

# ON ADAPTIVE WAVELET ESTIMATION OF A CLASS OF WEIGHTED DENSITIES

FABIEN NAVARRO<sup>1,2</sup>, CHRISTOPHE CHESNEAU<sup>1</sup> AND JALAL FADILI<sup>2</sup>

<sup>1,2</sup>*LMNO CNRS-Université de Caen, 14032 Caen Cedex, France*

<sup>2</sup>*GREYC CNRS-ENSICAEN-Université de Caen, 14050 Caen Cedex, France*

## Abstract

We investigate the estimation of a weighted density taking the form  $g = w(F)f$ , where  $f$  denotes an unknown density,  $F$  the associated distribution function and  $w$  is a known non-negative weight. Such a class encompasses many examples, including those arising in order statistics or when  $g$  is related to the maximum or the minimum of  $N$  (random or fixed) independent and identically distributed (*i.i.d.*) random variables. We here construct a new adaptive non-parametric estimator for  $g$  based on a plug-in approach and the wavelets methodology. For a wide class of models, we show that it attains fast rates of convergence under the  $\mathbb{L}_p$  risk with  $p \geq 2$  over Besov balls. Our estimator is also simple to implement and fast. We also report an extensive simulation study to support our findings.

**Key words and phrases:** Reliability, weighted density, density estimation, plug-in approach, wavelets, block thresholding, series system, parallel system.

## 1 Introduction

### 1.1 Problem statement

Let  $(\Omega, \mathcal{A}, \mathbb{P})$  be a probability space,  $X$  be a real random variable with unknown density  $f$  and  $Y$  be a random variable having the unknown weighted density

$$g(x) = w(F(x))f(x), \quad x \in \mathbb{R}, \quad (1.1)$$

where  $w$  denotes a known non-negative weight and  $F$  denotes the distribution function of  $f$ . The goal we pursue here is to estimate  $g$  from a  $n$ -*i.i.d.* sample  $X_1, \dots, X_n$  of  $X$ .

Such an estimation problem arises in many situations, typically when  $g$  is related to the maximum<sup>1</sup> of  $N$  *i.i.d.* random variables, where  $N$  is a discrete random number in  $\mathbb{N}^*$  which is independent of the  $X_i$ 's. Application fields cover hydrology, meteorology, reliability, investment, management science, insurance business, etc.. For example, when the  $X_i$  are non-negative, the random variable  $Y = \max(X_1, \dots, X_N)$  (or  $Y = \min(X_1, \dots, X_N)$ ) arises naturally in reliability theory as the lifetime of a parallel (series) system with a random number  $N$  of identical components with lifetimes  $X_1, \dots, X_N$ .

To make things clearer to the reader, we next give some illustrative examples.

### 1.2 Motivating examples

**Example 1.1** (Order statistics). Let  $X_1, \dots, X_m$  be  $m$  *i.i.d.* random variables with absolutely continuous distribution function  $F$  and probability density function (pdf)  $f$ . Let  $X_{(1)} \leq \dots \leq X_{(m)}$  denote the corresponding order statistics. Then, the pdf  $g_{X_{(j)}}$  of the  $j$ -th order statistic is

$$g_{X_{(j)}}(x) = \frac{m!}{(j-1)!(m-j)!} (F(x))^{j-1} (1-F(x))^{m-j} f(x), \quad x \in \mathbb{R}.$$

Thus,  $X_{(m)}$ , for example, is the random variable representing the largest observation of a sample of  $n$  and corresponds to the sample maximum and the density  $g_{X_{(m)}}$  of  $X_{(m)} = \max(X_1, \dots, X_m)$  is given by

$$g_{X_{(m)}}(x) = m (F(x))^{m-1} f(x), \quad x \in \mathbb{R}.$$

---

<sup>1</sup>Since  $\min(X_1, \dots, X_N) = -\max(-X_1, \dots, -X_N)$  the results can be easily reformulated for the sample minimum.

The aim is to estimate  $g_{X^{(j)}}$  from a  $n$ -i.i.d. sample  $X_1, \dots, X_n$  of  $X$ .

**Example 1.2** (Maximum of a random number  $N$  of i.i.d. random variables). Let  $X$  be a random variable with density  $f$ ,  $\{X_i\}_{i \in \mathbb{N}^*}$  be a sequence of i.i.d. random variables with density  $f$  and  $N$  be a discrete random variable taking values in  $\mathbb{N}^*$  with a known probability mass function. Then the density of  $Y = \max(X_1, \dots, X_N)$  is

$$g(x) = w(F(x))f(x), \quad x \in \mathbb{R}, \quad (1.2)$$

where

$$w(u) = \sum_{k=1}^{\infty} ku^{k-1} \mathbb{P}(N = k), \quad u \in [0, 1].$$

The goal is again to estimate  $g$  from a  $n$ -i.i.d. sample  $X_1, \dots, X_n$  of  $X$ .

**Example 1.3** (Pile-up model). Let us now present the ‘‘pile-up model’’. Let  $\{Y_i\}_{i \in \mathbb{N}^*}$  be a sequence of i.i.d. random variables with density  $g$ ,  $N$  be a discrete random variable in  $\mathbb{N}^*$  as in the previous example, and let  $X = \min(Y_1, \dots, Y_N)$  with density  $f$ . Then the density of  $Y_1$  is

$$g(x) = w(F(x))f(x), \quad x \in \mathbb{R},$$

where

$$w(u) = \frac{1}{M'(M^{-1}(1-u))}, \quad u \in [0, 1],$$

$M(u) = \mathbb{E}(u^N)$  and  $M'(u) = \mathbb{E}(Nu^{N-1})$ . We are seeking an estimate of  $g$  from a  $n$ -i.i.d. sample  $X_1, \dots, X_n$  of  $X$ .

### 1.3 Previous work

Some distributional properties of the maximum and minimum of random variables have been extensively studied in the literature (see, e.g., [Raghunandan and Patil \(1972\)](#), [Shaked \(1975\)](#) and [Shaked and Wong \(1997\)](#)). In addition, the literature on order statistics contains a huge work about the maximum. In the context of extreme value theory, various statistical properties and (real data) applications can be found in [Adamidis et al. \(2005\)](#), [Louzada et al. \(2012\)](#) and the references therein.

The estimation of the density function of the maximum of two independent random variables has been considered by [Chen and Hsu \(2004\)](#) via kernel methods. The Pile-up model has been considered by [Comte and Rebafka \(2010\)](#) via model selection methods.

