Computational Statistics and Data Analysis 79 (2014) 149–164

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Random average shifted histograms

M. Bourel [a,c], R. Fraiman [b], B. Ghattas [c,*]

[a] *IMERL, Facultad de Ingenieria, Universidad de la República, Montevideo, Uruguay*
[b] *CMAT, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay*
[c] *Université d'Aix Marseille, Institut de Mathématiques de Marseille, Marseille, France*

ARTICLE INFO

Article history:
Received 27 August 2013
Received in revised form 7 March 2014
Accepted 7 May 2014
Available online 27 May 2014

Keywords:
Machine learning
Histogram
Density estimation
Bootstrap
Bagging
Average shifted histograms

ABSTRACT

A new density estimator called *RASH*, for *Random Average Shifted Histogram*, obtained by averaging several histograms as proposed in *average shifted histograms*, is presented. The principal difference between the two methods is that in RASH each histogram is built over a grid with random shifted breakpoints. The asymptotic behavior of this estimator is established for the one-dimensional case and its performance through several simulations is analyzed. *RASH* is compared to several classic density estimators and to some recent ensemble methods. Although *RASH* does not always outperform the other methods, it is very simple to implement, being also more intuitive. The two dimensional case is also analyzed empirically.

1. Introduction

There is no doubt that, in regression and classification, ensemble learning, which consists on combining several models, gives rise to more complex models that largely outperform the classical simple methods. Algorithms like Bagging (Breiman, 1996a), Boosting (Freund and Schapire, 1997), Stacking (Breiman, 1996b; Wolpert, 1992) and Random Forests (Breiman, 2001) have been deeply studied both from the standpoint of theory and that of applications and they have evolved into many variants achieving very high performances when tested over tens of different data sets from the machine learning benchmark. These algorithms have been designed for supervised learning, initially restricted to regression or binary classification. Several extensions are actually under study: multivariate regression and multi-class learning, among others.

Nevertheless, there exist very few extensions of ensemble methods for unsupervised learning such as clustering analysis or density estimation. In this work we present a contribution to this last case, which is an important problem in statistics. Some extensions of Boosting (Di Marzio and Taylor, 2004), Bagging (Ridgeway, 2002; Rosset and Segal, 2002) and Stacking (Smyth and Wolpert, 1999) to density estimation have been already considered. Other approaches inspired from Bagging and Stacking have also been studied empirically (Bourel and Ghattas, 2013).

We suggest a new simple algorithm, *Random Average Shifted Histogram* (*RASH*) for density estimation aggregating histograms which are "weak learners" in this context. Our idea arises from the average shifted histogram introduced by Scott (1992): to avoid the problem of the histogram's origin choice, this method averages several histograms built using different shifts of the breakpoints over a fixed grid. The main difference is that in *RASH* the breakpoints are randomly shifted. We introduce thus a random breakpoints histogram (RH-estimate) and study its asymptotic properties. Then we build up our aggregation estimate by averaging M RH-estimates. We show by extensive simulations that this kind of aggregation gives

[*] Corresponding author.
E-mail addresses: mbourel@fing.edu.uy (M. Bourel), rfraiman@cmat.edu.uy (R. Fraiman), badih.ghattas@univ-amu.fr (B. Ghattas).

http://dx.doi.org/10.1016/j.csda.2014.05.004
0167-9473/© 2014 Elsevier B.V. All rights reserved.

rise to better estimates. We compare our algorithm to Average Shifted Histogram (*ASH*) and to several classic algorithms used in the literature. Besides being simple, our approach seems to be more intuitive and shows very high accuracy.

Section 2 gives a brief description of the histogram and its main asymptotic properties. Our algorithm is presented in Section 3. In Section 4 we provide asymptotic results for the RH-estimate in the one-dimensional case: consistency, asymptotic normality and rates of convergence. Section 5 describes the simulation study, where we compare our proposal with several competitors and provide an extension to the two dimensional configurations. All proofs are given in the Appendix.

2. Some density estimators

We start fixing some notations and describing the algorithms we will compare with *RASH*. We also recall some important results about the histogram, kernel density estimator, average shifted histogram, and an aggregated model selection introduced in Samarov and Tsybakov (2007).

2.1. Histogram

We consider, for an i.i.d. sample $X_1, \dots, X_n$ of random variables with density f, L_n intervals $I_{1,n} \dots, I_{L_n,n}$ where $|I_{j,n}| = \left|\left[a_j(n), a_{j+1}(n)\right)\right| = h_n$ for all n, and $h_n \downarrow 0$ as $n \to +\infty$.

The ordinary histogram (*Hist*) is defined as:

$$\widehat{f}_{n,0}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \sum_{j=1}^{L} \mathbb{1}_{I_{j,n}}(X_i)\mathbb{1}_{I_{j,n}}(x).$$

If $x \in I_{j,n}$, we have that:

$$\mathbb{E}\big(\widehat{f}_{n,0}(x)\big) = \frac{1}{nh_n} \sum_{i=1}^{n} \sum_{j=1}^{L} \mathbb{P}(X_i \in I_{j,n})\mathbb{1}_{I_{j,n}}(x) = \frac{1}{nh_n} n\mathbb{P}(X_i \in I_{j,n}) = \frac{1}{|I_{j,n}|} \int_{I_{j,n}} f(t)dt$$

$$Var\big(\widehat{f}_{n,0}(x)\big) = \frac{1}{n^2h_n^2} nVar\left(\mathbb{1}_{I_{j,n}}(X_i)\right) \le \frac{1}{nh_n^2}\mathbb{P}(X_1 \in I_{j,n}) = \frac{1}{nh_n}\frac{\mathbb{P}(X_1 \in I_{j,n})}{h_n}.$$

When $h_n \to 0$ and $nh_n \to \infty$ we get the classical properties for the histogram:

$$\mathbb{E}\big(\widehat{f}_{n,0}(x)\big) \to f(x), \qquad Var\big(\widehat{f}_{n,0}(x)\big) \to 0.$$

The histogram depends on two parameters: the bin width h_n and the origin x_0. There is a huge literature that proposes several optimal choices for h_n. If we suppose that the underlying density f is Gaussian, it can be shown (see Scott, 1979) that an optimal choice for h is:

$$h_{\text{opt}} = 3.5\widehat{\sigma}n^{-1/3},$$

where $\widehat{\sigma}$ is an estimate of the standard deviation.

2.2. Average shifted histogram

The histogram estimate may change significantly when the origin x_0 changes, even if h_n is fixed. Scott (1985), introduced the Average Shifted Histogram (*ASH*) algorithm which aims to avoid choosing x_0. It is a nonparametric density estimator which averages several histograms with different origins. Consider for example an histogram with origin x_0, bin width h, and support $\left\{[jh, (j+1)h)\right\}_{j\in\mathbb{Z}}$. For $M > 0$, let $\delta = \frac{h}{M}$ and divide each interval $\left[jh, (j+1)h\right)$ into M new subintervals $B_k = \left[k\delta, (k+1)\delta\right)$ obtaining a finer grid. For instance, if $k = 0$, we divide $[0, h)$ into M intervals $B_0 = [0, \delta), B_1 = [\delta, 2\delta), \dots, B_{M-1} = \left[(M-1)\delta, M\delta\right) = \left[(M-1)\delta, h\right)$. Let ν_k be the number of observations falling in B_k and $x \in B_0 = [0, \delta)$. Then there are M "shifted" histograms with bin width $h = M\delta$ which cover B_0. The value of the first one at x is:

$$\widehat{f}_1(x) = \frac{\nu_{1-M} + \nu_{2-M} + \dots + \nu_0}{nh}.$$

The value of the second one at x is

$$\widehat{f}_2(x) = \frac{\nu_{2-M} + \nu_{3-M} + \dots + \nu_0 + \nu_1}{nh}.$$

The Mth final shifted histogram which covers $[0, \delta)$ takes at x the value:

$$\widehat{f}_M(x) = \frac{\nu_0 + \dots + \nu_{M-1}}{nh}.$$

The *ASH* estimate at x is defined as the average of these M estimators:

$$\widehat{f_{ASH}}(x) = \frac{1}{M}\sum_{m=1}^{M}\widehat{f_m}(x).$$

Since the frequency of ν_{1-M} is $\frac{1}{M}$, the frequency of ν_{2-M} is $\frac{2}{M}, \ldots,$ the frequency of ν_{M-1-M} is $\frac{M-1}{M}$, while that of ν_0 is $\frac{M}{M} = 1$, that of ν_1 is $\frac{M-1}{M}, \ldots,$ that of ν_{M-1} is $\frac{1}{M}$, we can rewrite the estimator as

$$\widehat{f_{ASH}}(x) = \frac{1}{M}\sum_{j=1-M}^{M-1}\left(\frac{M-|j|}{nh}\right)\nu_{k+j} \quad \forall\, x \in B_k,$$

which may be written:

$$\widehat{f_{ASH}}(x) = \frac{1}{nh}\sum_{j=1-M}^{M-1}\left(1-\frac{|j|}{M}\right)\nu_{k+j} \quad \forall\, x \in B_k.$$

For further details see Scott (1985, 1992), Scott and Härdle (1992) and Scott (2009). In particular, it is proved that if we have a sample $\{X_1, \ldots, X_n\}$, as $M \to +\infty$, the *ASH* estimator converges to a kernel density estimator where the kernel is triangular,

$$\lim_{M\to+\infty}\widehat{f_{ASH}}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left(1-\left|\frac{x-X_i}{h}\right|\right)\mathbb{1}_{(-1,1)}\left(\frac{x-X_i}{h}\right).$$

Fig. 1 shows the plot of the *ASH* estimate for $n = 500$ for different values of M together with *Hist* and the kernel density estimator *Kde* with a triangular kernel using Silverman's rule of thumb (Silverman, 1986) to find an optimal bandwidth h, for a three Gaussian mixture density.

