Bagging of Density Estimators

Mathias Bourel, Jairo Cugliari

Bagging of Density Estimators

Mathias Bourel
IMERL, Facultad de Ingeniería
Universidad de la República, Uruguay
mbourel@fing.edu.uy

Jairo Cugliari
Laboratoire ERIC
Université Lumire Lyon 2, France
Jairo.Cugliari@univ-lyon2.fr

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. We prove the L^2-consistency of these new estimators and compare them to several similar approaches by extensive simulations. Based on them, we give also a way to construct non parametric pointwise confidence intervals for the target density.

Keywords: density estimation; aggregation; bagging; histogram; polygon frequency; kernel density estimator.

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 point wise estimate $\hat{f}(x)$ of $f(x)$ for all $x \in \mathbb{R}^d$, and second, by assessing the uncertainty of the point wise estimate through the construction of a confidence interval for f. Of course both problems implies 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 the constant level for at each partition, on one hand 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 added 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. section 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 (1985a) 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). However, a reasonable question is to ask whether further improvement can be achieved from these construction 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 (principally binary classification) or regression from the theoretical viewpoint and have very high 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 for 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 most simplest density estimator, the histograms, each one constructed over several different deterministic grids in Average Shifted Histograms, ASH (Scott, 1985b). 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 show 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 follow. Section 2 introduces notation and reviews the framework context we need about density estimators. In Section 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 Section 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 Section 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:

$$\hat{f}_n^{\mathrm{H}}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j \in \mathbb{Z}} \mathbb{1}_{B_j}(x_i) \mathbb{1}_{B_j}(x) = \frac{\nu_j}{nh} \tag{1}$$

  where $\nu_j \sim \mathrm{Bin}(n, p_j)$, 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:

$$\mathbb{E}\big(\hat{f}_n^{\mathrm{H}}(x)\big) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j\in\mathbb{Z}} \mathbb{P}(x_i \in B_j)\mathbb{1}_{B_j}(x) = \frac{1}{nh} n\mathbb{P}(x_i \in B_j) = \frac{p_j}{h} = f(\xi_j)$$

$$\mathrm{Var}\big(\hat{f}_n^{\mathrm{H}}(x)\big) = \frac{1}{n^2h^2}\,\mathrm{Var}(\nu_j) = \frac{p_j(1-p_j)}{nh^2} \leq \frac{p_j}{nh^2} = \frac{f(\xi_j)}{nh}$$

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:

$$\mathbb{E}\big(\hat{f}_n^{\mathrm{H}}(x)\big) \to f(x),\ \mathrm{Var}\big(\hat{f}_n^{\mathrm{H}}(x)\big) \to 0.$$

and, moreover, if f is locally Lipschitz the histogram is mean square consistent, i.e $\mathrm{MSE}\left(\hat{f}_n^{\mathrm{H}}(x)\right) = \mathrm{Bias}^2\left(\hat{f}_n^{\mathrm{H}}(x)\right) + \mathrm{Var}\left(\hat{f}_n^{\mathrm{H}}(x)\right) \to 0$.

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 this value the histogram has a rate of convergence of order $n^{-2/3}$ with respect to the Mean Integrated Squared Error (MISE).

- *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

$$\hat{f}_n^{\mathrm{FP}}(x) = \left(\frac{1}{2} + j - \frac{x}{h}\right)\frac{\nu_j}{nh} + \left(\frac{1}{2} - j + \frac{x}{h}\right)\frac{\nu_{j+1}}{nh} \tag{2}$$

The frequency polygon was deeply studied in Scott (1985a). With respect to the histogram, it presents the 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 order $n^{-4/5}$ with respect to the MISE (Scott (1985a)).

- *Kernel Density Estimators.* A *Kernel Density Estimator, KDE,* is a function defined by

$$\hat{f}_n^{\mathrm{KDE}}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3}$$

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 $\hat{f}_n^{\mathrm{KDE}}$, 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

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

$$\text{Var}\left(\hat{f}_n^{\text{KDE}}(x)\right) = \frac{||K||_2^2 f(x)}{nh} + o\left(\frac{1}{nh}\right)$$

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:

$$\mathbb{E}\big(\hat{f}_n^{\text{KDE}}(x)\big) \to f(x),\ \text{Var}\big(\hat{f}_n^{\text{KDE}}(x)\big) \to 0.$$

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

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

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

$$\begin{aligned}\mathbb{E}\left(\hat{f}^*(x)|\mathcal{L}\right) =& \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left(K\left(\frac{x-x_i^*}{h}\right)\right)\\ =& \frac{1}{h}\mathbb{E}\left(K\left(\frac{x-x_i^*}{h}\right)\right) = \frac{1}{h}\sum_{y\in R_{x_i^*}} K\left(\frac{x-y}{h}\right)\mathbb{P}(x_i^*=y)\\ =& \frac{1}{h}\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)\frac{1}{n} = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right) = \hat{f}(x).\end{aligned}$$

This simple results holds also for the histogram and for the frequency polygon too (see demonstration of theorem 1), that is $\mathbb{E}(\widehat{f}^*(x)|\mathcal{L}) = \widehat{f}$. 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".

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 draw 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 follows (Figure 1):

Let $\mathcal{L} = \{x_1, \ldots, x_n\}$ be a sample with unknown distribution F admitting a density f. Also, considerer $\widehat{f}_n$ 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.

