. Then there exists an l and thresholds $\lambda_{j,k}$ such that*

$$\sup_{\substack{\|a\|_{B_{p,q}^m} \leq A \\ \sum_i |f_i - f_{i-1}|^2 \leq B/n}} E \sum_{j,k} |T_{\lambda_{j,k}}^S(W_n(\text{med}(X, 2l+1))_{j,k}) - a_{j,k}|^2 = O\left(n^{-\frac{2m}{2m+1}}\right).$$

We impose the condition $\sum_i |f_i - f_{i-1}|^2 \leq B/n$ to have control over the ℓ_2 -norm of the bias

$$\sum_{i=1}^n (E \text{med}(X_i, 2l + 1) - f_i)^2, \quad (35)$$

which we introduce by filtering the initial data with a median filter. This condition is not strong, most times it follows from the Besov norm condition. But we will take a look at this later. Note that the ℓ_2 -norm of the bias in the wavelet coefficients

$$\sum_{j,k} (EW_n(\text{med}(X, 2l+1))_{j,k} - a_{j,k})^2$$

is the same as (35).

Proof: Since the distribution of the z_i is symmetric $E\text{med}(z_i, 2l+1) = 0$. In the following $b_{j,k}$ denotes the bias and $e_{j,k}$ the pure noise in the wavelet coefficient with index j, k , i.e.

$$W_n(\text{med}(X, 2l+1)) = W_n(f) + (b_{j,k}) + (e_{j,k})$$

and $(e_{j,k}) = W_n(\text{med}(z, 2l+1))$. First we prove that the influence of the bias of our estimation is not too large.

$$\begin{aligned} & E(T_{\lambda_{j,k}}^S (a_{j,k} + b_{j,k} + e_{j,k}) - a_{j,k})^2 \\ &= E(T_{\lambda_{j,k}}^S (a_{j,k} + b_{j,k} + e_{j,k}) - T_{\lambda_{j,k}}^S (a_{j,k} + e_{j,k}) + (T_{\lambda_{j,k}}^S (a_{j,k} + e_{j,k}) - a_{j,k}))^2 \\ &\leq E(T_{\lambda_{j,k}}^S (a_{j,k} + e_{j,k}) - a_{j,k})^2 + b_{j,k}^2 + 2|b_{j,k}| \sqrt{E(T_{\lambda_{j,k}}^S (a_{j,k} + e_{j,k}) - a_{j,k})^2}, \end{aligned}$$

since $|T_\lambda^S(x+a) - T_\lambda^S(x)| \leq |a|$. Thus

$$\begin{aligned} & \sum_{j,k} E(T_{\lambda_{j,k}}^S (a_{j,k} + b_{j,k} + e_{j,k}) - a_{j,k})^2 \\ &\leq \sum_{j,k} E(T_{\lambda_{j,k}}^S (a_{j,k} + e_{j,k}) - a_{j,k})^2 + \sum_{j,k} b_{j,k}^2 \\ &\quad + 2 \sqrt{\sum_{j,k} b_{j,k}^2} \sqrt{\sum_{j,k} E(T_{\lambda_{j,k}}^S (a_{j,k} + e_{j,k}) - a_{j,k})^2}. \end{aligned}$$

Note that $\sum_{j,k} b_{j,k}^2 = \sum_{i=1}^n |E\text{med}(X_i, 2l+1) - f_i|^2$, but for $l < i \leq n-l$,

$$\begin{aligned} & |E\text{med}(X_i, 2l+1) - f_i| \\ &= |E(\text{med}(X_i, 2l+1) - (\text{med}(z_i, 2l+1) + f_i))| \\ &\leq E|\text{med}(z_{i-l} + f_{i-l} - f_i, \dots, z_{i+l} + f_{i+l} - f_i) - \text{med}(z_i, 2l+1)| \\ &\leq E \max_{j=-l, \dots, l} |f_{i+j} - f_i| \\ &\leq \sum_{j=-l+1}^l |f_{i+j} - f_{i+j-1}|. \end{aligned}$$

If $i \leq l$ or $i > n-l$ then $|E\text{med}(X_i, 2l+1) - f_i| \leq \sum_{j=2}^{2l+1} |f_j - f_{j-1}|$ respectively $\leq \sum_{j=n-2l+1}^n |f_j - f_{j-1}|$. Hence

$$\sum_{i=1}^n |E\text{med}(X_i, 2l+1) - f_i|^2 \leq \sum_{i=l+1}^{n-l} 2l \sum_{j=-l+1}^l |f_{i+j} - f_{i+j-1}|^2$$

$$\begin{aligned}
& + 2l^2 \sum_{j=2}^{2l+1} |f_j - f_{j-1}|^2 + 2l^2 \sum_{j=n-2l+1}^n |f_j - f_{j-1}|^2 \\
& \leq 8l^2 \sum_{i=2}^n |f_i - f_{i-1}|^2 = O(1/n).
\end{aligned}$$

This implies if we choose a fixed median filter, then it suffices to prove

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} E \sum_{j,k} (T_{\lambda_{j,k}}^S(W_n(f_i + \tilde{z}_{j,k}) - a_{j,k}))^2 = O\left(n^{-\frac{2m}{2m+1}}\right)$$

where $\tilde{z}_i := \text{med}(z_i, D)$ and D is chosen such that $E|\tilde{z}_1|^L < \infty$ and L satisfies (depending on p) the moment conditions of Theorem 4.1. So we could apply Theorem 4.1 if the \tilde{z}_i were independent. The following two lemmas are adaptations to the new situation.

Lemma 4.9 *Let X_1, \dots, X_n be bounded random variables, $\|X_i\|_\infty < K$ and X_i, X_j are independent if $|i - j| \geq D$. Let $S_j = \sum_{i=0}^{\lfloor (n-1)/D \rfloor} X_{iD+j}$, $j = 1, \dots, D$, $\sigma_j = \sqrt{ES_j^2}$, and $\sigma_{max} = \max_{j=1, \dots, D} \sigma_j$ (we simply set $X_k = 0$ for $k > n$). Then*

$$P\left(\sum_{i=1}^n X_i > \lambda\right) \leq D \begin{cases} \exp\left(\frac{\lambda^2}{2D^2\sigma_{max}^2} \frac{1}{2}\right) & : \lambda \leq \frac{\sigma_{max}^2 D}{K} \\ \exp\left(\frac{-\lambda}{4KD}\right) & : \lambda \geq \frac{\sigma_{max}^2 D}{K} \end{cases}. \quad (36)$$

Proof: Note that the S_j are sums of independent random variables.

$$\begin{aligned}
P\left(\sum_{i=1}^n X_i > \lambda\right) & \leq \sum_{i=1}^D P(S_i > \lambda/D) \\
& = \sum_{i=1}^D P(S_i/\sigma_i > \lambda/(\sigma_i D)) \\
& \leq \sum_{i=1}^D \begin{cases} \exp\left(\frac{\lambda^2}{2D^2\sigma_i^2} \left(1 - \frac{\lambda K}{2D\sigma_i^2}\right)\right) & : \lambda \leq \frac{\sigma_i^2 D}{K} \\ \exp\left(\frac{-\lambda}{4KD}\right) & : \lambda \geq \frac{\sigma_i^2 D}{K} \end{cases} \\
& \leq \sum_{i=1}^D \begin{cases} \exp\left(\frac{-\lambda^2}{2D^2\sigma_i^2} \frac{1}{2}\right) & : \lambda \leq \frac{\sigma_i^2 D}{K} \\ \exp\left(\frac{-\lambda}{4KD}\right) & : \lambda \geq \frac{\sigma_i^2 D}{K} \end{cases} \\
& \leq D \begin{cases} \exp\left(\frac{-\lambda^2}{4D^2\sigma_{max}^2}\right) & : \lambda \leq \frac{\sigma_{max}^2 D}{K} \\ \exp\left(\frac{-\lambda}{4KD}\right) & : \lambda \geq \frac{\sigma_{max}^2 D}{K} \end{cases},
\end{aligned}$$

where the last inequality holds since if $\lambda \leq \sigma_{max}^2 D/K$ then $\lambda^2/(4D^2\sigma_{max}^2) \leq \lambda/(4KD)$. \square

Lemma 4.10 *Let X_1, \dots, X_n be random variables with $EX_i = 0$, $E|X_i|^3 \leq Cn^{-3/2}$, $i = 1, \dots, n$ and X_i, X_j are independent if $|i - j| \geq D$. Let $S_j = \sum_{i=0}^{(n-1)/D} X_{iD+j}$, $j = 1, \dots, D$, $\sigma_j = \sqrt{ES_j^2}$ and $\sigma_{max} = \max_{j=1, \dots, D} \sigma_j$ (again $X_k := 0$ for $k > n$). Then for all $\varepsilon < 1$*

$$\limsup_{n \rightarrow \infty} \sup_{0 \leq x \leq \varepsilon D \sigma_{max} \sqrt{2 \log n}} \frac{P(\sum_{i=1}^n X_i \leq -x)}{\Phi\left(\left(-\infty, \frac{-x}{\sigma_{max} D}\right)\right)} \leq D$$

and

$$\limsup_{n \rightarrow \infty} \sup_{0 \leq x \leq \varepsilon D \sigma_{max} \sqrt{2 \log n}} \frac{P(\sum_{i=1}^n X_i \geq x)}{\Phi\left(\left(\frac{x}{\sigma_{max} D}, \infty\right)\right)} \leq D$$

where the rate only depends on ε and C .

Proof: If $x \geq 0$ then,

$$\begin{aligned} P\left(\sum_{i=1}^n X_i \geq x\right) &\leq \sum_{j=1}^D P(S_j > x/D) \\ &\leq \sum_{j=1}^D P(S_j/\sigma_j > x/(\sigma_{max} D)). \end{aligned}$$

A similar inequality holds for $x \leq 0$. Since the S_j are sums of independent random variables, the assertion follows now with Lemma 4.5. \square

It is clear that in the situation of Lemma 4.10 for $x \geq 0$

$$\frac{P(\sum_{i=1}^n X_i \leq x)}{\Phi\left(\left(-\infty, \frac{x}{\sigma_{max} D}\right)\right)} \leq 2 \text{ and } \frac{P(\sum_{i=1}^n X_i \geq -x)}{\Phi\left(\left(\frac{-x}{\sigma_{max} D}, \infty\right)\right)} \leq 2.$$

With the help of Lemma 4.10 it is possible to prove a similar variation of Lemma 4.4.

