ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Butucea, Cristina

Working Paper

Numerical results concerning a sharp adaptive density estimator

SFB 373 Discussion Paper, No. 1999,34

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Butucea, Cristina (1999) : Numerical results concerning a sharp adaptive density estimator, SFB 373 Discussion Paper, No. 1999,34, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10046225

This Version is available at:
https://hdl.handle.net/10419/61783

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Numerical results concerning a sharp adaptive density estimator

Cristina BUTUCEA
Humboldt Universität zu Berlin
SFB 373, Spandauer Strasse 1; D 10178 Berlin, Germany
& Paris 6 University, France
(E-mail butucea@ensae.fr, butucea@wiwi.hu-berlin.de)

Abstract

We give here a simulation study of a density estimator, issued from sharp adaptive estimation. This nonparametric estimator was previously proved to have interesting theoretical properties. In this paper we describe the method and apply it successfully to i.i.d. simulated data issued from different densities.

Keywords: pointwise density estimation, adaptivity, kernel estimator, Lepski's criterion, simulation study

1 Introduction

We consider $X_1, \ldots, X_n$, n independent, identically distributed observations with a common probability density $f : I\!R \to [0, +\infty)$. We want to estimate f at a real point x, by a modified kernel estimator, which was proved to be adaptive to the local smoothness of the density, in the sense described below, in Butucea (1999) [4].

Kernel estimators for densities were introduced by Rosenblatt (1956) and Parzen (1962). For a kernel function K (usually a symmetric density function) and a bandwidth $h > 0$, we define the kernel estimator

$$f_n(x, h, K) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{1}$$

For an introductory study in density estimation and further developments we refer to Devroye and Györfi (1985), Silverman (1986) and Scott (1992). It is well known that drastic improvements of kernel estimator are obtained by a good choice of the bandwidth, rather than of the kernel.

A modification in that sense of (1) consists in choosing variable bandwidths kernel estimators. There are two possibilities to vary the bandwidth. The first is to write $h = h(x)$, as a function of the estimation point. The corresponding estimator is known in the literature as the balloon estimator (see Terrell and Scott 1992 for a study of its improvements). A second possibility is the data-dependent variable bandwidth, giving the estimator

$$\widetilde{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{X_i - x}{h_i}\right). \tag{2}$$

Thus, Breiman, Meisel and Purcell (1977) proposed a bandwidth h_i proportional to $(f(X_i))^{-1}$. Abramson (1982) proposed h_i proportional to $(f(X_i))^{-1/2}$, for multivariate observations, and found better behavior of this estimator in pointwise estimation. Abramson's estimator works for densities that are bounded away from 0 and feasible estimators of $f(x)$ use preliminary estimators clipped away from 0 in the bandwidth expression.

A global implementation of Abramson's estimator was done by Hall and Marron (1988), but deficiencies of this methods were found in Terrell and Scott (1992), precisely for the Gaussian density a bias of $O\left((h/\log h)^2\right)$ instead of the expected $O\left(h^4\right)$. Those results were confirmed by Hall, Hu and Marron (1995), where the bias was explicitly quantified for laws with exponentially decreasing tails and was found to be much greater than for the case of polynomially decreasing tails of the probability distribution.

Various methods were explored, concerning the data-based variable bandwidth selectors. The cross-validation was much developed since it was introduced by Rudemo (1982). Stone (1984) proved that, for one fixed density, the adaptive kernel estimator having bandwidth obtained by cross-validation is asymptotically equivalent to the best kernel estimator having fixed bandwidth.

Other data-dependent bandwidths are based on cross-validation (see Hall and Marron 1988), local cross-validation (Hall and Schucany 1989, Mielniczuk, Sarda and Vieu 1989), smoothed cross-validation (Hall, Marron and Park 1992).

Different methods include solve-the-equation bandwidth (Sheather 1986 and Sheather and Jones 1991), bootstrap bandwidth (Sain and Scott 1996) and smoothed bootstrap bandwidth (Hazelton 1996). For a review of these methods and comparative theoretical and numerical studies see Berlinet and Devroye (1994), Wand and Jones (1995) and Jones, Marron and Sheather (1996).

An adaptive kernel estimator which aims at detecting underlying features in curves by looking at estimators with different levels of resolution (i.e. bandwidths) is introduced by Chaudhuri and Marron (1997).

In another approach, Devroye and Lugosi (1996) found a data-based bandwidth whose performance, in $\mathbb{L}_1$ error with respect to the optimal non random, global bandwidth and over the set of all densities (universal smoothing factor) is less than a factor 3. The method was improved in Devroye and Lugosi (1997) and Devroye, Lugosi and Udina (1998) to that it provides non asymptotic bounds and it allows us to select automatically the bandwidth and the kernel order. This proves the supremacy of data-driven smoothing factor over the fixed bandwidth.

Recently, Devroye and Lugosi (1998) showed that optimization between all bandwidths depending on data can not be solved. Even when restriction over the class of estimated densities is done, the class of variable kernel estimates is too large to be optimized when comparing their $\mathbb{L}_1$ errors.

Here, we implement the exact adaptive estimation procedure (adaptive to the smoothness of the density) described in next section. This estimator is issued from kernel estimators via the Lepski type adaptation procedure (introduced by Lepskii 1990). The bandwidth selector is both local, i.e. for estimation at a point x the bandwidth depends on x, and data-dependent. Note also that in our method the choice of K depends on x and on the data, but it is matched with the choice of h. In this paper, we give a simulation study of this sharp pointwise adaptive density estimator.

2 Estimation procedure

The estimation framework is nonparametric, which means that the density to estimate belongs to a very large class of functions. In our case, these functions are mainly described by a smoothness parameter β (e.g., the number of continuous derivatives). More precisely, we consider, for β a positive integer and $L > 0$ the class

$$W(\beta, L) = \{f : I\!R \to (0, \infty) : \int_{I\!R} f = 1, \int_{I\!R} \left(f^{(\beta)}(x)\right)^2 dx \leq L^2\}.$$

This class can be generalized for non-integer $\beta > 1/2$, by introducing the Fourier transform $\mathcal{F}(f)(x) = \int_{I\!R} f(y)e^{-ixy}\, dy$, for any x in $I\!R$. The $\mathbb{L}_2$ Sobolev class, allowing non-integer values of $\beta > 1/2$ can be written as

$$W(\beta, L) = \{f : I\!R \to (0, \infty) : \int_{I\!R} f = 1, \int_{I\!R} |\mathcal{F}(f)(x)|^2 |x|^{2\beta} dx \leq 2\pi L^2\}.$$

In a slightly different way to Abramson (1982), we need the additional assumption that the density is bounded away from 0 at the estimation point. We shall write that our densities belong to $W_n(\beta, L)$,

$$W_n(\beta, L) = \{f \in W(\beta, L) : f(x) \geq \rho_n\},$$

where ρ_n is a sequence of positive real numbers that satisfies

$$\lim_{n\to\infty} \rho_n = 0 \text{ and } \liminf_{n\to\infty} (\rho_n \log n) > 0.$$

As we study the estimator asymptotically, this means that nearer to 0 is the estimated value, more observations are needed. In practice we take $\rho_n = 1/\log n$.

The quality of the approximation of an estimator $f_n(x)$ of $f(x)$ at fixed x shall be quantified by the **maximal risk** over $W_n(\beta, L)$

$$R_{n,\beta}(f_n, \psi_{n,\beta}) = \sup_{f \in W_n(\beta,L)} E_f \left[\psi_{n,\beta}^{-q} |f_n(x) - f(x)|^q\right], \tag{3}$$

where $q > 0$.

We are interested in finding the asymptotically best estimator f_n^* independent of β, provided that β belongs to a set B_{N_n}. More precisely, the set $B_{N_n} = \{\beta_1, ..., \beta_{N_n}\}$ is such that $1/2 < \beta_1 < \ldots < \beta_{N_n} < +\infty$ and more technical conditions have to be verified. For example, an equidistant grid of points with the largest value β_{N_n} tending to ∞ when $n \to \infty$, slower than $\log n$, is fulfilling this conditions.

