*Journal of Nonparametric Statistics*
2012, 1–18, iFirst article

Taylor & Francis
Taylor & Francis Group

# Combining thresholding rules: a new way to improve the performance of wavelet estimators

F. Autin[a], J.-M. Freyermuth[b]* and R. von Sachs[c]

[a]*LATP, Aix-Marseille Université, Marseille, France;* [b]*ORSTAT and Leuven Statistics Research Center, K.U.Leuven, Leuven, Belgium;* [c]*ISBA, Université Catholique de Louvain, Louvain la Neuve, Belgium*



(*Received 12 September 2011; final version received 2 July 2012*)

In this paper, we address the situation where we cannot differentiate wavelet-based threshold procedures because their sets of *well-estimated* functions (maxisets) are not nested. As a generic solution, we propose to proceed via a combination of these procedures in order to achieve new procedures which perform better in the sense that the involved maxisets contain the union of the previous ones. Throughout the paper we propose illuminating interpretations of the maxiset results and provide conditions to ensure that this combination generates larger maxisets. As an example, we propose to combine vertical- and horizontal-block thresholding procedures that are already known to perform well. We discuss the limitation of our method, and we check our theoretical results through numerical experiments.

**Keywords:** curve estimation; wavelet methods; maximal spaces; rate of convergence; thresholding rules

*AMS Subject Classifications*: 62G05, 62G20, 41A25, 42C40, 65T60

## 1. Introduction

The literature about wavelet-based nonparametric function estimation is very large. Many thresholding procedures have been proposed and compared from both a practical and a theoretical point of view, particularly thanks to the minimax approach. In the last decade, a new theoretical way, dual to the minimax approach, has been proposed to assess and compare their performances. This approach consists of determining the maxiset of a thresholding procedure that is the maximal functional space for which the quadratic risk of the procedure reaches a given rate of convergence. As previously discussed in Cohen, De Vore, Kerkyacharian, and Picard (2001b), Kerkyacharian and Picard (2000, 2002), Autin (2004, 2008a,b), Autin, Le Pennec, Loubes, and Rivoirard (2010), Autin, Freyermuth, and von Sachs (2011a) and Autin, Freyermuth, and von Sachs (2011b), this approach can be successful at differentiating between minimax-equivalent procedures whenever their maxisets are nested. Without such embeddings, the comparison would be impossible. Hence, the best procedure within a family of thresholding rules – that is, the one with the largest

*Corresponding author. Email: jean-marc.freyermuth@econ.kuleuven.be

ISSN 1048-5252 print/ISSN 1029-0311 online

http://dx.doi.org/10.1080/10485252.2012.709854
http://www.tandfonline.com

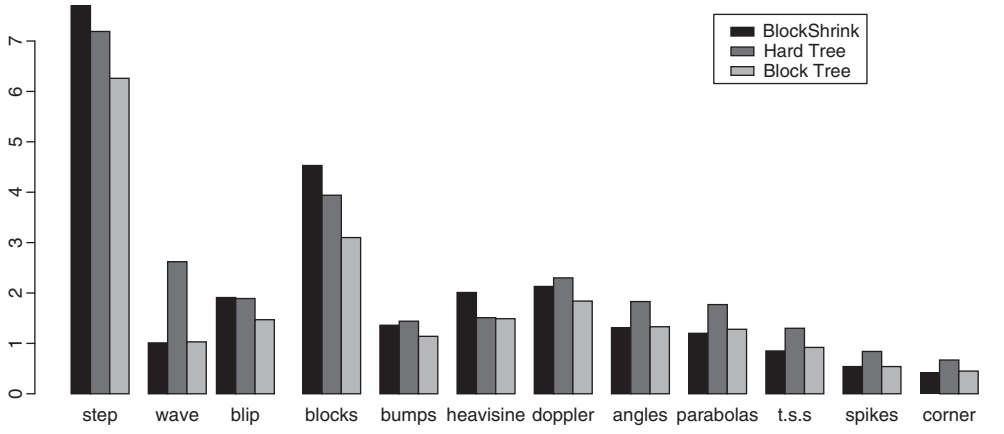


Figure 1. Quadratic risk of estimators Blockshrink, Hard Tree and their combination Block Tree.



maxiset – does not always exist. Even if it has often been viewed as a problem, this just reveals the fact that the procedures are well suited to estimate different classes of functions.

In this paper, we address that situation by taking a different road. Our message is: 'if you cannot differentiate between thresholding rules then combine them'. This is similar to approaches such as aggregation, and model or basis averaging (see Kohn, Marron, and Yau 2000; Barber and Nason 2004; Fryzlewicz 2007). However, in our framework, considering maxisets of thresholding rules, it takes a particular flavour which results in a very specific way of combining these methods (see Proposition 5.1). The resulting new thresholding rule is proved to borrow strength from other well-chosen ones (those with non-nested maxisets) to yield better maxiset. Numerical experiments allow to check that the maxiset approach successfully explains what can be observed in a practical setting.

Taken from the results of our numerical simulations given in detail in Section 6, Figure 1 shows an example of our method that combines two block thresholding rules: horizontal- and vertical-block thresholding procedures which, to the best of our knowledge, are among thresholding rules those with the largest but not nested maxisets encountered in the literature. We recall that estimators induced by these rules are, respectively, the *Blockshrink estimator* studied by Cai (1997) and Autin et al. (2011b) and the *Hard Tree estimator* studied by Autin (2004, 2008a) and Autin et al. (2011a). Our numerical results clearly illustrate the need to use the combination of the previous estimators, called the Block Tree estimator, rather than the Blockshrink estimator or the Hard Tree estimator, since this Block Tree estimator behaves well over all the 12 functions considered here. More information about these numerical experiments can be found in Section 6.

The rest of the paper is organised as follows: Sections 2 and 3 describe our theoretical model and the maxiset approach. In Section 4, we define and give the maxiset properties of the thresholding rules. In Section 5, we describe our method to combine thresholding rules and its limitation. Finally, the detailed proofs of our theoretical results are given in the appendix.

## 2. Background of study

Let us consider a compactly supported wavelet basis of $L_2([0,1])$ with $V$ vanishing moments ($V \in \mathbb{N}^*$) which has been previously periodised $\{\phi, \psi_{jk},\ j \in \mathbb{N},\ k \in \{0,\dots,2^j-1\}\}$. Examples of such bases are given in Daubechies (1992). Any function $f \in L_2([0,1])$ can be written as

follows:

$$f(\cdot) = \alpha\phi(\cdot) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}\psi_{jk}(\cdot). \tag{1}$$

The coefficient $\alpha$ and the components of $\theta = (\theta_{jk})_{j,k}$ are the scaling and wavelet coefficients of $f$, respectively. They correspond to the $L_2$-scalar products between $f$ and the scaling and wavelet functions $\phi$ and $\psi_{jk}$, respectively.

In the prominent denoising context of nonparametric regression, which we also adopt in this work, one considers disposing of $N$ noisy observations $Y_i$ with variance $\sigma^2$, which are modelled as

$$Y_i = f\left(\frac{i}{N}\right) + \sigma\zeta_i, \quad 1 \le i \le N, \ \ \zeta_i \text{ are i.i.d. } \mathcal{N}(0,1). \tag{2}$$



Motivated from Equations (1) and (2), in order to focus on the essentials of developing our results in a general abstract framework, in the sequel, we concentrate on the sequential version of the Gaussian white noise model in the coefficient domain. That is, we assume to dispose of noisy observations $\hat{\theta}_{jk}$ of the wavelet coefficients $\theta_{jk}$ of the target function $f$ and hence picture those as realisations of independent random variables:

$$\hat{\alpha} = \alpha + \epsilon\xi \quad \text{and} \quad \hat{\theta} = (\hat{\theta}_{jk})_{j,k} = (\theta_{jk} + \epsilon\xi_{jk})_{j,k}, \tag{3}$$

where again $\xi$ and $\xi_{jk}$ are i.i.d. $\mathcal{N}(0,1)$, and $0 < \epsilon < \exp(-1)$ is now the abstract noise level. It is well known that this sequence model (3) and the nonparametric regression model (2) are equivalent with the calibration $\epsilon = \sigma/\sqrt{N}$. Hence, considerations which let the noise level $\epsilon$ tend to zero amount to letting the sample size $N$ tend to infinity.

We focus on the performances of *keep-or-kill* estimators (KK-estimators) which are wavelet estimators that can be written as follows:

$$\hat{f}(\cdot) = \hat{\alpha}\phi(\cdot) + \sum_{(j,k)\in\mathcal{K}_\epsilon}\hat{\theta}_{jk}\psi_{jk}(\cdot), \tag{4}$$

where $\mathcal{K}_\epsilon$ is a finite set of indices that may be random or deterministic.