The rest of the proof is quite similar to the proof of Theorem 4.1. Now let $\tilde{z}_i := \text{med}(z_i, D)$ and D is chosen such that $E|\tilde{z}_1|^L < \infty$ and L satisfies (depending on p) the moment conditions of Theorem 4.1. Thus by the same reasoning as in the proof of Theorem 4.1 we can assume that the \tilde{z}_i are bounded by n^ε for some $\varepsilon > 0$, for this the independence of the random variables was not needed.

Important in the proof of Theorem 4.1 was the distribution of the noise in the wavelet coefficients. We denote the coefficients of the wavelet transform by $(c_{j,k,i})$, i.e. $a_{j,k} = \sum_i c_{j,k,i} f_i$. At the boundary we have the problem that $\tilde{z}_1 = \dots = \tilde{z}_{(D+1)/2}$ and $z_{n-(D-1)/2} = \dots = \tilde{z}_n$. But

$$e_{j,k} = \left(\tilde{z}_1 \sum_{i=1}^{(D+1)/2} c_{j,k,i} \right) + \sum_{i=(D+1)/2+1}^{n-(D-1)/2-1} c_{j,k,i} \tilde{z}_i + \left(\tilde{z}_n \sum_{i=n-(D-1)/2}^n c_{j,k,i} \right)$$

and the last sum satisfies the conditions of the Lemmas 4.10 and 4.9. Anyway only about $O(\log n)$ wavelet coefficients are affected by this problem. If we do not threshold these coefficients, the risk would increase at most by $O((\log n)/n)$ and this is small compared to the minimax risk. Let $e_{j,k,l} = \sum_{i=1}^{n/D} c_{j,k,iD+l} \tilde{z}_{iD+l}$, $\sigma_{j,k,l}^2 = Ee_{j,k,l}^2$ and $\sigma_{max,j,k}^2 = \max_l \sigma_{j,k,l}^2$, clearly $\sigma_{max,j,k}^2 \leq E\tilde{z}_1^2$. Given this and the Lemmas 4.9 and 4.10 we can proceed now as in the proof of Theorem 4.1. Hence with the right thresholds we can achieve D times the performance of the risk as for Gaussian noise with variance $2\sigma_{max}^2 D^2$, where σ_{max}^2 is the maximum of the $\sigma_{max,j,k}$ where the corresponding wavelet coefficient is not discarded. \square

Important in the Lemmas 4.10 and 4.9 was the term σ_{max} , we want to show now that in general

$$\sigma_{max,j,k} \approx E\tilde{z}_1^2/D \quad (37)$$

Let β be the Hölder continuity of the wavelet base. Let $h = \log_2 n$, from (10) we already know that

$$|2^{(h-j)/2} c_{j,k,i} - \psi(2^{j-h}i - k)| \leq C_1 2^{(j-h)\beta},$$

and further

$$|\psi(2^{j-h}i - k) - \psi(2^{j-h}(i-1) - k)| \leq C_2 2^{(j-h)\beta},$$

for some constants C_1, C_2 since ψ is Hölder continuous with exponent β . Thus

$$|c_{j,k,i} - c_{j,k,i+1}| \leq (2C_1 + C_2) 2^{(j-h)(\beta+1/2)}$$

and

$$|c_{j,k,i}^2 - c_{j,k,i+1}^2| = |c_{j,k,i} - c_{j,k,i+1}| |c_{j,k,i} + c_{j,k,i+1}| \leq C_3 2^{(j-h)/2} 2^{(j-h)(\beta+1/2)},$$

since the $|c_{j,k,i}| = O(2^{(j-h)/2})$, C_3 is a constant. Because of $\sigma_{j,k,l}^2 = E\tilde{z}_1^2 \sum_i c_{j,k,iD+l}^2$ and $|\{i : c_{j,k,i} \neq 0\}| = O(2^{h-j})$ which is a consequence of using compactly supported wavelets (see relations (1, 2)),

$$\begin{aligned} |\sigma_{j,k,l}^2 - \sigma_{j,k,l+1}^2| &= |E\tilde{z}_1^2 \sum_i (c_{j,k,iD+l}^2 - c_{j,k,iD+l+1}^2)| \\ &= O(E\tilde{z}_1^2 2^{(j-h)\beta}). \end{aligned}$$

Since $\sum_l \sigma_{j,k,l}^2 = E\tilde{z}_1^2 \sum_i c_{j,k,i}^2 = E\tilde{z}_1$, the $\sigma_{j,k,l}^2$ are all about of the same size and thus $\sigma_{max,j,k}^2$ is of the size $\approx E\tilde{z}_1^2/D$.

Now we turn to the question when

$$\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 = O(1/n) \quad (38)$$

follows from $\|a\|_{B_{p,q}^m} \leq A$. If $m \leq 1$ then assume $\beta \geq m$, where β is the Hölder continuity of the wavelet, since otherwise the characterization of smoothness via wavelets does not make sense. Again $h = \log_2 n$. Since for a constant $C_1 > 0$,

$$|c_{j,k,i} - c_{j,k,i+1}| \leq C_1 2^{(j-h)(1/2+\beta)}$$

and $|\{i : c_{j,k,i} \neq 0\}| = O(2^{h-j})$, it follows

$$\sum_{i=1}^{n-1} |c_{j,k,i} - c_{j,k,i+1}|^2 \leq C_2 2^{2\beta(j-h)},$$

where C_2 is another constant. Note that since the wavelet transform is orthogonal, $f_i = \sum_{j,k} a_{j,k} c_{j,k,i}$. Thus

$$\begin{aligned} \sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 &= \sum_{i=1}^{n-1} \left(\sum_{j,k} a_{j,k} (c_{j,k,i} - c_{j,k,i+1}) \right)^2 \\ &\leq \sum_{i=1}^{n-1} h \sum_{j=0}^{h-1} \left(\sum_k a_{j,k} (c_{j,k,i} - c_{j,k,i+1}) \right)^2 \\ &\leq \sum_{i=1}^{n-1} h \sum_{j=0}^{h-1} C_3 \sum_k (a_{j,k} (c_{j,k,i} - c_{j,k,i+1}))^2 \end{aligned}$$

since $\#\{k : c_{j,k,i} \neq 0 \text{ or } c_{j,k,i+1} \neq 0\} = O(1)$, see (1, 2), C_3 is a constant

$$\begin{aligned} &= hC_3 \sum_{j=0}^{h-1} \sum_k a_{j,k}^2 \sum_i (c_{j,k,i} - c_{j,k,i+1})^2 \\ &\leq hC_3 C_2 \sum_{j=0}^{h-1} 2^{2\beta(j-h)} \sum_k a_{j,k}^2 \\ &\leq hC_3 C_2 A^2 \sum_{j=0}^{h-1} 2^{2\beta(j-h)} \begin{cases} 2^{-2jm} & : p \geq 2 \\ 2^{-2js} & : p \leq 2 \end{cases} \end{aligned}$$

where the last inequality follows from the proof of Lemma 4.2. Thus if $\beta = 1$ and $m \geq 1$ respectively $s \geq 1$ then the last term is equal to $O(hn^{-2})$. If $m \leq \beta$ respectively $s \leq \beta$ then the last term is equal to $O(hn^{-2m})$ respectively $O(n^{-2s})$. Hence for $p \geq 2$ we obtain

$$\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 = O(\log n / n^{-(2m \wedge 2)}), \quad (39)$$

and for $p \leq 2$ we obtain

$$\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 = O(\log n / n^{-(2s \wedge 2)}). \quad (40)$$

Because of the additional condition $m > 1/p$, $2s > 1$, for $p \leq 2$ condition 38 always follows. If $p \geq 2$ then $m > 1/2$ is necessary for condition (38).

Remark 4.11 *This last proof shows how to deal with noise that is not independent, but where e_i and e_j are independent if $|i - j| \geq D$, where D is a fixed*

constant. The Lemmas 4.9 and 4.10 are applicable and then it is easy to follow the line of the proof of Theorem 4.8. So if the tail behavior of the noise is nice then the minimax rate for this problem is again the same as for Gaussian noise. It is not necessary that the noise in the random variables converges in distribution to a normal random variable, only the bounds in the Lemmas 4.10 and 4.9 are needed. The approach via large deviation results for the wavelet coefficients is also possible for other kind of correlated noise. In [24] Johnstone and Silverman investigated thresholding for stationary Gaussian noise. They considered the following model, we observe $Y_i = f_i + e_i$ where (e_i) is a stationary Gaussian noise process with covariance function $r(k) \sim (|k|^{-\alpha})$, $\alpha > 1/2$. If $\alpha > 1$ then one can show that the performance of soft thresholding is not that different from the iid case. The noise in the wavelet coefficients has a normal distribution and the variance is bounded from above by a fixed constant which is smaller than $\sum_{k \in \mathbb{Z}} |r(k)|$. Of course the wavelet coefficients are not independent, but this is not necessary. If $\alpha < 1$ the situation is different, then it can be shown that the variance of the wavelet coefficients is not uniformly bounded.

