Bagging of density estimators

Mathias Bourel$^{1,2}$ · Jairo Cugliari$^3$

Received: 21 August 2018 / Accepted: 29 March 2019 / Published online: 15 April 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
In this work we give new density estimators by averaging classical density estimators such as the histogram, the frequency polygon and the kernel density estimators obtained over different bootstrap samples of the original data. Using existent results, we prove the $L^2$-consistency of these new estimators and compare them to several similar approaches by simulations. Based on them, we give also a way to construct non-parametric pointwise variability band for the target density.

Keywords Aggregation · Bagging · Density estimation · Histogram · Kernel density estimator · Polygon frequency

1 Introduction

Ubiquitous in data analysis, density estimation techniques are certainly the most used unsupervised learning technique on low dimension. Whether for studying asymmetry, normality, residual diagnostic or bump hunting among others, one usually relies on a visual inspection of a plot of the density to take a primary decision, mostly in one or two dimensions.

The general aim is to gather basic information about the unobserved data generation mechanism out of a sample of $n$ observations say $x_1, \ldots, x_n$. One usually supposes that the observations are realizations of a random variable $X$ that admits a probability density function $f$ (i.e. $f$ is non-negative and integrates 1). Then, the learning task is to estimate $f$ as accurately as possible. First, by obtaining a pointwise estimate $\hat{f}(x)$ of...
\( f(x) \) for all \( x \in \mathbb{R}^d \), and second, by assessing the uncertainty of the pointwise estimate through the construction of a confidence interval for \( f \). Of course both problems imply different difficulties and involves specific techniques. In what follows, we focus on nonparametric approaches for the first objective and give a possible way to construct a pointwise confidence interval for the true density. In particular, we center our attention on three classes of base or individual estimators for the density: histograms, frequency polygons and the kernel density estimator. We postpone the formal definition up to the next section but we provide a discussion guided by intuitive descriptions here.

Histograms are undoubtedly the most popular construction for density estimation. They rely on the best constant-wise approximation of \( f \) given by the data using a binning argument. The power of this simple construction combined with the ease of its interpretation makes them accessible to non-technical users. Besides, theoretical properties can be derived showing that histograms are consistent estimators. Some of the lacks of the histograms are inherent to one hand the constant level for at each partition, and their discontinuities and on the other hand, they have null derivative everywhere. Frequency polygons are constructed on top of histograms (and so take profit of the binning advantages) providing a linear piecewise estimator. Although the additional regularity was once pointed out as a flaw (see Fisher 1932) it is now well known that it increases the quality of the estimator. Theoretically, this is shown by a faster rate of convergence (cf. Sect. 2). Further regularity can be gained using kernel density estimators. Essentially one first picks a kernel function with the desired degree of regularity for the final estimate. Then, the empirical measure is convolved to produce the kernel density estimation of \( f \). It has been proved in Scott (1985b) that the rates of convergence of the frequency polygon are similar to the kernel density estimator. All the three approaches are exhaustively studied both from on practice and theory as individual estimators and a general reference for the subject is Scott (2015). Also extensions exists, like in Botev et al. (2010) where the authors introduce an adaptive kernel density estimation method based on the smoothing properties of linear diffusion processes. However, a reasonable question is to ask whether further improvement can be achieved from these constructions by means of aggregation schemes.

Ensemble learning or aggregation methods are increasingly used in the supervised framework: these methods combine intermediate predictors to obtain an aggregated model with the aim to obtain a better estimator. Bagging (Breiman 1996) (Bootstrap and AGGRegatING), Boosting (Freund and Schapire 1997), Stacking (Wolpert 1992), and Random Forests (Breiman 2001) have been broadly studied in the case of classification (mainly binary classification) or regression from the theoretical viewpoint and have strong performances when tested over tens of various datasets selected from the machine learning benchmark. Several extensions are still under study: multivariate regression, multiclass classification, and adaptation to functional data or time series. Very few developments exist for ensemble learning applied on unsupervised techniques such as clustering analysis and density estimation. Only on few works, several authors look at the adaptation of the aggregation procedure to estimate a density under somehow restrictive conditions. One of the first is the mean of the simplest density estimator, the histograms, each one constructed over several different deterministic
grids in Average Shifted Histograms, ASH (Scott 1985a). With a combination of several kernel density estimators with different bandwidths, often in a normal context, and varying the form of the aggregation we can cite Ridgeway (2002), Glodek et al. (2013), Song et al. (2004), Rosset and Segal (2002), Smyth and Wolpert (1999) and Rigollet and Tsybakov (2007).

Another kind of aggregation can be obtained by introducing randomness in the individual estimators. In Bourel and Ghattas (2013) the authors include randomness in the construction of the intermediate histograms using then different aggregation schemes: AggregHist, using simple aggregation, BagHist using Bagging or StackHist using Stacking. Mathematically well sound, these approaches were explored thorough empirical simulation without theoretical framework. The Random Average Shifted Histogram, RASH (Bourel et al. 2014) is constructed as the mean average of different histograms, each one constructed over a random translated grid of the initial breakpoints of the initial histogram. RASH is shown to be consistent and perform well in comparison to the precedent aggregation schemes.

In this work, we explore the contribution of Bagging to the density estimation task (like for BagHist) where the intermediate estimators are either histograms, frequency polygons or kernel density estimators. This article is organized as follows. Section 2 introduces notation and reviews the framework context we need about density estimators. In Sect. 3 we present our three methods: BagHist, BagFP and BagKDE and establish a theoretical result about their consistency. The results of comprehensive simulations are the object of Sect. 4. For this, we use several target densities available on literature to evaluate the performance of our estimators and comparing them to classical density estimators. We also explore the construction of pointwise confidence intervals as an approximation of a confidence band using the bootstrap procedure. The work concludes with a discussion in Sect. 5.

2 Some density estimators

We look at the principal results using different density estimators. While for the histogram and the kernel density estimator these results are quite popular we give some detail in order to fix notation. For a detailed exposition see Scott (2015).

In all cases, the starting point is an independent and identically distributed sample $x_1, \ldots, x_n$ of a real random variable with density $f$.

- **Histogram.** Let $B_j = [jh, (j + 1)h]$ be a set of intervals defined over the support of $f$, and $h \to 0$ as $n \to +\infty$.