2.3. A model selection algorithm by Samarov and Tsybakov

We now describe briefly the algorithm proposed in Samarov and Tsybakov (2007), denoted *ST*, in what follows. This one is in fact a special case of a two-fold cross validation model selection method. Suppose we have a sample $\aleph = \{X_1, \ldots, X_n\}$ with common probability density f. We split the data into two disjoint subsets $\aleph = \aleph_1 \cup \aleph_2$ with $|\aleph_1| = n_1$, $|\aleph_2| = n_2$ and $n = n_1 + n_2$. With the first part we build M estimators $\hat{f}_1, \ldots, \hat{f}_M$ and with the other we select $\widehat{f_{\tilde{M}}}$ such that $\tilde{M} = \text{Argmin}_{1\le m\le M} J_m$ where

$$J_m = -\frac{2}{n_2}\sum_{X_i\in\aleph_2}\hat{f}_m(X_i) + \int \hat{f}_m^2.$$

It is well known that J_m is an unbiased estimator of the *Mean Integrated Squared Error, (MISE)* of $\hat{f}_m$ up to an additive constant $\|f\|^2$ (since $\mathbb{E}(J_m) = \mathbb{E}\big(\|f - f_m\|^2\big) - \|f\|^2$).

3. The random average shifted histogram algorithm

Our algorithm *RASH* aggregates M histograms and works as follows. Let $\widehat{f_{n,0}}$ be the histogram obtained with the data set at hand using equally spaced breakpoints $a_0, \ldots, a_L$. We define a sequence of histograms $\left\{\widehat{f_n^{(m)}}\right\}_{m=1,\ldots,M}$ constructed over the same initial data set but using a randomly shifted set of equally spaced breakpoints $a_0 + e_n^{(m)}, a_1 + e_n^{(m)}, \ldots, a_L + e_n^{(m)}$ where $e_n^{(m)}$ is a random variable with density g_n. If for $j = 1, \ldots, L$, we denote $I_{j,n}\left(e_n^{(m)}\right) = I_{j,n} + e_n^{(m)} = \left[a_{j-1} + e_n^{(m)}, a_j + e_n^{(m)}\right)$ we have that:

$$\widehat{f_n^{(m)}}(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\sum_{j=1}^{L}\mathbb{1}_{I_{j,n}\left(e_n^{(m)}\right)}(X_i)\mathbb{1}_{I_{j,n}\left(e_n^{(m)}\right)}(x). \tag{1}$$

Our final estimator is defined to be the average of the histograms $\hat{f}_n^{(1)}, \ldots, \hat{f}_n^{(M)}$:

$$\widehat{f_M}(x) = \frac{1}{M}\sum_{m=1}^{M}\widehat{f_n^{(m)}}(x),$$

where the $e_n^{(m)}$ are independent random variables with density g_n for each $m = 1, \ldots, M$. The algorithm is detailed in Fig. 2.

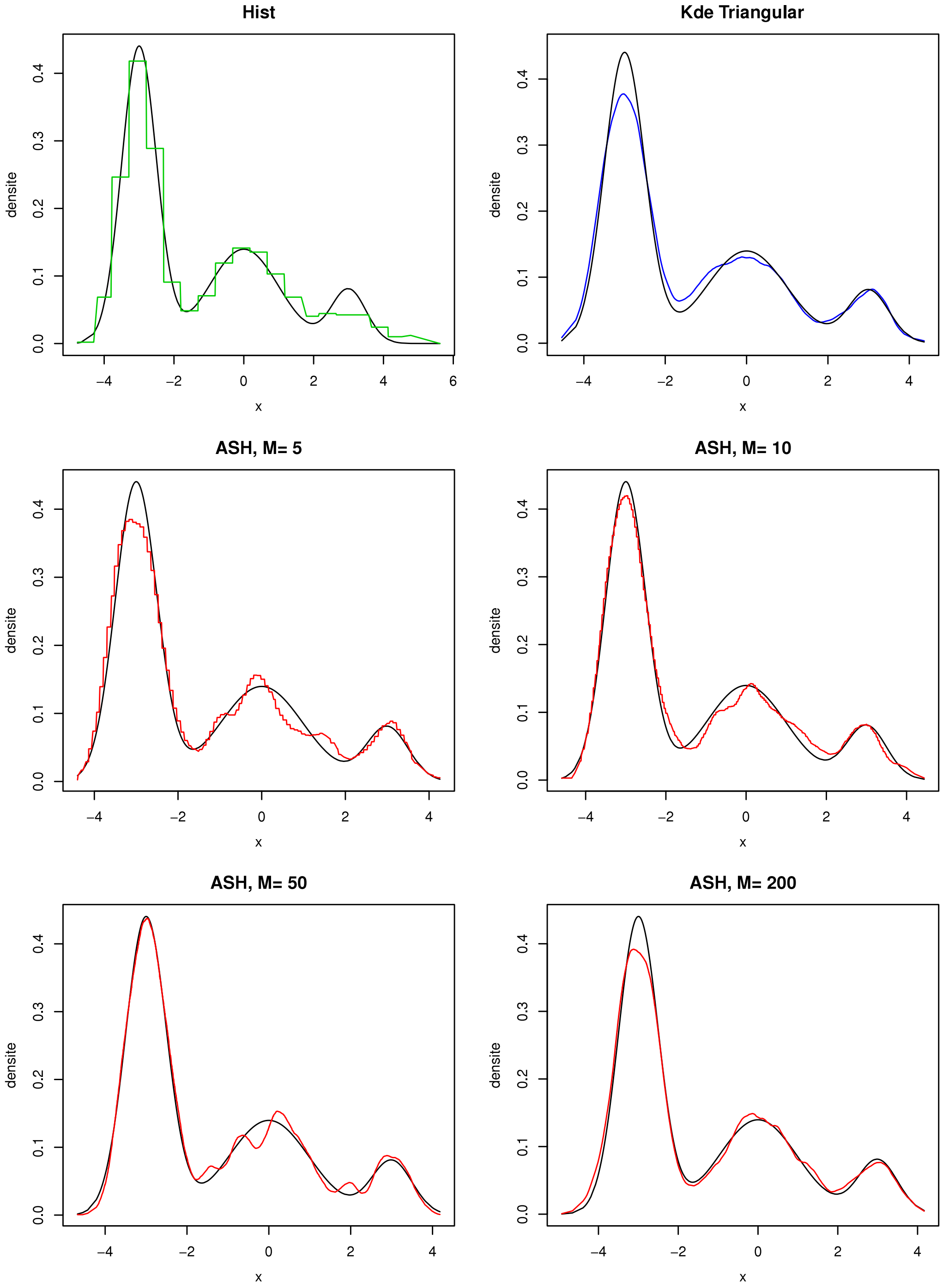

Fig. 1. Adjustment of the classical methods for a three Gaussian mixture and $n = 500$. At the top panel we plot *Hist* (in green) and *Kde* with a triangular kernel (in blue) together with the true density (in black). In the middle and bottom panels we plot the *ASH* estimators (in red) for different values of M. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Asymptotic results

As most of the results in this section will refer to only one RH-estimator we will write $\widehat{f}_n$ instead of $\widehat{f}_n^{(m)}$ and e_n instead of $e_n^{(m)}$. We will first show the consistency of the RH-estimate $\widehat{f}_n$ and then its asymptotic normality. Through all the manuscript $\{X_1, \ldots, X_n, \ldots\}$ will be a sequence of i.i.d. random variables with density f. We use the notations $\xrightarrow[n]{P}$, $\xrightarrow[n]{D}$ and $\xrightarrow[n]{a.s}$ for the different types of convergence: in probability, in distribution and almost sure respectively.

```
1. Let $\aleph$ be the original sample, $\widehat{f}_{n,0}$ be the histogram constructed over $\aleph$ and
   $I_{1,n}, \dots, I_{L,n}$, where $I_{j,n} = [a_{j-1}, a_j)$, the set of the intervals of $f_{n,0}$.
2. For $m = 1$ to $M$:
   (a) Let $e_n^{(m)}$ be a real random variable with density $g_n$.
   (b) Set $\mathcal{I}^m = \left\{ I_{1,n}\left(e_n^{(m)}\right), I_{2,n}\left(e_n^{(m)}\right), \dots, I_{L,n}\left(e_n^{(m)}\right) \right\}$ the modified intervals obtained by
       setting
           $I_{j,n}\left(e_n^{(m)}\right) = \left[a_j + e_n^{(m)}, a_{j+1} + e_n^{(m)}\right)$
       for all $j = 1, \dots, L$.
   (c) Set $\widehat{f}_n^{(m)}$ to be the histogram constructed over $\aleph$ using the intervals in $\mathcal{I}^m$.
3. Output: $\widehat{f}_M(x) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}_n^{(m)}(x)$.
```

Fig. 2. Aggregating histograms using randomly shifted breakpoints (*RASH*).

4.1. Consistency

In order to be able to perform aggregation consistently we need to show that each estimate $\widehat{f_n}(x)$ is a consistent estimate of f. The following theorem states the L^2 convergence of $\widehat{f_n}$.

Theorem 1 (*Consistency*). *Assume that f is bounded in a neighborhood of x, and that the sequence h_n fulfills*

$$h_n \to 0, \quad nh_n \to +\infty, \quad \textit{and} \quad \frac{\mathbb{E}(|e_n|)}{h_n^2} \to 0.$$

Then $\mathbb{E}\big(\widehat{f_n}(x) - f(x)\big)^2 \to 0.$

The last condition of Theorem 1 relates the mass concentration of g_n around 0, to the parameter h_n. The following examples illustrate this assumption when g_n is the uniform distribution and the normal distribution.