### 1.4 Contributions and relation to prior work

In this paper, we develop a new *non-linear* adaptive estimator for  $g$  in model (1.1) based on a plug-in method, wavelets and the block thresholding rule introduced by [Cai \(1999\)](#). Wavelet-based thresholding estimators are attractive for non-parametric function estimation because of their virtues from the viewpoints of spatial adaptivity, computational efficiency and asymptotic optimality properties. In the case of simple density estimation, wavelet thresholding is probably one of the most attractive nonlinear methods. We refer to e.g., [Antoniadis \(1997\)](#), [Härdle et al. \(1998\)](#) and [Vidakovic \(1999\)](#) for a detailed discussion of the performances of wavelet estimators and some of their advantages over traditional methods such as kernel-based or projection estimators.

We here explore the theoretical performance of our estimator under the  $\mathbb{L}_p$  risk with  $p \geq 2$  over a very rich class of function spaces, namely Besov spaces. Sharp rates of convergence are obtained. Application of our estimator to Example 1.2 above is described in detail. Finally, extensive simulation experiments are carried out to illustrate the practical performance of our estimator. In particular, the numerical tests indicate that our block thresholding estimator, which is simple to implement and fast, compares very favorably to standard kernel-based methods.

### 1.5 Paper organization

The paper is structured as follows. Our wavelet estimator is described in Section 2. Section 3 presents our estimator convergence rates. Simulations are detailed in Section 4.

## 2 Wavelet estimators

First of all, we briefly recall some key facts on wavelets and Besov spaces that will be essential to us in the sequel. Then we develop our nonlinear adaptive wavelet block thresholding estimator.

### 2.1 Wavelets and Besov balls

Let  $b > 0$ ,  $p > 0$  and  $\mathbb{L}_p([-b, b]) = \left\{ h : [-b, b] \rightarrow \mathbb{R}; \|h\|_p^p = \int_{-b}^b |h(x)|^p dx < +\infty \right\}$ .

For the purposes of this paper, we use compactly supported wavelet bases on  $[-b, b]$ . More precisely, we consider the Daubechies family  $\text{db}_{2T}$  with the scaling and wavelet functions  $\phi$  and  $\psi$ , where  $R \geq 2$  is a fixed integer, see e.g., Mallat (2009). Define the scaled and translated version of the  $\phi$  and  $\psi$

$$\phi_{j,k}(x) = 2^{j/2} \phi(2^j x - k), \quad \psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k).$$

Then there exists an integer  $\tau$  and a set of consecutive integers  $\Lambda_j$  such that  $\text{Card}(\Lambda_j) = C2^j$  for a  $C > 0$  and, for any integer  $\ell \geq \tau$ , the collection

$$\mathcal{B} = \{\phi_{\ell,k}, k \in \Lambda_\ell; \psi_{j,k}; j \in \mathbb{N} - \{0, \dots, \ell - 1\}, k \in \Lambda_j\},$$

is an orthonormal basis of  $\mathbb{L}_2([-b, b])$ .

Consequently, for any integer  $\ell \geq \tau$ , any  $h \in \mathbb{L}_2([-b, b])$  can be expanded on  $\mathcal{B}$  as

$$h(x) = \sum_{k \in \Lambda_\ell} \alpha_{\ell,k} \phi_{\ell,k}(x) + \sum_{j=\ell}^{\infty} \sum_{k \in \Lambda_j} \beta_{j,k} \psi_{j,k}(x),$$

where

$$\alpha_{\ell,k} = \int_{-b}^b h(x) \phi_{\ell,k}(x) dx, \quad \beta_{j,k} = \int_{-b}^b h(x) \psi_{j,k}(x) dx. \quad (2.1)$$

As is traditional in the wavelet estimation literature, we will investigate the performance of our estimator by assuming that the unknown density  $f$  belongs to a Besov ball  $B_{q,r}^s(M)$ . The Besov norm for a function can be related to a sequence space norm on its wavelet coefficients. More precisely, for  $M > 0$ ,  $s \in (0, T)$ ,  $q \geq 1$  and  $r \geq 1$ , a function  $h \in \mathbb{L}_p([-b, b])$  belongs to  $B_{q,r}^s(M)$  if and only if there exists a constant  $M^* > 0$  (depending on  $M$ ) such that the associated wavelet coefficients (2.1) satisfy

$$\left( \sum_{k \in \Lambda_\tau} |\alpha_{\tau,k}|^q \right)^{1/q} + \left( \sum_{j=\tau}^{\infty} \left( 2^{j(s+1/2-1/q)} \left( \sum_{k \in \Lambda_j} |\beta_{j,k}|^q \right)^{1/q} \right)^r \right)^{1/r} \leq M^*,$$

with the usual modifications if  $q = \infty$  or  $r = \infty$ .

In this expression,  $s$  is a smoothness parameter and  $q$  and  $r$  are norm parameters. They include many traditional smoothness spaces such as Hölder and Sobolev spaces. A comprehensive account on Besov spaces can be found in e.g., Devore and Popov (1988); Meyer (1992); Härdle et al. (1998).

### 2.2 Plug-in block wavelet estimator

Let us consider the general statistical framework described in Section 1 with a  $n$ -i.i.d. sample  $X_1, \dots, X_n$  of  $X$ . First of all, we investigate the estimation of  $f$  via the so-called wavelet block hard thresholding estimator. We suppose that  $\text{supp}(f) \subseteq [-b, b]$  with  $b > 0$ .

Let  $p \geq 2$ , and  $j_1$  and  $j_2$  be the integers corresponding to the finest and coarsest scales defined as

$$j_1 = \lfloor (p/2) \log_2(\log n) \rfloor, \quad j_2 = \lfloor \log_2(n / \log n) \rfloor,$$

where  $\lfloor a \rfloor$  denotes the whole number part of  $a \in \mathbb{R}^+$ . For any integer  $j \in \{j_1, \dots, j_2\}$ , let  $A_j$  and  $U_{j,K}$  be given such that  $(U_{j,K})_{K \in A_j}$  form a partition of  $\Lambda_j$  and all the  $U_{j,K}$ 's are defined with  $L = \lfloor (\log n)^{p/2} \rfloor$  consecutive integers  $k$ . In a nutshell, at each scale  $j$ , each  $U_{j,K}$  is the set containing position indices of the wavelet coefficients inside block  $K \in A_j$ . Note that the number of elements in  $A_j$  is proportional to  $\lfloor 2^j / L \rfloor$ .

We define the wavelet block hard thresholding estimator of  $f$  by

$$\widehat{f}(x) = \sum_{k \in \Lambda_{j_1}} \widehat{\alpha}_{j_1, k} \phi_{j_1, k}(x) + \sum_{j=j_1}^{j_2} \sum_{K \in A_j} \sum_{k \in U_{j, K}} \widehat{\beta}_{j, k} \mathbb{1} \left\{ \left( \sum_{k \in U_{j, K}} |\widehat{\beta}_{j, k}|^p / L \right)^{1/p} \geq \kappa n^{-1/2} \right\} \psi_{j, k}(x), \quad x \in [-b, b], \quad (2.2)$$

where  $\mathbb{1}_{\{\cdot\}}$  denotes the indicator function, and

$$\widehat{\alpha}_{j_1, k} = \frac{1}{n} \sum_{i=1}^n \phi_{j_1, k}(X_i), \quad \widehat{\beta}_{j, k} = \frac{1}{n} \sum_{i=1}^n \psi_{j, k}(X_i)$$

and  $\kappa > 0$  is a threshold parameter to be discussed later.