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

Figure 1: Bagging of density estimators

To obtain the bagged histogram (BagHist) estimator $\hat{f}_n^{\text{BAGHIST}}$ we simply use at each step b, histograms $\hat{f}_n^{\text{H}}$ 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):

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

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

Proof. We will give a global proof, inspired by 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 $\mathbb{E}(\hat{f}(x)) = \mathbb{E}\left(\mathbb{E}[\hat{f}(x)|\mathcal{L}]\right)$

(2) and to calculate the variance we use the decomposition

$$\text{Var}(\hat{f}(x)) = \underbrace{\mathbb{E}(\text{Var}(\hat{f}(x)|\mathcal{L}))}_{(A)} + \underbrace{\text{Var}(\mathbb{E}(\hat{f}(x)|\mathcal{L}))}_{(B)}$$

Without loss of generality, in some calculation 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

$$\mathbb{E}[\hat{f}_n^{\text{BAGHIST}}(x)|\mathcal{L}] = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x)|\mathcal{L}\right] = \mathbb{E}\left[\hat{f}_b^*(x)|\mathcal{L}\right] = \mathbb{E}\left[\frac{\nu_j^*}{nh}\right] = \frac{\nu_j}{nh} = \hat{f}_n^{\text{H}}(x)$$

and if f is locally Lipschitz

$$\left|\mathbb{E}(\mathbb{E}[\hat{f}(x)|\mathcal{L}]) - f(x)\right| = \left|\frac{p_j}{h} - f(x)\right| = \left|\frac{hf(\xi_j)}{h} - f(x)\right| \leq \gamma_j|\xi_j - x| \leq \gamma_j h \to 0$$

• *BagFP.* If $x \in B_j = \left[(j-\frac{1}{2})h, (j+\frac{1}{2})h\right]$, then we have

$$\mathbb{E}[\hat{f}_n^{\text{BAGFP}}(x)|\mathcal{L}] = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x)|\mathcal{L}\right] = \mathbb{E}\left[\hat{f}_b^*(x)|\mathcal{L}\right] = \mathbb{E}\left[\left(\frac{1}{2}+j-\frac{x}{h}\right)\frac{\nu_j^*}{nh} + \left(\frac{1}{2}-j+\frac{x}{h}\right)\frac{\nu_{j+1}^*}{nh}|\mathcal{L}\right]$$

$$= \left(\frac{1}{2}+j-\frac{x}{h}\right)\frac{\nu_j}{nh} + \left(\frac{1}{2}-j\ +\frac{x}{h}\right)\frac{\nu_{j+1}}{nh} = \hat{f}_n^{\text{FP}}(x)$$

and, if f has second derivative, if $x \in B_0$:

$$\left|\mathbb{E}\left(\mathbb{E}[\hat{f}_n^{\text{BAGFP}}(x)|\mathcal{L}]\right) - f(x)\right| \approx \left|f''(\xi_0)(h^2 - 3x^2)\right| \leq 4h^2 f''(\xi_0) \to 0$$

- *BagKDE.* For BagKDE we have:

$$\mathbb{E}[\hat{f}_n^{\text{BAGKDE}}(x)|\mathcal{L}] = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x)|\mathcal{L}\right] = \mathbb{E}\left[\hat{f}_b^*(x)|\mathcal{L}\right] = \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i^*}{h}\right)\right] = \hat{f}_n^{\text{KDE}}$$

and this 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 formula defined above.

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

$$\text{Var}[\hat{f}_n^{\text{BAGHIST}}(x)|\mathcal{L}] = \text{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x)|\mathcal{L}\right] = \frac{1}{B}\text{Var}\left[\hat{f}_b^*(x)|\mathcal{L}\right] = \frac{1}{B}\frac{np_0^*(1-p_0^*)}{(nh)^2}$$

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

$$\begin{aligned}\frac{1}{Bnh^2}\mathbb{E}\left(p_0^*(1-p_0^*)\right) =& \frac{1}{Bnh^2}\left[\mathbb{E}\left(\frac{\nu_0}{n}\right) - \text{Var}\left(\frac{\nu_0}{n}\right) - \left[\mathbb{E}\left(\frac{\nu_0}{n}\right)\right]^2\right] \\ =& \frac{1}{Bnh^2}\left(p_0 - \frac{1}{n}p_0 - \frac{1}{n}p_0^2 - p_0^2\right) \\ =& \frac{1}{Bnh^2}\left(hf(\xi_0) - \frac{h}{n}f(\xi_0) - \frac{h^2}{n}f(\xi_0)^2 - h^2f(\xi_0)^2\right) \\ =& \frac{1}{Bnh}f(\xi_0) - \frac{1}{Bn^2h}f(\xi_0) - \frac{1}{Bn^2}f(\xi_0)^2 - \frac{1}{Bn}f(\xi_0)^2 \to 0\end{aligned}$$

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

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

$$\begin{aligned}\text{Var}[\hat{f}_n^{\text{BAGFP}}(x)|\mathcal{L}] &= \text{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b^*(x)|\mathcal{L}\right] = \frac{1}{B}\text{Var}\left[\hat{f}_b^*(x)|\mathcal{L}\right] \\ &= \frac{1}{B}\left\{\left(\frac{1}{2} - \frac{x}{h}\right)\text{Var}(\hat{f}_0^*) + \left(\frac{1}{2} + \frac{x}{h}\right)\text{Var}(\hat{f}_1^*) + 2\left(\frac{1}{4} - \frac{x^2}{h^2}\right)\text{Cov}(\hat{f}_0^*, \hat{f}_1^*)\right\}\end{aligned}$$

where $\hat{f}_0^*$ and $\hat{f}_1^*$ are the histogram estimations over $[-h, 0]$ and $[0, h]$ respectively. As $\text{Var}(\hat{f}_0^*) = \frac{np_o^*(1-p_o^*)}{n^2h^2}$, then taking expectation:

$$\begin{aligned}\mathbb{E}(\mathrm{Var}(\hat{f}_0^*)) =& \mathbb{E}\left(\frac{n\frac{\nu_0}{n}(1-\frac{\nu_0}{n})}{n^2h^2}\right) = \frac{1}{n^2h^2}\,\mathbb{E}\left(\nu_0\left(1-\frac{\nu_0}{n}\right)\right)\\ =&\frac{1}{n^2h^2}\left(\mathbb{E}(\nu_0) - \frac{1}{n}\,\mathbb{E}(\nu_0^2)\right)\\ =&\frac{1}{n^2h^2}\left(\mathbb{E}(\nu_0) - \frac{1}{n}(\mathrm{Var}(\nu_0) + [\mathbb{E}(\nu_0)]^2)\right)\\ =&\frac{1}{n^2h^2}\left[np_0 - \frac{1}{n}(np_0(1-p_0) + (np_0)^2)\right]\\ =&\frac{nhf(\xi_0)}{n^2h^2} - \frac{hf(\xi_0)}{n^2h^2} + \frac{h^2f(\xi_0)^2}{n^2h^2} - \frac{nh^2f(\xi_0)^2}{n^2h^2} \to 0\end{aligned}$$

In the same way $\mathbb{E}(\mathrm{Var}(\hat{f}_1^*)) \to 0$.