  The ordinary histogram is defined as:

  \[
  \hat{f}_n^H(x) = \frac{1}{nh} \sum_{i=1}^n \sum_{j \in \mathbb{Z}} 1_{B_j}(x_i) 1_{B_j}(x) = \frac{v_j}{nh}
  \]  

  where $v_j \sim \text{Bin}(n, p_j)$, with $p_j = \int_{B_j} f(t)dt$, is the number of observations of the sample that fall in bin $B_j$. If $x \in B_j$, we have that:
Kernel density estimators

A Kernel Density Estimator (KDE), is a function defined by

$$
\hat{f}_n^{KDE}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right)
$$

for all $x \in \mathbb{R}$ where $K$ is a kernel function, i.e. a non-negative, symmetric and unimodal function such that $\int K(u) \, du = 1$. Parameter $h$ is called the bandwidth of the estimator $\hat{f}_n^{KDE}$, who inherits all the mathematical properties of $K$. The function $K$ indicates the weight that observation $x_i$ has in the estimation of $x$: observations close to $x$ are weighted more important. It can be shown that

$$
\mathbb{E} \left( \hat{f}_n^{KDE}(x) \right) = f(x) + \frac{f''(x)}{2} \mu_2(K) h^2 + o(h^2)
$$

$$
\text{Var} \left( \hat{f}_n^{KDE}(x) \right) = \frac{||K||^2 f(x)}{nh} + o \left( \frac{1}{nh} \right)
$$

Frequency polygons

Frequency polygons are constructed on top of histograms connecting with straight lines the midpoint of two consecutive bin values. The expression of the frequency polygon for an $x \in B_j = [(j-1/2)h, (j+1/2)h]$ is

$$
\hat{f}_{np}^{FP}(x) = \frac{1}{2} + j - \frac{x}{h} \frac{v_j}{nh} + \left( \frac{1}{2} - j + \frac{x}{h} \right) \frac{v_{j+1}}{nh}
$$

The frequency polygon was deeply studied in Scott (1985b). With respect to the histogram, it presents the advantages of being continuous and smooth. Under weak conditions, an optimal choice for $h$ is of order $n^{-1/5}$ and it achieves a rate of convergence of order $n^{-4/5}$ with respect to the MISE (Scott 1985b).

- Frequency polygon. Frequency polygons are constructed on top of histograms connecting with straight lines the midpoint of two consecutive bin values. The expression of the frequency polygon for an $x \in B_j = [(j-1/2)h, (j+1/2)h]$ is

$$
\hat{f}_{np}^{FP}(x) = \frac{1}{2} + j - \frac{x}{h} \frac{v_j}{nh} + \left( \frac{1}{2} - j + \frac{x}{h} \right) \frac{v_{j+1}}{nh}
$$

(2)

- Kernel density estimators. A Kernel Density Estimator (KDE), is a function defined by

$$
\hat{f}_n^{KDE}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x-x_i}{h} \right)
$$

(3)

for all $x \in \mathbb{R}$ where $K$ is a kernel function, i.e. a non-negative, symmetric and unimodal function such that $\int K(u) \, du = 1$. Parameter $h$ is called the bandwidth of the estimator $\hat{f}_n^{KDE}$, who inherits all the mathematical properties of $K$. The function $K$ indicates the weight that observation $x_i$ has in the estimation of $x$: observations close to $x$ are weighted more important. It can be shown that

$$
\mathbb{E} \left( \hat{f}_n^{KDE}(x) \right) = f(x) + \frac{f''(x)}{2} \mu_2(K) h^2 + o(h^2)
$$

$$
\text{Var} \left( \hat{f}_n^{KDE}(x) \right) = \frac{||K||^2 f(x)}{nh} + o \left( \frac{1}{nh} \right)
$$

for some $\xi_j \in B_j$. For a $x \in B_j$ a fixed point, when $h \to 0$, $n \to \infty$ and $nh \to \infty$, we get the classical properties for the histogram:

$$
\mathbb{E} \left( \hat{f}_n^{H}(x) \right) \to f(x), \quad \text{Var} \left( \hat{f}_n^{H}(x) \right) \to 0,
$$

and, moreover, if $f$ is locally Lipschitz the histogram is mean square consistent, i.e. MSE$(\hat{f}_n^{H}(x)) = \text{Bias}^2(\hat{f}_n^{H}(x)) + \text{Var}(\hat{f}_n^{H}(x)) \to 0$. It is widely used in many fields, because of its computational simplicity. The histogram depends on two parameters: the bin width $h$ and an origin $x_0$ to fix the grid. There is a huge literature that proposes several optimal choices for $h$ using different criteria. If we suppose that the underlying density $f$ is Gaussian, it can be shown (see Scott 1979) that an optimal choice for $h$ is of order $n^{-1/3}$. With these values, the histogram has a rate of convergence of order $n^{-2/3}$ with respect to the Mean Integrated Squared Error (MISE).
where $\mu_r(K) = \int u^r K(u) \, du$. When $h \to 0$ and $nh \to \infty$, we get the classical properties as for the histogram:

$$E(\hat{f}_{n}^{\text{KDE}}(x)) \to f(x), \quad \text{Var}(\hat{f}_{n}^{\text{KDE}}(x)) \to 0.$$

As for the histogram, if $h$ is large the variance decreases but the bias is large. On the other hand, if $h$ is small, the bias is small but the variance is large. The optimal rate of convergence of KDE is of order $n^{-4/5}$ as for the frequency polygon.

### 3 Bagging of estimators

The bootstrap method was introduced in Efron (1979) and has the purpose of doing statistical inference using resamples of the original set of data. The bootstrap has three big applications: bias correction, construction of confidence interval and hypothesis testing (Efron and Tibshirani 1993). More precisely, if we have a data set $L = \{x_1, x_2, \ldots, x_n\}$ drawn from a population with distribution $F$, the non-parametric bootstrap procedure consists in drawing, with replacement, a new sample $L^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$ from $L$ with the same size. Then, the sample $L^*$ has a distribution $F_n$ which is the empirical distribution of $L$. Informally, the key idea is to think the sample $L$ as a new hypothetical population from which a (bootstrap) sample $L^*$ is draw. Then, the method treats the empirical distribution $F_n$ of the original sample $L$ as the true distribution with respect to bootstrap sample. Reiterating this procedure several times and obtaining many bootstrap samples is a cornerstone to the construction of several bootstrap based approaches.

However, the use of bootstrap in nonparametric density estimation requires some caution, particularly concerning the bias of the estimators. In our setting, let $\hat{f}$ be a nonparametric density estimator for $f$ obtained from the sample $L$.

Now we draw a bootstrap sample $L^*$ of $L$.

In the case of the kernel density estimator, the estimation over $L^*$ is without bias,

$$E \left( \hat{f}^*(x) \mid L \right) = \frac{1}{nh} \sum_{i=1}^{n} E \left( K \left( \frac{x - x_i^*}{h} \right) \right) = \frac{1}{h} E \left( K \left( \frac{x - y}{h} \right) \right) = \frac{1}{h} \sum_{y \in \text{supp}(x_i^*)} K \left( \frac{x - y}{h} \right) P(x_i^* = y) = \frac{1}{h} \sum_{j=1}^{n} K \left( \frac{x - x_j}{h} \right) \frac{1}{n} = \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{x - x_j}{h} \right) = \hat{f}(x),$$

where $\text{supp}(x_i^*)$ is the support of $x_i^*$. This simple result holds also for the histogram and for the frequency polygon too (see the demonstration of Theorem 1), that is $E(\hat{f}^*(x) \mid L) = \hat{f}(x)$. The fact is treated in Hall (1997) in this terms: “far from accurately estimating the substantial bias of $f$, the bootstrap sets the bias of this kind of density estimator equal to zero”.