The estimator  $\widehat{f}$  was initially developed by [Cai \(1999\)](#) for the regression model under the  $\mathbb{L}_2$ -risk with equispaced deterministic samples. The  $\mathbb{L}_p$  risk version was studied in [Picard and Tribouley \(2000\)](#) for the standard density estimation problem and by [Chesneau \(2010\)](#) for the biased density estimation problem.

The idea underlying  $\widehat{f}$  (2.2) is to operate a group/block selection: it keeps intact the large groups of unknown wavelet coefficients of  $f$  (2.1) and removes the others. Wavelet block thresholding is one of the most attractive non-linear thresholding methods, since it is both numerically straightforward to implement and asymptotically optimal for a large variety of Sobolev or Besov classes. Detailed references on the subject for various models include, but are not limited to, [Cai \(1999, 2002\)](#), [Li and Xiao \(2008\)](#); [Li \(2008\)](#), [Picard and Tribouley \(2000\)](#) and [Chesneau \(2008, 2010\)](#).

The performance of Block thresholding estimators depends on the threshold level  $\kappa$ . In the non-parametric regression setting, in order to choose this key parameter, [Cai and Zhou \(2009\)](#) proposed an adaptive James-Stein block thresholding estimator whose parameters (including the threshold) minimize the Stein's unbiased risk estimate (SURE) and established its minimax rates of convergence under the mean squared error over Besov balls. Other selection strategies have been developed in the literature (see e.g. [Nason \(1996\)](#) which considered wavelet estimators based on cross-validation to choose the thresholding parameter in practice). In this work, we focus on the universal threshold proposed by [Donoho et al. \(1996\)](#). The reason for this choice is twofold. First, it is the one consistent with the theoretical convergence rates established in Section 3. Secondly, it allows to remain fair when comparing to the other methods of the literature tested in Section 4.

Finally, plugging (2.2) into (1.1) leads to the following estimator of  $g$ :

$$\widehat{g}(x) = w(\widehat{F}(x)) \widehat{f}(x), \quad x \in [-b, b], \quad (2.3)$$

where

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}. \quad (2.4)$$

The rest of the paper explores the theoretical and practical performances of  $\widehat{g}$ .

### 3 Estimator convergence rates

In this section, we discuss the asymptotic properties of the proposed estimator. Rates of  $\mathbb{L}_p$  convergence are investigated under the following assumptions:

**(A.1)** Compact support:  $\text{supp}(f) \subseteq [-b, b]$  with  $b > 0$ .

**(A.2)** Uniform boundedness: there exists a constant  $C_1 > 0$  such that

$$\sup_{x \in [-b, b]} f(x) \leq C_1. \quad (3.1)$$

**(A.3)** Lipschitz continuity of  $w$ : there exist two constants  $\theta \in (0, 1]$  and  $C_2 > 0$  such that

$$|w(u) - w(v)| \leq C_2 |u - v|^\theta, \quad \text{for all } (u, v) \in [0, 1]^2, \quad (3.2)$$

Assumption A.1 is a usual one in the wavelet density estimation framework (see e.g. [Donoho et al. \(1996\)](#)). Extension to non-compactly supported densities might be possible and ideas from [Juditsky and Lambert-Lacroix \(2004\)](#) might be inspiring, although these authors considered a model different from ours, and their results were valid only for the case where  $f$  is in the Hölder class. Such an extension is however beyond the scope of this paper and we leave it for a future work.

Theorem 3.1 studies the  $\mathbb{L}_p$  risk of  $\widehat{g}$  (2.3) over Besov balls and Assumptions A.1-A.3 on  $f$  and  $w$ .

**Theorem 3.1.** *Consider the general statistical framework described in Section 1 (the estimation of  $g$  (1.1) is of interest). Suppose that Assumptions A.1-A.3 hold. Let  $p \geq 2$  and  $\widehat{g}$  be given by (2.3). Then, for any  $f \in B_{q,r}^s(M)$ ,  $q \geq 1$ ,  $r \geq 1$  and  $s \in (1/q, T)$ , there exists a constant  $C > 0$  such that*

$$\mathbb{E}(\|\widehat{g} - g\|_p^p) \leq C \max(\varphi_{n,p}, n^{-\theta p/2}),$$

where

$$\varphi_{n,p} = \begin{cases} n^{-sp/(2s+1)}, & \text{if } q \geq p, \\ \left(\frac{\log n}{n}\right)^{sp/(2s+1)}, & \text{if } \{p > q, qs > (p-q)/2\}, \\ \left(\frac{\log n}{n}\right)^{(s-1/q+1/p)p/(2(s-1/q)+1)}, & \text{if } qs < (p-q)/2 \text{ or } \{qs = (p-q)/2, p \leq q/r\}, \\ \left(\frac{\log n}{n}\right)^{(s-1/q+1/p)p/(2(s-1/q)+1)} (\log n)^{p-q/r}, & \text{if } \{qs = (p-q)/2, p > q/r\}. \end{cases} \quad (3.3)$$

Note that the rate  $\varphi_{n,p}$  in (3.3) is the near optimal (or optimal in some regimes) one in the minimax sense for  $f$ . See, e.g., Donoho et al. (1996) and Härdle et al. (1998).

In the case  $\theta \geq \max(2s/(2s+1), 2(s-1/q+1/p)p/(2(s-1/q)+1))$  (as  $\theta = 1$ ), the near optimality of the estimator of  $f$  is transferred to that of  $g$  through the plug-in principle. In the other case, the regularity of  $w$  deteriorated the rate of convergence; it becomes  $n^{-\theta p/2}$ .

The proof of Theorem 3.1 uses a suitable decomposition of the  $\mathbb{L}_p$  risk and capitalizes on results on the performances of  $\widehat{f}$  (2.2) and  $\widehat{F}$  (2.4) established in Chesneau (2010).