Remark 1.
- Suppose that $e_n \sim U[-\eta_n, \eta_n]$. Then $\mathbb{E}(|e_n|) = \frac{\eta_n}{2}$, and the condition of the previous theorem holds if $\frac{\eta_n}{h_n^2} \to 0$.
- Suppose that $e_n \sim N(0, \sigma_n^2)$. Then $\mathbb{E}(|e_n|) = \frac{\sqrt{2}\sigma_n}{\sqrt{\pi}}$, and the condition of the previous theorem holds if $\frac{\sigma_n}{h_n^2} \to 0$.

It is straightforward that *RASH* is consistent as it is an average of a finite number of consistent estimators.

4.2. Asymptotic normality of the shifted histogram

The following theorem states asymptotic normality of the RH-estimate, $\widehat{f_n}$.

Theorem 2 (*Asymptotic Normality*). *Assume that f is bounded in a neighborhood of x, and that the sequence h_n fulfills*

$$h_n \to 0, \quad nh_n \to \infty, \qquad \sqrt{\frac{n}{h}}\mathbb{E}(|e_n|) \to 0.$$

Then

$$\sqrt{nh_n}\big(\widehat{f_n}(x) - f(x)\big) \underset{n}{\overset{D}{\to}} N\big(0, f^2(x)\big). \tag{2}$$

The proof is a consequence of the classical result for regular histograms like $\widehat{f}_{n,0}$, the Slutsky Theorem and the following proposition:

Proposition 1. *Under the assumptions of Theorem 2 we have that*

$$\sqrt{nh_n}\big(\widehat{f_n}(x) - \widehat{f}_{n,0}(x)\big) \underset{n}{\overset{P}{\to}} 0. \tag{3}$$

Remark 2.
- In the case of $e_n \sim U[-\eta_n, \eta_n]$, if $\eta_n\sqrt{\frac{n}{h_n}} \to 0$ Theorem 2 holds.
- In the case of $e_n \sim N(0, \sigma_n^2)$, if $\sigma_n\sqrt{\frac{n}{h_n}} \to 0$ Theorem 2 holds.

4.3. Rates of convergence

In this section we provide strong consistency and strong rates of convergence results for the RH-estimate. The results are based on the following Lemma.

Lemma 1. *Assume that* $\frac{nh_n^2}{\log n} \underset{n}{\rightarrow} \infty$. *Then we have that:*

$$\widehat{f_n}(x) - \mathbb{E}\big(\widehat{f_n}(x)\big) \underset{n}{\overset{a.s}{\longrightarrow}} 0. \tag{4}$$

Theorem 3 (*Strong Consistency*)**.** *If* $h_n \rightarrow 0$ *and* $\frac{nh_n^2}{\log n} \underset{n}{\rightarrow} \infty$, *we have that:*

$$\widehat{f_n}(x) - f(x) \underset{n}{\overset{a.s}{\longrightarrow}} 0. \tag{5}$$

Then RASH estimate is strongly consistent. Finally the following theorem states the strong rates of convergence of the RH-estimate.

Theorem 4 (*Strong Rates of Convergence*)**.** *If* $\beta_n \rightarrow \infty$, $\frac{nh_n^2}{\beta_n^2 \log n} \underset{n}{\rightarrow} \infty$, $\beta_n h_n \rightarrow 0$, *and the density* f *is Lipschitz in a neighborhood of* x *we have that:*

(i)

$$\beta_n \left|\widehat{f_n}(x) - f(x)\right| \underset{n}{\overset{a.s}{\longrightarrow}} 0.$$

(ii) *In particular, if* $\beta_n = n^\beta$,

$$\forall \beta < \frac{1}{4} \quad \text{we have that } \beta_n \left|\widehat{f_n}(x) - f(x)\right| \underset{n}{\overset{a.s}{\longrightarrow}} 0.$$

5. Experiments

First, we give some simulations to compare our method with the other algorithms described in Section 2 over 12 one-dimensional simulation models. Next, with the aim to extend and generalize them, we give some preliminary results obtained by direct extension of this method to the two-dimensional case.

5.1. The one dimensional case

We consider several data generating models often used in the literature. We first show how our algorithm adjusts for the different models, and that the adjustment error decreases monotonically when increasing the number of histograms used in the aggregation. We then compare our method with the methods described in Section 2: *Hist*, *Kde*, *ASH* and *ST*.

5.1.1. Models used for the simulations

Twelve models found in the papers we have referenced are used in our simulations. We denote them by $\mathcal{M}1, \ldots, \mathcal{M}12$ and we group them according to their difficulty level.

- Some standard densities used in Di Marzio and Taylor (2004) and Rigollet and Tsybakov (2007):
  ($\mathcal{M}1$): the standard Gaussian density $N(0, 1)$.
  ($\mathcal{M}2$): the Exponential density with parameter one.
  ($\mathcal{M}3$): the Chi-square density χ^2_{10}.
  ($\mathcal{M}4$): the Student density with four degrees of freedom t_4.
- Some Gaussian mixtures used in Di Marzio and Taylor (2004) and Smyth and Wolpert (1999):
  ($\mathcal{M}5$): $0.5N(-1, 0.3) + 0.5N(1, 0.3)$.
  ($\mathcal{M}6$): $0.25N(-3, 0.5) + 0.5N(0, 1) + 0.25N(3, 0.5)$.
  ($\mathcal{M}7$): $0.55N(-3, 0.5) + 0.35N(0, 1) + 0.1N(3, 0.5)$.
- Gaussian mixtures used in Rigollet and Tsybakov (2007) and Marron and Wand (1992):
  ($\mathcal{M}8$): the Claw density, $0.5N(0, 1) + \sum_{i=0}^{4} \frac{1}{10} N\left(\frac{i}{2} - 1, \frac{1}{10}\right)$.
  ($\mathcal{M}9$): the Smooth Comb Density, $\sum_{i=0}^{5} \frac{2^{5-i}}{63} N\left(\mu_i, \sigma_i^2\right)$ where $\mu_i = \frac{65-96\frac{1}{2^i}}{21}$ and $\sigma_i^2 = \frac{\left(\frac{32}{63}\right)^2}{2^{2i}}$.

- Mixture density with highly inhomogeneous smoothness used in Rigollet and Tsybakov (2007):
  $(\mathcal{M}10)$: $0.5N(0,1)+0.5\sum_{i=1}^{14}\mathbf{1}_{\left(\frac{2(i-1)}{T},\frac{2i-1}{T}\right]}$.
- Finally we include in our study two simple models known to be challenging for density estimators:
  $(\mathcal{M}11)$: a triangular density with support $[0,2]$ and maximum at 1.
  $(\mathcal{M}12)$: the Beta density with parameters 2 and 5.

All the simulations were performed with the R software, and for models $\mathcal{M}8$ and $\mathcal{M}9$ we use the benchden package. Fig. 3 shows the shape of the densities we have used to generate the data sets together with their estimation obtained from *Kde* (with a Gaussian kernel and a bandwidth chosen with Silverman's rule of thumb) and the *RASH* algorithm for $M=200$ and for $n=500$ observations.

5.1.2. Tuning the algorithms

We give here our choices for the parameters needed to be tuned for each method used in our comparisons.

For the *Kernel density estimator* we use two different types of kernel, Gaussian (*Kde*) and triangular (*Kdet*), and some common data-driven bandwidth selectors:

- Silverman's rules of thumb (Silverman, 1986) using factor 1.06 (*Nrd*) and factor 0.9 (*Nrd*0),
- the unbiased cross-validation rule (*UCV*, Bowman, 1984),
- the Sheather Jones plug-in method (*SJ*, Sheather, 2004).

A detailed description of these selectors are given in the appendix of Bourel and Ghattas (2013).

For *ST* we fix a grid with 10, 20 and 50 equally spaced breakpoints, i.e $\mathcal{G}=\{10,20,50\}$. For each $L\in\mathcal{G}$, we consider M RH-estimators $\widehat{f_n}^{(m)}$ with Gaussian perturbation $N(0,h)$ as defined in Section 3. We recall that h is the distance between two consecutive breakpoints. We calculate the value of J_m for each $1\leq m\leq M$ and each $g\in\mathcal{G}$, and the histogram with the minimum value of J_m is the selected predictor.

For the other methods that involve histograms we fix a grid of numbers of equally spaced breakpoints $\mathcal{G}=\{5,10,15,20,25,30,35,40,45,50\}$.

For *Hist*, the number of breaks $L\in\mathcal{G}$ is optimized choosing the value which maximizes the log-likelihood over 200 test samples.

We also use two variants of *Hist*. The difference with the classic histogram is over the grid used to construct them. Another way to fix the grid is to select on the range of the data L points which maximize the log-likelihood of the histogram built over them. This method proposed in Klemelä (2009) for the multivariate case, provides greedy partitions and for this simulation we adapted it to the one-dimensional (*HistG*) case. Such partitions give splits with axes parallel to the coordinate axes and are similar to the one obtained by Classification And Regression Trees (CART, Breiman et al., 1984). For this reason, we also construct an histogram over a grid obtained by a partition of the data made by CART using the package rpart of **R**. We denote this kind of histogram by *HistC*.

For *ASH* the number of breaks $L\in\mathcal{G}$ is optimized again choosing the value which maximizes the log likelihood of the estimator over 200 test samples drawn from the same distribution as the learning sample. For each model, we aggregate $M=150$ histograms with L equally spaced breakpoints.

