Comparison of kernel density estimators with assumption on number of modes

Raphaël Coudret, Gilles Durrieu, Jerôme Saracco

HAL Id: hal-00748790
https://hal.archives-ouvertes.fr/hal-00748790

Submitted on 6 Nov 2012

COMPARISON OF KERNEL DENSITY ESTIMATORS WITH ASSUMPTION ON NUMBER OF MODES

RAPHAËL COUDRET, GILLES DURRIEU, AND JÉRÔME SARACCO

ABSTRACT. A data-driven bandwidth choice for a kernel density estimator called critical bandwidth is investigated. This procedure allows the estimation to have as many modes as assumed for the density to estimate. Both Gaussian and uniform kernels are considered. For the Gaussian kernel, asymptotic results are given. For the uniform kernel, an argument against these properties is mentioned. These theoretical results are illustrated with a simulation study which compare the kernel estimators that rely on critical bandwidth with another one which uses a plug-in method to select its bandwidth. An estimator that consists in estimates of density contour clusters and takes assumptions on number of modes into account is also considered. Finally, the methodology is illustrated using environment monitoring data.

Keywords: Bandwidth, Kernel density estimator, Mode, Numerical study, Valvometry

1. INTRODUCTION

Since the seminal papers of Parzen (1962) and Rosenblatt (1956), the use of kernels to find an estimate $\hat{f}_{K,h}$ of a density function f of a random variable X is widely studied because of the advantages of the nonparametric point of view. Let $(X_1, \ldots X_n)$ be a vector of independent and identically distributed random variables generated from f. For $t \in \mathbb{R}$, the kernel density estimator $\hat{f}_{K,h}(t)$ of $f(t)$ can be defined as

$$\hat{f}_{K,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) \tag{1}$$

where K is the kernel and h is the bandwidth on which the amount of smoothness of $\hat{f}_{K,h}$ relies. Most of the time, K is a probability density function and h is a positive real. The larger the bandwidth, the smoother the estimate is. The choice of the bandwidth h is an important area in kernel estimators research field. Even if it exists a sufficiently large interval around the optimal bandwidth where $\hat{f}_{K,h}$ stays roughly the same (Scott (1992), p. 161), h needs to be carefully determined. To perform this

choice, one can use biased or unbiased cross-validation (Rudemo (1982); Scott and Terrell (1987)) as well as plug-in methods (Sheather and Jones (1991)) among other approaches.

In practical situations, the scientist that brings the data to analyze is able to determine if the estimated density function is smooth enough or not. In this paper, we are interested in using this information on the necessary amount of smoothing in order to set the corresponding bandwidth h for the estimator $\hat{f}_{K,h}$. More precisely, we will assume a fixed number $N(f)$ of modes of f. We will introduce and study the bandwidth $h_{crit,k}$ which is the smallest one such that the estimator $\hat{f}_{K,h}$ has $k \geq N(f)$ modes. Thus, the definition of $h_{crit,k}$ is:

$$(2) \qquad\qquad h_{crit,k} := \min_{N(\hat{f}_{K,h}) \leq k} h, \text{ with } k \geq N(f).$$

We will precise why this definition is available for the different kernels we will consider. The link between h and $N(\hat{f}_{K,h})$ has been studied by several authors. With a Gaussian kernel (i.e. K is the density function of the standardized normal distribution), according to Silverman (1981), the function $h \mapsto N(\hat{f}_{K,h})$ is decreasing, which allows him and Mammen et al. (1991) to test the number of modes of f. For many other kernels among those with bounded support, we do not have these kind of properties, but we have at our disposal a visualization tool called the "mode tree" (see for details Minnotte and Scott (1993) or Minnotte et al. (1998)). Other theoretical results are also available in the literature, see for instance Hall et al. (2004). Let also define $h_{crit} := h_{crit,N(f)}$. A method to find an estimate of f based on this kind of assumptions already exists (Polonik (1995a)). This estimate will be compared with $\hat{f}_{K,h_{crit}}$, in a simulation study.

The paper is organized as follows. In Section 2, we give asymptotic results for the density kernel estimator $\hat{f}_{K,h_{crit,k}}$ where K is the Gaussian kernel. We also present theoretical results for the uniform kernel. In Section 3, we present a simulation study in order to compare numerical performances of various density estimators based or not on the assumption on the number of modes. Then, in Section 4, we describe how to use $h_{crit,k}$ in the context of mixture models. We apply it to environment monitoring data in Section 5. Lastly, concluding remarks are given in Section 6.

2. Estimating a density with $N(f)$ modes

In this section we study the kernel density estimator $\hat{f}_{K,h}$ given in (1) with the bandwidth $h_{crit,k}$ defined in (2). For our purpose, we only consider two kernels:

- the uniform kernel defined for $t \in \mathbb{R}$ as $K(t) = \mathbf{1}_{[-\frac{1}{2}, \frac{1}{2}]}(t)$,

- and the Gaussian kernel defined for $t \in \mathbb{R}$ as $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$.

The first kernel has a bounded support, this is not the case for the Gaussian kernel. We also describe two alternatives of $\hat{f}_{K,h_{crit,k}}$ which are respectively $\hat{f}_{K,h_{SJ}}$, where h_{SJ} is the bandwidth given by Sheather and Jones (1991) plug-in method (see Section 2.4) and Polonik (1995a) estimator based on density contour clusters (see Section 2.5).

2.1. Assumptions on the density f of X.

We need the following assumptions on the density f of X in order to have a density with $N(f)$ modes which can be properly estimated.

($\mathcal{H}1$) f is uniformly continuous on $\mathbb{R}$.

($\mathcal{H}2$) $\exists (r,s), \ -\infty < r < s < +\infty, \ f(x) \neq 0 \Rightarrow x \in [r,s]$ and $x \in]r,s[\Rightarrow f(x) \neq 0$.

($\mathcal{H}3$) $\exists! \ (z_1, z_2, \ldots, z_{2N(f)-1}) \in]r,s[^{2N(f)-1}, \quad \forall i \in \{1, 2, \ldots, 2N(f) - 1\}$,

$$f^{(1)}(z_i) = 0 \text{ and } sign(f^{(2)}) = (-1)^i,$$

where $f^{(q)}$ is the qth derivative of f.

($\mathcal{H}4$) $f \in \mathcal{C}^2(]r,s[)$.

($\mathcal{H}5$) $\lim_{t \downarrow r} f^{(1)}(t) > 0$ and $\lim_{t \uparrow s} f^{(1)}(t) < 0$.

($\mathcal{H}6$) $\forall x \in]r,s[$ that verifies $f^{(1)}(x) = 0, \ f^{(2)}(x) \neq 0$.

Remark 1. ($\mathcal{H}1$) follows Devroye and Wagner (1980) that gives an asymptotic result with the L_∞ norm, that we discuss in Section 2.3. ($\mathcal{H}2$) - ($\mathcal{H}6$) are taken from Mammen et al. (1991).

2.2. A computable bandwidth.

For the Gaussian kernel, some interesting results on $h_{crit,k}$ already exist. They underline that the bandwidth $h_{crit,k}$ is easily computable. Indeed, Silverman (1981) shows that the function $h \mapsto N(\hat{f}_{K,h})$ is decreasing and right continuous. This ensures computability of $h_{crit,k}$ with the desired accuracy by a dichotomous search. With the assumption that $h_{crit,k} \in [h_1, h_2]$, and if we want to obtain it with an error less than $\frac{h_2 - h_1}{2^m}$, we have to compute $N(\hat{f}_{K,h})$, m times, for various h. If for each h, to determine $N(\hat{f}_{K,h})$, $\hat{f}_{K,h}$ is computed in $\tilde{n}$ points, then the computational complexity of the whole algorithm to find $h_{crit,k}$ is equal to $O(n\tilde{n}m)$. In our simulations, we often take $\tilde{n} = 10000$ and $m \leq 30$.

For the uniform kernel, we provide a similar result in Propositions 1 and 2 by explaining that $h \mapsto N(\hat{f}_{K,h})$ is piecewise constant and has at most $\frac{n(n-1)}{2}$ jumps. Besides, if $\{X_{(i)}\}_{i \in \{1,\ldots,n\}}$ is a set of ordered random variables from which we want to compute $\hat{f}_{K,h_{crit,k}}$, the locations of the jumps

are in $\left\{X_{(j)} - X_{(i)}\right\}_{i \in \{1,\ldots,n-1\}, j \in \{i+1,\ldots,n\}}$. This means that we are able to find $h_{crit,k}$ by analyzing values of $\hat{f}_{K,h}$ between jumps.

Let us first introduce some additional notations. Let $A_h = \cup_{i=1}^{n}\{X_i - \frac{h}{2}\} = \left\{a_{h,(i)}\right\}_{i \in \{1,\ldots,\mathrm{card}(A_h)\}}$ and $B_h = \cup_{i=1}^{n}\{X_i + \frac{h}{2}\} = \left\{b_{h,(i)}\right\}_{i \in \{1,\ldots,\mathrm{card}(B_h)\}}$. In order to deduce the value of $N(\hat{f}_{K,h})$, we only need to investigate how the points in $A_h \cup B_h$ are ordered because of Proposition 1 below. Note that $w := \mathrm{card}(A_h \cup B_h) \leq 2n$. We set $\left\{c_{h,(i)}\right\}_{i \in \{1,\ldots,w\}}$ as the ordered points in $A_h \cup B_h$. Let us also write $c_{h,(0)} = -\infty$ and $c_{h,(w+1)} = +\infty$.

Proposition 1. *Let* $(X_1, \ldots X_n)$ *be a vector of independent random variables generated from* f. *Let* $\hat{f}_{K,h}$ *be the kernel estimator of* f *for the uniform kernel* K. *Then,* $\forall h > 0$, $\forall i \in \{0, 1, \ldots, w\}$, *the function* $\hat{f}_{K,h}$ *is constant on* $]c_{h,(i)}, c_{h,(i+1)}[$.

The proof is given in Appendix A.1.