Let us define $\overline{B} = B_{N_n} \setminus \{\beta_{N_n}\}$ and for all β in $\overline{B}$

$$\psi_{n,\beta} = \begin{cases} a\left(\frac{\log n}{n}\right)^{\frac{\beta-1/2}{2\beta}}, \text{ for } \beta \in \overline{B} \\ \left(\frac{1}{n}\right)^{\frac{\beta-1/2}{2\beta}}, \text{ for } \beta = \beta_{N_n} \end{cases}, \tag{4}$$

where the constant a is function of β, L, q and $f(x)$, $a = a(\beta, L, q, f(x)) > 0$ and satisfies

$$0 < \liminf_{\beta\to\infty} a(\beta, L, q, f(x_0)) \sqrt{\beta} \leq \limsup_{\beta\to\infty} a(\beta, L, q, f(x_0)) \sqrt{\beta} < \infty.$$

Theorem 1 *(Butucea, 1999) There exist an estimator f_n^* independent of β in B_{N_n} and an explicit constant $a = a(\beta, L, q, f(x)) > 0$ associated to $\psi_{n,\beta}$ in (4), such that*

$$\lim_{n\to\infty} \inf_{f_n} \sup_{\beta\in\overline{B}} R_{n,\beta}\left(f_n, \psi_{n,\beta}\right) = \lim_{n\to\infty} \sup_{\beta\in\overline{B}} R_{n,\beta}\left(f_n^*, \psi_{n,\beta}\right) = 1$$

and

$$\limsup_{n\to\infty} R_{n,\beta_{N_n}}(f_n^*, \psi_{n,\beta_{N_n}}) < \infty.$$

This result translates in density estimation the results for Gaussian white noise model by Tsybakov (1998). Differently from the Gaussian white noise model, the density model that we considered is heteroscedastic. In particular, the variance of the kernel estimators described below is proportional to the unknown value $f(x)$, value that appears consequently in the expression of the asymptotic uniform risk and in the adaptive estimator $f_n^*(x)$. This is the reason for introducing a preliminary estimator in our estimation procedure and for considering the clipping of our densities in the class $W_n(\beta, L)$.

The theorem means that for all densities in our class, we can find the estimator $f_n^*(x)$, described in the next subsection, and the proper normalization such that its maximal risk converges to 1, uniformly in β over $\overline{B}$. The same procedure attains the faster optimal rate at the last point β_{N_n}. This is a sharp asymptotic result, because it distinguishes among estimators attaining the rate $\psi_{n,\beta}$. Theoretical properties of adaptation in β were briefly described in Butucea (1999) [5] and studied in more details in Butucea (1999) [4].

As in the most part of the literature, we shall consider $q = 2$ that corresponds to the mean squared-error criterion.

2.1 Adaptive estimator

We describe here the simulation algorithm of estimation at each point x of the support of the density. We define the set $B_{N_n} = \{1, 2, \ldots, 6\}$ of integer regularities β. For each β in B_{N_n} we compute the expressions:

$$k_\beta = \left(\frac{1}{\beta(2\beta-1)L^2}\right)^{1/(2\beta)},$$

$$b_\beta^2 = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{|u|^{2\beta}}{\left(1+|u|^{2\beta}\right)^2} du \text{ and } \nu_\beta^2 = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{1}{\left(1+|u|^{2\beta}\right)^2} du.$$

The pilot estimator:

We have to choose a preliminary estimator of $f(x)$. In simulations, the final estimator proves to be very little dependent on the choice of this pilot estimator. We use the kernel estimator $f_n(x)$ with bandwidth $h_n = 1/\log n$ and Gaussian kernel $K(x) = \exp\left\{-x^2/2\right\}/\sqrt{2\pi}$:

$$f_n(x) = f_n(x, h_n, K) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right),$$

see formula (1). For technical reasons, this value is clipped away from 0 at ρ_n and we set here $\rho_n = 1/\log n$:

$$a_n(x) = \max\left\{ f_n(x), \frac{1}{\log n} \right\},$$

(cf. Abramson's (1982) clipped estimator). The last quantity appears in the expression of the random bandwidth below.

The kernel estimator with regularity β:

The kernel is defined for each integer β as the function

$$K_\beta(x) = \frac{1}{\pi} \int_0^\infty \frac{\cos(xu)}{1+|u|^{2\beta}} du.$$

We remark that $\|K_\beta\|_2^2 = \nu_\beta^2 = (2\beta - 1) b_\beta^2$ and that we can write it as a finite sum, (see Gradshteyn and Ryzhik 1994, formula 3.738.2):

$$K_\beta(x) = \frac{1}{2\beta} \sum_{k=1}^{\beta} \exp\left\{ -|x| \sin\left[\frac{(2k-1)\pi}{2\beta} \right] \right\}$$
$$\cdot \sin\left\{ \frac{(2k-1)\pi}{2\beta} + |x| \cos\left\{ \frac{(2k-1)\pi}{2\beta} \right\} \right\}.$$

In particular, for $\beta = 1$ we get $K_\beta = \exp(-|x|)/2$ and for $\beta = 2$ we get the familiar Silverman's kernel. Figure 1 shows that the kernel is fluctuating more as β grows from 1 (the sharp peaked function) to 6 (where it may also take negative values). We introduce the bandwidth

$$\widehat{h}_{n,\beta}(x) = \widehat{h}_{n,\beta}(x, X_1, \ldots, X_n) = k_\beta \left(\frac{a_n(x) \log n}{n} \right)^{1/(2\beta)}$$

and define the kernel estimator

$$\widehat{f}_{n,\beta}(x) = f_n\left(x, \widehat{h}_{n,\beta}(x), K_\beta \right) = \frac{1}{n \widehat{h}_{n,\beta}(x)} \sum_{i=1}^{n} K_\beta\left(\frac{X_i - x}{\widehat{h}_{n,\beta}(x)} \right).$$

The Lepski type estimator of β and the final adaptive estimator:

At each estimation point x we choose among these kernel estimators as follows. Let

$$\widehat{\eta}_{n,\beta}(x) = \nu_\beta \sqrt{\frac{1}{\beta k_\beta}} \left(\frac{a_n(x) \log n}{n} \right)^{(\beta - 1/2)/(2\beta)}$$

and define

$$\hat{\beta} = \max\left\{ \beta \in B_{N_n} : \left| \widehat{f}_{n,\gamma}(x) - \widehat{f}_{n,\beta}(x) \right| \le \widehat{\eta}_{n,\gamma}(x), \ \forall \gamma \in B_{N_n}, \ \gamma \le \beta \right\}.$$

This is an iterative algorithm which was introduced by Lepskii (1990). It starts with the first value of our set B_{N_n}. It suffices to choose among the estimators $\left\{ \widehat{f}_{n,\beta}(x), \beta \in B_{N_n} \right\}$ the one corresponding to regularity $\widehat{\beta}$. The adaptive estimator is

$$f_n^*(x) = \widehat{f}_{n,\hat{\beta}}(x).$$

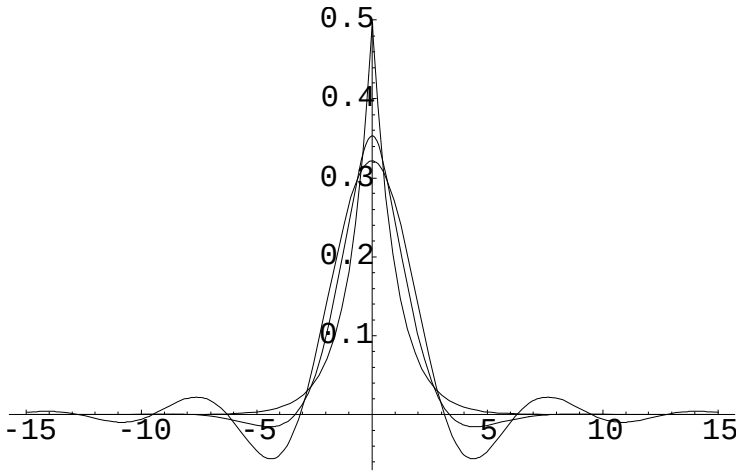

Figure 1: Kernels for $\beta = 1$, 2 and 6

In theory, we supposed L known, in simulations we put $L = 10$, but it may also be tuned in practice. Its choice plays a role only in the expression of the asymptotic constant, it does not affect the estimation rate.