**Proof of Theorem 3.1.** Observe that

$$\widehat{g}(x) - g(x) = w(\widehat{F}(x))(\widehat{f}(x) - f(x)) + f(x)(w(\widehat{F}(x)) - w(F(x))).$$

By Assumption A.3 implying  $\sup_{x \in [0,1]} w(x) \leq C$ , together with Assumption A.2, we have

$$\begin{aligned} |\widehat{g}(x) - g(x)| &\leq C(|\widehat{f}(x) - f(x)| + |w(\widehat{F}(x)) - w(F(x))|) \\ &\leq C(|\widehat{f}(x) - f(x)| + |\widehat{F}(x) - F(x)|^\theta). \end{aligned}$$

By the Jensen inequality, we have

$$\mathbb{E}(\|\widehat{g} - g\|_p^p) \leq C \left( \mathbb{E}(\|\widehat{f} - f\|_p^p) + \mathbb{E}(\|\widehat{F} - F\|_{\theta p}^{\theta p}) \right).$$

It follows from (Chesneau, 2010, Theorem 4.1 with  $w(x) = 1 = \mu$ ) that

$$\mathbb{E}(\|\widehat{f} - f\|_p^p) \leq C\varphi_{n,p},$$

where  $\varphi_{n,p}$  is given by (3.3).

By Assumption A.1, we have

$$\mathbb{E}(\|\widehat{F} - F\|_{\theta p}^{\theta p}) \leq C \sup_{x \in [-b,b]} \mathbb{E}(|\widehat{F}(x) - F(x)|^{\theta p}).$$

In order bound the last term, note that

$$\widehat{F}(x) - F(x) = \frac{1}{n} \sum_{i=1}^n U_i(x),$$

with  $U_i(x) = \mathbb{1}_{\{X_i \leq x\}} - F(x)$ ,  $U_1(x), \dots, U_n(x)$  are *i.i.d.* with  $\mathbb{E}(U_1(x)) = 0$ ,  $|U_1(x)| \leq 2$  and  $\mathbb{E}((U_1(x))^2) \leq 4$ .

The case  $\theta < 2/p$ : it follows from the Hölder inequality that

$$\mathbb{E}(|\widehat{F}(x) - F(x)|^{\theta p}) \leq \left( \mathbb{E}((\widehat{F}(x) - F(x))^2) \right)^{\theta p/2} = \left( \frac{1}{n} \mathbb{V}(U_1(x)) \right)^{\theta p/2} \leq C n^{-\theta p/2}.$$

The case  $\theta \geq 2/p$ : the Rosenthal inequality (see Rosenthal (1970)) applied with the exponent  $\theta p \geq 2$  yields

$$\mathbb{E} \left( |\widehat{F}(x) - F(x)|^{\theta p} \right) = n^{-\theta p} \mathbb{E} \left( \left| \sum_{i=1}^n U_i(x) \right|^{\theta p} \right) \leq C n^{-\theta p} n^{\theta p/2} \leq C n^{-\theta p/2}.$$

Combining the inequalities above, we obtain the desired result, i.e.,

$$\mathbb{E}(\|\widehat{g} - g\|_p^p) \leq C(\varphi_{n,p} + n^{-\theta p/2}) \leq C \max(\varphi_{n,p}, n^{-\theta p/2}).$$

□

### 3.1 An illustrative application

Let's recall Example 1.2, where  $\{X_i\}_{i \in \mathbb{N}^*}$  is a sequence of *i.i.d.* random variables with pdf  $f$  and  $N$  be a discrete random variable of values in  $\mathbb{N}^*$  independent of this sequence. The density of  $Y = \max(X_1, \dots, X_N)$  is given by (1.2).

Suppose that Assumptions A.1-A.2 hold. Thus, several examples for the distribution of  $N$  can be considered.

(a) Degenerate distribution.  $\mathbb{P}(N = m) = 1$ . Then

$$w(u) = m u^{m-1}, \quad u \in [0, 1], \quad (3.4)$$

(b) Geometric distribution.  $N \sim G(\eta)$  ( $\mathbb{P}(N = k) = \eta(1 - \eta)^{k-1}$ ,  $k \in \mathbb{N}^*$ ). Then

$$w(u) = \frac{\eta}{(1 - u(1 - \eta))^2}, \quad u \in [0, 1], \quad (3.5)$$

(c) Poisson plus 1 distribution.  $N = P + 1$  with  $P \sim \mathcal{P}(\lambda)$  ( $\mathbb{P}(N = k) = e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!}$ ,  $k \in \mathbb{N}^*$ ). Then

$$w(u) = e^{-\lambda} e^{\lambda u} (1 + \lambda u), \quad u \in [0, 1].$$

**Remark 3.2.** In examples (a)–(c) above it is clear that Assumption A.3 is satisfied with  $\theta = 1$ ; more precisely, in example (a), we have  $C_2 = m(m - 1)$ ; in example (b), we have  $C_2 = 2(1 - \eta)/\eta^2$ ; in example (c), we have  $C_2 = \lambda(2 + \lambda)$ .

In this context, Theorem 3.1 can be applied. Let  $p \geq 2$  and  $\widehat{g}$  be the estimator given in (2.3). Then, for any  $f \in B_{q,r}^s(M)$ ,  $q \geq 1$ ,  $r \geq 1$  and  $s \in (1/q, T)$  there exists a constant  $C > 0$  such that

$$\mathbb{E}(\|\widehat{g} - g\|_p^p) \leq C \varphi_{n,p},$$

where  $\varphi_{n,p}$  is given by (3.3).

**Remark 3.3.** Taking  $m = 2$ , the obtained rate is similar to the one attained by the kernel estimators developed by Chen and Hsu (2004); the only difference is the extra-logarithmic term  $(\log n)^{2s/(2s+1)}$ . However, unlike kernel estimators Chen and Hsu (2004), our procedure  $\widehat{g}$  is adaptive and our rate of convergence holds for a wider class of functions  $f$  including Hölder class, Sobolev class, etc..

## 4 Simulation results

We now illustrate these theoretical results by a simulation study within the context described in Section 3.1. That is, we consider the problem of estimating the density  $g$  of the maximum of a random number  $N$  of *i.i.d.* random variables. From a reliability study standpoint, this problem corresponds to a parallel system with  $N$  identical components. Thereby, we have considered two numerical examples. They complement the asymptotic results of 3.1.

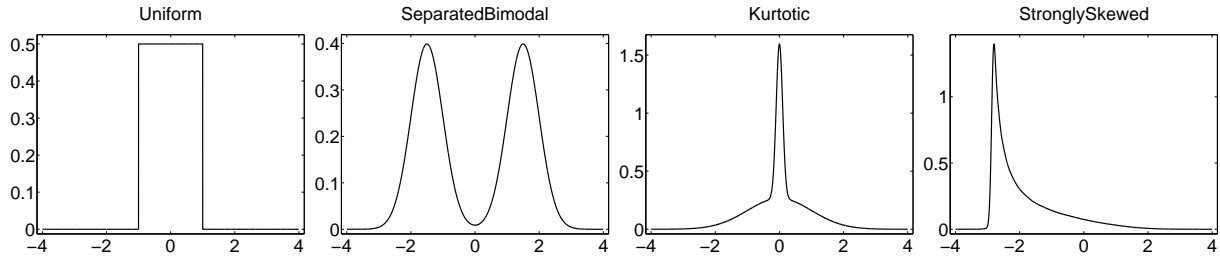


Figure 1: Test densities.

