ESAIM: PS
January 2008, Vol. 12, p. 173–195
DOI: 10.1051/ps:2007047

ESAIM: Probability and Statistics
www.esaim-ps.org

RANDOM THRESHOLDS FOR LINEAR MODEL SELECTION

Marc Lavielle[1] and Carenne Ludeña[2]

Abstract. A method is introduced to select the significant or non null mean terms among a collection of independent random variables. As an application we consider the problem of recovering the significant coefficients in non ordered model selection. The method is based on a convenient random centering of the partial sums of the ordered observations. Based on L-statistics methods we show consistency of the proposed estimator. An extension to unknown parametric distributions is considered. Simulated examples are included to show the accuracy of the estimator. An example of signal denoising with wavelet thresholding is also discussed.

Mathematics Subject Classification. 62F12, 62G05, 62P99.

Received October 9, 2006.

1. Introduction

Consider the following model

$$y_i = \mu_i + \varepsilon_i,\ \ i = 1, \dots, n$$

where (μ_i) is a sequence of unknown constants some of which are zero and (ε_i) are centered, independent random variables with common cumulative distribution F_ε. The problem we study in this article is choosing the significant, non zero mean, coefficients based on the observations (y_i). Based on the data, it only seems reasonable to choose index i if

$$|y_i| > \tau, \tag{1}$$

for some appropriate threshold. Thus the problem is the choice of τ in (1), which will depend on the distribution of the sequence (ε_i).

In practice τ must be calibrated in terms of the data. A usual technique is to consider a sequence of thresholds (τ_j) for values ranging from very small (many significant coefficients) to very big (few significant coefficients) and study the point where a substantial decrease in the number of significant coefficients occurs. Of course choosing the "right" τ is equivalent to choosing the "right" number k of significant coefficients, with the advantage that this can be done independently of the choice of (τ_j) by looking at the relative size of the observations. Indeed, a jump in the relative size of the observations should indicate the existence of significant (not noise) coefficients.

This has been considered by a number of authors (see, for example, [12,13,17]) and many adaptive procedures aimed at studying the correct "jump point" have been developed.

Keywords and phrases. Adaptive estimation, linear model selection, hard thresholding, random thresholding, L-statistics.

[1] University Paris-Sud and University René Descartes, France; Marc.Lavielle@math.u-psud.fr

[2] IVIC, Venezuela; cludena@ivic.ve

One method that has proved to work remarkably well in practice in many settings including non ordered model selection for regression problems is to consider not only the individual observations, but instead the partial sums of the absolute (or squared) observations ordered decreasingly, and study the fluctuations of these partial sums around their expectation conditional to the total sum. Even when this conditional expectation cannot be computed in a closed form, an exponential change of variable makes this centering possible. The estimated "jump point" is obtained by minimization of a certain functional of the conditionally centered partial sums. Intuitively, each ordered observation from the null mean subset should account for a certain expected proportion of the observable total sum. If non null mean terms exist, this proportion changes and can thus be detected. Centering the partial sums by their non conditional expectations introduces an additional variance term.

Our main result, Theorem 2.2 proves the method will consistently select this "jump point" under mild assumptions over the gap between significant and non significant coefficients.

Our method requires previous knowledge of F_ε. In a parametric setting, $F_\varepsilon = F_\varepsilon(\,\cdot\,;\theta^\star)$, we show that the case $\theta^\star$ unknown can also be successfully considered introducing a consistent estimator of $\theta^\star$. When $\theta^\star$ is a scale parameter, an appropriate modification of the estimating procedure yields a scale free statistic, which is also shown to be consistent.

We then apply our method to the problem of estimating the significant coefficients for the regression problem

$$y_i = f(x_i) + \eta_i,\ i = 1, \ldots, n$$

where f is an unknown function in some function space S and η_i are independent random variables with variance σ^2. A usual estimation procedure is to consider $f \in L^2(\mu)$ and a finite orthonormal system $\{\phi_\lambda\}_\Lambda$, with $|\Lambda| = M_n$. Denoting by $\langle y, \phi_j \rangle_n = 1/n \sum_{i=1}^n y_i \phi_\lambda(x_i)$ the empirical coefficients, Donoho and Johnstone [10] in their seminal article proposed choosing only those coefficients whose absolute value exceeded a certain threshold $u = \sqrt{\frac{\tau\sigma^2 \log(n)}{n}}$. This procedure has since been refereed to as hard thresholding.

In a very interesting reinterpretation, Barron *et al.* [4] study the problem of hard thresholding in the context of non ordered model selection based on the addition of a penalization term. Their arguments are combinatorial based on the complexity of the underlying linear spaces: the size of the set of all possible models of size k out of K is bounded by $(eK/k)^k$ and a logarithmic factor depending on K must be introduced in order to bound the probabilities. In terms of (1) our observations would now be the empirical coefficients $\langle y, \phi_j \rangle_n$. Of course, except for the case $\eta_i \sim N(0, \sigma^2)$, the empirical coefficients will not be necessarily independent, although uncorrelated, so that the problem does not comply to our assumptions. However, in practice the method works well. As discussed in Section 6.1, our method can be interpreted as a random threshold procedure.

Although closely related to the problem of estimating the proportion of false null hypothesis [9,14], our goals are different. We are more interested in finding a consistent estimator for the jump point as well as convergence rates than in establishing an overall test for the proportion of false null hypothesis. In the latter case the main goal is establishing lower bounds that assure a specified confidence level. In terms of the proposed methodology, in our procedure we look at the fluctuations of the conditionally centered ordered data and not at the number of individual rejected tests. However, the connection between both approaches remains an interesting research topic.

The article is organized as follows: in Section 2 we introduce the problem and basic notation as well as the proposed test procedure. In Section 3 we state and prove theoretical results that justify our procedure, namely consistency of the selected subset of significant coefficients. In Section 4 we consider certain extensions which include the parametric distribution case, an application to the problem of non ordered linear model selection for the regression setting and interpret out testing scheme in terms of a random penalization procedure. In Section 5 we present simulated examples. An application to signal denoising with wavelet thresholding is proposed Section 6.

2. Describing the procedure

2.1. A first hypothesis testing procedure

Assume we observe $y_i = \mu_i + \varepsilon_i$. Variables ε_i are assumed to be centered, independent and identically distributed with common cumulative distribution F_ε. We begin by assuming that the cumulative distribution function $F_{|\varepsilon|}$ of the $|\varepsilon_i|$'s is known. In Section 4 we will deal with the unknown $F_{|\varepsilon|}$ case.

Given the collection $(y_i; 1 \leq i \leq n)$, we are interested in this section in testing if all the μ_i's are null or not. Thus, we introduce the following hypothesis:

Null hypothesis:

$\boldsymbol{H_0}$: $\mu_i \equiv 0$ for $i = 1, \ldots, n$.

Alternative hypothesis:

$\boldsymbol{H_1}$: there exists a non empty subset I of $\{1, 2, \ldots, n\}$ such that $\mu_i \neq 0$ for $i \in I$.

Then, the test procedure is defined as follows:

i) Order the observations $|y_{(1)}| \geq |y_{(2)}| \geq \ldots \geq |y_{(n)}|$.

ii) For $i = 1, \ldots, n$, let $X_{(i)} = -\log\left(1 - F_{|\varepsilon|}(|y_{(i)}|)\right)$.

iii) Let $T_j = \sum_{i=1}^{j} X_{(i)}$ and $Q_j = \mathbb{E}_{H_0}(T_j | T_n)$.

iv) Define the test statistic $D_n = \max_j |T_j - Q_j| / \sqrt{n}$. We will reject the null hypothesis if $D_n > d_\alpha$, where d_α is defined in Section 3.

Remark 1. Under the null hypothesis, the sequence $(X_{(i)})$ is a decreasing sequence of exponential random variables with parameter 1. Then, the conditional expectation $\mathbb{E}_{H_0}(T_j | T_n)$ can easily be computed using the following proposition (the proof is given in the Appendix):

Proposition 2.1. *Assume $X_{(1)}$, $X_{(2)}$, ... , $X_{(n)}$ is an ordered sequence of Exp(1) random variables, with $X_{(1)} \geq X_{(2)} \geq \ldots X_{(n)}$. For any $1 \leq j \leq n$, let $T_j = \sum_{i=1}^{j} X_{(i)}$. Then, for any $j \leq K \leq n$,*

$$\mathbb{E}\left(X_{(i)}\right) = \sum_{\ell=1}^{n} \frac{1}{\ell} \quad (2)$$

$$\mathbb{E}(T_j) = j + j \sum_{i=j+1}^{n} \frac{1}{i} \quad (3)$$

$$\mathbb{E}(T_j | T_K) = \frac{\mathbb{E}(T_j)}{\mathbb{E}(T_K)} T_K. \quad (4)$$

Remark 2. The distribution of the test statistic D_n cannot be computed in a closed form. Nevertheless, the following standard result will allow us to construct probability tables (the proof is given in the Appendix):

Theorem 2.1. *Assume $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ is an ordered sequence of Exp(1) random variables, with $X_{(1)} \geq X_{(2)} \geq \ldots X_{(n)}$. For any $1 \leq j \leq n$, let $T_j = \sum_{i=1}^{j} X_{(i)}$. Introduce for $t \in [0,1]$ the random process $d_n(t) = T_{[nt]} - \mathbb{E}\left(T_{[nt]} | T_n\right)$. Then, $\frac{1}{\sqrt{n}} d_n(t)$, as a stochastic process indexed on $t \in [0,1]$, converges in distribution to a zero mean Gaussian process Δ with covariance function defined by*

$$\begin{aligned}\mathbb{E}(\Delta(t)\Delta(s)) = & \int_0^1 \int_0^1 [(1-u) \wedge (1-v) - (1-u)(1-v)][\mathbb{1}_{[0,t]}(u) - t + t\log(t)] \\ & \times [\mathbb{1}_{[0,s]}(v) - s + s\log(s)] \mathrm{d}G^{-1}(u) \mathrm{d}G^{-1}(v),\end{aligned}$$

where $G(x)$ is the distribution function of an exponential r.v.

Using Theorem 2.1, we can conclude that statistic D_n defined in the test procedure converges weakly to $\Delta_\infty = \sup_t \Delta(t)$. Then, d_α is defined as the α quantile of Δ_∞.

Remark 3. Instead of assuming that the distribution of the $|\varepsilon_i|$ is known, we can assume that there exists an increasing continuous function $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that the cumulative distribution function F_h of $h(|\varepsilon|)$ is known. Then, $X_{(i)}$ is defined as $-\log\left(1 - F_h(h(|y_{(i)}|))\right)$. Without any loss of generality, we will consider the case $h = id$ in the following.