Remark 2. Applying arguments similar to that used in the proof of Proposition 1, we obtain the following results:

- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \in A_h \Leftrightarrow \forall u \in]c_{h,(i-1)}, c_{h,(i)}[$, $f(u) = f(c_{h,(i)}) - \frac{\gamma}{nh}$, with $\gamma \in \mathbb{N}^*$,
- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \in B_h \Leftrightarrow \forall u \in]c_{h,(i)}, c_{h,(i+1)}[$, $f(u) = f(c_{h,(i)}) - \frac{\gamma}{nh}$, with $\gamma \in \mathbb{N}^*$,
- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \notin A_h \Leftrightarrow \forall u \in]c_{h,(i-1)}, c_{h,(i)}[$, $f(u) = f(c_{h,(i)})$,
- $\forall i \in \{1, \ldots, w\}$, $c_{h,(i)} \notin B_h \Leftrightarrow \forall u \in]c_{h,(i)}, c_{h,(i+1)}[$, $f(u) = f(c_{h,(i)})$.

Proposition 2. *Let* $(X_1, \ldots X_n)$ *be a vector of independent random variables generated from* f. *Let* $\hat{f}_{K,h}$ *be the kernel estimator of* f *for the uniform kernel* K. *The number of modes* $N(\hat{f}_{K,h})$ *of* $\hat{f}_{K,h}$ *is such that*

$$N(\hat{f}_{K,h}) = card\left(\left\{(i,j) \ : \ a_{h,(i)} \in]b_{h,(j-1)}, b_{h,(j)}] \ and \ b_{h,(j)} \in [a_{h,(i)}, a_{h,(i+1)}[\right\}\right).$$

The proof is given in Appendix A.2.

For h small enough, the sequence $\left\{c_{h,(k)}\right\}_{k \in \{1,\ldots,w\}}$ is equal to $\left\{a_{h,(1)}, b_{h,(1)}, \ldots, a_{h,(n)}, b_{h,(n)}\right\}$. For h large enough, this sequence is equal to $\left\{a_{h,(1)}, \ldots, a_{h,(n)}, b_{h,(1)}, \ldots, b_{h,(n)}\right\}$. Between a bandwidth h and a bandwidth $h + \varepsilon$, with $\varepsilon > 0$, the only change in the order that can occur is that: for a set $J \subset \{1, \ldots n-1\}$ and a nonempty set $I \subset \{j+1, \ldots, n\}$, $\forall j \in J$, $\forall i \in I$, $b_{h,(j)} < a_{h,(i)}$ and $a_{h+\varepsilon,(i)} \leq b_{h+\varepsilon,(j)}$. Because of Proposition 2, differences in the order of the $\left\{a_{h,(i)}\right\}_{i \in \{1,\ldots,\mathrm{card}(A_h)\}} \cup \left\{b_{h,(i)}\right\}_{i \in \{1,\ldots,\mathrm{card}(B_h)\}}$ is the only cause of differences between $N(\hat{f}_{K,h})$ and $N(\hat{f}_{K,h+\varepsilon})$. Even for h

very small and for a given $b_{h,(j)}$, this event occurs at most $(n - j)$ times. This implies that for all $b_{h,(j)}$, we can observe it at most $\frac{n(n-1)}{2}$ times and that is why the function $h \mapsto N(\hat{f}_{K,h})$ has at most $\frac{n(n-1)}{2}$ jumps.

Remark 3. The number of jumps in $h \mapsto N(\hat{f}_{K,h})$ is not bounded by $\frac{n(n-1)}{2}$ for every kernel. Indeed Hall et al. (2004) studied the set of points $X(\omega) = (-1, 0, 1)$ and drew $N(\hat{f}_{K_\theta,h})$, with $K_\theta(x) = C_\theta(1 - x^2)^\theta \mathbf{1}_{[-1,1]}(x)$ in function of h and θ, where C_θ ensures that $\|K_\theta\|_{L_1} = 1$. For example for $\theta = 1.5$, one can find 4 different values in $h \mapsto N(\hat{f}_{K_\theta,h})$, which is greater than $\frac{n(n-1)}{2} = 3$.

When events like those previously described occur, it means that for some $(i, j) \in \{j + 1, \ldots, n\} \times \{1, \ldots, n - 1\}$, $X_{(j)} + \frac{h}{2} < X_{(i)} - \frac{h}{2}$ and $X_{(j)} + \frac{h+\varepsilon}{2} \geq X_{(i)} - \frac{h+\varepsilon}{2}$, or equivalently $h < X_{(i)} - X_{(j)} \leq h + \varepsilon$. This leads to a procedure to plot $h \mapsto N(\hat{f}_{K,h})$. We write $H = \cup_{j=1}^{n-1} \cup_{i=j+1}^{n} \{X_{(i)} - X_{(j)}\}$. Let $h_{(\gamma)}$ be an element of the ordered sequence of the elements of H. Choose any $h_{(0)}$ such that $h_{(0)} < h_{(1)}$. Set $l := \mathrm{card}\,(H)$. With the sequence $\left\{ N(\hat{f}_{K,h_{(\gamma)}}) \right\}_{\gamma \in \{0,\ldots,l\}}$, we can deduce the value of $h \mapsto N(\hat{f}_{K,h})$ for any h. Consequently, we can find $h_{crit,k}$. Note that because of the inequality $h < X_{(i)} - X_{(j)} \leq h + \varepsilon$, the function $h \mapsto N(\hat{f}_{K,h})$ is right continuous.

Remark 4. If we consider the example from Hall et al. (2004) with the uniform kernel, we have $X(\omega) = (-1, 0, 1)$. We have $h_{(1)} = 1$ and $h_{(2)} = 2$. We choose $h_{(0)} = 0.5$. Then, $\left\{ N(\hat{f}_{K,h_{(i)}}) \right\}_{i \in \{0,\ldots,l\}} = (3, 2, 1)$, and

$$N(\hat{f}_{K,h}) = \begin{cases} 3 & \text{for } h \in [0, 1[, \\ 2 & \text{for } h \in [1, 2[, \\ 1 & \text{for } h \in [1, \infty[. \end{cases}$$

The number of jumps of $h \mapsto N(\hat{f}_{K,h})$ is equal to $2 \leq \frac{n(n-1)}{2} = 3$.

2.3. **Asymptotic results on $\hat{f}_{K,h_{crit,k}}$.**

Proof of consistency for this estimator toward f is not trivial since $h_{crit,k}$ is data-driven. However, for the Gaussian kernel, we have the pointwise convergence in probability, among others. To explain this result, we first find conditions for a given data-driven bandwidth h_n under which some asymptotic properties can be shown for $\hat{f}_{K,h_n}$. This is realized in Theorem A by combining Theorem 2 of Devroye and Wagner (1980) about the L_∞ distance between $\hat{f}_{K,h_n}$ and f and Theorem 3.3 from Devroye (1987, p. 38), concerning the L_1 distance.

Theorem A (Devroye and Wagner (1980), Devroye (1987)). *Let f be a probability density satisfying $(\mathcal{H}1)$, h_n a random bandwidth depending on X. If we assume the following hypotheses:*

($\mathcal{F}$1) K is a Riemann integrable probability density,

($\mathcal{F}$2) $\sup_{x \in \mathbb{R}} K(x) < \infty$,

($\mathcal{F}$3) $\int_0^\infty \sup_{|x| \geq z} K(x) dz < \infty$,

($\mathcal{F}$4) $\forall \varepsilon > 0$, $\mathbb{P}(h_n > \varepsilon) \to 0$, when $n \to \infty$,

($\mathcal{F}$5) $\forall A > 0$, $\mathbb{P}(nh_n > A) \to 1$, when $n \to \infty$,

then, we have, for $n \to \infty$,

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \hat{f}_{K,h_n}(t) - f(t) \right| > \varepsilon \right) \to 0,$$

and

$$\mathbb{P}\left(\int_{\mathbb{R}} \left| \hat{f}_{K,h_n}(t) - f(t) \right| dt > \varepsilon \right) \to 0.$$

Mammen et al. (1991) prove ($\mathcal{F}$4) and ($\mathcal{F}$5) for $h_{crit,k}$, and we have the following theorem on the consistency of $\hat{f}_{K,h_{crit}}$ for a Gaussian kernel.

Theorem 1. *Let f be a density satisfying $(\mathcal{H}1) - (\mathcal{H}6)$ and let $\hat{f}_{K,h_{crit,k}}$ be the estimator of f with the Gaussian kernel K and the bandwidth $h_{crit,k}$ given in (2). Then we have, for $n \to \infty$,*

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \hat{f}_{K,h_{crit,k}}(t) - f(t) \right| > \varepsilon \right) \to 0,$$

and

$$\mathbb{P}\left(\int_{\mathbb{R}} \left| \hat{f}_{K,h_{crit,k}}(t) - f(t) \right| dt > \varepsilon \right) \to 0.$$

The proof is given in Appendix A.3.

Assuming some regularity conditions on the kernel, Hall et al. (2004) proved similar results in their Theorems 3.1 and 3.2 (pp. 2130–2131). These conditions on the kernel are stronger than continuity on $\mathbb{R}$ and thus the uniform kernel does not satisfy them. In that case, we prove that we cannot have ($\mathcal{F}$4) in the following theorem.

Theorem 2. *For any probability density function f of X, let $\hat{f}_{K,h_{crit,k}}$ be the estimator of f when K is the uniform kernel with $h_{crit,k}$ given in (2). Then we have $h_{crit,k}$ increasing with n.*

The proof is given in Appendix A.4.

