WAVELET THRESHOLDING FOR NON (NECESSARILY) GAUSSIAN NOISE: FUNCTIONALITY[*†]

R. Averkamp[‡] and C. Houdré[§]

December 4, 1999

Abstract

For signals belonging to balls in smoothness classes and noise with higher moments, asymptotically minimax results are obtained. A median filtering method to tackle noise with heavy tails is presented. Unbiased risk estimation, à la Stein, is also studied for infinitely divisible noise. Some comparisons of wavelet thresholding and of kernel estimators are also briefly discussed.

1 Introduction

A well known approach to obtaining asymptotic results for wavelet thresholding estimators is to assume that the function to estimate 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 [6], [8] or [5]. These results were then

---

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

†Key words: Wavelets, Thresholding, Minimax.

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

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

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

extended to non-Gaussian noise by Delyon and Juditsky [4] and Neumann and Spokoiny [15].

Here we further extend the results of Delyon and Juditsky and Neumann and Spokoiny. Delyon 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. It will be shown here that under lesser conditions on the noise, for soft thresholding the ratio of the minimax risk for Gaussian noise and other types of noises also tends to 1. For noise with heavy tails, by first median filtering the data and then applying wavelet thresholding it is still possible to have the same minimax rate as for Gaussian noise. For noise with thin tails, we present some examples showing that other types of non–linear estimators can outperform wavelet tresholding. To complete our study of wavelet methods in a non–Gaussian framework, we also present unbiased risk estimates à la Stein for the general class of infinitely divisible noise with finite variance. These notes conclude with some comparison of thresholding with kernel estimators with varying bandwidth.

2 Moment conditions

The model considered throughout these notes (unless otherwise specified) is the familiar one. 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 as $f_i = f_{n,i} = f(i/n)/\sqrt{n}$, 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 apply a wavelet transform W_n to the data, so we then have noisy wavelet coefficients.

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

where j_0 is a fixed constant, i.e., we assume that we stop the wavelet transformation at the level j_0 (of course, the coefficients also depend 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) (this has already been discussed in [1]). The difference with the model studied in [1] is that the performance of the estimators is not compared to a benchmark, but that rather the possible values of $(a_{\cdot,\cdot})$ are restricted, i.e., it is assumed,

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

for some constant A, where $s := m + 1/2 - 1/p$ and $m > 1/p$. Indeed, if the $(a_{\cdot,\cdot})$ are the wavelet coefficients of f and the wavelet basis is sufficiently smooth then $\|f\|_{B^m_{p,q}} \leq C_2 A$, where $C_2 = C_2(m,p,q)$ is a constant and $1 \leq p, q \leq +\infty$. As already mentioned in [1], the condition $m > 1/p$ is necessary to have well defined point values of functions in the Besov space $B^m_{p,q}$ (and not generalized functions). This implies that really "nice" functions like $\text{sign}(x)$ do not belong to any of the Besov spaces we consider here. Also considering quasi–norms rather than norms, we could obtain similar results for the cases $0 < p < 1$, or $0 < q < 1$. Again, we refer the reader to the books by Daubechies [3], Meyer [14], and Vidakovic [18] for a much more extensive and precise list of references and further information on wavelets (and functions spaces and statistics).

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

$$\inf_{\widehat{a}} \sup_{a, \|a\|_{B^m_{p,q}} \leq C} E\|\widehat{a} - a\|_2^2 \sim c n^{-\frac{2m}{2m+1}}, \tag{2.1}$$

where the infimum is for all estimators and c is a positive constant (see [15]).

Further, estimators based on soft thresholding achieve this rate. Later it was also shown by Neumann and Spokoiny ([15]) that for noise with all moments, soft thresholding achieves the same rate as in the Gaussian case. In both cases the actual performance, not just the rate, is the same, i.e. the c in relation (2.1) is the same. If the noise satisfies certain regularity conditions, then even the minimax rate is the same. Delyon and Juditsky ([4]) extended these results by showing that even for more general distributions, soft thresholding can achieve the same rate as in the Gaussian case. This result will be complemented here, 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 Delyon and Juditsky, namely that it is possible to assume the noise is bounded. In the sequel Φ denotes the standard normal distribution, and E_Φ is integration with respect to it. We also note here that the Besov assumption can also be replaced by a Triebel–Lizorkin one.

Theorem 2.1 *Assume the model as above, where $p, q \geq 1$ and $m > 1/p$. Let the z_i have finite moments of order L where L satisfies the conditions*

$$L > \frac{6}{2m/(2m+1)}, \quad \textit{if } p \geq 2 \tag{2.2}$$

and

$$L > \frac{6(m+1/2-1/p)(2m+1)}{(m+1/2-1/p)(2m+1)-m}, \quad \textit{if } 1 \leq p \leq 2. \tag{2.3}$$

Further let z_i be symmetric with continuous distribution. Then

$$\liminf_{n\longrightarrow\infty} \frac{\inf_{(\lambda)} \sup_{a,\|a\|_{B^m_{p,q}}\leq A} E_\Phi \sum_{j,k} |T^S_{\lambda_{j,k}}(w_{j,k}-a_{j,k})|^2}{\inf_{(\widetilde{\lambda})} \sup_{a,\|a\|_{B^m_{p,q}}\leq A} E \sum_{j,k} |T^S_{\widetilde{\lambda}_{j,k}}(w_{j,k}-a_{j,k})|^2} \geq 1. \tag{2.4}$$

Above, the requirement of symmetry is imposed for technical reasons (we need $Ez_i\mathbf{1}_{\{|z_i|>\lambda\}} = 0$ for all λ). The continuity requirement is made to simplify partial integration. Both requirements can be circumvented by more technical efforts in the proof.

To prove this theorem a fair amount of technical preparation is needed. We start with some lemmas.

Lemma 2.2 *Let $s := m + 1/2 - 1/p$ and let*

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

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 any α such that $2^l \geq n^\alpha = 2^{\alpha h}$.

Proof: As usual, at 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

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

$\square$

The previous lemma shows that if we want to achieve the same minimax rate as for Gaussian noise, we do not need to worry about the 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 $1 \leq p \leq 2$. This is due to the fact that 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

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

We will want to use the following result of Kolmogorov (see [17, p.855]).