Remark 4. A uniform change of variable can also be used, by setting $X_{(i)} = F_{|\varepsilon|}(|y_{(i)}|)$. Indeed, the conditional expectation of T_j can also be computed here:

$$\mathbb{E}\left(T_j|T_K\right) = \frac{j(K-j)}{K+1} + \frac{j(j+1)}{K(K+1)}T_K.$$

2.2. Choosing the right coefficients

If we reject the null hypothesis, the next step is to select the significant coefficients. For this we define the following procedure:

i) For $i = 1, \ldots, n$, let $X_{(i)} = -\log\left(1 - F_{|\varepsilon|}(|y_{(i)}|)\right)$.

ii) Let K_n be some positive integer. For $1 \le k \le n - K_n$ and $1 \le j \le K_n$, set

$$B_{k,j,n} = \frac{j\left(1 + \sum_{i=j+1}^{n-k} 1/i\right)}{K_n\left(1 + \sum_{i=K_n+1}^{n-k} 1/i\right)} \tag{5}$$

and compute

$$\begin{aligned} T_{k,j} &= \sum_{i=k+1}^{k+j} X_{(i)}, & (6)\\ Q_{k,j} &= B_{k,j,n}\, T_{k,K_n}, & (7)\\ \eta_k &= \max_{1\le j\le K_n} \frac{|T_{k,j} - Q_{k,j}|}{\sqrt{n}}. & (8)\end{aligned}$$

iii) Let

$$\hat{k} = \mathrm{Arg}\min_{1\le k\le n-K_n} \eta_k.$$

Remark 2.1. The ℓ_1 or the ℓ_2 norms can be used instead of the ℓ_∞ norm to define η by setting

$$\eta_k = n^{-\frac{3}{2}} \sum_{j=1}^{K_n} |T_{k,j} - Q_{k,j}|$$

or

$$\eta_k = n^{-2} \sum_{j=1}^{K_n} (T_{k,j} - Q_{k,j})^2.$$

In the spirit of Section 2.1, it is possible to reinterpret the above procedure in a data-dependent "Hypothesis testing" setting. Let s be the (data-dependent) one to one mapping from $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, n\}$ defined by $Y_{s(i)} = Y_{(i)}$ (recall that $(Y_{(i)})$ is a decreasing sequence).

Then, define the set of alternative hypotheses:

Alternative hypothesis:

$\boldsymbol{H_1(k)}$: there exists a subset $I_k \subset \{1, 2, \ldots, n\}$, with $|I_k| = k$, such that,

– for any $i \in I_k$, $s(i) \le k$ and $\mathbb{E}_{H_1(k)}(Y_i) \ne 0$,

– for any $i \notin I_k$, $s(i) \ge k + 1$ and $\mathbb{E}_{H_1(k)}(Y_i) = 0$.

Under $\boldsymbol{H_1(k)}$, there are k significant coefficients and $|y_{(k+1)}|, \ldots, |y_{(n)}|$ have distribution $F_{|\varepsilon|}$. We will denote $\mathbb{E}_k\left(\right)$ the expectation under $\boldsymbol{H_1(k)}$ (instead of $\mathbb{E}_{H_1(k)}\left(\right)$). That is,

$$\mathbb{E}_k\left(T_{k,j}\right) := j\left(1 + \sum_{i=j+1}^{n-k} 1/i\right). \tag{9}$$

So that,

$$B_{k,j,n} = \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}$$

and $Q_{k,j}$ can be thought of as $Q_{k,j} = \mathbb{E}_k\left(T_{k,j}|T_{k,K_n}\right)$ using the results of the previous section and Proposition 2.1.

Our main consistency result requires bounding deviations of the $k-$th order statistic, among n observations, for $k = n$ and $k = n^d$, for some $d < 1/2$. For this, two general results are required. The first concerning the asymptotic behavior of the maximum of n independent variables, and the second related to the limiting Gaussian behavior of intermediate order statistics. In order to unify notation and simplify the presentation both results will be given under the assumption that there exist $w = w(F_\varepsilon)$ and $x_0 < w$ such that the distribution of the errors, F_ε has a differentiable density f_ε over (x_0, w). We also assume f_ε satisfies one of the Von-Misses type conditions we give below.

Condition **VM**.

1. We assume $w = \infty$ that f_ε is positive near infinity and that there exists $\alpha > 0$ such that

$$\lim_{x\to\infty} \frac{x f_\varepsilon(x)}{[1 - F_\varepsilon(x)]} = \alpha.$$

2. We assume $w < \infty$, f_ε is positive near infinity and that there exists $\alpha > 0$ such that

$$\lim_{x\uparrow w} \frac{(w - x) f_\varepsilon(x)}{1 - F_\varepsilon(x)} = \alpha.$$

3. We assume that $f_\varepsilon(x)$ is positive and that

$$\lim_{x\uparrow w} \frac{f(x) \int_x^w (1 - F_\varepsilon(t))\mathrm{d}t}{[1 - F_\varepsilon(x)]^2} = 1,$$

for $x \in (x_0, w)$.

We then have the following result

Proposition 2.2.

1. *(Convergence of extremes, de Haan* [8], *Ths. 2.7.1, 2.7.2, 2.7.3.) Assume one of the* **VM** *conditions hold. Then, $F_\varepsilon(a_n + b_n x)^n \to G$, for $G = G(i, \theta)$ the generalized extreme value function, and some choice of constants a_n, b_n. The parameters of the GEV function G depend on the* **VM** *condition that holds.*
2. *(Convergence of intermediate extreme values, Falk* [11], *Th.2.1.) Assume one of the* **VM** *conditions hold. Assume $k \to \infty$, $k/n \to 0$ and set*

$$a_{n,k} = F_\varepsilon^{-1}(1 - k/n); \quad b_{n,k} = k^{1/2}/(n f_\varepsilon(a_{n,k})).$$

Then, if $|\varepsilon|_{(k,n)}$ is the $n - k$ order statistic

$$P(|\varepsilon|_{(k,n)} \leq d_n + c_n x) \to \Phi(x),$$

where Φ stands for the standard normal distribution function and c_n, d_n are any sequences that satisfy

$$\lim_n c_n/b_{n,k} = 1 \text{ and } \lim_n (d_n - a_{n,k})/b_{n,k} = 0.$$

We now state our asymptotic framework:

AF1. There exists $t^\star \in (0,1)$ and a subset $I_{k_n^\star}$ of $\{1,2,\ldots,n\}$, with $k_n^\star = [t^\star n]$ and $|I_{k_n^\star}| = k_n^\star$, such that $\mu_i \neq 0$ if $i \in I_{k_n^\star}$. For all other index, $\mu_i = 0$.

AF2. For any $i \in I_{k_n^\star}$, $|\mu_i| \geq \alpha_n$, where $\alpha_n \to \infty$ is such that

$$\alpha_n = 2a_n + r_n b_n \tag{10}$$

for $r_n \to \infty$.

AF3. $K_n/n \to c$ such that $0 < c < 1 - t^\star$.

We have the following result

Theorem 2.2. *Let (u_n) be any positive and decreasing sequence such that $\sqrt{n}\, u_n \to \infty$. Then, under the asymptotic framework defined by* **VM, AF1, AF2, AF3***,*

$$P\left(\left|\frac{\hat{k}}{n} - t^\star\right| > u_n\right) \to 0. \tag{11}$$

Moreover, for $a > 0$ there exist constants c_1, c_2 which depend on a such that if

$$u_n = \frac{c_1 \alpha_n \sqrt{\log n}}{2\sqrt{n}} + \frac{c_2 \alpha_n \log(n)}{2n},$$

then

$$\mathbb{P}_{H_1(k_n^\star)}\left(\left|\frac{\hat{k}}{n} - t^\star\right| > u_n\right) \leq 2\mathrm{e}^{-a\log(n)} + 2\mathbb{P}\left(\max_{1\leq i\leq n} |\varepsilon_i| > \alpha_n\right). \tag{12}$$

The proof of Theorem 2.2 is given in Section 3.

Remark 2.2. No overlapping is ensured with high probability when $n \to \infty$ under Condition **AF2**. This condition can be weakened assuming that the minimum gap between both subsets is such that the overlap between both ordered sequences is not any larger than n^d, for some $d < 1/2$. In this case we would require the alternative condition

AF2'. For any $i \in I_{k_n^\star}$, $|\mu_i| \geq \alpha_n$, where $\alpha_n \to \infty$ is such that

$$\alpha_n = a_n + r_{1,n} b_n + a_{n^d, n-k_n^\star} + r_{2,n} b_{n^d, n-k_n^\star},$$

for $r_{i,n} \to \infty$ for $i = 1,2$ and some $d < 1/2$.

Thus the condition over α_n in [**AF2'**] yields a slight improvement with respect to [**AF2**] in terms of the constant since $a_{n^d,n-k_n^\star} + r_{2,n} b_{n^d,n-k_n^\star}$ is smaller than $2a_n + r_{1,n} b_n$. However it is not possible to have $\alpha_n \leq a_n + r_{1,n} b_n + a_{n^{1/2},n-n^{1/2}}$ if rates as in Theorem 2.2 are sought. This fact yields an important insight as to the minimal signal to noise ratio that should be required.

3. Proof of Theorem 2.2

Our procedure is based on two facts: a) under our assumptions over the error distribution, if the null hypothesis is rejected, that is, if there is a group of significant coefficients and one of non significant coefficients, both groups of observations will not mix with high probability under assumption **AF2**) and b) if both groups are separated at index $k_n^\star$, then $T_{k_n,j} - Q_{k_n,j}$ will only converge at rate $\sqrt{n}$ for index k_n such that $|k_n - k_n^\star| = o(\sqrt{n})$.

Set $u_i = y_i$ for $i \in I_{k_n^\star}$ and $v_i = y_i$ for $i \notin I_{k_n^\star}$. Thus (v_i) is an i.i.d. sequence with distribution F_ε.

We have the following lemma that assures that both collections are stochastically in order with high probability:

Lemma 3.1. *Let $(u_{(i)})$ and $(v_{(i)})$ be the sequences $(|u_i|)$ and $(|v_i|)$ in a decreasing order. Then*

$$\mathbb{P}\big(v_{(1)} > u_{(k_n^\star)}\big) \to 0$$

and

$$\mathbb{P}\Big(v_{(1)} > \frac{\alpha_n}{2}\Big) \to 0.$$

Proof. Let $(\tilde{v}_i)$ be a sequence of i.i.d. r.v. with distribution F_ε. Then,

$$\begin{aligned}\mathbb{P}\big(v_{(1)} > u_{(k_n^\star)}\big) &\le \mathbb{P}\big(v_{(1)} + \tilde{v}_{(1)} > \alpha_n\big)\\ &\le \mathbb{P}\big(v_{(1)} > \alpha_n/2\big) + \mathbb{P}\big(\tilde{v}_{(1)} > \alpha_n/2\big)\\ &\le 2\mathbb{P}\big(v_{(1)} > \alpha_n/2\big) \to 2W\left(\frac{\alpha_n/2 - a_n}{b_n}\right) \to 0. \qquad \square\end{aligned}$$

Lemma 3.1 yields the $(u_{(i)})$ and $(v_{(i)})$ are stochastically in order with high probability. Let Ω_n be the subset of Ω where $v_{(1)} < \alpha_n/2$ and $u_{(k_n^\star)} > \alpha_n/2$. Clearly $P(\Omega_n) \to 1$. In what follows we will restrict our proof to Ω_n.

Let $\mathbb{E}_k\left(T_{k,j}\right)$ be as defined in equation (9). Also let $a_i = \mathbb{E}_0\left(X_{(i)}\right) = \sum_{\ell=1}^n 1/\ell$.