It is simple to deal with stationary Gaussian noise, the distribution of the noise is determined solely by the covariance function. It is not clear what is meant if one talks about correlated non-Gaussian noise. Stationary Gaussian processes with $\sum_{k \in \mathbb{Z}} |r(k)| < \infty$ can be presented as moving average of a Gaussian white noise process. Thus the corresponding concept for correlated non-Gaussian noise is a moving average of a sequence of iid random variables. So we assume

$$e_k = \sum_i a_{i-k} w_i, \quad k \in \mathbb{Z},$$

where the w_i are iid random variables, additionally we assume $\sum_{k \in \mathbb{Z}} |a_k| < \infty$. If $z_{j,k}$ is the noise in the wavelet coefficient with index (j, k) then

$$z_{j,k} = \sum_i c_{j,k,i} e_i = \sum_i c_{j,k,i} \sum_l a_{l-i} w_l = \sum_l \left(\sum_i a_{l-i} c_{j,k,i} \right) w_l.$$

Note that

$$\begin{aligned} \sum_l \left(\sum_i a_{l-i} c_{j,k,i} \right)^2 &= \sum_l \sum_u \sum_v c_{j,k,u} c_{j,k,v} a_{l-u} a_{l-v} \\ &= \sum_u \sum_v c_{j,k,u} c_{j,k,v} \sum_l a_{l-u} a_{l-v} \\ &= \sum_m \sum_u c_{j,k,m} c_{j,k,m+u} r(u) \\ &\leq \sum_u |r(u)| \|c_{j,k,\cdot}\|_2^2 = \sum_u |r(u)|. \end{aligned}$$

Further $\max_l |\sum_i a_{l-i} c_{j,k,i}| \leq \max_i |c_{j,k,i}| \sum_i |a_i|$. Hence if the tail of the w_l decays fast enough, i.e. the density has exponential decay or has compact support,

then the $z_{j,k}$ converge in distribution to a normal distribution fast enough and soft thresholding should work as well as for Gaussian noise which has the same covariance structure. However one has to note one thing, the whole setup is rather artificial, since knowledge of the a_i and the distribution of the w_j is required. Another point is, why should the covariance structure in the noise remain the same if the sampling rate is increased? Further, if the noise is filtered white noise, why is the signal not filtered?

4.3 Very thin tails

We want to see now what can be achieved with non-linear filtering for other types of noise. We give two examples where thresholding of wavelet coefficients is not the best method. We assume the same model as usual, $X_i = f_i + z_i/\sqrt{n}$, $i = 1, \dots, n = 2^h$ where the z_i are iid bounded random variables. Let $a = W_n(f)$ where W_n is a discrete wavelet transform based on a wavelet base with Hölder continuity 1. We look at the minimax risk for the set $\|a\|_{B_{p,q}^m} \leq A$ with $m \geq 1/2$ and $m > 1/p$, again $s = m + 1/2 - 1/p$. By the same reasoning as in the proof of Theorem 3.15, it is easy to show that for this type of noise, soft thresholding has the same minimax rate as it would have for Gaussian noise with the same variance.

Our estimator for f_i based on the X_i will be

$$\hat{f}_i := \max_{j=0, \dots, D-1} X_{i+j} - \frac{c_D}{\sqrt{n}}, \quad i = 1, \dots, n - D + 1,$$

and for $i > n - D + 1$, $\hat{f}_i = \hat{f}_{n-D+1}$, here $c_D := E \max_{i=0, \dots, D-1} z_i$. Thus for $i \leq n - D + 1$,

$$\hat{f}_i - f_i = \max_{j=0, \dots, D-1} \left(f_{i+j} - f_i + \frac{1}{\sqrt{n}}(z_{i+j} - c_D) \right),$$

hence

$$\hat{f}_i - f_i \leq \max_{j=0, \dots, D-1} |f_{i+j} - f_i| + \max_{j=0, \dots, D-1} \frac{1}{\sqrt{n}}(z_{i+j} - c_D)$$

and

$$\hat{f}_i - f_i \geq - \max_{j=0, \dots, D-1} |f_{i+j} - f_i| + \max_{j=0, \dots, D-1} \frac{1}{\sqrt{n}}(z_{i+j} - c_D).$$

It follows that

$$|\hat{f}_i - f_i| \leq \max_{j=0, \dots, D-1} |f_{i+j} - f_i| + \frac{1}{\sqrt{n}} \left| \max_{j=0, \dots, D-1} z_{i+j} - c_D \right|,$$

and

$$\begin{aligned} E|\hat{f}_i - f_i|^2 &\leq 2 \left(\sum_{j=1, \dots, D-1} |f_{i+j} - f_{i+j-1}| \right)^2 + \frac{2}{n} E \left(\max_{j=0, \dots, D-1} z_{i+j} - c_D \right)^2 \\ &\leq 2D \sum_{j=1, \dots, D-1} |f_{i+j} - f_{i+j-1}|^2 + \frac{2}{n} E \left(\max_{j=0, \dots, D-1} z_{i+j} - c_D \right)^2. \end{aligned}$$

A similar computation for $i > n - D + 1$ gives

$$E|\widehat{f}_i - f_i|^2 \leq 2D \sum_{j=m-D+1}^n |f_j - f_{j-1}|^2 + \frac{2}{n} E \left(\max_{j=0, \dots, D-1} z_{i+j} - c_D \right)^2.$$

Hence

$$E \sum_{i=1}^n |\widehat{f}_i - f_i|^2 \leq 4D^2 \sum_{i=1}^{n-1} |f_{i+1} - f_i|^2 + 4\text{var}(\max_{j=0, \dots, D-1} z_j).$$

From the previous section (equations (39) and (40)) we know that $\sum_i |f_{i+1} - f_i|^2 = O(\log nn^{-(2m \wedge 2)})$ or $O(\log nn^{-(2s \wedge 2)})$, depending on p . If the distribution of the z_i is $(\delta_{-1} + \delta_1)/2$, then $E \max_{j=0, \dots, D-1} z_j = 1 - 1/2^{D-1}$ and $E(\max_{j=0, \dots, D-1} z_j)^2 = 1$ and thus $\text{var}(\max_{j=0, \dots, D-1} z_j) = 1/2^{D-2} - 1/(2^{2D-2})$. Hence for $p \geq 2$

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} E \sum_{i=1}^n |\widehat{f}_i - f_i|^2 \leq C(D^2 \log nn^{-(2m \wedge 2)} + 2^{-D}), \quad (41)$$

where C is a constant. The rate in (41) is minimized by choosing $D = 2 \log_2 n$ and thus

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} E \sum_{i=1}^n |\widehat{f}_i - f_i|^2 = O\left(\frac{(\log n)^3}{n^{2m \wedge 2}}\right), \quad (42)$$

If $p \leq 2$ then the right side of identity (42) must be replaced by $O\left(\frac{(\log_2 n)^3}{n^{2s \wedge 2}}\right)$.

The second example is if the z_i are random variables which are uniformly distributed on $[-1, 1]$. The variance of $\max_{j=1, \dots, D} z_j$ is of size $O(1/D^2)$. Now we have

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} E \sum_{i=1}^n |\widehat{f}_i - f_i|^2 \leq C(D^2 \log nn^{-(2m \wedge 2)} + D^{-2}), \quad (43)$$

(respectively $\leq C(D^2 \log nn^{-(2s \wedge 2)} + D^{-2})$), again C is a constant. Assume now that $m \geq 1$, if $p \geq 2$ and $s \geq 1$ if $p \leq 2$. So if we take $D = \sqrt{n}$ then

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} E \sum_{i=1}^n |\widehat{f}_i - f_i|^2 = O\left(\frac{\log n}{n}\right), \quad (44)$$

which is clearly better than the minimax rate for soft thresholding. The case $m < 1$ or $s < 1$ is more complicated, since then

$$\sup_{\|a\|_{B_{p,q}^m} \leq A} \sum_{i=1}^{n-1} |f_i - f_{i+1}| = \begin{cases} O(n^{-2m}) & : p \geq 2 \\ O(n^{-2s}) & : p \leq 2 \end{cases}.$$

Hence we have to minimize

$$\frac{D^2 \log n}{n^{2m}} + \frac{1}{D^2} \text{ if } p \geq 2 \text{ and } \frac{D^2 \log n}{n^{2s}} + \frac{1}{D^2} \text{ if } p \leq 2.$$

If $p \geq 2$ choosing $D = n^{m/2}/\sqrt[4]{\log n}$ leads to a rate of $O(\sqrt{\log n}/n^m)$. This is better than the soft thresholding minimax rate $O(n^{2m/(2m+1)})$ in the Gaussian model for $m > 1/2$. If $p \leq 2$ then choosing $D = n^{s/2}/\sqrt[4]{\log n}$ gives a rate $O(\sqrt{\log n}/n^s)$ and this is better than the minimax rate for soft thresholding $O(n^{2m/(2m+1)})$ in the Gaussian model if $p > 1/(m + 1/2 - 2m/(2m + 1))$.

5 Unbiased risk estimation

In the function space approach, the thresholds depend not only on n , but also on the Besov space to which the target functions belong and also on the Besov norm of the function. In practice it is often not known which threshold is appropriate, we do not know to which function space the function belongs nor its function space norm. Donoho and Johnstone developed a method called sureshrink where the threshold is chosen automatically (see [16]). Their method is based on Stein's unbiased risk estimate: Let X_i , $i = 1, \dots, n$, be iid $N(0, \sigma^2)$ random variables and $g : \mathbb{R}^n \rightarrow \mathbb{R}$, where $g = (g_1, \dots, g_n)$ is weakly differentiable, then for all $a \in \mathbb{R}^n$,

$$E\|X + a + g(X + a) - a\|^2 = n\sigma^2 + E\|g(X + a)\|^2 + 2\sigma^2 \sum_{i=1}^n \frac{\partial}{\partial x_i} g_i(X + a). \quad (45)$$

Thus the risk of the estimator $x + g(x)$ can be estimated unbiasedly by $n\sigma^2 + g(x)^2 + 2\sigma^2 \sum_{i=1}^n \partial/\partial x_i g_i(x)$. This estimate is useful only if the variance of the risk estimate is small compared to the actual risk. This is especially the case if $g_i(\cdot)$ only depends on X_i , then the strong law of large numbers ensures this.

The sureshrink method works now as follows: for each level (except the highest levels) in the noisy wavelet transform, the largest threshold smaller than $\sqrt{2 \log n}$ is chosen which minimizes the unbiased risk estimate. For soft thresholding finding this minimum is simple and takes $O(n \log n)$ time.