As $\mathrm{Cov}(\hat{f}_0^*, \hat{f}_1^*) = \frac{-np_0^*p_1^*}{n^2h^2}$, then taking expectation we have

$$\begin{aligned}\mathbb{E}\left(\mathrm{Cov}(\hat{f}_0^*, \hat{f}_1^*)\right) =&\frac{1}{nh^2}\,\mathbb{E}\left(\frac{\nu_0}{n}\frac{\nu_1}{n}\right) \le \frac{1}{n^3h^2}\,\mathbb{E}(\nu_0^2)\,\mathbb{E}(\nu_1^2) = \frac{1}{n^3h^2}[np_0(1-p_0)+p_0^2][np_1(1-p_1)+p_1^2]\\ =&\frac{1}{n^3h^2}(nhf(\xi_0) - nh^2f(\xi_0)^2 + h^2f(\xi_0)^2)(nhf(\xi_1) - nh^2f(\xi_1)^2 + h^2f(\xi_1)^2)\\ =&\frac{1}{n^3h^2}(n^2h^2f(\xi_0)f(\xi_1) - n^2h^3f(\xi_0)f(\xi_1)^2 + nh^3f(\xi_0)f(\xi_1)^2\\ &+ \frac{1}{n^3h^2}(n^2h^4f(\xi_0)^2f(\xi_1)^2 - n^2h^3f(\xi_0)^2f(\xi_1) - nh^4f(\xi_0)^2f(\xi_1)^2)\\ &+ \frac{1}{n^3h^2}(nh^3f(\xi_0)^2f(\xi_1) - nh^4f(\xi_0)^2f(\xi_1)^2 + h^4f(\xi_0)^2f(\xi_1)^2)\\ =&\frac{1}{n}f(\xi_0)f(\xi_1) - \frac{h}{n}f(\xi_0)f(\xi_1)^2 + \frac{h}{n^2}f(\xi_0)f(\xi_1)^2\\ &+ \frac{h^2}{n}f(\xi_0)^2f(\xi_1)^2 - \frac{h}{n}f(\xi_0)^2f(\xi_1) - \frac{h^2}{n^2}f(\xi_0)^2f(\xi_1)^2\\ &+ \frac{h}{n^2}f(\xi_0)^2f(\xi_1) - \frac{h^2}{n^2}f(\xi_0)^2f(\xi_1)^2 + \frac{h^2}{n^3}f(\xi_0)^2f(\xi_1)^2 \to 0\end{aligned}$$

So we conclude that $\mathbb{E}\left(\mathrm{Var}[\hat{f}_n^{\mathrm{BAGFP}}(x)|\mathcal{L}]\right) \to 0$

(B) We recall from Scott (2015) pag. 103 that

$$\mathrm{Var}(\mathbb{E}(\hat{f}_n^{\mathrm{BAGFP}}(x)|\mathcal{L})) = \mathrm{Var}(\hat{f}_n^{\mathrm{FP}}(x)) = \left(\frac{2x^2}{nh^3} + \frac{1}{2nh}\right)f(\xi_0) - \frac{f(\xi_0)^2}{n} + o\left(\frac{1}{n}\right)$$

Then if $nh \to \infty$ and $n \to +\infty$:

$$|\,\mathrm{Var}(\mathbb{E}(\hat{f}_n^{\mathrm{BAGFP}}(x)|\mathcal{L}))| \to 0$$

- *BagKDE.*

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n^3h^2}\sum_{i=1}^{n}K^2\left(\frac{x-x_i^*}{h}\right)\right) =& \frac{1}{n^3h^2}\sum_{i=1}^{n}\mathbb{E}\left(K^2\left(\frac{x-x_i^*}{h}\right)\right)\\
=& \frac{1}{n^2h^2}\sum_{j=1}^{n}K^2\left(\frac{x-x_{ij}}{h}\right)\mathbb{P}(x_i^*=x_{ij})\\
=& \frac{1}{n^3h^2}\sum_{j=1}^{n}K^2\left(\frac{x-x_{ij}}{h}\right) \leq \frac{1}{n^3h^2}\sum_{j=1}^{n}\tilde{C}\\
=& \frac{\tilde{C}}{(nh)^2} \to 0
\end{aligned}$$

because since K is bounded, K^2 also.
(b)

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n^2}\left(\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i^*}{h}\right)\right)^2\right) =& \frac{1}{n^4h^2}\,\mathbb{E}\left(\sum_{i=1}^{n}K\left(\frac{x-x_i^*}{h}\right)\right)^2\\
=& \frac{1}{n^4h^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}K\left(\frac{x-x_i^*}{h}\right)\right)^2\mathbb{P}(x_i^*=x_{ij})\\
\leq& \frac{1}{n^5(nh)^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}C\right)^2\\
=& \frac{C^2}{n^2(nh)^2} \to 0
\end{aligned}$$

So we conclude that $\mathbb{E}\left(\text{Var}[\hat{f}_n^{\text{BAGKDE}}(x)|\mathcal{L}]\right) \to 0$

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

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

□

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. The impact of the aggregation is analyzed. We also pay attention at the repartition between integrated variance (IV) and integrated square bias (ISB). Finally, we use the bootstrapped version of the density estimator to construct confidence interval. We study the empirical covering of these confidence bands that result.