For *RASH* we aggregate $M=150$ histograms with L breakpoints. Each of them is built using the sorted values of a mesh random grid obtained by adding a Gaussian noise $e_n^{(m)}\sim N(0,h)$ where h is the distance between two consecutive breakpoints of the original histogram (this one has equidistant breakpoints). The number of breakpoints is optimized by testing different values for each of them over a fixed grid with $L\in\mathcal{G}$ of equally spaced breakpoints for each case. The number of breakpoints retained for each model is the one which maximizes the log-likelihood over 200 independent test samples drawn from the corresponding model.

We also consider the aggregation by RASH with M random perturbations of an histogram obtained by maximization of log-likelihood as in *HistG* and M random perturbations of an histogram obtained by CART as in *HistC*. These methods are called *RASHG* and *RASHC* respectively.

The optimal values of L we found for these methods which involves histograms are given in Table 1 for three different sample sizes $n=100$, 500 and $n=1000$.

5.1.3. Results

The performance of each model is evaluated using the *MISE*. It is estimated as the average of the integrated squared error over 200 Monte Carlo simulations. Fig. 4 shows for $n=500$ how the MISE varies when increasing the number of histograms in *RASH* for the 12 models. These graphics show clearly that the contribution of the aggregation to the reduction of the MISE is significant. For most of the models, the error seems to be stable after about 50 iterations.

In Tables 2–4 we show the results we obtain for each model and each value of n. In these tables we give the values of 100× MISE for each method and the simulation model for the three values of n. For the *Kde* in the two versions (*Kde* for the Gaussian kernel and *Kdet* for the triangular kernel) we kept the best result among the four choices of bandwidth selectors (Nrd, Nrd0, UCV and SJ), the best choice being between brackets. The best performance is indicated in boldface and the second one in blue. We use the same number of observations for the learning and the test samples.

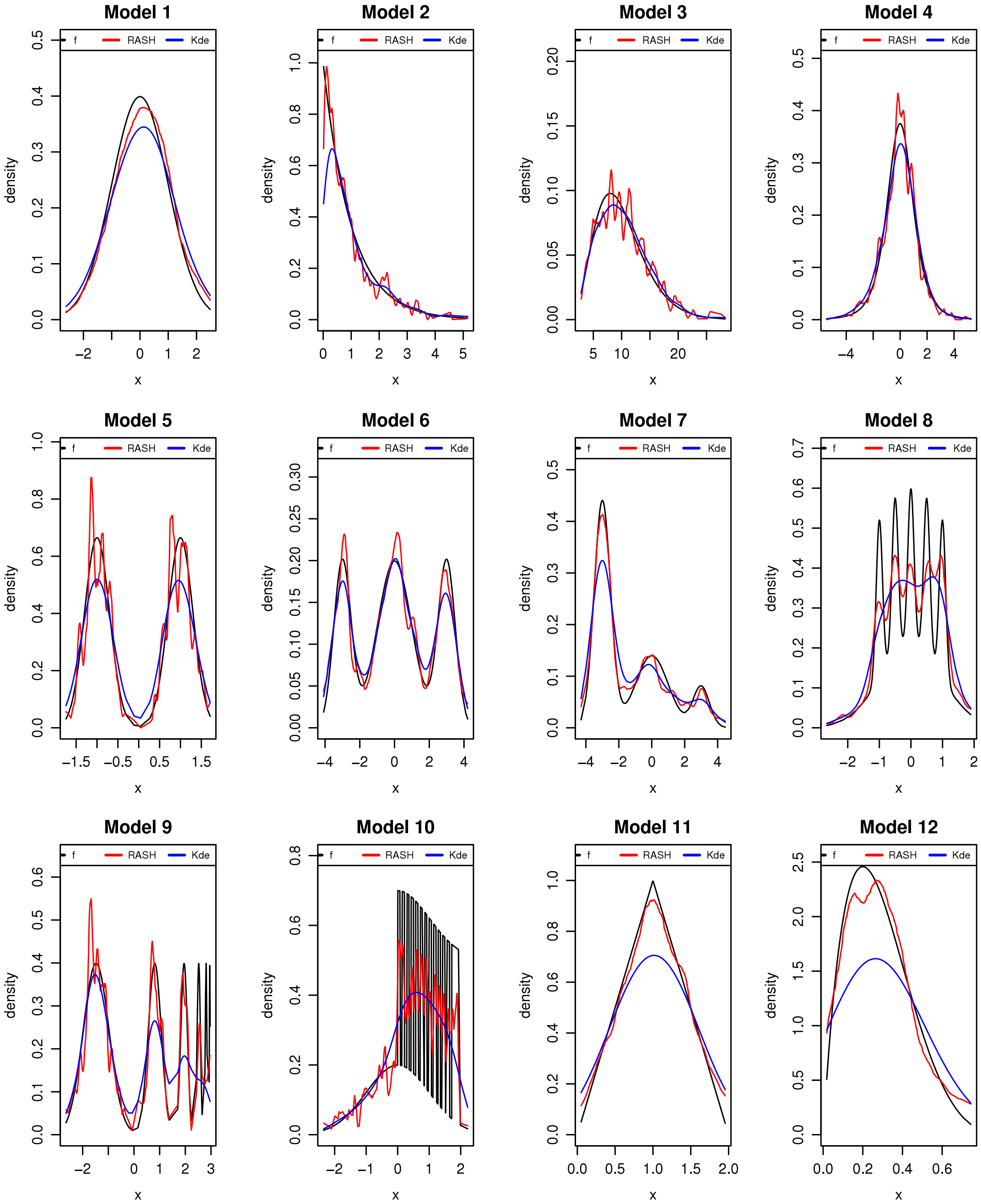

Fig. 3. Adjustment of RASH (in red) and Kde (in blue) estimates to the true density (in black) for the 12 models and for $n = 500$ and $M = 200$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

It is clear that no method outperforms completely all the others in all the cases. Nevertheless, we can see that in general when *RASH* or one of its variants is not the best method, it is very often the second one with better performance and when it comes in the third position, its accuracy is very close to that of *ASH*. The ensemble methods estimators considered in this work outperform well in general the classical histogram and the kernel density estimator in most cases. *RASH* has similar performance as *ASH* on difficult models ($\mathcal{M}_8$, $\mathcal{M}_9$ and $\mathcal{M}_{10}$) but outperforms clearly the others for the simple ones, in particular for models $\mathcal{M}_{11}$ and $\mathcal{M}_{12}$. Moreover *RASH* adjusts clearly better than *Kde* in non classical models such as mixing models ($\mathcal{M}_5$ to $\mathcal{M}_7$) and the more complicated models ($\mathcal{M}_8$ to $\mathcal{M}_{12}$). As expected, for all the methods the error decreases when increasing the sample size.

Table 1
Optimal number of breakpoints used for *Hist* (H), *HistG* (HG), *RASH*, *RASHG* and *ASH* for each model and for each value of n.

Model	$n = 100$					$n = 500$					$n = 1000$				
	H	HG	RASH	RASHG	ASH	H	HG	RASH	RASHG	ASH	H	HG	RASH	RASHG	ASH
$\mathcal{M}_1$	50	45	10	10	10	50	45	10	10	10	50	45	10	10	10
$\mathcal{M}_2$	50	50	50	50	10	50	50	50	50	10	50	50	50	50	10
$\mathcal{M}_3$	50	45	45	45	15	50	45	45	45	15	50	45	45	45	15
$\mathcal{M}_4$	50	35	45	15	20	50	35	45	15	20	50	35	45	15	20
$\mathcal{M}_5$	50	50	45	20	20	50	50	45	20	20	50	50	45	20	20
$\mathcal{M}_6$	50	50	10	10	10	50	50	10	10	10	50	50	10	10	10
$\mathcal{M}_7$	50	40	50	50	15	50	40	50	50	15	50	40	50	50	15
$\mathcal{M}_8$	50	50	45	15	15	50	50	45	15	15	50	50	45	15	15
$\mathcal{M}_9$	50	50	50	30	50	50	50	50	30	50	50	50	50	30	50
$\mathcal{M}_{10}$	50	50	50	50	20	50	50	50	50	20	50	50	50	50	20
$\mathcal{M}_{11}$	50	50	5	5	5	50	50	5	5	5	50	50	5	5	5
$\mathcal{M}_{12}$	50	15	15	15	5	50	15	15	15	5	50	15	15	15	5