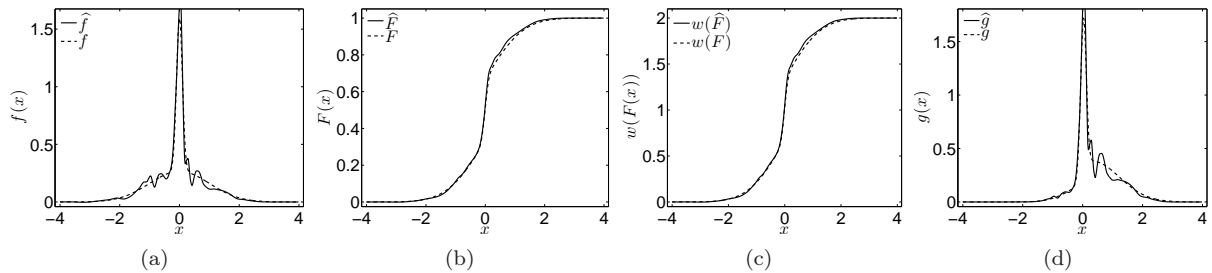


Figure 2: Typical reconstructions from a single simulation with  $n = 1000$  for the Kurtotic density. The dashed line depicts the original density and the solid one depicts its wavelet block estimate. (a):  $\hat{f}(x)$ . (b):  $\hat{F}(x)$ . (c):  $w(\hat{F}(x))$ . (d):  $\hat{g}(x) = w(\hat{F}(x))\hat{f}(x)$

**Computational aspects.** In the sequel, we will refer to our adaptive wavelet estimator (2.3) simply as Block. We have compared its performance to alternatives from the literature on several densities. We have considered the uniform distribution, as well as a family of normal mixture densities (“SeparatedBimodal”, “Kurtotic” and “StronglySkewed”, initially introduced in Marron and Wand (1992)) representing different degrees of smoothness (see Figure 1). We assumed that the density function  $f$  of the  $X_i$ ’s has a compact support included in  $[-b, b]$ . We have used the formulae given by Marron and Wand (1992) to simulate such densities so that

$$\min_l(\mu_l - 3\sigma_l) = -3, \quad \max_l(\mu_l + 3\sigma_l) = 3,$$

where  $l = 1, \dots, d$  with  $d$  the number of densities in the mixture (see, (Marron and Wand, 1992, Section 4, Table 1), for the values of the parameters). Thereby, it is very unlikely to have values outside the interval  $[-4, 4]$  and we loose little by assuming compact support. The same kind of assumption was made in the context of wavelet density estimation by Vannucci and Vidakovic (1997). In order to simplify the presentation of the results, one can simply rescale the data such that they fall into  $[-b, b]$  (which covered the full range of all observed data). Thus, the density was evaluated at  $T = 2^J$  equispaced points  $t_i = 2ib/T$ ,  $i = -T/2, \dots, T/2 - 1$  between  $-b$  and  $b$ , where  $J$  is the index of the highest resolution level and  $T$  is the number of discretization points. The primary level  $j_1 = 3$ ,  $T = 512$  and the Symmlet wavelet with 6 vanishing moments were used throughout all experiments. All simulations have been implemented under Matlab.

**Results and discussion.** In order to illustrate Theorem 3.1, we study the influence of  $p$  on the numerical performances of the Block and the term-by-term ( $L = 1$ ) thresholding estimator. Let us first consider a parallel system with  $m = 2$  identical independent components. Then, the corresponding weighted function is (3.4) and the goal is to estimate  $g$  in (1.2) from  $X_1, \dots, X_n$  sample simulated from one of the test densities. A typical example of estimation for the Kurtotic density (for  $p = 2$ ), with  $n = 1000$  is given in Figure 2. The mean  $L_p$  risk of  $\hat{g}$  i.e.,  $R_p(\hat{g}, g) = (1/T) \sum_{i=-T/2}^{T/2-1} |\hat{g}(t_i) - g(t_i)|^p$ , is obtained with 50 samples for  $n = 1000$ , and it is plotted as a function of  $p$  in Figure 3. As predicted by Theorem 3.1, the larger  $p$ , the smaller  $L_p$  risk of  $\hat{g}$ . We can see that our estimation procedure provides better results than the term-by-term thresholding ( $L = 1$ ) in all cases. In particular, the risk improvement achieved by the block estimators upon the term-by-term estimator is significant for the non-smooth Uniform density. This is in agreement with the predictions of our theoretical findings.

To conclude this first example, we illustrate from a single simulation, the fact that the parameters

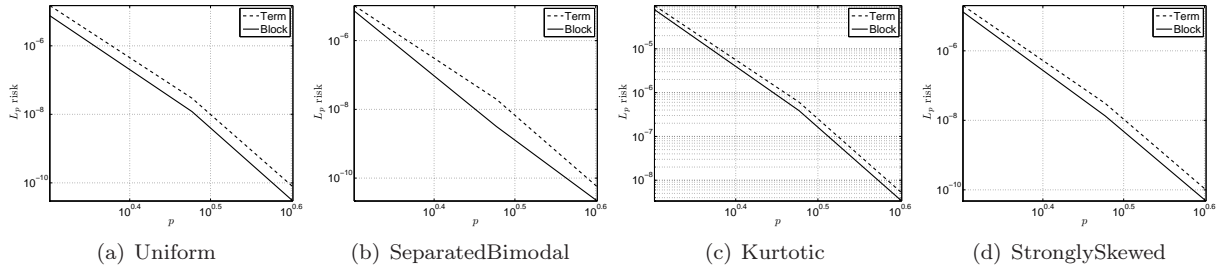


Figure 3: The influence of  $p$  in the numerical values of the  $\mathbb{L}_p$  risk (in a log-log scale) of Block (solid) and term-by-term (dashed) thresholding ( $L = 1$ ).