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 $\mathcal{M}1$ to $\mathcal{M}8$. Their definition is the object of Table 1 and Figure 2 shows a graphical display of the densities.

Model	Description
$(\mathcal{M}1)$: Normal Standard	Standard Gaussian density $\mathcal{N}(0,1)$
$(\mathcal{M}2)$: Chi 10	Chi-square density χ^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	$0.5\mathcal{N}(0,1)+0.5\sum_{i=1}^{10}\mathbf{1}_{\left(\frac{2(i-1)}{10},\frac{2i-1}{10}\right]}$ (Rigollet and Tsybakov, 2007)
$(\mathcal{M}8)$: Mix3	Mixture of uniforms $0.5\mathcal{U}[-2,-1]+0.5\mathcal{U}[1,2]$

Table 1: Simulated univariate densities.

The notation $\mathcal{N}(\mu,\sigma^2)$ is used to refer to a normal distribution with mean equal to μ and variance equal to σ^2, $\mathcal{U}[a,b]$ is the uniform density over the support $[a,b]$, and χ^2_ν is a Chi squared density with ν degrees of freedom. Models 3, 4, 7 and 8 are mixtures of normal densities or normal densities with indicators. Models 5, 6 and 8 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 leaved for evaluation (either MISE or empirical covering).

All the simulations are done with the **R** software, and for models $\mathcal{M}4$ we use the benchden package.

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. Also we include the RASH estimator. 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

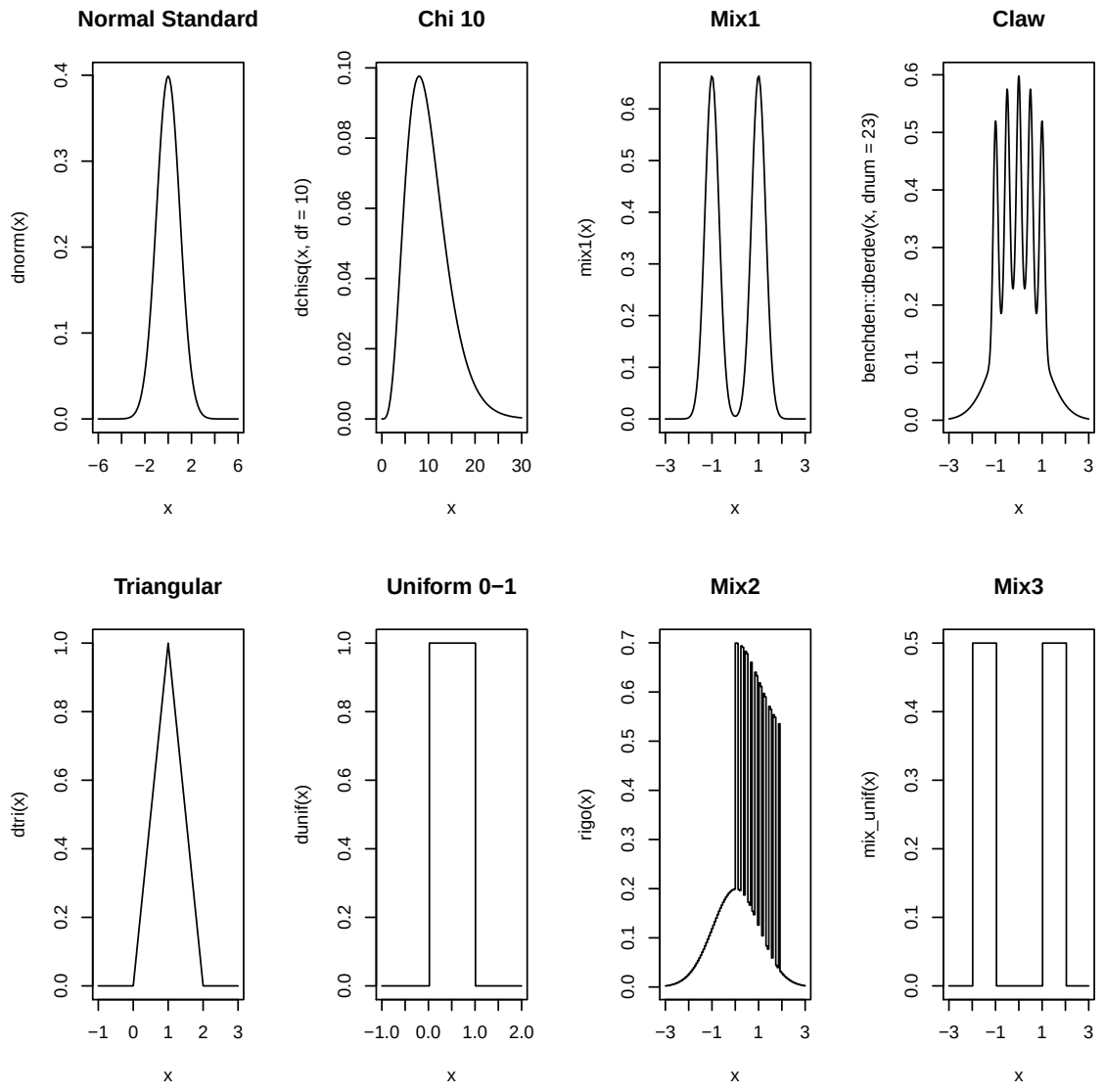


Figure 2: Densities used for the simulations.

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 rates as adjusted straight lines. Individual values of these plots are presented in AppendixA.