As we have seen in one of the previous sections, the central limit theorem works fast for wavelet coefficients, so it is reasonable to expect that this approach works for other types of noises too. But it is still an interesting problem to find unbiased risk estimates for other types of distributions.

For Gaussian random variables the existence of such estimates is based on the following identity

$$\int_{\mathbb{R}} g'(x) e^{-x^2/2} dx = \int_{\mathbb{R}} x g(x) e^{-x^2/2} dx,$$

which is sometimes called Stein's identity. We try below to find a similar identity for non-Gaussian random variables. Then g' is replaced by $K(g)$ and K is an operator which commutes with translations. For simplicity we concentrate on the one-dimensional case.

Let f be a density on \mathbb{R} , with variance σ^2 and mean 0. Let $d(x) = x + g(x)$ be an estimator in the location model induced by f . Let $F = \{f * \delta_a : a \in \mathbb{R}\}$, $L^2(F)$ and $L^1(F)$ have the canonical meaning. We want to estimate the risk of d unbiasedly:

$$\begin{aligned} & \int_{\mathbb{R}} (d(x + a) - a)^2 f(x) dx \\ &= \int_{\mathbb{R}} g(x + a)^2 f(x) dx + \int_{\mathbb{R}} x^2 f(x) dx + 2 \int_{\mathbb{R}} x g(x + a) f(x) dx. \end{aligned}$$

The first summand can be estimated unbiasedly, the second is a constant, so we need a function $h \in L^1(F)$ such that

$$\int_{\mathbb{R}} h(x+a)f(x)dx = \int_{\mathbb{R}} xg(x+a)f(x)dx. \quad (46)$$

If g is a polynomial the right-hand side of (46) is a polynomial in a . It is then well-known that there exists an h such that (46) holds. But if for example $g(x) = T_{\lambda}^S(x)$ then g does not even have a power series expansion. On the other hand h does not have to be unique, this is possible if there is a function q with $q * f = 0$, which is possible if \widehat{f} has zeros.

Assume \widehat{f} does not have zeros. By computing the generalized Fourier transform of both sides of

$$\int_{\mathbb{R}} g(-x+a)(-x)f(-x)dx = \int_{\mathbb{R}} h(-x+a)f(-x)dx, \quad (47)$$

we get:

$$\widehat{g}(w)\widehat{f}'(-w)/i = \widehat{h}(w)\widehat{f}(-w). \quad (48)$$

This identity shows that if \widehat{g} converges to 0 fast enough, e.g., if \widehat{g} has compact support, then there exists an h such that (46) holds. Since \widehat{f} has no zeros, h is uniquely determined. Hence the set

$$U_f := \left\{ g \in L^2(F) : \exists h \in L^1(F), \int_{\mathbb{R}} h(x+a)f(x)dx = \int_{\mathbb{R}} xg(x+a)f(x)dx, \forall a \in \mathbb{R} \right\}$$

is a vector space and clearly there is a unique linear mapping $K_f : U_f \rightarrow L^1(f)$ with

$$\int_{\mathbb{R}} K_f(g)(x+a)f(x)dx = \int_{\mathbb{R}} g(x+a)xf(x)dx.$$

Let us note some properties of the operators K_f :

Theorem 5.1 *Let f, f_1, f_2 be densities with finite second moment, assume K_f, K_{f_1} and K_{f_2} are well defined in the sense of the previous paragraph, then for all $b \in \mathbb{R}$ and $g \in U_f$, respectively $g \in U_{f_1 * f_2}$:*

1. $K_f(g(\cdot + b)) = K_f(g)(\cdot + b)$
2. $K_{f * \delta_b}(g) = K_f(g) + bg(\cdot)$
3. $K_{f_1 * f_2}(g) = K_{f_1}(g) + K_{f_2}(g)$
4. $K_{bf(\cdot/b)}(g) = K_f(g(\cdot/b))(\cdot/b)/b$ for $b > 0$.

Proof:

1.: $\int_{\mathbb{R}} g(x+a+b)xf(x)dx = \int_{\mathbb{R}} K_f(g)(x+a+b)f(x)dx$ and hence $K_f(g(\cdot+b)) = K_f(g)(\cdot+b)$.

2.:

$$\begin{aligned} \int_{\mathbb{R}} xg(x+a)f(x-b)dx &= \int_{\mathbb{R}} (K_f(g)(x+a+b) + bg(x+a+b))f(x)dx \\ &= \int_{\mathbb{R}} (K_f(g)(x+a) + bg(x+a))f(x-b)dx. \end{aligned}$$

3.: let h_1, h_2 be such that $\int_{\mathbb{R}} g(x+a)xf_i(x)dx = \int_{\mathbb{R}} h_i(x+a)f(x)dx$, $i = 1, 2$, then

$$\begin{aligned} &\int_{\mathbb{R}} (h_1 + h_2)(z+a)(f_1 * f_2)(z)dz \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (h_1(x+y+a) + h_2(x+y+a))f_1(x)f_2(y)dx dy \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} h_1(x+y+a)f_1(x)dx f_2(y)dy + \int_{\mathbb{R}} \int_{\mathbb{R}} h_2(y+x+a)f_2(y)dy f_1(x)dx \\ &= \int_{\mathbb{R}} g(z+a)z(f_1 * f_2)(z)dz. \end{aligned}$$

4.:

$$\begin{aligned} \int_{\mathbb{R}} g(x+a)x(bf(bx))dx &= \int_{\mathbb{R}} g(x/b+a)xf(x)/bdx \\ &= \int_{\mathbb{R}} g((x+ba)/b)x/bf(x)dx \\ &= \int_{\mathbb{R}} K_f(g(\cdot/b))(x+ba)/bf(x)dx \\ &= \int_{\mathbb{R}} K_f(g(\cdot/b))((x+a)b)/b(bf(xb))dx, \end{aligned}$$

and thus $K_{bf(\cdot/b)}(g) = K_f(g(\cdot/b))(\cdot/b)/b$. \square

Note that the third property is very useful for wavelet analysis, since the noise in a wavelet coefficient is a convolution of the noise in the original data.

For the normal distribution with variance 1 this operator K is defined by $K(g) = g'$, i.e. $K = D$, where D is the differential operator. In general it is quite complicated to compute K , but from equation (48) we deduce formally

$$\widehat{K_f(g)}(w) = \widehat{g}(w)\widehat{f}'(-w)/(\widehat{f}(-w)i).$$

This gives a hint, that h can be computed by a convolution of the estimator and a function or measure, which is the inverse Fourier transform of $\widehat{f}'(-w)/(\widehat{f}(-w)i)$. Of course this is a rather handwaving argument. Assume $K_f(g) := K_f * g$ where

$K_f \in L^1(\mathbb{R})$ and $\widehat{K}_f = \widehat{f}'(-\cdot)/(\widehat{f}(-\cdot)i)$. If $g \in L^\infty(\mathbb{R})$, then $K_f * g$ does what it is supposed to do:

$$\begin{aligned}
\int_{\mathbb{R}} (K_f * g)(x+a)f(x)dx &= \int_{\mathbb{R}} \int_{\mathbb{R}} K_f(x-t)g(t+a)dt f(x)dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} K_f(x-t)f(x)dx g(t+a)dt \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} K_f(-(t-x))f(x)dx g(t+a)dt \\
&= \int_{\mathbb{R}} (K_f(-\cdot) * f(\cdot))(t)g(t+a)dt \\
&= \int_{\mathbb{R}} tf(t)g(t+a)dt,
\end{aligned}$$

where the last equality follows from the construction of K :

$$\widehat{K}(-\cdot)\widehat{f} = \frac{\widehat{f}'}{i} = \widehat{(f)\text{id}}.$$

If K_f is known, then it is still a problem to compute $K_f(g)$, $K_f(g)$ is not necessarily simple like g' . If $g = \sum_i g_i$ and the $K_f(g_i)$ are easy to compute, then we can compute $K_f(g)$ because K_f is linear. For example we can take $g_a^+(x) = (x-a)_+$ and $g_a^-(x) := (x-a)_-$ as simple building blocks for functions. Note that $K_f(g_a^+)(x) = K_f(g_0^+)(x-a)$ and $K_f(g_0^-(x)) = \sigma^2 - K_f(g_0^+)$ if $\int_{\mathbb{R}} xf(x)dx = 0$ and $\int_{\mathbb{R}} x^2f(x)dx = \sigma^2$. For example the soft thresholding estimator and T_λ^M have the following decompositions:

$$T_\lambda^S(x) = x - g_0^+(x) + g_\lambda^+(x) - g_0^-(x) + g_{-\lambda}^-(x), \quad (49)$$

$$T_\lambda^M(x) = x - g_0^+(x) + 2g_{\lambda/2}^+(x) - g_\lambda^+(x) - g_0^-(x) + 2g_{-\lambda/2}^-(x) - g_{-\lambda}^-(x).$$

Another example is obtained if $g : \mathbb{R}^+ \rightarrow \mathbb{R}$, $g \in C^2$, and $g(0) = 0$, then $g(x) = g'(0_+)x_+ + \int_0^\infty (x-a)_+g''(a)da$.

A very simple example are compound Poisson distributions, let F be a compound Poisson distribution with Fourier transform $\exp((\Psi(x) - 1)\lambda)$, where Ψ is the characteristic function of the density f . Then $\widehat{K}_F := \lambda\Psi'(-w)/i$ and thus $K_F(x) = -\lambda f(-x)x$, i.e. $K_F(g) = K_F * g$.