## 3. Maxiset approach

In order to assess the theoretical efficiency of estimators, Cohen et al. (2001b) suggested the maxiset point of view. This new setting offers a complementary approach to the minimax one and was successfully applied in order to differentiate between *minimax-optimal* estimators (see among others (Kerkyacharian and Picard 2002; Autin 2004, 2008b).

In the following, we consider KK-estimators $\hat{f}_{\mu,m}$ associated with a given thresholding rule $\mu$ (for a more general definition, see Definition 4.1). Roughly speaking, these are estimators which set to zero all the empirical wavelet coefficients which are associated with a *score* that is below a threshold of size $mt$ (with $t$ depending on the noise level $\epsilon$), keeping the remaining ones unchanged. Here, the choice of how to construct the threshold rule will be important, for example, taking the maximum or a certain $\ell_p$-norm of the empirical wavelet coefficients.

Generally speaking, providing the *maxiset performance of an estimator* $\hat{f}_{\mu,m}$ means determining the largest functional space (maxiset) $F_{\mu,m}$ over which the $L_2$-risk of this estimator converges at

a prespecified rate $v = v_\epsilon$, that is,

$$\sup_{0<\epsilon<\exp(-1)} v_\epsilon^{-1}\mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 < \infty \iff f \in F_{\mu,m}.$$

Naturally, the rate $v_\epsilon$ tending to zero as $\epsilon \to 0$ translates into a rate of convergence of the KK-estimator for $N \to \infty$, as indicated above. We also refer the reader to the formulation of Theorem 4.6 for an example of a rate $v_\epsilon$ which is classically used in the maxiset framework.

There exists some cases where the spaces $F_{\mu,m}$ generated by a thresholding estimator $\hat{f}_{\mu,m}$ are different for different values of $m$ (this aspect will be exemplified in Section 4.3). Hence, it is worth paying attention to the role of the parameter $m$ that is frequently passed out for asymptotic theory but is crucial for practical purposes. To reduce this gap between the theoretical and the practical setting, we choose to present our maxiset results for thresholding rules instead of thresholding estimators, thereby we shall focus on the largest functional space over which the quadratic risk of a wide range of estimators based on the same thresholding rule converges at a prespecified rate. More specifically, for a positive real number $M$ to be specified hereafter, we shall say that the functional space $\mathcal{G}_{\mu,M}$ is the *maxiset of the thresholding rule* $\mu$ for the rate of convergence $v$ and the $L_2$-risk if and only if

$$\sup_{m\geq 2M}\sup_{0<\epsilon<\exp(-1)} v_\epsilon^{-1}\mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 < \infty \iff f \in \mathcal{G}_{\mu,M}.$$

In other words, the space $\mathcal{G}_{\mu,M}$ can be viewed as the intersection of the functional spaces $F_{\mu,m}$ over all $m \geq 2M$.

Otherwise, from the maxiset point of view, *the larger the maxiset the better the rule*. Obviously, the size of the maxiset depends on the chosen rate; the slower the rate the larger the maxiset. When comparing distinct rules of reconstruction, we say that one is better than the other if the maxiset of the one contains the maxiset of the other, for the same given rate.

The first maxiset results were provided by Cohen et al. (2001b) and Kerkyacharian and Picard (2000, 2002) who determined the maximal functional spaces for estimators based on Hard and Soft thresholding rules, respectively. They also proved that estimators built from the local bandwidth selection rule of Lepski (1991) were at least as efficient as the latter ones.

As discussed in Autin (2004), thresholding rules with larger maxisets can be constructed from rules that are *not elitist* – that is, rules that do not systematically kill all the 'small' empirical wavelet coefficients. As examples, we cite estimators that rely on vertical-block thresholding rules (see Cohen et al. 2001b; Autin 2008a,b; Autin et al. 2011a) or horizontal-block thresholding rules (see, among others, Cai 1997, 1999, 2008; Hall, Kerkycharian, and Picard 1998a,b; Cai and Silverman 2001; Cai and Zhou 2009; Autin et al. 2011b). When looking at these procedures, their maxisets contain those of procedures based on elitist rules, including Hard and Soft thresholding estimators, and also many Bayes procedures (see Autin, Picard, and Rivoirard 2006).

Nevertheless the following open question arises from these previous works: in order to estimate a signal what is the best choice among vertical- and horizontal-block thresholding rules?

As emphasised in Section 1, the maxisets of vertical- and horizontal-block thresholding estimators are not embedded and thus these estimators cannot be differentiated from one another. Even in the practical setting, as shown by the quadratic risks of the estimators in the Figure 1 for several test functions, it seems to be difficult to identify a winning method. Hence, from both a theoretical and a practical point of view, the answer to the question is not clear.

As a way out, in this paper, we propose to combine existing thresholding rules so as to get a new well-performing rule which reconstructs at least as many functions as the ones generated by the vertical- and horizontal-block thresholding rules. To reach this goal, we first introduce a large family of wavelet estimators built from thresholding rules.

## 4. Maxiset properties of thresholding rules

From now on, we adopt the following choices:

- $t_\epsilon = \epsilon\sqrt{\ln(\epsilon^{-1})}$,
- for any given $\lambda > 0$, $j_\lambda$ is the integer such that $2^{-j_\lambda} \leq \lambda^2 < 2^{1-j_\lambda}$.

Note that the first of the two above choices determines notably the order of the threshold value as a function of the noise level (see Definition 4.1). The reader can immediately remark that, in terms of the parameters of the nonparametric regression model (2), it is about the order of the classical universal threshold as detailed in Section 6.2. The second choice will be used, with $\lambda = mt_\epsilon$, in Definition 4.1 to specify the index set $\mathcal{K}_\epsilon$ (as introduced by Equation (4)) of the specific KK-estimator $\hat{f}_{\mu,m}$ of our interest.

### 4.1. *Estimators built from thresholding rules*

Let us now introduce a family of wavelet estimators built from thresholding rules. The following definition is a slight generalisation of the one given by Autin (2008b) in that condition (2.3) therein is not required here.

DEFINITION 4.1 *Let $m > 0$ and consider the sequential model* (3). *An estimator $\hat{f}_{\mu,m}$ is called $(\mu, m)$ thresholding estimator if there exists a thresholding rule $\mu$ that is a sequence of non-negative functions $(\mu_{jk}(m, t_\epsilon, \cdot))_{j,k}$ that are monotonically nonincreasing with respect to $m$, and such that*,

$$\hat{f}_{\mu,m}(\cdot) = \hat{\alpha}\phi(\cdot) + \sum_{j\in\mathbb{N},\, j<j_{mt_\epsilon}} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk} \mathbf{1}\{\mu_{jk}(m, t_\epsilon, \hat{\theta}) > mt_\epsilon\}\psi_{jk}(\cdot)$$

$$:= \hat{\alpha}\phi(\cdot) + \sum_{(j,k)\in\mathcal{K}_{\epsilon,m,\mu}} \hat{\theta}_{jk}\psi_{jk}(\cdot).$$

In the previous expression, $\mathcal{K}_{\epsilon,m,\mu}$ also denotes the set of couple of indices kept by the $(\mu, m)$ thresholding estimator. We remark that a $(\mu, m)$ thresholding estimator $\hat{f}_{\mu,m}$ does not use empirical wavelet coefficients $\hat{\theta}_{jk}$ with $j \geq j_{mt_\epsilon}$.

For a given sequence of empirical wavelet coefficients $\hat{\theta} = (\hat{\theta}_{j,k})_{j,k}$, some examples of thresholding rules $\mu$ and of the $(\mu, m)$ thresholding estimator $(m > 0)$ associated with are the following:

- The Hard thresholding rule $\mu^{\mathrm{H}}$ : $\mu^{\mathrm{H}}_{jk}(m, t_\epsilon, \hat{\theta}) := |\hat{\theta}_{jk}|$.

The $(\mu^{\mathrm{H}}, m)$ thresholding estimator relies on a basic elitist rule which keeps in the signal reconstruction only the empirical wavelet coefficients strictly greater than the threshold value $mt_\epsilon$ in absolute value. The other ones are killed.

- The Hard Tree thresholding rule $\mu^{\mathrm{T}}$ : $\mu^{\mathrm{T}}_{jk}(m, t_\epsilon, \hat{\theta}) := \max_{(j',k')\in\mathcal{T}_{j,k}(mt_\epsilon)} |\hat{\theta}_{j'k'}|$.