Lemma 2.3 *Let X_i, $i = 1, \ldots, n$, be zero mean, independent random variables. Let $s_n := \sqrt{\sum_{i=1}^n EX_i^2}$, let $|X_i| \leq K$, $i = 1, \ldots, n$ and let $S_n = \sum_{i=1}^n X_i$. Then*

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

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

Lemma 2.4 *Let $n \in \mathbb{N}$, and let $K_n = O(n^{-\varepsilon})$, for some $\varepsilon > 0$. Let $X_{n_1}, \ldots, X_{n_{k(n)}}$ be independent random variables, with mean zero and such that $\|X_i\|_\infty \le K_n$, $i = n_1, \ldots, n_{k(n)}$, with moreover $\sum_{i=1}^{k(n)} EX_{n_i}^2 = 1$. Let F be the distribution function of $\sum_{i=1}^{k(n)} X_{n_i}$, let $\lambda_n = \log n$ and let $c_n := (1 - \lambda_n K_n/2)$. Then, for any a with $0 < a < \lambda_n$,*

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

Proof: Using Lemma 2.3 we obtain

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

□

Further we need the following large deviation result which is a simple extension of Lemma 5.8 in [16], the difference with this lemma is that the requirement of identical distributions is dropped. The proof with the help of Esseen's inequality ([16, Theorem 5.4]) is essentially the same as for Lemma 5.8 in [16].

Lemma 2.5 *Let $C > 0$ and $\varepsilon \in (0,1)$, then there exists an increasing sequence (β_n) converging to 1 such that for independent, zero mean, random variables $X_{n_1}, \ldots, X_{n_{k(n)}}$ with $\sum_{i=1}^{k(n)} EX_{n_i}^2 = 1$ and $\max_{i=1,\ldots,k(n)} E|X_{n_i}|^3 \le$*

$Cn^{-3/2}$, *and all* x *with* $|x| \le (1-\varepsilon)\sqrt{2\log n}$,

$$\beta_n \le \frac{P(\sum_{i=1}^k X_i \le x)}{\Phi((-\infty, x])} \le 1/\beta_n \text{ and } \beta_n \le \frac{P(\sum_{i=1}^k X_i > x)}{\Phi((x,\infty))} \le 1/\beta_n. \quad (2.5)$$

Lemma 2.6 *Let* X_i, $i = 1, \ldots, n$ *be iid random variables with* $m_4 := EX_i^4 < \infty$, $EX_i = 0$, $EX_1^2 = 1$. *Let* $Y := \sum_{i=1}^n a_i X_i$, *where* $\sum_{i=1}^n a_i^2 = 1$. *Then* $\min(3, m_4) \le EY^4 \le \max(3, m_4)$.

Note that 3 is the fourth moment of a standard normal random variable.

Proof:

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

The assertion now follows from $\sum_i a_i^4 \le 1$. □

Proof of Theorem 2.1: 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, let l such that $2^l \le n^\alpha < 2^{l+1}$ and let (c_n) be 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 $\widetilde{z}_i := \min(\max(z_i, -c_n), c_n)$ and $\widetilde{e}_{j,k} := (W(\widetilde{z}/\sqrt{n}))_{j,k}$.

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

using the elementary inequality $(a+b)^4 \le 8(a^4+b^4)$, and since by Lemma 2.6, $Ee_{j,k}^4 \le M := \max(3, Ez_1^4)/n^2$.
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 without loss of generality, we can assume that 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 and above with $2^l \le n^\alpha < 2^{l+1}$. Let w be the noise part in one of these coefficients, then with $n = 2^h$

$$
w = \sum_{i=1}^n v_i \frac{z_i}{\sqrt{n}} \text{ and } \max_i |v_i| \le C_1 2^{-(h-l)/2} \le C_1\sqrt{n^{\alpha-1}}, \tag{2.6}
$$

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 Lemma 2.3 and 2.4, we will need that $P(\max_{1\le i\le n} |z_i 2^{-(h-l)/2}| \ge n^{-\varepsilon}) = O(1/n^2)$ for some $\varepsilon > 0$. The z_i have moments of order L, hence

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

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

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

Now let α be such that $\sup_{a,\|a\|_{B^m_{p,q}}\le A}\sum_{j\ge \alpha h}|a_{j,k}|^2=o(n^{-2m/(2m+1)})$, if $p\ge 2$, then Lemma 2.2 implies that $1-\alpha<2m/(2m+1)$ and thus

$$L>\frac{6}{2m/(2m+1)-2\varepsilon}. \tag{2.7}$$

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

$$L>\frac{6}{\frac{(m+1/2-1/p)(2m+1)-m}{(m+1/2-1/p)(2m+1)}-2\varepsilon}. \tag{2.8}$$

Since ε can be arbitrarily small, these are the same conditions as in the theorem.

Here are some examples of moment conditions for $1\le p<2$:

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

As already known, the numerator in equation (2.4) is $\sim$ *constant* $n^{-2m/(2m+1)}$. From now on we assume $\alpha\in(0,1)$ with $2^l\le n^\alpha<2^{l+1}$ such that

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

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

In view of the previous computations, without loss of generality, we can assume that $\|z_i 2^{(h-l)/2}\|_\infty\le n^{-\varepsilon}$, $i=1,\dots,n$, provided thresholds smaller than $(\log n)/\sqrt{n}$ are used. 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, $\lambda\ge 0$ and $a\in$

$\mathbb{R}$, let $p_\mu(\lambda, a) := \int_{\mathbb{R}} (T^S_\lambda(x+a) - a)^2 \mu(dx)$. If the e_i are iid $N(0, 1/n)$ random variables, then the 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 $Ee^2_{j,k} = 1/n$. Of course l, $\mu_{j,k}$ and $\lambda_{j,k}$ depend on n, but for simplicity we choose not to indicate this in the notations.