2.2 Test densities

We consider the following test densities:

1. Standard Gaussian density $\varphi(0,1)$;
2. Mixed Gaussian densities $f(x) = 0.7\varphi(-2,1.5) + 0.3\varphi(2,0.5)$;
3. Cauchy density $f(x) = 1/\left(\pi\left(1+x^2\right)\right)$;
4. The extreme value distribution $f(x) = \exp\{-\exp\{-x\} - x\}$;
5. The logistic density $f(x) = \exp\{-x\}/(1+\exp\{-x\})^2$;
6. Laplace density (or the symmetrized exponential) $f(x) = \exp\{-|x|\}/2$;
7. The claw density (see Marron and Wand 1992, Berlinet and Devroye 1994)
$$f(x) = \frac{1}{10}\left[5\varphi(0,1) + \varphi(-1,0.1) + \varphi(-0.5,0.1) + \varphi(0,0.1) + \varphi(0.5,0.1) + \varphi(1,0.1)\right];$$
8. The smooth comb (see Marron and Wand 1992, Berlinet and Devroye 1994, Marron and Tsybakov 1995)
$$f(x) = \frac{32}{63}\varphi\left(-\frac{31}{21},\frac{32}{63}\right) + \frac{16}{63}\varphi\left(\frac{17}{21},\frac{16}{63}\right) + \frac{8}{63}\varphi\left(\frac{41}{21},\frac{8}{63}\right) + \frac{4}{63}\varphi\left(\frac{53}{21},\frac{4}{63}\right) + \frac{2}{63}\varphi\left(\frac{59}{21},\frac{2}{63}\right) + \frac{1}{63}\varphi\left(\frac{62}{21},\frac{1}{63}\right);$$

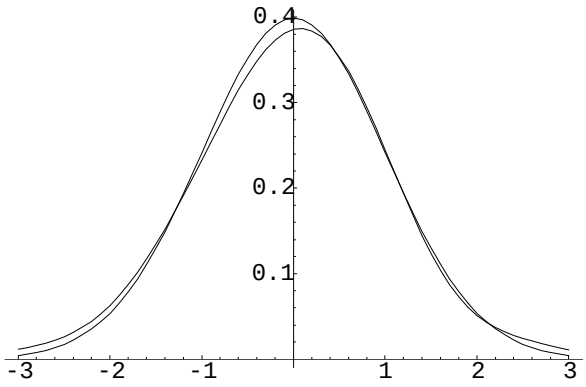

Figure 2: Adaptive estimator of the standard Gaussian density

9. The triangular density $f(x) = (1 - |x|)_+$

10. The saw tooth (see Berlinet and Devroye 1994),

$$g(x) = f(x+9) + f(x+7) + \ldots + f(x-7) + f(x-9),$$

where f is the previous triangular density, number 9.

3 Numerical results

We consider samples of size $n = 1000$ from each test density and a grid of estimation of step 0.1. The set $B_{N_n} = B$ is $\{1, \ldots, 6\}$. At each estimation point x, we compute the preliminary estimation and for each β in B the variable bandwidths and the corresponding kernel estimators. Then we estimate regularity $\widehat{\beta}$ of our density and give the adaptive estimator.

3.1 The Gaussian density

For the Gaussian density, $\widehat{\beta}(x)$ is constantly equal to 6 and the adaptive estimator in Figure 2 coincides with $\widehat{f}_{n,6}(x) = f_n\left(x, \widehat{h}_{n,6}(x), K_6\right)$. We shall denote from now on $\mathbb{L}_1 error$, $\mathbb{L}_2 error$ and $\mathbb{L}_\infty error$ the discrete norms $\mathbb{L}_1$, $\mathbb{L}_2$ and, respectively, $\mathbb{L}_\infty$ of the difference between the true density and its estimated value over the estimation grid. For $\widehat{f}_{n,6}(x)$ we get:

$$\mathbb{L}_1 error = 0.045,\ \mathbb{L}_2 error = 0.022 \text{ and } \mathbb{L}_\infty error = 0.0197.$$

Numerical comparison with Hall and Marron's (1988) variable kernel estimator was performed next. This method also uses a preliminary kernel estimator, f_n^1, with fixed bandwidth computed, for example, by Silverman's rule of thumb. The variable bandwidth is proportional to $(1/n)^{1/9} \sqrt{f_n^1(X_i)}$ at each observed point X_i and the kernel is Gaussian. Its estimated errors are

$$\mathbb{L}_1 error = 0.0764,\ \mathbb{L}_2 error = 0.0411 \text{ and } \mathbb{L}_\infty error = 0.0389.$$

As we know from Silverman (1986), we can compute precisely the $MISE\,(h)$ (Mean Integrated Squared Error) for kernel estimators with fixed, non random bandwidth of Gaussian densities. We shall write in a very similar way the Mean Squared Error $MSE(x,h)$, for a kernel estimator with square integrable kernel K and bandwidth h, at a point x, as follows

$$MSE\,(x,h) = \frac{\left(\left(K^h\right)^2 * f\right)(x)}{n} + \left(1-\frac{1}{n}\right)\left(K^h * f\right)^2(x) - 2f\,(x)\left(K^h * f\right)(x) + f^2\,(x)\,,$$

where $K^h\,(x) = K\,(x/h)\,/h$. This is precisely computed and plotted, for $n = 1000$ and standard Gaussian density number 1, in Figure 3. For each x of the estimation support, we proceed to minimization in h, and get the "oracle" bandwidth (which is supposed to know the right amount of smoothing corresponding to our optimality criterion)

$$h_{MSE}\,(x) = \arg\min_{h>0.01} MSE\,(x,h)\,.$$

We see in Figure 3 and more detailed on its superposed sections in Figure 4, that those functions have local minima. Our procedure takes for $h_{MSE}\,(x)$ the least $h > 0.01$ corresponding to the first local minimum of the $MSE\,(x,h)$. The value of $MSE\,(x, h_{MSE}\,(x))$ and the bandwidth $h_{MSE}\,(x)$ are plotted with continuous line in Figures 5 and 6, respectively. We remark that

$$\max_x h_{MSE}\,(x) = 1.1.$$

The value of $h_{MSE}\,(x)$ explodes at the inflection points $x = \pm 1$ of the Gaussian density. This corresponds well to the fact that the bandwidth which is optimal for the $AMSE$ (Asymptotic Mean Squared Error) criterion contains $f''\,(x)$ in the denominator (under the hypotheses that f'' is continuous).

We compare the "oracle" with the ideal adaptive bandwidth (containing the true density instead of the preliminary estimator):

$$h_{AD}\,(x) \overset{Def}{=} h_{n,6}\,(x) = k_6\left(\frac{\max\left\{f\,(x)\,, 1/\log n\right\}\log n}{n}\right)^{1/12},$$

since $\widehat{\beta}\,(x) = 6$, in simulations, for all x, and where $f\,(x)$ is the Gaussian density at point x. The $MSE\,(x, h_{AD}\,(x))$ is computed for a kernel estimator having Gaussian kernel instead of K_6 which is considered in the simulations. This function and $h_{AD}\,(x)$ are plotted with dots in Figures 5 and 6, respectively. Our bandwidth selector, $h_{AD}\,(x)$, and $h_{MSE}\,(x)$ are quite different and the mean squared error of our bandwidth does not always descend to the theoretic minimum. It is interesting however that the visual quality of our estimator does not suffer from that. Moreover, outside the neighborhoods of the inflection points, our selector reproduces remarkably well the optimal MSE curve.

The random bandwidth $\widehat{h}_{n,6}$ plotted in Figure 7 used in the adaptive estimator is quite close to the ideal $h_{AD}\,(x)$, despite the rough preliminary estimator that it contains.