1) Consider first the case $k > k_n^\star$. On Ω_n,

$$\begin{aligned}T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n}\, T_{k,K_n}\\ &= \big(T_{k,j} - \mathbb{E}_{k_n^\star}\left(T_{k,j}\right)\big) - B_{k,j,n}\left(T_{k,K_n} - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)\right)\\ &\quad + \mathbb{E}_{k_n^\star}\left(T_{k,j}\right) - B_{k,j}\mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)\\ &= R_{k,j} + S_{k,j}.\end{aligned}$$

We have decomposed the statistics $T_{k,j} - Q_{k,j}$ into a random part $R_{k,j}$ and a deterministic part $S_{k,j}$. Remark over Ω_n, $R_{k,j}$ is a function of $v_{(k)}, \ldots, v_{(K_n)}$. First, let $k = [tn]$ and $j = [sn]$ for $t \le s$. As in Theorem 2.1, $R_{k,j}\mathbb{I}_{\Omega_n}$ (normalized by $\sqrt{n}$) as a process indexed by $(t,s) \in (0,1)^2$ converges in distribution to a zero-mean Gaussian process $\Gamma_{t,s} = (1-t^\star)[\Delta(s-t^\star) - \Delta(t-t^\star)]$.

On the other hand,

$$\begin{aligned}S_{k,j} &= \mathbb{E}_{k_n^\star}\left(T_{k,j}\right) - \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)\\ &= \mathbb{E}_{k_n^\star}\left(T_{k,j}\right) - \mathbb{E}_k\left(T_{k,j}\right) - \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\big(\mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right) - \mathbb{E}_k\left(T_{k,K_n}\right)\big)\\ &= \sum_{i=j+1}^{j+k-k_n^\star} a_i - \sum_{i=1}^{k-k_n^\star} a_i + B_{k,j,n}\left(\sum_{i=K_n+1}^{K_n+k-k_n^\star} a_i - \sum_{i=1}^{k-k_n^\star} a_i\right)\\ &= \sum_{i=1}^{k-k_n^\star}\big(a_{i+j} - a_i + B_{k,j,n}(a_{i+K_n} - a_i)\big).\end{aligned}$$

Thus, there exists a constant, $\gamma > 0$, which depends on c in **[AF3]**, such that $\sup_j |S_{k,j}| \geq \gamma(k - k_n^\star)$ and

$$\begin{aligned}
\mathbb{P}_{k_n^\star}\left(k_n^\star - \widehat{k} > n\, u_n\right) &\leq \mathbb{P}\left(\eta_{k_n^\star} > \sup \eta_k \ , \ (k - k_n^\star) > n\, u_n\right) \qquad (13)\\
&\leq \mathbb{P}\left(2 \sup_k \sup_j R_{k,j} > \inf_k \sup_j |S_{k,j}| \ , \ (k - k_n^\star) > n\, u_n\right) + \mathbb{P}(\Omega_n^c)\\
&\leq \mathbb{P}\left(2 \sup_k \sup_j R_{k,j} > \gamma\, n\, u_n\right) + \mathbb{P}(\Omega_n^c)\,.
\end{aligned}$$

Because of the weak convergence of $R_{k,j}\, \mathbb{1}_{\Omega_n}$ the above probability tends to zero when n goes to infinity.

2) Consider now the case $k < k_n^\star$. On Ω_n,

$$\begin{aligned}
T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n}\, T_{k,K_n}\\
&= (1 - B_{k,j,n}) T_{k,k_n^\star - k} + \left(T_{k_n^\star,j} - \mathbb{E}\left(T_{k_n^\star,j}\right)\right) - B_{k,j,n}\left(T_{k_n^\star,K_n} - \mathbb{E}\left(T_{k_n^\star,K_n}\right)\right)\\
&\quad + \mathbb{E}\left(T_{k_n^\star,j}\right) - B_{k,j,n} \mathbb{E}\left(T_{k_n^\star,K_n}\right)\\
&= A_{k,j} + R_{k_n^\star,j} + U_{k,j}
\end{aligned}$$

where $A_{k,j} = (1 - B_{k,j,n}) T_{k,k_n^\star - k}$. Observe that over Ω_n, $|y_{(i)}| > \alpha_n/2$. Then, there exists $c(\alpha_n) > 0$ such that $T_{k,k_n^\star - k} > c(\alpha_n)(k_n^\star - k)$, thus $A_{k,j} = \mathcal{O}(k_n^\star - k)$. On the other hand, $R_{k_n^\star,j}\mathbb{1}_{\Omega_n}$ converges in distribution to a zero-mean Gaussian process $\Gamma_{t^\star,s}$. Consider now the bias term $U_{k,j}$:

$$\begin{aligned}
U_{k,j} &= \mathbb{E}_{k_n^\star}\left(T_{k_n^\star,j}\right) - \frac{\mathbb{E}_k\left(T_{k,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} \mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right)\\
&= \mathbb{E}_{k_n^\star}\left(T_{k_n^\star,j}\right) - \mathbb{E}_k\left(T_{k_n^\star,j}\right) - \frac{\mathbb{E}_k\left(T_{k_n^\star,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)}\left(\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right) - \mathbb{E}_k\left(T_{k,K_n}\right)\right)\\
&\quad + \frac{\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} \mathbb{E}_k\left(T_{k,k_n^\star}\right)\\
&= \sum_{i=j+k-k_n^\star+1}^{j-k} a_i - \sum_{i=1}^{k_n^\star - k} a_i + \frac{\mathbb{E}_k\left(T_{k_n^\star,j}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} \sum_{i=K_n+k-k_n^\star+1}^{K_n} a_i - \frac{\mathbb{E}_{k_n^\star}\left(T_{k_n^\star,K_n}\right)}{\mathbb{E}_k\left(T_{k,K_n}\right)} \sum_{i=1}^{k_n^\star - k} a_i.
\end{aligned}$$

Thus, there exists a constant $\delta > 0$, which depends on c in **[AF3]**, such that $\sup_j |U_{k,j}| \geq \delta(k - k_n^\star)$ and $\mathbb{P}_{k_n^\star}\left(\widehat{k} - k_n^\star > n\, u_n\right) \to 0$ when $n \to \infty$. The latter together with (13) shows (11).

In order to show (12), sharper bounds on $\mathbb{P}_{k_n^\star}\left(\sup_k \sup_j R_{k,j} > C\, n\, u_n\right)$ are required for any given constant C. As above, we will restrict our attention to the set $\Omega_n \cap \Omega_n'$ and drop this fact from the notation. Consider first as above the case $k > k_n^\star$. Write,

$$\begin{aligned}
R_{k,j} &= \left(T_{k,j} - \mathbb{E}_{k_n^\star}\left(T_{k,j}\right)\right) - B_{k,j,n}\left(T_{k,K_n} - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)\right)\\
&= R_{k,j}^{(1)} + R_{k,j}^{(2)}.
\end{aligned}$$

Since $\sup_j B_{k,j,n} = 1$, we have that $\sup_j |R_{k,j}^{(2)}| = |T_{k,K_n} - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)|$.

Let G denote the common distribution function of the collection (X_i). We can rewrite

$$T_{k,K_n} - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right) = \sum_i [X_i - \mathbb{E}_{k_n^\star}\left(X_i\right)] \mathbb{1}_{\{G^{-1}(1-K_n/n) < X_i\}}.$$

Thus, recalling $w_n = \alpha_n - (a_n - r_{1,n} b_n)$,

$$\frac{T_{k,K_n} - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)}{w_n}$$

is the sum of independent bounded r.v. with variance bounded by 1, so that by Bennet's inequality

$$\begin{aligned}\mathbb{P}\left(\frac{\sup_j |R^{(2)}_{k,j}|}{w_n} > \frac{\gamma/2nu_n}{w_n}\right) &\leq \mathbb{P}\left(\frac{|T_{k,K_n} - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}\right)|}{w_n} > \frac{c_1\gamma}{4\sqrt{2}}\sqrt{2n\log n} + \frac{3c_2\gamma}{4}\frac{\log n}{3}\right)\\ &\leq \mathrm{e}^{-(a+1)\log(n)},\end{aligned}$$

choosing $c_2 \geq \frac{4(a+1)}{3\gamma}$ and $c_1 \geq 4\frac{\sqrt{2}\sqrt{a+1}}{\gamma}$.

Hence, adding in k

$$\mathbb{P}\left(\sup_k \sup_{j<k+K_n} R^{(2)}_{k,j} > \frac{\gamma}{2nu_n}\right) \leq \mathrm{e}^{-a\log n}.$$

For $R^{(1)}_{k,j}$ we have

$$\frac{T_{k,j} - \mathbb{E}_{k_n^\star}\left(T_{k,j}\right)}{w_n} = \sum_i \frac{[X_i - \mathbb{E}_{k_n^\star}\left(X_i\right)]\mathbb{1}_{\{G^{-1}(1-j/n)<X_i\}}}{w_n}.$$

Hence in this case we must use a functional version of Bennet's inequality (Th. 7.3 in [6]) which yields

$$\mathbb{P}\left(\frac{\sup_j |R^{(2)}_{k,j}|}{w_n} > \mathbb{E}\left(\frac{\sup_j |R^{(2)}_{k,j}|}{w_n}\right) + \sqrt{2xv} + \frac{x}{3}\right) \leq \mathrm{e}^{-x},$$

for $v \geq n + 2\mathbb{E}\left(\frac{\sup_j |R^{(2)}_{k,j}|}{w_n}\right)$. Thus it remains to bound $A = \mathbb{E}\left(\frac{\sup_j |R^{(2)}_{k,j}|}{w_n}\right)$. This can be done using standard symmetrization and entropy techniques to obtain, $A \leq 4\sqrt{n\log n}$, as the random entropy of the class $\mathcal{A} = \{\mathbb{1}_{G^{-1}(1-t)}, t \in [0,1]\}$ (as it is a collection of increasing functions) is bounded by $2\log n$.

As above,

$$\begin{aligned}\mathbb{P}\left(\frac{\sup_j |R^{(1)}_{k,j}|}{w_n} > \frac{\gamma/2nu_n}{w_n}\right) &\leq \mathbb{P}\Bigg(\frac{|T_{k,j} - \mathbb{E}_{k_n^\star}\left(T_{k,j}\right)|}{w_n} > 4\sqrt{n\log n}\\ &\qquad + \sqrt{2(a+1)\log n(n + 4\sqrt{n\log n})} + (a+1)\frac{\log n}{3}\Bigg)\\ &\leq \mathrm{e}^{-(a+1)\log(n)},\end{aligned}$$

choosing $c_i, i = 1, 2$ appropriately.

The case $k < k_n^\star$ follows analogously.

4. Some extensions

4.1. **Unknown distribution**

Assume now that the distribution F_ε of the ε_i's is a parametric distribution $F_\varepsilon(\cdot\ ;\ \theta^\star)$, but where the parameter $\theta^\star$ is unknown. For any $0 \leq k \leq n-1$, let $\widehat{\theta}_k = \widehat{\theta}(y_{(k+1)}, y_{(k+2)}, \ldots, y_{(n)})$ be an estimator of θ. Let $F_{|\varepsilon|}(\cdot\ ;\ \theta^\star)$ be the distribution of the $|\varepsilon_i|$'s. We will consider the following assumptions:

F1. The cumulative distribution function $F_{|\varepsilon|}$ is two times differentiable as a function of θ with *a.e.* strictly positive derivative at $\theta = \theta^\star$.

