Final Report

of the Work Done in

UGC Minor Project

F.No.39-938/2010 (SR)

DATE OF IMPLEMENTATION: 01.02.2011 to 31.01.2013

Annexure VIII

UNIVERSITY GRANTS COMMISSION
BAHADUR SHAH ZAFAR MARG
NEW DELHI 110 002

PROFORMA FOR SUBMISSION OF INFORMATION AT THE TIME OF SENDING THE FINAL REPORT OF THE WORK DONE ON THE PROJECT

1. NAME AND ADDRESS OF THE PRINCIPAL INVESTIGATOR: SANTANU DUTTA

2. NAME AND ADDRESS OF THE INSTITUTION: Mathematical Sciences department TEZPUR UNIVERSITY, Napaam 784028, Tezpur Assam.

3. UGC APPROVAL NO. AND DATE : F. No. 39-938/2010 (SR), Date 01.02.2011

4. DATE OF IMPLEMENTATION: 01.02.2011 to 31.01.2013
5. TENURE OF THE PROJECT: 2 years w.e.f. 01.02.2011

6. TOTAL GRANT ALLOCATED: Rs1,20,000/- (Rupees One lakh and Twenty Thousand only)

7. TOTAL GRANT RECEIVED : Rs1,00,000/- (Rupees One Lakh only)

8. FINAL EXPENDITURE: Rs 99,840/- (Rupees Ninety Nine thousand Eight hundred and Forty only)

9. TITLE OF THE PROJECT: Smooth Bootstrap Estimation of Various Measures of Accuracy and Band Width Selection for Kernel Density Estimators.

10. OBJECTIVES OF THE PROJECT: The kernel density estimator is one of the most widely used statistical tools. Practical implementation of this estimator depends crucially on data based choice of a smoothing parameter. Appropriate amount of smoothing depends on the data, the type of problem and the measures of accuracy. We consider three types of problems in the context of density estimation, viz. global estimation, local estimation of the density function and also interval estimation of the density at a design point. In global estimation we construct a function based on the data which can serve as a proxy for the unknown density, and the accuracy of such an estimator is usually measured by mean integrated squared error (MISE). The L_1 distance between the kernel estimator and the actual density is also another important measure of accuracy. In local estimation, the goal is to estimate the value of a continuous density at a specific point, and the accuracy of such an estimator is measured by mean squared error (MSE). In interval estimation, the aim is to construct a confidence interval for the unknown value of the density at a point with a specific coverage probability. The accuracy of such a confidence interval is measured in terms of coverage error . MISE, L_1 distance, MSE and the coverage error are unknown. Reliable estimator of these measures are essential in the context of bandwidth selection for local, global or interval estimation of the unknown density. In this project we aim to

1. We propose new bootstrap based estimator of MISE of a kernel density estimator.

2. Propose new bandwidth selection algorithm for global estimation, minimizing the estimator of MISE.

3. We propose new bootstrap based estimator of MSE of a kernel density estimator.

4. Propose new bandwidth selection algorithm for local estimation of the value of a continuous density at a given point.

5. Propose an algorithm for bandwidth selection for kernel based interval estimation of the value of a continuous density at a given point.

6. Estimate the L_1 distance between the kernel density estimator and the density.

7. Propose an algorithm for choice of smoothing parameter by minimizing the estimated L_1 distance.

8. Obtain asymptotic properties the resulting estimators.

9. Use extensive simulations and analysis of real data to get insight into the finite sample performance of the proposed bandwidth selectors.

11. WHETHER OBJECTIVES WERE ACHIEVED : Yes

(GIVE DETAILS) The objectives of the project were achieved in terms of five articles, three of which are published in peer reviewed SCI indexed journals and two are unpublished monographs, The summary of the five articles are as follows.

Summary of the first article: Smoothing methods for kernel density estimators struggle when the shape of the reference density differs markedly from the actual density. The smooth bootstrap bandwidth selector minimizes an estimator of the Mean Integrated Squared Error (MISE) of a kernel density estimator. The smooth bootstrap based MISE estimator depends crucially on a pilot bandwidth λ. The earlier bandwidth selectors used some reference distribution to estimate the unknown constants in the pilot bandwidth λ used in the MISE estimator. When the shape of the density generating the data and the reference density differ widely, the resulting estimates perform poorly. We propose a new smooth bootstrap method where the choice of λ does not involve any pilot estimate, and no reference distribution is used at any stage. The proposed bootstrap performs reliably in difficult cases and asymptotically outperforms well known automatic bandwidths.

Summary of the Second article: We consider the problem of data-based choice of the bandwidth of a kernel density estimator, with an aim to estimate the density optimally at a given design point. The existing local bandwidth selectors seem to be quite sensitive to the underlying density and location of the design point. For instance, some bandwidth selectors perform poorly while estimating a density, with bounded support, at the median. Others struggle to estimate a density in the tail region or at the trough between the two modes of a multimodal density. We propose a scale invariant bandwidth selection method such that the resulting density estimator performs reliably irrespective of the density or the design point. We choose bandwidth by minimizing a bootstrap estimate of the mean squared error (MSE) of a density estimator. Our bootstrap MSE estimator is different in the sense that we estimate

the variance and squared bias components separately. We provide insight into the asymptotic accuracy of the proposed density estimator

Summary of the Third article: In the third paper we address the problem of point-wise and uniform convergence of kernel density estimators using random bandwidths, viz. plug-in, smooth bootstrap or cross validation bandwidths. Most of the known asymptotic properties of a kernel density estimator were obtained assuming that the bandwidth sequence is a non-random positive sequence. However, practical applications of kernel density estimation depend crucially on efficient algorithms for data-based choice of the bandwidth. See Park and Marron (1990), Cao et al. (1994), Bose and Dutta (2013) for a review and comparison of different data-based bandwidth selection algorithms. Such bandwidths are functions of the data, and it is hard to compute the bias of a kernel density estimator using such random bandwidths. We obtain the rates of point-wise and uniform convergence of kernel density estimators using random bandwidths under i.i.d. as well as strongly mixing dependence assumptions. Point-wise rates are faster and not affected by the tail of the density.

Summary of the Fourth article: In the next article, we consider the problem of construction of confidence interval for $f(x_0)$, where f is the unknown density generating the given data and x_0 is a given design point. A density function may be arbitrarily specified at a point x_0. This technical difficulty is overcome by assuming that f is continuous. We propose a bandwidth selection method for kernel based interval estimation of a density at a design point, with an aim to minimize the coverage error. The bandwidth is chosen by minimizing a bootstrap estimate of the coverage error. The proposed algorithm seems to be the first bandwidth selector for kernel based interval estimation of a density.

Summary of the Fifth article: Finally in the last article, we present a new method for automatic selection of the bandwidth matrix for a multivariate kernel density estimate, under weak conditions. The existing multivariate methods for data based choice of a bandwidth matrix aim to minimize some L_2 measure of accuracy, and impose a number of assumptions on the underlying density and its derivatives. In contrast we suggest to choose the bandwidth matrix with an aim to minimize a suitable L_1 distance, and we impose no conditions on the density function at all. The asymptotic result obtained in the paper provides insight into accuracy of the density estimator, using our automatic bandwidth. Simulations and analysis of real data confirm that this new method is not merely of academic interest, but compares well with the existing sophisticated bandwidth selectors, such as the plug-in method based on 2 stage of pilot estimation (Duong and Hazelton (2003)).

12. ACHIEVEMENTS FROM THE PROJECT : In this project we have addresses several important problems in the context of nonparametric density estimation. In global density estimation the aim is to estimate the unknown density function. In local estimation the goal is to estimate the value of a continuous density at a given design point. These are different problems. In global estimation the parameter space is the class of all densities on the real line. In local estimation the parameter space is $(0,\infty)$. In interval estimation of the value of a density at a given point, the aim is to construct a confidence interval with coverage probability close to a desired level, minimizing the coverage error. In multivariate density estimation the goal is to estimate the unknown joint density based on vector valued data.

The five monographs prepared during this project are substantial contributions to the theory of kernel density estimation, in the context of local, global estimation of a density and intervals estimation of a density at given point. Algorithms for data based smoothing based on resampling have been developed, and asymptotic properties of the resulting estimators have been obtained. Numerical simulations provide detailed insight into the performance of the proposed algorithms.

13. SUMMARY OF THE FINDINGS. Following are the findings-

1. A smooth bootstrap based estimator of the mean integrated squared error (MISE) of a kernel density estimator is proposed. The estimator has closed form expression, and re-sampling is required to implement the MISE estimator. (See Bose and Dutta (2013))

2. Asymptotic properties of the bootstrap MISE estimator are obtained (See Bose and Dutta (2013))

3. A bandwidth selector is proposed which minimizes the bootstrap MISE Estimator. In this bandwidth elector no reference distribution is used and resulting density estimator performs reliably in difficult cases and asymptotically outperforms well known automatic bandwidths. (See Bose and Dutta (2013))

4. A smooth bootstrap based estimator of the mean squared error (MSE) of a kernel density estimator is proposed. Our bootstrap MSE estimator is different in the sense that we estimate the variance and squared bias components separately. (see Dutta (2014))

5. A bandwidth selector is proposed which minimizes the proposed bootstrap MSE estimator. The proposed bandwidth selector performs reliably and can be recommended safely, especially when not much prior information on f is available (see Concluding remarks in Dutta 2014).

6. Asymptotic properties of the resulting local density estimator is obtained. Using extensive simulations the (see Dutta 2014).

7. A bandwidth selection method for kernel based interval estimation of a density at a design point, with an aim to minimize the coverage error. The bandwidth is chosen by minimizing a bootstrap estimate of the coverage error. The proposed algorithm seems to be the first bandwidth selector for kernel based interval estimation of a density.

8. An estimator of the L_1 distance between the kernel density estimator and the unknown density is proposed. A new algorithm to select the bandwidth matrix of a multivariate KDE is proposed. Asymptotic property of the resulting density estimator obtained.

14. CONTRIBUTION TO THE SOCIETY. (GIVE DETAILS)

Kernel density estimators are one of the most widely used tools for data analysis by practitioners. However application of kernel density estimators depend crucially on the choice of a smoothing parameter, or a matrix of such parameters for multivariate data. The amount of smoothing not only depends on the data, but also on the context of the problems. Global estimation of the density function, local estimation of the value of the density at a given point or interval estimation of a density at a given point are different problems. In this

project new algorithms for data choice of the smoothing parameter have been developed for local and global estimation of the unknown density and also for construction of confidence interval for the unknown density. A new algorithm for data based choice of the bandwidth matrix for multivariate density estimator has also been developed. These algorithms are expected to be of great help for practitioners.

15. WHETHER ANY PH.D. ENROLLED/PRODUCED : None
OUT OF THE PROJECT

16. NO. OF PUBLICATIONS OUT OF THE PROJECT: 03
(PLEASE ATTACH RE-PRINTS)Enclosed herewith

Santanu Dutta

(PRINCIPAL INVESTIGATOR)

(REGISTRAR/PRINCIPAL)

तेजपुर विश्वविद्यालय
Registrar
Tezpur University

Statistics and Probability Letters

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

Density estimation using bootstrap bandwidth selector

Arup Bose [a], Santanu Dutta [b,*]

[a] Statistics and Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India

[b] Mathematical Sciences Department, Tezpur University Napaam, 784028, Tezpur, Assam, India

ARTICLE INFO

Article history:
Received 14 May 2012
Received in revised form 29 August 2012
Accepted 30 August 2012
Available online 17 September 2012

MSC:
62G07
62G09
62G20

Keywords:
Kernel density estimator
Bootstrap
Plug-in
Cross-validation
Automatic bandwidth

ABSTRACT

Smoothing methods for density estimators struggle when the shape of the reference density differs markedly from the actual density. We propose a bootstrap bandwidth selector where no reference distribution is used. It performs reliably in difficult cases and asymptotically outperforms well known automatic bandwidths.