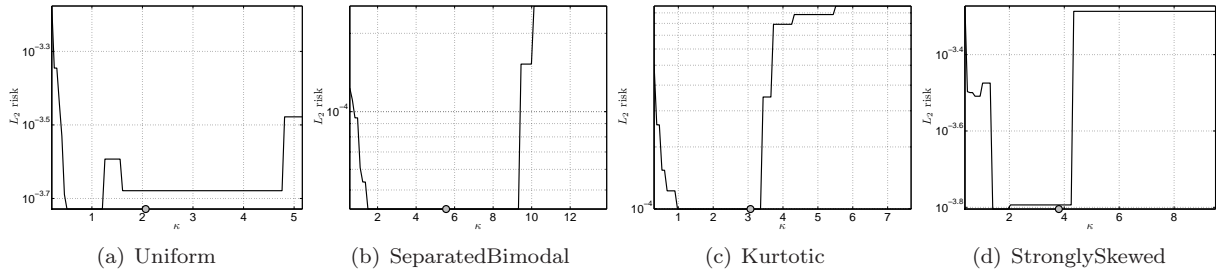


Figure 4:  $\mathbb{L}_2$  risk as a function of the threshold level  $\kappa$  (in a semi-log scale), the gray circle represents the universal threshold.

dictated by the theory yield the expected performance. We display in Figure 4 the empirical  $\mathbb{L}_2$  risk as a function of the threshold level  $\kappa$ , where the vertical dashed line represents the universal threshold. One can see that the minimum of the  $\mathbb{L}_2$  risk is close to the universal threshold for all test densities, thus supporting the choice dictated by our theoretical procedure, although derived in an asymptotic setting.

In our second example, the adaptive estimator described in Section 3.1 is tested when  $N$  follows a Geometric distribution, so that the weight function is that given by (3.5). This example is devoted to a simulation study comparing the performance of the block hard thresholding estimator with that of the traditional kernel defined as follows

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (4.1)$$

where the positive kernel  $K$  satisfies  $\int K(x)dx = 1$  and the smoothing parameter  $h$  is known as the bandwidth.

Many procedures of bandwidth selection for kernel density estimation have been developed in the literature (see, e.g., Silverman (1986)). We use least-squares cross-validation (LSCV) (Rudemo (1982), Bowman (1984)) where the bandwidth is defined as

$$h_{\text{LSCV}} = \underset{h}{\operatorname{argmin}} \int_{-b}^b \hat{f}_h(x)^2 dx - 2n^{-1} \sum_{i=1}^n \hat{f}_{-i}(X_i),$$

and  $\hat{f}_{-i}$  is the *leave-one out* kernel estimator constructed from the data without the observation  $X_i$ . It is motivated by the fact that for independent data

$$\text{LSCV}(h) = \int_{-b}^b \hat{f}_h(x)^2 dx - 2n^{-1} \sum_{i=1}^n \hat{f}_{-i}(X_i)$$

is an unbiased estimator of  $\text{MISE}(h) = \int_{-b}^b f^2(x)dx$ . One frequently used cross-validation (CV) procedure is the K-fold CV (as described e.g. in (Hastie et al., 2009, Section 7.10)) in which the data set  $X_1, \dots, X_n$  is randomly partitioned into K approximately equal-sized and non-overlapping subsets  $S_1, \dots, S_K$ . To obtain the bandwidths  $h_{\text{LSCV}}$ , we have performed a 10-fold CV, using a Gaussian kernel, with a simple “rule-of-thumb” pilot bandwidth  $h_{\text{ROT}}$ . Figure 5(b) contains a plot of the LSCV function versus the kernel bandwidth  $h$  and Figure 5(c) the estimated MISE as a function of  $h$ . For each density, it is clear



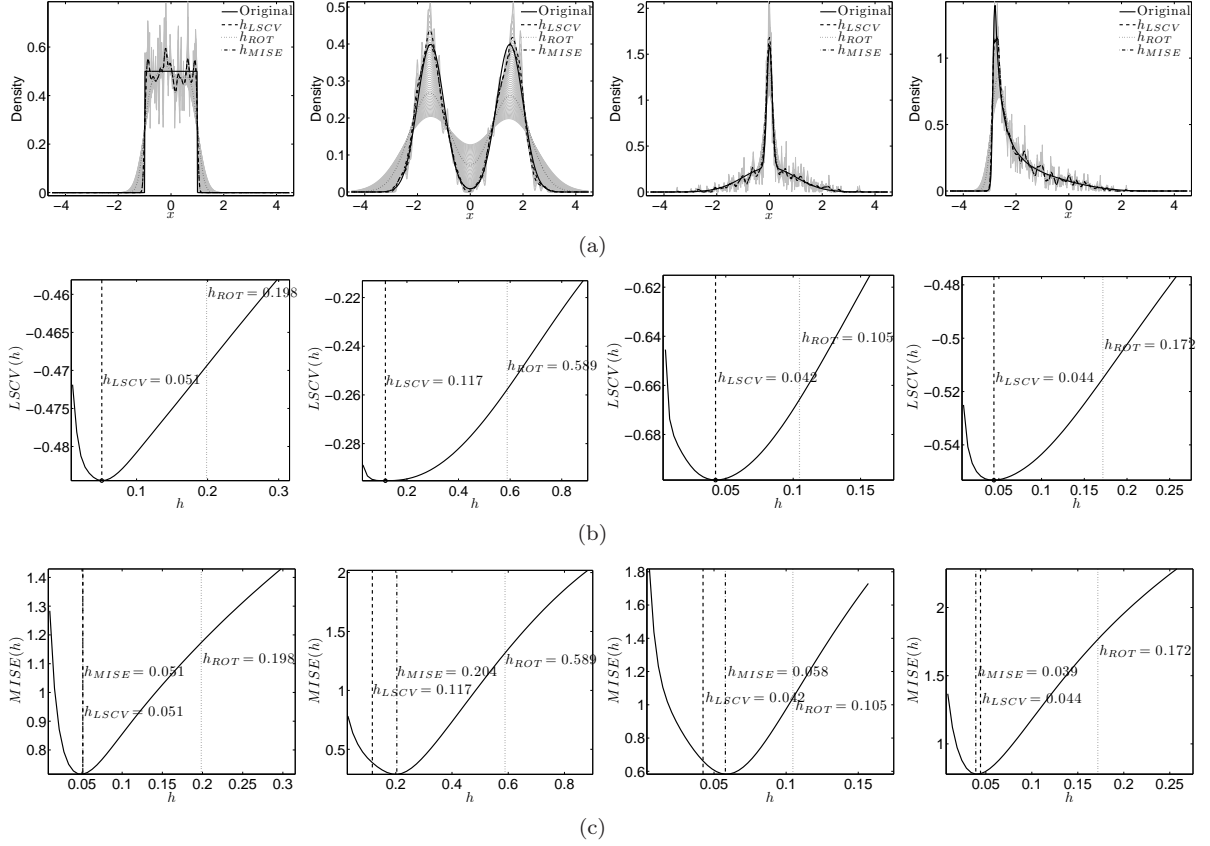


Figure 5: (a): Density estimates. (b): Graph of the the LSCV function versus the kernel bandwidth  $h$  for each of the tested densities, the vertical dashed lines represent the value of  $h$  that minimizes  $LSCV(h)$ . (c): The solid line depicts the estimated MISE as a function of  $h$ , the vertical dashed-dotted lines represent the true MISE-minimizing bandwidth  $h_{MISE}$  and the vertical dotted lines represent the pilot bandwidth  $h_{ROT}$ .

from this figure (Figure 5(b)), that the value of  $h_{LSCV}$  is the unambiguous minimizer of  $LSCV(h)$ . We see that  $h_{LSCV}$  provides a decent approximation, close to  $h_{MISE}$  for all test densities. For the StronglySkewed density, the bandwidth which minimizes  $MISE(h)$  in this case is  $h_{MISE} = 0.039$  and  $h_{LSCV} = 0.044$ . In this case, for the Uniform density,  $h_{MISE} = h_{LSCV} = 0.051$ .