© Springer
Let $\mathcal{L} = \{x_1, \ldots, x_n\}$ be a sample with unknown distribution $F$ admitting a density $f$. Also, consider $\hat{f}_n$ a density estimator evaluated in $\mathcal{L}$.

For $b \in \{1, \ldots, B\}$:

1. obtain $\mathcal{L}_b^* = \{x_1^*, \ldots, x_n^*\}$ a bootstrap sample from $\mathcal{L}$;
2. construct $\hat{f}_b^*$ the density estimator obtained over this bootstrap sample. In particular, the bandwidth $h$ is calculated over $\mathcal{L}_b^*$.

**Output:** The final estimator is the simple pointwise average of the individual estimators i.e.

$$\hat{f}^*(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b^*(x).$$

Fig. 1 Bagging of density estimators

In supervised learning, the main application of bootstrap is definitely the Bagging. It is a parallel aggregation method of individual entities that at each step $b$ draws a bootstrap sample $\mathcal{L}_b^*$ of the original sample $\mathcal{L}$ and compute an estimator (a classifier in classification or a predictor in regression) over $\mathcal{L}_b^*$. For an input $x$, the output of the Bagging method is the average in regression or the majority rule in classification of the intermediate estimators at $x$. We follow this aggregation strategy to construct new density estimators of a density function $f$. Our procedures run as described in Fig. 1.

To obtain the bagged histogram (BagHist) estimator $\hat{f}_n^{\text{BAGHIST}}$ we simply use at each step $b$, histograms $\hat{f}_n^H$ defined in (1) as $f_b^*$. Analogously for bagged frequency polygons (BagFP) $\hat{f}_n^{\text{BAGFP}}$ and bagged kernel density estimators (BagKDE) $\hat{f}_n^{\text{BAGKDE}}$, we replace $f_b^*$ with frequency polygon estimator $\hat{f}_n^{\text{FP}}$ [cf. Eq. (2)] or kernel density estimator $\hat{f}_n^{\text{KDE}}$ [cf. Eq. (3)] respectively.

### 3.1 $L^2$ consistency of the Bagging of estimators

Here we prove $L^2$ consistency, for three estimators BagHist, BagFP and BagKDE which correspond to bagging of histograms, bagging of polygon frequencies and bagging of kernel density estimators proving that, since $h \to 0$, $n \to +\infty$ and $nh \to +\infty$, for all $x$ in bin $B_j$ (in case of histogram or frequency polygon) or for all $x \in \mathbb{R}$ (in case of kernel density estimator):

$$\mathbb{E}[(\hat{f}_n^{\text{BAGHIST}}(x) - f(x))^2] \to 0, \quad \mathbb{E}[(\hat{f}_n^{\text{BAGFP}}(x) - f(x))^2] \to 0,$$

$$\mathbb{E}[(\hat{f}_n^{\text{BAGKDE}}(x) - f(x))^2] \to 0.$$

**Theorem 1** BagHist, BagFP and BagKDE are $L^2$-consistent.

**Proof** We will give a global proof, using existent results of Hall (1997) and Scott (2015) to encompass the different methods, because the demonstration for all these estimators follows the same steps. We have to compute:

(1) for the expectation $\mathbb{E}(\hat{f}(x)) = \mathbb{E}\left(\mathbb{E}[\hat{f}(x)|\mathcal{L}]\right)$

Springer
and to calculate the variance we use the decomposition
\[
\text{Var}(\hat{f}(x)) = \mathbb{E}(\text{Var}(\hat{f}(x)|\mathcal{L})) + \text{Var}(\mathbb{E}(\hat{f}(x)|\mathcal{L})).
\]

Without loss of generality, in some calculations for histogram or frequency polygon, and with the aim to simplify notations, we will assume that \(x \in B_0\).

\(1\) – **BagHist.** If \(x \in B_j = [jh, (j+1)h]\), then we have
\[
\mathbb{E}[\hat{f}_{n}\text{baghist}(x)|\mathcal{L}] = \mathbb{E}\left[\frac{1}{B} \sum_{b=1}^{B} \hat{f}^*_b(x)|\mathcal{L}\right] = \mathbb{E}\left[\hat{f}^*_b(x)|\mathcal{L}\right] = \mathbb{E}\left[\frac{v^*_j}{nh}\right] = \hat{f}^H_n(x)
\]
and if \(f\) is locally Lipschitz
\[
\left|\mathbb{E}(\mathbb{E}[\hat{f}(x)|\mathcal{L}]) - f(x)\right| = \left|\frac{D_j}{h} - f(x)\right| = \left|\frac{hf(\xi_j)}{h} - f(x)\right| 
\leq \gamma |\xi_j - x| \leq \gamma_j h \rightarrow 0
\]

\(2\) – **BagFP.** If \(x \in B_j = [(j - \frac{1}{2})h, (j + \frac{1}{2})h]\), then we have
\[
\mathbb{E}[\hat{f}_{n}\text{bagfp}(x)|\mathcal{L}] = \mathbb{E}\left[\frac{1}{B} \sum_{b=1}^{B} \hat{f}^*_b(x)|\mathcal{L}\right] = \mathbb{E}\left[\hat{f}^*_b(x)|\mathcal{L}\right] = \mathbb{E}\left[\frac{v^*_j}{nh}\right] = \hat{f}^F_n(x)
\]
and, if \(f\) has second derivative, if \(x \in B_0\):
\[
\left|\mathbb{E}\left(\mathbb{E}[\hat{f}_{n}\text{bagfp}(x)|\mathcal{L}]\right) - f(x)\right| \approx \left|f''(\xi_0)(h^2 - 3x^2)\right| \leq 4h^2 f''(\xi_0) \rightarrow 0
\]

\(3\) – **BagKDE.** For BagKDE we have:
\[
\mathbb{E}[\hat{f}_{n}\text{bagkde}(x)|\mathcal{L}] = \mathbb{E}\left[\frac{1}{B} \sum_{b=1}^{B} \hat{f}^*_b(x)|\mathcal{L}\right] = \mathbb{E}\left[\hat{f}^*_b(x)|\mathcal{L}\right] = \mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\right] = \hat{f}^\text{KDE}_n(x)
\]
and it is well known that
\[ \left| \mathbb{E} \left( \mathbb{E} [ \hat{f}_n^{\text{BAGKDE}}(x) | \mathcal{L} ] \right) - f(x) \right| \to 0 \]

(2) For variance we use the well known decomposition formula defined above.