1. Introduction

Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with an unknown density $f(\cdot)$. The *kernel density estimator* (KDE) of f, based on the kernel $K(\cdot)$ and bandwidth $h \equiv h_n$, is defined as

$$K_n(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - X_i}{h}\right) \tag{1.1}$$

where $h \to 0$ and $nh \to \infty$ as $n \to \infty$. The *mean integrated squared error (MISE)* of $K_n(\cdot)$ is a global measure of accuracy of $K_n(\cdot)$. It has enjoyed great popularity, especially in the context of optimal bandwidth selection of a KDE. See for instance, Taylor (1989), Faraway and Jhun (1990) and Hall et al. (1992). In this article we consider the problem of *bandwidth selection* with a view to achieve the minimum possible value of the MISE (call it M).

Bandwidth selection procedures with this goal in mind have been widely studied over the past decade and several procedures to choose this bandwidth have been proposed in the literature. In particular, the Sheather and Jones (1991) plug-in bandwidth (say h_{SJPI}) and the smooth bootstrap bandwidth proposed by Cao et al. (1994) (say h_{Cao}), have been suggested as new standard methods. See Cao et al. (1994) and Jones et al. (1996) for a detailed comparison of a number of automatic bandwidths. The latter have suggested that bandwidths such as h_{SJPI} be considered as the benchmark of good performance. However, Loader (1999) observed that h_{SJPI} often over-smooths and misses important features when given difficult problems. As we shall see later this criticism is also relevant for h_{Cao}.

* Corresponding author.
E-mail addresses: bosearu@gmail.com (A. Bose), tezpur1976@gmail.com (S. Dutta).

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.08.027

A common feature in these bandwidth selectors is that any unknown functional $T(f)$ is approximated by $T(f_\alpha)$, where f_α is another KDE using the same kernel K and a "pilot bandwidth" λ. Loader (1999) pointed out that these bandwidth selectors are heavily dependent on the specification of λ. For instance in the smooth bootstrap method of Cao et al. (1994), λ is chosen with an aim to estimate $\int [f^{(2)}(x)]^2 dx$ accurately. In Jones et al. (1991), λ is selected with a view to minimize asymptotic (relative) MSE for the selected bandwidth. In all these methods, the best choice of λ depends on some functional of the density or its derivatives. For instance, Cao (1993) and Cao et al. (1994) have proposed the choice $\lambda = \frac{C}{n^{\frac{1}{7}}}$ where C depends on $\int [f^{(3)}(y)]^2 dy$. The unknown constants in λ are usually estimated by approximating the underlying density using a reference distribution. If this reference distribution is far removed from f, the smooth bootstrap bandwidths struggle. For instance, Jones et al. (1991, p. 1925) have observed that for densities which are somewhat far from the Gaussian in terms of shape, the performance of their bootstrap bandwidth selector is not so good.

The plug-in bandwidth selectors, such as h_{SJPI}, also exhibit this demerit. In this method, the optimal choice of h is expressed as a function of $\int [f^{(2)}(x)]^2 dx$ (see Loader, 1999), which is approximated using $\int [f_\eta^{(2)}(x)]^2 dx$. By varying λ, a wide range of "optimal" values of h can be selected. The plot of h against a broad range of values of λ is referred to as the "actual" relation between λ and h. To choose an appropriate value of λ, a common approach is to "assume" a relation between λ and h. Plug-in methods differ with respect to the choice of this relation (see for example, Sheather and Jones, 1991). The Sheather and Jones method uses a complicated "assumed" relation, based on estimating the density derivatives using a reference normal distribution. As a consequence, if f is substantially different from a normal distribution in shape, h_{SJPI} suffers.

The above mentioned bandwidth selectors use some reference distribution to estimate the unknown constants in λ. When the shape of f and the reference density differ widely, the resulting estimates perform poorly. We propose a new smooth bootstrap method where the choice of λ does not involve any pilot estimate, and no reference distribution is used at any stage. A smooth bootstrap bandwidth $\hat{h}$ equals

$$\hat{h} = \text{minimizer of } M^*(h), \quad h \in I,$$

where I is a compact interval and $M^* \equiv M^*(h)$ is a smooth bootstrap estimator of M. It is defined using (another) KDE K_n^0 with kernel K^0 and bandwidth λ. See (3.2) for the definition of M^*.

From (A.7) in the Appendix it is easy to see that for $n\lambda \to \infty$ and $h \in I$,

$$E|M^*(h)/M(h) - 1| = O\left(\frac{1}{n^{1/(2s+1)}}\sqrt{\frac{1}{n\lambda} + \lambda^{2p}} + \sqrt{\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy}\right).$$

Hence the asymptotic accuracy of M^* depends on the accuracy of $K_n^{0(s)}$ in estimating $f^{(s)}$. Our choice of λ is motivated by the following inequality, established in Lemma 1 in the Appendix. Here p, C_1, C_2 are constants which do not depend on f, but depend on the kernel K^0 and the order s of the original kernel K.

$$\int E[K_n^{0(s)}(y) - f^{(s)}(y)]^2 dy \le \frac{C_1}{n\lambda^{1+2s}} + C_2\lambda^{2p}\int [f^{(s+p)}(y)]^2 dy.$$

The minimizer of the right side of the above inequality equals

$$\lambda = \frac{C_3}{\left[\int [f^{(s+p)}(y)]^2 dy\right]^{1/(2s+2p+1)}} n^{-1/(2s+2p+1)},$$

where C_3 is a constant which depends on K and K^0. The coefficient $C_3/[\int [f^{(s+p)}(y)]^2]^{1/(2s+2p+1)}$ varies widely depending on the choice of f. We observe that within a class of mixed normal densities, this coefficient varies approximately from $\frac{1}{9}$ to 1.3 depending on the choice of f. Through extensive simulations we find that

$$\lambda = \frac{1}{8} n^{-1/(2s+2p+1)}, \quad \text{where } s, p \ge 2,$$

works very well. With this choice of λ, let $\hat{h}^*$ be the bandwidth minimizing M^* in I. This is our recommended bootstrap bandwidth and it works well in capturing important features of a wide variety of densities. In particular, for a second order kernel K, $p = s = 2$.

In Section 2 we report a detailed simulation study and analysis of a real data set. Simulations demonstrate that for a second order kernel, our bootstrap bandwidth can perform much better than h_{SJPI} and h_{Cao} bandwidths in a number of difficult problems – especially when f exhibits a number of peaks and sample size is moderate. In Theorem 1 of Section 3, we obtain the L_1 rate at which $\hat{h}^*$ succeeds in minimizing the M as sample size is increased. Its proof is given in the Appendix.

2. Simulation and data analysis

In Cao et al. (1994) the performance of a number of automatic bandwidths, including the versions of bootstrap bandwidth proposed by Faraway and Jhun (1990), Jones et al. (1991) and Cao (1993), were compared via simulations. They found that h_{SJPI} and h_{Cao} outperformed other automatic bandwidths. Loader (1999) showed that in some difficult examples, where f possesses several peaks and n is moderate, the unbiased cross-validation bandwidth selector reveals the actual features of f much more effectively than the estimate using h_{SJPI} based on moderate sample size. So we compare the performance of $\hat{h}^*$ with these two data based bandwidths and also with the unbiased cross-validation bandwidth (call it h_{UCV}).

Loader (1999) has mentioned two important criteria for assessing the performance of any automatic bandwidth $\tilde{h}$:

(a) If a bandwidth selector is to be useful, it must perform reliably in difficult cases.

(b) How well does the density estimate, using $\tilde{h}$, approximate the true f? This can be measured by the MISE of the automatic density estimate.

We compare the performance of $\hat{h}^*$ with h_{SJPI}, h_{Cao} and h_{UCV} using the above guideline. Let $N(\mu, \sigma^2)$ denote normal distribution with mean μ and variance σ^2. Let $\phi_{\sigma^2}(\cdot)$ denote the density of $N(0, \sigma^2)$ distribution. We draw samples of size 50 and 500 from five test densities, namely (a) the claw density, $\frac{1}{2}N(0, 1) + \frac{1}{10}\sum_{i=0}^{4} N(i/2 - 1, 0.01)$, (b) an equal mixture of ten normal distributions, i.e. $\frac{1}{10}\sum_{i=1}^{10} N(10i - 5, 1)$, (c) the asymmetric bimodal normal mixture density $0.75N(0, 1) + 0.25N(3/2, 1/9)$, (d) the outlier density i.e. $\frac{1}{10}N(0, 1) + \frac{9}{10}N(0, 0.01)$ and (e) the standard Cauchy density. These test densities are drastically different from a normal density in shape. Throughout, we use

$$\lambda = \frac{1}{8n^{1/(2s+2p+1)}} \quad \text{where } s = p = 2$$

and K, K^0 as the standard normal densities. For these kernels, M^* has a simple expression

$$\frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\left\{\left(1 - \frac{1}{n}\right)\phi_{2\lambda^2+2h^2}(X_i - X_l) - 2\phi_{2\lambda^2+h^2}(X_i - X_l) + \phi_{2\lambda^2}(X_i - X_l)\right\}.$$

(a) To compare the bandwidths according to our first criterion, consider the two examples in Loader (1999) where the goal was to capture the important features of the underlying density based on samples of size $n = 193$ and $n = 100$, from the claw density and the mixture of ten normal densities. Loader (1999, p. 423) observed that, while under a theoretical MISE criterion, the five peaks of a claw density should be detectable for $n = 193$ in practice an estimate using h_{SJPI} fails to capture the peaks and over-smooths. Similarly while the ten-modal structure of the underlying density is quite obvious in the data from the mixture of ten normal densities, the h_{SJPI} clearly over-smooths.

In Figs. 1 and 2 we plot the four estimates, using $\hat{h}^*$, h_{SJPI}, h_{Cao} and h_{UCV}, based on samples of size $n = 100$ and $n = 193$ from the mixture of ten normal densities and the claw density. The estimate, using $\hat{h}^*$, captures the peaks of the underlying test densities while h_{SJPI} and h_{Cao} over-smooth and miss important features. The performance of the cross-validation density estimate varies depending on f. While for the claw density h_{UCV} clearly helps to captures the peaks, for the mixture of ten normal densities the estimate over-smooths.

We repeat this experiment 100 times, and observe the same pattern – h_{SJPI}, h_{Cao} consistently over-smooth and the h_{UCV} based estimate fluctuates widely. In contrast for most of the samples, $\frac{\hat{h}^*}{h^*}$ is close to 1, where h^* is the bandwidth minimizing the exact MISE. So $\hat{h}^*$ performs reliably.

We also consider the problem of estimating the standard Cauchy density based on a sample of size 500. This is a difficult problem since it is known that fixed bandwidth estimates are generally inadequate for estimating heavy tailed densities (see Loader, 1999, p. 435). The estimates using $\hat{h}^*$, h_{SJPI}, h_{Cao} and h_{UCV} are displayed in Fig. 3. We find that estimate using $\hat{h}^*$ is closest (in terms of integrated squared error (ISE)) to the standard Cauchy density compared to the other estimates. h_{SJPI} clearly under-smooths and the estimates using h_{Cao} and h_{UCV} are much flatter than the Cauchy density.

For a replicated study, we generate 100 samples, each of size 500, from the standard Cauchy density. h_{SJPI} and h_{UCV} are highly variable (their ISE values vary widely). In contrast, h_{Cao} is least variable but consistently over-smooths. Our $\hat{h}^*$ exhibits much lower sampling fluctuation, and for most of the samples the ISE of the estimate using $\hat{h}^*$ is lower than the same for the other estimates.

In Figs. 4–7, we plot the four estimates based on samples of size $n = 50$ and 500, from each one of the four test densities (a)–(d). In each figure, estimate 1 uses $\hat{h}^*$, and estimates 2, 3 and 4 use h_{SJPI}, h_{Cao} and h_{UCV} respectively.