The $(\mu^{\mathrm{T}}, m)$ thresholding estimator was already studied by Autin (2008a) and Autin et al. (2011a). It relies on a rule which keeps empirical wavelet coefficients with level strictly less than $j_{mt_\epsilon}$ that are larger in absolute value than the threshold $mt_\epsilon$ and keeps their ancestors in the dyadic tree rooted a $(j_0, k_0) := (0, 0)$ too. Here, for any $j < j_{mt_\epsilon}$, $\mathcal{T}_{j,k}(mt_\epsilon)$ corresponds to the dyadic tree rooted at $(j, k)$ and being reduced to indices with level strictly less than $j_{mt_\epsilon}$. For $j \geq j_{mt_\epsilon}$, it is

the singleton $\{(j,k)\}$. This estimator is tree structured (i.e. the empirical wavelet coefficients that have been kept for the signal reconstruction satisfy the hereditary constraint of Engel 1994). They can be viewed as both a hybrid wavelet version of Lepski's kernel method (see Autin 2008a) and a vertical-block thresholding method (see Autin et al. 2011a).

- The BlockShrink rule $\mu^{\mathrm{B}}$ : $\mu^{\mathrm{B}}_{jk}(m,t_\epsilon,\hat{\theta}) := (\sum_{k'\in\mathcal{P}_{j,k}(\epsilon)}\hat{\theta}^2_{jk'})^{1/2}$.

The $(\mu^{\mathrm{B}},m)$ thresholding estimator was studied by Cai (1997) and Autin et al. (2011b). It relies on a rule which keeps empirical wavelet coefficients if the $l_2$-norm of the empirical wavelet coefficients from their block is larger than the threshold value $mt_\epsilon$. Here, $\mathcal{P}_{j,k}(\epsilon)$ denotes the block that contains $(j,k)$. Blocks are non-overlapping with common size $\lceil \ln(\epsilon^{-1})\rceil = \lceil -\ln(F^{-1}(t_\epsilon))\rceil$, where $F^{-1}$ is the inverse function of $F:\epsilon\longrightarrow F(\epsilon):=t_\epsilon$ and $\lceil x\rceil$ denotes the smallest integer bigger than or equal to $x$. A precise description is given in Cai (1997) and Autin et al. (2011b) in particular for handling boundaries.

More generally, for wise choices of $\mu$, the resulting estimators show good theoretical and practical performance. In particular, we recall in the next section that the sets of functions they are able to *well estimate* are quite large for near minimax rates (see Cohen, Dahmen, Daubechies, and DeVore 2001a; Autin 2004, 2008b).

### 4.2. *Maxiset results*

In this section, our aim is twofold. We provide the maxisets of thresholding rules in Theorem 4.6, and we recall sufficient conditions to guarantee large maxisets such as the cautiousness of a thresholding rule (see Definition 4.7 and Corollary 4.9). To begin, let us define the functional spaces that shall appear in our future maxiset results.

DEFINITION 4.2 *Let $0<u<V$, where $V$ is the number of vanishing moments of the chosen wavelet basis. A function $f\in L_2([0,1])$ belongs to the Besov space $\mathcal{B}^u_{2,\infty}$ if and only if*:

$$\sup_{J\geq 0} 2^{2Ju}\sum_{j\geq J}\sum_{k=0}^{2^j-1}\theta^2_{jk}<\infty.$$

Following Autin (2004), for any chosen rate $v$, Besov spaces $\mathcal{B}^u_{2,\infty}$ usually appear when studying the maxisets of wavelet estimators that kill any empirical wavelet coefficient with a level greater than or equal to a maximum resolution level $j_\epsilon = O(\ln(v_\epsilon^{-1}))$ $(0<\epsilon<\exp(-1))$.

DEFINITION 4.3 *Let $m'\geq 1$, $0<r<2$ and a thresholding rule $\mu$ be given. A function $f\in L_2([0,1])$ belongs to the space $W_{\mu,m'}(r)$ if and only if*

$$\sup_{m\geq m'}\sup_{0<\lambda<\exp(-1)}(m\lambda)^{r-2}\sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1}\theta^2_{jk}\mathbf{1}\{\mu_{jk}(m,\lambda,\theta)\leq m\lambda\}<\infty.$$

The spaces $W_{\mu,m'}(r)$ contain functions for which there is a control of the energy of their wavelet coefficients that do not survive the thresholding rule $\mu$.

We now give sufficient conditions in order to prepare our future maxiset results. As usual in the maxiset setting, we shall suppose that a *large-deviation* property (LD-property) will hold to derive our results. This kind of property ensures that large-deviation quantities that naturally appear when bounding the decomposition of the risk of a $(\mu,m)$ thresholding estimator are small enough.

DEFINITION 4.4 *We say that a thresholding rule $\mu$ satisfies the LD-property if and only if for any given $\nu > 0$ there exists $m_{\mu,\nu} \geq 1$ such that for any $m \geq m_{\mu,\nu}$, any $(j,k)$ and any sequence of real numbers $\theta$ and Gaussian random variables $\hat{\theta}$ connected to $\theta$ via model* (3),

$$\sup_{0<\epsilon<\exp(-1)} \epsilon^{-\nu}\mathbb{P}(|\mu_{jk}(m,t_\epsilon,\hat{\theta}) - \mu_{jk}(m,t_\epsilon,\theta)| > m_{\mu,\nu}t_\epsilon) \leq \frac{1}{2}.$$

*Remark 1* Note that for the examples of thresholding rules we gave, the LD-property is satisfied for

- $\mu^{\mathrm{H}}$, when choosing $m_{\mu^{\mathrm{H}},\nu} = \sqrt{2\nu + 4\ln(2)}$ (due to the concentration inequality for standard Gaussian variables: $\mathbb{P}(|Z| > t) \leq 2\exp(-t^2/2)$, $t > 0$, $Z \sim \mathcal{N}(0,1)$ see Gordon 1941),
- $\mu^{\mathrm{T}}$, when choosing $m_{\mu^{\mathrm{T}},\nu} = \sqrt{2(\nu + 2 + 2\ln(2))}$ (due to the concentration inequality for standard Gaussian variables),
- $\mu^{\mathrm{B}}$, when choosing $m_{\mu^{\mathrm{B}},\nu}$ such that $m^2_{\mu^{\mathrm{B}},\nu} - 2\ln(m_{\mu^{\mathrm{B}},\nu}) = 2\nu + 1$ (obtained from inequality (9.9) given in Cai 1999).

Another definition given hereafter states that for any sequence $\theta = (\theta_{jk})_{j,k}$, the number of coefficients kept by the method must not be too large to hope for large maxisets (see Autin 2004, 2008b). That is the reason why we shall focus on the rules that satisfy the *sparsity* property (S-property).

DEFINITION 4.5 *We say that a thresholding rule $\mu$ satisfies the S-property if and only if there exists $C_\mu > 0$ such that for any $0 < \epsilon < \exp(-1)$, any $m > 0$ and any sequence of real numbers $\theta = (\theta_{jk})_{j,k}$ :*

$$\sum_{j<j_{mt_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\mu_{jk}(m,t_\epsilon,\theta) > \frac{mt_\epsilon}{2}\right\}$$

$$\leq C_\mu \ln(\epsilon^{-1}) \sum_{n\in\mathbb{N}} (m2^n t_\epsilon)^{-2} \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\{\mu_{jk}(m,t_\epsilon,\theta) \leq m2^n t_\epsilon\}.$$

Note that rules $\mu^{\mathrm{H}}$, $\mu^{\mathrm{T}}$ and $\mu^{\mathrm{B}}$ satisfy the S-property (see also Autin 2008b or Autin et al. 2011a,b).

THEOREM 4.6 *Let $s > 0$. Consider a thresholding rule $\mu$ such that the LD-property and the S-property hold. Then, for any $m' \geq m_{\mu,4}$, we have the following equivalence*:

$$\sup_{m\geq 2m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{-4s/(1+2s)} \mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 < \infty \iff f \in \mathcal{G}_{\mu,m'},$$

*with*

$$\mathcal{G}_{\mu,m'} := \mathcal{B}^{s/(1+2s)}_{2,\infty} \cap W_{\mu,m'}(2/(1+2s)). \tag{5}$$

*Remark 2* There is a natural and interesting interpretation of Theorem 4.6. The *S*-property ensures that the functions to be estimated have a sufficient degree of sparsity to be able to control the variance of our thresholding rules at the level required by the prespecified rate. Here, we are naturally interested in thresholding rules that outperform, in the maxiset sense, at least the Hard thresholding one. In order to construct such rules, we got inspired from the definition of the spaces $W_{\mu,m'}(r)$. We find as natural candidates those rules for which the set $\mathcal{K}_{\epsilon,m,\mu}$ contains $\mathcal{K}_{\epsilon,m,\mu^H}$, for any $0 < \epsilon < \exp(-1)$ and any $m > 0$.