2.4. **Sheather and Jones' plug-in method to choose a bandwidth.** In the more general context of estimating a density without assumption on the number of its modes, algorithms that provide a

suitable bandwidth h are of particular interest. Among a large selection of procedure, we focus on the plug-in method developed by Sheather and Jones (1991) (see also Jones and Sheather (1991)), which leads to the bandwidth h_{SJ}. We chose it because h_{SJ} has good asymptotic properties and is easy to compute. This bandwidth is designed to minimize the asymptotic mean integrated squared error ($AMISE$) between $\hat{f}_{K,h}$ and f, defined as:

$$AMISE(h) = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f^{(2)}),$$

where for any function ψ, $R(\psi) = \int_{-\infty}^{\infty} \psi(x)^2 dx$ and σ_K^2 is the variance of a random variable of density K. To minimize $AMISE(h)$, $R(f^{(2)})$ must be estimated. For this purpose, modifying an estimator studied by Hall and Marron (1987), Sheather and Jones (1991) used the following one:

$$(3) \qquad \hat{R}(f^{(2)}) = \frac{1}{n(n-1)\tilde{h}^5} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{K}^{(4)}\left(\frac{X_i - X_j}{\tilde{h}}\right),$$

where $\tilde{K}$ is allowed to be different from the kernel K used in the estimate of f, $\tilde{K}^{(4)}$ is the fourth derivative of $\tilde{K}$ and $\tilde{h} = \left[\frac{2K^{(4)}(0)}{n\sigma_K^2 \hat{R}(f^{(3)})}\right]^{1/7}$. Estimator $\hat{R}(f^{(3)})$ of $R(f^{(3)})$ is similar to the one in (3). It requires a new bandwidth $\breve{h}$ chosen to be equal to $0.912\hat{\lambda}n^{-1/9}$, where $\hat{\lambda}$ is the sample interquartile range. Finally,

$$h_{SJ} = \arg\min_h \left(\frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 \hat{R}(f^{(2)})\right).$$

Let $h_{opt} = \arg\min_h(MISE(h))$, Sheather and Jones (1991) showed that

$$\frac{h_{SJ}}{h_{opt}} = 1 + O_P(n^{-5/14}).$$

This means that h_{SJ} is close to the bandwidth that minimizes the expected L_2 distance between a kernel density estimator and the true density. Note that, for the Gaussian kernel, this result is valid for a density with three derivatives and which verifies for all x and y:

$$\exists Z > 0, \; \left|f^{(3)}(x) - f^{(3)}(y)\right| \leq Z\,|x-y|^{\frac{1}{4}}.$$

2.5. **Polonik's estimator based on excess mass location.** The kernel density estimator $\hat{f}_{K,h}$ aims to associate a fixed point t with a value $\hat{f}_{K,h}(t)$ as close as possible to $f(t)$. Another approach, described in this subsection, tries to determine for every given $\lambda \in [0,\infty[$, the set $\hat{\Gamma}_{n,\mathbb{C}}(\lambda)$ which is the most similar to $\Gamma(\lambda) := \{t \; : \; f(t) \geq \lambda\}$, where $\mathbb{C}$ is a set of unions of disjoint intervals of $\mathbb{R}$ chosen such that for every $\lambda \in [0,\infty[$, $\Gamma(\lambda)$ lies in $\mathbb{C}$. In our case, $\mathbb{C}$ is made of every unions of

at most $N(f)$ disjoint intervals. This procedure was developed by Polonik (1995a,b) and is related to the test for multimodality introduced by Müller and Sawitzki (1991). It leads to an estimator of $f(t) = \int_0^\infty \mathbf{1}_{\Gamma(\lambda)}(t)d\lambda$ defined by:

$$\hat{f}_P(t) := \int_0^\infty \mathbf{1}_{\hat{\Gamma}_{n,\mathbb{C}}(\lambda)}(t)d\lambda,$$

where $\hat{\Gamma}_{n,\mathbb{C}}(\lambda)$, the so-called empirical generalized λ-cluster in $\mathbb{C}$, verifies

$$\hat{\Gamma}_{n,\mathbb{C}}(\lambda) = \arg \max_{\{C(i)\}_{i\in\{0,\ldots n\}}} \left(\frac{i}{n} - \lambda\mu(C(i)) \right).$$

For every $C \in \mathbb{C}$, $\mu(C)$ is the sum of the length of all disjoint closed intervals in C. $C(i)$ is one of the narrowest elements of $\mathbb{C}(i)$, which means that $\forall C \in \mathbb{C}(i)$, $\mu(C(i)) \leq \mu(C)$, where $\mathbb{C}(i)$ is the subset of $\mathbb{C}$ such that $\forall C \in \mathbb{C}(i)$, $\sum_{j=1}^n \mathbf{1}_C(X_j) = i$. Then, if $\int_{\mathbb{R}} \hat{f}_P(t)dt = 1$, Polonik (1995a, Theorem 3.1) gives that:

$$\mathbb{P}\left(\lim_{n\to\infty} \int_{\mathbb{R}} \left| \hat{f}_P(t) - f(t) \right| dt = 0 \right) = 1.$$

We have $\int_{\mathbb{R}} \hat{f}_P(t)dt = \frac{n-1}{n}$, but, the proof of Theorem 3.1 from Polonik (1995a) still works for $\int_{\mathbb{R}} \hat{f}_P(t)dt \leq 1$. Thus, to obtain a value for $\hat{f}_P(t)$, $C(i)$ should be determined for every $i \in \{0, \ldots, n\}$. Then, because $\hat{\Gamma}_{n,\mathbb{C}}(\lambda)$ is the $C(i)$ that maximizes $n+1$ linear functions, it only changes for a finite number k_n of λ, with $k_n \leq n$. Let $\left\{ \lambda_{(i)} \right\}_{i\in\{1,\ldots,k_n\}}$ be the set of this ordered change points and take $\lambda_{(0)} = 0$. When $\lambda \geq \lambda_{(k_n)}$, $\hat{\Gamma}_{n,\mathbb{C}}(\lambda) = C(0)$ and $\mu(C(0)) = 0$, and we can set $\hat{\Gamma}_{n,\mathbb{C}}(\lambda) = C(0) = \emptyset$. Hence, another expression for $\hat{f}_P$ is given by

$$\forall t \in \mathbb{R}, \ \hat{f}_P(t) = \sum_{i=0}^{k_n-1} \left[\left(\lambda_{(i+1)} - \lambda_{(i)} \right) \mathbf{1}_{\hat{\Gamma}_{n,\mathbb{C}}(\lambda_{(i)})}(t) \right],$$

which can be used to computed $\hat{f}_P(t)$.

3. SIMULATION STUDY

In this simulation study, we compare four density estimators: $\hat{f}_{K,h_{crit}}$ based on the bandwidth h_{crit} with both Gaussian and uniform kernels, $\hat{f}_{K,h_{SJ}}$ based on the Sheather and Jones' bandwidth with the Gaussian kernel, and Polonik's estimator $\hat{f}_P$. We specifically show numerical convergences of $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel and illustrate consequences of the unsatisfied requirement of the uniform kernel.

3.1. **Simulated data and quality assessment of the estimates.** To generate simulated datasets, we use both a beta and a Gaussian mixture. The beta mixture model is defined by:

$$(4) \qquad X \sim \begin{cases} \mathcal{B}(\alpha_1, \beta_1) & \text{with probability } p_1, \\ \\ \mathcal{B}(\alpha_2, \beta_2) & \text{with probability } p_2 = 1 - p_1. \end{cases}$$

Note that the corresponding density of X is:

$$f_1(t) = p_1 \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} t^{\alpha_1 - 1}(1 - t)^{\beta_1 - 1} + p_2 \frac{\Gamma(\alpha_2 + \beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)} t^{\alpha_2 - 1}(1 - t)^{\beta_2 - 1}.$$

In this section, we used the parameters $\alpha_1 = 2$, $\beta_1 = 5$, $\alpha_2 = 10$, $\beta_2 = 2$, $p_1 = \frac{2}{3}$. Graphically, we observe that $N(f_1) = 2$, which is theoretically confirmed in Section 4.

The Gaussian mixture we chose is the asymmetric claw density introduced by Marron and Wand (1992) :

$$(5) \qquad X \sim \begin{cases} \mathcal{N}(0, 1) & \text{with probability } \frac{1}{2}, \\ \\ \mathcal{N}\left(l + \frac{1}{2}, \left(\frac{2^{-l}}{10}\right)^2\right) & \text{with probability } \frac{2^{1-l}}{31}, \text{ for } l \in \{-2, -1, 0, 1, 2\}. \end{cases}$$

Despite the fact that this mixture has 6 components, the underlying density has 5 modes (see Minnotte et al. (1998)). Its expression is:

$$f_2(t) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}t^2} + \sum_{l=-2}^{2} \frac{20}{31\sqrt{2\pi}} e^{-50\left(\frac{t-l-\frac{1}{2}}{2^{-l}}\right)^2}.$$

Because most of theoretical results we present in this paper concern the L_1 distance between an estimator $\hat{f}$ and the true density f, it makes sense to use the following criterion, often called integrated absolute error (IAE), defined as:

$$IAE = \left\| \hat{f} - f \right\|_{L_1} = \int_{\mathbb{R}} \left| \hat{f}(t) - f(t) \right| dt.$$

We are also interested in another criterion which is the absolute error committed by the estimator $\hat{z}$ of

$$z := \arg \min_{z_i \in \{z_{2j}\}_{j \in \{1, \ldots, N(f)-1\}}} f(z_i),$$

with z_{2j} defined in ($\mathcal{H}3$). The estimator $\hat{z}$ is chosen to verify

$$\hat{z} := \arg \min_{z_i \in \hat{Z}} \hat{f}(z_i),$$

where $\hat{Z}$ is made of the points $\tilde{z}$ such that,

$$\exists \varepsilon > 0, \ \forall \tilde{\varepsilon} < \varepsilon, \ \hat{f}(\tilde{z} - \tilde{\varepsilon}) > \hat{f}(\tilde{z}) \text{ and } \hat{f}(\tilde{z} + \tilde{\varepsilon}) > \hat{f}(\tilde{z}).$$

3.2. **Simulation results.** In Figure 1, we first draw an example of density estimation with the four considered estimators, using a sample of size $n = 1600$ generated from model (4). Apart from $\hat{f}_{K,h_{crit}}$ with the uniform kernel, estimates of the density seem to be close to f, considering the shape. We observe a lot of peaks for $\hat{f}_P$. This feature is directly related to the estimation method as it has been already noticed by Müller and Sawitzki (1991). Polonik (1995a) wrote that $N(\hat{f}_P)$ can be different from $N(\hat{f})$. In Figure 1(b), the estimation of z related to $\hat{f}_P$ is close to 0 and far from $z \approx 0.6219$. For a given estimation $\hat{f}$ of f, finding $\hat{z}$ in $]-\infty, z_1[\cup]z_{2N(f)-1}, \infty[$ only happens if there is a mode of $\hat{f}$ in this interval that is sufficiently far from z_1 and from $z_{2N(f)-1}$. This event is especially likely to arise when $N(\hat{f}) > N(f)$ but it can also happen when $N(\hat{f}) = N(f)$ if the modes of f are not clearly separated. In addition, when K is the Gaussian kernel, estimations of z for $\hat{f}_{K,h_{crit}}$ (Figure 1(a)) and for $\hat{f}_{K,h_{SJ}}$ (Figure 1(b)) are close to z.