- **BagHist**
  (A) Because of the independence and identical distribution of the bootstrap samples, if \( x \in B_0 \):

\[
\text{Var}[\hat{f}_n^{\text{BAGHIST}}(x) | \mathcal{L}] = \text{Var} \left[ \frac{1}{B} \sum_{b=1}^{B} \hat{f}^*_b(x) | \mathcal{L} \right] = \frac{1}{B} \text{Var} \left[ \hat{f}^*_b(x) | \mathcal{L} \right]
\]

\[
= \frac{1}{B} np_0^*(1 - p_0^*)
\]

where \( p_0^* \) is equal to \( \frac{\nu_0}{n} \). Taking expectation over \( \mathcal{L} \) we have

\[
\frac{1}{Bnh^2} \mathbb{E} \left( p_0^*(1 - p_0^*) \right) = \frac{1}{Bnh^2} \left[ \mathbb{E} \left( \frac{\nu_0}{n} \right) - \text{Var} \left( \frac{\nu_0}{n} \right) - \left[ \mathbb{E} \left( \frac{\nu_0}{n} \right) \right]^2 \right]
\]

\[
= \frac{1}{Bnh^2} \left( p_0 - \frac{1}{n} p_0 - \frac{1}{n} p_0^2 - p_0^2 \right)
\]

\[
= \frac{1}{Bnh^2} \left( hf(\xi_0) - \frac{h}{n} f(\xi_0) - \frac{h^2}{n} f(\xi_0)^2 - h^2 f(\xi_0)^2 \right)
\]

\[
= \frac{1}{Bnh} f(\xi_0) - \frac{1}{Bn^2 h} f(\xi_0) - \frac{1}{Bn^2 h} f(\xi_0)^2 - \frac{1}{B} f(\xi_0)^2
\]

which tends to 0 as \( n \to \infty, nh \to \infty \) and \( h \to 0 \).

(B) \( \text{Var} \left[ \mathbb{E}(\hat{f}_n^{\text{BAGHIST}}(x) | \mathcal{L}) \right] = \text{Var}(\hat{f}_n^H(x)) \leq \frac{p_0}{nh^2} = \frac{f(\xi_0)}{nh} \to 0 \)

- **BagFP**
  (A) Because of the independence and identical distribution of the bootstrap samples:

\[
\text{Var}[\hat{f}_n^{\text{BAGFP}}(x) | \mathcal{L}] = \text{Var} \left[ \frac{1}{B} \sum_{b=1}^{B} \hat{f}^*_b(x) | \mathcal{L} \right] = \frac{1}{B} \text{Var} \left[ \hat{f}^*_b(x) | \mathcal{L} \right]
\]

\[
= \frac{1}{B} \left\{ \left( \frac{1}{2} - \frac{x}{h} \right) \text{Var}(\hat{f}_0^*) + \left( \frac{1}{2} + \frac{x}{h} \right) \text{Var}(\hat{f}_1^*) \right\}
\]

\[
+ \frac{2}{B} \left\{ \left( \frac{1}{4} - \frac{x^2}{h^2} \right) \text{Cov}(\hat{f}_0^*, \hat{f}_1^*) \right\}
\]
where \( \hat{f}_0^* \) and \( \hat{f}_1^* \) are the histogram estimations over \([-h, 0]\) and \([0, h]\) respectively. As \( \text{Var}(\hat{f}_0^*) = \frac{n_p^2(1-p_0^2)}{n^2h^2} \), then taking expectation:

\[
\mathbb{E}(\text{Var}(\hat{f}_0^*)) = \mathbb{E}\left( \frac{nv_0(1-v_0)}{n^2h^2} \right) = \frac{1}{n^2h^2}\mathbb{E}\left( v_0 \left( 1 - \frac{v_0}{n} \right) \right)
\]

\[
= \frac{1}{n^2h^2} \left( \mathbb{E}(v_0) - \frac{1}{n} \mathbb{E}(v_0^2) \right)
\]

\[
= \frac{1}{n^2h^2} \left( \mathbb{E}(v_0) - \frac{1}{n} (\text{Var}(v_0) + [\mathbb{E}(v_0)]^2) \right)
\]

\[
= \frac{1}{n^2h^2} \left[ n p_0 - \frac{1}{n} (n p_0 (1 - p_0) + (n p_0)^2) \right]
\]

\[
= \frac{nhf(\xi_0)}{n^2h^2} - \frac{hf(\xi_0)}{n^2h^2} + \frac{h^2f(\xi_0)^2}{n^2h^2} - \frac{nh^2f(\xi_0)^2}{n^2h^2} \rightarrow 0
\]

In the same way \( \mathbb{E}(\text{Var}(\hat{f}_1^*)) \rightarrow 0 \). As \( \text{Cov}(\hat{f}_0^*, \hat{f}_1^*) = \frac{np_0p_1^*}{n^2h^2} \), then taking expectation we have:

\[
\mathbb{E}\left( \text{Cov}(\hat{f}_0^*, \hat{f}_1^*) \right) = \frac{1}{nh^2} \mathbb{E}\left( \frac{v_0 v_1}{n} \right) \leq \frac{1}{n^2h^2} \mathbb{E}(v_0^2) \mathbb{E}(v_1^2)
\]

\[
= \frac{1}{n^3h^2} [np_0(1 - p_0) + p_2^2][np_1(1 - p_1) + p_1^2]
\]

\[
= \frac{1}{n^3h^2} (nhf(\xi_0) - nh^2f(\xi_0)^2 + h^2f(\xi_0)^2)(nhf(\xi_1)
\]

\[
- nh^2f(\xi_1)^2 + h^2f(\xi_1)^2)
\]

\[
= \frac{1}{n^3h^2} (n^2h^2f(\xi_0) f(\xi_1) - n^2h^3f(\xi_0)f(\xi_1)^2
\]

\[
+ nh^3f(\xi_0) f(\xi_1)^2
\]

\[
+ \frac{1}{n^3h^2}(n^2h^4f(\xi_0)^2 f(\xi_1)^2 - n^2h^3f(\xi_0)^2 f(\xi_1)
\]

\[
- nh^4f(\xi_0)^2 f(\xi_1)^2)
\]

\[
+ \frac{1}{n^3h^2}(nh^3 f(\xi_0)^2 f(\xi_1) - nh^4 f(\xi_0)^2 f(\xi_1)^2
\]

\[
+ h^4 f(\xi_0)^2 f(\xi_1)^2)
\]

\[
= \frac{1}{n} f(\xi_0) f(\xi_1) - \frac{h}{n} f(\xi_0) f(\xi_1)^2 + \frac{h}{n^2} f(\xi_0) f(\xi_1)^2
\]

\[
+ \frac{h^2}{n} f(\xi_0)^2 f(\xi_1)^2 - \frac{h}{n} f(\xi_0)^2 f(\xi_1) - \frac{h^2}{n^2} f(\xi_0)^2 f(\xi_1)^2
\]

\[
+ \frac{h}{n^2} f(\xi_0)^2 f(\xi_1) - \frac{h^2}{n^2} f(\xi_0)^2 f(\xi_1)^2
\]

\[
+ \frac{h^2}{n^2} f(\xi_0)^2 f(\xi_1)^2 \rightarrow 0
\]