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

$$\liminf_{n \longrightarrow \infty} \inf_{j \le l, k} \inf_{a} \frac{p_{\Phi_n}(\lambda_{j,k}, a) + 1/n^2}{p_{\mu_{j,k}}(\widetilde{\lambda}_{j,k}, a)} \ge 1, \tag{2.9}$$

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

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

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

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

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

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

Thus

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

Hence without loss of generality we assume that $\sup_{j \le l, k} \lambda_{j,k} \le \lambda_n \sim \dfrac{\sqrt{2 \log n}}{\sqrt{n}}$ and to complete the proof, it is enough to show that for suitable thresholds

$\widetilde{\lambda}_{j,k}$,

$$\liminf_{n\longrightarrow\infty} \inf_{j\leq l,k} \inf_{a} \frac{p_{\Phi_n}(\lambda_{j,k},a)}{p_{e_{j,k}}(\widetilde{\lambda}_{j,k},a)} \geq 1. \tag{2.10}$$

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 inconvenient, we will multiply the random variables and thresholds by $\sqrt{n}$, and the risks by n. The size of the fraction in (2.10) is not changed by this transformation.*

We will need the following inequality,

$$\begin{aligned}
p_\Phi(\lambda,a) &= a^2\mu(-\lambda < x+a < \lambda) + \int_{\lambda-a}^{\infty} (x-\lambda)^2\Phi(dx) \\
&\quad + \int_{-\infty}^{-\lambda-a} (x+\lambda)^2\Phi(dx) \\
&\geq a^2\Phi(-\lambda-a < x < \lambda-a) + \int_{\lambda}^{\infty} (x-\lambda)^2\Phi(dx) \\
&\quad + \int_{-\infty}^{-\lambda-a} (x+\lambda+a)^2\Phi(dx) \\
&\geq a^2\Phi(-\lambda-a < x < \lambda-a) + \frac{p_\Phi(\lambda,0)+p_\Phi(\lambda+a,0)}{2}
\end{aligned} \tag{2.11}$$

since Φ is symmetric. We also need another inequality: if $\lambda \geq 1$ then

$$\begin{aligned}
p_\Phi(\lambda,0) &\geq 2\Phi(\{x > \lambda+1\}) \\
&\geq \frac{2}{\sqrt{2\pi}}\left(\frac{1}{\lambda+1} - \frac{1}{(1+\lambda)^3}\right)\exp(-(\lambda+1)^2/2) \\
&\geq \frac{1}{\sqrt{2\pi}\lambda}\exp(-(\lambda+1)^2/2),
\end{aligned} \tag{2.12}$$

where the second inequality follows from a lower bound for the error function (see [17, p.850]). From μ_n we know that we can apply Lemma 2.3 and Lemma 2.4. By Lemma 2.5 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 (2.6)) independent of the index of the wavelet coefficient such that for $\alpha_n := \sqrt{\varepsilon_1 2\log n}$, and for all c with $|c| < \alpha_n$

$$\beta_n \leq \frac{\Phi((c,\infty))}{\mu_n((c,\infty))} \text{ and } \beta_n \leq \frac{\Phi((-\infty,c))}{\mu_n((-\infty,c))}. \tag{2.13}$$

We now make a distinction of cases to prove (2.10), the first case is $\lambda < \alpha_n/2$, the other is $\lambda \geq \alpha_n/2$.

Assume $\lambda < \alpha_n/2$. For fixed λ, let $r_a(x) := (T_\lambda^S(x+a)-a)^2$, 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 (2.13) if easily follows that

$$\frac{\int_{-\infty}^{\infty} r_a(x)d\Phi(x)}{\int_{-\alpha_n}^{\alpha_n} r_a(x)d\mu_n(x)} \geq \beta_n. \tag{2.14}$$

Further because of Lemma 2.4,

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

with $c_n = 1 - n^{-\varepsilon}\log n/2$. The last identity follows from $\lambda < \alpha_n/2$ using (2.12). Since $p_\Phi(\lambda,0) \leq p_\Phi(\lambda, a)$, the assertion (2.10) 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 2.4 and relation (2.12) that $\lambda/\tilde{\lambda} \longrightarrow 1$ 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 $p_n(\widetilde{\lambda}, a) \leq a^2 + p_n(\widetilde{\lambda}, 0)$ and (2.11), it follows

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

where 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 $p(\widetilde{\lambda}, a) > 1/2$ if α_n is sufficiently large, further

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

since $\widetilde{\lambda} \sim \lambda \leq \lambda_n \sim \sqrt{2\log n}$. Like in the paragraph above for the equation (2.14), it is easy to see that

$$\inf_{\lambda\geq\alpha_n/2} \inf_{a\geq 1} \frac{\int_{-\infty}^{\infty}(T^S_{\widetilde{\lambda}}(x+a) - a)^2\Phi(dx)}{\int_{-\alpha_n}^{\alpha_n}(T^S_{\widetilde{\lambda}}(x+a) - a)^2\mu_n(dx)} \longrightarrow 1,$$

thus

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

To finish the proof we show that

$$\begin{aligned}\liminf_{n\longrightarrow\infty} \inf_{\lambda\geq\alpha_n/2} \inf_{a>1} \frac{p_\Phi(\lambda, a)}{p_\Phi(\widetilde{\lambda}, a)} &= \liminf_{n\longrightarrow\infty} \inf_{\lambda\geq\alpha_n/2} \inf_{a>1} \frac{\int_{-\infty}^{\infty}(T^S_{\lambda}(x+a) - a)^2\Phi(dx)}{\int_{-\infty}^{\infty}(T^S_{\widetilde{\lambda}}(x+a) - a)^2\Phi(dx)} \\ &\geq 1. \qquad (2.15)\end{aligned}$$

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

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

If $a > \alpha_n/4$, then

$$\inf_{a>\alpha_n/4} \inf_{x\leq\widetilde{\lambda}-\alpha_n/4} \frac{T^S_\lambda((x+a)-a)^2}{T^S_{\widetilde{\lambda}}((x+a)-a)} \geq \left(\frac{\alpha_n/4-\lambda}{\alpha_n/4-\widetilde{\lambda}}\right) \longrightarrow 1.$$

Since

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

and $p_\Phi(\lambda, a) > 1/2$, the relation (2.15) holds for $a > \alpha_n/4$. □

3 Very heavy tails

The conditions for Theorem 2.1 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 used is simple, we apply a median filter to our data and then apply wavelet thresholding to this data.

In the paper of Delyon and Juditsky conditions are given for soft thresholding to achieve the same minimax rate as for the Gaussian case. In some instances a finite third moment is enough. Kovac and Silverman ([12]) used median based filtering for detecting outliers.