Then, we generate 100 replicates from model (4) for various sample sizes $n \in \left\{100 \times 2^i\right\}_{i \in \{0,\ldots,9\}}$. For each sample and each density estimation procedure, IAE is computed. The corresponding values are represented in Figure 2 with boxplots. Not surprisingly, for $\hat{f}_{K,h_{crit}}$ with the uniform kernel, we observe in Figure 2(a) that IAE increases with n, which is compatible with Theorem 2. Performances of $\hat{f}_{K,h_{crit}}$ for the Gaussian kernel, shown in Figure 2(b), are better. In this case, boxplots exhibit the L_1 convergence of Theorem 1. They reach a similar precision to those obtained for $\hat{f}_{K,h_{SJ}}$ which are drawn in Figure 2(c). Polonik's method (Figure 2(d)) needs extensive computational time. That is why we are not able to draw boxplots for the greatest sample sizes, but we still observe convergence of this procedure despite of the many peaks of the estimates. Values of IAE appear to be slightly greater for this estimator than those for $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel.

For each replicate we made, we also compute estimations of z. In Figure 3 we draw various values of $\hat{z}$ for the four previously considered estimators. For $\hat{f}_{K,h_{crit}}$ with the uniform kernel, in Figure 3(a), $\hat{z}$ values move away from the position of the local minimum of f_1 (which is equal to 0.6219) when sample size increases. This is not surprising because of the poor quality of this estimator which, according to Figure 1, provides two modes close to each other, but far from the true ones. In Figure 3(b), we notice a convergence of the boxplots toward the aimed location for $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel. For $\hat{f}_{K,h_{SJ}}$ with the Gaussian kernel, in Figure 3(c), a similar convergence can be observed. For $\hat{f}_P$, in

Figure 3(d), values of $\hat{z}$ seem to tend toward 0 or 1 when sample size increases. A possible explanation of this phenomenon is that a spurious mode close to 0 or 1 is sometimes created by the procedure. Because this mode is located in an interval of t where $f_1(t)$ is small, the local minimum near the mode is the minimum over all local minima. This occurs in Figure 1(b), for example.

These simulation results lead us to focus only on the estimators $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel because both exhibit IAE convergence toward 0 and convergence of $\hat{z}$ toward z. To study them lengthier, we use the model (5). Moreover, this selection is also made because of the costs in computational time of the different methods. For example, for a given sample of size $n = 1600$, we measured computational time of the methods we consider in this study, with our Intel Core 2 Quad Q9505 processor. To obtain h_{crit} with a Gaussian kernel we need about 1.6 seconds using the `density()` R function. Finding h_{SJ} requires 0.004 seconds, with the `KernSmooth` R package while our R implementation of $\hat{f}_P$ needs 425 seconds to be computed. Our R algorithm finds h_{crit} in 4 seconds for the uniform kernel.

In Figure 4, we draw another example of density estimation using the estimators $\hat{f}_{K,h_{SJ}}$ and $\hat{f}_{K,h_{crit}}$, and a sample of size $n = 1600$ generated from model (5). We observe very similar results for both procedures even if $\hat{f}_{K,h_{crit}}$ seems to produce an estimate slightly more precise than $\hat{f}_{K,h_{SJ}}$ for the estimation of z.

Boxplots presenting IAE values for these methods are drawn in Figure 5. The dispersion of the IAE values does not seem to decrease with n in Figure 5(a). This result could come from the fact that for the asymmetric claw density, $(\mathcal{H}2)$ does not hold and Theorem 1 cannot be applied, but IAE values of Figure 5(a) globally decrease. Thus, another practical explanation is that the bandwidth h_{crit} cannot adapt to the various sharpness of the modes of f_2, while $\hat{f}_{K,h_{SJ}}$ may compensate this phenomenon by creating a new mode such that $N(\hat{f}_{K,h_{SJ}}) > N(f_2)$. Indeed, in Figure 5(b), IAE values converge toward 0 like those in Figure 2(c).

To conclude this simulation study, we notice that for the asymmetric claw density, estimates of z can be located at the extrema of the sample, for both $\hat{f}_{K,h_{crit}}$ and $\hat{f}_{K,h_{SJ}}$, in Figure 6. This previously occurs for $\hat{f}_P$ and model (4) in Figure 3(d). For $\hat{f}_{K,h_{crit}}$, in Figure 6(a), samples that produce this kind of estimate are not sufficiently frequent for the estimates to be considered as outliers. In Figure 6(b), for $\hat{f}_{K,h_{SJ}}$ and when $n \geq 6400$, every $\hat{z}$ is in the tails of the estimated distribution. Thus, for the estimation of z, h_{crit} seems to perform better than h_{SJ} despite of the fact that $(\mathcal{H}2)$ does not hold for f_2.

4. Assuming the number of modes of a mixture density

Estimation of a density of a mixture model with a known maximum number of components is the main task that $\hat{f}_{K,h_{crit,k}}$ can realize. Indeed, if each component of the mixture model is an unimodal density, there are various cases where the number of modes of the density of the mixture is at most equal to the number of components. However, it is not true for every mixture model. For instance, densities made from components $f_{\mu,\theta}$ that verify $f_{\mu,\theta}(x) = C_\theta(1 - (x - \mu)^2)^\theta \mathbf{1}_{[-\mu,\mu]}(x)$, where μ and θ should be chosen for each component, can have a number of modes greater than its number of components (see Remark 3 and Hall et al. (2004)).

Notice that a function which is convex on an open interval does not have any mode on this interval and that the density of a mixture of densities which are convex on an interval is convex on this interval too. Thus, let f be a density of a mixture model with m components. For $i \in \{1, \ldots, m\}$, let g_i be the density of the component i. If it exists $\mu_i \in \mathbb{R}$ such that g_i is convex on $]-\infty, \mu_i[$ and on $]\mu_i, \infty[$, then $N(f) \leq m$ and $\forall j \in \{1, \ldots, N(f)\}, \; z_{2j-1} \in \cup_{i=1}^m \{\mu_i\}$. For example one can take for g_i the standard two-sided power distribution of van Dorp and Kotz (2002):

$$
g_i(t) := \left\{
\begin{array}{ll}
\gamma_i \left(\frac{t}{\theta_i}\right)^{\gamma_i-1} & \text{for } t \in [0, \theta_i[, \\
\gamma_i \left(\frac{1-t}{1-\theta_i}\right)^{\gamma_i-1} & \text{for } t \in [\theta_i, 1],
\end{array}
\right.
$$

with $\theta_i \in [0, 1]$ and $\gamma_i > 2$ If we allow ourselves not to respect $(\mathcal{H}2)$, the Laplace distribution is another valid choice for g_i.

Mixture model densities with $N(f) \leq m$ are not restricted to those that verify the previous condition of convexity. Considering the beta mixture model of Section 3 with $\alpha_1 = 2$ and $\beta_2 = 2$, the density f_3 of (4) can be written

$$
f_3(t) := p_1 g_1(t) + (1 - p_1) g_2(t),
$$

with $g_1(t) := \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} t(1 - t)^{\beta_1 - 1}$ and $g_2(t) := \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} t^{\alpha_2 - 1}(1 - t)$. When $\beta_1 > 2$, for $q \in \{1, 2, 3\}$, the qth derivative of g_1 verifies:

$$
g_1^{(q)} \geq 0 \Leftrightarrow (-1)^{q-1} x \leq \frac{(-1)^{q-1} q}{\beta_1},
$$

and we also have for $q \in \{1, 2, 3\}$:

$$
g_2^{(q)} \geq 0 \Leftrightarrow x \leq 1 - \frac{q}{\alpha_2}.
$$

These relations hold for strict inequalities and imply that f_3 has no mode in $[0, \frac{1}{\beta_1}[$ and in $]1 - \frac{1}{\alpha_2}, 1]$ because $g_1^{(1)}$ and $g_2^{(1)}$ have the same sign on these intervals on which this sign is constant. Assume now that $\frac{3}{\beta_1} < 1 - \frac{2}{\alpha_2}$ and that $\frac{2}{\beta_1} < 1 - \frac{3}{\alpha_2}$, which is true for the values of β_1 and α_2 chosen in Section 3 for f_1. Then, f_3 has at most one mode on $\left[\frac{1}{\beta_1}, \min\left\{\frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2}\right\}\right[$. This can be proven by assuming that f_3 has two modes located in z_1 and z_3 and an antimode located in z_2 with $\frac{1}{\beta_1} \leq z_1 < z_2 < z_3 < \min\left\{\frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2}\right\}$. This implies that $f_3^{(2)}(z_2) = p_1 g_1^{(2)}(z_2) + (1 - p_1) g_2^{(2)}(z_2) \geq 0$, and because $g_1^{(3)}(t) > 0$ and $g_2^{(3)}(t) > 0$ for $t \in \left]z_2, \min\left\{\frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2}\right\}\right[$, $p_1 g_1^{(2)}(t) + (1 - p_1) g_2^{(2)}(t) > 0$ which negates the fact that $f_3^{(2)}(z_3) = p_1 g_1^{(2)}(z_3) + (1 - p_1) g_2^{(2)}(z_3) \leq 0$. A demonstration of the same type leads to the property that f_3 has at most one mode on $\left]\max\left\{\frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2}\right\}, 1 - \frac{1}{\alpha_2}\right]$.

Thus, in order to demonstrate that $N(f_3) \leq 2$, we have to show that f_3 has no mode on Δ, with:

$$\Delta := \left[\max\left\{\frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2}\right\}, \min\left\{\frac{3}{\beta_1}, 1 - \frac{3}{\alpha_2}\right\}\right].$$

Because of assumptions we made on β_1 and α_2, Δ is included in $\left]\frac{2}{\beta_1}, 1 - \frac{2}{\alpha_2}\right[$ on which both $g_1^{(2)}$ and $g_2^{(2)}$ are positive. This implies that f_3 is convex on this interval and then has no mode on it. Thus f_3 has no mode on Δ and $N(f_3) \leq 2$. Although we anticipate that generalizing this result to a wider set of mixtures is feasible, the demonstration would probably be tedious.