So we conclude that \( \mathbb{E}\left( \text{Var}\left[ \hat{f}_n^{\text{BAGFF}}(x) | \mathcal{L} \right] \right) \rightarrow 0 \)
(B) We recall from Scott (2015) pag. 103 that

\[
\text{Var}(\mathbb{E}(f_n^{\text{Bagfp}}(x)|\mathcal{L})) = \text{Var}(f_n^{\text{pp}}(x)) = \left(\frac{2x^2}{nh^3} + \frac{1}{2nh} \right) f(\xi_0) - \frac{f(\xi_0)^2}{n} + o\left(\frac{1}{n}\right)
\]

Then if \(nh \to \infty\) and \(n \to +\infty\):

\[|\text{Var}(\mathbb{E}(f_n^{\text{Bagfp}}(x)|\mathcal{L}))| \to 0\]

\(-BagKDE\)

(A) Because of the independence and identical distribution of the bootstrap samples:

\[
\text{Var}[f_n^{\text{BagKDE}}(x)|\mathcal{L}] = \text{Var}\left[\frac{1}{B} \sum_{b=1}^{B} \hat{f}_b^*(x)|\mathcal{L}\right] = \frac{1}{B} \text{Var}\left[\hat{f}_b^*(x)|\mathcal{L}\right]
\]

So we compute \(\text{Var}\left[\hat{f}_b^*(x)|\mathcal{L}\right]\):

\[
\text{Var}\left[\hat{f}_b^*(x)|\mathcal{L}\right] = \frac{1}{n} \left[\sum_{i=1}^{n} \frac{1}{(nh)^2} K^2 \left(\frac{x - x_i^*}{h}\right)\right] - \frac{1}{n^2} \left[\frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - x_i^*}{h}\right)\right]^2
\]

(a)

and therefore \(\text{Var}\left[\hat{f}_b^*(x)|\mathcal{L}\right] \leq (a) + (b)\). Taking expectation:

\[
\mathbb{E} \left(\frac{1}{n^3h^2} \sum_{i=1}^{n} K^2 \left(\frac{x - x_i^*}{h}\right)\right) = \frac{1}{n^3h^2} \sum_{i=1}^{n} \mathbb{E} \left(K^2 \left(\frac{x - x_i^*}{h}\right)\right)
\]

\[= \frac{1}{n^2h^2} \sum_{j=1}^{n} K^2 \left(\frac{x - x_{ij}}{h}\right) \mathbb{P}(x_i^* = x_{ij})
\]

\[= \frac{1}{n^3h^2} \sum_{j=1}^{n} K^2 \left(\frac{x - x_{ij}}{h}\right) \leq \frac{1}{n^3h^2} \sum_{j=1}^{n} \tilde{C}
\]

\[= \tilde{C} \frac{1}{(nh)^2} \to 0
\]

since \(K\) is bounded, \(K^2\) also.

(b)

\[
\mathbb{E} \left(\frac{1}{n^2} \left(\frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - x_i^*}{h}\right)\right)^2\right)
\]
Bagging of density estimators

\[ E \left( \frac{1}{n} \sum_{i=1}^{n} K \left( \frac{x - x_i^*}{h} \right) \right)^2 \]

\[ = \frac{1}{n^4 h^2} \sum_{j=1}^{n} \left( \sum_{i=1}^{n} K \left( \frac{x - x_i^*}{h} \right) \right)^2 P(x_i^* = x_{ij}) \]

\[ \leq \frac{1}{n^2 (nh)^2} \sum_{j=1}^{n} \left( \sum_{i=1}^{n} C \right)^2 \]

\[ = \frac{C^2}{n^2 (nh)^2} \to 0 \]

So we conclude that \( E \left( \text{Var}[\hat{f}_{bag}^{KDE}(x)|\mathcal{L}] \right) \to 0 \)

(b) It is a well known result that \( \text{Var} \left( E[\hat{f}_{bag}^{KDE}(x)|\mathcal{L}] \right) = \text{Var}(\hat{f}_n^{KDE}) \to 0 \)

So, with the usual assumption of \( n \to \infty, h \to 0, nh \to \infty \) this implies \( L^2 \) convergence for \( \hat{f}_n^{BAGHIST}, \hat{f}_n^{BAGPP} \) and \( \hat{f}_n^{BAGKDE} \).

4 Experiments

We describe in this section a series of numerical experiments aiming to show the practical performance of the bagged versions of the classical density estimators. First, we obtain a numerical estimate of the MISE on simulated data sets created following baseline densities. Then, impact of the aggregation is analyzed. We also use the bootstrapped version of the density estimators to construct a confidence interval for it and study its performance.

4.1 Simulations

Among the numerous possibilities of univariate densities, we choose eight simulation models partially following the work of Bourel et al. (2014). This choice presents a different degree of difficulty related to the number of modes, asymmetry, tail behavior and regularity. We denote them by \( M_1 \) to \( M_8 \). Their definition is the object of Table 1 and Fig. 2 shows a graphical display of the densities. The notation \( \mathcal{N}(\mu, \sigma^2) \) is used to refer to a normal distribution with mean equal to \( \mu \) and variance equal to \( \sigma^2 \), \( \mathcal{U}[a, b] \) is the uniform density over the support \( [a, b] \), and \( \chi^2_v \) is a Chi squared density with \( v \) degrees of freedom. Models 3, 4, 7 and 8 are mixtures of densities. Models 2 and 7 are asymmetrical.

At each replication we draw two datasets following each density. The first one is used for estimation purposes while the second one is left-out for evaluation (either MISE or empirical covering). All the simulations are done with the \( R \) software (R Core Team 2018).
Table 1  Simulated univariate densities

| Model | Description |
|-------|-------------|
| ($\mathcal{M}_1$) : Normal standard | Standard Gaussian density $\mathcal{N}(0, 1)$ |
| ($\mathcal{M}_2$) : Chi 10 | Chi-square density $\chi^2_{10}$ |
| ($\mathcal{M}_3$) : Mix1 | $0.5\mathcal{N}(-1, 0.3) + 0.5\mathcal{N}(1, 0.3)$ |
| ($\mathcal{M}_4$) : Claw | the Claw density (Marron and Wand 1992) |
| ($\mathcal{M}_5$) : Triangular | Symmetric triangular density with support on $[0,2]$ |
| ($\mathcal{M}_6$) : Uniform 0–1 | Uniform density $\mathcal{U}(0, 1)$ |
| ($\mathcal{M}_7$) : Mix2 | $0.5\mathcal{N}(0, 1) + 0.5 \sum_{i=1}^{10} I \left( \frac{2(i-1)}{10}, \frac{2i-1}{10} \right)$ (Rigollet and Tsybakov 2007) |
| ($\mathcal{M}_8$) : Mix3 | Mixture of uniforms $0.5\mathcal{U}[-2, -1] + 0.5\mathcal{U}[1, 2]$ |

![Fig. 2 Densities used for the simulations](image)

4.2 Quality of the estimation

We compare density estimators of different nature. On one hand side we use three individual estimators: histograms (H), frequency polygons (FP) and kernel density estimators (KDE), on the other hand, their bagged versions, respectively BagHist, BagFP and BagKDE. We also include the RASH estimator. We choose a Gaussian
kernel for the KDE. We use cross validation to calibrate the bandwidth $h$ at each step of all the intermediate estimation methods. An alternative would be to use maximum likelihood as in Bourel et al. (2014). In our framework cross validation has, in general, a better computation behavior. Also it is more general and may be used for example with dependent data as in time series.