DEFINITION 4.7 *We say that a thresholding rule $\mu$ is* cautious *if and only if the following property holds for any $m > 0$, any $(j,k)$ and any $0 < \epsilon < \exp(-1)$ :*

$$\mu_{jk}(m, t_\epsilon, \hat{\theta}) \geq |\hat{\theta}_{jk}| \quad \forall\, \hat{\theta}.$$

Note that a cautious rule $\mu$ does not kill any of the coefficients that are kept by the Hard thresholding rule. In particular,

$$\mathcal{K}_{\epsilon,m,\mu} \supset \mathcal{K}_{\epsilon,m,\mu^{\mathrm{H}}} \quad \text{for any } 0 < \epsilon < \exp(-1) \text{ and any } m > 0.$$

The thresholding rules $\mu^{\mathrm{H}}$, $\mu^{\mathrm{T}}$ and $\mu^{\mathrm{B}}$ are clearly cautious. According to Definition 4.3 and following Remark 2 of Section 4, one gets the following result.

PROPOSITION 4.8 *Let $\mu$ be a cautious rule. Then, for any $m' \geq 1$ and any $0 < r < 2$,*

$$W_{\mu,m'}(r) \supset W_{\mu^{\mathrm{H}},m'}(r).$$

As a direct consequence of Theorem 4.6 and Proposition 4.8, we get the following corollary.

COROLLARY 4.9 *Let $\mu$ be a cautious rule such that the LD-property and the S-property hold. Consider, as in Theorem* 4.6, *$m' \geq m_{\mu,4}$. Then, the set of functions well estimated by the thresholding rule $\mu$ is quite large. Indeed, for any $s > 0$,*

$$f \in \mathcal{G}_{\mu^H,m'} \Longrightarrow \sup_{m \geq 2m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{-4s/(1+2s)} \mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 < \infty,$$

*with $\mathcal{G}_{\mu^{\mathrm{H}},m'} := \mathcal{B}_{2,\infty}^{s/(1+2s)} \cap W_{\mu^{\mathrm{H}},m'}(2/(1+2s))$.*

*Remark 3* We recall that the functional set $\mathcal{G}_{\mu^{\mathrm{H}},m'}$ can be considered as a large functional space since it contains the space $\mathcal{B}_{2,\infty}^s$ (see among others Autin 2004).

### 4.3. *A note for some particular thresholding rules*

Before going further, we pay attention to particular thresholding rules for which the associated spaces $W_{\mu,m'}(r)$ $(0 < r < 2)$ are identical for any value of $m' \geq 1$.

DEFINITION 4.10 *A thresholding rule $\mu$ is said to satisfy the connection property (C-property) if and only if, for any $(j,k)$, any $(m,\epsilon)$ and any sequence of real numbers $\theta$,*

$$\mu_{jk}(m, t_\epsilon, \theta) \text{ only depends on parameters } mt_\epsilon \text{ and } \theta.$$

*In the sequel, we use $\tilde{\mu}_{jk}(mt_\epsilon, \theta) := \mu_{jk}(m, t_\epsilon, \theta)$ to denote a thresholding rule $\mu$ satisfying the C-property.*

PROPOSITION 4.11 *Consider a thresholding rule $\mu$ satisfying the C-property. Then, for any $0 < r < 2$ and any $m' \geq 1$*

$$W_{\mu,m'}(r) = W_{\tilde{\mu}}(r),$$

*where $W_{\tilde{\mu}}(r)$ is the set of functions $f \in L_2([0,1])$ such that*

$$\sup_{\lambda>0} \lambda^{r-2} \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\{\tilde{\mu}_{jk}(\lambda,\theta) \leq \lambda\} < \infty.$$

*Remark 4*

(a) The proof of the previous proposition is obvious by considering the required change of variables.

(b) Note that both $\mu^{\mathrm{H}}$ and $\mu^{\mathrm{T}}$ satisfy the $C$-property, whereas $\mu^{\mathrm{B}}$ does not. For the latter case, various values of $m'$ generate distinct functional spaces $W_{\mu^{\mathrm{B}},m'}(r)$. In particular, note that $W_{\mu^{\mathrm{B}},m'}(r)$ can be rewritten as the space of functions $f$ such that

$$\sup_{m\geq m'}\sup_{0<\lambda<m\exp(-1)}\lambda^{r-2}\sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\left(\sum_{k'\in\mathcal{P}_{j,k}(F^{-1}(\lambda/m))}\theta_{jk'}^2\right)^{1/2}\leq\lambda\right\}<\infty.$$

## 5. Combining thresholding rules to get larger maxisets

We present a method to get more powerful thresholding rules in the maxiset sense and to address the problem of not nested maxisets. Following Remark 2, we are naturally interested in combining many thresholding rules that have non-nested maxisets. Nevertheless, we must keep in mind that too *greedy* thresholding rules, in the sense of rules being associated with a set $\mathcal{K}_{\epsilon,m,\mu}$ with too large cardinality, could induce poor maxisets too. A way to ensure large maxisets is to preserve the $S$-property (see Definition 4.5) when combining the thresholding rules. This leads to a limitation of our method that will be detailed in Section 5.2.

### 5.1. *Asking for the maximum of thresholding rules*

The next proposition gives the way to combine thresholding rules in an appropriate manner. It can be viewed as a special case of model averaging in the coefficient domain over thresholding rules. As our objective is to enlarge the space $W_{\mu,m'}$, we have to increase the score of the combination. Therefore, we consider the maximum of the scores over the thresholding rules as described hereafter.

PROPOSITION 5.1 *Let $\mu^{(1)}$ and $\mu^{(2)}$ be two thresholding rules which satisfy the LD-property. Consider, for any $m>0$,*

$$\begin{aligned}\hat{f}_{\mu^{(3)},m}(\cdot)&=\hat{\alpha}\phi(\cdot)+\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\hat{\theta}_{jk}\mathbf{1}\{\mu_{jk}^{(3)}(m,t_\epsilon,\hat{\theta})>mt_\epsilon\}\psi_{jk}(\cdot)\\&=\hat{\alpha}\phi(\cdot)+\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\hat{\theta}_{jk}\mathbf{1}\{\max(\mu_{jk}^{(1)}(m,t_\epsilon,\hat{\theta}),\mu_{jk}^{(2)}(m,t_\epsilon,\hat{\theta}))>mt_\epsilon\}\psi_{jk}(\cdot).\end{aligned}$$

*Then, $\mu^{(3)}$ is a thresholding rule satisfying the LD-property, with $m_{\mu^{(3)},\nu}=\max(m_{\mu^{(1)},\nu+1},m_{\mu^{(2)},\nu+1})$, for any $\nu>0$.*

Proposition 5.1 reflects the key point of our method to get thresholding rules with larger maxisets. Indeed, the following corollary of Theorem 4.6 and Proposition 5.1 holds.

COROLLARY 5.2 *Let $s>0$ and $\mu^{(1)}$ and $\mu^{(2)}$ be two thresholding rules which satisfy the LD-property and the S-property. Consider estimators $\hat{f}_{\mu^{(3)},m}$ ($m>0$) defined as in the previous lemma. If $\mu^{(3)}$ satisfies the S-property too, then for any $m'\geq m_{\mu^{(3)},4}$,*

(a) $\sup_{m\geq 2m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{-4s/(1+2s)} \mathbb{E}\|\hat{f}_{\mu^{(3)},m} - f\|_2^2 < \infty \iff f \in \mathcal{G}_{\mu^{(3)},m'}$.
(b) $\mathcal{G}_{\mu^{(3)},m'} \supset \mathcal{G}_{\mu^{(1)},m'} \cup \mathcal{G}_{\mu^{(2)},m'}$.

*Definitions of the spaces* $\mathcal{G}_{\mu^{(i)},m'}$ $(i \in \{1,2,3\})$ *are done in Equation* (5).

In Corollary 5.2, the equivalence given in (a) means that considering the maximum of two thresholding rules generates a new thresholding rule for which maxiset have been determined, provided that the *LD*-property and the *S*-property are satisfied. The embedding property (b) is quite interesting since it proves that from two chosen thresholding rules $\mu^{(1)}$ and $\mu^{(2)}$ with possibly not nested maxisets and satisfying the *LD*-property and the *S*-property, we are able to construct a new rule $\mu^{(3)}$ which is at least as efficient as the two previous ones in the maxiset sense, provided that $\mu^{(3)}$ satisfies the *S*-property.