(b) To compare the bandwidths according to our second criterion we approximate the exact MISE of the automatic density estimates, for different sample sizes and test densities. We draw 100 samples, of size $n = 50$, 500 from the five test densities (a)–(e) and we compute the integrated squared error (ISE) of the estimate. The average ISE (we call it AISE) is a Monte Carlo estimate of the MISE of the automatic estimate. These values are provided in Table 1. We also report the minimum of the 100 ISEs, which represents the "best estimate" by a particular bandwidth selector.

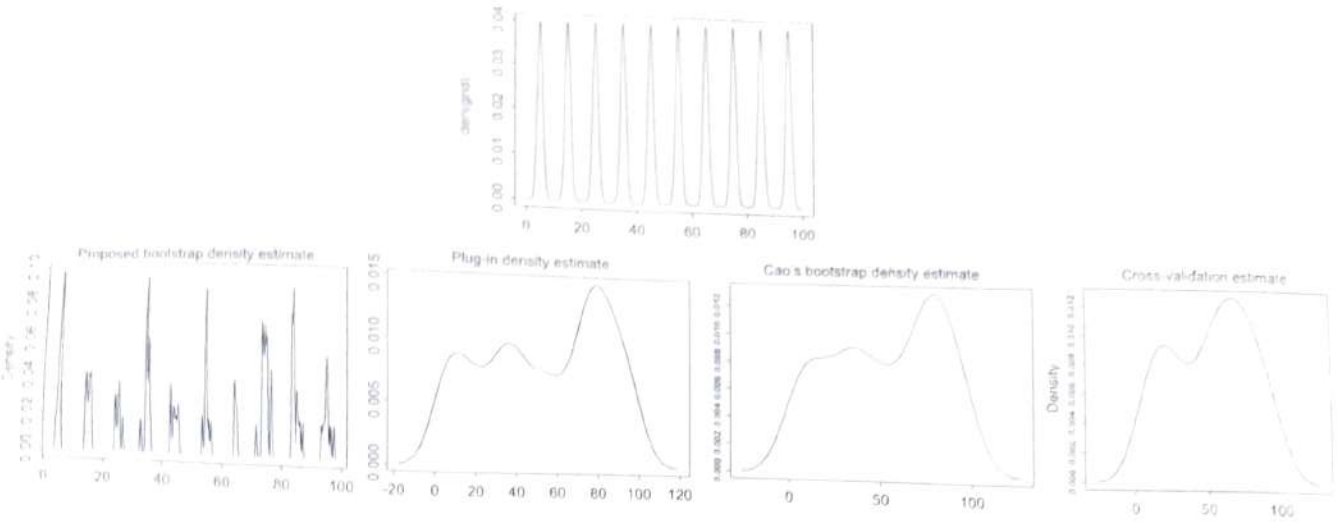


Fig. 1. The exact density and the four density estimates of mixture of ten normal densities $n = 100$.

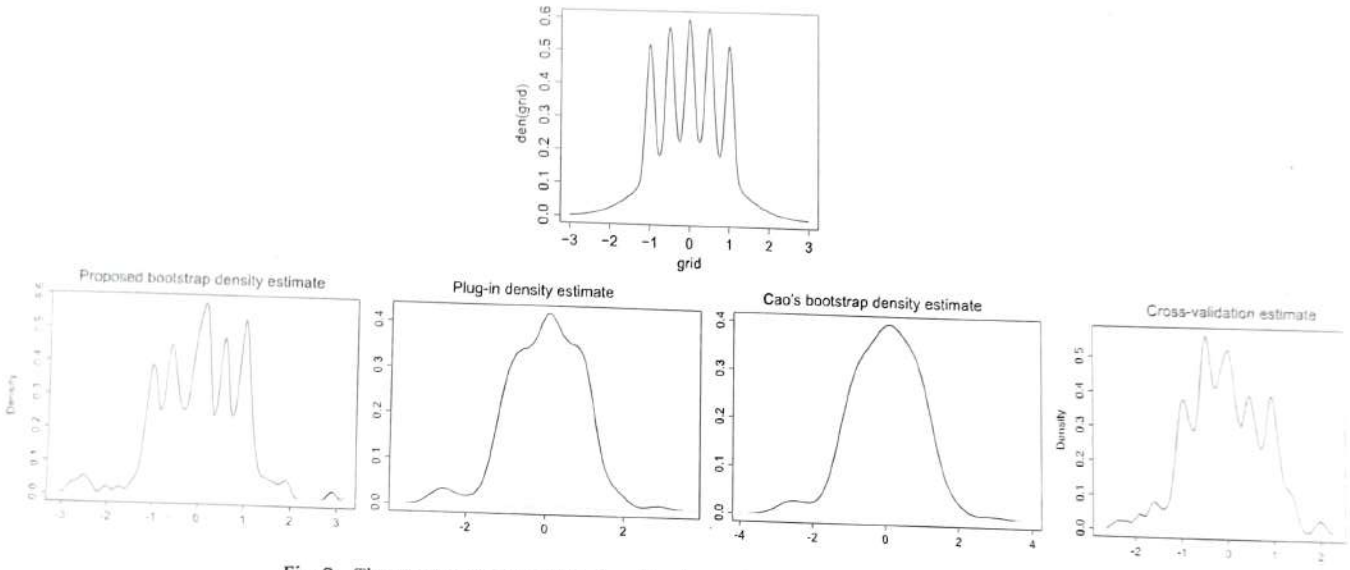


Fig. 2. The exact density and the four density estimates of the claw density $n = 193$.

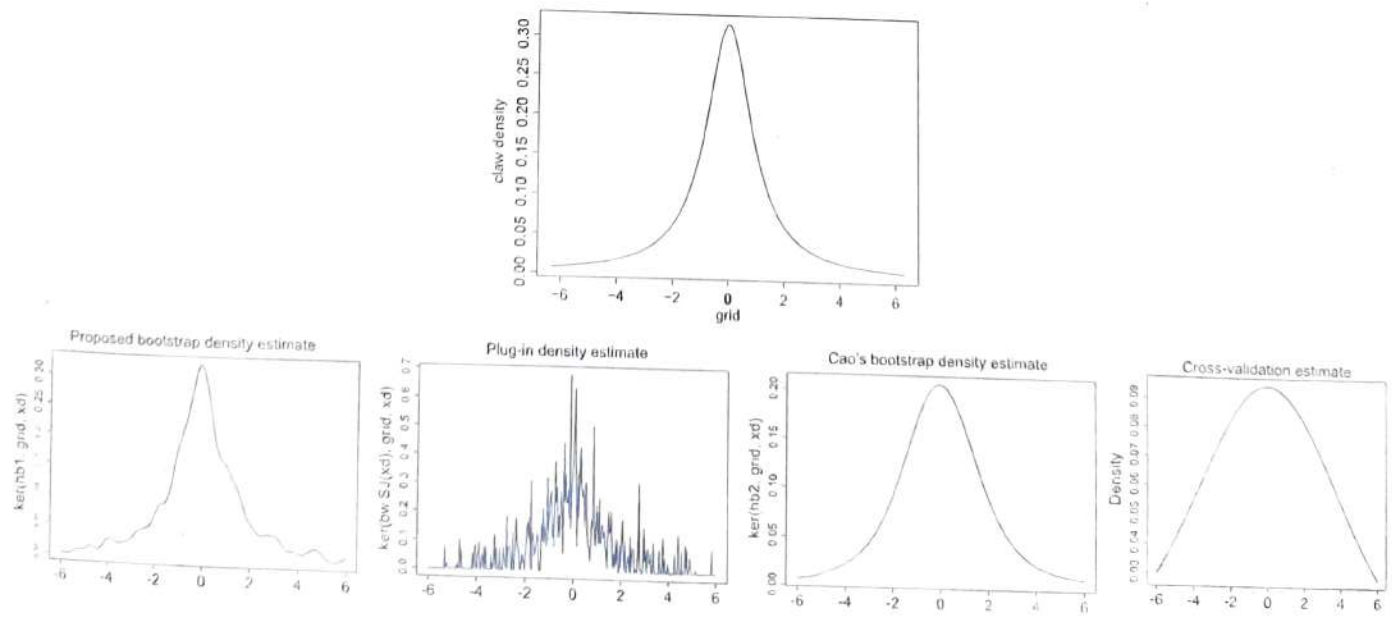


Fig. 3. The exact density and the four density estimates of the standard Cauchy density $n = 500$.

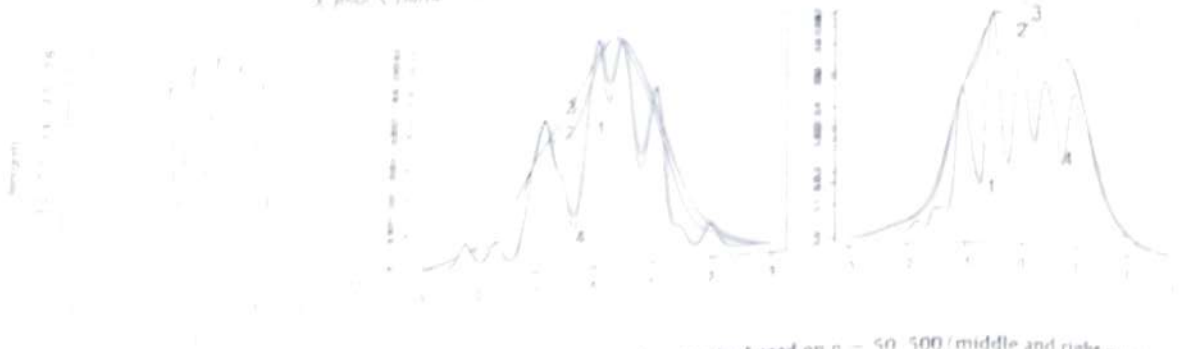


Fig. 4. The exact density (left panel) and four density estimates of the claw density based on $n = 50, 500$ (middle and right panel).

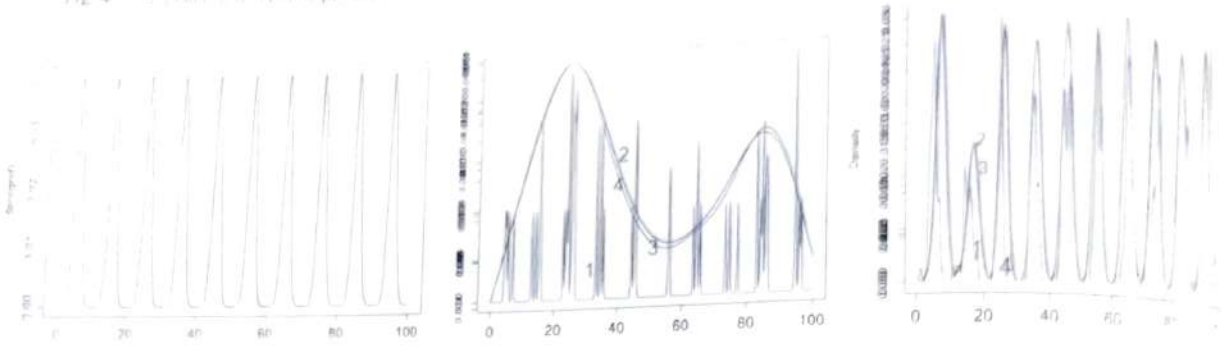


Fig. 5. The exact density (left panel) and four density estimates of mixture of ten normal densities based on $n = 50, 500$ (middle and right panel).