Let us comment these plots. First, the adjusted lines are of relative good quality since the points for each combination density-estimator are almost aligned. Remember that each point is the mean

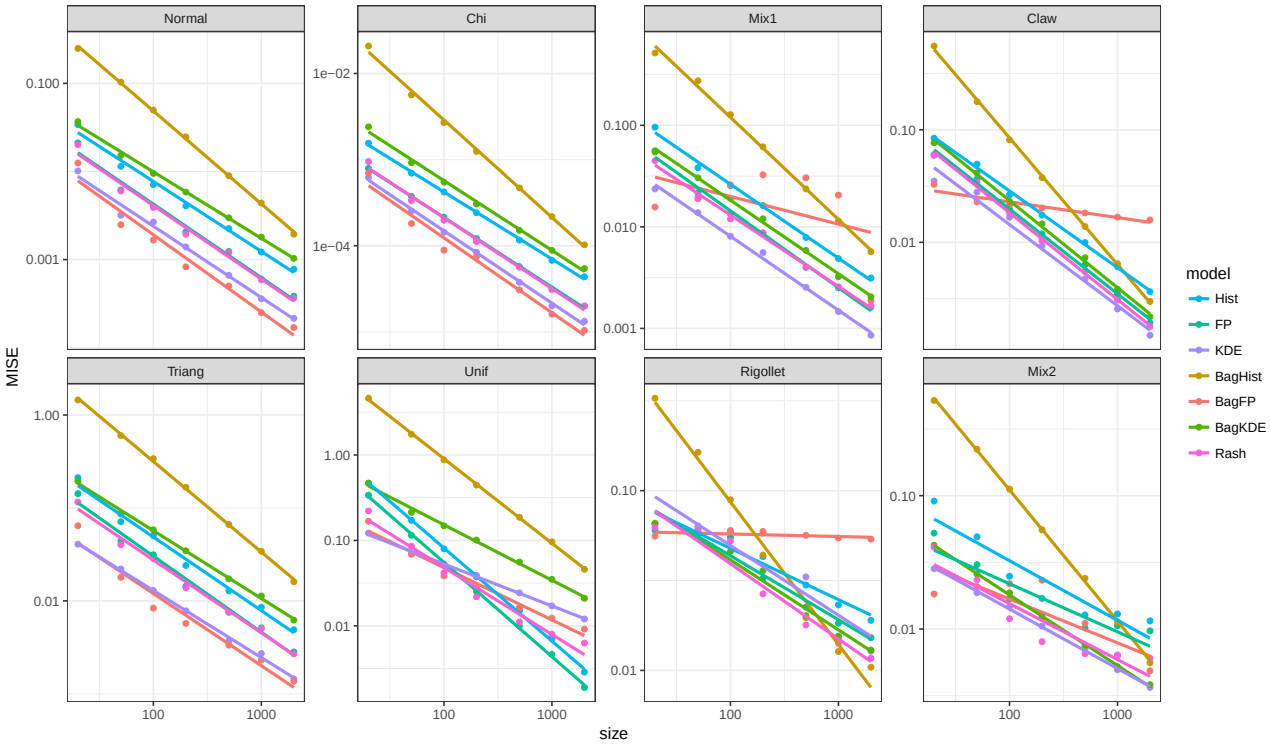


Figure 3: MISE by estimation method for the six simulation data sets in scale log.

average of $M = 100$ replicates and so the inner replicate variability is reduced even for a few points determining each line. Now, for each panel most of the adjusted lines are almost parallels which means the methods share a similar convergence behavior. The exception are KDE and BagKDE which seems to have more difficulties than the other ones in some situations. 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. On unimodal targets, KDE (or at least its bagged version) shows a very competitive performance. However, in presence of multi-modality they are competitive only for relatively small sample sizes. The fact that the results are not entirely satisfactory for the bagged version of kde may be because kde is a good and stable density estimator (more stable in any case than histogram) and, according with Breiman (1996), bagging kernel density estimators may be 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 result of experiments are

presented in Figure 4 in a log-log scale with $n = 500$ observations.

Globally we observe that MISE decreases with increases values of B until some point between 20 and 50 bootstrap samples after which more samples does not produce further enhancement. Similarly to the precedent figure, bagging KDE produce less improvement than bagging H and FP when the underlying target is multimodal.

4.4. Decomposition of MISE into ISB and ISV

Bagging succeeds in reducing MISE even for a modest number of bootstrap samples. We study now how this reduction affects the two well known components of the MISE, that is the integrated squared bias (ISB) and the integrate variance (IV). In mathematical terms, the decomposition is written as

$$\text{MISE}(\widehat{f}) = \int \mathbb{E}\big(\widehat{f}(x) - f(x)\big)^2 \, dx = \underbrace{\int \Big(\mathbb{E}\big(\widehat{f}(x)\big) - f(x)\Big)^2 \, dx}_{ISQ} + \underbrace{\int \Big(\widehat{f}(x) - \mathbb{E}\Big(\widehat{f}(x)\Big)\Big)^2 \, dx}_{IV}. \quad (4)$$

The terms are estimated using Monte Carlo simulation. We replicate M times the simulation, i.e estimation and prediction steps for each of the density models introduced before and the density estimators. Then, the estimator of MISE can be written as

$$\text{MISE}(\widehat{f}) = \frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{K} \sum_{k=1}^{K} \Big(\widehat{f}^{(k)}(x_m) - f(x_m) \Big)^2 \right].$$

The empirical counterpart of (4) is then

$$\widehat{\text{MISE}}(\widehat{f}) = \underbrace{\frac{1}{M} \sum_{m=1}^{M} \left[\left(\frac{1}{K} \sum_{k=1}^{K} \widehat{f}^{(k)}(x_m) - f(x_m) \right)^2 \right]}_{\widehat{ISB}} + \underbrace{\frac{1}{M} \sum_{m=1}^{M} \left[\left(\widehat{f}(x_m) - \frac{1}{K} \sum_{k=1}^{K} \widehat{f}^{(k)}(x_m) \right)^2 \right]}_{\widehat{IV}}$$

where K is the number of replicate of the training sample and M is the number of train/test divisions. We use $n = 500$ observations, $B = 200$ as the number of individual estimators in the aggregation, $K = 100$ and $M = 100$.