F2. $\theta^\star$ belongs to some compact set Θ and there exists, under $H_{k_n^\star}$, a consistent estimator $\widehat{\theta}_{k_n^\star} = \widehat{\theta}(y_{(k_n^\star+1)}, y_{(k_n^\star+2)}, \ldots, y_{(n)})$ of $\theta^\star$.

F3. There exists (a, b) such that $0 < a < t^\star < b < 1$ and a Lipschitz continuous function $\tilde{\theta}$ defined on $[a, b]$ such that, under $H_{k_n^\star}$, $(\widehat{\theta}_{[tn]})$ converges uniformly on $[a, b]$ in probability to $(\tilde{\theta}(t))$.

Remark 1. Under hypothesis **F2** and **F3**, $\widehat{\theta}_{k_n^\star}$ is a consistent estimator of $\tilde{\theta}(t^\star) = \theta^\star$.

Remark 2. When $t < t^\star$, convergence of $\widehat{\theta}_{[tn]}$ can be difficult to check with any estimator, since $\widehat{\theta}_{[tn]}$ depends on some y_i's that are not distributed under distribution F_ε. Nevertheless, it is possible to use an estimator based on some empirical quantiles and that only depends on the smallest observations, that is, that depends only on the observations distributed under F_ε.

For any $\theta \in \Theta$, let $X_i(\theta) = -\log\left(1 - F_{|\varepsilon|}(|y_i|, \theta)\right)$, and $T_{k,j}(\theta) = \sum_{i=k+1}^{k+j} X_{(i)}(\theta)$.
Then, we define the following procedure:

i) Let $K_n \leq [(1-b)\ n]$ be some positive integer. For $[a\ n] \leq k \leq n - K_n$;

1. let $\widehat{\theta}_k = \widehat{\theta}(y_{(k+1)}, y_{(k+2)}, \ldots, y_{(n)})$;
2. for $i = 1, \ldots, n$, let $X_{(i)}(\widehat{\theta}_k) = -\log\left(1 - F_{|\varepsilon|}(|y_{(i)}|; \widehat{\theta}_k)\right)$;
3. for $1 \leq j \leq K_n$, compute

$$\begin{aligned} T_{k,j}(\widehat{\theta}_k) &= \sum_{i=k+1}^{k+j} X_{(i)}(\widehat{\theta}_k), \\ Q_{k,j}(\widehat{\theta}_k) &= B_{k,j,n}\, T_{k,K_n}(\widehat{\theta}_k), \\ \eta_k(\widehat{\theta}_k) &= \max_{k+1 \leq j \leq n} \frac{|T_{k,j}(\widehat{\theta}_k) - Q_{k,j}(\widehat{\theta}_k)|}{\sqrt{n}}; \end{aligned}$$

ii) let

$$\hat{k} = \text{Arg} \min_{an \leq k \leq bn} \eta_k(\widehat{\theta}_k).$$

Remark. Here, $Q_{k,j}(\widehat{\theta}_k) = B_{k,j,n}\, T_{k,K_n}(\widehat{\theta}_k)$ is the conditional expectation of $T_{k,j}$, conditionally to T_{k,K_n}, assuming that $k_n^\star = k$ and that $\theta^\star = \widehat{\theta}_k$.

Then, we have the following result,

Theorem 4.1. *Assume* **F1, F2, F3**.

i) Introduce for $t \in [0,1]$ and $s \in [a,b]$ the random process

$$\hat{d}_n(t,s) = T_{k_n^\star,[K_n t]}(\widehat{\theta}_{[ns]}) - \mathbb{E}_{H_{k_n^\star}}\left(T_{k_n^\star,[K_n t]}(\widehat{\theta}_{[ns]}) | T_{k_n^\star,K_n}(\widehat{\theta}_{[ns]})\right).$$

Then, $\hat{d}_n(t,s)/\sqrt{n}$, as a stochastic process indexed on $[0,1] \times [a,b]$, converges in distribution, under $\boldsymbol{H_{k_n^\star}}$, to a zero mean Gaussian process $(\Lambda(t,s))$.

ii) Let (u_n) be any positive and decreasing sequence such that $\sqrt{n}\, u_n \to \infty$. Then, under the asymptotic framework defined by **AF1, AF2, AF3***,*

$$\mathbb{P}_{H_1(k_n^\star)}\left(\left|\frac{\hat{k}}{n} - t^\star\right| > u_n\right) \to 0. \tag{14}$$

Proof. We first show *i)*.

With the above notation, for any $\theta \in \Theta$, let

$$\begin{aligned}\Psi_j(\theta) &= T_{k_n^\star,j}(\theta) - Q_{k_n^\star,j}(\theta)\\ \Psi'_j(\theta) &= \frac{\partial \Psi_j}{\partial\theta}(\theta).\end{aligned}$$

For $t \in [0,1]$ and $s \in [a,b]$, let

$$\begin{aligned}\hat{d}_n(t,s) &= \Psi_{[nt]}(\widehat{\theta}_{[ns]})\\ &= \Psi_{[nt]}(\tilde{\theta}(s)) + (\widehat{\theta}_{[ns]} - \tilde{\theta}(s))\Psi'_{[nt]}(\tilde{\theta}(s)) + \mathcal{O}((\tilde{\theta}(s) - \widehat{\theta}_{[ns]})^2).\end{aligned}$$

Using the same proof used for the convergence of $(\Psi_{[nt]}(\theta^\star))/\sqrt{n}$ (see the Appendix), we show that, for $s \in [a,b]$, $(\Psi_{[nt]}(\tilde{\theta}(s)))/\sqrt{n}$ and $(\Psi'_{[nt]}(\tilde{\theta}(s)))/\sqrt{n}$ also converge to two zero-mean Gaussian processes. Then, using hypothesis **F3**, $\widehat{\theta}_{[ns]} \to \tilde{\theta}(s)$ uniformly over $[a,b]$, and then, $\hat{d}_n(t,s)$ to a zero mean Gaussian process $\Lambda(t,s)$.

We show now *ii)*. For $s \in [a,b]$, let $a_i(\tilde{\theta}(s)) = \mathbb{E}_{H_0}\left(X_{(i)}(\tilde{\theta}(s))\right)$. Following the proof of Theorem 2.2, consider first the case $k > k_n^\star$. On Ω_n, for $s \in [a,b]$,

$$\begin{aligned}T_{k,j}(\tilde{\theta}(s)) - Q_{k,j}(\tilde{\theta}(s)) &= \left(T_{k,j}(\tilde{\theta}(s)) - \mathbb{E}_{k_n^\star}\left(T_{k,j}(\tilde{\theta}(s))\right)\right) - B_{k,j,n}\left(T_{k,K_n}(\tilde{\theta}(s)) - \mathbb{E}_{k_n^\star}\left(T_{k,K_n}(\tilde{\theta}(s))\right)\right)\\ &\quad + \mathbb{E}_{k_n^\star}\left(T_{k,j}(\tilde{\theta}(s))\right) - B_{k,j}\mathbb{E}_{k_n^\star}\left(T_{k,K_n}(\tilde{\theta}(s))\right)\\ &= R_{k,j}(\tilde{\theta}(s)) + S_{k,j}(\tilde{\theta}(s)).\end{aligned}$$

As in Theorem 2.2, $R_{k,j}(\tilde{\theta}(s))\mathbb{1}_{\Omega_n}$ (normalized by $\sqrt{n}$) as a process indexed by $(t,w,s) \in (0,1)^2 \times (a,b)$ converges in distribution to a zero-mean Gaussian process. On the other hand,

$$S_{k,j}(\tilde{\theta}(s)) = \sum_{i=1}^{k-k_n^\star}\left(a_{i+j}(\tilde{\theta}(s)) - a_i(\tilde{\theta}(s)) + B_{k,j,n}(a_{i+K_n}(\tilde{\theta}(s)) - a_i(\tilde{\theta}(s))\right).$$

Thus, there exists a constant, $\gamma > 0$, which depends on a,b in [F3], such that $\sup_j |S_{k,j}(\tilde{\theta}(s))| \geq (k - k_n^\star)\gamma$. We conclude that $\mathbb{P}_{k_n^\star}\left(k_n^\star - \widehat{k} > n\,u_n\right) \to 0$ using the arguments used for Theorem 2.2. The case $k < k_n^\star$ is identical. □

4.2. The unknown variance case

When $\theta^\star$ is a scale parameter, *i.e.* $F_\varepsilon(y;\theta^\star) = F_\varepsilon(y/\theta^\star;1)$, we introduce the following procedure which is scale invariant:

i) for $i = 1,\ldots,n$, let $X_{(i)} = |y_{(i)}|$;
ii) let K_n be some positive integer. For $1 \leq k \leq n - K_n$ and $1 \leq j \leq K_n$, compute

$$T_{k,j} = \sum_{i=1}^{k+j} X_{(i)}, \tag{15}$$

$$Q_{k,j} = \frac{\mathbb{E}_{H_1(k)}\left(\sum_{i=k}^{k+j} X_{(i)}\right)}{\mathbb{E}_{H_1(k)}\left(\sum_{i=k}^{k+K_n} X_{(i)}\right)} T_{k,K_n}, \tag{16}$$

$$\eta_k = \max_{1\leq j\leq K_n} \frac{|T_{k,j} - Q_{k,j}|}{\sqrt{n}}; \tag{17}$$

iii) let

$$\hat{k}_u = \text{Arg} \min_{1 \le k \le n-K_n} \eta_k.$$

Remark. Notice, the minimization problem at hand is not changed if we consider $\frac{|T_{k,j}-Q_{k,j}|}{\sigma\sqrt{n}}$, so the procedure is indeed scale invariant. We have the following result, whose proof is omitted as it resembles quite closely that of Theorem 2.2.