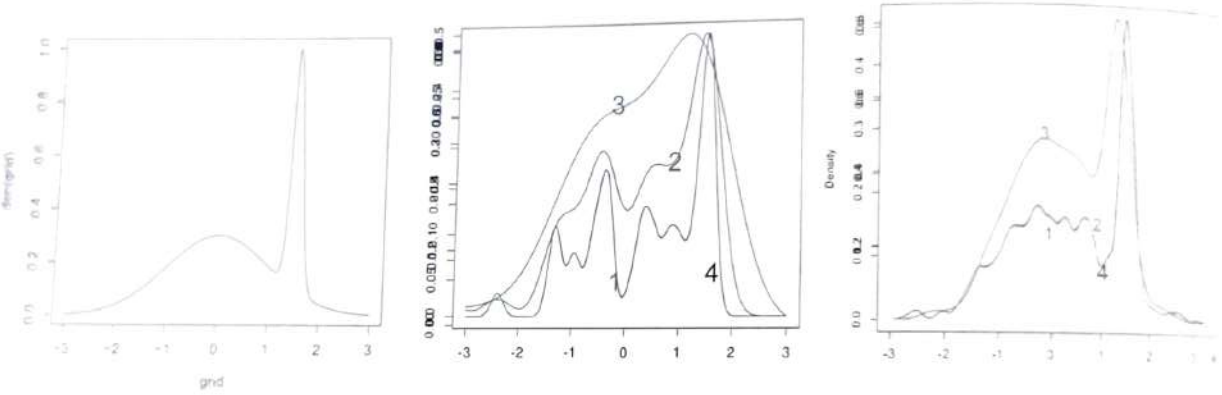


Fig. 6. The exact density (left panel) and four density estimates of $0.75N(0, 1) + 0.25N(3/2, 1/9)$ based on $n = 50, 500$ (middle and right panel).

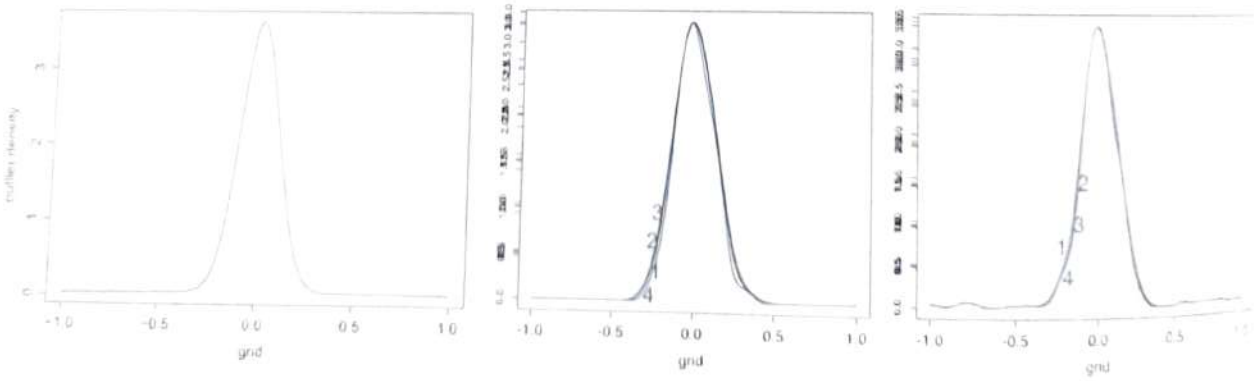


Fig. 7. The exact density (left panel) and four density estimates of outlier density based on $n = 50, 500$ (middle and right panel).

The main observations are as follows.

1. In Figs. 4 and 6, for claw and asymmetric bimodal normal mixture densities, Estimate 3, using h_{Cao}, is over-smoothed even for sample size $n = 500$. In Figs. 3 and 4, the estimate using h_{SJPI} bandwidth completely fails to capture the main features of f when it is the standard Cauchy or the claw density, even for $n = 500$.

2. From Figs. 1–7, the estimate with bandwidth $\hat{h}^*$ captures the main features of f, even when f has a complicated structure, across different sample sizes. Incidentally, the estimation of the claw density for a sample of size as small as

Table 1

AISE and minimum ISE of the density estimates using h_{SJPI}, $\hat{h}^*$, h_{Cao} and h_{UCV}

n	Distributions					
	$\hat{h}$	Outlier	Asymmetric bimodal	Ten normal mix	Claw	Standard Cauchy
50	h_{SJPI}	0.139 (0.049)	0.061 (0.034)	0.019 (0.018)	0.060 (0.048)	0.017 (0.003)
	$\hat{h}^*$	**0.098 (0.018)**	**0.072 (0.032)**	**0.016 (0.009)**	**0.057 (0.047)**	**0.028 (0.006)**
	h_{Cao}	0.212 (0.035)	0.111 (0.091)	0.007 (0.004)	0.056 (0.048)	0.015 (0.005)
	h_{UCV}	0.142 (0.017)	0.069 (0.027)	0.0167(0.008)	0.064 (0.012)	0.025 (0.003)
500	h_{SJPI}	0.02 (0.006)	0.014 (0.006)	0.003 (0.002)	0.110 (0.038)	0.022 (0.001)
	$\hat{h}^*$	**0.02 (0.007)**	**0.014 (0.006)**	**0.002 (0.001)**	**0.017 (0.007)**	**0.003 (0.001)**
	h_{Cao}	0.03 (0.006)	0.056 (0.045)	0.003 (0.002)	0.044 (0.041)	0.011 (0.009)
	h_{UCV}	0.026 (0.005)	0.012 (0.004)	0.0025 (0.002)	0.011 (0.004)	0.021(0.001)

$n = 50$ is a very difficult problem. Even under a theoretical MISE criterion, the claws show up only at sample size exceeding 53 (see p. 726 of Marron and Wand, 1992). Therefore it is not surprising that the estimates using h_{SJPI} and h_{Cao} completely miss the claws (see Fig. 4, middle panel) for $n = 50$. However it is encouraging to note that even for this small sample size, Estimate 1, using $\hat{h}^*$, captures four out of the five peaks at the expense of some spurious wiggles near the tail. For mixed normal densities, Estimates 1 and 4, using $\hat{h}^*$ and h_{UCV}, are almost indistinguishable, especially for large n (see Figs. 4–7).

We may conclude that $\hat{h}^*$ performs reliably in difficult cases while the performance of h_{Cao} and h_{SJPI} vary from one test density to another. Moreover, the performance of the estimate using $\hat{h}^*$ improves drastically as n increases, irrespective of the shape of f. This is not true for h_{Cao} and h_{SJPI}.

3. From Table 1 we see that the AISE, using $\hat{h}^*$, decreases much faster than the AISE using h_{Cao} and h_{SJPI}, for all five test densities. For f equal to the claw density and standard Cauchy, the AISE using h_{SJPI} does not seem to decrease at all even when n is increased from 50 to 500. In Fig. 4, the h_{SJPI} density estimate (numbered 2) is over-smoothed and completely misses the peaks of the claw density, even for $n = 500$. This is reflected in the AISE values. The same is true for h_{Cao} as well (see Fig. 4).

4. From Table 1 we see that the AISE values using h_{UCV} and $\hat{h}^*$ are close for the mixed normal densities. In general, the estimate using h_{UCV} seems to perform well for f equal to a mixed normal density, especially for large n. However for f equal to the Cauchy density and $n = 500$, the AISE using h_{UCV} is much larger than the AISE using $\hat{h}^*$. Moreover, for the Cauchy distribution, the AISE using h_{UCV} decreases only marginally (less than 20%) even when n is increased from 100 to 500. As mentioned above, this observation is also true for the estimates using h_{Cao} and h_{SJPI}. In contrast, for the same distribution, the AISE using $\hat{h}^*$ is reduced by nine times as n is increased to 500. This rate of improvement in the estimate using $\hat{h}^*$ is quite remarkable, as a KDE using a fixed bandwidth is generally considered inadequate for estimating heavy tailed densities (see Loader, 1999, p. 435). In general, the density estimate using $\hat{h}^*$ seems to perform reliably irrespective of f.

The comparison of the minimum ISE values for the estimators using $\hat{h}^*$, h_{SJPI} and h_{Cao} lead to similar conclusions as above. For a complicated f, such as the claw density, the accuracy of the estimate, using h_{SJPI} and h_{Cao}, improve very slowly even when the sample size is increased drastically. The minimum ISE values of the density estimates using h_{UCV} and $\hat{h}^*$ are similar for all the test densities and sample size. For f equal to the standard Cauchy density, the h_{UCV} bandwidth exhibits very high sampling fluctuation. Overall $\hat{h}^*$ seems to perform much more reliably, in comparison to the other bandwidth selectors.

2.1. Application to real data

A popular data set is the eruption durations of the Old Faithful geyser. There are several versions of the Old Faithful data (see Loader, 1999) – we consider the one that comes from Azzalini and Bowman (1990) and is also available in the MASS package in the software R, using the command "geyser$duration".

Loader (1999) pointed out that the comparison of the plot of density estimates, using different bandwidth selectors, conveys a one sided view of the bias–variance trade off. High variance can be seen in terms of a wiggly estimate, but there is no way to visualize bias. In simulations one has the advantage of comparing an estimate with the actual density. But for real data, the bias cannot be seen. Therefore while modeling real data sets, the author emphasized using additional criteria, such as comparison of Akaike-style criterion to decide the appropriate bandwidth for a given data set.

He has provided strong evidence that estimates using classical bandwidth selectors, such as the cross-validation bandwidth, can reveal important features present in the data. Therefore, we also consider the unbiased cross-validation bandwidth (h_{UCV}) along with $\hat{h}^*$, h_{SJPI} and h_{Cao}. The corresponding density estimate plots are displayed in Fig. 8. We also calculate the Akaike-style criterion (AIC), defined in Loader (1999), for the four bandwidths. The lower the AIC value, the more appropriate the estimate is.

In Fig. 8, the estimates using h_{UCV} and $\hat{h}^*$ clearly reveal three peaks, located close to 2, 4 and between 4 and 5. The curve using h_{SJPI} also indicates three peaks. In contrast the estimate using h_{Cao} exhibits two peaks, located close to 2 and 4. The

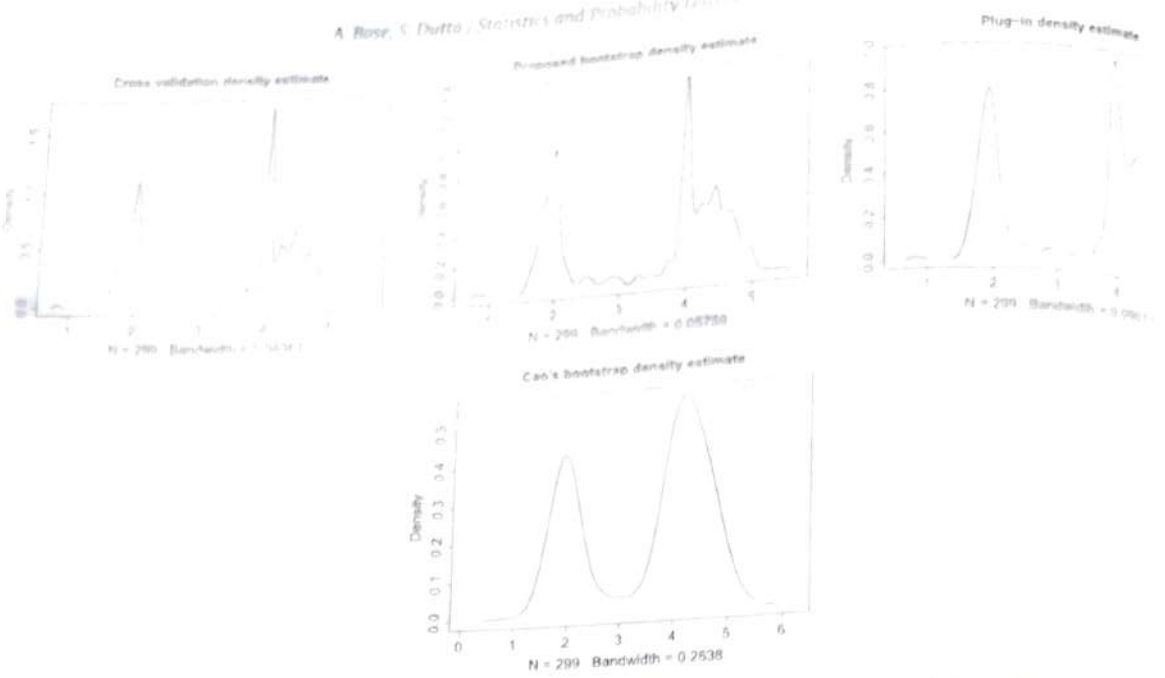


Fig. 8. Density estimates of Old Faithful geyser eruption durations, using h_{UCV}, $\hat{h}^*$, h_{SJPI} and h_{Cao} (left to right) respectively

two peaks, located close to 2 and 4, are much taller in the estimates using h_{UCV} and $\hat{h}^*$, than the same peaks in the other estimates. With real data one cannot be sure. So we compare AIC values for the four bandwidths.