This example leads to infinitely divisible distributions:

Theorem 5.2 *Let f be an infinitely divisible density with finite second moment, i.e.*

$$\widehat{f}(t) = \exp\left(\int_{\mathbb{R}} \left(\frac{\exp(ixt) - 1 - ixt}{x^2}\right) M(dx) + ibt\right),$$

where M is a finite positive measure. Assume $M(\{0\}) = 0$ and $b = 0$. If g is Lipschitz continuous then

$$K(g)(t) := \int_{\mathbb{R}} \frac{g(t+x) - g(t)}{x} M(dx)$$

is well defined and $K(g)$ is bounded and continuous. Then

$$\int_{\mathbb{R}} K(g)(x+a)f(x)dx = \int_{\mathbb{R}} g(x+a)xf(x)dx.$$

Proof: It is clear that $K(g)$ is well defined, bounded and continuous. If g has compact support then $\int_{\mathbb{R}} |g(x+y) - g(x)|/y|M(dy)$ and $K(g)$ are in $L^1(\mathbb{R})$ and

$$\begin{aligned} \widehat{K(g)}(t) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{g(y+x) - g(y)}{x} M(dx) \exp(ity) dy \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{g(y+x) - g(y)}{x} \exp(ity) dy M(dx) \\ &= \widehat{g}(t) \int_{\mathbb{R}} \frac{\exp(-ixt) - 1}{x} M(dx). \end{aligned}$$

Since $\int_{\mathbb{R}} (\exp(-ixt) - 1)/x M(dx) = \widehat{f}'(-t)/(\widehat{f}(-t)i)$, the Fourier transforms of $\int_{\mathbb{R}} K(g)(x+a)f(x)dx$ and $\int_{\mathbb{R}} g(x+a)xf(x)dx$ are equal and thus the two terms themselves are equal for all $a \in \mathbb{R}$.

If g is Lipschitz continuous but does not have compact support, then let $g_n(x) := (1 - |x|/n)_+ g(x)$. Note that the set of the Lipschitz constants of the g_n is bounded! Clearly g_n and $K(g_n)$ respectively converge to g and $K(g)$ pointwise and $\|K(g_n)\|_{\infty}$ is bounded. Hence $\lim_{n \rightarrow \infty} \int_{\mathbb{R}} g_n(x+a)xf(x)dx = \int_{\mathbb{R}} g(x+a)xf(x)dx$ and $\lim_{n \rightarrow \infty} \int_{\mathbb{R}} K(g_n)(x+a)f(x)dx = \int_{\mathbb{R}} K(g)(x+a)f(x)dx$ for all a . Thus

$$\int_{\mathbb{R}} K(g)(x+a)f(x)dx = \int_{\mathbb{R}} g(x+a)xf(x)dx.$$

□

Remark 5.3 *The assumption that b is 0, is not serious, b is a location parameter of the density and thus we can use Theorem 5.1. Also the condition $M(\{0\}) = 0$ is not restrictive, if $M(\{0\}) = \sigma^2$ then the distribution is the convolution of a mean zero normal distribution with variance σ^2 and a infinitely divisible distribution with Lévy measure $M(\cdot \cap \mathbb{R} \setminus \{0\})$. Again we can use Theorem 5.1 for this situation. For the normal distribution the operator K is $\sigma^2 D$. Hence in the general case we obtain:*

$$K(g)(t) := bg(t) + M(\{0\})g'(t) + \int_{\mathbb{R} \setminus \{0\}} \frac{g(t+x) - g(t)}{x} M(dx)$$

Remark 5.4 *Let f be a density with finite second moment σ^2 and zero mean, then without loss of generality $K_f(1) = 0$. Let $f_n = \star_{i=1}^n \sqrt{n}f(\cdot/\sqrt{n})$. By the central limit theorem f_n converges in distribution to a normal density. So one would expect that K_{f_n} converges in some sense to $\sigma^2 D$. Assume that $K_f(g)(x) =$*

$\int_{\mathbb{R}} (g(x+y) - g(x))/y M(dy)$. Note that if $K_f(g) = Q * g$, where Q is a measure, then with $Q^-(A) := Q(-A)$,

$$\begin{aligned} (Q * g)(x) &= \int_{\mathbb{R}} g(x-y) - g(x) Q(dy) \\ &= \int_{\mathbb{R}} \frac{g(x+y) - g(x)}{y} y Q^-(dy), \end{aligned}$$

where the first equality holds since $K_f(1) = 0$, i.e. $\int_{\mathbb{R}} 1 Q(dx) = 0$. Since $\int_{\mathbb{R}} x(x+a)f(x)dx = \int_{\mathbb{R}} x^2 f(x)dx$, taking $g(x) = x$ gives $M(\mathbb{R}) = K(x)$ and $\int K(x)f(x)dx = \int_{\mathbb{R}} x^2 f(x)dx = \sigma^2$. As we already know $K_{f_n}(g)(x) = K_f(g(\cdot/\sqrt{n}))(x\sqrt{n})/\sqrt{n}$. Using the form of K_f , we have now

$$K_{f_n}(g)(x) = \int_{\mathbb{R}} \frac{g(x+y/\sqrt{n}) - g(x)}{y/\sqrt{n}} M(dy).$$

Thus if g is Lipschitz continuous and differentiable then $\lim_{n \rightarrow \infty} K_{f_n}(g)(x) = \sigma^2 g'(x)$.

Examples:

1. Let $f(x) = \exp(-\sqrt{2}|x|)/\sqrt{2}$ be the variance normalized Laplace density. It is easy to see that, $\hat{f}(w) = 2/(2+w^2)$. Thus

$$\frac{\hat{f}'(w)}{\hat{f}(w)i} = \frac{2iw}{2+w^2}$$

and hence

$$\hat{K}(w) = \frac{-2iw}{2+w^2} = -iw\hat{f}(w) = \hat{f}'.$$

This is because

$$\begin{aligned} \hat{f}'(w) &= \int_{\mathbb{R}} f'(x) \exp(ixw) dx \\ &= f(x) \exp(ixw) \Big|_{-\infty}^{\infty} - \int_{\mathbb{R}} f(x) iw \exp(iwx) dx \\ &= -iw\hat{f}(w). \end{aligned}$$

Thus $K = -\exp(-\sqrt{2}|x|)\text{sgn}(x) \in L^1(\mathbb{R})$. Tedious but simple computations yield now that

$$K * x_+ = \left\{ \begin{array}{ll} \frac{\exp(\sqrt{2}x)}{2} & : x \leq 0 \\ 1 - \frac{\exp(-\sqrt{2}x)}{2} & : x > 0 \end{array} \right\} =: h(x).$$

Using identity (49) we obtain

$$\begin{aligned} K(T_\lambda^S(x) - x) &= -h(x) - (1-h(x)) + h(x-\lambda) + (1-h(x+\lambda)) \\ &= h(x-\lambda) - h(x+\lambda), \end{aligned}$$

taking this all together we have now for X with Laplace distribution

$$E(T_\lambda^S(X+a) - a)^2 = 1 + E \min((X+a)^2, \lambda^2) + 2(h(X+a-\lambda) - h(X+a+\lambda)).$$

2. Let $f_t(x) = \exp(-x)x^{t-1}/\Gamma(t)\mathbf{1}_{\mathbb{R}^+}(x)$ the density of the Gamma distribution.

Since the mean of this distribution is t we want to compute $K_{f_t * \delta_{-t}}$. Then by Feller [18, p.567], $\log(\widehat{f}_t(x)) = t \int_0^\infty (\exp(iyx) - 1)/y \exp(-y) dy$ and thus $(\log \widehat{f}_t)'(x) = ti \int_0^\infty \exp(iyx) \exp(-y) dy$. Thus $K_{f_t}(g) = Q * g$ where $Q \in L^1(\mathbb{R})$ and

$$\widehat{Q}(x) = t \int_0^\infty \exp(-iyx) \exp(-y) dy = t \int_{-\infty}^0 \exp(ixy) \exp(y) dy.$$

Hence $K_{f_t * \delta_{-t}}(g)(x) = t \int_{-\infty}^0 \exp(y) g(x-y) dy - tg(x)$.

3. Another example is the cosine hyperbolic density, $f(x) = 1/\cosh(\pi x/2)$, again by [18, p.567]

$$\log(\widehat{f}(x)) = \int_{-\infty}^\infty \frac{\exp(ixy) - 1 - iyx}{y^2} \frac{y}{\exp(y) - \exp(-y)} dy$$

and thus

$$K_f(g)(x) = \int_{\mathbb{R}} \frac{g(x+y) - g(x)}{y} \frac{y}{\exp(y) - \exp(-y)} dy.$$

All these examples were infinitely divisible distributions. For the uniform distribution with density $\mathbf{1}_{(-1,1)}(x)/2$, K does not have a nice form. Assume $g : [-1, 1] \rightarrow \mathbb{R}$ and $\int_{-1}^1 g(x)/2 dx = 0$. If \bar{g} is the 2 periodic continuation of g on \mathbb{R} then $\bar{g} * \mathbf{1}_{(-1,1)}/2 = 0$. Thus unbiased risk estimators are not uniquely determined. Let

$$r(a) = \int_{-1}^1 (x+a)_+ x / 2 dx = \begin{cases} 0 & : a \leq -1 \\ 1/6 + a/4 - a^3/12 & : a \in (-1, 1) \\ 0 & : a \geq 1 \end{cases}.$$

After some guessing one finds that with

$$h(x) = \begin{cases} 0 & : x \leq 0 \\ -(x - [x/2]2)(x - [x/2]2 - 2)/2 & : x \geq 0 \end{cases}.$$

(h is the 2 periodic continuation of $-x(x-2)/2$ defined on $[0, 2]$ to \mathbb{R}^+ .) $\int_{-1}^1 h(x+a)/2 dx = r(a)$. So with the help of (49) we can compute now an unbiased risk estimator for soft thresholding. The Figure 12 shows the unbiased risk estimators for soft thresholding with threshold 2 for the normal distribution, the Laplace distribution, the gamma distribution with $t = 2$ and the uniform distribution. The distributions were transformed to have unit variance and zero mean.