Theorem 4.2. *Let (u_n) be any positive and decreasing sequence such that $\sqrt{n}\, u_n \to \infty$. Then, under the asymptotic framework defined by* **AF1, AF2,AF3**,

$$P\left(\left| \frac{\hat{k}_u}{n} - t^\star \right| > u_n \right) \to 0.$$

Moreover, for $a > 0$ there exist constants c_1, c_2 which depend on a such that if $u_n = \frac{c_1\alpha_n\sqrt{\log n}}{2\sqrt{n}} + \frac{c_2\alpha_n \log(n)}{2n}$ then

$$P\left(\left| \frac{\hat{k}_u}{n} - t^\star \right| > u_n \right) \le 2\mathrm{e}^{-a\log n} + 2P\left(\max_{1\le i \le n} \frac{|\varepsilon_i|}{\sigma} > \alpha_n \right).$$

4.3. Random thresholding

It is interesting we can link this procedure to a random thresholding one, or as in [4,5] in terms of penalized estimation. This link clearly appears when we use the ℓ_2-norm to define η_k:

$$\eta_k = n^{-2} \sum_{j=k+1}^{k+K_n} (T_{k,j} - Q_{k,j})^2.$$

Lemma 2.2 ensures that good choice for the cutpoint between significant and non significant coefficients is $\widehat{k} = \arg\min \eta_k$. Thus, it is reasonable to assume we are looking from left to right to the first k such that $\eta_k > \eta_{k-1}$. We will assume coefficients are significant while $\eta_k < \eta_{k-1}$. In order to develop this idea we must understand how $\eta_k - \eta_{k-1}$ looks like. We have

$$\begin{aligned} n^2(\eta_k - \eta_{k-1}) &= \sum_{j=1}^{K_n} (T_{k,j} - B_{k,j}T_{k,K_n})^2 - \sum_{j=1}^{K_n-1} T_{k-1,j} - B_{k-1,j}T_{k-1,n})^2 \\ &= (X_k - B_{k-1,k}T_{k-1,k-1+K_n})^2 + \sum_{j=1}^{K_n-1} (T_{k,j} - B_{k,j}T_{k,K_n}))^2 \\ &\quad - \sum_{j=1}^{K_n-1} \Big(X_k + T_{k,j} - B_{k,j}(X_k + T_{k,K_n})(1+o(1)\Big)^2 \\ &\approx (X_k - B_{k-1,k}T_{k-1,k-1+K_n})^2 \\ &\quad + \sum_{j=1}^{K_n-1} X_k^2(1 - B_{k,j,n})^2 + 2X_k(1-B_{k,j,n})(T_{k,j} - B_{k,j,n}T_{k,k+K_n}). \end{aligned}$$

Hence coefficients will be significant approximatively until the first k such that

$$X_k \le \tau_{k,n} := \frac{\sum_{j=1}^{K_n-1} (T_{k,j} - B_{k,j}T_{k,K_n})(1 - B_{k,j,n})}{\sum_{j=1}^{K_n-1} (1 - B_{k,j,n})^2}.$$

TABLE 1. Estimated percentiles of D_n under H_0 obtained with different values of n.

$n \setminus \alpha$	0.50	0.90	0.95	0.99
20	0.27	0.55	0.67	0.93
50	0.29	0.55	0.65	0.82
500	0.29	0.56	0.65	0.83
5000	0.30	0.55	0.64	0.79

5. Numerical experiments

We consider here the model

$$y_i = \mu_i + \varepsilon_i \tag{18}$$

where (ε_i) is a collection of i.i.d. r.v.

5.1. Testing the null hypothesis H_0

The distribution of $D_n = \max_j |T_j - \widehat{T}_j|/\sqrt{n}$ under H_0 is estimated by Monte-Carlo (using 5000 simulated samples). Here, the $(y_i; 1 \leq i \leq n)$ are i.i.d. N(0,1) r.v. Using $h(y_i) = y_i^2$, we set $X_{(i)} = -\log(1 - F_h(y_{(i)}^2))$ where F is the cumulative distribution of a $\chi^2(1)$ distribution. Then, (T_j), $(\widehat{T}_j)$ and D_n are computed as described in Section 2.

Table 1 displays the estimated percentiles of order 0.50, 0.90, 0.95 and 0.99 obtained with different values of n. We see in this table that the distribution of D_n (except the tail) does not depend on n for $n \geq 20$. In particular, $\mathbb{P}_{H_0}(D_n > 0.65) \approx 0.05$ for any $n \geq 20$.

Using a level $\alpha = 5\%$, the test consists in rejecting the null hypothesis H_0 if $D_n > 0.65$. We estimated the power of this test, by simulating data under H_1. Here, the $(y_i; 1 \leq i \leq n/5)$ are i.i.d. $\mathcal{N}(\mu, 1)$ r.v. Figure 1 displays the estimated probability to reject the null hypothesis H_0 for different values of μ and n.

5.2. Estimating the number of significant coefficients

5.2.1. *A Gaussian example*

In the following experiment, we have simulated 500 Gaussian random variables, with $\mu_i = 4$ for $1 \leq i \leq 100$, and $\mu_i = 0$ for $101 \leq i \leq 500$. (ε_i) is a collection of $\mathcal{N}(0,1)$ i.i.d. r.v.

Assuming that the variance of the ε_i's is known, we set

$$X_i = -\log(1 - F(y_i^2))$$

where F is the cumulative distribution function of a χ^2 r.v.

Then, we used the procedure described in Section 2.1. Figure 2 displays the two sequences (T_k) and (Q_k). We find $D_n = 14.95$ and reject the null hypothesis.

After rejecting the null hypothesis, we will estimate the number of significant coefficients, following the procedure described in Section 2.2. We use here $K = 200$. Then, for $k = 1, 2, \ldots, 300$, we computed the sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$. Figure 3 displays these two sequences for $k = 70$, $k = 100$ and $k = 130$. We see that $(T_{k,j})$ concentrates around its conditional expected value $(\mathbb{E}_{H_1(k)}(T_{k,j}|T_{k,200}))$ only for $k = 100$. A bias is clearly present for $k = 70$ and $k = 130$. The sequence (η_k) defined by $\eta_k = \sum_{j=1}^{200}(T_{k,j} - Q_{k,j})^2/\sqrt{n-k}$ is displayed Figure 4. A minimum at $\hat{k} = 97$ is obvious.

Repeating the same procedure with 100 simulated sequences, we obtained 100 values of $\hat{k}$. The mean value of $\hat{k}$ is 97.6 and the standard deviation is 4.8.

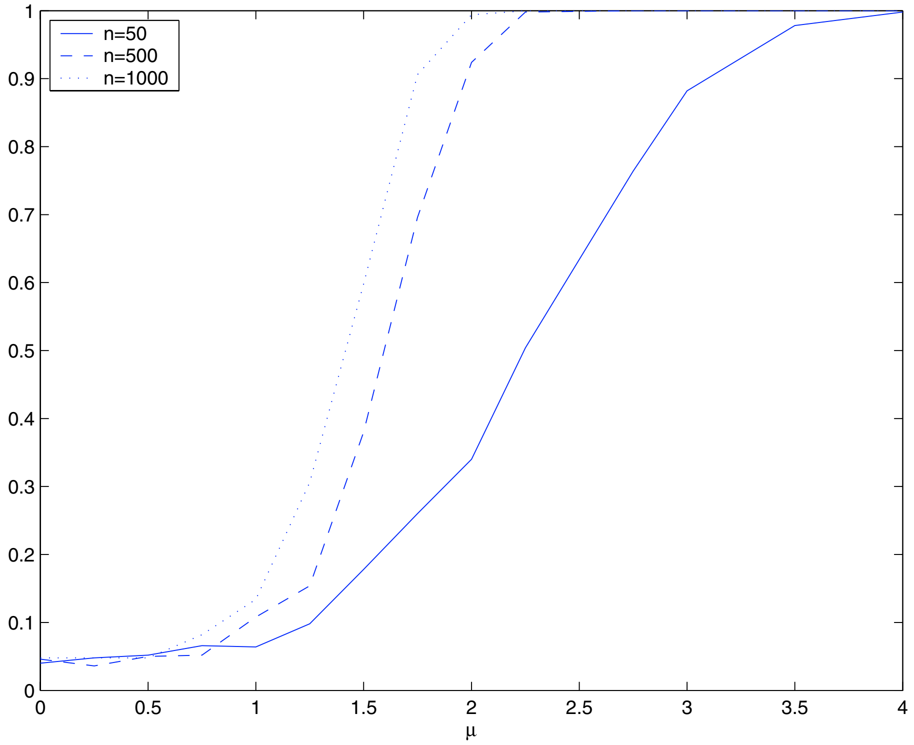

FIGURE 1. The estimated power of the 5% level test, for different values of μ and n.

If we consider now that the variance is unknown, we use the procedure described Section 4.1, estimating the variance under $\mathbf{H}_1(k)$ by

$$\widehat{\theta}_k = \frac{1}{n-k} \sum_{i=k+1}^{n} y_{(i)}^2 .$$

The results obtained when the variance is unknown are very similar than those obtained when the variance is known. The mean value of $\hat{k}$ is 97.3 and the standard deviation is 5.1.

5.2.2. *An exponential example*

In this second example, $n = 500$ again, but (ε_i) is a collection of *Expo(1)* i.i.d. r.v. Here, μ_i is uniformly distributed in $[3,6]$ for $1 \leq i \leq 100$, and $\mu_i = 0$ for $101 \leq i \leq 500$.

When the parameter of the exponential distribution is known, we use the procedure described Section 2.1, setting $X_i = y_i$. Figure 5 displays the two sequences (T_k) and (Q_k). We find $D_n = 3.72$ and reject the null hypothesis.

The number of significant coefficients is estimated as before. Figure 6 displays these two sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$ for $k = 70$, $k = 100$ and $k = 130$. In this example, the sequence (η_k) displayed Figure 7 is defined by $\eta_k = \sum_{j=1}^{200} |T_{k,j} - Q_{k,j}| / \sqrt{(n-k)^3}$. A minimum at $\hat{k} = 99$ is obvious.

Repeating the same procedure with 100 simulated sequences, we obtained 100 values of $\hat{k}$. The mean value of $\hat{k}$ is 103.8 and the standard deviation is 6.2.

The results obtained using the procedure described Section 4.1 when $\theta^\star$ is unknown are very similar: the mean value of $\hat{k}$ is 102.9 and the standard deviation is 5.6.

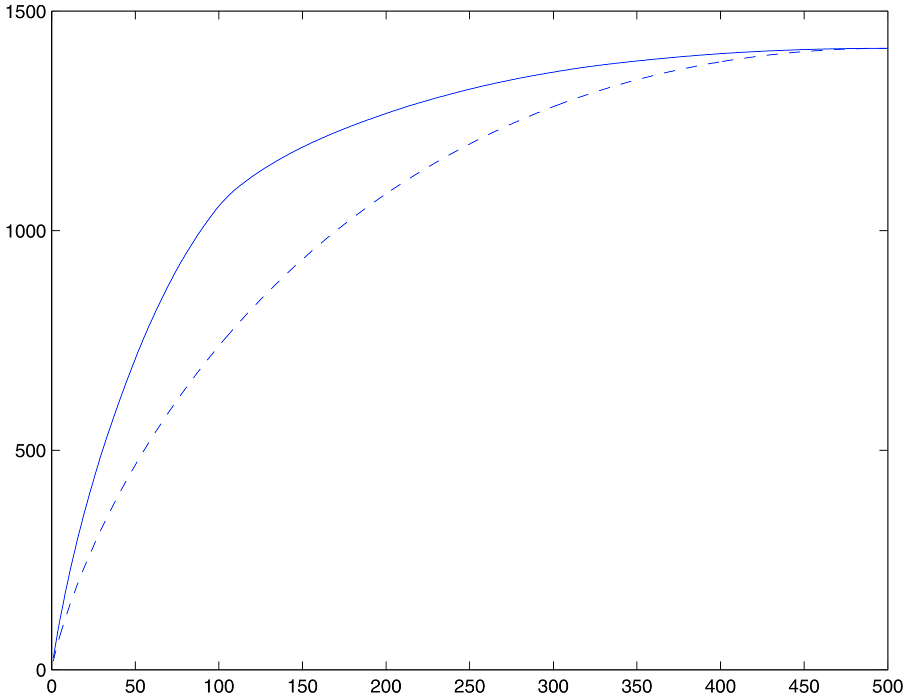

FIGURE 2. Example 1. The two sequences (T_k) and (Q_k).