Except for the vicinity of the inflection points, h_{AD} is greater than h_{MSE}, which is quite natural in view of the theoretical results: $\widehat{\beta}$ (respectively h_{AD}) with high probability

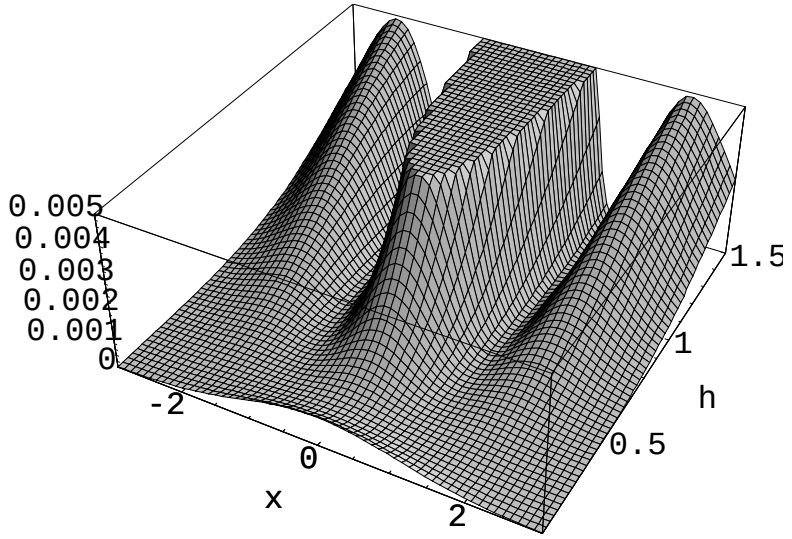

Figure 3: $MSE(x,h)$

overshoots the true value β (respectively h_{MSE}). The reason why this rule does not hold near the inflection points is that our set B is too small. We have to include into B the values that are much greater than 6 in order to track correctly the h_{MSE} near these points. It is not very profitable to do this since it would make the numerical computations very long for a very modest gain in the estimation accuracy.

We proceed to a Monte Carlo study. Let us consider 500 samples of size $n = 150$ of i.i.d. observations of Gaussian law. For each sample, we compute the kernel estimator having Gaussian kernel K and the "oracle" bandwidth h_{MSE} (corresponding to $n = 150$). We obtain vectors of size 500 of estimation errors for $f_n\left(x, h_{MSE}\left(x\right), K\right)$, see formula (1), and we consider the median value of each vector:

$$M\left(\mathbb{L}_1 error\right) = 0.094388,\ M\left(\mathbb{L}_2 error\right) = 0.0508582,\ M\left(\mathbb{L}_\infty error\right) = 0.048295.$$

We compute, for each sample, the adaptive estimator, $f_n\left(x, \widehat{h}_{n,\widehat{\beta}}\left(x\right), K_{\widehat{\beta}}\left(x\right)\right)$ and because of the low sample size these estimators are less regular, i.e. the estimated regularity $\widehat{\beta}\left(x\right)$ is no more constantly equal to 6. The median values of estimation errors are close to the previously obtained median errors:

$$M\left(\mathbb{L}_1 error\right) = 0.125978,\ M\left(\mathbb{L}_2 error\right) = 0.0721679,\ M\left(\mathbb{L}_\infty error\right) = 0.0954043.$$

Marron and Wand (1992) gave exact formulae for $MISE\left(h\right)$ of a kernel estimator, with Gaussian kernel function, as a function of the bandwidth h and in the case where the estimated densities are Gaussian mixtures. Similar studies can obviously be done on $MSE\left(x,h\right)$, for Gaussian mixtures and other densities. The practical problem that was already observed for $MISE\left(h\right)$ is that, as a function of h, $MSE(x,h)$ has more local minima and the study becomes tedious.

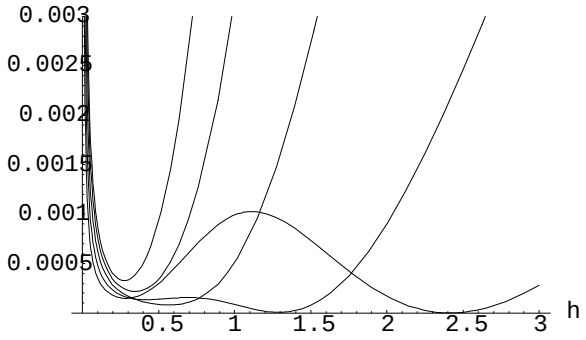

Figure 4: $MSE\left(x,h\right)$, for $x \in \{0.5, 0.75, 1, 1.25, 1.5\}$

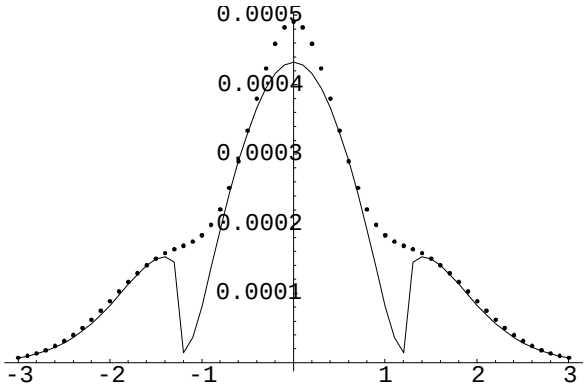

Figure 5: $MSE\left(x,h_{MSE}\left(x\right)\right)$ and $MSE\left(x,h_{AD}\left(x\right)\right)$

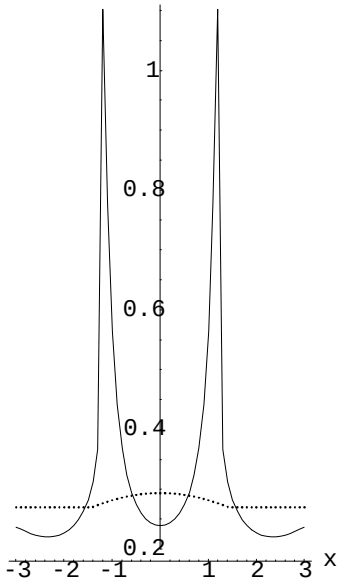

Figure 6: $h_{MSE}\left(x\right)$ and $h_{AD}\left(x\right)$

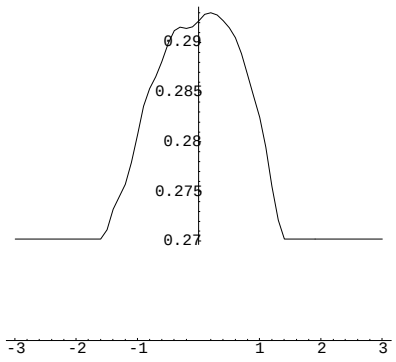

Figure 7: Random bandwidth of $\widehat{f}_{n,6}(x)$

3.2 Gaussian mixtures

For the Gaussian mixture density number 2 in Section 2.2, we compare on the same simulated sample, two preliminary kernel estimators with different bandwidths in Figures 8 and 11. We consider first a relatively "good" preliminary estimation and an obviously "bad" one. They have the following errors, respectively:

$$\mathbb{L}_1 error = 0.0865,\ \mathbb{L}_2 error = 0.0378 \text{ and } \mathbb{L}_\infty error = 0.0335$$
$$\mathbb{L}_1 error = 0.3116,\ \mathbb{L}_2 error = 0.127 \text{ and } \mathbb{L}_\infty error = 0.12.$$

The estimated regularities $\widehat{\beta} = \widehat{\beta}(x)$ are plotted as functions of x in Figures 9 and 12. We conclude that a change in the preliminary estimator has small influence on the quality of the final adaptive estimators in Figures 10 and 13, respectively, that have errors:

$$\mathbb{L}_1 error = 0.091,\ \mathbb{L}_2 error = 0.0369 \text{ and } \mathbb{L}_\infty error = 0.028$$
$$\mathbb{L}_1 error = 0.0855,\ \mathbb{L}_2 error = 0.0345 \text{ and } \mathbb{L}_\infty error = 0.0384.$$

Nevertheless, in order to get a smoother adaptive estimator, the pilot estimator should be chosen with some care.