In the following given $a_1, \ldots, a_{2k+1}$, $\mathrm{med}(a_1, \ldots, a_{2k+1})$ denotes 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 $\mathrm{med}(a_i, 2k+1)$ for $\mathrm{med}(a_{i-k}, \ldots, a_{i+k})$. If i is smaller than k, then $\mathrm{med}(a_i, 2k+1) := \mathrm{med}(a_1, \ldots, a_{2k+1})$, a similar boundary correction is performed for the largest indexes. The following lemma, whose proof is clear, makes the advantage of the median filter also clear:

Lemma 3.1 *Let $X_1, \ldots, X_{2k-1}$ be independent random variables and $c, p > 0$ such that*

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

then

$$P(\mathrm{med}(X_1,\ldots,X_{2k-1}) > x) \le \binom{2k-1}{k}\frac{c}{x^{kp}},$$

and thus $\mathrm{med}(X_1,\ldots,X_{2k-1})$ *has moments of order* $kp-\varepsilon$*, for all* $\varepsilon > 0$*.*

Thus for example the median of seven Cauchy 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. As usual, we are given $X_i = f_i + n^{-1/2}z_i$, $i = 1,\ldots,n$, $n = 2^h$, 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 $m > 0$, $1 \le p,q \le \infty$, $m > 1/p$ and $s = m + 1/2 - 1/p$.

Theorem 3.2 *Let* $P(z_1 > x) = O(1/x^c)$ *for some* $c > 0$ *and let* $A, B > 0$*. Then there exists an* l *and thresholds* $\lambda_{j,k}$ *such that*

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

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

$$\sum_{i=1}^n (E\mathrm{med}(X_i, 2l+1) - f_i)^2, \qquad (3.1)$$

which we introduce by filtering the initial data with a median filter. This condition is not that strong and in most cases it follows from the Besov norm condition. We will take another look at this later (or rather below). Note that the ℓ_2-norm of the bias in the wavelet coefficients

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

is the same as (3.1).

Proof: Since the distribution of the z_i is symmetric $Emed(z_i, 2l+1) = 0$. 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.

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

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

$$\begin{aligned}
&E(T^S_{\lambda_{j,k}}(a_{j,k}+b_{j,k}+e_{j,k})-a_{j,k})^2\\
&\quad = E(T^S_{\lambda_{j,k}}(a_{j,k}+b_{j,k}+e_{j,k})-T^S_{\lambda_{j,k}}(a_{j,k}+e_{j,k})+(T^S_{\lambda_{j,k}}(a_{j,k}+e_{j,k})-a_{j,k}))^2\\
&\quad \le E(T^S_{\lambda_{j,k}}(a_{j,k}+e_{j,k})-a_{j,k})^2+b_{j,k}^2+2|b_{j,k}|\sqrt{E(T^S_{\lambda_{j,k}}(a_{j,k}+e_{j,k})-a_{j,k})^2},
\end{aligned}$$

since $|T^S_\lambda(x+a)-T^S_\lambda(x)| \le |a|$. Thus

$$\begin{aligned}
&\sum_{j,k} E(T^S_{\lambda_{j,k}}(a_{j,k}+b_{j,k}+e_{j,k})-a_{j,k})^2\\
&\quad \le \sum_{j,k} E(T^S_{\lambda_{j,k}}(a_{j,k}+e_{j,k})-a_{j,k})^2+\sum_{j,k} b_{j,k}^2\\
&\qquad +2\sqrt{\sum_{j,k} b_{j,k}^2}\sqrt{\sum_{j,k} E(T^S_{\lambda_{j,k}}(a_{j,k}+e_{j,k})-a_{j,k})^2}.
\end{aligned}$$

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

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

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

$$\begin{aligned}
\sum_{i=1}^n |E\mathrm{med}(X_i,2l+1)-f_i|^2 &\le \sum_{i=l+1}^{n-l} 2l \sum_{j=-l+1}^{l}|f_{i+j}-f_{i+j-1}|^2\\
&\quad + 2l^2\sum_{j=2}^{2l+1}|f_j-f_{j-1}|^2+2l^2\sum_{j=n-2l+1}^{n}|f_j-f_{j-1}|^2\\
&\le 8l^2\sum_{i=2}^n|f_i-f_{i-1}|^2 = O(1/n).
\end{aligned}$$

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

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

where $\tilde{z}_i := \mathrm{med}(z_i, D)$ and D is chosen such that $E|\tilde{z}_1|^L < \infty$ and L satisfies (depending on p) the moment conditions of Theorem 2.1. So we could apply Theorem 2.1 if the $\tilde{z}_i$ were independent. The following two lemmas are adapted to the new situation (the D–dependent case).

Lemma 3.3 *Let $X_1, \ldots, X_n$, be zero mean, bounded random variables, with $\|X_i\|_\infty \leq K$, and let X_i, X_j be independent if $|i-j| \geq D$. Let $S_j = \sum_{i=0}^{[(n-1)/D]} X_{iD+j}$, $j = 1, \ldots, D$, $\sigma_j = \sqrt{ES_j^2}$, and $\sigma_{max} = \max_{j=1,\ldots,D} \sigma_j$ (we simply set $X_k = 0$ for $k > n$). Then*

$$P\left(\sum_{i=1}^n X_i > \lambda\right) \leq D \begin{cases} \exp\left(\frac{-\lambda^2}{4D^2\sigma_{max}^2}\right) & : \lambda \leq \frac{\sigma_{max}^2 D}{K} \\ \exp\left(\frac{-\lambda}{4KD}\right) & : \lambda \geq \frac{\sigma_{max}^2 D}{K} \end{cases}. \tag{3.2}$$

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 $\lambda \leq \dfrac{\sigma_{max}^2 D}{K}$, then $\dfrac{\lambda^2}{4D^2\sigma_{max}^2} \leq \dfrac{\lambda}{4KD}$.
□

Lemma 3.4 *Let* $X_1, \ldots, X_n$ *be random variables with* $EX_i = 0$, $E|X_i|^3 \leq Cn^{-3/2}$, $i = 1, \ldots, n$ *and let* X_i, X_j *be independent if* $|i-j| \geq D$. *Let* $S_j = \sum_{i=0}^{(n-1)/D} X_{iD+j}$, $j = 1, \ldots, D$, $\sigma_j = \sqrt{ES_j^2}$ *and* $\sigma_{max} = \max_{j=1,\ldots,D} \sigma_j$ *(again* $X_k := 0$ *for* $k > n$*). Then for all* $\varepsilon < 1$,

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

and

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

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 from Lemma 2.5. □

It is clear that in the context of Lemma 3.4 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 3.4 it is possible to prove a similar variation of Lemma 2.4.
The rest of the proof is quite similar to the proof of Theorem 2.1. 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 2.1. Thus by the same reasoning as in the proof of Theorem 2.1 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.