The AIC value is minimum for h_{UCV}, indicating that the estimate using h_{UCV} seems to be more appropriate than the bandwidths for this particular data. The value of AIC is lower for $\hat{h}^*$, than those for h_{SJPI} and h_{Cao}.

The density estimates using h_{UCV} and $\hat{h}^*$ seem to exhibit almost the same features. AIC values suggest that for this Faithful geyser data, features revealed by h_{UCV} and $\hat{h}^*$ are more reliable than those captured by h_{Cao}. There seem to be prominent peaks, and the peak near 4 seems to have the highest density.

3. Asymptotic properties

The MISE of a KDE, using kernel K and bandwidth h, is defined as

$$M = \int_{-\infty}^{\infty} E\big[K_n(y) - f(y)\big]^2 dy = V + B, \quad \text{where}$$

$$V = \frac{1}{nh}\int K^2(v)dv - \frac{1}{n}\int\left\{\int K(v)f(y-hv)dv\right\}^2 dy,$$

$$\text{and} \quad B = \int\left[\int K(u)f(y-hu)du - f(y)\right]^2 dy.$$

A smooth bootstrap MISE estimator, say M^*, of M is then defined as:

$$M^*(h) \equiv M^* = V^* + B^*, \quad \text{where}$$

$$V^* = \frac{1}{nh}\int K^2(u)du - \frac{1}{n}\int\left[\int K(u)K_n^0(y-hu)du\right]^2 dy \quad \text{and}$$

$$B^* = \int\left[\int K(u)K_n^0(y-hu)du - K_n^0(y)\right]^2 dy, \quad \text{where}$$

K^0 is another kernel and K_n^0 is the corresponding KDE using the bandwidth λ. Different choices of K^0 and λ yield different versions of M^*.

In the sequel we assume that M^*, M are minimized with respect to h, for $h \in I = \left[\frac{\epsilon_1}{n^{1/(2s+1)}}, \frac{\epsilon_2}{n^{1/(2s+1)}}\right]$ and $0 < \epsilon_1 < \epsilon_2$

Remark 1. Restricting h to constant multiples of $\frac{1}{n^{1/(2s+1)}}$ is not too demanding (see Assumption 2.3 in Park and Marron 1990). Different versions of $\hat{h}$ minimize different versions of M^*, using different choices of λ and/or K^0. We mention a few such bootstrap bandwidth selectors.

Taylor's (1989) bandwidth selector h_T is a minimizer of M^*, using $\lambda = h$. Faraway and Jhun (1990) proposed to select h by minimizing M^*, where λ is chosen by least-squares cross-validation. Jones et al. (1991) proposed yet another version of h (say h_{JMP}) where K, K^0 have eight bounded continuous derivatives and $\lambda = Cn^p h^m$, where C is a constant having a complicated expression depending on the integrated squared derivatives of f.

Cao's (1993) bandwidth selector (say h_{Cao}) is obtained by minimizing M^* using $K^0 = K$, where K is a second order kernel with six derivatives and λ is independent of h.

Let us collect below all the assumptions that we shall require on the two kernels and the bandwidths. A function H is said to be uniformly bounded if $\|H\| = \sup_{-\infty<y<\infty} |H(y)| < \infty$. Let s, $p \geq 2$ and s, p denote the orders (defined later) of the kernels K and K^0 respectively.

Assumption A (*On Density f*).

(i) The density $f(\cdot)$ is uniformly bounded, and possesses $(s+p)$ continuous derivatives.
(ii) The jth density derivative $f^{(j)}$ is uniformly bounded and square integrable, for $j = s,\ p,\ (s+p)$.

Assumption B (*On Kernel K*). $K(\cdot)$ is the sth order square integrable, symmetric kernel, i.e. $K(-x) = K(x)$, $\int K(x)dx = 1$, $\int K(x)x^j dx = 0$, $j = 1, 2, \ldots, s-1$ and $\int |K(x)x^s|dx < \infty$. Also let $\int |K(x)x^{s+1}|dx < \infty$.

Assumption C (*On Kernel K^0*).

(i) The pilot kernel $K^0(\cdot)$ is an absolutely integrable pth order kernel, i.e. $\int K^0(x)dx = 1$, $\int K^0(x)x^j dx = 0$, $j = 1, 2, \ldots, p-1$, and $\int |K^0(x)x^p|dx < \infty$, $p \geq 2$, such that
  (a) $K^0(\cdot)$ is symmetric, continuous and uniformly bounded.
  (b) $K^0(x) \to 0$ as $|x| \to \infty$.
(ii) $K^0(\cdot)$ has s continuous derivatives on $(-\infty, \infty)$ and its sth derivative $K^{0(s)}(\cdot)$ satisfies the above conditions (a) and (b) and also the following.
  (c) $\int |K^{0(s)}(x)|dx < \infty$.
  (d) $\int K^{0(s)}(x)x^j dx = 0$, where $j = 0, 1, 2, \ldots, s-1, s+1, \ldots, s+p-1$, $\frac{(-1)^s}{s!}\int K^{0(s)}(x)x^s dx = 1$ and $\int |K^{0(s+p)}(x)x^{s+p}|dx < \infty$.

Remark 2. (i) The choice of p in A (ii) depends on K^0. If K^0 is the standard normal density and K is any second order kernel, then Assumption C is satisfied for $s = p = 2$.

(ii) Assumption A(i)–(ii) on f are valid for a wide class of densities which include the mixed normal, the Cauchy, the beta(m, n) $(m, n > 2)$ and the gamma(n) $(n > 2)$. For a second order kernel K, Cao (1993) obtained asymptotic properties of his bootstrap bandwidth selector assuming that f is six times differentiable, the derivatives are bounded and the first four derivatives are integrable. But for a second order kernel and for $p = 2$, we require assumptions only on the first four derivatives of f, i.e. we impose fewer assumptions on f.

(iii) In Remark 3.3, Hall et al. (1992) suggested that the asymptotic accuracy of a smoothed cross-validation or a smoothed bootstrap bandwidth can be improved by using K^0 to be a higher order kernel. As we shall see, this observation is also true for our proposal. $K^0(x) = \frac{(3-x^2)}{2}\phi(x)$, where $\phi(x)$ is the standard normal density, is a fourth order kernel. For K equal to any second order kernel and $K^0(x) = \frac{(3-x^2)}{2}\phi(x)$, the Assumption C is satisfied for $s = 2$ and $p = 4$.

The following result provides a bound on the L_1 accuracy of $\hat{h}^*$.

Theorem 1. *Suppose s, $p \geq 2$, Assumptions A–C hold, $h \to 0$ and $nh \to \infty$, as $n \to \infty$.*
Let $\lambda = \frac{C}{n^{1/(2s+2p+1)}}$ and $h \in I = \left[\frac{\epsilon_1}{n^{1/(2s+1)}}, \frac{\epsilon_2}{n^{1/(2s+1)}}\right]$ where $0 < \epsilon_1 < \epsilon_2$ and C is a positive constant. Then

$$E\left|\frac{M(\hat{h})}{M(h^*)} - 1\right| = O\left(\frac{1}{n^{p/(2s+2p+1)}}\right), \quad \text{where}$$

$M(h^*)$ is the minimum value of M for $h \in I$.

Suppose K is a second order kernel, satisfying Assumption B. If K^0 is the Gaussian kernel, $s = p = 2$ and Assumption C on K^0 is satisfied. So under Assumption A on f, using $\lambda = \frac{1}{8n^{1/9}}$ in Theorem 1,

$$E\left|\frac{M(\hat{h}^*)}{M(h^*)} - 1\right| = O\left(\frac{1}{n^{2/9}}\right).$$

Incidentally, for symmetric second order kernels with finite support (see p. 70 Park and Marron, 1990),

$$n^{1/5}\left(\frac{M(\bar{h})}{M(h^*)}-1\right)\to^{L} 2\sigma^2\chi_1^2,$$

where $\bar{h}$ is the biased cross-validation (BCV) or the unbiased cross-validation (UCV) bandwidth and σ^2 is variance of the asymptotic distribution of $n^{1/10}(\bar{h}/h^*-1)$. A similar result also holds for $\bar{h}$ equal to the plug-in bandwidth by Park and Marron (1990) (they call it h_{PI}). See p. 70 in Park and Marron (1990). On the other hand, Theorem 1 implies that for any second order kernel (i.e. $s=2$), $p=2$ and $\lambda=\frac{1}{8n^{1/9}}$,

$$n^{1/5}\left(\frac{M(\hat{h}^*)}{M(h^*)}-1\right)=o_P(1).$$

Therefore under the above conditions, $\hat{h}^*$ is asymptotically more accurate than the UCV and BCV bandwidths and also the plug-in bandwidth h_{PI}.

Remark 3. (i) For fixed s, the term $\frac{1}{n^{p/(2s+2p+1)}}$ (in the right side of Theorem 1) goes to zero faster as p is increased. So the rate at which $E\left|\frac{M(\hat{h}^*)}{M(h^*)}-1\right|$ goes to zero can be improved further by using K^0 equal to a higher order kernel satisfying Assumption C. For example, if K is a second order kernel and $K^0(x)=\frac{(3-x^2)}{2}\phi(x)$ (so that $s=2$ and $p=4$) the rate at which $E\left|\frac{M(\hat{h}^*)}{M(h^*)}-1\right|=O\left(\frac{1}{n^{4/13}}\right)$.

(ii) For fixed p, $\frac{1}{n^{p/(2s+2p+1)}}\to 0$ at a slower rate as s is increased. So from Theorem 1, it is not advisable to take $s>2$. It is well known that a higher order K can lead to a negative valued density estimate. Moreover, Marron and Wand (19 provide substantial evidence that the use of higher order K does not lead to significant improvement even for very la samples. Simulations in Section 2 confirm that K, $K^0=\phi$ work well.

Some further improvement may be achieved using $K^0(x)=\frac{(3-x^2)}{2}\phi(x)$, especially for large sample size. However, a density estimate using the proposed method, with both K, $K^0=\phi$ and $\lambda=\frac{1}{8n^{1/9}}$, seems to perform reliably in a number difficult examples so these choices of K, K^0 and λ remain our recommendation.

Acknowledgments

We are grateful to Anil Ghosh for his detailed comments and to R. Cao and J.S. Marron for sharing their work. We grateful to Peter Hall for his encouraging comments. Finally, we thank the Referee for her/his encouraging and construct comments.

The first author's research was supported by J.C.Bose Fellowship, Dept. of Science and Technology, Govt. of India. second author's research was supported by UGC minor research project F. No. 39-938/2010 (SR).

Appendix

Let C denote some positive constant independent of n, h or λ, and DCT stands for Dominated Convergence Theor Lemma 1 is used in the proof of Theorem 1 and also in the introduction of this paper.