We then compared the performance of the Block estimator  $\hat{g}$  with that of the plug-in kernel estimator, say  $\hat{g}_{LSCV}$ , given by  $\hat{g}_{LSCV} = w(\hat{F}(x))\hat{f}_{LSCV}(x)$ , where  $\hat{F}$  is defined by (2.4) and  $\hat{f}_{LSCV}$  is given by (4.1) with  $h_{LSCV}$ . Figure 6 shows the results of  $\hat{g}$  and  $\hat{g}_{LSCV}$  for  $N \sim G(\eta)$ , with  $\eta = 0.9$ ,  $\eta = 0.5$  and  $\eta = 0.1$  respectively. Table 1 presents the MISE for samples sizes  $n = 200, 1000, 2000$  and  $5000$ . For virtually all cases, the Block estimator consistently showed lower  $L_2$  risk than  $\hat{g}_{LSCV}$ , with the exception of the (very smooth) SeparatedBimodal density for which the kernel estimator performs slightly better. This comes at no surprise given that this density is very smooth. Additionally, small discrepancies in the estimate of  $f$  may lead to substantial discrepancies for the estimate of  $g$  at the locations overweighted by  $w(\hat{F}(\cdot))$ . It turns out that this is the case for the Geometric distribution where the weights evolve in  $O(1/\eta)$  at high values of  $x$ , and thus the discrepancies in  $\hat{g}$  increase as  $\eta$  gets smaller. However, the kernel estimator  $\hat{g}_{LSCV}$  seems to be more concerned (see, Figure 6(c)), confirming that Block generally provides a better estimate of  $f$ . Furthermore, as expected, for both methods, and in all cases, the MISE is decreasing as the sample size increases. Without any prior smoothness knowledge on the unknown density, the Block estimator provides very competitive results in comparison to  $\hat{g}_{LSCV}$ .

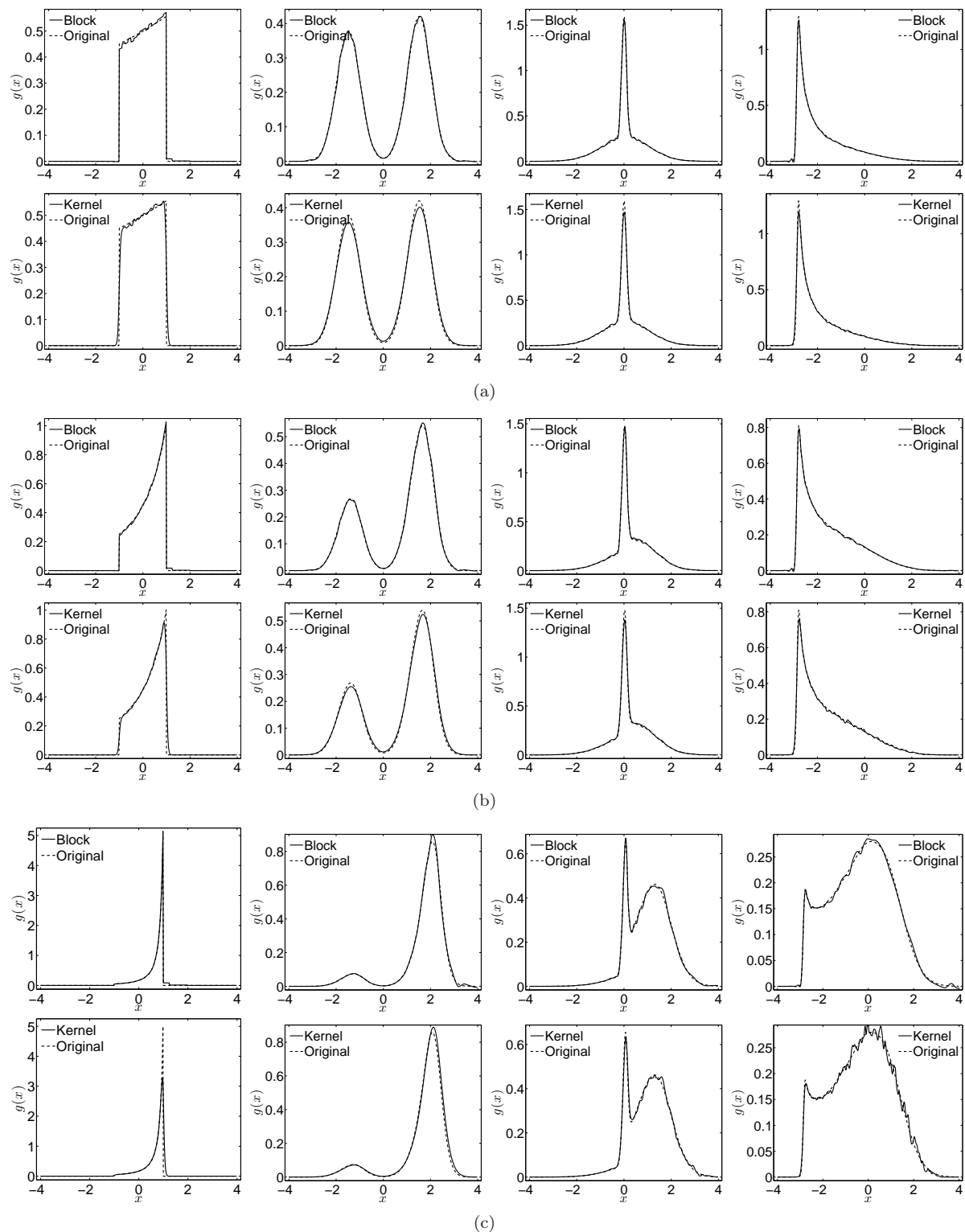


Figure 6: Original densities (dashed), Block thresholding estimator  $\hat{g}$  (solid) (1st row), kernel estimator  $\hat{g}_{\text{LSCV}}$  (solid) (2nd row) from 50 replications of  $n = 1000$  samples  $X_1, \dots, X_n$ . From left to right Uniform, SeparatedBimodal, Kurtotic and StronglySkewed.  $N \sim G(\eta)$ , with (a)  $\eta = 0.9$ , (b)  $\eta = 0.5$  and (c)  $\eta = 0.1$ .