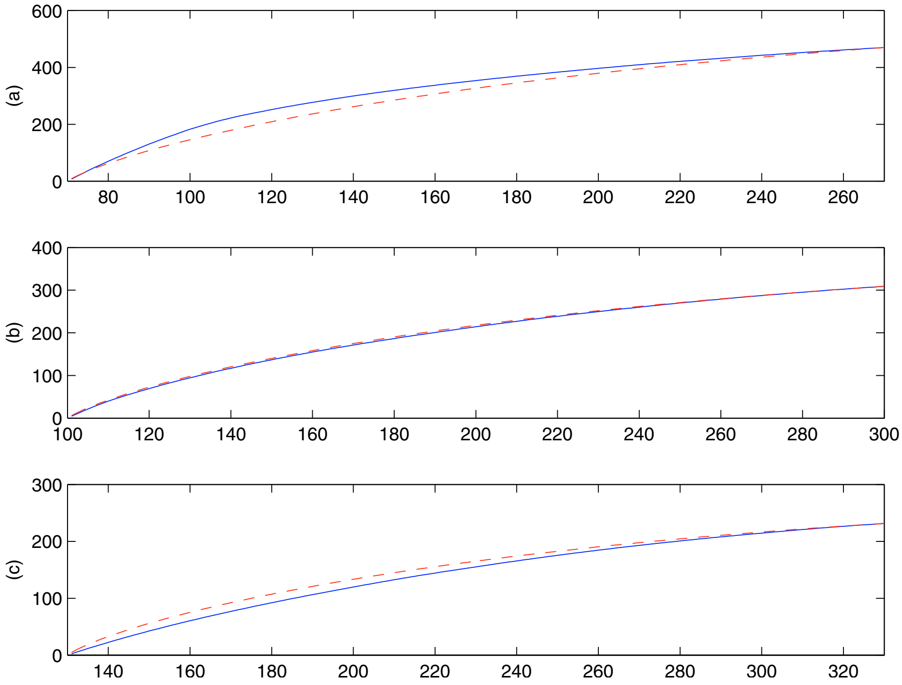

FIGURE 3. Example 1. The two sequences $(T_{k,j}, 1 \le j \le 200)$ and $(Q_{k,j}, 1 \le j \le 200)$.

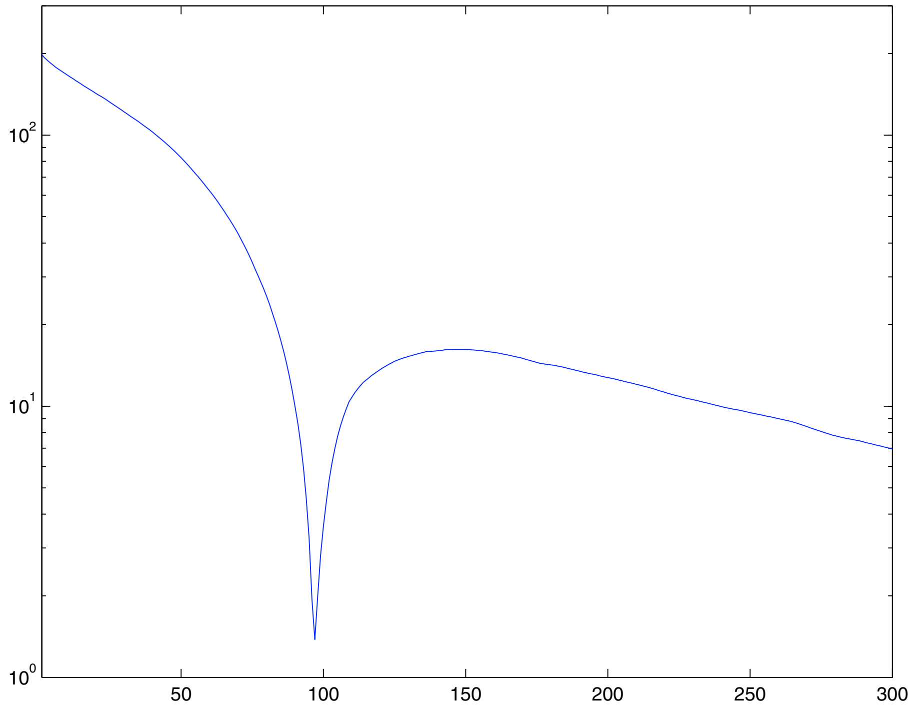

FIGURE 4. Example 1. The sequence (η_k) (in a semilog scale).

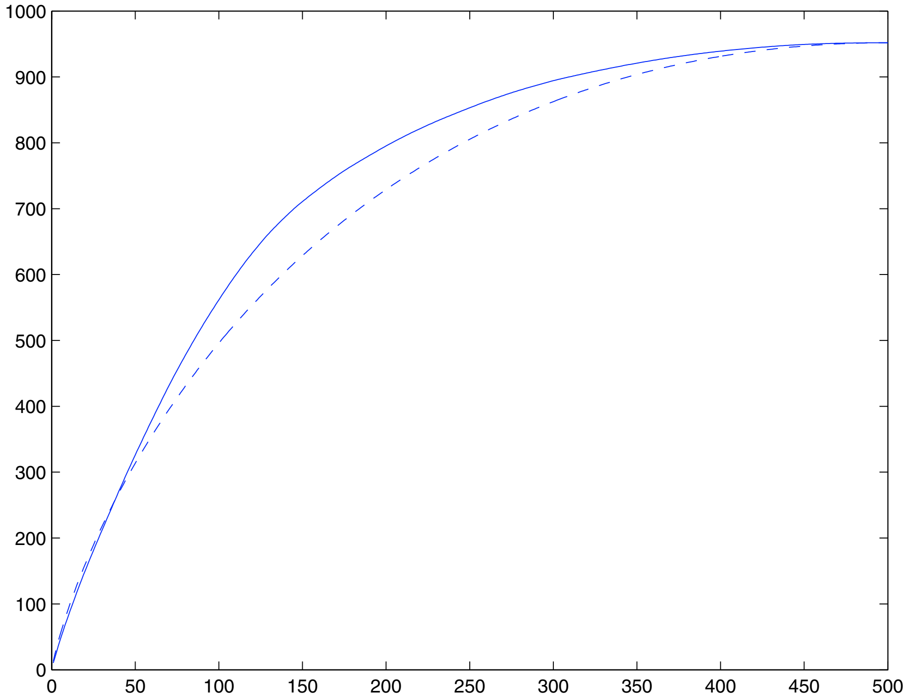

FIGURE 5. Example 2. The two sequences (T_k) and (Q_k).

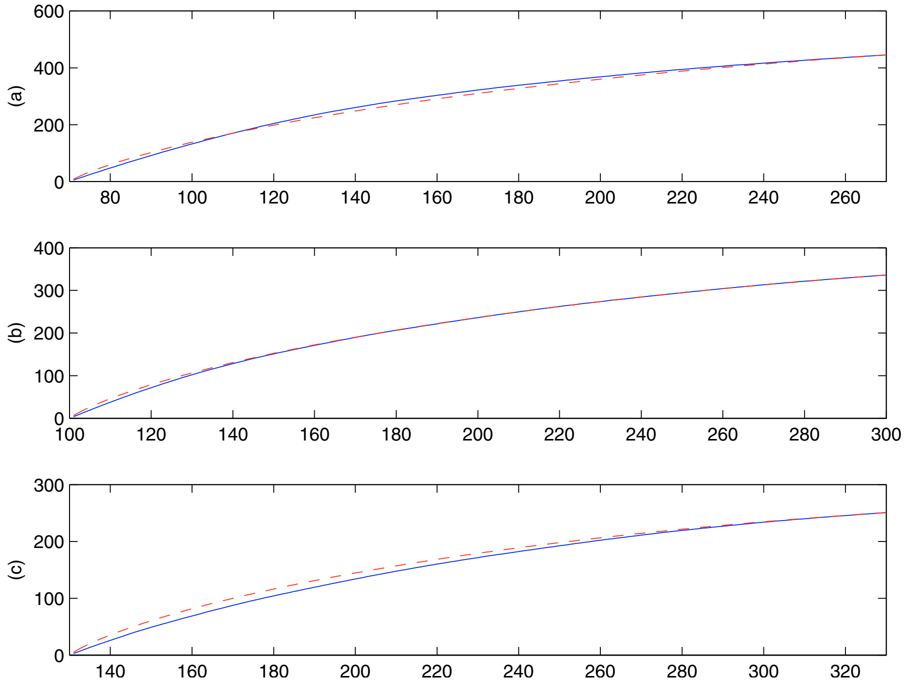

FIGURE 6. Example 2. The two sequences $(T_{k,j}, 1 \leq j \leq 200)$ and $(Q_{k,j}, 1 \leq j \leq 200)$.

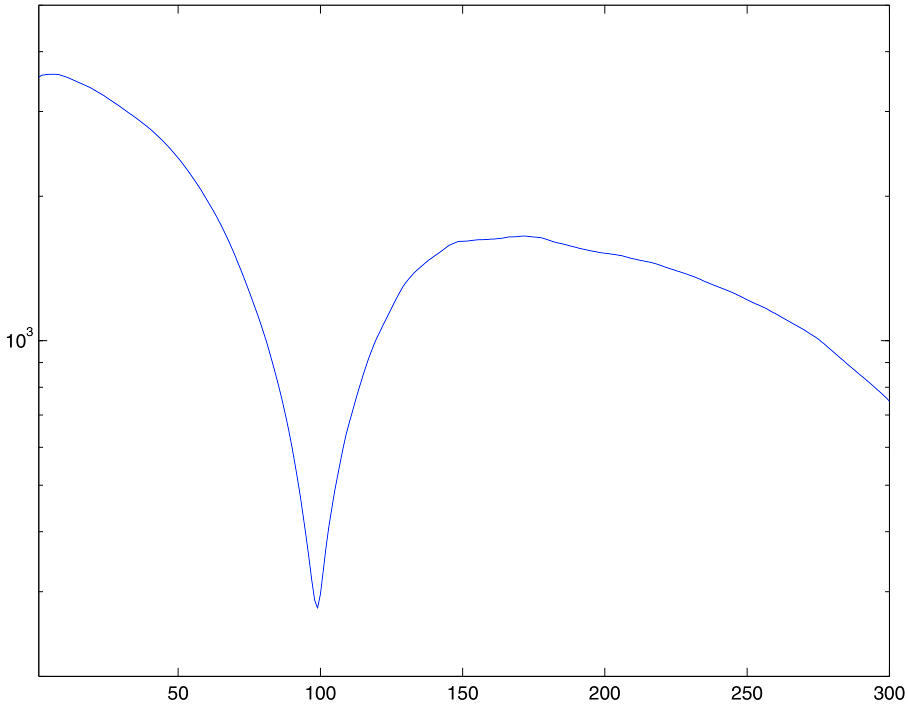

FIGURE 7. Example 2. The sequence (η_k) (in a semilog scale).