Table 2 reports the mean variation (in percentage) of MISE, ISQ and IV for each estimator and density. In all cases, MISE present reductions (negative variations). Both BagH and BagFP reduce MISE by reducing square bias and variance, while for KDE there is as systematic an increment on the variance of the estimator. However, the bias reduction is such that more than compensates the variance increment and produce a reduction of the MISE.

	Histogram			Frequency Polygon			KDE		
	MISE	Sq. Bias	Var.	MISE	Sq. Bias	Var.	MISE	Sq. Bias	Var.
normal	-34,5%	-32,9%	-37,3%	-12,8%	-20,1%	-22,0%	-37,8%	-82,4%	189%
chi2	-34,8%	-22,5%	-35,7%	-12,4%	-6,8%	-15,4%	113%	-70,0%	141%
mezcla1	-32,7%	-21,2%	-24,8%	-13,5%	-8,5%	-5,4%	-62,0%	-64,5%	101%
mezcla2	-37,9%	-13,6%	-40,7%	-14,8%	-6,4%	-15,5%	-60,9%	-67,5%	84%
bart	-34,4%	-19,7%	-22,6%	-13,6%	-17,3%	6,9%	-16,5%	-19,2%	108%
triangular	-23,1%	-19,4%	-30,8%	-11,9%	-11,1%	-22,4%	-68,9%	-79,4%	131%

Table 2: MISE, squared bias and variance reduction (in percentage) due to bagging for each estimator by density model.

4.5. Variability bands

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

$$\mathbb{P}\{\hat{l}_n(x) \leq f(x) \leq \hat{u}_n(x)\} \geq 1-\alpha, \quad \forall\, x$$

that is, the quantities are the borders of an interval that covers at 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 $\hat{f}_n(x) \pm c\hat{\sigma}_n(x)$ for all x, with $c > 0$. However, since nonparametric density estimators are biased, the usual construction does not yields on a really a confidence band for f. Indeed, for a fixed x, due to the bias $\mathbb{E}\left(\hat{f}_n(x)\right) - f(x)$, 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 $\overline{f}_n(x) = \mathbb{E}\left(\hat{f}_n(x)\right)$ 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 $\hat{f}_n^{\mathrm{H}}(x)$ 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 $\mathrm{Var}\left[\sqrt{\hat{f}_n^{\mathrm{H}}(x)}\right]$ which is approximately $1/(4nh)$ and thus independent of $f(x)$. We define $\bar{f}_n^{\mathrm{H}} = \mathbb{E}[\hat{f}_n^{\mathrm{H}}(x)]$ 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):

$$l_n(x) = \left(\max\left\{\sqrt{\hat{f}_n^{\mathrm{H}}(x)} - c, 0\right\}\right)^2, \qquad u_n(x) = \left(\sqrt{\hat{f}_n^{\mathrm{H}}(x)} + c\right)^2$$

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

2. *Variability Band for KDE.* As we have shown with the histogram, since the variance of $\hat{f}_n^{\mathrm{KDE}}(x)$ also involves the true density f, it is more suitable to use the square root (see Bowman and Azzalini (1997)). In the case of the kernel density estimator $\mathrm{Var}\left(\sqrt{\hat{f}_n^{\mathrm{KDE}}(x)}\right) \approx \frac{||K||_2^2}{4nh_n}$ 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

$$l_n(x) = \left(\hat{f}_n^{\mathrm{KDE}}(x) - \frac{||K||_2}{\sqrt{4nh_n}}\right)^2, \qquad u_n(x) = \left(\hat{f}_n^{\mathrm{KDE}}(x) + \frac{||K||_2}{\sqrt{4nh_n}}\right)^2.$$

   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 band* The bootstrapped sample induces a distribution that can be used to asses 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 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 $x \in \mathbb{R}^d$. Then we consider the set of bootstrapped density estimators evaluated at that point, i.e. $\{\hat{f}_1^*(x), \ldots, \hat{f}_B^*(x)\}$. Note that the bagged estimator is the "middle" of the tube generated. This set is a collection of B univariate measures. Then, a $(1-\alpha)\times 100\%$ confidence interval can be obtained by considering the empirical quantiles at $\alpha/2$ and $1-\alpha/2$ for this ensemble.

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 $\hat{l}_n(t_i)$ 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 $f(x)$ the quantity

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{\{\hat{l}_n(t_i) \leq f(t_i) \leq \hat{u}_n(t_i)\}},$$

where $\mathbb{I}_A$ is the indicator function of the set A. The mean width of the interval is defined by

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

We give as reference the variability band constructed through the kernel density estimator as explained before (we denote this method as Kde-sm). The ones that we obtained for the histograms are too large covering always the true density, for this reason they are not presented in Table 3.

	Coverage				Mean width			
Density	Hist	FP	Kde	Kde-sm	Hist	FP	Kde	Kde-sm
Normal	95.45	92.54	92.83	96.54	0.24	0.18	0.19	0.09
Chi2	95.75	92.88	92.96	94.79	0.06	0.04	0.04	0.02
Mix1	95.55	93.11	90.34	95.27	0.40	0.30	0.25	0.19
Claw	95.05	89.79	87.07	91.33	0.30	0.22	0.25	0.22
Triangular	95.79	92.74	92.69	94.94	0.66	0.51	0.45	0.23
Uniform	92.20	89.40	88.61	90.56	1.06	0.84	0.73	0.52
Mix2	77.78	41.00	63.14	47.73	0.36	0.26	0.26	0.21
Mix3	88.86	85.59	83.41	89.84	0.37	0.29	0.22	0.27

Table 3: Mean empirical coverage and mean interval widths for the densities and estimators considered.

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 algorithms and do several simulations over densities with different characteristics. In future work, these ideas and simulations should be extended to the multivariate case. Also, we bring a way to compute a kind of confidence band, which is more close to a point wise variability band in the sense that the authors who studied on this subject give. This construction needs a deeper study to be able to draw more conclusive conclusions about it.