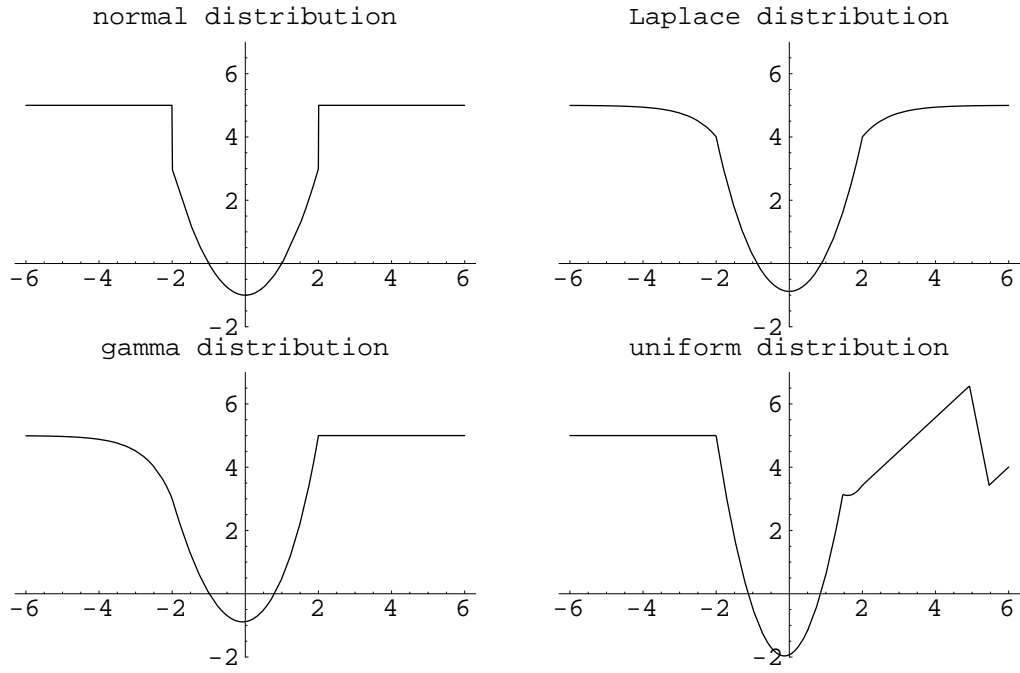


Figure 12: The unbiased risk estimators for soft thresholding

Remark 5.5 *As we have seen in equation (45), for Gaussian random variables, unbiased risk estimation is possible for multivariate means, even if the estimators for coordinates are not independent. This is also possible for other types of distributions, one has to apply the operator K coordinate wise. Let X_i , $i = 1, \dots, n$ be random variables, X_i has the distribution F_i and $EX_1 = 0$, $EX_1^2 = \sigma_1^2$. Assume that an operator K_1 exists such that $EX_1g(X_1 + a_1) = EK_1(g)(X_1 + a_1)$ for some g . If $g : \mathbb{R}^n \rightarrow \mathbb{R}$ and $a \in \mathbb{R}^n$ then $E(X_1 + g(X + a) - a_1)^2 = \sigma_1^2 + Eg(X + a)^2 + 2EX_1g(X + a)$. Given that g satisfies some necessary conditions,*

$$\begin{aligned}
& EX_1g(X + a) \\
&= \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} x_1g(x_1 + a_1, \dots, x_n + a_n)F_1(dx) \otimes_{i=2}^n F_i(d(x_2, \dots, x_n)) \\
&= \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} K_1(g(\cdot, x_2 + a_2, \dots, x_n + a_n))(x_1 + a_1)F_1(dx) \\
&\quad \otimes_{i=2}^n F_i(d(x_2, \dots, x_n)).
\end{aligned}$$

Thus $E(X_1 + g(X + a) - a_1)^2 = \sigma_1^2 + Eg(X + a)^2 + 2EK_1(g(\cdot, X_2 + a_2, \dots, X_n + a_n))(X_1 + a_1)$.

6 Some last thoughts

6.1 Comparison of thresholds in the two approaches

In this thesis we considered two approaches to wavelet thresholding. For both approaches wavelet thresholding is able to achieve the minimax rate, but different thresholds are used. We take a closer look at the Gaussian noise case. For the ideal estimator approach the optimal minimax rate is achieved with thresholds of uniform size $\sim \sqrt{2 \log n} \sigma$, although we saw later that we can choose thresholds level-wise and this was still a minimax method. For the level j the thresholds were of size $\sim \sqrt{2^j \log 2} \sigma$. In the function space approach almost similar thresholds are chosen. If we assume our signal is member of a bounded subset of the Besov space $B_{p,q}^m$ then optimal minimax thresholds are of the size $C \sqrt{(j - j_0)_+}$, where C and j_0 depend on m, p and q (see [12] and [37]). Using the thresholds of the function space approach in the ideal estimator context and vice versa does not achieve the minimax rate. But using thresholds of size *constant* \sqrt{j} for the level j in the function space approach achieves almost the minimax rate, it is worse by a factor $O(\log n)$. The natural question now is, is it possible to reconcile both approaches, i.e. is there a set of thresholds which achieves the optimal minimax rate in both contexts?

This is not possible, we will show this for $p \geq 2$. We assume now our usual situation, let $X_i = f_i + z_i$, $i = 1, \dots, n$, where the f_i are parameters of interest and the z_i are iid normal random variables with mean zero and variance $1/n$. Let W be a wavelet transform and $a = W(f)$, assume $\|a\|_{B_{p,q}^m} \leq A$, $m > 1/p$, $p \geq 2$ and $q \geq 1$. Let $p(\cdot, \cdot)$ have the usual meaning. If $\lambda \geq a \geq 0$, then

$$p(\lambda, a) \geq a^2 \Phi((-\lambda - a, \lambda - a)) + \int_{-\infty}^{-\lambda - a} (x + \lambda)^2 \Phi(dx) \geq \frac{a^2}{2}. \quad (50)$$

Now let $\lambda_{n,j}$ be a set of thresholds which achieve the optimal minimax rate in the ideal estimator context. For a fixed $\alpha \in (0, 1)$, the optimal thresholds for the level $j = \alpha \log_2 n$ have to be at least of size $\sim \sqrt{2^j \log 2} / \sqrt{n} = C \sqrt{j} / \sqrt{n}$, where C is a constant. The reason is that $2^j p(\lambda_j, 0) = O(\log n/n)$ is needed to achieve the minimax rate for the ideal estimator approach. Let

$$j_0 := \min_j \left\{ j : \frac{C \sqrt{j}}{\sqrt{n}} \geq 2A \sqrt{2^{-j(2m+1)}} \right\},$$

simple calculations yield that $j_0 \sim (\log_2 n)/(2m + 1)$. If $a_{j_0,k} = A \sqrt{2^{-j_0(2m+1)}}$, $k = 0, \dots, 2^{j_0} - 1$ and $a_{j,k} = 0$ elsewhere, then clearly $\|a\|_{B_{p,q}^m} \leq A$. If n tends to infinity, then for n larger than a certain bound, $\lambda_{n,j_0} > A \sqrt{2^{-j_0(2m+1)}}$. Now it follows from (50) that the risk for thresholding the signal (a) at level j_0 with thresholds λ_{n,j_0} is at least as large as

$$2^{j_0} A^2 2^{-j_0(2m+1)} / 2 = A^2 2^{-j_0 2m} / 2. \quad (51)$$

Using the definition of j_0 we obtain

$$2^{2m+3} A^2 2^{-j_0(2m+1)} \geq \frac{C^2(j_0 - 1)}{n}. \quad (52)$$

Combining the relations (51) and (52) yields that the risk for estimating $(a.)$ is as large as

$$A^2 2^{-j_0 2m-1} \geq A^2 \left(\frac{C}{A} \right)^{\frac{4m}{(2m+1)}} n^{\frac{-2m}{(2m+1)}} (j_0 - 1)^{\frac{2m}{(2m+1)}} 2^{-\frac{2m(2m+3)}{(2m+1)}-1}.$$

Since $j_0 \sim \log_2 n / (2m + 1)$, this is worse than the minimax rate for the Besov spaces $B_{p,q}^m$ in the function space context. Thus it is not possible to chose thresholds which yield in both contexts the minimax rate.

6.2 Block thresholding and kernel estimators

When applying soft or hard thresholding, the estimation of one wavelet coefficient is based on a noisy version of this wavelet coefficient alone. There are other methods where a kind of hard thresholding is applied to a whole block of wavelet coefficients. A block of noisy wavelet coefficients $a_1 + e_1, \dots, a_k + e_k$, ($a.$ represents the signal and $e.$ the noise) is kept if $\sum (a_i + e_i)^2$ is larger than a threshold, otherwise the whole block is set to zero. So hard thresholding is applied to a whole block of coefficients. This estimator which is called block thresholding has been investigated by Hall, Kerkyacharian and Picard ([22]) and Cai ([5]). It was shown that block thresholding shares the minimax properties of soft thresholding, as well in the ideal estimator approach as in the function space approach. It is also possible to change this thresholding policy, by keeping a block if one of the coefficients in it is larger than a threshold, this estimator has the same properties as the other block thresholding estimator.

In block thresholding the blocks are horizontal, i.e. composed of the coefficients with the indexes $(j, k), \dots, (j, k + L)$, now we consider another type of block thresholding estimator, the blocks are vertical and not disjunct. First we have to introduce a notation, we will say an index (j', k') (or the wavelet coefficient with this index) is above the index (j, k) if $j' \leq j$ and $|[2^{j'-j}k] - k'| \leq C$ where $C \in \mathbb{N}_0$ is a constant. In this method, if the modulus of a coefficient $a_{j,k} + e_{j,k}$ is larger than a threshold, then not only the coefficient itself but also all coefficients above it are kept. A variation of this method is to keep also the coefficients with the indexes (j, k') , $|k - k'| \leq C_2$ and all coefficients above them too, where C_2 is another constant.