*Remark 1* The corollary 5.2 shows that this way to combine thresholding rules yields larger maxisets. We would like to remark that this methodology holds for shrinkage rules, too.



### 5.2. *Limitation of the method*

In this paragraph, we point out a limitation of our method to achieve enlargements of maxisets. In fact, we will see that we cannot treat the case of a function for which any cautious rule would fail to estimate with the prespecified rate. We deduce this limitation from the facts that, first, we naturally want that maxiset to contain one of the Hard thresholding; second, if at least one of the thresholding rules is cautious, then, its combination with other rules is cautious too. We give this result in Theorem 5.4 but we need first to define a new important functional space.

DEFINITION 5.3 *Let* $m' \geq 1$, $0 < r < 2$ *and* $F^{-1}$ *be the function defined in Section* 4.1. *A function* $f$ *belongs to the space* $W_{\star,m'}(r)$ *if and only if*

$$\sup_{0<\lambda<4m'\exp(-1)} \lambda^{r-2} \left(F^{-1}\left(\frac{\lambda}{4m'}\right)\right)^2 \sum_{j\leq j_\lambda+1} \sum_{k=0}^{2^j-1} \mathbf{1}\{|\theta_{jk}| > \lambda\} < \infty.$$

The spaces $W_{\star,m'}(r)$ play a crucial role in the limitation in our method as we shall see Theorem 5.4. These spaces characterise functional spaces that are upper bounds for the maxisets of cautious rules. From their definition, we deduce that functions that are not sparse enough – that is, possess too many large wavelet coefficients – will not be estimated at the prespecified rate by using a cautious rule.

THEOREM 5.4 (Limitation of the method for enlarging maxisets) *Let* $s > 0$ *and* $\mu$ *be a cautious rule satisfying the LD-property. Then, for any* $m' \geq m_{\mu,4}$,

$$\sup_{m\geq 2m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{-4s/(1+2s)} \mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 < \infty \Longrightarrow f \in \mathcal{G}_{\star,m'},$$

*with* $\mathcal{G}_{\star,m'} := \mathcal{B}_{2,\infty}^{s/(1+2s)} \cap W_{\star,m'}(2/(1+2s))$.

As a conclusion, we proved first that larger maxisets can be obtained by combining existing thresholding rules that satisfy both the LD- and the S-properties. Second, there exists a well-defined limitation to this method, meaning that thresholding rules emerging from our procedure fail whenever we are dealing with functions that cannot be estimated with the prespecified rate by any cautious rule.

## 6. Example: Combining block thresholding rules

### 6.1. *Maxiset for the Block Tree rule*

In this section, we provide an example of our method by combining the Blockshrink and the Hard Tree rules, as we suggested in Section 1. To the best of our knowledge, the largest maxisets of thresholding rules which have been provided up to now are those of the thresholding rules $\mu^{\mathrm{T}}$ and $\mu^{\mathrm{B}}$ and they are known not to be embedded.

As previously precised, these two thresholding rules satisfy the LD-property and the S-property. When combining these two rules, we get the following rule, called Block Tree rule,

$$\mu_{jk}^{\mathrm{BT}}(m, t_\epsilon, \hat{\theta}) := \max\left(\max_{(j',k')\in\mathcal{T}_{j,k}(mt_\epsilon)} |\hat{\theta}_{j'k'}|, \left(\sum_{k'\in\mathcal{P}_{j,k}(\epsilon)} \hat{\theta}_{jk'}^2\right)^{1/2}\right)$$

that clearly satisfies the S-property. From Corollary 5.2, we get the following.

COROLLARY 6.1 *Let $s > 0$ and $m' \geq \max(m_{\mu^{\mathrm{T}},5}, m_{\mu^{\mathrm{B}},5})$. Then,*

$$\sup_{m\geq 2m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{-4s/(1+2s)} \mathbb{E}\|\hat{f}_{\mu^{\mathrm{BT}},m} - f\|_2^2 < \infty \iff f \in \mathcal{G}_{\mu^{\mathrm{BT}},m'},$$

*with $\mathcal{G}_{\mu^{\mathrm{BT}},m'} := \mathcal{B}_{2,\infty}^{s/(1+2s)} \cap W_{\mu^{\mathrm{BT}},m'}(2/(1+2s))$.*

### 6.2. *Numerical experiments*

We propose to illustrate our theoretical results with the following numerical experiments. Let us recall, from Equation (2), the notations of the nonparametric model we are dealing with:

$$Y_i = f\left(\frac{i}{N}\right) + \sigma\zeta_i, \quad 1 \leq i \leq N, \quad \zeta_i \text{ are i.i.d. } \mathcal{N}(0,1),$$

as well as the classical calibration $\epsilon = \sigma/\sqrt{N}$.

Using this model, we generate the data sets from a large panel of functions often used in wavelet estimation studies (see Antoniadis, Bigot, and Sapatinas 2001), the number of observations is $N = 2048$ and the signal-to-noise ratio, defined as the logarithmic decibel scale of the ratio of the standard deviation of the function values to the standard deviation of the noise, is set to 10. We use the Daubechies least asymmetric wavelets with eight vanishing moments. We use the universal threshold value $\hat{\sigma}\sqrt{2\ln(N)}$ for the Hard Tree estimator $\hat{f}_{\mathrm{HT}}$ and $\hat{\sigma}\sqrt{5\ln(N)}$ for the Blockshrink estimator $\hat{f}_{\mathrm{B}}$ as suggested in Cai (1997). We adopt the standard approach to estimate $\sigma$ by computing the median absolute deviation over the thresholded wavelet coefficients at the finest wavelet scale (see e.g. Vidakovic 1999).

The integrated squared error of the estimator $\hat{f}$ at the $u$th Monte Carlo replication ($1 \leq u \leq U$) ($\mathrm{ISE}^{(u)}(\hat{f})$) is computed as follows:

$$\mathrm{ISE}^{(u)}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}^{(u)}\left(\frac{i}{N}\right) - f\left(\frac{i}{N}\right)\right)^2.$$

The mean ISE (MISE) is computed over $U = 200$ Monte Carlo replications:

$$\mathrm{MISE}(\hat{f}) = \frac{1}{U}\sum_{u=1}^{U} \mathrm{ISE}^{(u)}(\hat{f}).$$

Table 1. MISE ($10^{-4}$), number of false positives/negatives and average size of the number of non-zero empirical wavelet coefficients in the estimator.

| | $\hat{f}_{\mathrm{B}}$ | $\hat{f}_{\mathrm{HT}}$ | $\hat{f}_{\mathrm{BT}}$ | $\hat{f}_{\mathcal{O}}$ | $\hat{f}_{\mathrm{B}}$ | $\hat{f}_{\mathrm{HT}}$ | $\hat{f}_{\mathrm{BT}}$ | $\hat{f}_{\mathcal{O}}$ |
|---|---|---|---|---|---|---|---|---|
| | Function: Step | | | | Function: Doppler | | | |
| MISE | 7.70 | 7.19 | 6.26 | 2.57 | 2.13 | 2.30 | 1.84 | 1.06 |
| False positives | 25.9 | 0.9 | 26.5 | 0.0 | 24.5 | 7.9 | 24.9 | 0.0 |
| False negatives | 15.7 | 22.3 | 12.1 | 0.0 | 11.3 | 20.7 | 10.1 | 0.0 |
| Size | 60.2 | 28.6 | 64.4 | 50.0 | 75.2 | 49.2 | 76.8 | 62.0 |
| | Function: Wave | | | | Function: Angles | | | |
| MISE | 1.01 | 2.62 | 1.03 | 0.77 | 1.31 | 1.83 | 1.33 | 0.77 |
| False positives | 5.3 | 4.5 | 9.8 | 0.0 | 4.8 | 1.2 | 5.9 | 0.0 |
| False negatives | 6.2 | 20.4 | 2.2 | 0.0 | 6.7 | 13.8 | 6.2 | 0.0 |
| Size | 53.1 | 38.0 | 61.6 | 53.0 | 33.0 | 22.4 | 34.7 | 35.0 |
| | Function: Blip | | | | Function: Parabolas | | | |
| MISE | 1.91 | 1.89 | 1.47 | 0.76 | 1.20 | 1.77 | 1.28 | 0.82 |
| False positives | 16.9 | 1.0 | 17.9 | 0.0 | 8.0 | 1.0 | 8.9 | 0.0 |
| False negatives | 6.6 | 13.0 | 5.1 | 0.0 | 2.0 | 7.2 | 1.7 | 0.0 |
| Size | 47.3 | 25.0 | 49.8 | 37.00 | 30.0 | 17.8 | 31.2 | 24.0 |
| | Function: Blocks | | | | Function: time.shift.sine | | | |
| MISE | 4.53 | 3.94 | 3.10 | 1.40 | 0.85 | 1.30 | 0.92 | 0.56 |
| False positives | 47.3 | 0.6 | 47.7 | 0.0 | 11.8 | 1.1 | 12.8 | 0.0 |
| False negatives | 59.3 | 80.5 | 49.4 | 0.0 | 0.2 | 5.4 | 0.2 | 0.0 |
| Size | 139.9 | 72.0 | 150.3 | 152.0 | 36.6 | 20.7 | 37.6 | 25.0 |
| | Function: Bumps | | | | Function: Spikes | | | |
| MISE | 1.36 | 1.44 | 1.14 | 0.56 | 0.54 | 0.84 | 0.54 | 0.34 |
| False positives | 104.4 | 2.2 | 104.6 | 0.0 | 17.6 | 1.6 | 18.1 | 0.0 |
| False negatives | 32.4 | 72.8 | 28.2 | 0.0 | 9.2 | 21.6 | 8.8 | 0.0 |
| Size | 241.0 | 98.4 | 245.4 | 169.0 | 75.5 | 47.0 | 76.3 | 66.0 |
| | Function: Heavisine | | | | Function: Corner | | | |
| MISE | 2.01 | 1.51 | 1.49 | 0.76 | 0.43 | 0.67 | 0.45 | 0.25 |
| False positives | 7.8 | 0.8 | 8.4 | 0.0 | 5.3 | 1.1 | 6.3 | 0.0 |
| False negatives | 13.0 | 15.6 | 10.6 | 0.0 | 3.9 | 7.4 | 3.7 | 0.0 |
| Size | 23.7 | 14.2 | 26.8 | 28.0 | 23.4 | 15.7 | 24.6 | 22.0 |