An interesting example of mixture, difficultly estimated with kernel estimators, is the claw density number 7. Among the 6 kernel estimators $\widehat{f}_{n,\beta}$ with fixed regularity β, there is one that minimizes simultaneously the $\mathbb{L}_1$, $\mathbb{L}_2$ and $\mathbb{L}_\infty$ errors. It corresponds to $\beta = 2$, with respective errors

$$\mathbb{L}_1 error = 0.0902,\ \mathbb{L}_2 error = 0.0519 \text{ and } \mathbb{L}_\infty error = 0.0577.$$

The estimated $\widehat{\beta}$ is plotted in Figure 14. The corresponding adaptive estimator has the errors:

$$\mathbb{L}_1 error = 0.102,\ \mathbb{L}_2 error = 0.055 \text{ and } \mathbb{L}_\infty error = 0.0689.$$

This adaptive estimator is given in Figure 15 and is visually very satisfactory. It looks better than the estimator with fixed kernel $\beta = 2$, in spite of the fact that its $\mathbb{L}_1$, $\mathbb{L}_2$ and

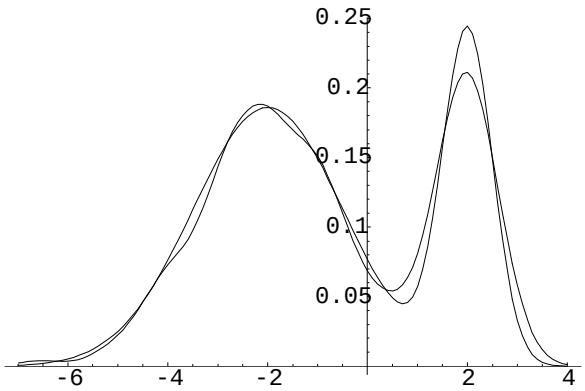

Figure 8: Good pilot estimator (1)

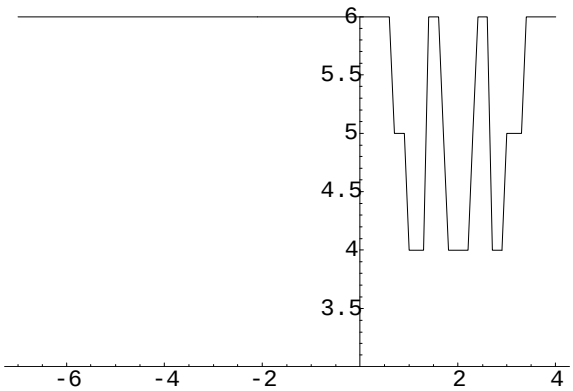

Figure 9: Regularity estimator (1)

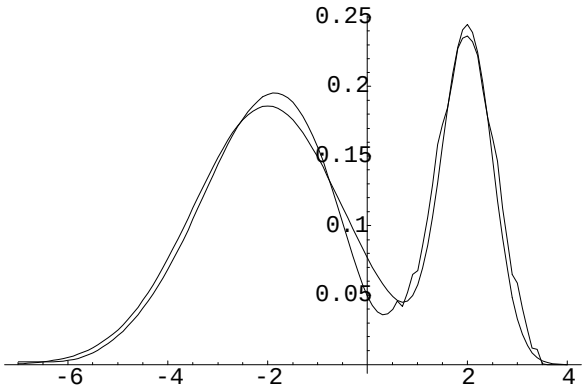

Figure 10: Adaptive estimator (1)

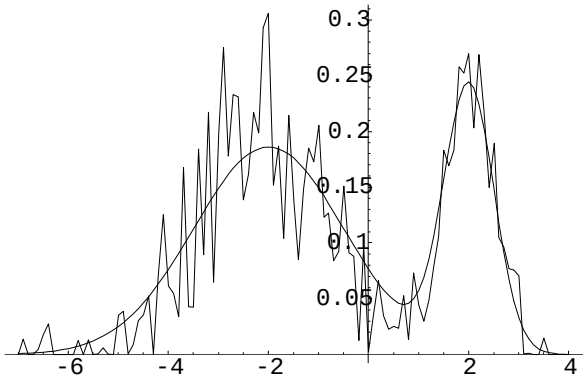

Figure 11: Bad pilot estimator

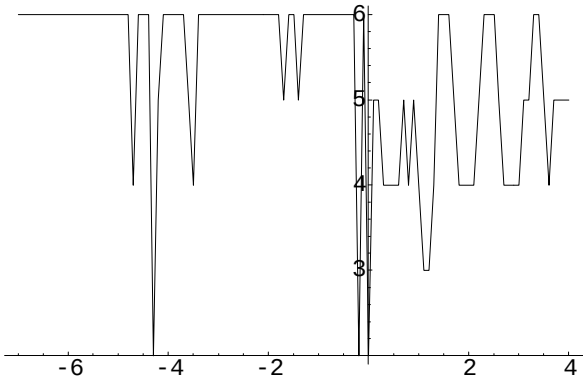

Figure 12: Regularity estimator (2)

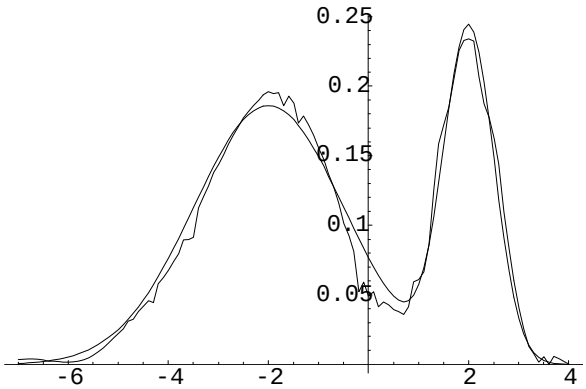

Figure 13: Adaptive estimator (2)

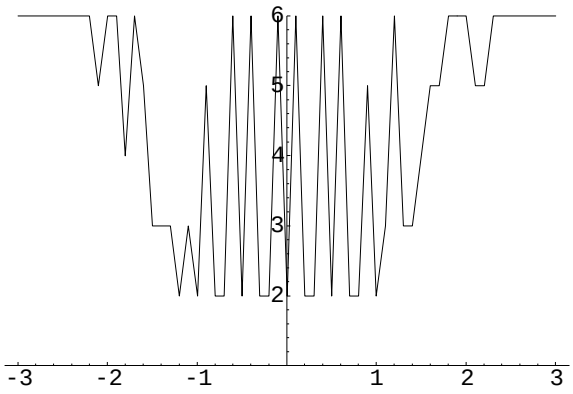

Figure 14: Regularity estimator - claw density

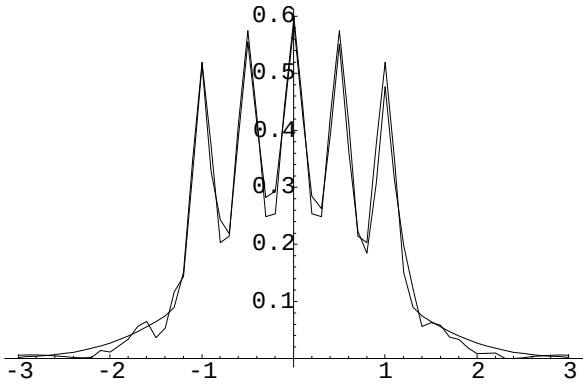

Figure 15: Adaptive estimator - claw density

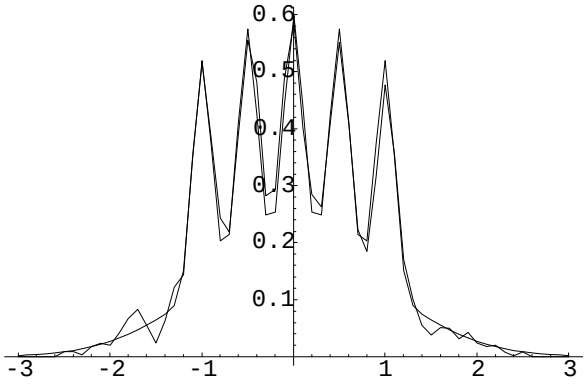

Figure 16: Kernel estimator with $\beta = 2$, $\widehat{f}_{n,2}$

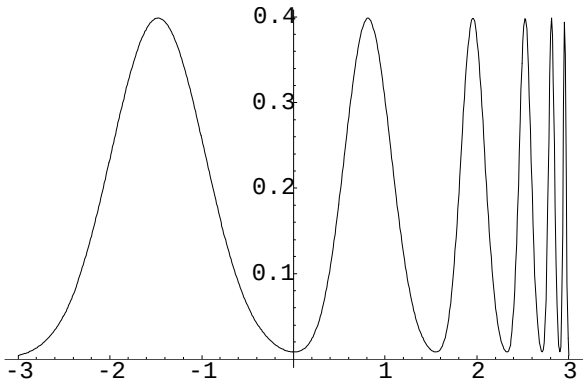

Figure 17: Smooth comb density, number 8

$\mathbb{L}_\infty$ errors are slightly greater. The kernel estimator with $\beta = 2$, $\widehat{f}_{n,2}$, is plotted in Figure 16.