Of important in the proof of Theorem 2.1 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 $\tilde{z}_1 = \ldots = \tilde{z}_{(D+1)/2}$ and $z_{n-(D-1)/2} = \ldots = \tilde{z}_n$. 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 Lemma 3.4 and 3.3. 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 $e_{j,k,l} = \sum_{i=1}^{n/D} c_{j,k,iD+l} \tilde{z}_{iD+l}$, $\sigma_{j,k,l}^2 = E e_{j,k,l}^2$ and $\sigma_{max,j,k}^2 = \max_l \sigma_{j,k}^2$, clearly $\sigma_{max,j,k}^2 \leq E\tilde{z}_1^2$. Given this and Lemma 3.3 and 3.4 we can now proceed as in the proof of Theorem 2.1. 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. □

Also of importance in Lemma 3.4 and 3.3 is the term σ_{max}, we show next that in general

$$\sigma_{max,j,k} \approx E\tilde{z}_1^2 / D \tag{3.3}$$

Let β be the Hölder continuity of the wavelet base. Let $h = \log_2 n$. Since ψ is compactly supported and Hölder continuous with exponent β, we know that (see, e.g., (4.2) in [1])

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

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. 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. But $\sigma^2_{j,k,l} = E\tilde{z}_1^2 \sum_i c^2_{j,k,iD+l}$ and $|\{i : c_{j,k,i} \neq 0\}| = O(2^{h-j})$ which is a consequence of using compactly supported wavelets. Indeed, recall that

$$\phi_{j,k} = \sum_{i=0}^{2^{j_0-j}(N-1)} u_{j_0-j,i+2^{j_0-j}k}\phi_{j_0,i}, \tag{3.4}$$

and

$$\psi_{j,k} = \sum_{i=0}^{2^{j_0-j}(N-1)} v_{j_0-j,i+2^{j_0-j}k}\phi_{j_0,i}, \tag{3.5}$$

where $u_{\cdot,\cdot}$ and $v_{\cdot,\cdot}$ only depend on the filter (h). This claim about the length of the filters (u_{j_0-j}) and (u_{j_0-j}) can be proved via a simple induction argument. Actually (see [1]), $\max_i |u_{j_0-j,i}| = O(2^{(j_0-j)/2})$ and $\max_i |v_{j_0-j,i}| = O(2^{(j_0-j)/2})$.

Thus,

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

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

We now turn to the problem to know when

$$\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 = O(1/n) \tag{3.6}$$

follows from $\|a\|_{B^m_{p,q}} \leq A$. If $m \leq 1$ then assume $\beta \geq m$, where β is the Hölder continuity exponent 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

$$\begin{aligned}
\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 &= \sum_{i=1}^{n-1} \left(\sum_{j,k} a_{j,k}(c_{j,k,i} - c_{j,k,i+1}) \right)^2 \\
&\leq \sum_{i=1}^{n-1} h \sum_{j=0}^{h-1} \left(\sum_{k} a_{j,k}(c_{j,k,i} - c_{j,k,i+1}) \right)^2 \\
&\leq \sum_{i=1}^{n-1} h \sum_{j=0}^{h-1} C_3 \sum_{k} (a_{j,k}(c_{j,k,i} - c_{j,k,i+1}))^2 \\
\text{since } |\{k : c_{j,k,i} \neq 0 \text{ or } c_{j,k,i+1} \neq 0\}| = O(1),&\ \text{see (3.4, 3.5), } C_3 \text{ is a constant} \\
&= hC_3 \sum_{j=0}^{h-1} \sum_{k} a_{j,k}^2 \sum_{i} (c_{j,k,i} - c_{j,k,i+1})^2 \\
&\leq hC_3C_2 \sum_{j=0}^{h-1} 2^{2\beta(j-h)} \sum_{k} a_{j,k}^2 \\
&\leq hC_3C_2A^2 \sum_{j=0}^{h-1} 2^{2\beta(j-h)} \begin{cases} 2^{-2jm} & : p \geq 2 \\ 2^{-2js} & : p \leq 2, \end{cases}
\end{aligned}$$

where the last inequality follows from the proof of Lemma 2.2. 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

$$\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 = O(\log n / n^{-(2m \wedge 2)}), \tag{3.7}$$

and for $p \leq 2$ we obtain

$$\sum_{i=1}^{n-1} |f_i - f_{i+1}|^2 = O(\log n / n^{-(2s \wedge 2)}). \tag{3.8}$$

Because of the additional condition $m > 1/p$, $2s > 1$, for $p \leq 2$, the condition 3.6 always follows. If $p \geq 2$, then $m > 1/2$ is necessary for the condition (3.6).

Remark 3.5 *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 so–called D–dependent condition). Lemma 3.3 and 3.4 are applicable and then it is easy to follow the line of the proof of Theorem 3.2. 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 for the noise in the random variables to converge in distribution to a normal random variable, only the bounds in Lemma 3.4 and 3.3 are needed. The approach via large deviation results for the wavelet coefficients is also possible for other kind of correlated noise. In [11] 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.*

Again, for stationary Gaussian noise, the distribution of the noise is determined solely by the covariance function which makes life easier. It is not that clear what is meant by correlated non-Gaussian noise. Stationary Gaussian processes with $\sum_{k\in\mathbb{Z}} |r(k)| < \infty$, can be re presented as moving average of a Gaussian white noise process. Thus the corresponding concept for correlated non-Gaussian noise could be 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

$$\sum_l \left(\sum_i a_{l-i} c_{j,k,i} \right)^2 \quad = \quad \sum_l \sum_u \sum_v c_{j,k,u} c_{j,k,v} a_{l-u} a_{l-v}$$

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

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 isn't the signal?*

4 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. The model is the usual one, $X_i = f_i + z_i/\sqrt{n}$, $i = 1, \ldots, n = 2^h$ 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 $\|a\|_{B^m_{p,q}} \leq A$ 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 4.1 in [1], 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