5. Oyster opening amplitudes modeled with a bimodal density

In this section we describe a real data application. We apply the estimator $\hat{f}_{K, h_{crit, k}}$ to opening amplitudes of oysters. These animals are studied by a laboratory called Environnements et Paléoenvironnements Océaniques et Continentaux (http://molluscan-eye.epoc.u-bordeaux1.fr, EPOC) in order to derive water quality indicators. Their approach is based on the assumption that a water of poor quality leads to perturbations in the oysters' behavior. It consists in measurements of the distance between the two parts of the shell of the oysters with a frequency of 0.625 Hz. The procedure carried out to obtain the dataset is non invasive. It relies on electrodes stuck on the shell of the oysters and on the GSM/GPRS service to transfer the data. The animals studied in this article are immersed in the Bay of Arcachon, in France.

We aim at estimating the density f of the distances of the parts of the shell of these animals. During a day, following the tide, an oyster is either open or closed (see for instance Sow et al. (2011)). For each of this state, the density of the opening amplitudes is assumed to be unimodal. If we also assume that

these densities behave similarly to the two-sided power distribution or to the beta mixture detailed in Section 4, we have $N(f) \leq 2$.

Because of these assumptions, we estimate f with $\hat{f}_{K,h_{crit,2}}$ with the Gaussian kernel and a sample of size $n \approx 50000$. This leads to Figure 7. In Figure 7(a), the data come from an oyster that does not exhibit any feature of sickness. In Figure 7(b), the oyster analyzed died one week after these measures were recorded. Generally, when they are in the open state, dying oysters produce opening amplitudes with a wide variability which implies a larger mode for the corresponding density. Thus, the local minimum z_2 of f located between its two modes comes close to the location z_1 of the mode related to the close state, when oysters' health becomes poor. This feature can also be observed for $\hat{f}_{K,h_{crit,2}}$ in Figure 7. Indeed, let $\hat{z}_1$ and $\hat{z}_3$ be the local maxima of $\hat{f}_{K,h_{crit,2}}$ and let $\hat{z}_2$ be its local minimum such that $\hat{z}_1 < \hat{z}_2 < \hat{z}_3$. Then, $\frac{\hat{z}_2 - \hat{z}_1}{\hat{z}_3 - \hat{z}_1}$ is greater in Figure 7(a) than in Figure 7(b) with respective values 0.3073 and 0.1566. This observation could lead to a detection of oysters in poor health. Notice that the choice of the estimator $\hat{f}_{K,h_{crit,2}}$ is important here in order for the estimate to have exactly two modes which allows $\frac{\hat{z}_2 - \hat{z}_1}{\hat{z}_3 - \hat{z}_1}$ to make sense. Maximum amplitude of the openings of the animal is also linked to its health and we can see that it is approximately 4 times greater in Figure 7(a) than in Figure 7(b). However, this quantity may also vary with the size of the animal, and with the position of the electrodes on it. That is why we prefer not to rely on it.

6. Concluding remarks

The estimator $\hat{f}_{K,h_{crit,k}}$, when K is a Gaussian kernel, is able to estimate a density f that has a known number of modes because when sample size is large enough, $\hat{f}_{K,h_{crit,k}}$ is close to f from both L_1 and L_∞ points of view. This is not the case when K is the uniform kernel in our simulation study and theoretical reasons were explored (Theorem 2).

An interesting feature of $\hat{f}_{K,h_{crit,k}}$ is that it has a deterministic number of modes and can have as many modes as f. In real data analysis, this allows to seek the positions of the various modes and antimodes of the estimate of a density and to derive indicators from them in order to compare densities (see Section 5). This is not possible with Polonik's estimator $\hat{f}_P$ or with $\hat{f}_{K,h_{SJ}}$ because both $N(\hat{f}_P)$ and $N(\hat{f}_{K,h_{SJ}})$ are random variables.

When K is the Gaussian kernel, the constraint $N(\hat{f}_{K,h_{crit,k}}) = k$ does not imply a great loss of convergence rate for $\hat{f}_{K,h_{crit}}$ in simulation, compared with results of $\hat{f}_{K,h_{SJ}}$. Actually, $\hat{f}_{K,h_{crit}}$ exhibits slightly better results than $\hat{f}_P$ and appears to produce the most precise estimate of the location of the

minimum of the local minima of a density, among the four studied estimators. From a practical point of view, note that the bandwidth h_{crit} required more time to compute than h_{SJ} but less than $\hat{f}_P$.

To conclude, asymptotic properties of $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel, together with its behavior in simulation, and its deterministic number of modes allow this estimator to be applied to real datasets that are assumed to come from mixture model densities. For that matter, an implementation of this work is included in numerical procedures daily performed on the environmental data of EPOC laboratory.

Appendix A. Proofs

A.1. **Proof of Proposition 1.** Let $(u,v) \in]c_{h,(i)}, c_{h,(i+1)}[\times]u, c_{h,(i+1)}[$ and $\{X_{(i)}\}_{i \in \{1,\dots n\}}$ be the ordered sequence of the elements of X. We will show that $\hat{f}_{K,h}(u)$ is neither greater nor lesser than $\hat{f}_{K,h}(v)$ with a proof by contradiction. Note that for the uniform kernel we have $\hat{f}_{K,h}(u) = \frac{1}{nh}\text{card}\left(\{X_k \in [u - \frac{h}{2}, u + \frac{h}{2}]\}\right)$.

If $\hat{f}_{K,h}(u) > \hat{f}_{K,h}(v)$, this implies that it exists at least one $k \in \{1,\dots,n\}$, for which we have $X_{(k)} \in [u - \frac{h}{2}, v - \frac{h}{2}[$, which means that it exists $k' \in \{1,\dots,w\}$ which verifies $c_{h,(k')} = b_{h,(k)} \in [u,v[$. Because $[u,v[\subset]c_{h,(i)}, c_{h,(i+1)}[$, $c_{h,(k')} \in]c_{h,(i)}, c_{h,(i+1)}[$, but this is impossible.

Conversely, $\hat{f}_{K,h}(v) > \hat{f}_{K,h}(u)$ implies that it exists $X_{(k)} \in]u + \frac{h}{2}, v + \frac{h}{2}]$. Then it exists $k' \in \{1,\dots,w\}$ such that $c_{h,(k')} = a_{h,(k)} \in]u,v] \subset]c_{h,(i)}, c_{h,(i+1)}[$ and it is also impossible.

A.2. **Proof of Proposition 2.** We will show the equivalence between the presence of a mode between $a_{h,(i)}$ and $b_{h,(j)}$ and the inequality $b_{h,(j-1)} < a_{h,(i)} \le b_{h,(j)} < a_{h,(i+1)}$.

At first, we notice that ordered like this, there is no element of A_h or B_h that can be between $a_{h,(i)}$ and $b_{h,(j)}$. This is why the last inequality is equivalent to $\exists k \in \{1,\dots,w-1\}$, $a_{h,(i)} = c_{h,(k)}$ and $b_{h,(j)} = c_{h,(k+1)}$, in the case where $a_{h,(i)} \ne b_{h,(j)}$.

From Proposition 1, $\hat{f}_{K_u,h}$ is constant on $]a_{h,(i)}, b_{h,(j)}[$, and thanks to Remark 2, it is equivalent to: $\hat{f}_{K,h}$ is constant on $[a_{h,(i)}, b_{h,(j)}] = [c_{h,(k)}, c_{h,(k+1)}]$. In order for this interval to be a mode, we must prove that it exists $\varepsilon > 0$ for which $\hat{f}_{K,h}$ is increasing on $[c_{h,(k)} - \varepsilon, c_{h,(k)}[$ and decreasing on $]c_{h,(k+1)}, c_{h,(k+1)} + \varepsilon]$, which is also made in Remark 2.

When $a_{h,(i)} = b_{h,(j)} = c_{h,(k)}$, $\hat{f}_{K,h}$ is increasing on $[c_{h,(k)} - \varepsilon, c_{h,(k)}[$ too and decreasing on $]c_{h,(k)}, c_{h,(k)} + \varepsilon]$. The mode is reduced to a single point.

A.3. **Proof of Theorem 1.** The main idea of the proof is to apply Theorem A of Section 2.3. When K is the Gaussian kernel, $(\mathcal{F}1)$ and $(\mathcal{F}2)$ are verified as well as $(\mathcal{F}3)$ because K is defined and decreasing

on $[0, \infty]$ and then $\int_0^\infty \sup_{|x| \geq z} K(x) dz = \int_0^\infty K(x) dx < \infty$. It remains to prove that $(\mathcal{F}4)$ and $(\mathcal{F}5)$ are also verified. Every convergence written in this proof is for $n \to \infty$.

Because of Theorem 2 from Mammen et al. (1991), we have

$$\forall A > 0, \ \mathbb{P}(N(\hat{f}_{K,n^{-1/4}}) > A) \to 1.$$

If we set $A = N(\hat{f}_{K,h_{crit,k}})$, we have:

$$\mathbb{P}(N(\hat{f}_{K,n^{-1/4}}) > N(\hat{f}_{K,h_{crit,k}})) \to 1.$$

Because $h \to N(\hat{f}_{K,h})$ is decreasing and piecewise constant, we can write

$$\mathbb{P}(h_{crit,k} > n^{-1/4}) \to 1, \qquad \text{or equivalently} \qquad \mathbb{P}(n h_{crit,k} > n^{3/4}) \to 1.$$

Then, for all $A > 0$, there exists an integer $\tilde{n} \in \mathbb{N}$ such that for all $n \geq \tilde{n}$, we have $n^{3/4} > A$ and $\mathbb{P}(n h_{crit,k} > n^{3/4}) < \mathbb{P}(n h_{crit,k} > A)$. Consequently we have

$$\forall A > 0, \ \mathbb{P}(n h_{crit,k} > A) \to 1$$

and $(\mathcal{F}5)$ is verified.