Acknowledgements

We would like to thank project ECOS-2015 *Aprendizaje Automático para la Modelización y el Análisis de Recursos Naturales*, n^o U14E02 and the *ANII* -Uruguay for their financial support.

6. References

M. Bourel and B. Ghattas. Aggregating density estimators: an empirical study. *Open Journal of Statistics*, 3(5), 2013.

M. Bourel, B. Ghattas, and R. Fraiman. Random average shifted histograms. *Computational Statistics & Data Analysis*, 79:149–164, November 2014.

A.W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford Statistical Science Series. OUP Oxford, 1997.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Monographs on statistics and applied probability. Chapman & Hall, 1993.

R.A. Fisher. *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd, 1932.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

M. Glodek, M. Schels, and F. Schwenker. Ensemble gaussian mixture models for probability density estimation. *Computational Statistics*, 28(1):127–138, 2013.

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer New York, 1997.

J.S Marron and M.P. Wand. Exact mean integrated square error. *The Annals of Statistics*, 20(2): 712–736, 1992.

G. Ridgeway. Looking for lumps: Boosting and bagging for density estimation. *Comput. Stat. Data Anal.*, 38(4):379–392, 2002.

P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

S. Rosset and E. Segal. Boosting density estimation. In *In Advances in Neural Information Processing Systems 15*, pages 641–648. MIT Press, 2002.

D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

David W. Scott. Frequency polygons: Theory and application. *Journal of the American Statistical Association*, 80(390):348–354, 1985a. ISSN 01621459. URL `http://www.jstor.org/stable/2287895`.

D.W Scott. Averaged shifted histogram: Effective nonparametric density estimators inseveral dimensions. *The Annals of Statistics*, 13(3):1024–1040, 1985b.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2015.

P. Smyth and D. Wolpert. Linearly combining density estimators via stacking. *Mach. Learn.*, 36(1-2): 59–83, 1999.

X. Song, K. Yang, and M. Pavel. Density boosting for gaussian mixtures. *Neural Information Processing*, 3316:508–515, 2004.

L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006.

D.H. Wolpert. Stacked Generalization. *Neural Networks*, 5:241–259, 1992.

AppendixA. Additional results

Quality of the estimation

For sake of completeness we present in this annexe the individual values of plot 3. In the following tables, values are 100×MISE obtained as mean average over 100 replicates. At each line, best results are shown in blue, while worst are shown in red.

- $n = 50$

	H	FP	Kde	BagHist	BagFP	BagKde	Rash
Normal	1.1447	0.6171	0.3181	10.3243	0.2487	1.5195	0.6025
Chi2	0.0696	0.0374	0.0254	0.5635	0.0181	0.0919	0.0334
Mixture	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	17.2112	11.4841	6.8795	175.0100	6.9725	21.2978	8.5623
Tsybakov	6.2811	6.0856	6.3647	16.3470	5.8546	5.7913	5.9907
Uniform Mixt.	4.9186	3.0408	1.8700	22.3239	2.0799	2.5847	2.3357

- $n = 100$

	H	FP	Kde	BagHist	BagFP	BagKde	Rash
Normal	0.7098	0.4016	0.2665	4.9699	0.1665	0.9510	0.3908
Chi2	0.0422	0.0215	0.0144	0.2689	0.0089	0.0549	0.0198
Mixture	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	7.9732	5.2470	5.2962	87.8440	3.8460	14.8139	4.1718
Tsybakov	5.8523	5.5056	6.0074	8.9022	5.9432	4.5841	5.2529
Uniform Mixt.	2.4888	1.6908	1.4754	11.2250	2.1891	1.8702	1.1925

- $n = 200$

	H	FP	Kde	BagHist	BagFP	BagKde	Rash
Normal	0.4077	0.2059	0.1394	2.4724	0.0827	0.5840	0.1937
Chi2	0.0242	0.0121	0.0084	0.1246	0.0075	0.0305	0.0111
Mixture	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	3.7493	2.5354	3.9036	44.4120	2.7891	10.1004	2.1922
Tsybakov	4.2920	3.2531	5.7958	4.3796	5.9281	3.5878	2.6659
Uniform Mixt.	1.6959	1.2517	1.0518	5.5502	2.3213	1.2049	0.8032

- $n = 500$

	H	FP	Kde	BagHist	BagFP	BagKde	Rash
Normal	0.2254	0.1236	0.0663	0.8958	0.0500	0.2977	0.1203
Chi2	0.0116	0.0057	0.0038	0.0470	0.0031	0.0151	0.0056
Mixture	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	1.5125	0.9942	2.4329	18.6821	1.6281	5.5846	1.1093
Tsybakov	2.9831	2.2515	3.3141	1.9641	5.6468	2.0181	1.7904
Uniform Mixt.	1.2756	1.0160	0.7029	2.3993	1.1044	0.7520	0.6518

- $n = 1000$

	H	FP	Kde	BagHist	BagFP	BagKde	Rash
Normal	0.1220	0.0599	0.0360	0.4380	0.0249	0.1804	0.0591
Chi2	0.0068	0.0031	0.0020	0.0218	0.0016	0.0088	0.0031
Mixture	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	0.7052	0.4630	1.7129	9.6299	1.2285	3.5032	0.7965
Tsybakov	2.3140	1.8291	1.5237	1.2728	5.4538	1.5519	1.4157
Uniform Mixt.	1.2970	1.0580	0.4948	1.1314	0.6291	0.5210	0.6376