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

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

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

hence

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

and

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

It follows that

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

and

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

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

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

Hence

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

From the previous section (equations (3.7) and (3.8)) we know that $\sum_i |f_{i+1} - f_i|^2 = O(\log n n^{-(2m\wedge 2)})$ or $O(\log n n^{-(2s\wedge 2)})$, 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 $var(\max_{j=0,\dots,D-1} z_j) = 1/2^{D-2} - 1/(2^{2D-2})$. Hence for $p \ge 2$

$$\sup_{\|a\|_{B^m_{p,q}} \le A} E \sum_{i=1}^{n} |\widehat{f}_i - f_i|^2 \le C(D^2 \log n n^{-(2m\wedge 2)} + 2^{-D}), \tag{4.1}$$

where C is a constant. The rate in (4.1) is minimized by choosing $D = 2\log_2 n$ and thus

$$\sup_{\|a\|_{B^m_{p,q}}\le A} E\sum_{i=1}^n |\widehat{f}_i - f_i|^2 = O\left(\frac{(\log n)^3}{n^{2m\wedge 2}}\right), \tag{4.2}$$

If $p \le 2$ then the right hand–side of (4.2) 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

$$\sup_{\|a\|_{B^m_{p,q}}\le A} E\sum_{i=1}^n |\widehat{f}_i - f_i|^2 \le C(D^2 \log n n^{-(2m\wedge 2)} + D^{-2}), \tag{4.3}$$

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

$$\sup_{\|a\|_{B^m_{p,q}}\le A} E\sum_{i=1}^n |\widehat{f}_i - f_i|^2 = O\left(\frac{\log n}{n}\right), \tag{4.4}$$

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

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

Hence we have to minimize

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

If $p \ge 2$ 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 $p \le 2$ 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 [7]). Their method is based on Stein's unbiased risk estimate: Let X_i, $i = 1, \ldots, n$, be iid $N(0, \sigma^2)$ random variables and $g : \mathbb{R}^n \longrightarrow \mathbb{R}$, where $g = (g_1, \ldots, g_n)$ is weakly differentiable, then for all $a \in \mathbb{R}^n$,

$$E\|X+a+g(X+a)-a\|^2 = n\sigma^2+E\|g(X+a)\|^2+2\sigma^2\sum_{i=1}^{n}\frac{\partial}{\partial x_i}g_i(X+a). \quad (5.1)$$

Thus the risk of the estimator $x + g(x)$ can be estimated unbiasedly by $n\sigma^2 + g(x)^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial g_i}{\partial x_i}(x)$. 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 previously seen ([1]), 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 of interest 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

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

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:

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

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

$$\int_{\mathbb{R}} h(x+a)f(x)dx = \int_{\mathbb{R}} xg(x+a)f(x)dx. \tag{5.2}$$

If g is a polynomial the right-hand side of (5.2) is a polynomial in a. It is then well-known that there exists an h such that (5.2) holds. But if for example $g(x) = T^S_\lambda(x)$ 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

$$\int_{\mathbb{R}} g(-x+a)(-x)f(-x)dx = \int_{\mathbb{R}} h(-x+a)f(-x)dx, \tag{5.3}$$

we get:

$$\widehat{g}(w)\widehat{f}'(-w)/i = \widehat{h}(w)\widehat{f}(-w). \tag{5.4}$$

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 (5.2) 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 map $K_f : U_f \longrightarrow L^1(f)$ 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

$$\begin{aligned}&\int_{\mathbb{R}} (h_1 + h_2)(z+a)(f_1 * f_2)(z)dz \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (h_1(x+y+a) + h_2(x+y+a))f_1(x)f_2(y)dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} h_1(x+y+a)f_1(x)dxf_2(y)dy + \int_{\mathbb{R}}\int_{\mathbb{R}} h_2(y+x+a)f_2(y)dyf_1(x)dx \\ &= \int_{\mathbb{R}} g(z+a)z(f_1 * f_2)(z)dz.\end{aligned}$$

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 differentiation operator. In general it is quite complicated to compute K, but from equation (5.4) 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:

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

where the last equality follows from the construction of K:

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

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 $\int_{\mathbb{R}} x^2 f(x)dx = \sigma^2$. For example the soft thresholding estimator and T_λ^M have the following decompositions:

$$T_\lambda^S(x) = x - g_0^+(x) + g_\lambda^+(x) - g_0^-(x) + g_{-\lambda}^-(x), \tag{5.5}$$

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

Another example is obtained if $g : \mathbb{R}^+ \longrightarrow \mathbb{R}$, $g \in C^2$, 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 $\widehat{K_F} = \frac{\lambda\Psi'(-w)}{i}$ 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.*

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

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

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

Since $\int_{\mathbb{R}} (\exp(-ixt) - 1)/xM(dx) = \hat{f}'(-t)/(\hat{f}(-t)i)$, 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 $\lim_{n\longrightarrow\infty} \int_{\mathbb{R}} g_n(x+a)xf(x)dx = \int_{\mathbb{R}} g(x+a)xf(x)dx$ and $\lim_{n\longrightarrow\infty} \int_{\mathbb{R}} K(g_n)(x+a)f(x)dx = \int_{\mathbb{R}} K(g)(x+a)f(x)dx$ 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 $b = 0$, 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 $M(\cdot \cap \mathbb{R}\backslash\{0\})$. 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:*

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

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

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

where the first equality holds since $K_f(1) = 0$, i.e. $\int_{\mathbb{R}} 1Q(dx) = 0$. 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, $\lim_{n\longrightarrow\infty} K_{f_n}(g)(x) = \sigma^2 g'(x)$.

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

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

and hence

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

This is because

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

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

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

Using (5.5) 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 [9, 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 [9, 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 the above examples are infinitely divisible distributions. For the uniform distribution with density $2^{-1}\mathbf{1}_{(-1,1)}(x)$, K does not have a nice form. Assume $g : [-1,1] \longrightarrow \mathbb{R}$ and $2^{-1}\int_{-1}^1 g(x)dx = 0$. If $\bar{g}$ is the 2 periodic continuation of g on $\mathbb{R}$ then $\bar{2}^{-1}g * \mathbf{1}_{(-1,1)} = 0$. Thus unbiased risk estimators are not uniquely determined. Let

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

After some guessing one finds that with

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

(h is the 2 periodic continuation of $-x(x-2)/2$ defined on $[0,2]$ to $\mathbb{R}^+$.) $2^{-1}\int_{-1}^1 h(x+a)dx = r(a)$. So with the help of (5.5) we can now compute an unbiased risk estimator for soft thresholding. The Figure 1 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 mean zero.