Lemma 1. *Suppose $s, p\geq 2$, Assumptions A–C hold, $\lambda\to 0$ and $n\lambda^{1+2s}\to\infty$, as $n\to\infty$. Then*

$$\int E[K_n^{0(s)}(y)-f^{(s)}(y)]^2dy\leq\frac{C_1}{n\lambda^{1+2s}}+C_2\lambda^{2p}\int[f^{(s+p)}(y)]^2dy,$$

where C_1 and C_2 are purely functions of K^0 and K respectively.

Proof. Let us recall that

$$K_n^0(y)=\frac{1}{n\lambda}\sum_{i=1}^{n}K^0\left(\frac{y-X_i}{\lambda}\right)\Rightarrow K_n^{0(s)}(y)=\frac{1}{n\lambda^{1+s}}\sum_{l=1}^{n}K^{0(s)}\left(\frac{y-X_l}{\lambda}\right).$$

Therefore, $E(K_n^{0(s)}(y))=\frac{1}{\lambda^s}\int K^{0(s)}(u)f(y-\lambda u)du$. Expanding $f(y-\lambda u)$ under Assumption C on $K^{0(s)}$, $E[K_n^{0(s)}(y)$ $f^{(s)}(y)+b(y)$, where

$$b(y)=\frac{(-1)^{s+p}\lambda^p}{(s+p-1)!}\int K^{0(s)}(u)u^{s+p}\int_0^1(1-t)^{s+p-1}f^{(s+p)}(y-t\lambda u)dtdu.$$

Applying Cauchy–Schwartz inequality it is easy to verify that

$$b^2(y) \le \frac{C'\lambda^{2p}}{[(s+p-1)!]^2} \int\int_0^1 \left|K^{0(s)}(u)u^{s+p}\right| (1-t)^{s+p-1} \left[f^{(s+p)}(y-t\lambda u)\right]^2 dtdu,$$

where $C' = \frac{\int \left|K^{0(s)}(u)u^{s+p}\right| du}{s+p}$. Consequently, under Assumption A on $f^{(s+p)}(\cdot)$

$$\left[E[K_n^{0(s)}(y)] - f^{(s)}(y)\right]^2 = b^2(y) \le \frac{C'\lambda^{2p}}{[(s+p-1)!]^2} g(y) \tag{A.1}$$

$$\Rightarrow \int \left[E[K_n^{0(s)}(y)] - f^{(s)}(y)\right]^2 dy \le \frac{C'\lambda^{2p}}{[(s+p-1)!]^2} \int g(y)dy.$$

Therefore

$$\int \left[E[K_n^{0(s)}(y)] - f^{(s)}(y)\right]^2 dy \le C^2\lambda^{2p} \int [f^{(s+p)}(y)]^2 dy,$$

where $C = \frac{C'}{(s+p-1)!}$ and $g(y) = \int |K^{0(s)}(u)u^{s+p}| \int_0^1 (1-t)^{s+p-1}[f^{(s+p)}(y-t\lambda u)]^2 dtdu$.

It is easy to verify that

$$\int \mathrm{Var}\left[K_n^{0(s)}(y)\right] dy = \frac{1}{n\lambda^{2+2s}} \int \mathrm{Var}\left[K^{0(s)}\left(\frac{y-X_1}{\lambda}\right)\right] dy$$

$$\le \frac{1}{n\lambda^{1+2s}} \iint \left[K^{0(s)}(u)\right]^2 f(y-u\lambda)dudy = \frac{\int \left[K^{0(s)}(u)\right]^2 du}{n\lambda^{1+2s}}. \tag{A.2}$$

Now

$$\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy = \int \mathrm{Var}\left[K_n^{0(s)}(y)\right] dy + \int \left[E[K_n^{0(s)}(y)] - f^{(s)}(y)\right]^2 dy.$$

Therefore from (A.1) and (A.2) we see that

$$\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy \le \frac{\int \left[K^{0(s)}(u)\right]^2 du}{n\lambda^{1+2s}} + C^2\lambda^{2p} \int [f^{(s+p)}(y)]^2 dy,$$

where $C = \frac{1}{(s+p)!} \int |K^{0(s)}(u)u^{s+p}| du$. This completes the proof of the lemma. □

Proof of Theorem 1. Under the assumption $h \in I$, h^* and $\hat{h}^*$ are minimizers of M and M^*, with respect to h, in $I = \left[\frac{\epsilon_1}{n^{1/(2s-1)}}, \frac{\epsilon_2}{n^{1/(2s-1)}}\right]$. Therefore $\hat{h}^*, h^* \in I$. Recalling the definitions of M^* and M it is easy to verify that, almost surely,

$$|M^* - M| \le L_{1n} + L_{2n} \quad \text{(say), where} \tag{A.3}$$

$$L_{1n} = \frac{1}{n} \left| \int \left\{ \int K(v)K_n^0(y-hv)dv \right\}^2 dy - \int \left\{ \int K(v)f(y-hv)dv \right\}^2 dy \right|,$$

$$L_{2n} = \left| \int \left[\int K(u)K_n^0(y-h.u)du - K_n^0(y) \right]^2 dy - \int \left[\int K(u)f(y-h.u)du - f(y) \right]^2 dy \right|.$$

Using $|a^2 - b^2| \le (|a| + b)|a - b|$, for any a and $b > 0$, it is easy to see that (writing $y^* = y - hv$)

$$L_{1n} \le \frac{1}{n} \int \left[\left\{ \int K(v)(|K_n^0(y^*)| + f(y^*))dv \right\} \left\{ \int K(v)|K_n^0(y^*) - f(y^*)|dv \right\} \right] dy.$$

Now $|K_n^0(y^*)| + f(y^*) \le |K_n^0(y^*) - f(y^*)| + 2f(y^*)$ (as f is non-negative). Using this inequality it is easy to see that

$$L_{1n} \le \frac{1}{n}\left[\int \left\{K_n^0(y) - f(y)\right\}^2 dy + 2\int f(y)|K_n^0(y) - f(y)|dy\right]$$

$$\frac{1}{n}\left[\int \left\{K_n^0(y) - f(y)\right\}^2 dy + 2\sqrt{\int f^2(y)dy \int \left\{K_n^0(y) - f(y)\right\}^2 dy}\right] = e_{1n} \quad \text{(say)}. \tag{A.4}$$

Now under the smoothness Assumptions A and C on f and K^0, using Taylor's expansion with integral remainder we get

$$L_{2n} = \frac{h^{2s}}{((s-1)!)^2}\left|\int \left\{\iint_0^1 (1-t)^{s-1}K(u)u^s K_n^{0(s)}(y - thu)dtdu\right\}^2 dy\right.$$
$$\left. - \int \left\{\iint_0^1 (1-t)^{s-1}K(u)u^s f^{(s)}(y - thu)dtdu\right\}^2 dy\right|.$$

Further using $|a^2 - b^2| \le (a-b)^2 + 2b|a-b|$, $a, b > 0$, we see that

$$L_{2n} \le \frac{h^{2s}}{[(s-1)!]^2}\left[\int f_{2y}^2 dy + 2\int \left[f_{2y}f_{3y}\right]dy\right] \quad \text{where}$$

$$f_{2y} = \int |K(u)u^s| \int_0^1 (1-t)^{s-1}\left|f^{(s)}(y - thu) - K_n^{0(s)}(y - thu)\right| dtdu \quad \text{and}$$

$$f_{3y} = \int |K(u)u^s| \int_0^1 (1-t)^{s-1}\left|f^{(s)}(y - thu)\right| dtdu.$$

Further it is easy to see that

$$\int f_{2y}^2 dy \le C_1 \int \left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy, \quad \text{and} \quad \int \left[f_{2y}f_{3y}\right]dy \le C_2\sqrt{\int \left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy},$$

where C_1, C_2 are positive constants (free of n and h). Therefore the above inequalities imply that

$$L_{2n} \le \frac{h^{2s}}{[(s-1)!]^2}\left[C_1 \int \left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy + C_2\sqrt{\int \left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy}\right].$$

Therefore for $h \in I$, we see that

$$L_{2n} \le \frac{\epsilon_2^{2s}}{n^{2s/(2s+1)}[(s-1)!]^2}\left[C_1 \int \left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy + C_2\sqrt{\int \left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy}\right]$$
$$= e_{2n} \quad \text{(say)}.$$

From (A.3)–(A.5) we get

$$|M - M^*| \le e_{1n} + e_{2n} \quad \forall h \in I.$$

We note that e_{1n} and e_{2n} are independent of h. Therefore

$$\|M - M^*\| = \sup_{h \in I} |M - M^*| \le e_{1n} + e_{2n}.$$

Hence using, $|\inf f - \inf g| \le \|f - g\|$, we see that

$$E|M(\hat{h}^*) - M(h^*)| \le E|M(h^*) - M^*(\hat{h}^*)| + E|M(\hat{h}^*) - M^*(\hat{h}^*)|$$
$$\le 2E\|M - M^*\| \le 2E(e_{1n} + e_{2n}). \tag{A.6}$$

Since K^0 is a pth order kernel, under the Assumptions A–C and for $\lambda = \frac{C}{n^{1/(2s+2p+1)}}$, from Rao (1983, p. 45) we see that

$$\int E\{f(y) - K_n^0(y)\}^2 dy = O\left(\frac{1}{n\lambda} + \lambda^{2p}\right).$$

Now recalling the formula of c_{1n} and using $E(\sqrt{X}) \le \sqrt{E(X)}$, where X is a nonnegative random variable, we see that

$$E(c_{1n}) = O\left(\frac{1}{n}\sqrt{\frac{1}{n\lambda} + \lambda^{2p}}\right).$$

$\lambda = \frac{C}{n^{1/(2s+2p+1)}}$ satisfies the conditions on λ in Lemma 1. Therefore under the stated conditions, using Lemma 1, we get

$$\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy = o(1).$$

Again using $E(\sqrt{X}) \le \sqrt{E(X)}$, where X is a nonnegative random variable, we see that

$$E(c_{2n}) = O\left(\frac{1}{n^{2s/(2s+1)}}\sqrt{\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy}\right).$$

Therefore from (A.6) we see that

$$\begin{aligned} E|M(\hat{h}^*) - M(h^*)| &\le 2E\|M - M^*\| \\ &= O\left(\frac{1}{n}\sqrt{\frac{1}{n\lambda} + \lambda^{2p}} + \frac{1}{n^{2s/(2s+1)}}\sqrt{\int E\left[K_n^{0(s)}(y) - f^{(s)}(y)\right]^2 dy}\right). \end{aligned} \tag{A.7}$$

Now using Lemma 1, with $\lambda = \frac{C}{n^{1/(2s+2p+1)}}$, $s, p \ge 2$, in the right side of (A.7) we get

$$E|M(\hat{h}^*) - M(h^*)| = O\left(\frac{1}{n^{1+p/(2s+2p+1)}} + \frac{1}{n^{2s/(2s+1)+p/(2s+2p+1)}}\right).$$

Further we note that $M \ge \frac{\int K^2}{nh} - \frac{C\int f^2(y)dy}{n}$, $\forall h \in I$ and hence

$$M(h^*) \ge \frac{\int K^2}{\epsilon_2 \cdot n^{(2s)/(2s+1)}} + o\left(\frac{1}{n^{2s/(2s+1)}}\right).$$

Therefore, under the stated conditions,

$$\begin{aligned} E\left|\frac{M(\hat{h}^*)}{M(h^*)} - 1\right| &= O\left(\frac{1}{n^{1/(2s+1)+p/(2s+2p+1)}} + \frac{1}{n^{p/(2s+2p+1)}}\right) \\ &= O\left(\frac{1}{n^{p/(2s+2p+1)}}\right), \quad \text{where } s, p \ge 2. \end{aligned}$$

So Theorem 1 is proved completely. □

References

Azzalini, A., Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. Appl. Stat. 39, 357–365.
Cao, R., 1993. Bootstrapping the mean integrated squared error. J. Multivariate Anal. 45, 137–160.
Cao, R., Cuevas, A., González-Manteiga, W., 1994. A comparative study of several smoothing methods in density estimation. Comput. Statist. Data Anal. 17, 153–176.
Faraway, J.J., Jhun, M., 1990. Bootstrap choice of bandwidth for density estimation. J. Amer. Statist. Assoc. 85 (412), 1119–1122.
Hall, P., Marron, J.S., Park, B.U., 1992. Smoothed cross-validation. Probab. Theory Related Fields 92, 1–20.
Jones, M.C., Marron, J.S., Park, B.U., 1991. A simple root n bandwidth selector. Ann. Statist. 19, 1919–1932.
Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Assoc. 91, 401–407.
Loader, C.R., 1999. Bandwidth selection: classical or plug-in? Ann. Statist. 27 (2), 415–438.
Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20 (2), 712–736.
Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. J. Amer. Statist. Assoc. 85 (409), 66–72.
Rao, B.L.S.P., 1983. Nonparametric Functional Estimation. Academic Press Inc., New York.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B 53, 683–690.
Taylor, C.C., 1989. Bootstrap choice of the smoothing parameter in kernel density estimation. Biometrika 76 (4), 705–712.