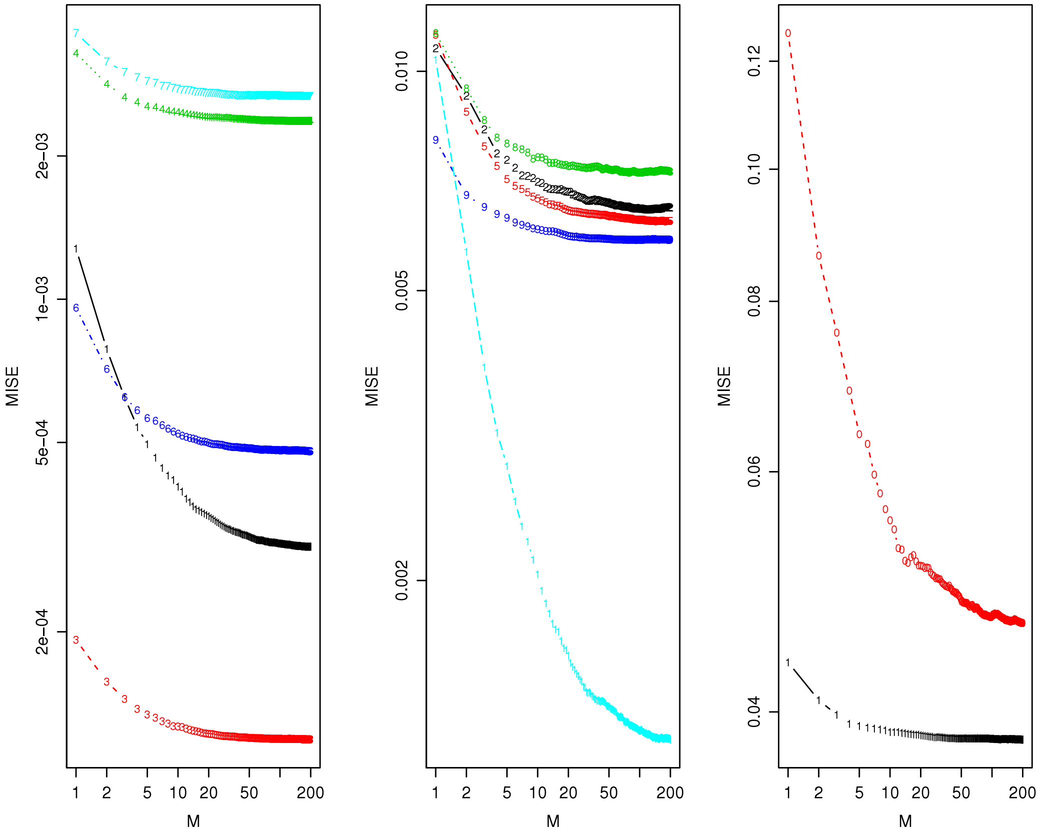

Fig. 4. MISE error versus number of aggregated histograms in *RASH* plotted on a log scale for models 1–12. The models are grouped according to the MISE scale: the left panel for models 1, 3, 4, 6 and 7, the central one for models 2, 5, 8, 9 and 11, and the right panel for models 10 and 12.

5.2. The two dimensional setting

In this section we consider some possible extensions of *RASH* to the bivariate case. This part is very incipient yet but we think that it can be of interest to extend the RASH method presented in the previous sections.

A two dimensional histogram is based over a partition of $\mathbb{R}^2$. We will compare different methods where such partition may be regular (equally spaced breakpoints in both directions) or optimized using a classical criterion on a greedy partition as in Klemelä (2009). Another use of greedy partition and histogram built over it can be found for example in Iacono and Irpino (2011).

- Hist2D: a two-dimensional histogram, with a regular equispaced grid in each of its coordinates.
- KDE2D: a bivariate kernel density estimator using optimized bandwidth following the plug-in method of Wand and Jones (1994).

Table 2
100 × MISE for each model and each method with $n = 100$, $M = 150$.

Model	Hist	HistG	HistC	ASH	RASH	RASHG	RASHC	ST	Kde	Kdet
$\mathcal{M}_1$	2.79	3.56	0.535	0.281	0.279	0.667	0.232	0.996	**0.179**(nrd0)	0.373(nrd)
$\mathcal{M}_2$	5.15	9.14	1.43	5.69	2.72	1.58	**1.26**	4.5	4.38(ucv)	2.22(nrd0)
$\mathcal{M}_3$	0.157	0.211	0.031	0.092	0.092	0.06	0.015	0.062	**0.01**(nrd)	0.021(nrd)
$\mathcal{M}_4$	1.41	2.56	0.409	0.808	0.236	0.923	**0.178**	0.803	0.187 (nrd0)	0.297(nrd)
$\mathcal{M}_5$	6.69	9.44	5.99	3.8	1.31	3.5	2.88	2.69	3.17(ucv)	**0.63**(nrd0)
$\mathcal{M}_6$	0.893	0.928	0.264	0.097	0.095	0.207	0.186	0.273	0.154(ucv)	**0.087**(nrd0)
$\mathcal{M}_7$	1.26	2.98	0.917	0.8	0.782	0.757	0.605	0.777	0.517 (ucv)	**0.189**(nrd0)
$\mathcal{M}_8$	4.02	5.52	2.68	2.31	1.6	2.62	2.17	2.75	2.13(ucv)	**1.59**(nrd0)
$\mathcal{M}_9$	2.67	2.34	2.27	1.73	1.09	1.58	1.71	1.69	1.32(ucv)	**0.938**(sj)
$\mathcal{M}_{10}$	6.27	9.03	5.9	**4.79**	4.81	5.55	5.71	5.57	6.7(nrd0)	5.39(nrd0)
$\mathcal{M}_{11}$	19.4	24.7	2.76	**0.957**	1	1.65	1.19	5.83	2.18(nrd0)	2.11(nrd)
$\mathcal{M}_{12}$	135	64.9	21.1	23.6	16.8	10.6	**7.87**	35.4	58(nrd0)	15.1(nrd)

Table 3
100 × MISE for each model and each method with $n = 500$, $M = 150$.

Model	Hist	HistG	HistC	ASH	RASH	RASHG	RASHC	ST	Kde	Kdet
$\mathcal{M}_1$	0.455	1.05	0.188	0.077	0.078	0.574	**0.058**	0.207	0.091(nrd0)	0.108(nrd)
$\mathcal{M}_2$	0.772	2.52	**0.573**	4.07	0.603	1.33	0.943	0.945	2.91(ucv)	1.04(sj)
$\mathcal{M}_3$	0.024	0.063	0.013	0.012	0.004	0.031	0.004	0.012	**0.003**(ucv)	0.006(nrd)
$\mathcal{M}_4$	0.179	0.858	0.208	0.102	0.103	0.453	**0.056**	0.173	0.101(nrd0)	0.082 (nrd)
$\mathcal{M}_5$	1.12	3.35	7.21	0.248	0.401	1.75	3.44	0.955	2.03(ucv)	**0.196**(nrd0)
$\mathcal{M}_6$	0.089	0.319	0.323	**0.03**	0.034	0.165	0.236	0.118	0.087(ucv)	**0.03**(nrd0)
$\mathcal{M}_7$	0.247	0.958	1.02	**0.055**	0.061	0.396	0.667	0.227	0.29(ucv)	0.061 (nrd0)
$\mathcal{M}_8$	0.859	2.22	2.21	**0.427**	0.433	1.35	1.83	0.906	1.77(ucv)	0.53(sj)
$\mathcal{M}_9$	0.619	1.1	2.47	0.414	0.412	0.8	1.75	0.63	0.885(ucv)	**0.397**(sj)
$\mathcal{M}_{10}$	4.79	7.73	6.24	**4.39**	4.82	5.31	5.78	4.73	6.15(ucv)	4.71 (ucv)
$\mathcal{M}_{11}$	1.1	2.83	0.959	0.847	0.817	1.44	**0.363**	1.08	1.42(nrd0)	0.622 (nrd)
$\mathcal{M}_{12}$	6.95	30.1	8.07	6.69	3.75	14.5	**3.32**	9.61	40.9(ucv)	4.09(nrd)

Table 4
100 × MISE for each model and each method with $n = 1000$, $M = 150$.

Model	Hist	HistG	HistC	ASH	RASH	RASHG	RASHC	ST	Kde	Kdet
$\mathcal{M}_1$	0.218	0.622	0.145	**0.029**	0.03	0.346	0.037	0.135	0.071(nrd0)	0.064(nrd)
$\mathcal{M}_2$	**0.378**	0.897	0.461	4.42	0.502	0.555	0.952	0.512	2.53(ucv)	0.778(sj)
$\mathcal{M}_3$	0.01	0.034	0.011	0.007	0.006	0.022	0.461	0.008	**0.002**(nrd)	0.003 (nrd)
$\mathcal{M}_4$	0.095	0.449	0.176	**0.034**	0.035	0.237	0.041	0.115	0.074(nrd0)	0.047(nrd)
$\mathcal{M}_5$	0.625	2.05	7.66	0.129	0.237	1.16	3.66	0.567	1.67(ucv)	**0.12**(nrd0)
$\mathcal{M}_6$	0.066	0.176	0.348	0.018	0.019	0.103	0.262	0.071	0.066(ucv)	**0.017**(nrd0)
$\mathcal{M}_7$	0.117	0.531	1.12	**0.034**	0.036	0.29	0.768	0.143	0.235(ucv)	0.037(nrd0)
$\mathcal{M}_8$	0.569	1.43	2.19	**0.255**	0.257	0.917	1.81	0.605	1.67(ucv)	0.329(sj)
$\mathcal{M}_9$	0.405	0.865	2.57	0.286	**0.285**	0.693	1.79	0.436	0.762(ucv)	**0.285**(sj)
$\mathcal{M}_{10}$	4.75	7.19	6.29	4.5	4.65	5.46	5.86	4.65	5.59(ucv)	**3.2**(ucv)
$\mathcal{M}_{11}$	0.638	2.71	0.723	0.439	0.322	1.44	**0.231**	0.691	1.16(nrd0)	0.384(nrd)
$\mathcal{M}_{12}$	4.5	19.2	6.15	5.53	2.83	10.2	2.8	5.7	33(ucv)	**2.55**(nrd)

- BagHist2D: the two-dimensional version of the algorithm described in Bourel and Ghattas (2013). At each step we build a histogram with *Hist*2*D* over a bootstrap sample of the data, and the final estimator is an average of M histograms.
- ASH2D: the average shifted histogram for two-dimensional data (Scott, 1992).
- RASH2D: it is a direct extension of our algorithm to the bivariate case. At each step the breakpoints given by *Hist*2*D* are shifted randomly using a Gaussian perturbation in each direction.
- GrHist2D: a greedy two dimensional histogram following the idea of Klemelä (2009). At each step, a binary partition like in Classification and Regression Trees (*CART*) is performed and used to estimate the histogram.
- BagHistGr2D: as in *BagHist*2*D*, this algorithm constructs several histograms with greedy partitions as in *GrHist*2*d* over bootstrap samples of the data and averages them. This method has been used in Klemelä (2009).
- RASHGr2D: an extension of our algorithm using greedy partitions as in *GrHist*2*D* and applying at each step a random Gaussian shift of the breakpoints for each direction.