This new method achieves the optimal minimax rate in the ideal estimator context. Let the observations $X_i = f_i + e_i$, $i = 1, \dots, n = 2^m$ be given, where the f_i are the parameters of interest and the e_i are iid random variables with distribution $N(0, 1)$. We apply a periodic wavelet transform W_n to these observations. Let $Y = W_n(X)$, $a = W_n(f)$ and $z = W_n(e)$, clearly the $z_{j,k}$ are iid random variables with distribution $N(0, 1)$. Now let λ_n be such $E \mathbf{1}_{\{|e_1| > \lambda_n - 1\}} (1 + e_1^2) = 1/n$,

with the background of this thesis it is easy to see that $\lambda_n \sim \sqrt{2 \log n}$. Let L be the constant which takes the role of C in the meaning of ‘‘above’’. Note that the number of coefficients above a coefficient is less than $(2L + 1) \log_2 n$, since in each level there are only $2L + 1$ coefficients above a fixed coefficient. We define the estimator $\widehat{a}_{j,k}$ for the coefficient $a_{j,k}$ by

$$\widehat{a}_{j,k} := \begin{cases} Y_{j,k} & : |Y_{j,k}| \geq \lambda_n \text{ or } \exists (j', k'), (j, k) \text{ is above } (j', k') \text{ and } |Y_{j',k'}| \geq \lambda_n \\ 0 & : \text{else} \end{cases}.$$

We have now

$$\begin{aligned} & \sum_{j,k} E(\widehat{a}_{j,k} - a_{j,k})^2 \\ & \leq \sum_{j,k} E \left(\mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} z_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}| < \lambda_n\}} a_{j,k}^2 + \mathbf{1}_{\{(j,k) \text{ above a } |Y_{j',k'}| \geq \lambda_n\}} z_{j,k}^2 \right) \\ & \leq \sum_{j,k} E \left(\mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} z_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}| < \lambda_n\}} a_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} \sum_{(j',k') \text{ above } (j,k)} z_{j',k'}^2 \right). \end{aligned} \quad (53)$$

If $|a_{j,k}| < 1$ then

$$\begin{aligned} E \mathbf{1}_{\{|a_{j,k} + z_{j,k}| \geq \lambda_n\}} z_{j,k}^2 & \leq E \mathbf{1}_{\{|z_{j,k}| \geq \lambda_n - 1\}} z_{j,k}^2 \leq \frac{1}{n}, \\ E \mathbf{1}_{\{|a_{j,k} + z_{j,k}| \leq \lambda_n\}} a_{j,k}^2 & \leq a_{j,k}^2 \end{aligned}$$

and

$$\begin{aligned} E \mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} \sum_{(j',k') \text{ above } (j,k)} z_{j',k'}^2 & \leq (2L + 1) \log_2 n E \mathbf{1}_{\{|z_{j,k}| \geq \lambda_n - 1\}} E z_{j,k}^2 \\ & \leq \frac{(2L + 1) \log_2 n}{n}. \end{aligned}$$

If $|a_{j,k} + z_{j,k}| < \lambda_n$ then $|a_{j,k}| < |\lambda_n| + |z_{j,k}|$, thus

$$E \mathbf{1}_{\{|a_{j,k} + z_{j,k}| < \lambda_n\}} a_{j,k}^2 \leq 2|\lambda_n|^2 + 2E z_{j,k}^2.$$

Further

$$E \mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} \sum_{(j',k') \text{ above } (j,k)} z_{j',k'}^2 \leq (2L + 1) \log_2 n$$

and

$$E \mathbf{1}_{\{|a_{j,k} + z_{j,k}| \geq \lambda_n\}} z_{j,k}^2 \leq 1.$$

Hence

$$\begin{aligned} & \frac{E \left(\mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} z_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}| < \lambda_n\}} a_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}| \geq \lambda_n\}} \sum_{(j',k') \text{ above } (j,k)} z_{j',k'}^2 \right)}{1/n + \min(a_{j,k}^2, 1)} \\ & \leq C \log n, \end{aligned} \quad (54)$$

where C is a constant. Combining (53) and (54) we obtain

$$\frac{\sum_{j,k} E(\hat{a}_{j,k} - a_{j,k})^2}{1 + \sum_{j,k} \min(a_{j,k}^2, 1)} \leq C \log n.$$

Remark 6.1 *A similar result holds for the function space approach. We will leave the proof out, but here is a simple heuristic explanation: In the function space approach all coefficients above a level j_0 are kept and to the other coefficients a soft (or hard) thresholding estimator is applied with threshold $C\sqrt{(j - j_0)_+}$ where j is the level of a coefficient and C is a constant. This method achieves the minimax rate. We compare now vertical block thresholding with soft and hard thresholding. When using the threshold $C\sqrt{(j - j_0)_+}$ one accepts at least a bias of size $C\sqrt{(j - j_0)_+}/2$ for coefficients larger than $C\sqrt{(j - j_0)_+}/2$. But the number of coefficients above a coefficient in level j and below the level j_0 is $(2L + 1)(j - j_0)_+$, so by keeping these coefficients too, the risk of the whole estimator is increased by the variance of the noise times $(2L + 1)(j - j_0)_+$. But this is a constant multiple of $C^2(j - j_0)_+/4$. If a coefficient is smaller than $C\sqrt{(j - j_0)_+}/2$, then if this coefficient is kept (because the noise term is too large), then we have a loss of $C^2(j - j_0)_+/4$, so the additional risk of size $(2L + 1)(j - j_0)_+$ times variance of the noise is again a constant multiple of the actual loss.*

The error estimates for the vertical block thresholding estimator, are rather rough. It is reasonable to expect that the real risk of this estimator compares more favorably with other estimators. Indeed, often “real world” signals exhibit a correlation of the size of their wavelet coefficients which are above each other. Irregularities, like a discontinuity, in a signal affect all wavelets in whose support this irregularity is.

Interesting about the vertical block thresholding method is that it is close to a kernel estimate with locally varying bandwidth. The first application of wavelets in non-parametric statistics was to build linear estimators. As was mentioned earlier, a simple estimation method is to compute a wavelet transform of the noisy data, and then to discard the lower levels. But this is a simple kernel estimate. Assume the situation at the beginning of the chapter, then a simple first order approximation of the noisy wavelet coefficients is (2^j is small compared to n):

$$\tilde{a}_{j,k} := \sum_i \frac{\psi_{j,k}(i/n)}{\sqrt{n}} X_i.$$

If we estimate f_i by discarding the levels below the level j_0 , then by a similar first order approximation,

$$\hat{f}_i := \sum_{j \geq j_0, k} \tilde{a}_{j,k} \frac{\psi_{j,k}(i/n)}{\sqrt{n}}$$

$$\begin{aligned}
&= \sum_{j \geq j_0, k} \left(\sum_l \frac{\psi_{j,k}(l/n)}{\sqrt{n}} X_l \right) \frac{\psi_{j,k}(i/n)}{\sqrt{n}} \\
&= \frac{1}{n} \sum_l X_l \sum_{j \geq j_0, k} \psi_{j,k}(l/n) \psi_{j,k}(i/n) \\
&= \frac{1}{n} \sum_l K(l/n, i/n) X_l,
\end{aligned}$$

where $K(x, y) = \sum_{j \geq j_0, k} \psi_{j,k}(x) \psi_{j,k}(y)$. If we keep also level $j_0 + 1$, then $K(x, y)$ has to be replaced by $K(2x, 2y)/2$. Thus the parameter 2^{-j_0} corresponds to the bandwidth of a classical linear kernel estimator. Figure 6.2 shows for an artificial signal what wavelet coefficients are kept with different methods. The dark rectangles correspond to coefficients which are kept. The top graph shows this for a hard thresholding estimator. The lowest picture shows what coefficients are kept for a kernel estimator. The graph in the middle shows why the vertical block thresholding can be seen as a kernel estimator with locally varying bandwidth (we keep some neighbouring coefficients as well): it behaves locally like the simple kernel estimator based on wavelets.

This explanation of vertical block thresholding as a kernel estimate with locally varying bandwidth becomes clearer if the underlying wavelet basis is the Haar basis. Then for vertical block thresholding each estimate \hat{f}_i is the mean of some neighboring X_j .

For the Haar base the scaling equalities have the following form:

$$\phi_{j,k} = \frac{1}{\sqrt{2}}(\phi_{j+1,2k} + \phi_{j+1,2k+1}) \text{ and } \psi_{j,k} = \frac{1}{\sqrt{2}}(\phi_{j+1,2k} - \phi_{j+1,2k+1})$$

With this in mind it is easy to derive the explicit form of the discrete wavelet transform and its inverse. For an input signal X_0, \dots, X_{n-1} , $n = 2^h$ the discrete wavelet transform is defined by

$$c_0 = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \text{ and } d_{j,k} := \frac{1}{\sqrt{2^{h-j}}} \sum_{i=0}^{2^{h-j-1}-1} f_{2^{h-j}k+i} - \sum_{i=2^{h-j-1}}^{2^{h-j}-1} f_{2^{h-j}k+i}.$$

The inverse transformation is defined by

$$f_i = \frac{1}{\sqrt{n}} c_0 + \sum_j d_{j, [i/2^{h-j}]} \begin{cases} 1 & : i/2^{h-j} - [i/2^{h-j}] < 1/2 \\ -1 & : i/2^{h-j} - [i/2^{h-j}] \geq 1/2 \end{cases}.$$

To compute an estimation of f_i if we discard the levels below the level j_0 we compute

$$\begin{aligned}
\hat{f}_i &= \frac{1}{\sqrt{n}} c_0 + \sum_{j=0}^{j_0} d_{j,k} \begin{cases} 1 & : i/2^{h-j} - [i/2^{h-j}] < 1/2 \\ -1 & : i/2^{h-j} - [i/2^{h-j}] \geq 1/2 \end{cases} \\
&= \frac{1}{2^{h-j_0-1}} \sum_{l=[i/2^{h-j_0-1}]2^{h-j_0-1}}^{[i/2^{h-j_0-1}]2^{h-j_0-1}+2^{h-j_0-1}-1} X_l.
\end{aligned}$$

The last equality can be shown via a simple induction argument for j_0 . Thus if we discard the levels below j_0 then \widehat{f}_i is the mean of a block of 2^{h-j_0-1} X_l 's.