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Pointwise and uniform convergence of kernel density estimators using random bandwidths

Santanu Dutta [a,*], Alok Goswami [b]

[a] Mathematical Science Department, Tezpur University Napaam, 784028, Tezpur, Assam, India
[b] Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 18 February 2013
Received in revised form 11 September 2013
Accepted 11 September 2013
Available online 18 September 2013

MSC:
62G07
62G09
62G20

Keywords:
Density estimation
Random bandwidth
Point-wise
Sup-norm convergence

ABSTRACT

We obtain the rates of pointwise and uniform convergence of kernel density estim... using random bandwidths under i.i.d. as well as strongly mixing dependence assump... Pointwise rates are faster and not affected by the tail of the density.

1. Introduction

The estimation of the density of an absolutely continuous distribution has been an important problem in nonparame... statistics for a long time. Rosenblatt (1956) introduced the idea of a kernel-based density estimator which is defined ... follows.

Let $X_1, \ldots, X_n$ be identically distributed random variables with an unknown common density $f(\cdot)$. The *kernel den... estimator* (KDE) of f based on the kernel $K(\cdot)$ and bandwidth $h \equiv h_n$ is defined as

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

where the kernel K is a density function and $h \equiv h_n$ is the bandwidth which controls the smoothness of $\hat{f}_{n,h}$. A comm... assumption is that K is a second order kernel, i.e. it is a density satisfying $\int K(u)u\,du = 0$ and $\int K(u)u^2\,du < \infty$. Parz... (1962) proved pointwise convergence of such an estimator, assuming h to be a positive sequence satisfying $h \to 0$ a... $nh \to \infty$ as $n \to \infty$. Since then there has been extensive research on the asymptotic properties of $\hat{f}_{n,h}$. A detailed discussi... on the asymptotic properties of the KDE can be found in Rao (1983). Most of these asymptotic properties are obtain... assuming that the bandwidth sequence $\{h_n\}$ is a nonrandom positive sequence. However the practical application of kern...

* Corresponding author.
E-mail addresses: tezpur1976@gmail.com (S. Dutta), alok@isical.ac.in (A. Goswami).

0167-7152/$ – see front matter
http://dx.doi.org/10.1016/j.spl.2013.09.010

density estimation depends crucially on efficient algorithms for data-based choice of h. See Park and Marron (1990), Cao et al. (1994), Bose and Dutta (2013) for a review and comparison of different data-based bandwidth selection algorithms. If h is nonrandom, the convergence of the bias component $E[\hat{f}_{n,h}(y)] - f(y)$ is rather straightforward to obtain (see Rao (1983)). In that case, the convergence of the KDE is essentially determined by the asymptotic properties of the sequence $\{\hat{f}_{n,h}(y) - E[\hat{f}_{n,h}(y)]\}$, which is a sequence of the averages of a triangular array of mean zero random variables (see Rao, 1983 and Wied and Weißbach, 2010). Wied and Weißbach (2010) have reviewed different proofs of pointwise and uniform convergence of a KDE using a nonrandom bandwidth.

Far less seems to be known on point-wise or uniform convergence of $\hat{f}_{n,\hat{h}}(\cdot)$, where $\hat{h}$ is determined by some data-based bandwidth selection rule. In that case $\hat{h}$ is random, i.e. a function of $X_1, \ldots, X_n$, and $E[\hat{f}_{n,\hat{h}}(y)]$ is hard to compute. Therefore, it is difficult to prove the convergence of the bias component of the estimator in this case. Krieger and Pickands (1981), Mielniczuk (1990) have obtained the rate of pointwise convergence of $\hat{f}_{n,\hat{h}}(\cdot)$ where $\hat{h}$ is selected by the plug-in method. But for KDEs using other data-based bandwidth selectors, such results do not seem to be known. Under a number of assumptions on the kernel K, Einmahl and Mason (2005) proved that for any sequences $0 < a_n < b_n \leq 1$, satisfying $b_n = o(1)$ and $na_n / \log n \to \infty$,

$$\sup_{a_n \leq h \leq b_n} \|\hat{f}_{n,h} - E(\hat{f}_{n,h})\| = O\left(\sqrt{\frac{\max(\log(1/a_n), \log\log n)}{na_n}}\right) \text{ almost surely,}$$

and further, $\sup_{a_n \leq h \leq b_n} \|E(\hat{f}_{n,h}) - f\| = o(1)$ for any uniformly continuous f, where $\|\cdot\|$ denotes the sup-norm. These results naturally imply that $\|\hat{f}_{n,\hat{h}} - f\| = o(1)$ almost surely, where $\hat{h}$ is a random bandwidth satisfying $a_n \leq \hat{h} \leq b_n$. The work of Einmahl and Mason undoubtedly represents a rather significant achievement in the research on KDEs with data-based bandwidths. However, there seem to be some limitations. First of all, the results in Einmahl and Mason (2005) do not seem to provide any insight into the rate at which $P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon)$ goes to zero with increasing n, for arbitrary $\epsilon > 0$. But here is a more serious issue. Wied and Weißbach (2010) point out that the condition $\hat{h} \in [a_n, b_n]$, where a_n, b_n are nonrandom positive sequences, in Einmahl and Mason (2005) is quite restrictive. Ideally a random bandwidth $\hat{h}$ is expected to be scale invariant, i.e. $\hat{h}(CX_1, \ldots, CX_n) = C\hat{h}$, where $C > 0$. If a_n, b_n are nonrandom positive sequences, then a random bandwidth $a_n \leq \hat{h} \leq b_n$ cannot be scale invariant. As described below, our work addresses these issues.

We obtain the rates at which $r_{1n} = P(|\hat{f}_{n,\hat{h}}(x) - f(x)| > \epsilon)$ and $r_{2n} = P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon)$ converge to zero as $n \to \infty$, where $\hat{h}$ is a random bandwidth which optimizes some criterion on a compact interval. We are able to obtain sharper asymptotic upper bound for r_{1n} than r_{2n}. As a corollary, we prove that $\|\hat{f}_{n,\hat{h}} - f\|$ converges to zero completely under i.i.d. assumption. Complete convergence is stronger than the almost sure convergence. In general, while the rate of convergence of r_{2n} seems to depend on the tail of f, the convergence rate of r_{1n} seems to be unaffected by the same.

As for the issue of scale-invariance, we make a similar assumption as in Einmahl and Mason (2005), viz. $h \in H_n$ where $\{H_n\}$ is a sequence of compact intervals. But, the boundary points of H_n are chosen to be proportional to the sample interquartile range. This ensures that $\hat{h} \in H_n$ remains scale invariant. One can also use sample standard deviation to define H_n. But the use of standard deviation in H_n appears to impose more restrictive conditions on f for theoretical calculations, without any extra benefit (see a discussion on this issue in our final remarks). A wide variety of bandwidth selectors involve the optimization of some criterion with respect to h. One can always force the resulting bandwidth to be in H_n, by optimizing the criterion on H_n.

Finally, Wied and Weißbach (2010) also remark that the Einmahl and Mason (2005) use sophisticated mathematical techniques based on the paper by Talagrand (1994). In contrast, we use simple asymptotic calculations and some inequalities in Rao (1983), without requiring any sophisticated mathematical technique to obtain our results. Under strongly mixing type dependence assumption we use a Bernstein type inequality by Merlev'ede et al. (2009).

Bandwidth selectors are of two types, viz. local and global. In local bandwidth selection the aim is to estimate f at a given design point (assuming continuity of f). The global bandwidth selectors aim to capture all the important features of f, as far as possible. The pseudo-likelihood (PL) (Habbema et al., 1974) and the least squares cross validation (LSCV); Bowman (1984) and Stone (1984)), the biased cross validation (BCV) (Scott and Terrell (1987)), the smoothed cross validation (SCV) (Hall et al. (1992)), the different versions of the smooth bootstrap bandwidth selectors by Jones et al. (1991), Cao et al. (1994), Bose and Dutta (2013) and also the double kernel method by Devroye (1989) are well known global bandwidth selectors. All these methods involve the optimization of some function (based on $X_1, \ldots, X_n$) with respect to h. Among the local bandwidth selectors the bootstrap-based methods by Dutta (2014), Hazelton (1996, 1999) aim to minimize bootstrap estimate of the MSE of a density estimator.

While estimating f using a second order kernel it is quite common to assume that

$$h \in H_n = [c_1 n^{-1/5}, c_2 n^{-1/5}], \quad \text{where } c_1 < c_2.$$

Such an interval is well known to cover a wide range of reasonable bandwidths (see Park and Marron, 1990). Under this assumption, a random bandwidth $\hat{h}$ obtained by any one of the methods mentioned so far can be defined in general as

follows:

$$\hat{h} = \underset{h \in H_n}{\operatorname{argmin}} C(h), \tag{1.1}$$

where $C(h) = C(h, X_1, \ldots, X_n)$ is a function which is to be minimized for bandwidth selection.

It is natural to choose c_1, c_2 in H_n to be proportional to the sample standard deviation or the sample interquartile range. Dutta (2014) have suggested to use $c_1 = \frac{IQR}{2} 10^{-1}$ and $c_2 = \frac{IQR}{2} 10^{1}$, where $IQR = Q_3 - Q_1$. $Q_i \equiv Q_i(X_1, \ldots, X_n)$ is the ith sample quartile and $i = 1, 3$. We pursue with these choices of c_1 and c_2 in the sequel.

There are four theorems and one corollary in this paper. In Theorems 1 and 2 we obtain the upper bounds of the rates at which r_{1n}, r_{2n} converge to zero as n is increased. In these theorems we assume that $X_1, \ldots, X_n$ are i.i.d. random variables. In Theorems 3 and 4, we extend the results in Theorems 1 and 2 to the case where $\{X_n\}$ is a strongly mixing process with the mixing coefficient $\alpha(n) = O(\rho^n)$, $0 < \rho < 1$. Under i.i.d. assumption r_{2n} converge to zero at exponential rate. Consequently using the Borel–Cantelli lemma we prove complete uniform convergence of $\hat{f}_{n,\hat{h}}$ to f under i.i.d. assumptions, where $\hat{h}$ is a random bandwidth as defined in (1.1). Our results hold for a wide variety of global and local bandwidth selectors mentioned earlier.

2. Main results

We state four theorems and one corollary in this section. Proofs are given in Section 3. Let us introduce some notation. For any function g, let $\|g\| = \sup_{x \in \mathbb{R}} |g(x)|$ and $g^{(l)}$ denote the lth derivative of g, where $l \geq 1$.