We use the following models, most of them are combinations of some models used in the one-dimensional framework:

- ($\mathcal{M}_2 1$): a standard bivariate normal $N(\mu, \Sigma)$ with $\mu = (0, 0)$ and $\Sigma = Id$
- ($\mathcal{M}_2 2$): the Dumbbell density which is a mixture of bivariate normals.
- ($\mathcal{M}_2 3$): a random vector (X, Y) where $X, Y \sim 0.5N(-1, 0.3) + 0.5N(1, 0.3)$ are independent random variables.

Table 5
100 × MISE, $n = 100$, number of cells = 16, number of leaves = 15.

	Hist2D	Kde2D	BagHist2D	ASH2D	RASH2D	GrHist2D	BagHistGr2D	RASHGr2D
$\mathcal{M}_2 1$	0.316	0.046	0.130	**0.045**	0.153	0.359	0.199	**0.053**
$\mathcal{M}_2 2$	1.137	**0.029**	0.770	0.089	0.710	0.140	**0.091**	0.101
$\mathcal{M}_2 3$	2.286	2.696	1.804	0.601	**0.558**	15.99	12.74	**0.564**
$\mathcal{M}_2 4$	2.376	0.072	2.130	**0.017**	1.165	0.369	0.349	**0.025**
$\mathcal{M}_2 5$	0.863	0.882	0.782	0.927	**0.742**	1.542	0.910	**0.834**
$\mathcal{M}_2 6$	0.322	0.391	0.284	0.433	**0.246**	0.400	**0.274**	0.374

Table 6
100 × MISE, $n = 100$, number of cells = 36, number of leaves = 30.

	Hist2D	Kde2D	BagHist2D	ASH2D	RASH2D	GrHist2D	BagHistGr2D	RASHGr2D
$\mathcal{M}_2 1$	0.142	0.047	0.078	0.048	**0.045**	0.922	0.500	**0.095**
$\mathcal{M}_2 2$	0.174	**0.026**	0.092	0.053	0.080	0.346	0.159	**0.071**
$\mathcal{M}_2 3$	4.179	2.723	3.348	**2.225**	3.594	44.54	21.43	**3.570**
$\mathcal{M}_2 4$	0.576	0.063	0.541	**0.062**	0.328	0.886	0.155	**0.690**
$\mathcal{M}_2 5$	0.896	0.867	0.844	0.819	**0.805**	3.562	1.973	**0.74**
$\mathcal{M}_2 6$	0.370	0.392	0.338	0.318	**0.311**	0.890	0.634	**0.292**

Table 7
100 × MISE, $n = 200$, number of cells = 36, number of leaves = 30.

	Hist2D	Kde2D	BagHist2D	ASH2D	RASH2D	GrHist2D	BagHistGr2D	RASHGr2D
$\mathcal{M}_2 1$	0.076	0.028	0.043	0.027	**0.025**	0.486	0.192	**0.055**
$\mathcal{M}_2 2$	0.18	**0.019**	0.111	0.055	0.111	0.165	0.080	**0.070**
$\mathcal{M}_2 3$	2.074	3.815	1.525	0.765	**0.759**	21.250	13.720	**3.593**
$\mathcal{M}_2 4$	0.447	0.105	0.452	**0.056**	0.388	0.650	0.473	**0.161**
$\mathcal{M}_2 5$	1.017	0.950	0.927	0.921	**0.918**	1.950	0.803	**0.755**
$\mathcal{M}_2 6$	0.335	0.319	0.323	0.314	**0.289**	0.616	0.379	**0.290**

- $(\mathcal{M}_2 4)$: a random vector (X, Y) where $X, Y \sim 0.25N(-3, 0.5) + 0.5N(0, 1) + 0.25N(3, 0.5)$ are independent random variables.
- $(\mathcal{M}_2 5)$: a random vector (X, Y) where X and Y are independent and follow the Claw density.
- $(\mathcal{M}_2 6)$: a random vector (X, Y) where X and Y are independent and follow the Smooth Comb density.

A plot of these densities is given in Fig. 5. Tables 5–7 give the results we obtained for each model for $n = 100$ and $n = 200$ respectively. The number of cells n_c we choose for *Hist2D, ASH2D* and *RASH2D* are 16 and 36, and the number of leaves n_l of *GrHist2D* and *RASHGr2D* are 15 and 30. These values are arbitrary and unlike the one dimensional case they are not optimized. The values of 100 × MISE are averaged over 50 Monte Carlo simulations. Table 5 is for $n = 100$, $n_c = 16$ and $n_l = 15$, Table 6 is for $n = 100$, $n_c = 16$ and $n_l = 30$, and Table 7 is for $n = 200$, $n_c = 36$ and $n_l = 30$. For each of them, on the left side we grouped the methods using classical histograms and on the right side methods which are based on greedy partitions. The best performances are indicated in boldface for each of these groups.

We observe first that our methods clearly outperform the single histograms, classical ones and those built over a greedy partition. The reduction of the MISE is in some cases very significant. For regular partitions, only on model $\mathcal{M}_2 2$ *ASH2D* and *RASH2D* cannot do better than the Kde2D for the values of n and n_c we selected. For models $\mathcal{M}_2 5$ and $\mathcal{M}_2 6$, *RASHGr2D* gives the best results. The results obtained by *RASH2D* and *RASHGr2D* are very encouraging, in particular on difficult models to estimate ($\mathcal{M}_2 3$ to $\mathcal{M}_2 6$), and will be the subject of a further study.

6. Conclusion

In this work we present a new algorithm, *RASH*, for density estimation, averaging different histograms built using a random shift of the initial breakpoints. We have proved, under some reasonable assumptions, the consistency of this estimator and the asymptotic normality of the RH-estimator for the one-dimensional case. Also we have shown using extensive simulations in one and two dimensions that this algorithm has in general better accuracy than classical methods as the histogram or Kde and similar or better performance as the *ASH* method. *RASH* is very simple to implement, depends only on the number of breakpoints we use, and its computation complexity is proportional to that of a histogram. The extension of our algorithm and its properties to the multivariate case and for adaptively used breakpoints for histograms are still under study.

Acknowledgments

We thank the reviewers for their constructive comments that have greatly improved the presentation of the actual manuscript.

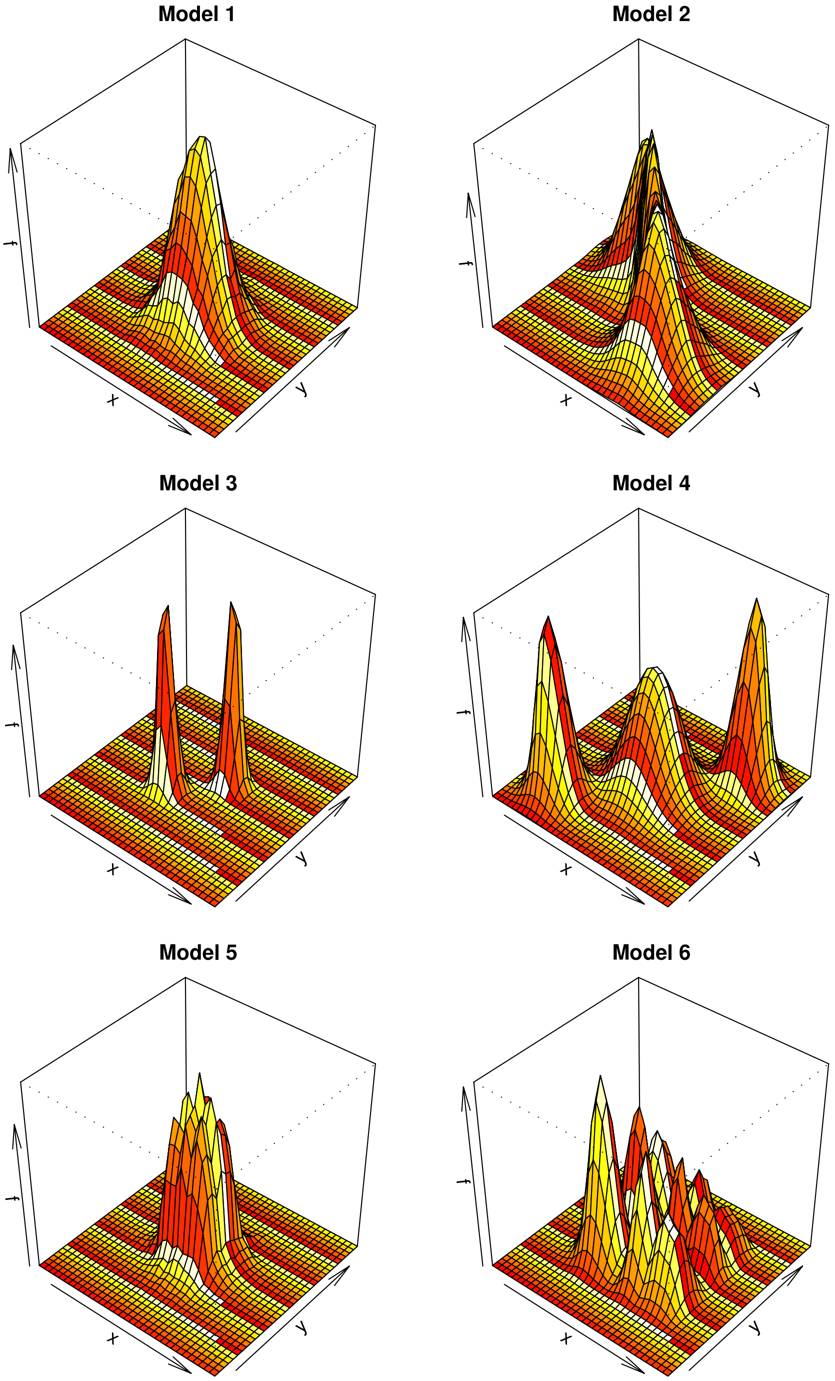

Fig. 5. Models used in the two-dimensional simulations.