Besides, Corollary 1.2 from Mammen et al. (1991) and $(\mathcal{H}3)$ imply that

$$(6) \qquad\qquad \mathbb{E}\left[N(\hat{f}_{K,n^{-1/6}})\right] \leq N(f) + o(1).$$

These authors follow Silverman (1983) and use the fact that, for $N(f) > 1$, assuming $(\mathcal{H}3)$, there is $h_0 > 0$ such that $\mathbb{P}(\tilde{h}_{crit} > h_0) \to 1$, where $\tilde{h}_{crit} = \min_{N(\hat{f}_{K,h})=N(f)-1} h$. Then, we have

$$\mathbb{P}(\tilde{h}_{crit} > n^{-1/6}) \to 1.$$

Because of the definition of $\tilde{h}_{crit}$, we have

$$\mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) > N(f) - 1) \to 1.$$

Thus, we have $\mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) \leq N(f) - 1) \to 0$, and

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[N(\hat{f}_{K,n^{-1/6}})\right] \\
&= \lim_{n\to\infty} \Big\{ \quad N(f)\mathbb{P}\left(N(\hat{f}_{K,n^{-1/6}}) = N(f)\right) \\
&\qquad\qquad + \left(1 - \mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) = N(f))\right) \mathbb{E}\left[N(\hat{f}_{K,n^{-1/6}})|N(\hat{f}_{K,n^{-1/6}}) \geq N(f) + 1\right] \quad \Big\}.
\end{aligned}
$$

By relation (6), we have

$$
\begin{aligned}
\lim_{n\to\infty} \Big\{ \quad &N(f)\mathbb{P}\left(N(\hat{f}_{K,n^{-1/6}}) = N(f)\right) \\
&+ \left(1 - \mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) = N(f))\right) \mathbb{E}\left[N(\hat{f}_{K,n^{-1/6}})|N(\hat{f}_{K,n^{-1/6}}) \geq N(f) + 1\right] \quad \Big\} \leq N(f).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\lim_{n\to\infty} \Big\{ \quad &N(f)\mathbb{P}\left(N(\hat{f}_{K,n^{-1/6}}) = N(f)\right) \\
&+ (N(f) + 1)\left(1 - \mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) = N(f))\right) \quad \Big\} \leq N(f).
\end{aligned}
$$

and we have

$$
\lim_{n\to\infty} \left(N(f) + 1 - \mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) = N(f))\right) \leq N(f).
$$

Finally we derive

$$
\mathbb{P}(N(\hat{f}_{K,n^{-1/6}}) = N(f) \leq k) \to 1, \qquad \text{and} \qquad \mathbb{P}(h_{crit,k} \leq n^{-1/6}) \to 1.
$$

For $N(f) = 1$, (6) and the fact that $N(\hat{f}_{K,n^{-1/6}}) \geq 1$ imply that

$$
\mathbb{P}\left(N(\hat{f}_{K,n^{-1/6}}) = 1 = N(f) \leq k\right) \to 1,
$$

and that, again,

$$
\mathbb{P}(h_{crit,k} \leq n^{-1/6}) \to 1.
$$

Thus, for every positive $N(f)$, for all $\varepsilon > 0$, there exists an integer $\tilde{n} \in \mathbb{N}$ such that for all $n \geq \tilde{n}$, $n^{-1/6} < \varepsilon$ and $\mathbb{P}(h_{crit,k} \leq n^{-1/6}) < \mathbb{P}(h_{crit,k} < \varepsilon)$. Consequently we have

$$
\forall \varepsilon > 0, \ \mathbb{P}(h_{crit,k} < \varepsilon) \to 1,
$$

and then $(\mathcal{F}4)$ is verified.

A.4. **Proof of Theorem 2.** First, note that for some h, we can know $N(\hat{f}_{K,h})$ by counting the number of variations of sign of the following function

$$
g'_{K,h,\varepsilon}(x) := \begin{cases} 1 & \text{for } x \in [a_{h,(i)} - \varepsilon, a_{h,(i)}[, \ \forall i \in \{1, \ldots, \text{card}(A_h)\}, \\ -1 & \text{for } x \in [b_{h,(i)}, b_{h,(i)} + \varepsilon[, \ \forall i \in \{1, \ldots, \text{card}(B_h)\}, \\ 0 & \text{elsewhere}, \end{cases}
$$

where ε is chosen in a way that ensure that $\forall (i,j) \in \{1, \ldots, \text{card}(A_h)\} \times \{1, \ldots, \text{card}(B_h)\}, (a_{h,(i)} - b_{h,(j)}) \in]-\infty, 0] \cup]\varepsilon, \infty[$, in order to obtain a unique value of $g'_{K,h,\varepsilon}(x)$ for each x. The aim of $g'_{K,h,\varepsilon}$ is to mimic the derivative of $\hat{f}_{K,h}$. It seems to be easier to use than dirac functions involved in $\hat{f}'_{K,h}$. Besides, one can see that $N(g_{K,h,\varepsilon}) = N(\hat{f}_{K,h})$, using the fact that Proposition 2 is valid for $g_{K,h,\varepsilon}$. That is why the number of variations of sign of $g'_{K,h,\varepsilon}$ is equal to $2N(\hat{f}_{K,h}) - 1$.

Let $C_{\varepsilon,n} := \{c_{h,\varepsilon,(i)}\}_{i \in \{1, \ldots, w\}}$ be the ordered sequence compound of the sets $\{a_{h,(i)} - \frac{\varepsilon}{2}\}_{i \in \{1, \ldots, \text{card}(A_h)\}}$ and $\{b_{h,(i)} + \frac{\varepsilon}{2}\}_{i \in \{1, \ldots, \text{card}(B_h)\}}$. Let

$$
\begin{aligned}
d_{h,\varepsilon,(i)} \ := \ & \mathbf{1}\left(c_{h,\varepsilon,(i)} \in \{a_{h,(i)} - \tfrac{\varepsilon}{2}\}_{i \in \{1, \ldots, \text{card}(A_h)\}}\right) \\
& - \ \mathbf{1}\left(c_{h,\varepsilon,(i)} \in \{b_{h,(i)} + \tfrac{\varepsilon}{2}\}_{i \in \{1, \ldots, \text{card}(B_h)\}}\right),
\end{aligned}
$$

and $D_{\varepsilon,n} := \{d_{h,\varepsilon,(i)}\}_{i \in \{1, \ldots, w\}}$. Every interval where $g'_{K,h,\varepsilon}(x) \neq 0$ is represented by a $c_{h,\varepsilon,(i)}$, then the number of variations of sign is the same for $g'_{K,h,\varepsilon}$ and for $D_{\varepsilon,n}$. We write $v(D_{\varepsilon,n})$ the number of variations of sign of $D_{\varepsilon,n}$ like Schoenberg (1950) did in his article.

Now, we prove that $v(D_{\varepsilon,n}) \geq v(D_{\varepsilon,n-1})$, for $n > 1$. This property is verified if $D_{\varepsilon,n-1} = JD_{\varepsilon,n}$ where J is a totally positive matrix, following Schoenberg (1950). To define J, we first focus on the case where the last point in the sample is different from the others. This means that if Ω is our sample space, we define Ω_1 as:

$$
\Omega_1 := \{\omega \ : \ \forall i \in \{1, \ldots, n-1\}, \ X_i(\omega) \neq X_n(\omega)\}.
$$

We remark that, when our sample comes from $\omega \in \Omega_1$, $D_{\varepsilon,n-1}$ is constructed by removing two points in $D_{\varepsilon,n}$. These points correspond to $c_{h,\varepsilon,(\gamma_1)} = X_n - h - \frac{\varepsilon}{2}$ and $c_{h,\varepsilon,(\gamma_2)} = X_n + h + \frac{\varepsilon}{2}$. This is why we

have :

$$
J = \begin{pmatrix}
 & & 0 & \cdots & 0 & \cdots & 0 \\
I_{\gamma_1-1} & & \vdots & & \vdots & & \vdots \\
 & & 0 & \cdots & 0 & \cdots & 0 \\
0 & \cdots & 0 & & 0 & \cdots & 0 \\
\vdots & & \vdots & I_{\gamma_2-\gamma_1-1} & \vdots & & \vdots \\
0 & \cdots & 0 & & 0 & \cdots & 0 \\
0 & \cdots & 0 & \cdots & 0 & & \\
\vdots & & \vdots & & \vdots & I_{n-\gamma_2} & \\
0 & \cdots & 0 & \cdots & 0 & &
\end{pmatrix}
$$

where I_γ is the $\gamma \times \gamma$ identity matrix. It is straightforward to show that J is a totally positive matrix since every minor of J is positive or equal to 0 (the details are left to the reader).

If $\omega \notin \Omega_1$, then $D_{\varepsilon,n} = D_{\varepsilon,n-1}$, because A_h and B_h stay the same if we build them with $(X_1, \dots, X_n)$ or with $(X_1, \dots, X_{n-1})$. Then $J = I_w$ and is totally positive.

To conclude, we write $\tilde{N}_{K,h} : n \mapsto \tilde{N}_{K,h}(n) = N(\hat{f}_{K,k})$. Recall that $\tilde{N}_{K,h}(n) = \frac{v\left(g'_{k,h,\varepsilon}\right)+1}{2} = \frac{v(D_{\varepsilon,n})+1}{2}$. Because $n \mapsto v(D_{\varepsilon,n})$ is increasing, $\tilde{N}_{K,h}$ is also an increasing function. Let $h_{crit,k,n}$ be the critical bandwidth defined in (2) for a sample of size n, then we have:

$$\forall h < h_{crit,k,n}, \ \tilde{N}_{K,h}(n) > N(f).$$

Because $\tilde{N}_{K,h}$ increases with n, it comes that,

$$\forall h < h_{crit,k,n}, \ \forall \gamma \in \mathbb{N}, \ \tilde{N}_{K,h}(n+\gamma) > N(f),$$

then,

$$\forall \gamma \in \mathbb{N}, \ \forall h < h_{crit,k,n}, \ \tilde{N}_{K,h}(n+\gamma) > N(f).$$

Thus, with the definition of $h_{crit,k}$,

$$\forall \gamma \in \mathbb{N}, \ h_{crit,k,n+\gamma} \geq h_{crit,k,n},$$

and Theorem 2 is proven.

Acknowledgements

References

Devroye, L. P. (1987). *A Course in Density Estimation*. Birkäuser, Boston.