Table 1:  $1000 \times \text{MISE}$  values from 50 replications of sample sizes  $n = 200, 1000, 2000$  and  $5000$ , when  $N$  follows a Geometric distribution of parameter  $\eta$ .

Uniform												
$\eta = 0.9$				$\eta = 0.5$				$\eta = 0.1$				
$n$	200	1000	2000	5000	200	1000	2000	5000	200	1000	2000	5000
Block	50.04	10.49	7.03	4.18	75.90	11.15	7.61	4.85	102.12	19.15	15.05	13.39
Kernel	47.31	13.76	11.37	9.32	77.52	17.62	14.99	12.60	109.00	70.50	63.24	55.94

SeparatedBimodal												
$\eta = 0.9$				$\eta = 0.5$				$\eta = 0.1$				
$n$	200	1000	2000	5000	200	1000	2000	5000	200	1000	2000	5000
Block	19.63	8.53	6.29	3.69	20.47	9.09	6.89	3.90	62.15	15.27	11.08	7.12
Kernel	14.56	6.30	4.98	3.54	16.30	7.00	5.32	3.82	44.01	13.26	10.86	7.79

Kurtotic												
$\eta = 0.9$				$\eta = 0.5$				$\eta = 0.1$				
$n$	200	1000	2000	5000	200	1000	2000	5000	200	1000	2000	5000
Block	49.73	11.31	8.03	5.52	52.37	11.93	8.04	5.57	98.03	18.08	9.22	7.01
Kernel	45.12	12.27	8.00	5.83	49.43	13.03	8.44	6.13	117.37	21.97	13.04	9.66

StronglySkewed												
$\eta = 0.9$				$\eta = 0.5$				$\eta = 0.1$				
$n$	200	1000	2000	5000	200	1000	2000	5000	200	1000	2000	5000
Block	43.73	9.91	7.78	5.12	36.39	8.69	7.02	4.70	88.49	10.12	9.03	5.93
Kernel	47.71	10.57	8.13	5.97	44.65	11.16	8.68	6.24	128.30	19.62	15.54	10.86

## References

- Adamidis, K., Dimitrakopoulou, T. and Loukas, S. (2005), "On an extension of the exponential-geometric distribution," *Statist. Probab. Lett.*, 73(3),259–269.
- Antoniadis, A. (1997), "Wavelets in statistics: a review (with discussion)," *J. Italian Statistical Society*, 6(2), 97–144.
- Bowman, M. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 71(2), 353–360.
- Cai, T. (1999), "Adaptive wavelet estimation: a block thresholding and oracle inequality approach," *The Annals of Statistics*, 27(3), 898–924.
- Cai, T. (2002), "On block thresholding in wavelet regression: Adaptivity, blocksize and threshold level," *Statistica Sinica*, 12, 1241–1273.
- Cai, T. and Zhou, H. (2009), "A data-driven block thresholding approach to wavelet estimation," *Ann. Stat.*, 37, 569–595.
- Chen, S.-M. and Hsu, Y.-S. (2004), "Kernel Density Estimations for Maximum of Two Independent Random Variables," *J. Nonparametr. Statist.*, 16(6), 901-924.
- Chesneau, C. (2008), "Wavelet estimation via block thresholding : a minimax study under  $L^p$  risk," *Statistica Sinica*, 18(3), 1007–1024.
- Chesneau, C. (2010), "Wavelet block thresholding for density estimation in the presence of bias," *Journal of the Korean Statistical Society*, 39(1), 43–53.
- Comte, F. and Rebafka, T. (2012), "Adaptive density estimation in the pile-up model involving measurement errors," *Electron. J. Statist.*, 2, 2002–2037.

- DeVore, R. and Popov, V. (1988), "Interpolation of Besov spaces," *Trans. Amer. Math. Soc.*, 305(1), 397–414.
- Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1996), "Density estimation by wavelet thresholding," *Ann. Statist.*, 24(2), 508–539.
- Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998), *Wavelet, Approximation and Statistical Applications*, Lectures Notes in Statistics, New York 129, Springer Verlag.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *Elements of Statistical Learning: Data Mining, Inference and Prediction* (Second Edition). Springer-Verlag, New York.
- Juditsky, A. and Lambert Lacroix, S. (2004), "On minimax density estimation on  $\mathbb{R}$ ," *Bernoulli*, 10, 187–220.
- Li, L. and Xiao, Y. (2008), "On the minimax optimality of block thresholded wavelet estimators with longmemory data," *Journal of Statistical Planning and Inference*, 137(9), 2850–2869.
- Li, L. (2008), "On the block thresholding wavelet estimators with censored data," *Journal of Multivariate Analysis*, 99(8), 1518–1543.
- Louzada, F., Bereta, E. and Franco .M (2012), "On the Distribution of the Minimum or Maximum of a Random Number of *i.i.d.* Lifetime Random Variables," *Applied Mathematics*, 3(4), 350–353.
- Mallat, S. (2009), *A Wavelet Tour of Signal Proc.: The Sparse Way, 3rd edition*, Academic Press.
- Marron, J.S. and Wand, M.P. (1992), "Exact Mean Integrated Squared Error," *Ann. Statist.*, 20(2), 712–736.
- Meyer, Y. (1992), *Wavelets and Operators*. Cambridge University Press, Cambridge.
- Nason, G. (1996). "Wavelet shrinkage using cross-validation," *J. R. Statist. Soc. B*, 50, 463–479.
- Picard, D. and Tribouley, K. (2000), "Adaptive confidence interval for pointwise curve estimation," *Annals of Statistics*, 28(1), 298–335.
- Raghunandan, K and Patil, S.A. (1972), "On order statistics for random sample size," *Stat. Neerl.*, 26(4), 121–126.
- Rosenthal, H.P. (1970). "On the subspaces of  $\mathbb{L}^p$  ( $p \geq 2$ ) spanned by sequences of independent random variables," *Israel Journal of Mathematics*, 8(3), 273–303.
- Rudemo, M. (1982), "Empirical choice of histograms and kernel density estimates," *Scand. J. Statist.*, 9(2), 65–78.
- Shaked, M. (1975), "On the distribution of the minimum and of the maximum of a random number of *i.i.d.* random variables," *Statistical Distributions in Scientific Work*, 1, ed. G. P Patil, S. Kotz and J. K. Ord. Reidel, Dordrecht. 363–380.
- Shaked, M. and Wong, T. (1997), "Stochastic Comparisons of Random Minima and Maxima," *J. Appl. Prob.*, 34(2), 420–425.
- Silverman, B. W. (1986), "Density estimation: for statistics and data analysis," Chapman & Hall.
- Vannucci, M. and B. Vidakovic, "Preventing the dirac disaster: Wavelet based density estimation," *J. Italian Statistical Society*, 6(2), 145–159.
- Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc., New York, 384 pp.