Appendix

To simplify the notations, we replace g_n by g and e_n by e, and, as mentioned in Section 4, the superscript (m) will be omitted in the proofs concerning the RH-estimators.

Proof of Theorem 1. Take $x \in I_{j,n}(e)$. We now analyze the bias and variance term of $\widehat{f_n}$.

1. For the bias we will show that

$$\mathbb{E}\big(\widehat{f_n}(x)\big) \to f(x).$$

The expectation of $\widehat{f_n}(x)$ is

$$\begin{aligned}\mathbb{E}\big(\widehat{f_n}(x)\big) &= \frac{1}{nh_n}\sum_{i=1}^{n}\sum_{j=1}^{L}\mathbb{P}\big(X_i \in I_{j,n}(e)\big)\mathbb{1}_{I_{j,n}(e)}(x)\\ &= \frac{1}{nh_n}n\mathbb{P}(X_i \in I_{j,n}(e)) = \int_{\mathbb{R}}\frac{1}{h_n}\underbrace{\int_{a_j+u}^{a_{j+1}+u} f(t)\,dt}_{q(u)}\ g(u)du.\end{aligned}$$

Since f is bounded in a neighborhood of x, we have that $q(u) = \int_{a_j+u}^{a_{j+1}+u} f(t)\,dt \le Ch_n$, and as $h_n \to 0$ the dominated convergence theorem entails that

$$\mathbb{E}\big(\widehat{f_n}(x)\big) \to \int_{\mathbb{R}} f(x)\ g(u)du = f(x).$$

2. Now we show that $Var\big(\widehat{f_n}(x)\big) \to 0$. We have:

$$Var\big(\widehat{f_n}(x)\big) = \frac{1}{n^2h_n^2}Var\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\right).$$

We use the following formula for conditional variance:

$$Var(Y) = \mathbb{E}\big(Var(Y|Z)\big) + Var\big(\mathbb{E}(Y|Z)\big)$$

with $Y = \sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)$ and $Z = e$, i.e:

$$Var(Y) = \underbrace{\mathbb{E}\left(Var\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right)}_{[I]} + \underbrace{Var\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right)}_{[II]}.$$

[I]. Because of the conditional independence we have:

$$\begin{aligned}Var\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right) &= \sum_{i=1}^{n}Var\left(\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\\ &= \sum_{i=1}^{n}\mathbb{P}(X_i \in I_{j,n}(e)\big|e)\big(1-\mathbb{P}(X_i \in I_{j,n}(e)\big|e)\big) \le \sum_{i=1}^{n}\mathbb{P}(X_i \in I_{j,n}(e)\big|e).\end{aligned}$$

Taking the expectation we obtain:

$$\begin{aligned}\mathbb{E}\left(Var\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right) &\le \sum_{i=1}^{n}\mathbb{E}\left(\mathbb{P}(X_i \in I_{j,n}(e)\big|e)\right) = \sum_{i=1}^{n}\mathbb{E}\left(\mathbb{E}\left(\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right)\\ &= \sum_{i=1}^{n}\mathbb{E}\left(\mathbb{1}_{I_{j,n}}(X_i)\right) = n\mathbb{P}(X_1 \in I_{j,n}).\end{aligned}$$

Then as $nh_n \to +\infty$, we have:

$$\frac{1}{n^2h_n^2}\mathbb{E}\left(Var\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right) \le \frac{1}{nh_n^2}\mathbb{P}(X_1 \in I_{j,n}) \to 0.$$

[II]. For the second term we return to the definition of the variance

$$Var\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right) = \underbrace{\mathbb{E}\left(\left[\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right]^2\right)}_{\text{(B)}} - \underbrace{\left[\mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right)\right]^2}_{\text{(A)}}. \tag{6}$$

(A): Since $\mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)|e\right)\right) = \sum_{i=1}^{n}\mathbb{E}\left(\mathbb{1}_{I_{j,n}}(X_i)\right) = n\mathbb{P}(X_1 \in I_{j,n})$ we have

$$\left[\mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)|e\right)\right)\right]^2 = n^2\left(\mathbb{P}(X_1 \in I_{j,n})\right)^2.$$

(B): We have from this previous calculation that

$$\mathbb{E}\left(\left[\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)|e\right)\right]^2\right) = n^2\mathbb{E}\left(\left(\mathbb{P}(X_1 \in I_{j,n}(e)|e)\right)^2\right).$$

We replace this expression in (6) and get:

$$\begin{aligned} Var\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)|e\right)\right) &= n^2\mathbb{E}\left(\left(\mathbb{P}(X_1\in I_{j,n}(e)|e)\right)^2\right) - n^2\left(\mathbb{P}(X_1\in I_{j,n})\right)^2 \\ &= n^2\left[\int_{\mathbb{R}}\left(\int_{I_{j,n}(u)} f(t)\,dt\right)^2 g(u)\,du - \left(\int_{I_{j,n}} f(t)\,dt\right)^2\right] \\ &= n^2\int_{\mathbb{R}}\left[\left(\int_{I_{j,n}(u)} f(t)\,dt\right)^2 - \left(\int_{I_{j,n}} f(t)\,dt\right)^2\right] g(u)\,du. \end{aligned}$$

Then

$$\begin{aligned} \frac{1}{n^2h_n^2}Var\left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\big|e\right)\right) &\le \frac{n^2}{2n^2h_n^2}\int_{\mathbb{R}}\left(-\int_{a_j}^{a_j+u} f(t)\,dt + \int_{a_{j+1}}^{a_{j+1}+u} f(t)\,dt\right) g(u)\,du \\ &\le \frac{C}{h_n^2}\int_{\mathbb{R}}|u|g(u)\,du = \frac{C\mathbb{E}(|e|)}{h_n^2} \to 0. \end{aligned}$$

Hence $\widehat{f_n}$ converges to f in L^2. □

Proof of Proposition 1.

$$\sqrt{nh_n}\left|\widehat{f_n}(x) - \widehat{f}_{n,0}(x)\right| = \sqrt{nh_n}\left|\frac{1}{nh_n}\sum_{j=1}^{L}\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}}(X_i)\mathbb{1}_{I_{j,n}}(x) - \frac{1}{nh_n}\sum_{j=1}^{L}\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}(e)}(X_i)\mathbb{1}_{I_{j,n}(e)}(x)\right|.$$

For a fixed j, suppose first that $x \in I_{j,n}\cap I_{j,n}(e)$, then

$$\sqrt{nh_n}\left|\widehat{f_n}(x) - \widehat{f}_{n,0}(x)\right| \le \sqrt{nh_n}\left(\underbrace{\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{1}_{I_{j,n}}(X_i)\mathbb{1}_{I^c_{j,n}(e)}(X_i)}_{(T_1)} + \underbrace{\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{1}_{I^c_{j,n}}(X_i)\mathbb{1}_{I_{j,n}(e)}(X_i)}_{(T_2)}\right).$$

Therefore,

$$\sqrt{nh_n}\mathbb{E}(|\widehat{f_n}(x) - \widehat{f}_{n,0}(x)|) \le \sqrt{nh_n}\left(\mathbb{E}\big((T_1)\big) + \mathbb{E}\big((T_2)\big)\right).$$

We analyze $\mathbb{E}\big((T_1)\big)$, the same arguments holds for $\mathbb{E}\big((T_2)\big)$.

$$\mathbb{E}\big((T_1)\big) = \frac{1}{h_n}\mathbb{P}\left(X_1\in I_{j,n}\cap I^c_{j,n}(e)\right) = \frac{1}{h_n}\int_{\mathbb{R}}\mathbb{P}\left(X_1\in I_{j,n}\cap I^c_{j,n}(u)\right)g(u)\,du.$$

If $u > 0$, then

$$\mathbb{E}\big((T_1)\big) = \frac{1}{h_n}\int_{\mathbb{R}} u\frac{\mathbb{P}(a_j\le X_1\le a_j+u)}{u}g(u)\,du.$$

Since $\frac{\mathbb{P}(a_j\le X_1\le a_j+u)}{u} \to f(x)$ if $u \to 0$, f is bounded, and $\frac{\mathbb{P}(a_j\le X_1\le a_j+u)}{u} \le \frac{1}{u_0}$ this quotient is bounded and there exist a constant C such that:

$$\mathbb{E}\big((T_1)\big) \le \frac{C}{h_n}\int_{\mathbb{R}}|u|g(u)\,du = \frac{C}{h_n}\mathbb{E}(|e|),$$

(for $u < 0$ the same argument holds replacing $\mathbb{P}(a_j\le X_1\le a_j+u)$ by $\mathbb{P}(a_j+u\le X_1\le a_j)$).