Devroye, L. P. and Wagner, P. J. (1980). The strong uniform consistency of kernel density estimates. In P. R. Krishnaiah (ed.), *Multivariate Analysis V: Proceedings of the fifth International Symposium on Multivariate Analysis, Volume 5*, North-Holland Pub. Co., 59–77.

Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115.

Hall, P., Minnotte, M. C., and Zhang, C. (2004). Bump hunting with non-gaussian kernels. *Ann. Statist.* **32** 2124–2141.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11** 511–514.

Mammen, E., Marron, J. S., and Fisher, N. J. (1991). Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Related Fields* **91** 115–132.

Marron, J. and Wand, M. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

Minnotte, M. C., Marchette, D. J., and Wegman, E. J. (1998). The Bumpy Road to the Mode Forest. *J. Comput. Graph. Statist.* **7** 239–251.

Minnotte, M. C. and Scott, D. W. (1993). The mode tree: A tool for visualization of nonparametric density features. *J. Comput. Graph. Statist.* **2** 51–68.

Müller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

Parzen, E. (1962). On estimation of probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

Polonik, W. (1995a). Density estimation under qualitive assumptions in higher dimensions. *J. Multivariate Anal.* **55** 61–81.

— (1995b). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Ann. Statist.* **23** 855–881.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **42** 43–47.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SCHOENBERG, I. J. (1950). On pólya frequency functions II: Variation diminishing integral operators of the convolution type. *Acta. Sci. Math. (Szeged)* **12** 97–106.

SCOTT, D. W. (1992). *Multivariate Density Estimation : Theory, Practice and Visualization*. John Wiley & Sons.

SCOTT, D. W. and TERRELL, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **62** 1131–1146.

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.

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99.

— (1983). *Probability, Statistics and Analysis*, chap. Some properties of a test for multimodality based on kernel density estimates. Cambridge University Press, 248–259.

SOW, M., DURRIEU, G., BRIOLLAIS, L., CIRET, P., and MASSABUAU, J.-C. (2011). Water quality assessment by means of HFNI valvometry and high-frequency data modeling. *Environmental Monitoring and Assessment* **182** 155–170.

VAN DORP, J. R. and KOTZ, S. (2002). The standard two-sided power distribution and its properties: with applications in financial engineering. *The American Statistician* **56** 90–99.

Raphaël Coudret : Inria cqfd & Imb, umr cnrs 5584, University of Bordeaux, 351, cours de la Libération, 33405 Talence, France,

E-mail address: rcoudret@math.u-bordeaux1.fr

Gilles Durrieu : Laboratoire de Mathématiques de Bretagne Atlantique, umr cnrs 6205, University of South Brittany, Campus de Tohannic, 56017 Vannes, France,

E-mail address: gilles.durrieu@univ-ubs.fr

Jérôme Saracco : Inria cqfd & Imb, umr cnrs 5584, University of Bordeaux, 351, cours de la Libération, 33405 Talence, France,

E-mail address: jerome.saracco@math.u-bordeaux1.fr

(a) Estimations $\hat{f}_{K,h_{crit}}$ with Gaussian kernel (dashed line) and $\hat{f}_{K,h_{crit}}$ with uniform kernel (mixed line).
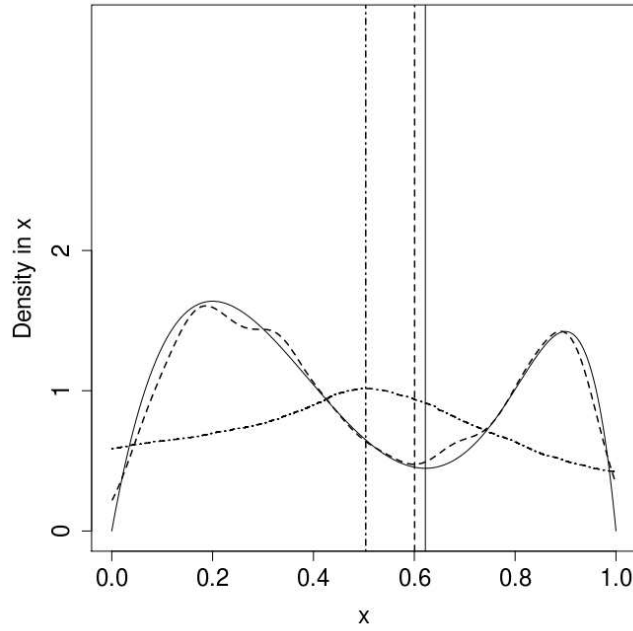
(b) Estimations $\hat{f}_P$ (long dashed line) and $\hat{f}_{K,h_{SJ}}$ (dotted line).
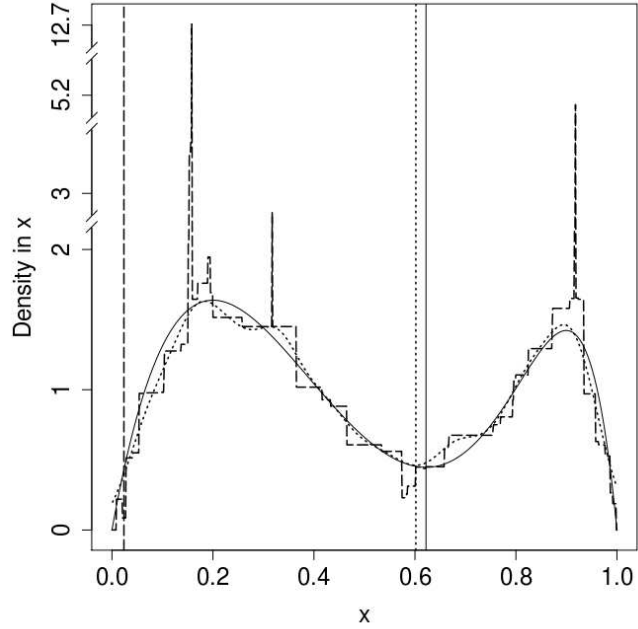
FIGURE 1. Beta mixture density f_1 (solid line), and various estimations; for $n = 1600$. Vertical lines: positions of the minimum of local minima of each plotted density. In Figure (b), vertical axis is broken between 2.2 and 2.8, between 3.5 and 5 and between 5.5 and 12.5 to be able to clearly see the shape of all curves.

(a) Kernel density estimator with h_{crit} and the uniform kernel.
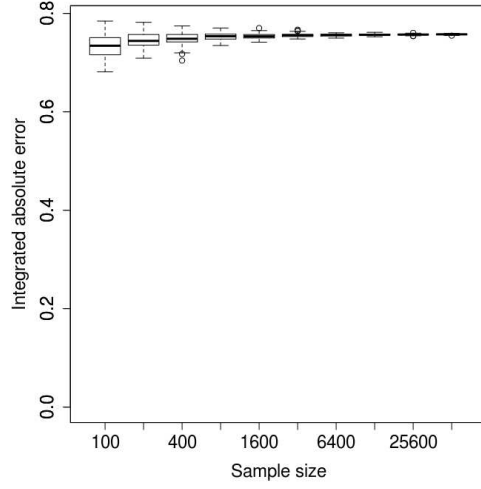
(b) Kernel density estimator with h_{crit} and the Gaussian kernel.
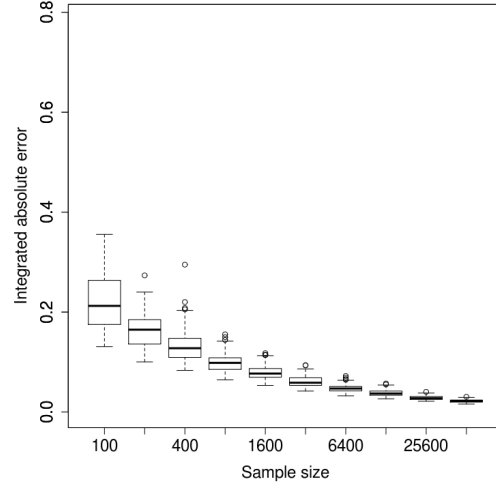
(c) Kernel density estimator with h_{SJ} and the Gaussian kernel.
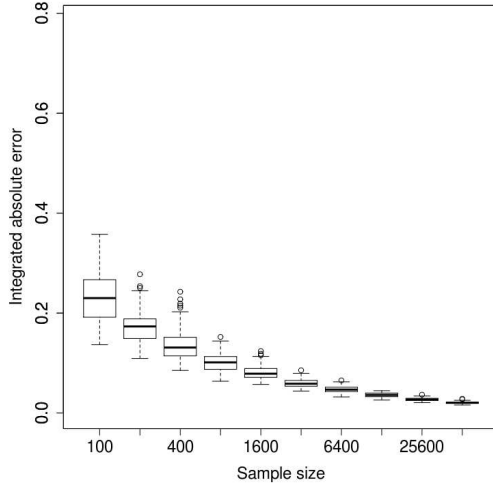
(d) Polonik's estimator.
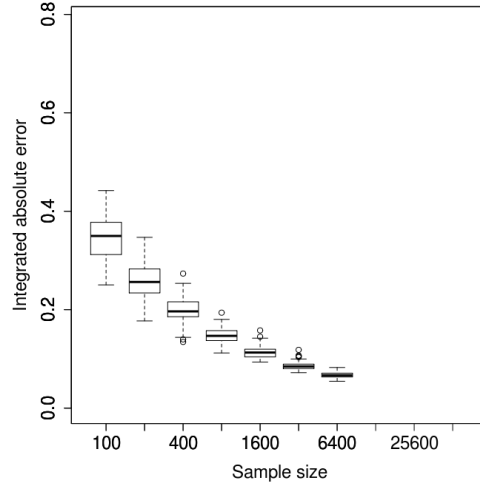
FIGURE 2. *IAE* for the beta mixture model (4) and various density estimators.