6. Wavelet thresholding for signal denoising

Assume that we observe

$$z_i = s(t_i) + e_i \ , \quad i = 1, \ldots, N \tag{19}$$

where $t_i = i/N$ and where e is a random noise. We aim to estimate the unknown signal s from the observed sequence $(z_i, 1 \leq i \leq N)$ without any particular assumption on s.

Then, denoising by wavelet thresholding involves the following three steps (see [10] for example):

1. choose a level J and perform a suitable wavelet transform of the data computing the wavelet coefficients $(c_{jk}, 0 \leq j \leq J, 1 \leq k \leq n_j)$;
2. perform a (hard or soft) thresholding of the wavelet coefficients:

$$\hat{c}_{jk} = \begin{cases} \tilde{c}_{jk}, & \text{if} |c_{jk}| \geq \lambda \\ 0, & \text{if} |c_{jk}| < \lambda. \end{cases} \tag{20}$$

   Hard thresholding means that $\tilde{c}_{jk} = c_{jk}$ and soft thresholding that $\tilde{c}_{jk} = sign(c_{jk})(|c_{jk}| - \lambda)_+$;
3. perform the inverse wavelet transform of the thresholded coefficients $(\hat{c}_{jk})$ to obtain the signal estimate $\hat{s}$.

Indeed, most signals have sparse wavelet series. In other words, a very small number of wavelet coefficients yields a very accurate reconstruction of most signals. Then, we will consider that the largest wavelet coefficients contain information about the unknown signal s, while the smallest ones are due to the random noise e. The problem here is the choice of the threshold λ, that is the number of significant coefficients to extract and to use for estimating the signal.

We use here the same global threshold λ for all the levels $j = 0, 1, \ldots, J$. Let (y_i) be the sequence of the n wavelet coefficients (with $n = \sum n_j$). If (e_i) is a Gaussian white noise with variance σ^2, then

$$y_i = \mu_i + \varepsilon_i \ , \quad i = 1, \ldots, n \tag{21}$$

where (ε_i) is also a Gaussian white noise with variance σ^2.

We are exactly in the context described Section 4.1: we can use our procedure for selecting the non zero coefficients μ_i when the distribution of ε depends on an unknown parameter (the variance σ^2 here).

6.1. An approximation result

More precisely consider the following setting.

1. Assume we observe $z_i = s(t_i) + e_i$ for a fixed collection t_i. Variables (e_i) are assumed to be independent and identically distributed with variance $\text{Var}(\varepsilon) = \sigma^2$.
2. Associated to the collection (t_i), we introduce the empirical inner product $\langle w, s \rangle_n = \frac{1}{n} \sum_i w(t_i) s(t_i)$ and its associated empirical norm $\| \cdot \|_n$.
3. We are interested in approximating s in terms of a certain orthonormal basis $\{\phi_\lambda\}_\lambda$. We assume that the basis is such that it is also orthonormal in the empirical norm $\langle , \rangle_n$.
4. Given the basis define the absolute empirical coefficients $y_j = \left| \langle y, \phi_j \rangle_n \right| \sqrt{n}$. More generally, we could consider the collection of the transformed coefficients $\gamma_j(h) = h(y_j/\sigma)$ for any given strictly increasing function h such that there exists β satisfying $h(ax) = a^\beta h(x)$ for any positive constant a.

Assume also the following assumptions are satisfied:

R1. e_i are an i.i.d. collection of centered normal r.v. with variance σ^2.

R2. $\{\phi_1, \ldots, \phi_n\}$ is orthonormal w.r.t. the empirical norm $\langle , \rangle_n$.

R3. For any $i \in I_{k_n^\star}$, $| \langle s, \phi_j \rangle_n | > a\sigma \sqrt{\log 2n} / \sqrt{n}$, with $a \geq 2\sqrt{2}$.

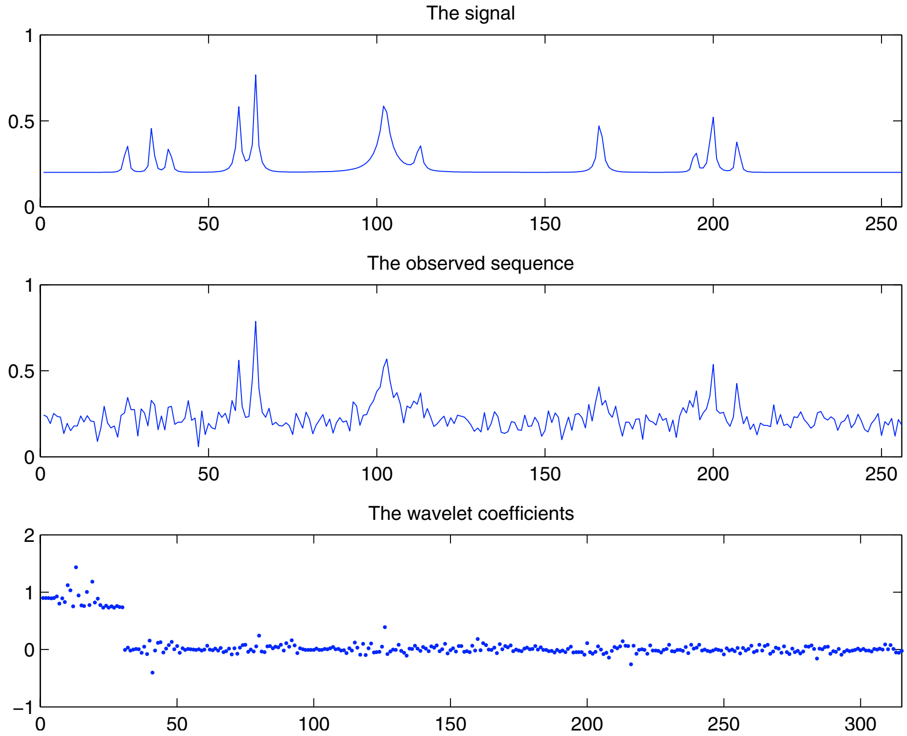

FIGURE 8. Denoising with wavelets. The unknown Bumps signal s, the observed series $(z_i, 1 \leq i \leq N)$ and the wavelet coefficients $(y_i, 1 \leq i \leq n)$.

We have the following result,

Lemma 6.1. *Assume* **AF1, AF3, R1, R2** *and* **R3** *hold true. Let* $\hat{k}_u$*, be the estimator defined in Section 4.2. Then, for* $b > 0$ *there exist constants* c_1, c_2 *which depend on* a *and* b *such that if* $u_n = \frac{c_1 \log(n)}{2\sqrt{n}} + \frac{c_2 \log^2(n)}{2n}$ *then*

$$P\left(\left|\frac{\hat{k}_u}{n} - t^\star\right| > u_n\right) \leq 2\mathrm{e}^{-b \log n} + 2\mathrm{e}^{-(a/2-\sqrt{2})\log(n)}.$$

Proof. It follows directly from Theorem 4.2 by checking that if $\varepsilon_i, i = 1, \ldots, n$ are independent standard normal random variables, then

$$P\left(\max_{1 \leq i \leq n} |\varepsilon_i| > a\sqrt{2n}\right) \leq \mathrm{e}^{-(a/2-\sqrt{2})\log(n)}. \qquad \square$$

6.2. Some numerical examples

Many different approaches have been proposed for this problem (see for example, [1, 2, 7, 10]). A complete comparison of all these methods is beyond the scope of the paper. We will illustrate our procedure with the Donoho and Johnstone's Bumps function. Figure 8 displays the sampled Bumps function of length 256, the observed series (z_i) and the wavelet coefficients (y_i). Here the SNR (ratio of the signal and noise variances) is 2.5. We used Daubechies' symmlet 8 wavelet in this example.

Figure 9 plots the sequence (T_k) of partial cumulative sums and the sequence (η_k). In this example, the sequence (η_k) attains its minimum for $\hat{k} = 39$.

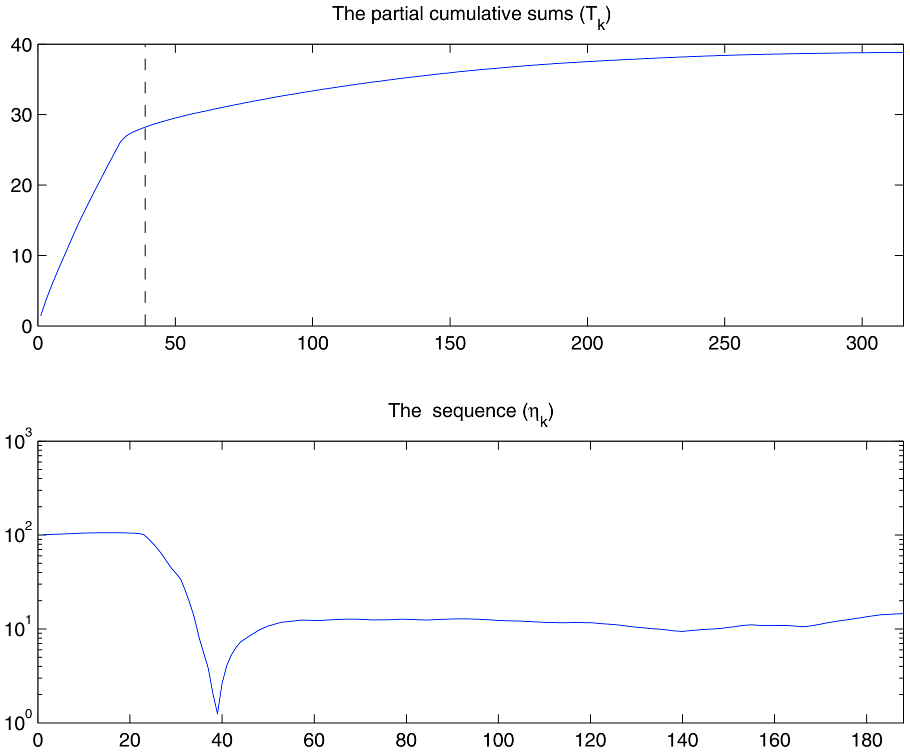

FIGURE 9. Denoising with wavelets. The partial cumulative sums (T_k) and the sequence (η_k) (in a semilog scale). The minimum of (η_k) is reached for $\hat{k} = 39$.

For any $0 \leq k \leq n$, let $\hat{s}_k$ be the reconstructed signal obtained by keeping the k largest (in absolute value) coefficients and padding the $n-k$ remaining coefficients to 0. The reconstruction of the signal $\hat{s}_{\hat{k}} = \hat{s}_{39}$ obtained by keeping only the 39 largest coefficients is displayed Figure 10. It is interesting to compare this reconstruction not only to the original signal s, but also to the ideal reconstruction that can be obtained with such a wavelet thresholding. The so-called *oracle* estimator minimizes the distance between the original s and the reconstruction. Using L_1 and L_2 distances, we define the L_1-oracle estimator $\hat{s}_{k_1^\star}$ and and the L_2-oracle estimator $\hat{s}_{k_2^\star}$ as follows:

$$\hat{s}_{k_1^\star} = \min_{0 \leq k \leq n} \sum_{i=1}^{N} |s(t_i) - \hat{s}_k(t_i)| \tag{22}$$

$$\hat{s}_{k_2^\star} = \min_{0 \leq k \leq n} \sum_{i=1}^{N} \left(s(t_i) - \hat{s}_k(t_i)\right)^2 . \tag{23}$$

In this example, our estimated signal obtained with $\hat{k} = 39$ coefficients is very close to the L_1-oracle estimator since $k_1^\star = 38$, while the L_2-oracle estimator is requires $k_2^\star = 54$ coefficients.