Note: $\hat{f}_{\mathrm{B}}, \hat{f}_{\mathrm{HT}}, \hat{f}_{\mathrm{BT}}$ and $\hat{f}_{\mathcal{O}}$ are, respectively, the Blockshrink, Hard Tree, Block Tree and Oracle estimators.



We use the connections between keep-or-kill estimation and hypothesis testing (see Abramovich, Benjamini, Donoho, and Johnstone 2006) in order to report in Table 1 the number of false positives/negatives (i.e. type I/II errors). This is obtained by comparing the set of indices of wavelet coefficients kept by each estimator with the set of indices kept by the keep-or-kill Oracle estimator

$$\hat{f}_{\mathcal{O}}(\cdot) = \hat{\alpha}\phi(\cdot) + \sum_{(j,k)\in\mathcal{S}^{\mathcal{O}}} \hat{\theta}_{jk}\psi_{jk}(\cdot),$$

where $\mathcal{S}^{\mathcal{O}} = \{(j,k); j \in \mathbb{N}, j < j_{\lambda_{\sigma/\sqrt{N}},p}; 0 \le k < 2^j; |\theta_{jk}| > \sigma/\sqrt{N}\}$.

When comparing the MISE results of the Blockshrink and of the Hard Tree estimators in Table 1, we understand that in practical situations we would not be able to decide which one to use. Indeed, according to the test function, it could be either the Blockshrink or the Hard Tree that performs the best. When not optimal, their MISE can be larger up to 33% (resp. 160%) compared with the other method. That is a potential huge loss for a practitioner who does not choose the method adapted to the target function we want to reconstruct. This observation is exactly what the maxiset approach suggests when the maxiset of these two methods are not nested. When looking at the results of the Block Tree estimator $\hat{f}_{\mathrm{BT}}$, it often provides the lowest MISE. If this is not the case, the deviation w.r.t. the MISE of the Blockshrink or of the Hard Tree does not pass over a reasonable 8%. There is no doubt that the Block Tree estimator is to be preferred

over the other two. Table 1 shows the impressive synergy when combining methods to increase the true discoveries at a comparatively low price in terms of false positives yielding these good performances of the Block Tree estimator.

*Remark 1* When comparing the behaviour of Blockshrink and Hard Tree estimators, they are quite sensitive to the choice of the wavelet family and regularity. Nevertheless, whatever the setting is, the Block Tree estimator remains the estimator to be preferred.

## Acknowledgements

The authors thank the associate editor and two anonymous referees for the helpful comments and suggestions which led to the considerable improvement of our article.

Financial support from the contract 'Projet d'Actions de Recherche Concertées' nr. 07/12/002 of the 'Communauté française de Belgique' granted by the 'Académie universitaire Louvain', the F+11/011 granted by the K.U. Leuven and the FWO G02470.12 are gratefully acknowledged.



## References

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2006), 'Adapting to Unknown Sparsity by Controlling the False Discovery Rate', *Annals of Statistics*, 34(2), 584–653.

Antoniadis, A., Bigot, J., and Sapatinas, T. (2001), 'Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study', *Journal of Statistical Software*, 6(6), 1–83.

Autin, F. (2004), 'Maxiset Point of View in Nonparametric Estimation', Ph.D. thesis, Université Paris 7 - Denis Diderot.

Autin, F. (2008a), 'On the Performances of a New Thresholding Procedure Using Tree Structure', *Electronic Journal of Statistics*, 2, 412–431.

Autin, F. (2008b), 'Maxisets for $\mu$-Thresholding Rules', *Test*, 17(2), 332–349.

Autin, F., Picard, D., and Rivoirard, V. (2006), 'Large Variance Gaussian Priors in Bayesian Nonparametric Estimation: A Maxiset Approach', *Mathematical Methods of Statistics*, 15(4), 349–373.

Autin, F., Le Pennec, E., Loubes, J.M., and Rivoirard, V. (2010), 'Maxisets for Model Selection', *Constructive Approximation*, 31(2), 195–229.

Autin, F., Freyermuth, J.-M., and von Sachs, R. (2011a), 'Ideal Denoising Within a Family of Tree-Structured Wavelet Estimators', *Electronic Journal of Statistics*, 5, 829–855.

Autin, F., Freyermuth, J.-M., and von Sachs, R. (2011b), 'Block-Threshold-Adapted Estimators via a Maxiset Approach'. Preprint 2011/17, ISBA, Université Catholique de Louvain.

Barber, S., and Nason, G.P. (2004), 'Real Nonparametric Regression Using Complex Wavelets', *Journal of the Royal Statistical Society Series B*, 66, 927–939.

Cai, T. (1997), 'On Adaptivity of Blockshrink Wavelet Estimator over Besov Spaces', Technical Report 97-05, Purdue University.

Cai, T. (1999), 'Adaptive Wavelet Estimation: A Block Thresholding and Oracle Inequality Approach', *Annals of Statistics*, 27(3), 898–924.

Cai, T. (2008), 'On Information Pooling, Adaptability and Superefficiency in Nonparametric Function Estimation', *Journal of Multivariate Analysis*, 99, 412–436.

Cai, T., and Silverman, B.W. (2001), 'Incorporating Information on Neighboring Coefficients into Wavelet Estimation', *Sankhya*, 63, 127–148.

Cai, T., and Zhou, H. (2009), 'A Data-Driven Block Thresholding Approach to Wavelet Estimation', *Annals of Statistics*, 37, 569–595.

Cohen, A., Dahmen, W., Daubechies, I., and DeVore, R. (2001a), 'Tree Approximation and Optimal Encoding', *Applied and Computational Harmonic Analysis*, 11(2), 192–226.

Cohen, A., De Vore, R., Kerkyacharian, G., and Picard, D. (2001b), 'Maximal Spaces with Given Rate of Convergence for Thresholding Algorithms', *Applied and Computational Harmonic Analysis*, 11, 167–191.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia, PA: SIAM.

Engel, J. (1994), 'A Simple Wavelet Approach to Nonparametric Regression from Recursive Partitioning Schemes', *Journal of Multivariate Analysis*, 49(2), 242–254.

Fryzlewicz, P. (2007), 'Bivariate Hard Thresholding in Wavelet Function Estimation', *Statistica Sinica*, 17, 1457–1481.

Gordon, R.D. (1941), 'Values of Mill's Ratio of Area to Bounding Ordinate of the Normal Probability Integral for Large Values of the Argument', *Annals of Mathematical Statistics*, 12, 364–366.

Hall, P., Kerkycharian, G., and Picard, D. (1998a), 'Block Threshold Rules for Curve Estimation Using Kernel and Wavelet Methods', *Annals of Statistics*, 26(3), 922–942.