(a) Estimator related to $\hat{f}_{K,h_{crit}}$ with the uniform kernel
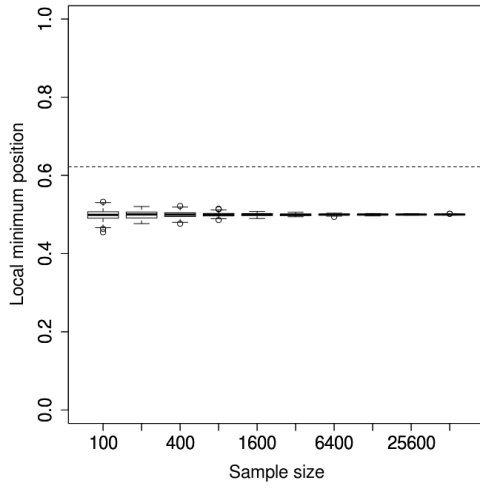
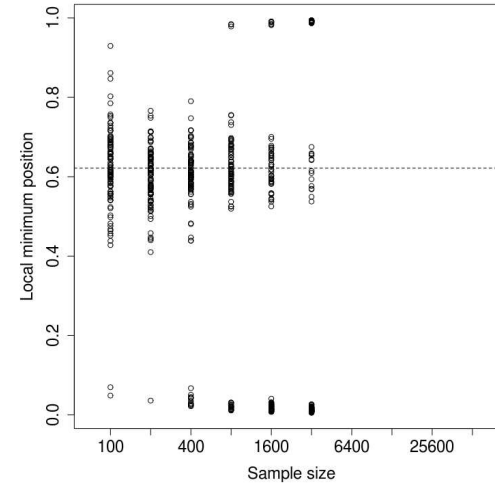(b) Estimator related to $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel
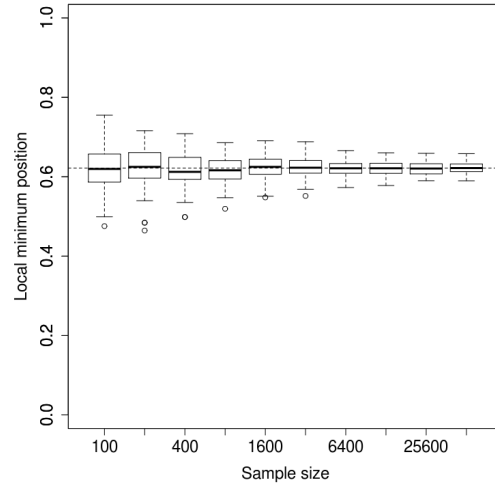
(c) Estimator related to $\hat{f}_{K,h_{SJ}}$ with the Gaussian kernel
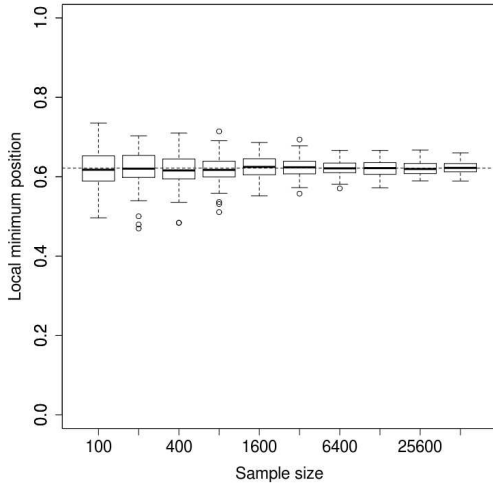
(d) Estimator related to $\hat{f}_P$

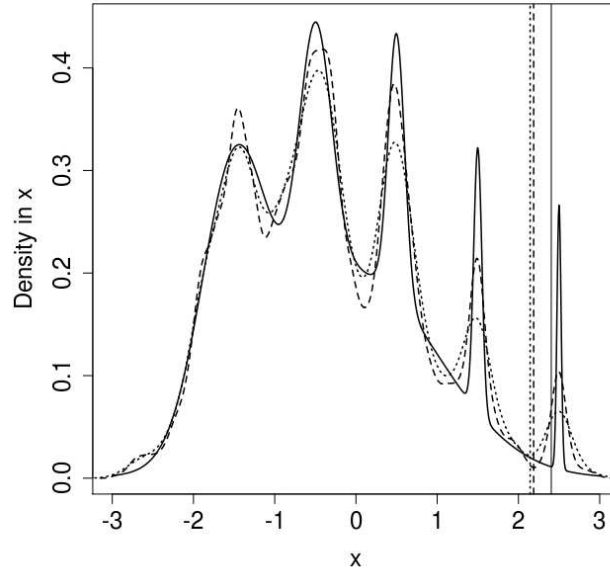FIGURE 3.   Estimations of the position of the local minimum of the density of the beta mixture model.

FIGURE 4. Asymmetric claw density f_2 (solid line). Estimations $\hat{f}_{K,h_{crit}}$ with Gaussian kernel (dashed line) and $\hat{f}_{K,h_{SJ}}$ (dotted line), for $n = 1600$. Vertical lines: positions of the minimum of local minima of each plotted density.

(a) Kernel density estimator with h_{crit} and the Gaussian kernel.

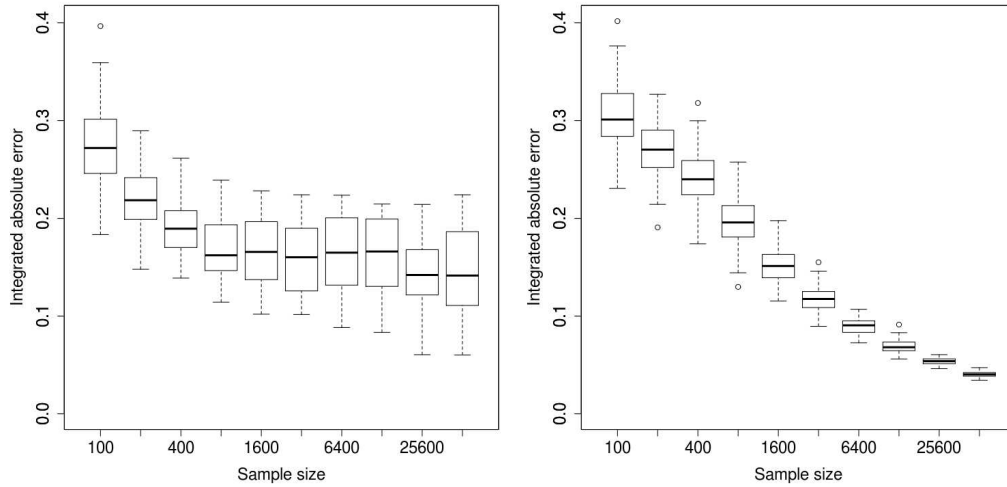(b) Kernel density estimator with h_{SJ} and the Gaussian kernel.

FIGURE 5. IAE for the asymmetric claw model (5) and various density estimators.

(a) Estimator related to $\hat{f}_{K,h_{crit}}$ with the Gaussian kernel.

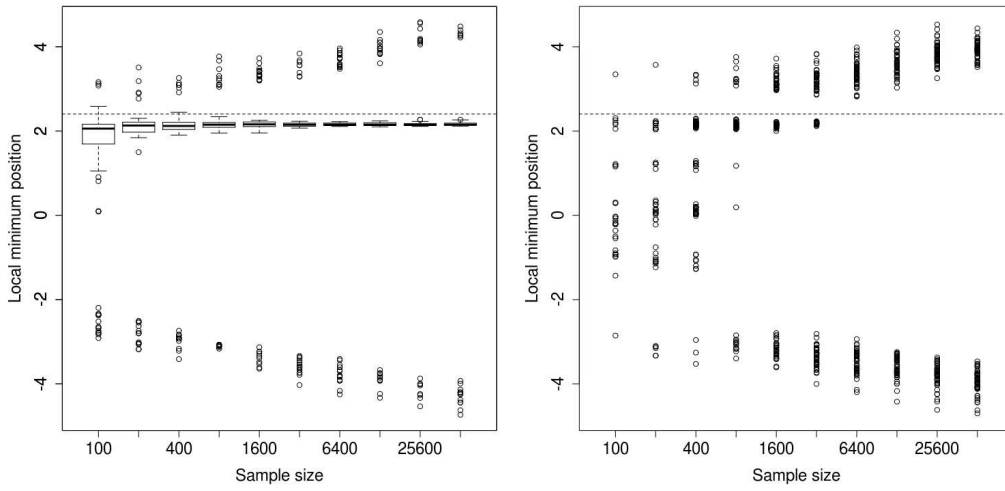(b) Estimator related to $\hat{f}_{K,h_{SJ}}$ with the Gaussian kernel.

FIGURE 6. Estimations of the position of the local minimum of the density of the asymmetric claw model.

(a) Oyster assumed to be healthy.

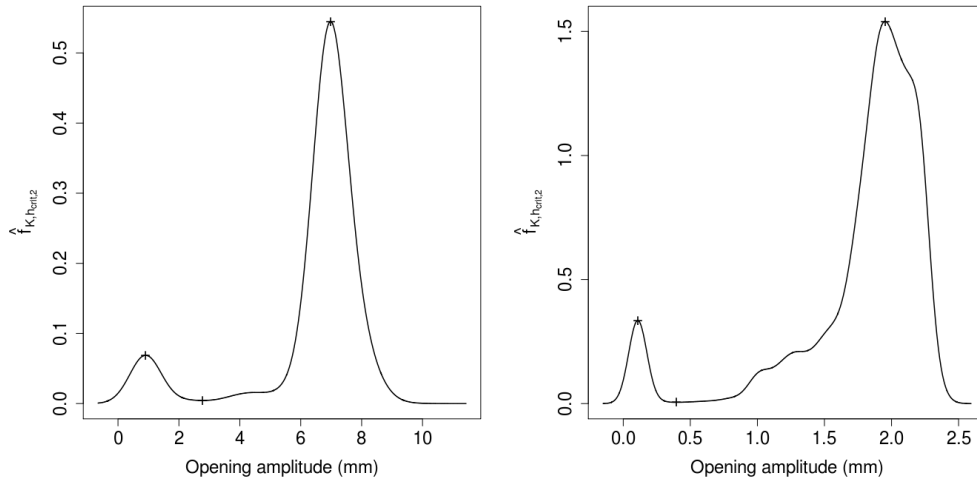(b) Oyster that dies the week that follows the day when measurements were taken.

FIGURE 7. Density estimations with $\hat{f}_{K,h_{crit,2}}$ of opening amplitudes of two oysters. Crosses indicate local extrema of the estimated densities.