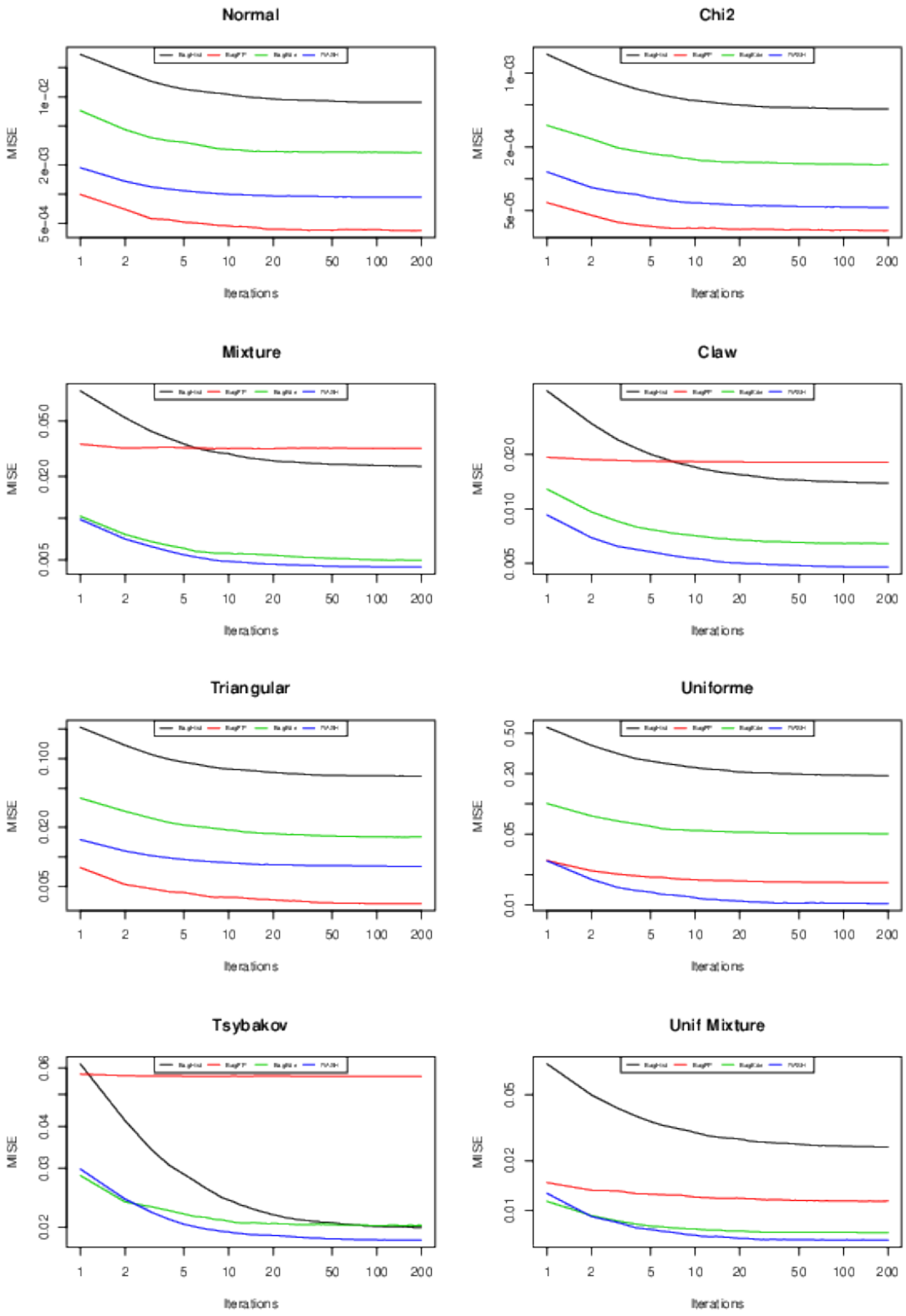
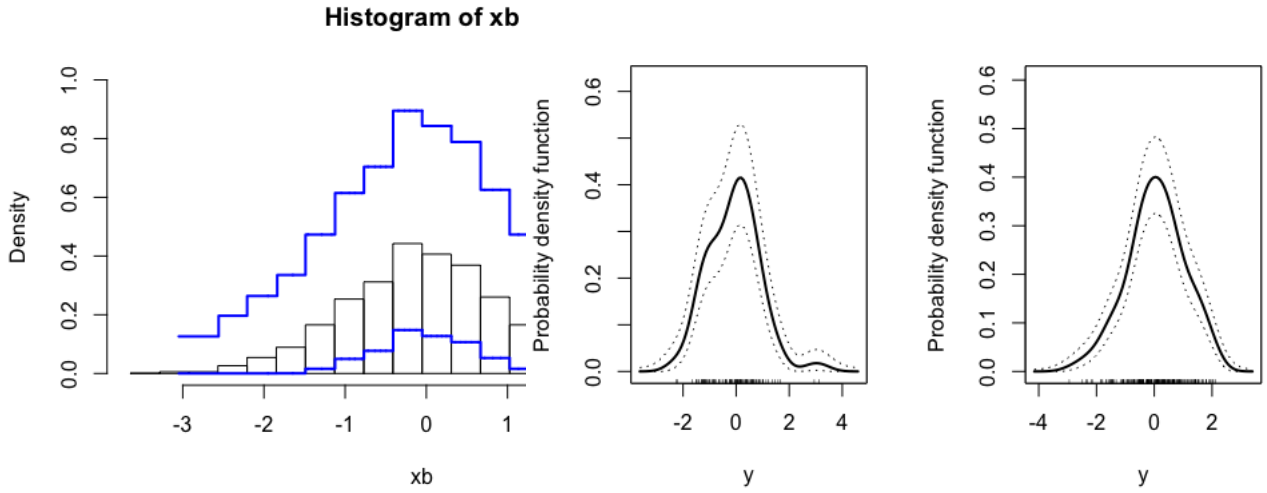


Figure 4: MISE error vs number of aggregates, n=500, M=100, B=200 in log-log scale


Histogram of xb
Density
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
xb
Probability density function
0.0
0.2
0.4
0.6
-2
0
2
4
y
Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-4
-2
0
2
y