Then if $e \sim g$ and as $\sqrt{nh_n}\frac{\mathbb{E}(|e|)}{h_n} = \sqrt{\frac{n}{h_n}}\mathbb{E}(|e|) \to 0$, which holds by hypothesis, we can conclude that

$$\sqrt{nh_n}\left|\widehat{f_n}(x) - \widehat{f}_{n,0}(x)\right| \underset{n}{\overset{P}{\to}} 0.$$

Now, if $x \notin I_{j,n} \cap I_{j,n}(e)$ then $x \in [a_j, a_j + e]$ (suppose that $e > 0$) and therefore $x - a_j < e$. Hence, from the Markov inequality, we can see that this case tends to zero in probability since:

$$\mathbb{P}\left(|e| > x - a_j\right) \leq \frac{\mathbb{E}(|e|)}{|x - a_j|} \leq \frac{\mathbb{E}(|e|)}{h_n} \leq \frac{\mathbb{E}(|e|)}{h_n}\sqrt{nh_n} \to 0,$$

where this last inequality holds from $\sqrt{nh_n} \to +\infty$. □

Proof of Lemma 1. Let $\delta > 0$ and suppose that $x \in I_{j,n}(e)$. Then:

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x))\right| > \delta\right) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{I_{j,n}(e)}(X_i) - \mathbb{P}\big(X_1 \in I_{j,n}(e)\big)\right| > \delta h_n\right) \\ &= \mathbb{E}\left(\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{I_{j,n}(e)}(X_i) - \mathbb{P}\big(X_1 \in I_{j,n}(e)\big)\right| > \delta h_n \Big| e\right)\right),\end{aligned}$$

and from Hoeffding's inequality we get

$$\mathbb{E}\left(\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{I_{j,n}(e)}(X_i) - \mathbb{P}\big(X_1 \in I_{j,n}(e)\big)\right| > \delta h_n \Big| e\right)\right) \leq \mathbb{E}\left(\exp\left(-2n\delta^2 h_n^2\right) \big| e\right) = e^{-2n\delta^2 h_n^2}.$$

Then the assumption entails that the series $\sum_n e^{-2\delta^2 h_n^2 n}$ is convergent which concludes the proof. □

Proof of Theorem 3. Let $\delta > 0$ and suppose that $x \in I_{j,n}(e)$:

$$\mathbb{P}\left(\left|\widehat{f}_n(x) - f(x)\right| > \delta\right) = \mathbb{P}\left(\left|\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x)) + \mathbb{E}(\widehat{f}_n(x)) - f(x)\right| > \delta\right).$$

Recall from the proof of Theorem 1 that as $\left|\mathbb{E}(\widehat{f}_n(x)) - f(x)\right| \to 0$ there exists a sequence $\eta_n \to 0$ such that

$$\left|\mathbb{E}(\widehat{f}_n(x)) - f(x)\right| \leq \eta_n.$$

Then:

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{f}_n(x) - f(x)\right| > \delta\right) &\leq \mathbb{P}\left(\left|\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x))\right| > \delta - \eta_n\right) \\ &\leq \mathbb{P}\left(\left|\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x))\right| > \frac{\delta}{2}\right),\end{aligned}$$

for all $n \geq n_0$. Using Hoeffding's inequality as in Lemma 1, we have that

$$\mathbb{P}\left(\left|\widehat{f}_n(x) - f(x)\right| > \frac{\delta}{2}\right) \leq e^{-\frac{\delta^2}{2}nh_n^2},$$

and since $\frac{nh_n^2}{\log n} \underset{n}{\to} +\infty$, we conclude that

$$\widehat{f}_n(x) - f(x) \xrightarrow[n]{a.s} 0. \quad \square$$

Proof of Theorem 4. First observe that if we replace δ by $\frac{\delta}{\beta_n}$ in the proof of Lemma 1, we obtain that

$$\mathbb{P}\left(\beta_n|\widehat{f}_n(x) - \mathbb{E}(\widehat{f}_n(x))| > \delta\right) \leq e^{-\frac{2\delta^2 nh_n^2}{\beta_n^2}}. \tag{7}$$

Next we deal with the bias term. As in the proof of Theorem 1 we have that

$$\begin{aligned}\mathbb{E}\big(\widehat{f}_n(x)\big) - f(x) &= \int_{\mathbb{R}} \frac{1}{h_n}\int_{a_j+u}^{a_{j+1}+u} f(t)\,dt\ g(u)du - f(x) \\ &= \int_{\mathbb{R}} \frac{1}{h_n}\int_{a_j+u}^{a_{j+1}+u} f(t)\,dt\ g(u)du - \frac{1}{h_n}\int_{\mathbb{R}} h_n f(x) g(u)\,du \\ &= \int_{\mathbb{R}} \frac{1}{h_n}\left(\int_{a_j+u}^{a_{j+1}+u} f(t)\,dt - h_n f(x)\right) g(u)du = \int_{\mathbb{R}} \frac{1}{h_n}\left(\int_{a_j+u}^{a_{j+1}+u} f(t)\,dt - \int_{a_j+u}^{a_{j+1}+u} f(x)\,dt\right) g(u)du \\ &= \int_{\mathbb{R}} \frac{1}{h_n}\left(\int_{a_j+u}^{a_{j+1}+u} \big[f(t) - f(x)\big]\,dt\right) g(u)du.\end{aligned}$$

Then, since f is Lipschitz in a neighborhood of x:

$$\eta_n := \left|\mathbb{E}(\widehat{f_n}(x)) - f(x)\right| \le \int_{\mathbb{R}} \frac{1}{h_n}\left(\int_{a_j+u}^{a_{j+1}+u} C\left|t-x\right| dt\right) g(u)du \le \int_{\mathbb{R}} \frac{1}{h_n} Ch_n^2\, g(u)du = Ch_n, \tag{8}$$

for some positive constant C.

Putting things together we have that:

$$\mathbb{P}\left(\beta_n \left|\widehat{f_n}(x) - f(x)\right| > \delta\right) = \mathbb{P}\left(\beta_n \left|\widehat{f_n}(x) - \mathbb{E}(\widehat{f_n}(x)) + \mathbb{E}(\widehat{f_n}(x)) - f(x)\right| > \delta\right) \tag{9}$$

$$\le \mathbb{P}\left(\left|\widehat{f_n}(x) - \mathbb{E}(\widehat{f_n}(x))\right| > \frac{\delta}{\beta_n} - \eta_n\right) \le \mathbb{P}\left(\left|\widehat{f_n}(x) - \mathbb{E}(\widehat{f_n}(x))\right| > \frac{\delta}{2\beta_n}\right), \tag{10}$$

for n large enough, since $\beta_n\eta_n \to 0$ if $\beta_n h_n \to 0$, as $n \to +\infty$ by (8). Finally, (7) implies that the right hand side of (10) is bounded by $e^{\frac{-n\delta^2 h_n^2}{2\beta_n^2}}$, the general term of a convergent series, which concludes the proof of (i).

In particular if we take $h_n = n^{-\alpha}$, $\beta_n = n^{\beta}$ in order to have $\frac{nh_n^2}{\beta_n^2 \log n} \xrightarrow[n]{} +\infty$, and $\beta_n\eta_n \to 0$:

- $\beta_n\eta_n \to 0 \Rightarrow n^{\beta} n^{-\alpha} = n^{\beta-\alpha} \to 0$, which holds if $\alpha > \beta$.
- $\frac{nh_n^2}{\beta_n^2 \log n} = \frac{nn^{-2\alpha}n^{-2\beta}}{\log n} = \frac{n^{1-2\alpha-2\beta}}{\log n} \xrightarrow[n]{} \infty \Rightarrow 1 - 2\alpha - 2\beta > 0 \Rightarrow \beta < \frac{1}{2} - \alpha$.

Then combining these two conditions, we have that all conditions hold if $\beta < \frac{1}{4}$, which proves (ii). □

References

Bourel, M., Ghattas, B., 2013. Aggregating density estimators: an empirical study. Open J. Stat. 3 (5).
Bowman, A., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71, 353–360.
Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24 (2), 123–140.
Breiman, L., 1996b. Stacked regression. Mach. Learn. 24 (1), 49–64.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.
Di Marzio, M., Taylor, C.C., 2004. Boosting kernel density estimates: a bias reduction technique? Biometrika 91 (1), 226–233.
Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55 (1), 119–139.
Iacono, M., Irpino, Antonio, 2011. Improving the MHIST-p Algorithm for Multivariate Histograms of Continuous Data. In: Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg, pp. 155–163 (Chapter 15).
Klemelä, J., 2009. Smoothing of Multivariate Data: Density Estimation and Visualization. In: Wiley Series in Probability and Statistics, Wiley.
Marron, J.S., Wand, M.P., 1992. Exact mean integrated square error. Ann. Statist. 20 (2), 712–736.
Ridgeway, G., 2002. Looking for lumps: boosting and bagging for density estimation. Comput. Statist. Data Anal. 38 (4), 379–392.
Rigollet, P., Tsybakov, A.B., 2007. Linear and convex aggregation of density estimators. Math. Methods Statist. 16 (3), 260–280.
Rosset, S., Segal, E., 2002. Boosting density estimation. Adv. Neural Inf. Process. Syst. 15, 641–648.
Samarov, A., Tsybakov, A.B., 2007. Aggregation of density estimators and dimension reduction. Adv. Stat. Model. Inference 233–251.
Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66, 605–610.
Scott, D.W., 1985. Averaged shifted histogram: effective nonparametric density estimators inseveral dimensions. Ann. Statist. 13 (3), 1024–1040.
Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. In: Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley.
Scott, D.W., 2009. Averaged shifted histogram. Comput. Statist. 2 (2), 160–164.
Scott, D.W., Härdle, W.K., 1992. Smoothing by weighted averaging of rounded points. Comput. Statist. 7, 97–128.
Sheather, S.J., 2004. Density estimation. Statist. Sci. 19 (4), 588–597.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Smyth, P., Wolpert, D., 1999. Linearly combining density estimators via stacking. Mach. Learn. 36 (1–2), 59–83.
Wand, P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Comput. Statist. 9, 97–116.
Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5, 241–259.