Hall, P., Kerkycharian, G., and Picard, D. (1998b), 'On the Minimax Optimality for Block Thresholded Wavelet Estimators', *Statistica Sinica*, 9, 33–49.

Kerkyacharian, G., and Picard, D. (2000), 'Thresholding Algorithms, Maxisets and Well Concentrated Bases', *Test*, 9(2), 283–344.
Kerkyacharian, G., and Picard, D. (2002), 'Minimax or Maxisets'? *Bernoulli*, 8(2), 219–253.
Kohn, R., Marron, J.S., and Yau, P. (2000), 'Wavelet Estimation Using Bayesian Basis Selection and Basis Averaging', *Statistica Sinica*, 10, 109–128.
Lepski, O.V. (1991), 'Asymptotically Minimax Adaptive Estimation I: Upperbounds. Optimally Adaptive Estimates', *Theory of Probability and Its Applications*, 36, 682–697.
Vidakovic, B. (1999), *Statistical Modelling by Wavelets*, New York: John Wiley & Sons, Inc., 384 pp.

# A. Appendix

This section aims at proving the results provided in our study. In the sequel, $C$ denotes a generic constant which does not depend on $\epsilon$ and that may be different from one line to another.

## A.1. *A technical lemma and its proof*

We begin by introducing a technical lemma that will be useful later.

LEMMA A.1 *Let $s > 0$ and $m' \geq 1$. Consider a thresholding rule $\mu$ that satisfies the S-property. Then,*

$$f \in W_{\mu,m'}(2/(1+2s))$$
$$\Downarrow$$
$$\sup_{m\geq m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{2/(1+2s)} (\ln(\epsilon^{-1}))^{-1} \sum_{j<j_{mt_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\mu_{jk}(m,t_\epsilon,\theta) > \frac{mt_\epsilon}{2}\right\} < \infty,$$

*where $\theta = (\theta_{jk})_{j,k}$ is connected to $f$, thanks to Equation* (1).

*Proof* Fix $s > 0$, $m \geq m' \geq 1$ and $0 < \epsilon < \exp(-1)$. Let $\mu$ be a thresholding rule that satisfies the S-property and consider $f \in W_{\mu,m'}(2/(1+2s))$.

Because of the S-property and the monotonicity of the functions $\mu_{jk}$,

$$\begin{aligned}
&\sum_{j<j_{mt_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\mu_{jk}(m,t_\epsilon,\theta) > \frac{mt_\epsilon}{2}\right\} \\
&\leq C_\mu \ln(\epsilon^{-1}) \sum_{n\in\mathbb{N}} (m2^n t_\epsilon)^{-2} \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\{\mu_{jk}(m,t_\epsilon,\theta) \leq m2^n t_\epsilon\} \\
&\leq C_\mu \ln(\epsilon^{-1}) \sum_{n\in\mathbb{N}} (m2^n t_\epsilon)^{-2} \sum_{j\in\mathbb{N}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\{\mu_{jk}(m2^n,t_\epsilon,\theta) \leq m2^n t_\epsilon\} \\
&\leq C \ln(\epsilon^{-1}) \sum_{n\in\mathbb{N}} (m2^n t_\epsilon)^{-2} (m2^n t_\epsilon)^{2-2/(1+2s)} \\
&\leq C \ln(\epsilon^{-1}) (mt_\epsilon)^{-2/(1+2s)}.
\end{aligned}$$

Therefore,

$$\sup_{m\geq m'} \sup_{0<\epsilon<\exp(-1)} (mt_\epsilon)^{2/(1+2s)} (\ln(\epsilon^{-1}))^{-1} \sum_{j<j_{mt_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\mu_{jk}(m,t_\epsilon,\theta) > \frac{m}{2}t_\epsilon\right\} < \infty.$$

■

## A.2. *Proof of Theorem* 4.6

*Proof* ($\Longrightarrow$) Let a thresholding rule $\mu$ satisfy the LD- and the S-properties and $m' \geq m_{\mu,4}$. Suppose that there exists $C > 0$ such that $\mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 \leq C\,(mt_\epsilon)^{4s/(1+2s)}$, for any $m \geq 2m'$ and any $0 < \epsilon < \exp(-1)$.

Fix $m \geq 2m'$.

$$\begin{aligned}\sum_{j\geq j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2 &\leq \mathbb{E}\|\hat{f}_{\mu,m}-f\|_2^2\\ &\leq C(mt_\epsilon)^{4s/(1+2s)}\\ &\leq C2^{-2s/(1+2s)j_{mt_\epsilon}}.\end{aligned}$$

Using the continuity of $F$ in $\epsilon$, we deduce that $f \in \mathcal{B}_{2,\infty}^{s/1+2s}$. Moreover,

$$\left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\mu_{jk}\left(\frac{m}{2},t_\epsilon,\theta\right)\leq\frac{m}{2}t_\epsilon\right\} = A_1+A_2+A_3,$$

with

$$\begin{aligned}A_1 &= \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\mu_{jk}\left(\frac{m}{2},t_\epsilon,\theta\right)\leq\frac{m}{2}t_\epsilon\right\}\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta})\leq mt_\epsilon\}\right]\\ &\leq \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta})\leq mt_\epsilon\}\right]\\ &\leq \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\mathbb{E}\|\hat{f}_{\mu,m}-f\|_2^2\\ &\leq C,\\ A_2 &= \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\mu_{jk}\left(\frac{m}{2},t_\epsilon,\theta\right)\leq\frac{m}{2}t_\epsilon\right\}\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta})> mt_\epsilon\}\right]\\ &\leq \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{|\mu_{jk}(m,t_\epsilon,\hat{\theta})-\mu_{jk}(m,t_\epsilon,\theta)|>\frac{m}{2}t_\epsilon\right\}\right]\\ &= \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbb{P}\left(|\mu_{jk}(m,t_\epsilon,\hat{\theta})-\mu_{jk}(m,t_\epsilon,\theta)|>\frac{m}{2}t_\epsilon\right)\\ &\leq C(mt_\epsilon)^{-4s/(1+2s)}\epsilon^4\\ &\leq C.\end{aligned}$$

The last inequalities use the monotonicity of the functions $\mu_{jk}$ with respect to the first variable, the LD-property and the fact that $m \geq 2m_{\mu,4}$.

Now

$$\begin{aligned}A_3 &= \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\sum_{j\geq j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\mu_{jk}\left(\frac{m}{2},t_\epsilon,\theta\right)\leq\frac{m}{2}t_\epsilon\right\}\\ &\leq \left(\frac{mt_\epsilon}{2}\right)^{-4s/(1+2s)}\sum_{j\geq j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\\ &\leq C\,(mt_\epsilon)^{-4s/(1+2s)}\,2^{-2s/(1+2s)j_{mt_\epsilon}}\\ &\leq C.\end{aligned}$$

The last inequality holds since we have already proved that $f \in \mathcal{B}_{2,\infty}^{s/(1+2s)}$. When combining the bounds of $A_1, A_2$ and $A_3$ and when using the continuity of $F$ in $\epsilon$, one deduces that $f \in W_{\mu,m'}(2/(1+2s))$. Finally, one gets $f \in \mathcal{G}_{\mu,m'}$.

($\Longleftarrow$) Suppose that $f \in \mathcal{B}^{s/(1+2s)}_{2,\infty} \cap W_{\mu,m'}(2/(1+2s))$ with $m' \geq m_{\mu,4}$. For any $m \geq 2m'$ and any $0 < \epsilon < \exp(-1)$, the $L_2$-risk of the estimator $\hat{f}_{\mu,m}$ can be decomposed as follows:

$$\begin{aligned}\mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 &= \mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta}) \leq mt_\epsilon\}\right] \\ &\quad + \mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}(\hat{\theta}_{jk}-\theta_{jk})^2\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta}) > mt_\epsilon\}\right] \\ &\quad + \sum_{j\geq j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2 + \epsilon^2 \\ &= A_4 + A_5 + A_6.\end{aligned}$$

Since $f \in \mathcal{B}^{s/(1+2s)}_{2,\infty} \cap W_{\mu,m'}(2/(1+2s))$ and due to the LD-property

$$\begin{aligned}A_4 &= \mathbb{E}\left[\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta}) \leq mt_\epsilon\}\right] \\ &\leq \sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\{\mu_{jk}(2m,t_\epsilon,\theta) \leq 2mt_\epsilon\} \\ &\quad + \sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbb{P}(|\mu_{jk}(m,t_\epsilon,\hat{\theta}) - \mu_{jk}(m,t_\epsilon,\theta)| > mt_\epsilon) \\ &\leq C[(mt_\epsilon)^{4s/(1+2s)} + \epsilon^4] \\ &\leq C(mt_\epsilon)^{4s/(1+2s)}.\end{aligned}$$