Figure 3 represents the dependence of MISE on the sample size $n$ for the different combinations of densities and estimators. Each point represents the average of $M = 100$ times the MISE$\times 100$ of the method using $B = 200$ intermediate estimators for the aggregating methods. Notice that these plots are in log–log scale which is useful to highlight the convergence behavior. Individual values of these plots are presented in Appendix.

Let us comment on these plots. First, the adjusted lines are of relatively good quality since the points corresponding to each combination density-estimator are almost aligned. Remember that each point is the mean average of $M = 100$ replicates and so the inner replicate variability is reduced even for a few points determining each line. However, in some few cases the quality of the fit is quite poor. Now, for each panel most of the adjusted lines are almost parallels which means the methods share a similar convergence behavior. Comparing the four simple densities (leftmost panels) and the four mixtures (rightmost panels) a difference in the behavior seems to appear, at least if one looks at some bagged version as for instance the BagFP. Second, if one compares each individual estimator (H, FP and KDE) with their bagged versions, the latter success to reduce the MISE in most of the situations. An important exception if BagHist which produces almost always worst results than the intermediate estimator, i.e. the histogram. However, the asymptotic behavior is such that with large sample sizes it is able to catch the quality of the histogram and even overwhelm it on the four mixtures. On regular targets, KDE (or at least its bagged version) shows a very
competitive performance. However, in presence of multi-modality they loss relative competitiveness with large sample sizes. The fact that the results are not entirely satisfactory for the bagged version of KDE may be because KDE it is already a good and stable density estimator (more stable in any case than histogram) and, according to Breiman (1996), bagging kernel density estimators maybe degrade the performance of this stable procedure.

### 4.3 Reduction of MISE due to aggregation

We concentrate now on aggregating methods. A natural matter to look at is the quality of the aggregation as the number of bootstrap samples increases. For this, we examine the MISE of the bagged versions for a range of increasing bootstrap samples. We replicate $M = 100$ times each combination of density simulation to construct the different curves. The results of experiments are presented in Fig. 4 in a log–log scale with $n = 500$ observations.

Globally we observe that MISE decreases monotonically with increasing values of $B$ until some point between 20 and 50 bootstrap samples after which more sample do not produce further enhancement.
4.4 Variability bands

A natural by-product of bootstrap samples is the construction of confidence bands. For some level \( \alpha \), one wants to estimate the quantities \( \hat{\ell}_n(x) \) and \( \hat{u}_n(x) \) that verify

\[
P\{ \hat{\ell}_n(x) \leq f(x) \leq \hat{u}_n(x) \} \geq 1 - \alpha, \quad \forall x
\]

that is, the quantities are the borders of an interval that covers the true density \( f(x) \) at some confidence level \( (1 - \alpha) \times 100\% \). We tackle here its construction for the density estimator. Generally, a confidence band for \( f \) is centered over an estimator \( \hat{f}_n \) of \( f \) and has the form \( \hat{f}_n(x) \pm c\hat{\sigma}(x) \) for all \( x \), with \( c > 0 \). However, since nonparametric density estimators are biased, the usual construction does not yield on proper confidence bands for \( f \). Indeed, for a fixed \( x \), due to the bias \( \mathbb{E}(\hat{f}_n(x)) - f(x) \), it is not easy to derive a confidence interval using the pivotal quantity \( \frac{\hat{f}_n(x) - f(x)}{\hat{\sigma}(x)} \). So, the interval is usually centered at \( \hat{f}_n(x) = \mathbb{E}(\hat{f}_n(x)) \) instead of being around \( f(x) \). For this reason, these confidence bands are often called variability bands.

We describe two popular constructions to compare with our procedure.

1. Variability Band for histograms. Under mild conditions (Wasserman 2006) the histogram estimator \( \hat{f}^H_n(x) \) is approximately unbiased for the target density \( f(x) \). But the approximate variance is \( f(x)/(nh) \) where \( h = 1/m \) is the inverse of the number of bins \( m \). Its dependence on the unknown target is an obstacle. To circumvent it, Scott (2015) looks at \( \text{Var}(\hat{f}^H_n(x))^{1/2} \) which is approximately \( 1/(4nh) \) and thus independent of \( f(x) \). We define \( \bar{f}^H_n = \mathbb{E}(\hat{f}^H_n(x)) \) as the target and as we say before the confidence band will not take account of the bias but only of the variability of the estimator. Then, using a normal approximation it is easy to show that (Wasserman 2006, p. 130):

\[
l_n(x) = \left( \max \left\{ \sqrt{\hat{f}^H_n(x)} - c, 0 \right\} \right)^2, \quad u_n(x) = \left( \sqrt{\hat{f}^H_n(x)} + c \right)^2
\]

where \( c = \frac{z_{\alpha/(2m)}}{2} \frac{\sqrt{m}}{n} \) give an approximate variability band for \( \bar{f}^H_n \) at \( (1 - \alpha) \times 100\% \) of confidence.

2. Variability band for KDE. As we have shown with the histogram, since the variance of \( \hat{f}^\text{KDE}_n(x) \) also involves the true density \( f \), it is more suitable to use its square root (see Bowman and Azzalini 1997). In the case of the kernel density estimator \( \text{Var}(\hat{f}^\text{KDE}_n(x))^{1/2} \approx \frac{||K||^2}{4nh_n} \) and again does not depend on the true unknown density \( f \). On this square root scale, for a fixed point \( x \) we consider the interval that back to the original scale is given by

\[
l_n(x) = \left( \hat{f}^\text{KDE}_n(x) - \frac{||K||^2}{\sqrt{4nh_n}} \right)^2, \quad u_n(x) = \left( \hat{f}^\text{KDE}_n(x) + \frac{||K||^2}{\sqrt{4nh_n}} \right)^2.
\]

As we said before this is not a confidence band for the true density \( f \), because of the bias so we will talk about a variability band.
3. **Bootstrap based confidence interval and resulting tube.** The bootstrapped sample induces a distribution that can be used to assess the variability of the estimator. Indeed, the simple superposition of the individual estimators (histogram, frequency polygon, kernel density estimator) gives a coarse idea of the uncertainty around the aggregate estimator. More the scatter of individual density estimators is dispersed, higher is the variance of the estimator. For the concrete construction of the confidence band we first fix the abscissa \( x \in \mathbb{R} \). Then we consider the set of bootstrapped density estimators evaluated at that point, i.e. \( \{ \hat{f}^*_1(x), \ldots, \hat{f}^*_B(x) \} \). This set is a collection of \( B \) univariate measures. Note that the bagged estimation is the mean average of this collection. Then, a \((1 - \alpha) \times 100\%\) confidence interval can be obtained by considering the empirical quantiles at \( \alpha/2 \) and \( 1 - \alpha/2 \) of this ensemble.