Using the same Bumps signal, the same SNR $= 2.5$ and the same symmlet 8 wavelet, we have compared our results with the different denoising procedures proposed in the Matlab Wavelet Toolbox. The results obtained with 1000 replications are summarized Table 2. For each method, we have estimated from these 1000 replications

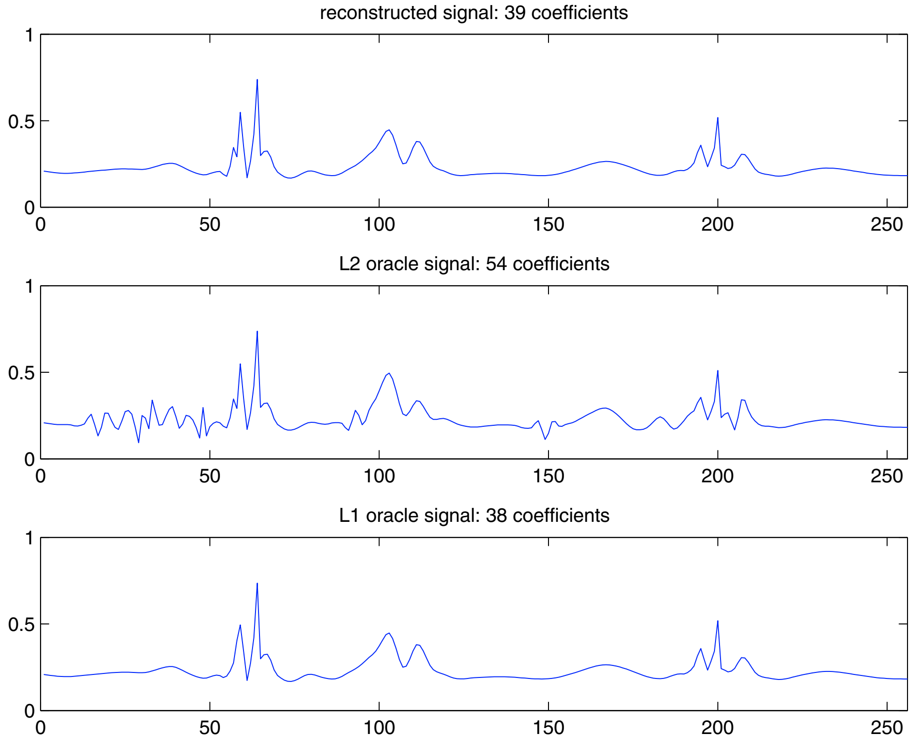

FIGURE 10. Denoising with wavelets. The reconstructed signal $\hat{s}_{39}$, the "ideal" reconstruction $\hat{s}_{54}$ considering the L_2-risk and the "ideal" reconstruction $\hat{s}_{38}$ considering the L_1-risk.

the expected risk ratios

$$\mathbb{E}\left(\frac{\sum_{i=1}^{N}\left|s(t_i)-\hat{s}_{\hat{k}}(t_i)\right|}{\sum_{i=1}^{N}\left|s(t_i)-\hat{s}_{k_1^\star}(t_i)\right|}\right)\quad\text{and}\quad\mathbb{E}\left(\frac{\sum_{i=1}^{N}\left(s(t_i)-\hat{s}_{\hat{k}}(t_i)\right)^2}{\sum_{i=1}^{N}\left(s(t_i)-\hat{s}_{k_1^\star}(t_i)\right)^2}\right),$$

and the expected differences of coefficient numbers

$$\mathbb{E}\left(|\hat{k}-k_1^\star|\right)\quad\text{and}\quad\mathbb{E}\left(|\hat{k}-k_2^\star|\right).$$

Rigrsure computes an adaptive threshold using principle of Stein's Unbiased Risk Estimate, *Heursure* is an heuristic variant of the first option, *Sqtwolog* threshold is $\sqrt{2 * log(n)}$, *Minimaxi* is a minimax thresholding and *Birgé Massart* uses the penalization $2\sigma^2 k(2+\log(n/k))$ proposed by Birgé and Massart.

We can remark that the *Heursure* estimator is the best one for the L_2 risk (the risk ratio is 1.06), but the L_1 risk is not so good. Furthermore, the number of wavelet coefficients used for estimating th signal with the first four methods are very different from the number of coefficients of the ideal L_1 and L_2 reconstructions. For this example, the proposed method and the penalization approach of Birgé and Massart yield almost the same very good results for the different criteria. The L_1 and L_2 risks are both almost optimal and the number of coefficients are well estimated.

This first result is very encouraging, but a more complete simulation study (using several signals of different lengths, with different SNR) will be considered in a further work.

TABLE 2. Results obtained from 1000 replications. Comparison of several thresholding methods.

	Risk ratio		Abs. diff. of coef. number	
Algorithm	L_1 Oracle	L_2 Oracle	L_1 oracle	L_2 oracle
Rigrsure	1.35	1.20	87.03	83.17
Heursure	1.13	1.06	34.20	30.59
Sqtwolog	1.10	1.16	12.67	16.43
Minimax	1.13	1.09	22.33	18.91
Birgé-Massart	1.05	1.07	7.73	10.06
Proposed method	1.05	1.07	7.03	9.79

7. Appendix

Proof of Proposition 2.1. For any $1 \leq i \leq n$, let

$$D_i = X_i - X_{i+1}$$

with $X_{n+1} = 0$. Thus, $X_i = \sum_{j=i}^n D_j$. Next, let $Z_j = jD_j$. As is well known, $(Z_j; 1 \leq j \leq n)$ is a sequence of i.i.d random variables (*Exp(1)*), so that, for any $1 \leq k \leq K \leq n$,

$$\begin{aligned} \mathbb{E}\left(T_k|T_K\right) &= \sum_{i=1}^{k}\sum_{j=i}^{n} \mathbb{E}\left(D_j|T_K\right) \\ &= \left(\sum_{i=1}^{k}\sum_{j=i}^{n}\frac{1}{j}\right)\mathbb{E}\left(Z_1|T_K\right). \end{aligned}$$

Since

$$\begin{aligned} \mathbb{E}\left(T_K|T_K\right) &= \left(\sum_{i=1}^{K}\sum_{j=i}^{n}\frac{1}{j}\right)\mathbb{E}\left(Z_1|T_K\right) \\ &= T_K \end{aligned}$$

and

$$\sum_{i=1}^{k}\sum_{j=i}^{n}\frac{1}{j} = k + k\sum_{j=k+1}^{n}\frac{1}{j}$$

we obtain

$$\mathbb{E}\left(T_k|T_K\right) = \frac{k + k\sum_{j=k+1}^{n} 1/j}{K + K\sum_{j=K+1}^{n} 1/j}\, T_K = \frac{B_{k,n}}{B_{K,n}}\, T_K. \tag{24}$$

□

Proof of Theorem 2.1.

Let $c_{nt,i} = \mathbb{1}_{[0,[nt]]}(i)$. By definition $\mathbb{E}\left(T_{[nt]}|T_n\right) = B_{[nt]}T_n$, so that $\mathbb{E}\left(T_{[nt]}\right) = \sum_i^n c_{nt,i}\mathbb{E}\left(X_{(i)}\right) = nB_{[nt]}$. Thus,

$$d_n(t) = \sum_i^n c_{nt,i}[X_{(i)} - \mathbb{E}\left(X_{(i)}\right)] - \frac{\sum_i^n c_{nt,i}\mathbb{E}\left(X_{(i)}\right)}{n}(T_n - n) = I_n(t) - II_n(t).$$

Let $G_n = \sum_{i=1}^{n} \zeta_{n-i}$ stand for the empirical sum of uniform r.v. ζ_i. Then, as in [16] it can be seen that

$$\frac{1}{\sqrt{n}} I_n(t) = -\frac{1}{\sqrt{n}} \int_0^1 [G_n - I](s) \mathbb{1}_{[0,t]}(s) \mathrm{d}F^{-1}(s) + o_p(1)$$

and

$$\frac{1}{\sqrt{n}} II_n(t) = -(t - t\log(t)) \frac{1}{\sqrt{n}} \int_0^1 [G_n - I](s) \mathrm{d}F^{-1}(s) + o_p(1),$$

where $o_p(1)$ is uniform for all $t \in [0,1)$. So that,

$$\frac{1}{\sqrt{n}} d_n(t) = \int_0^1 [R^t(u) - (t - t\log(t))F^{-1}(u)] \mathrm{d}[G_n(s) - (1-s)] + o_p(1),$$

with $R^t(u) = \int_0^u \mathrm{d}F^{-1}(s) \mathbb{1}_{[0,t]}(s)\mathrm{d}s$. The result now follows because

$$\mathcal{G} = \{R^t - (t - t\log(t))F^{-1}, t \in [0,1]\}$$

is a Donsker class.

□

References

[1] F. Abramovich and Y. Benjamini, Adaptive thresholding of wavelet coefficients. *CSDA* **4** (1996) 351–361.

[2] A. Antoniadis, J. Bigot and T. Sapatinas, Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Statist. Software* **6** (2001) 1–83.

[3] B. Arnold, N. Balakrishnan and H. Nagaraja, *A first course in order statistics.* Wiley series in probability (1993).

[4] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** (1999) 301–413.

[5] L. Birgé and P. Massart, Minimal penalties for Gaussian model selection. *Probab. Theor. Rel. Fields* **138** (2007) 33–73.

[6] O. Bousquet, Concentration inequalities for sub-additive functions using the entropy method, in *Stochastic inequalities and applications*, *Progr. Probab.* Birkhäuser, Basel **56** (2003) 213–247.

[7] T. Cai, Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Stat.* **3** (1999) 898–924.

[8] L. de Haan, *On regular variation and its application to the weak convergence of sample extremes.* 3rd ed., Mathematical Centre Tracts **32** Amsterdam (1975).

[9] D. Donoho and J. Jin, Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.* **32** (2004) 962–994.

[10] D. Donoho and I. Johnstone, Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** (1994) 425–455.

[11] M. Falk, A note on uniform asymptotic normality of intermidiate order statistics. *Ann. Inst. Statist. Math* **41** (1989) 19–29.

[12] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** (2001) 1348–1360.

[13] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning.* Springer, Series in statistics (2001).

[14] N. Meinhausen and J. Rice, Estimating the proportion of false null hypothesis among a large number of independently tested hypothesis. *Ann. Stat.* **34** (2006) 373–393.

[15] M.S. Pinsker, Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inform.* **2** (1980) 52–68.

[16] G. Shorak and J. Wellner, *Empirical processes with Applications to Statistics.* Wiley (1986).

[17] R. Tibshirani, Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B* **58** (1996) 267–288.