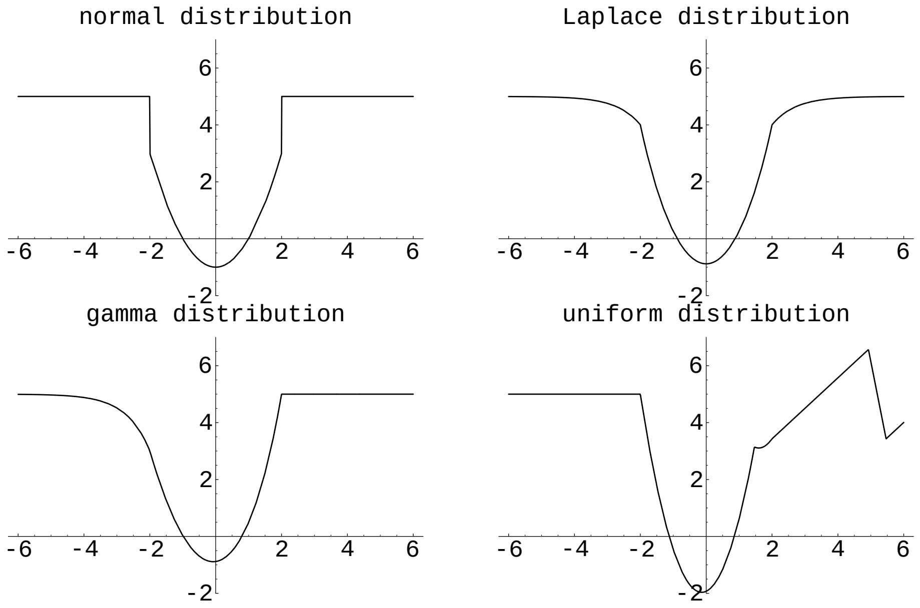

Figure 1: The unbiased risk estimators for soft thresholding

Remark 5.5 *As we have seen in equation (5.1), for normal 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, $i = 1, \ldots, n$ 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 $EX_1 g(X_1 + a_1) = EK_1(g)(X_1 + a_1)$ for some g. If $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $a \in \mathbb{R}^n$ then $E(X_1 + g(X + a) - a_1)^2 = \sigma_1^2 + Eg(X + a)^2 + 2EX_1 g(X + a)$. Given that g satisfies some necessary conditions,*

$$
\begin{aligned}
&EX_1 g(X + a) \\
&\quad = \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} x_1 g(x_1 + a_1, \ldots, x_n + a_n) F_1(dx) \otimes_{i=2}^n F_i(d(x_2, \ldots, x_n)) \\
&\quad = \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} K_1(g(\cdot, x_2 + a_2, \ldots, x_n + a_n))(x_1 + a_1) F_1(dx) \\
&\qquad\qquad\qquad\qquad\qquad \otimes_{i=2}^n F_i(d(x_2, \ldots, x_n)).
\end{aligned}
$$

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 paper and its companion we considered the two popular 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 $\sim \sigma\sqrt{2\log n}$, although we saw later that we can chose thresholds level-wise and this was still a minimax method. For the level j the thresholds were of size $\sim \sigma\sqrt{2j\log 2}$. In the function space approach almost similar thresholds are chosen. Assuming our signal is a member of a bounded subset of the Besov space $B^m_{p,q}$ 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 [4] and [15]). 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 $O(\log n)$. A natural question now is to know if 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, as shown next, for $p \geq 2$. Again, 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 $\|a\|_{B^m_{p,q}} \leq A$, $m > 1/p$, $p \geq 2$ and $q \geq 1$. Let $p(\cdot,\cdot)$ have the usual meaning. If $\lambda \geq a \geq 0$, then

$$p(\lambda, a) \geq a^2\Phi((-\lambda-a, \lambda-a)) + \int_{-\infty}^{-\lambda-a}(x+\lambda)^2\Phi(dx) \geq \frac{a^2}{2}. \qquad (6.1)$$

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 $\sim \sqrt{2j\log 2}/\sqrt{n} = C\sqrt{j}/\sqrt{n}$, 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 $a_{j_0,k} = A\sqrt{2^{-j_0(2m+1)}}$, $k = 0, \ldots, 2^{j_0} - 1$ and $a_{j,k} = 0$ elsewhere, then clearly $\|a\|_{B^m_{p,q}} \leq A$. 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 (6.1) that the risk for thresholding the signal $(a.)$ at level j_0 with thresholds λ_{n,j_0} is at least as large as

$$2^{j_0} A^2 2^{-j_0(2m+1)}/2 = A^2 2^{-j_0 2m}/2. \tag{6.2}$$

Using the definition of j_0 we obtain

$$2^{2m+3} A^2 2^{-j_0(2m+1)} \geq \frac{C^2(j_0 - 1)}{n}. \tag{6.3}$$

Combining the relations (6.2) and (6.3) 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 $B^m_{p,q}$ in the function space context. Thus it is not possible to choose 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 $a_1+e_1, \ldots, a_k+e_k$, ($a.$ 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

([10]) and Cai ([2]). 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 $(j,k),\ldots,(j,k+L)$, now we consider another type of block thresholding estimator, the blocks are vertical and not disjunct. First we have to introduce some 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 $C\in\mathbb{N}$ is a positive 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 $X_i=f_i+e_i$, $i=1,\ldots,n=2^m$ 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 paper and its companion, 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

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

We now have

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

$$\leq \sum_{j,k} E\left(\mathbf{1}_{\{|Y_{j,k}|\geq\lambda_n\}} z_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}|<\lambda_n\}} a_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}|\geq\lambda_n\}} \sum_{(j',k')\text{above}(j,k)} z_{j',k'}^2\right). \tag{6.4}$$

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

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

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

$$\begin{aligned} &\frac{E\left(\mathbf{1}_{\{|Y_{j,k}|\geq\lambda_n\}} z_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}|<\lambda_n\}} a_{j,k}^2 + \mathbf{1}_{\{|Y_{j,k}|\geq\lambda_n\}} \sum_{(j',k') \text{ above } (j,k)} z_{j',k'}^2\right)}{1/n + \min(a_{j,k}^2, 1)} \\ &\leq \ C\log n, \end{aligned} \tag{6.5}$$

where C is a constant. Combining (6.4) and (6.5) 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. The proof is left out, but here is a simple heuristic explanation: In the function space approach all the coefficients above a level* j_0 *are kept and a soft (or hard) thresholding estimator (with threshold* $C\sqrt{(j-j_0)_+}$ *were* j *is the level of a coefficient and* C *is a constant) is applied to the other coefficients. This method achieves the minimax rate. We now compare 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 then* $C\sqrt{(j-j_0)_+}/2$*, 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 in the size of their wavelet coefficients which are above each other. Irregularities, like a discontinuity, in a signal affect all the wavelets coefficients whose "support" contains this irregularity.