In Fig. 5 we show the different constructions generated by these methods for a standard normal target. Each panel corresponds to one of the three intermediate density estimators (Histogram, frequency polygons and KDE). At each time, the bootstrapped construction for the confidence interval is represented. For the histogram and KDE we also draw the constructions detailed above. The confidence interval for the histogram succeeds to cover the true density but produces an arguably too large band. For KDE the confidence for the intermediate estimator is very good, and the one obtained by bootstrap sampling presents a relatively high variability.

We compare the alternative constructions of the confidence band using two metrics. The aim is to obtain the narrowest band that warranties a given nominal coverage. For this, we consider the empirical coverage of the bands and its mean width. Let us call \( \hat{\ell}_n(t_i) \) and \( \hat{u}_n(t_i) \) the lower and upper bounds of the confidence bands, evaluated at points \( t_i, i = 1, 2, \ldots, N \). Then, we call the empirical mean coverage of the target \( f(t_i) \) the quantity
Table 2 Mean empirical coverage and mean interval widths for the densities and estimators considered

| Density   | Coverage | Mean width |
|-----------|----------|------------|
|           | Hist     | FP         | KDE        | KDE-sm     |
| Normal    | 95.45    | 92.54      | 92.83      | 96.54      |
| Chi10     | 95.75    | 92.88      | 92.96      | 94.79      |
| Mix1      | 95.55    | 93.11      | 90.34      | 95.27      |
| Claw      | 95.05    | 89.79      | 87.07      | 91.33      |
| Triangular| 95.79    | 92.74      | 92.69      | 94.94      |
| Uniform 0–1| 92.20   | 89.40      | 88.61      | 90.56      |
| Mix2      | 77.78    | 41.00      | 63.14      | 47.73      |
| Mix3      | 88.86    | 85.59      | 83.41      | 89.84      |

The mean width of the interval is defined by

\[
\frac{1}{N} \sum_{i=1}^{N} \left[ \hat{u}_n(t_i) - \hat{f}(t_i) \right],
\]

The mean width of the interval is defined by

\[
\frac{1}{N} \sum_{i=1}^{N} \left[ \hat{u}_n(t_i) - \hat{f}(t_i) \right].
\]

In this experiment, we set the confidence level at 95% to construct the variability bands. We give as reference the variability band constructed through the kernel density estimator as explained before (we denote this method as KDE-sm). The construction for the individual histograms produces too large bands which always cover the true density. For this reason, they are not presented in Table 2.

In general, the bands cover reasonably well the simplest densities (on the top rows) having more difficulties with the more exotic density models (on the bottom rows). Particularly, the last three densities in the table are too difficult targets producing very low empirical coverings. If we look at the methods, the construction using bootstrapped histograms is the best one among the bootstrap based ones, and give a fair competitor to KDE-sm.

5 Conclusions

In this work we present three univariate density estimators obtained by aggregation such as in Bagging. For each method, the intermediate estimators are histograms, frequency polygons or kernel density estimators. We prove the $L^2$ consistency of the three estimators and do several simulations over several densities with different characteristics. Also, we bring a way to compute a kind of confidence band, which is more close to a pointwise variability band in the sense that the authors who studied on this subject give. This construction needs a deeper study to be more conclusive.
conclusions about it. Another clue for future work is to investigate the natural extension of considering the bagged construction over multivariate densities.

Another continuation of this work is obtaining convergence results using metrics others than the mean-square-error. Also, convergence rates may be of interest in order to have a formal view of Fig. 3. Finally, the problem of density estimation could be overpassed by allowing the ideas detailed on this paper to be applied on other connected problems. For this, two key points must be recalled from our context. First, we use individual estimators that are coherent, that is each individual estimator converges to the target. Second, the average of the individual estimators is trivially also a density estimator. Then, one could image several situations where individual coherent estimators are bagged to construct a new aggregate estimator in both supervised or unsupervised learning tasks. In particular other functional estimation problems, like regression, copula or hazard rate, may benefit from the construction of variability bands.

Acknowledgements We would like to thank project ECOS-2014 Aprendizaje Automático para la Modelización y el Análisis de Recursos Naturales, no. U14E02, the LIA-IFUM and the ANII-Uruguay for their financial support.

Appendix: Additional results

Quality of the estimation

For sake of completeness we present in this appendix the individual values of Fig. 3. In the following tables (Tables 3, 4, 5, 6, 7), values are $100 \times \text{MISE}$ obtained as mean average over 100 replicates. At each line, best results are shown in bold.

| Table 3 MISE for sample size $n = 50$ |
|--------------------------------------|
| Hist | FP | KDE | BagHist | BagFP | BagKDE | RASH |
|------|----|-----|---------|-------|--------|------|
| Normal | 1.1447 | 0.6171 | 0.3181 | 10.3243 | **0.2487** | 1.5195 | 0.6025 |
| Chi10 | 0.0696 | 0.0374 | 0.0254 | 0.5635 | **0.0181** | 0.0919 | 0.0334 |
| Mix1 | 3.8260 | 2.1742 | **1.3801** | 27.4794 | 1.8937 | 3.0423 | 1.9464 |
| Claw | 4.9526 | 3.6531 | 2.7909 | 17.8066 | **2.2902** | 4.1301 | 3.4380 |
| Triangular | 7.1220 | 4.4168 | 2.2008 | 60.1150 | **1.7936** | 8.6147 | 4.0222 |
| Uniform 0-1 | 17.2112 | 11.4841 | **6.8795** | 175.0100 | 6.9725 | 21.2978 | 8.5623 |
| Mix2 | 6.2811 | 6.0856 | 6.3647 | 16.3470 | 5.8546 | **5.7913** | 5.9907 |
| Mix3 | 4.9186 | 3.0408 | **1.8700** | 22.3239 | 2.0799 | 2.5847 | 2.3357 |
### Table 4  MISE for sample size \( n = 100 \)

|        | Hist | FP   | KDE | BagHist | BagFP | BagKDE | RASH  |
|--------|------|------|-----|---------|-------|--------|-------|
| Normal | 0.7098 | 0.4016 | 0.2665 | 4.9699 | **0.1665** | 0.9510 | 0.3908 |
| Chi10  | 0.0422 | 0.0215 | 0.0144 | 0.2689 | **0.0089** | 0.0549 | 0.0198 |
| Mix1   | 2.5451 | 1.3417 | **0.8085** | 12.7401 | 2.5901 | 1.9115 | 1.1964 |
| Claw   | 2.6535 | 1.9604 | **1.6735** | 8.1400 | 2.1407 | 2.3149 | 1.8267 |
| Triangular | 5.0242 | 3.1358 | 1.2952 | 33.9450 | **0.8356** | 5.8185 | 2.8512 |
| Uniform 0-1 | 7.9732 | 5.2470 | 5.2962 | 87.8440 | **3.8460** | 14.8139 | 4.1718 |
| Mix2   | 5.8523 | 5.5056 | 6.0074 | 8.9022 | 5.9432 | **4.5841** | 5.2529 |
| Mix3   | 2.4888 | 1.6908 | 1.4754 | 11.2250 | 2.1891 | 1.8702 | **1.1925** |