Because of its very close peaks, the smooth comb density, number 8, was estimated on a grid with step 0.01 instead of 0.1, see Figure 17. We remark in this case that, among the estimators with fixed regularity β, the kernel estimator with $\beta = 2$ (Figure 18) minimizes errors

$$\mathbb{L}_1 error = 0.195 \text{ and } \mathbb{L}_2 error = 0.1178$$

and the kernel with $\beta = 3$ the error: $\mathbb{L}_\infty error = 0.17$. Because of the high irregularity of the estimator with $\beta = 1$ we started the regularity estimation from $\beta = 2$. For the estimation of the regularity also, we enlarged the confidence intervals associated to each kernel estimator by considering $1.3\eta_{n,\beta}$ instead of $\eta_{n,\beta}$ (see Section 2.1, The Lepski type estimator of β). This is done in order to increase the smoothness of the adaptive estimator. Its errors are

$$\mathbb{L}_1 error = 0.1487, \mathbb{L}_2 error = 0.088 \text{ and } \mathbb{L}_\infty error = 0.1823.$$

Plots of the estimated regularity and the corresponding adaptive estimator superposed to the theoretic underlying density (plotted with dots) are in Figure 19.

3.3 Other densities

We remark that the "worst-case" densities in the minimax sense (and Gaussian mixtures can reproduce them well) are not so badly estimated. On the other hand, functions like triangular distribution, number 9, (Figures 20 and 21) or saw tooth distribution, number 10, (Figure 22, $\widehat{\beta}(x) = 6$, for all x) have no particular difficulties except for a set of Lebesgue measure 0 and, however, our estimator seems not very preforming. Adaptive estimators errors are, respectively:

$$\mathbb{L}_1 error = 0.116, \mathbb{L}_2 error = 0.106, \mathbb{L}_\infty error = 0.211$$
$$\mathbb{L}_1 error = 0.14, \mathbb{L}_2 error = 0.0403, \mathbb{L}_\infty error = 0.0278.$$

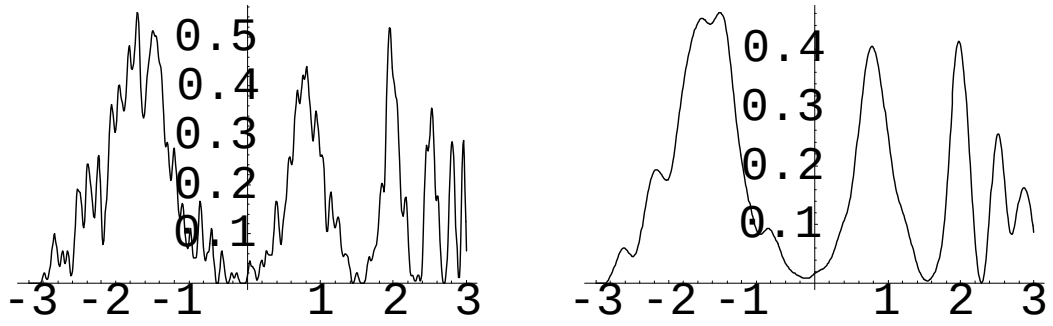

Figure 18: Kernel estimators of smooth comb density ($\beta = 2$ and $\beta = 3$, respectively)

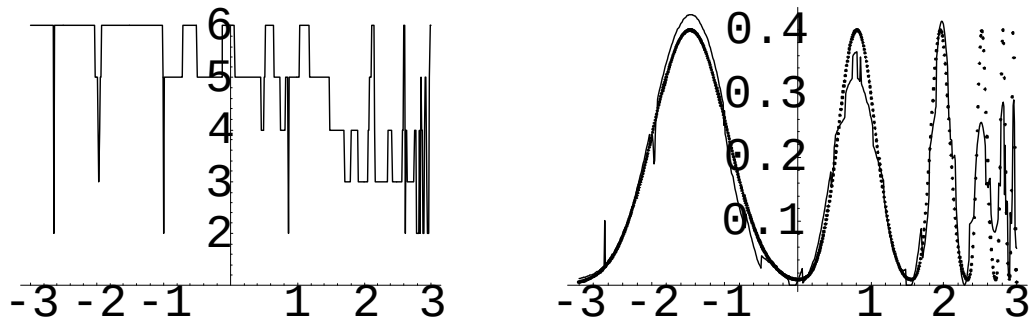

Figure 19: Regularity estimator and adaptive estmator, smooth comb density

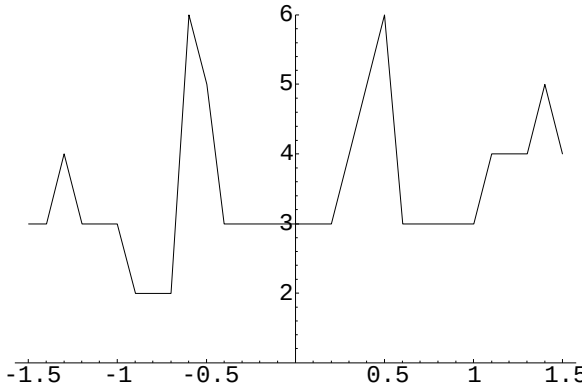

Figure 20: Regularity estimator, triangular density

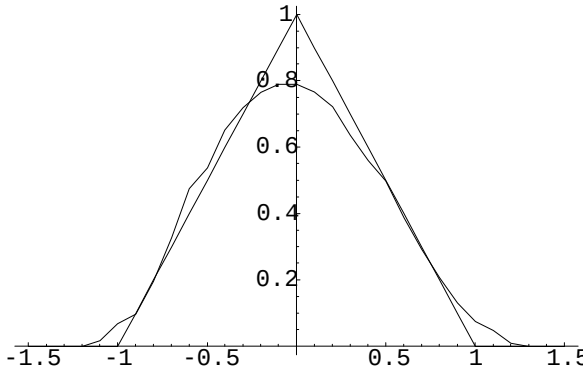

Figure 21: Adaptive estimator, triangular density

Such a behavior corresponds to the theoretical results since these functions have a low regularity β in a neighborhood of "difficult" points, and therefore their rate is smaller.

It is also recommendable that compactly supported functions, like these, or positive distributions should be estimated with one-sided kernels in order to avoid boundary effects. There are methods in the literature for obtaining such boundary kernels that we do not discuss here (see, e.g. Wand and Jones 1995). Also, one-sided kernels seem to be helpful in treating the vicinities of the "difficult" points where there is a jump of the derivative, in the case where these jump points of the derivative are known or expected.

Laplace density or the symmetric exponential, number 6, is also an extremely regular function except for a single point, a set of Lebesgue measure 0. This density is particularly hard to estimate since any usual method tends to oversmooth at the peak. Our method shares this drawback (see Figure 24), although the regularity plot, Figure 23, shows that in some neighborhood of 0 the empirical values of β have a slight tendency to drop down.