Of interest in the vertical block thresholding method is the fact 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 paper, 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,

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

where $K(x,y) = \sum_{j\geq j_0,k}\psi_{j,k}(x)\psi_{j,k}(y)$. If we also keep the 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 which wavelet coefficients are kept with different methods (the artificial signal is just a random signal; the coefficient with level j,k is a random variable with distribution $N(0,2^{-\alpha j}$. Nothing else but these coefficients are present in the signal). 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 an 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 $\widehat{f}_i$ 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 $X_0,\ldots,X_{n-1}$, $n=2^h$ 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

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

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 indices $(j_2, [i/2^{h-j_2}])$, $j_2 < j_1$ 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

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

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 [13]). There the local bandwidth is chosen from a set $a^{-j}h_1$, $a, h_1 > 0$ constants and $j = 0, 1, \ldots$. For the simple kernel estimator based on wavelets, the bandwidth is 2^{-j}, $j = 0, 1, \ldots$, and j is the last level of wavelet coefficients that is kept. This paper of Lepski Mammen and Spokoiny ([13]) shows that kernel estimates with a local varying bandwidth selection can be as good as wavelet thresholding in a minimax sense. The performance of 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 on 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 alround 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 automized curve estimator based on wavelet thresholding, which is included by default in major statistical packages, then one could really compare and decide which method is better for a particular application.

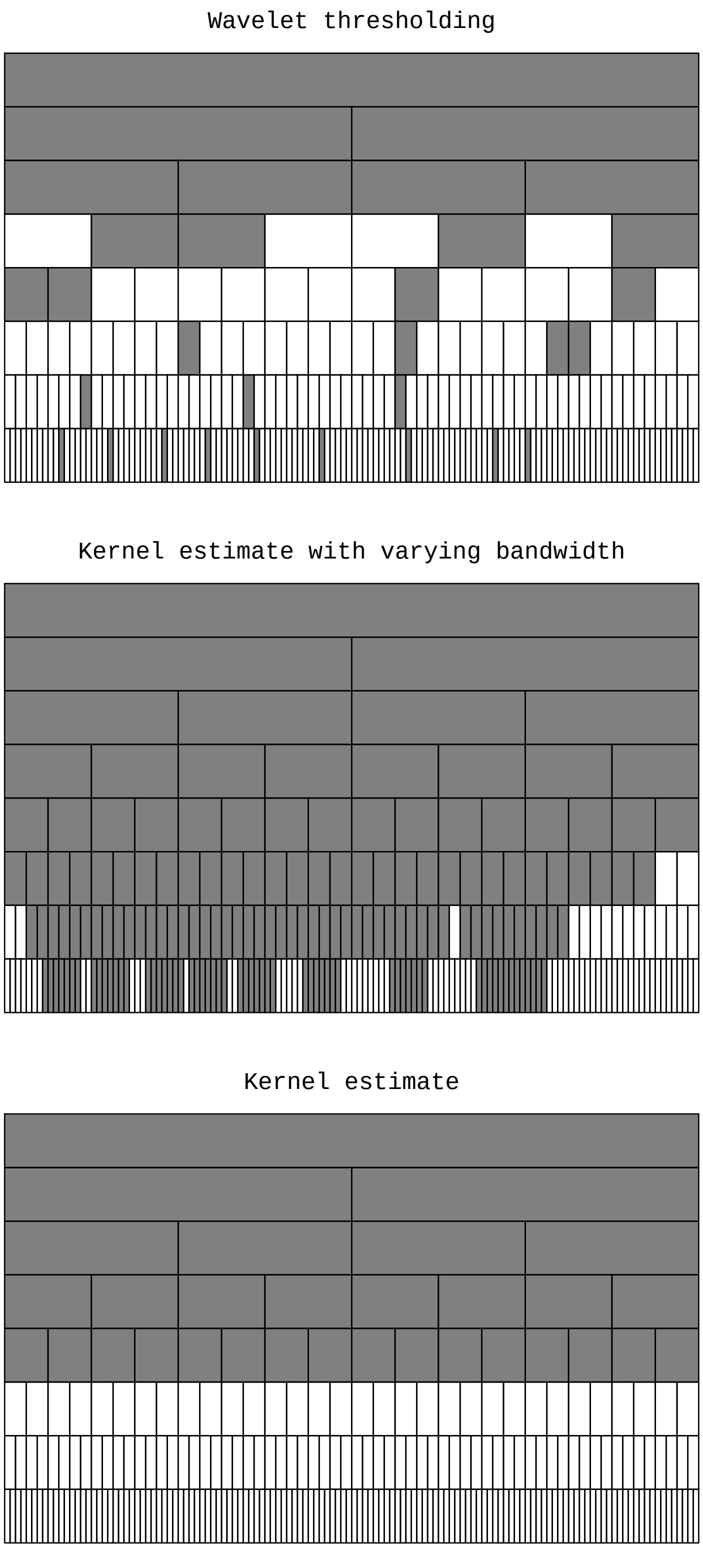

Figure 2: Hard thresholding, vertical block thresholding and kernel estimate

References

[1] Averkamp, R. and Houdré, C. Wavelet Thresholding for non (necessarily) Gaussian Noise: Idealism. Preprint.

[2] Cai, T., Adaptive Wavelet Estimation: A Block Thresholding And Oracle Inequality Approach. Annals of Statistics, in press

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

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

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

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

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

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

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

[10] 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).

[11] 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).

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

[13] 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).

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

[15] 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).

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

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

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

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

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