Using the Cauchy–Schwarz inequality, the LD-property and Lemma A.1,

$$\begin{aligned}A_5 &= \sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\mathbb{E}[(\hat{\theta}_{jk}-\theta_{jk})^2\mathbf{1}\{\mu_{jk}(m,t_\epsilon,\hat{\theta}) > mt_\epsilon\}] \\ &\leq \sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\mathbb{E}\left[(\hat{\theta}_{jk}-\theta_{jk})^2\mathbf{1}\left\{\mu_{jk}(m,t_\epsilon,\theta) > \frac{m}{2}t_\epsilon\right\}\right] \\ &\quad + C\epsilon^2\sum_{j<j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\mathbb{P}^{1/2}\left(|\mu_{jk}(m,t_\epsilon,\hat{\theta}) - \mu_{jk}(m,t_\epsilon,\theta)| > \frac{m}{2}t_\epsilon\right) \\ &\leq C((mt_\epsilon)^{4s/(1+2s)} + \epsilon^2) \\ &\leq C(mt_\epsilon)^{4s/(1+2s)}.\end{aligned}$$

Since $f \in \mathcal{B}^{s/(1+2s)}_{2,\infty}$

$$\begin{aligned}A_6 &= \epsilon^2 + \sum_{j\geq j_{mt_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2 \\ &\leq \epsilon^2 + C2^{-2sj_{mt_\epsilon}/(1+2s)} \\ &\leq C(mt_\epsilon)^{4s/(1+2s)}.\end{aligned}$$

When combining the bounds of $A_4$, $A_5$ and $A_6$ and using the continuity of $F$ in $\epsilon$, one deduces that

$$\sup_{m\geq 2m'}\sup_{0<\epsilon<\exp(-1)}(mt_\epsilon)^{-4s/(1+2s)}\mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 < \infty.$$

This ends the proof. ■

## A.3. *Proof of Proposition* 5.1

*Proof* It is obvious that $\mu^{(3)}$ is a thresholding rule that generates $(\mu^{(3)}, m)$ thresholding estimators. Suppose that $\mu^{(1)}$ and $\mu^{(2)}$ satisfy the LD-property and consider for any $\nu > 0$, $m_{\mu^{(3)},\nu} = \max(m_{\mu^{(1)},\nu+1}, m_{\mu^{(2)},\nu+1})$. Then, for any $0 < \epsilon < \exp(-1)$ and any $m \geq m_{\mu^{(3)},\nu}$,

$$\begin{aligned}
&\mathbb{P}(|\mu_{jk}^{(3)}(m,t_\epsilon,\hat{\theta}) - \mu_{jk}^{(3)}(m,t_\epsilon,\theta)| > m_{\mu^{(3)},\nu}t_\epsilon) \\
&\quad\leq \sum_{i=1}^{2} \mathbb{P}(|\mu_{jk}^{(i)}(m,t_\epsilon,\hat{\theta}) - \mu_{jk}^{(i)}(m,t_\epsilon,\theta)| > m_{\mu^{(3)},\nu}t_\epsilon) \\
&\quad\leq \sum_{i=1}^{2} \mathbb{P}(|\mu_{jk}^{(i)}(m,t_\epsilon,\hat{\theta}) - \mu_{jk}^{(i)}(m,t_\epsilon,\theta)| > m_{\mu^{(i)},\nu+1}t_\epsilon) \\
&\quad\leq \epsilon^{\nu+1} \\
&\quad\leq \frac{\epsilon^{\nu}}{2}.
\end{aligned}$$

Hence, $\mu^{(3)}$ satisfies the LD-property too. ■

## A.4. *Proof of Corollary* 5.2

*Proof*

(a) It is a direct consequence of Proposition 5.1 and Theorem 4.6.
(b) This point becomes obvious when looking at the definition of spaces $W_{\mu^i,m'}(2/(1+2s))$ (with $i \in \{1,2,3\}$). Indeed, for any $m \geq m'$, any $0 < \epsilon < \exp(-1)$ and any sequence of real numbers $\theta$,

$$\mu_{j,k}^{(3)}(m,t_\epsilon,\theta) = \max(\mu_{j,k}^{(1)}(m,t_\epsilon,\theta), \mu_{j,k}^{(2)}(m,t_\epsilon,\theta)) \geq \mu_{j,k}^{(i)}(m,t_\epsilon,\theta), \quad \text{for } i \in \{1,2\}.$$

■

## A.5. *Proof of Theorem* 5.4

*Proof* Consider a cautious rule $\mu$ that satisfies the LD-property and $m' \geq m_{\mu,4}$. Assume that there exists $C > 0$ such that, for any $0 < \epsilon < \exp(-1)$ and any $m \geq 2m'$,

$$\mathbb{E}\|\hat{f}_{\mu,m} - f\|_2^2 \leq C(mt_\epsilon)^{4s/(1+2s)}.$$

Then,

$$\begin{aligned}
\sum_{j \geq j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 &\leq \mathbb{E}\|\hat{f}_{\mu,2m'} - f\|_2^2 \\
&\leq C(2m't_\epsilon)^{4s/(1+2s)} \\
&\leq C2^{-2s/(1+2s)j_{2m't_\epsilon}}.
\end{aligned}$$

Using the continuity of $F$ in $\epsilon$, one gets $f \in \mathcal{B}_{2,\infty}^{s/(1+2s)}$.

Let us now prove that $f$ necessarily belongs to $W_{\star,m'}(2/(1+2s))$, that is,

$$\sup_{0<\lambda<4m'\exp(-1)} \lambda^{-4s/(1+2s)} \left(F^{-1}\left(\frac{\lambda}{4m'}\right)\right)^2 \sum_{j \leq j_\lambda+1} \sum_{k=0}^{2^j-1} \mathbf{1}\{|\theta_{jk}| > \lambda\} < \infty.$$

When considering the change of variables $t_\epsilon = \lambda(4m')^{-1}$ for $0 < \lambda < 4m'\exp(-1)$, one aims at proving that

$$\sup_{0<\epsilon<\exp(-1)} \epsilon^2(m't_\epsilon)^{-4s/(1+2s)} \sum_{j < j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\{|\theta_{jk}| > 4m't_\epsilon\} < \infty.$$

Since $\mu$ is a cautious rule, for any $0 < \epsilon < \exp(-1)$,

$$\begin{aligned}\epsilon^2 \sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\{|\theta_{jk}| > 4m't_\epsilon\} &\le \epsilon^2 \sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \theta) > 4m't_\epsilon\} \\ &= \mathbb{E}\left[\sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \theta_{jk})^2 \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \theta) > 4m't_\epsilon\}\right] \\ &= B_1 + B_2,\end{aligned}$$

with

$$\begin{aligned}B_1 &= \mathbb{E}\left[\sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \theta_{jk})^2 \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \theta) > 4m't_\epsilon\} \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \hat{\theta}) > 2m't_\epsilon\}\right] \\ &\le \mathbb{E}\left[\sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \theta_{jk})^2 \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \hat{\theta}) > 2m't_\epsilon\}\right] \\ &\le \mathbb{E}\|\hat{f}_{\mu,2m'} - f\|_2^2 \\ &\le C(m't_\epsilon)^{4s/(1+2s)},\end{aligned}$$

and because of the LD-property and the Cauchy–Schwarz inequality

$$\begin{aligned}B_2 &= \mathbb{E}\left[\sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \theta_{jk})^2 \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \theta) > 4m't_\epsilon\} \mathbf{1}\{\mu_{jk}(2m', t_\epsilon, \hat{\theta}) \le 2m't_\epsilon\}\right] \\ &\le C\epsilon^2 \sum_{j<j_{2m't_\epsilon}} \sum_{k=0}^{2^j-1} \mathbb{P}^{1/2}(|\mu_{jk}(2m', t_\epsilon, \hat{\theta}) - \mu_{jk}(2m', t_\epsilon, \theta)| > 2m't_\epsilon\}) \\ &\le C\epsilon^2 \\ &\le C(m't_\epsilon)^{4s/(1+2s)}.\end{aligned}$$

The last inequality is obtained because of $m' \ge m_{\mu,4}$.

Combining $B_1$ and $B_2$ and still using the continuity of $F$ in $\epsilon$, one gets $f \in W_{\star,m'}(2/(1+2s))$. Hence, $f \in \mathcal{G}_{\star,m'}$. This ends the proof. ■