Good results are obtained for the Cauchy density, number 3, logistic density, number 5 and extreme value distribution, number 4. Their estimation errors are, respectively,

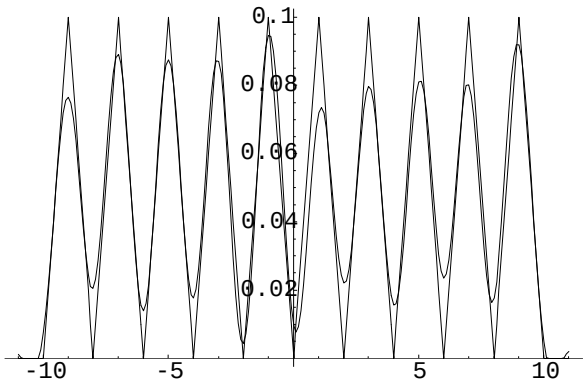

Figure 22: Adaptive estimator, saw tooth density

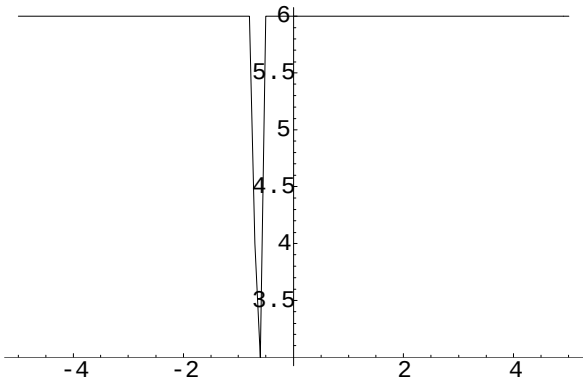

Figure 23: Regularity estimator, Laplace density

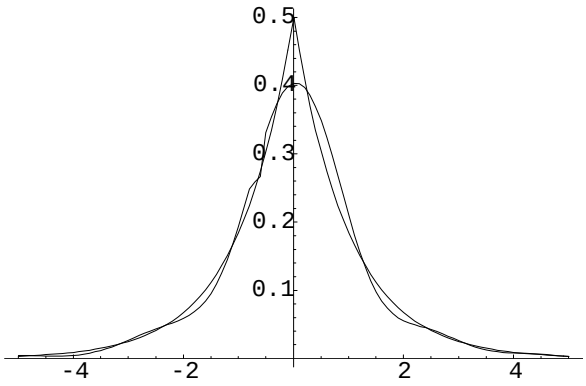

Figure 24: Adaptive estimator, Laplace density

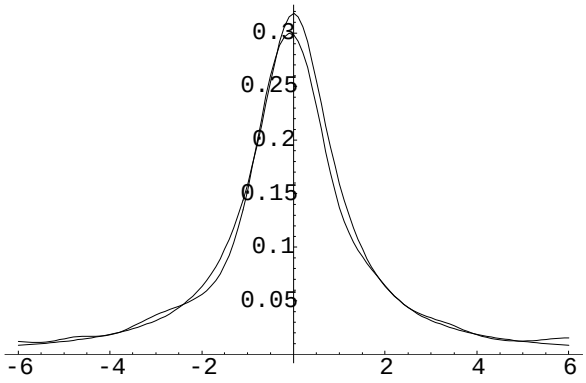

Figure 25: Adaptive estimator, Cauchy density

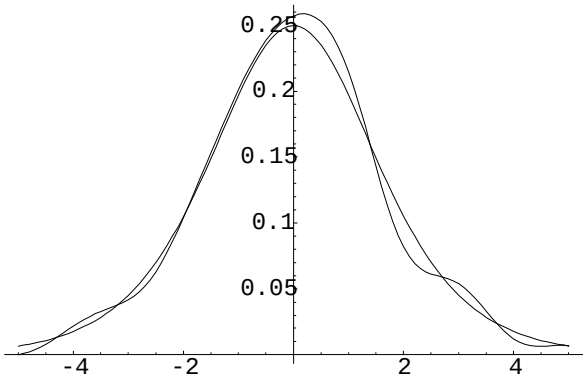

Figure 26: Adaptive estimator, logistic density

relatively small:

$$\mathbb{L}_1 error = 0.0672,\ \mathbb{L}_2 error = 0.0309,\ \mathbb{L}_\infty error = 0.0244$$
$$\mathbb{L}_1 error = 0.0702,\ \mathbb{L}_2 error = 0.0289,\ \mathbb{L}_\infty error = 0.0234$$
$$\mathbb{L}_1 error = 0.0295,\ \mathbb{L}_2 error = 0.0143,\ \mathbb{L}_\infty error = 0.0108.$$

Regularity estimators and adaptive estimators are in Figures 25, 26, 27, and 28, respectively. For the Cauchy and logistic density $\widehat{\beta}(x)$ is constantly equal to 6 and those graphics are skipped.

3.4 Further studies

We performed further studies and modifications of the sharp adaptive procedure, briefly detailed here. Adaptation was tried on larger sets of regularities, $B = \{1, \ldots, 10\}$, corresponding to choosing between more kernel estimators, on the Cauchy density.

Another modification was to consider a sharper grid on the set of regularities, like $B = \{0.75, 1, 1.25, 1.5, \ldots, 6\}$ for the logistic density, number 5 and Laplace density, number 6 (Figures 29 and 30). Here, we considered the kernel corresponding to rounded β (the

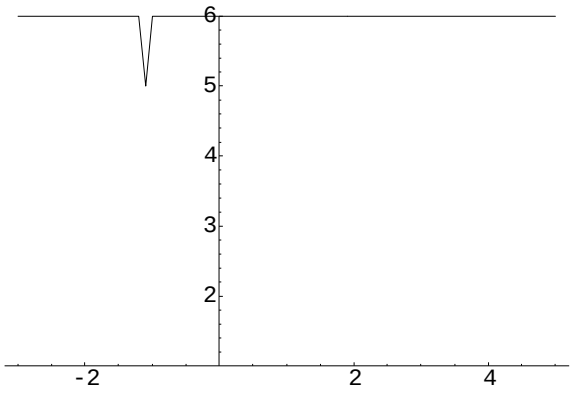

Figure 27: Regularity estimator, extreme value distribution

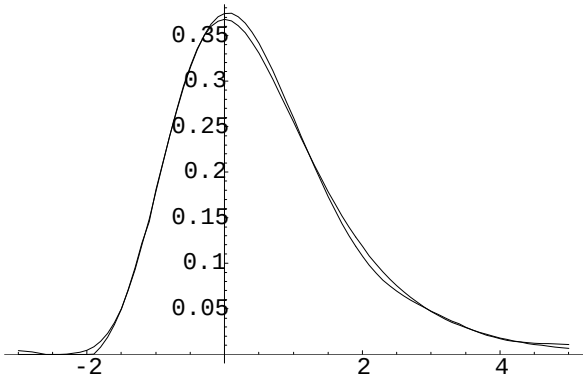

Figure 28: Adaptive estimator, extreme value distribution

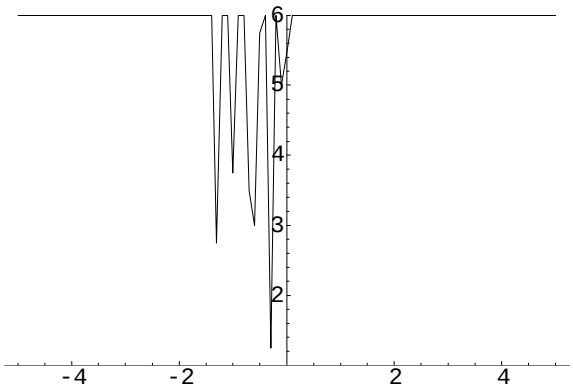

Figure 29: $\widehat{\beta}(x) \in B = \{0.75, 1, \ldots, 6\}$, Laplace density

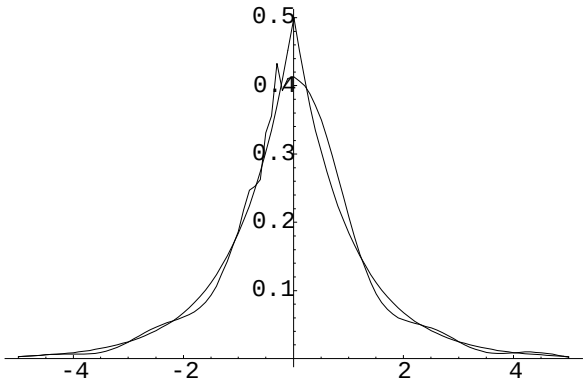

Figure 30: Adaptive estimator, Laplace density

closest integer value), since the sum expression for K_β in Section 4.2.1 holds for integer β. The kernel modification has, indeed, very small influence on the estimation, see Silverman (1986), Wand and Jones (1995).