### Table 5  MISE for sample size \( n = 200 \)

|        | Hist | FP   | KDE | BagHist | BagFP | BagKDE | RASH  |
|--------|------|------|-----|---------|-------|--------|-------|
| Normal | 0.4077 | 0.2059 | 0.1394 | 2.4724 | **0.0827** | 0.5840 | 0.1937 |
| Chi10  | 0.0242 | 0.0121 | 0.0084 | 0.1246 | **0.0075** | 0.0305 | 0.0111 |
| Mix1   | 1.6231 | 0.8703 | **0.5597** | 6.1511 | 3.2618 | 1.2028 | 0.8753 |
| Claw   | 1.7414 | 1.1899 | **0.9338** | 3.7563 | 2.0185 | 1.4601 | 1.0360 |
| Triangular | 2.4070 | 1.4362 | 0.7807 | 16.6818 | **0.5730** | 3.4612 | 1.3875 |
| Uniform 0-1 | 3.7493 | 2.5354 | 3.9036 | 44.4120 | 2.7891 | 10.1004 | **2.1922** |
| Mix2   | 4.2920 | 3.2531 | 5.7958 | 4.3796 | 5.9281 | 3.5878 | **2.6659** |
| Mix3   | 1.6959 | 1.2517 | 1.0518 | 5.5502 | 2.3213 | 1.2049 | **0.8032** |

### Table 6  MISE for sample size \( n = 500 \)

|        | Hist | FP   | KDE | BagHist | BagFP | BagKDE | RASH  |
|--------|------|------|-----|---------|-------|--------|-------|
| Normal | 0.2254 | 0.1236 | 0.0663 | 0.8958 | **0.0500** | 0.2977 | 0.1203 |
| Chi10  | 0.0116 | 0.0057 | 0.0038 | 0.0470 | **0.0031** | 0.0151 | 0.0056 |
| Mix1   | 0.7855 | 0.3997 | **0.2542** | 2.3672 | 3.0443 | 0.5857 | 0.4018 |
| Claw   | 0.9964 | 0.6338 | **0.4747** | 1.3815 | 1.8236 | 0.7305 | 0.5522 |
| Triangular | 1.2801 | 0.7536 | 0.3709 | 6.6334 | **0.3338** | 1.7276 | 0.7614 |
| Uniform 0-1 | 1.5125 | **0.9942** | 2.4329 | 18.6821 | 1.6281 | 5.846 | 1.1093 |
| Mix2   | 2.9831 | 2.2515 | 3.3141 | 1.9641 | 5.6468 | 2.0181 | **1.7904** |
| Mix3   | 1.2756 | 1.0160 | 0.7029 | 2.3993 | 1.1044 | 0.7520 | **0.6518** |
Table 7 MISE for sample size $n = 1000$

|       | Hist | FP  | KDE  | BagHist | BagFP | BagKDE | RASH |
|-------|------|-----|------|---------|-------|--------|------|
| Normal| 0.1220 | 0.0599 | 0.0360 | 0.4380 | **0.0249** | 0.1804 | 0.0591 |
| Chi10 | 0.0068 | 0.0031 | 0.0020 | 0.0218 | **0.0016** | 0.0088 | 0.0031 |
| Mix1  | 0.4877 | 0.2512 | **0.1465** | 1.1433 | 2.0536 | 0.3230 | 0.2569 |
| Claw  | 0.6064 | 0.3468 | **0.2559** | 0.6452 | 1.6762 | 0.3832 | 0.3082 |
| Triangular | 0.8526 | 0.5124 | 0.2713 | 3.4204 | **0.2305** | 1.1284 | 0.5016 |
| Uniform 0-1 | 0.7052 | **0.4630** | 1.7129 | 9.6299 | 1.2285 | 3.5032 | 0.7965 |
| Mix2  | 2.3140 | 1.8291 | 1.5237 | **1.2728** | 5.4538 | 1.5519 | 1.4157 |
| Mix3  | 1.2970 | 1.0580 | **0.4948** | 1.1314 | 0.6291 | 0.5210 | 0.6376 |

References

Botev ZI, Grotowski JF, Kroese DP (2010) Kernel density estimation via diffusion. Ann Stat 38(5):2916–2957
Bourel M, Ghattas B (2013) Aggregating density estimators: an empirical study. Open J Stat 3(5):344–355
Bourel M, Ghattas B, Fraiman R (2014) Random average shifted histograms. Comput Stat Data Anal 79:149–164
Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Oxford statistical science series. OUP Oxford, Oxford
Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
Breiman L (2001) Random forests. Mach Learn 45(1):5–32
Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26
Efron B, Tibshirani R (1993) An introduction to the bootstrap. Monographs on statistics and applied probability. Chapman & Hall, Boca Raton
Fisher R (1932) Statistical methods for research workers. Biological monographs and manuals. Oliver and Boyd, New York
Freund Y, Schapire R (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
Glodek M, Schels M, Schwenker F (2013) Ensemble gaussian mixture models for probability density estimation. Comput Stat 28(1):127–138
Hall P (1997) The bootstrap and Edgeworth expansion. Springer series in statistics. Springer, New York
Marron J, Wand M (1992) Exact mean integrated square error. Ann Stat 20(2):712–736
R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Ridgway G (2002) Looking for lumps: Boosting and bagging for density estimation. Comput Stat Data Anal 38(4):379–392
Rigollet P, Tsybakov AB (2007) Linear and convex aggregation of density estimators. Math Methods Stat 16(3):260–280
Rosset S, Segal E (2002) Boosting density estimation. In: Advances in neural information processing systems (NIPS), pp 641–648
Scott D (1985a) Averaged shifted histogram: effective nonparametric density estimators in several dimensions. Ann Stat 13(3):1024–1040
Scott D (1985b) Frequency polygons: theory and application. J Am Stat Assoc 80(390):348–354
Scott D (2015) Multivariate density estimation: theory, practice, and visualization. Wiley series in probability and statistics. Wiley, Hoboken
Scott DW (1979) On optimal and data-based histograms. Biometrika 66:605–610
Smyth P, Wolpert D (1999) Linearly combining density estimators via stacking. Mach Learn 36(1–2):59–83
Song X, Yang K, Pavel M (2004) Density boosting for gaussian mixtures. Neural Inf Process 3316:508–515
Wasserman L (2006) All of nonparametric statistics. Springer texts in statistics. Springer, New York
Wolpert D (1992) Stacked generalization. Neural Netw 5:241–259

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.