Now we claim that if in vertical block thresholding the coefficient with index $(j_1, [i/2^{h-j_1}])$ is kept because the coefficient with index (j, k) is larger than the threshold and $|[i/2^{h-j_1}] - [k/2^{j-j_1}]| \leq C$, then for all $j_2 < j_1$, $|[i/2^{h-j_2}] - [k/2^{j-j_2}]| \leq C$, i.e. also the coefficients with indexes $(j_2, [i/2^{h-j_2}])$, $j_2 < j_1$ are kept:

Note that if $x \in \mathbb{R}$ and $k \in \mathbb{N}$ then

$$x/k < [x]/k + 1/k \leq ([x]/k + (k-1)/k) + 1/k = [[x]/k] + 1,$$

thus $[x/k] = [[x]/k]$. It is clear that

$$|[i/2^{h-j_1}]/2^{j_1-j_2} - [k/2^{j-j_1}]/2^{j_1-j_2}| \leq C/2^{j_1-j_2},$$

hence we have now

$$C \geq -[-C/2^{j_1-j_2}] \geq |[[i/2^{h-j_1}]/2^{j_1-j_2}] - [[k/2^{j-j_1}]/2^{j_1-j_2}]| = |[i/2^{h-j_2}] - [k/2^{j-j_2}]|.$$

Thus for vertical block thresholding we also obtain

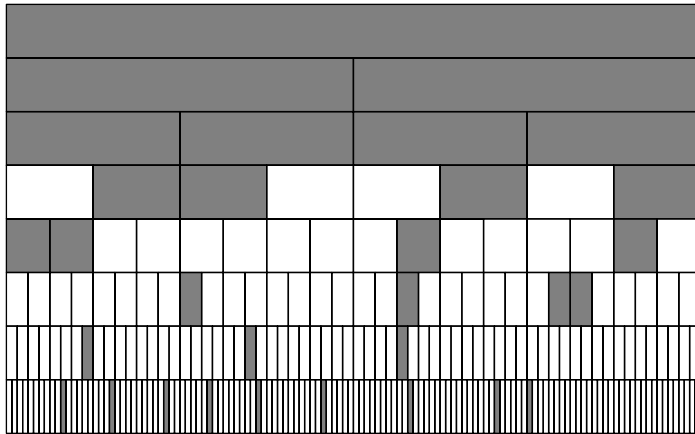
$$\widehat{f}_i = \frac{1}{2^{h-j_0-1}} \sum_{l=[i/2^{h-j_0-1}]2^{h-j_0-1}}^{[i/2^{h-j_0-1}]2^{h-j_0-1}+2^{h-j_0-1}-1} X_l,$$

but now j_0 depends on i and (X_l) , i.e. it is a kernel estimator with locally varying bandwidth.

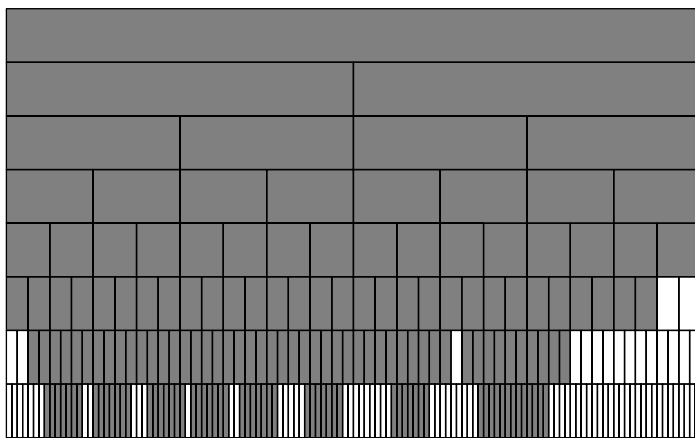
Note that there already exists a kernel estimator with locally varying bandwidth which achieves the same minimax rate as a wavelet thresholding estimator see [33]. There the local bandwidth is chosen from a set $a^{-j}h_1$, $a, h_1 > 0$ constants and $j = 0, 1, \dots$. For the simple kernel estimator based on wavelets, the bandwidth is 2^{-j} , $j = 0, 1, \dots$, and j is the last level of wavelet coefficients that is kept. This paper of Lepski Mammen and Spokoiny ([33]) shows that kernel estimates with a local varying bandwidth selection can be as good as wavelet thresholding in a minimax sense. The performance vertical block thresholding makes this also plausible.

Many examples used for Monte-Carlo experiments are just functions with finitely many discontinuities, and good kernel methods can cope with this type of functions. Often one knows what kind of function one expects, definitely not the kind of "worst case" functions in the minimax approaches. Then a special method might be better. But wavelets are a good allround method. Wavelets should perform better than kernel methods for images, images are much more complex, for example, a smooth sky and unsmooth trees. What is missing is a fully automatized curve estimator based on wavelet thresholding, which is included by default in major statistical packages, then people can really compare and decide what is better for them.

Wavelet thresholding



Kernel estimate with varying bandwidth



Kernel estimate

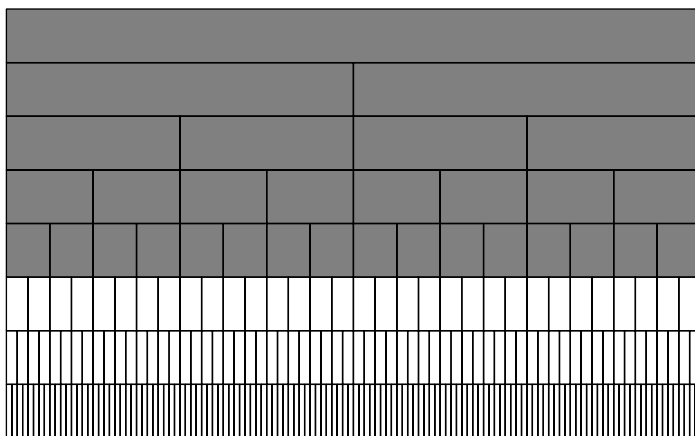


Figure 13: Hard thresholding, vertical block thresholding and kernel estimate

References

- [1] Barikov, N.K., Extrema of the distributions of quadratic forms of Gaussian variables. *Theory Probab. Appl.* 34, (1989).
- [2] Bruce, A.G. and Hong-Ye, G., WaveShrink with Semisoft Shrinkage. *StaSci Research Report No. 39* (1995).
- [3] Bruce, A. and Gao, H.Y. WaveShrink with firm shrinkage, *Research Report 39*, Statistical Science Division, MathSoft Inc. (1996).
- [4] Bickel, P., Minimax estimation of the mean of a normal distribution subject to doing well at the point. in *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi, D. Siegmund, eds.) Academic Press Inc., New York (1983).
- [5] Cai, T., Adaptive Wavelet Estimation: A Block Thresholding And Oracle Inequality Approach. *Annals of Statistics*, in press
- [6] 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).
- [7] 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).
- [8] Daubechies, I., Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* 41, 909-996 (1988).
- [9] Daubechies, I., Ten lectures on wavelets. No. 61 in *CBMS-NSF in Applied Mathematics*, Siam, Philadelphia (1992).
- [10] DeVore, R.A. and Lucier, B., Fast wavelet techniques for near optimal image processing. *1992 IEEE Military Communications Conference*, 2-12 (1992).
- [11] Delyon, B. and Juditsky A., Estimating Wavelet Coefficients. *Wavelets and Statistics*, Lecture Notes in Statistics 103, Springer Verlag (1995).
- [12] Delyon, B. and Juditsky A., On Minimax Wavelet Estimators. *Applied Computational Harmonic Analysis* 3, 215-228 (1996).
- [13] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D., Density estimation via wavelet thresholding. *Annals of Statistics* 24 (1996).
- [14] Donoho, D.L., De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* 41, No. 3, 613-627 (1994).

- [15] Donoho, D.L. and Johnstone, I.M., Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, No. 3, 425-455 (1994).
- [16] Donoho, D.L. and Johnstone, I.M., Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. Amer. Statist. Assoc.* 90, 1200-1224 (1995).
- [17] Donoho, D.L. and Johnstone, I.M., Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli* 2, No. 1, 39-62 (1996).
- [18] Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley & Sons (1966).
- [19] 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).
- [20] Efromovich, S.Y. and Pinsker, M.S., Estimation of a square integrable probability density of a random variable. *Problems of Information Transition* 18, (1982).
- [21] Gao, H.Y., *Wavelet Estimation of Spectral Densities in Time Series Analysis*. Ph.D. dissertation, University of California, Berkeley (1993).
- [22] Hall, P., Kerkyacharian, G. and Picard, D., On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica* Vol. 9, No. 1, 33-50 (1999).
- [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] Kolaczyk, 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] Krim, H. and Pesquet, J.-C., On the statistics of best bases criteria. *Wavelets and Statistics, Lecture Notes in Statistics* 103, Springer Verlag (1995).
- [32] Ledoux, M. and Talagrand, M., *Probability in Banach Spaces*. Springer Verlag (1991).
- [33] Lepski, O.V., Mammen, E. and Spokoiny V.G, Optimal spatial adaption to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics* Vol. 25, No. 3 (1997).
- [34] Mallat, S., Multiresolution approximation and wavelets. *Trans. Amer. Math. Soc.* 315 (1989).
- [35] Marshall, A.W. and Olkin I., *Inequalities: Theory of Majorization and its Applications*. Academic Press (1979).
- [36] Meyer, Y., *Ondelettes et opérateurs I: Ondelettes*. Paris: Hermann, Éditeurs des Sciences et des Arts (1990).
- [37] 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).
- [38] Petrov, V., *Limit Theorems of Probability Theory*. Clarendon Press Oxford (1995).
- [39] Strasser, H., *Mathematical Theory of Statistics*. de Gruyter, New York, Berlin (1985).
- [40] Sweldens, W., The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.* 29, No.2, 511-546 (1998).
- [41] Shorak, R.S. and Wellner J.A., *Empirical Processes with Applications to Statistics*, John Wiley & Sons (1986).
- [42] Wang, Y., Function estimation via Wavelet Shrinkage for Long-Memory Data. *Ann. Stat.* 24, No. 2, (1996).