Finally, the procedure was reiterated for the Gaussian mixture number 2, that means the procedure was started again having as a preliminary estimation the later adaptive estimator. All these modifications proved to be a plus in computing time, without significant improvements in the quality of the estimation.

4 Conclusion

The sharp adaptive density estimation method was designed in a framework of pointwise estimation. In theory, we search an estimation procedure which is uniformly well performing over Sobolev classes of functions and especially on their worst-case densities. In practice, the method works well for many densities, like Gaussian mixtures, extreme value distribution and other densities studied above. Besides, the adaptive estimator is particularly robust with respect to the preliminary estimation.

From a numerical point of view, we compared the $\mathbb{L}_1$, $\mathbb{L}_2$ and $\mathbb{L}_\infty$ distances between the true and the estimated function. There seems to be often a kernel which is the most performing in terms of those three considered distances. This kernel is sometimes the most regular kernel ($\beta = 6$) or Silverman's kernel ($\beta = 2$). The estimated regularity β varies around the very same value, which means that the adaptive procedure detects, indeed, the right kernel estimator.

Except for the Gaussian distribution, all our calculations were done for one given sample. Thus, the comparison with the ideal MSE are certainly subject to a random effect. However, they should be rather precise since the sample size is very large. A further study including Monte-Carlo simulations would be of interest here.

There are also other risk functions that may be used to measure the estimation quality. The $MISE$ was thoroughly studied in Silverman (1986), Wand and Jones (1995), the $MIAE$ (Mean Integrated Absolute Error) was studied by Devroye and Györfi (1985). Those risks are not always visually satisfactory and this has motivated the introduction of other risks, based on Hausdorff distances between graphs of functions, in Marron and

Tsybakov (1995). Here, we plotted the true and estimated functions and let to the reader's appreciation the visual estimation error.

The method is asymptotically efficient and it works well with high sample sizes. Nevertheless, for lower sample sizes, we should tune the parameters of our procedure. When looking for a rather smooth underlying density we may change our test between kernel estimators, for example, by considering $2\eta_{n,\beta}$ instead of $\eta_{n,\beta}$, in the definition of $\widehat{\beta}$. This will enlarge confidence intervals associated to each kernel estimator and allow us to stop at higher regularities β. On the contrary, less regular estimators will be obtained when this gauge is diminished. Adjustments of L go in the same direction, with less rapid results.

The clipping may be done at a lower level, for example at $1/n$ instead of $1/\log n$. This has an undersmoothing effect, since the bandwidth is proportional to this level.

In conclusion, the method works well for a large variety of densities and our results so far are satisfactory. We may think next of applying this method to real data, issued from different domains. Further theoretical studies should extend this method to the problem of dependent data like the estimation of marginal density of a discrete time mixing stationary process.

References

[1] Abramson, I.S. (1982) On bandwidth variation in kernel estimates-a square root law. *Ann. Statist.* **10** *1217-1223*

[2] Berlinet, A. and Devroye, L. (1994) A comparison of kernel density estimates. *Publ. Inst. Statist. Univ. Paris* **38** *3-59*

[3] Breiman, L., Meisel, W. and Purcell, E. (1977) Variable kernel estimates of multivariate densities. *Technometrics* **19** *135-144*

[4] Butucea, C. (1999) Nonparametric adaptive estimation of a probability density; rates of convergence, exact constant and numerical results *PhD Thesis, Paris 6 University*

[5] Butucea, C. (1999) Constante exacte adaptative dans l'estimation de la densité. *Comptes Rendus Acad. Sciences, to appear*

[6] Chaudhuri, P. and Marron, J.S. (1997) SiZer for exploration of structures in curves. *North Carolina Inst. of Statist. Mimeo Series 2355*

[7] Devroye, L. and Gyorfi, L. (1985) Nonparametric Density Estimation: The $\mathbb{L}_1$ View. *J. Wiley, New York*

[8] Devroye, L. and Lugosi, G. (1996) A universally acceptable smoothing factor for kernel density estimates. *Ann. Statist.* **24** *2499-2512*

[9] Devroye, L. and Lugosi, G. (1997) Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Ann. Statist.* **25** *2626-2637*

[10] Devroye, L. and Lugosi, G. (1998) Variable kernel estimates: on the impossibilty of tunning the parameters. Manuscript

[11] Devroye, L., Lugosi, G. and Udina, F. (1998) Inequalities for a new data-based method for selecting nonparametric density estimates. Manuscript

[12] Gradshteyn, I.S. and Ryzhik, I.M. (1994) Table of Integrals, Series and Products. *Academic Press 5th Edition*

[13] Hall, P., Hu, T.C. and Marron, J.S. (1995) Improved variable window kernel estimates of probabilty densities. *Ann. Statist.* **23** *1-10*

[14] Hall, P. and Marron, J.S. (1988) Variable window width kernel estimates of probabilty densities. *Probab. Th. Rel. Fields* **80** *37-49*

[15] Hall, P., Marron, J.S. and Park, B.U. (1992) Smoothed cross-validation. *Probab. Theory Rel. Fields* ***92*** *1-20*

[16] Hall, P. and Schucany W.R. (1989) A local cross-validation algorithm. *Statist. Probab. Letters* **8** *107-117*

[17] Hazelton, M. (1996) Bandwidth selection for local density estimators. *Scandin. Journal of Statist.* **23** *221-232*

[18] Jones, M.C., Marron, J.S. and Sheather, S.J. (1996) Progress in data-based bandwidth selection for kernel density estimation. *Comp. Statist.* ***11*** *337-381*

[19] Lepskii, O.V. (1990) On a problem of adaptive estimation in Gaussian white noise. *Theory Prob. Appl.* **35** *454-466*

[20] Marron, J.S. and Tsybakov, A.B. (1995) Visual Error Criteria for Qualitative Smoothing. *J. Amer. Statist. Assoc.* ***90*** *499-507*

[21] Marron, J.S. and Wand M.P. (1992) Exact mean integrated square error. *Ann. Statist.* **20** *712-713*

[22] Mielniczuk, J., Sarda, P. and Vieu Ph. (1989) Local data-driven bandwidth choice for density estimation. *J. Statist. Planning Infer.* **23** *53-69*

[23] Parzen, E. (1962) On the estimation of probability density function and mode. *Ann. Math. Statist.* **33** *1065-1076*

[24] Rosenblatt, M. (1956) On some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** *832-837*

[25] Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scandin. J. Statist.* ***9*** *65-78*

[26] Sain, S.R. and Scott, D.W. (1996) On locally adaptive density estimation. *J. Amer. Statist. Assoc.* **436** *1525-1534*

[27] Scott, D.W. (1992) Multivariate density estimation: theory, practice and visualization. *J. Wiley, New York*

[28] Sheather, S. (1986) An improved data-based algorithm for choosing the window width when estimating the density at a point. *Comp. Statist. Data Anal.* **4** *61-65*

[29] Sheather, S.J. and Jones M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc., Ser B* ***53*** *683-690*

[30] Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York

[31] Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* ***12*** *1285-1297*

[32] Terrell, G.R. and Scott, D.W. (1992) Variable kernel density estimation. *Ann. Statist.* **20** *1236-1265*

[33] Tsybakov, A.B. (1998) Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* ***26***

[34] Wand, M.P. and Jones, M.C. (1995) Kernel smoothing. *Chapman and Hall, London*