2.1. i.i.d. case

In this subsection we assume that $\{X_i\}_{i=1,2,\ldots}$ is a sequence of i.i.d. random variables with density f. In the first theorem we provide insight into the asymptotic accuracy of $\hat{f}_{n,\hat{h}}(x)$.

To state and prove the theorems we need some assumptions on the kernel K and the density f. They are as follows.

Assumption 1. K is a second order kernel, such that K is a bounded density satisfying $K^{(1)}$ continuous and

$$|z|K(z), \quad |zK^{(1)}(z)| \to 0, \quad \text{as } |z| \to \infty.$$

Assumption 2. There exists $\eta > 0$, such that f is strictly positive on the intervals $[Q_i^* - \eta, Q_i^* + \eta]$, $i = 1, 3$, where Q_i^* is the ith quartile of the underlying distribution.

Let us first introduce a lemma which will be used in the proof of Theorem 1.

Lemma 1. *Let K be a kernel satisfying Assumption 1 and $\|f^{(1)}\| < \infty$. Let $I_n = \left[\frac{a}{n^{1/5}}, \frac{b}{n^{1/5}}\right]$, where $0 < a < b$. Then*

$$P\left(\sup_{h \in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon\right) = O\left(n^{1/5} \exp(-Cn^{4/5}\epsilon^2)\right),$$

where C is a positive constant free of x.

In Lemma 1, a, b are positive constants. In Theorem 1 we extend this result to the case where I_n is replaced by H_n, i.e. a, b are replaced by the random variables c_1, c_2.

Theorem 1. *Let K be a kernel satisfying Assumption 1 and f be a density function satisfying Assumption 2 and $\|f^{(1)}\| < \infty$. Then for every $\epsilon > 0$,*

$$r_{1n} = P(|\hat{f}_{n,\hat{h}}(x) - f(x)| > \epsilon) = O\left(n^{1/5} \exp(-Cn^{4/5}\epsilon^2)\right), \text{ where}$$

C is a positive constant free of x.

Let us introduce another lemma which is an extension of Lemma 1 to sup-norm distance between $\hat{f}_{n,\hat{h}}(\cdot)$ and f.

Lemma 2. *Let K be a kernel satisfying Assumption 1 and $\|f\|$, $\|f^{(1)}\| < \infty$. Let $0 < \int |x|^{\gamma} f(x)dx < \infty$, for some $\gamma > 0$. Also let $I_n = \left[\frac{a}{n^{1/5}}, \frac{b}{n^{1/5}}\right]$, where $0 < a < b$. Then*

$$P\left(\sup_{h \in I_n} \|\hat{f}_{n,h} - f\| > \epsilon\right) = O\left(n^{\frac{(3+1/\gamma)}{5}} \exp(-Cn^{4/5}\epsilon^2)\right).$$

C is a positive constant.

In the next theorem we obtain the rate at which $P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon)$ goes to zero, for any positive ϵ.

Theorem 2. *Let K be a kernel satisfying Assumption 1 and f be a density satisfying Assumption 2 and $\|f\|, \|f^{(1)}\| < \infty$. Also let $\int |x|^{\gamma} f(x)dx < \infty$, for some $\gamma > 0$. Then for every $\epsilon > 0$,*

$$\tau_{2n} = P\left(\|\hat{f}_{n,h} - f\| > \epsilon\right) = O\left(n^{\frac{3\gamma+1}{5}} \exp(-Cn^{4/5}\epsilon^2)\right).$$

C is a positive constant.

Using the Borel–Cantelli lemma it is easy to see the following corollary.

Corollary 2.1. *Under the assumptions stated in Theorem 2, $\|\hat{f}_{n,h} - f\| = o(1)$ completely, as $n \to \infty$.*

2.2. Strongly mixing case

Suppose $\{X_t,\ t \in \mathbb{Z}\}$ is a $\mathbb{R}$-valued, strictly stationary process with marginal density f. Let $M_{-\infty}^{t}$ and M_{t+n}^{∞} denote σ-fields generated by $\{X_l,\ l \le t\}$ and by $\{X_l,\ l \ge t+n\}$ respectively. Then X_t is a strong mixing process if

$$\alpha(n) = \sup_t \sup\{|P(A \cap B) - P(A)P(B)| : A \in M_{t+n}^{\infty}, B \in M_{-\infty}^{t}\} \downarrow 0, \quad \text{as } n \to \infty.$$

Under very general dependence assumptions (that includes strong mixing condition), Lardjane (2007) has shown that the MSE of a KDE $\hat{f}_{n,h}(y)$ goes to zero at the rate similar (up to a logarithm) to the rate of convergence of the MSE under i.i.d. assumptions for h equal to a multiple of $(\log n/n)^{1/5}$ (see page 213, Lardjane (2007)). So under strong mixing condition we use $H_n = \left[c_1 (\log(n)/n)^{1/5},\ c_2 (\log(n)/n)^{1/5}\right]$, where c_1, c_2 are as defined earlier.

Lemma 3. *Let $\{X_n\}_{n=1,2,\ldots}$ be a strongly mixing process with marginal density f, satisfying $\|f^{(1)}\| < \infty$. Let $\alpha(n) \le \exp(-2cn)$, where $c > 0$. If $I_n = \left[a(\log(n)/n)^{1/5}, b(\log(n)/n)^{1/5}\right]$ where $0 < a < b$, then under Assumption 1 we see that*

$$P\left(\sup_{h \in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon\right) = O\left((n/\log(n))^{1/5} \exp\left(-(n/\log(n))^{3/5} C\epsilon^2\right)\right).$$

C is a positive constant free of x.

Lemma 4. *Let $\{X_t\}_{t=1,2,\ldots}$ be a strongly mixing process, with the common marginal density f satisfying Assumption 1. Let $\alpha(n) \le \exp(-2cn)$, where $c > 0$. Then*

$$P\left(|Q - Q_i^*| > \eta/4\right) = O\left(\sqrt{n}\exp(-\sqrt{n}s)\right), \quad i = 1, 3,$$

s is a positive constant.

The above lemma follows from inequality (3.7) in page 658 in Wang et al. (2011). The proof is given in the appendix.

Repeating the arguments used in the proof of Theorem 1, and using the Lemmas 3 and 4 we get the following theorem.

Theorem 3. *Let $\{X_n\}_{n=1,2,\ldots}$ be a strongly mixing process with marginal density f, satisfying Assumption 2 and $\|f^{(1)}\| < \infty$. Let $\alpha(n) \le \exp(-2cn)$ where $c > 0$. Then under Assumption 1*

$$P(|\hat{f}_{n,h}(x) - f(x)| > \epsilon) = O\left(\sqrt{n}\exp(-C(n/\log(n))^{3/5}\epsilon^2)\right),$$

C is a positive constant free of x.

Our next theorem, viz. Theorem 4, is an extension of our Theorem 2 to the strong mixing case, where the mixing coefficient $\alpha(n)$ decays at an exponential rate with the increase in n (we are thankful to the reviewer for suggesting this extension). To prove this theorem we need the following lemmas.

Lemma 5. *Let $\{X_t\}$ be a strongly mixing stationary process with marginal density f, satisfying $\|f\|, \|f^{(1)}\|, \int |x|^{\gamma} f(x)dx < \infty$, for some $\gamma > 0$ and $|f(x)| = o(1)$ as $|x| \to \infty$. The mixing coefficient α satisfies $\alpha(n) \le \exp(-2cn)$, for some $c > 0$. Further let K be a continuous density satisfying Assumption 1. For any $h \in \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$ and $\epsilon > 0$, $\exists\, C > 0$ such that*

$$P(\|\hat{f}_{n,h} - E(\hat{f}_{n,h})\| > \epsilon) = O\left(\left(\frac{n}{\log(n)}\right)^{\frac{(2+1\vee\gamma)}{5}} \exp(-C\sqrt{n}\epsilon^2)\right) \quad \text{as } n \to \infty.$$

In the above lemma h is a nonrandom bandwidth. The next lemma is an extension of Lemma 2 to a strongly mixing stationary process.

Lemma 6. *Let $\{X_t\}$ be a strongly mixing stationary process with marginal density f, satisfying $\|f\|$, $\|f^{(1)}\|$, $\int |x|^\gamma f(x)dx < \infty$, for some $\gamma > 0$ and $|f(x)| = o(1)$ as $|x| \to \infty$. The mixing coefficient α satisfies $\alpha(n) \le \exp(-2cn)$, for some $c > 0$. Further let K be a continuous density satisfying Assumption 1. Then as $n \to \infty$*

$$P\left(\sup_{h \in I_n} \|\hat{f}_{n,h} - f\| > \epsilon\right) = O\left(\left(\frac{n}{\log(n)}\right)^{\frac{(3+2/\gamma)}{5}} \exp(-C\sqrt{n}\epsilon^2)\right),$$

where $I_n = \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$ and a, b, C are positive constants.

Now we state Theorem 4, which holds for a kernel estimator based on strongly mixing stationary process and using a random bandwidth satisfying the stated condition.

Theorem 4. *Let $\{X_t\}$ be a strongly mixing stationary process with marginal density f, satisfying Assumption 2. Moreover let $\|f\|$, $\|f^{(1)}\|$, $\int |x|^\gamma f(x)dx < \infty$, for some $\gamma > 0$ and $|f(x)| = o(1)$ as $|x| \to \infty$. The kernel K is a continuous density satisfying Assumption 1. The mixing coefficient α satisfies $\alpha(n) \le \exp(-2cn)$, for some $c > 0$. Let $\hat{h} \in [c_1(\log(n)/n)^{1/5}, c_2(\log(n)/n)^{1/5}]$, where c_1, c_2 are as described in our paper. Then for every $\epsilon > 0$, as $n \to \infty$*

$$P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon) = O\left(\left(\frac{n}{\log(n)}\right)^{\frac{(3+1/\gamma)}{5}} \exp(-C\sqrt{n}\epsilon^2)\right),$$

where C is a positive constant. Consequently, $\|\hat{f}_{n,\hat{h}} - f\| \to 0$ almost surely as $n \to \infty$.

We note that the condition $\alpha(n) < D\rho^n$ implies that $\alpha(n) < \exp(-2cn)$ where $0 < \rho < 1$, $D > 0$ and $c = -\log(\rho)/2 - \log(D)/(2n)$. Clearly $c > 0$ for sufficiently large n. Therefore, under the stated conditions, Theorem 4 also holds for a strongly mixing sequence with $\alpha(n) = O(\rho^n)$, where $0 < \rho < 1$.

Final remarks.

1. Theorem 1 provides insight into the accuracy of the density estimators using the local bootstrap bandwidth selectors proposed by Hazelton (1996, 1999) and Dutta (2014). Theorem 3 ensures that these estimators remain consistent even in the presence of strong mixing type dependence.
2. Theorem 2 provides insight into the accuracy of a global density estimate in terms of the sup-norm distance. This result holds for the estimators obtained by the PL cross validation, the LSCV, the BCV, the SCV, the double kernel method and also the bootstrap bandwidths by Jones et al. (1991), Cao et al. (1994), Bose and Dutta (2013). Corollary 2.1 ensures complete uniform convergence of these estimators.
3. The range of bandwidths H_n can be widened to $\left[\frac{c_1}{n^{1/5+\delta}}, \frac{c_2}{n^{1/5-\delta}}\right]$, where $0 < \delta < 1/5$, to accommodate more values of h (we are thankful to Prof. J.S. Marron for this suggestion). In that case, under i.i.d. assumption, using similar calculations as in Section 3 we get that for $\hat{h} \in H_n$

$$P(|\hat{f}_{n,\hat{h}}(x) - f(x)| > \epsilon) = O\left(n^{\frac{4}{5}} \exp\left(-Cn^{4/5-\delta}\epsilon^2\right)\right)$$

and

$$P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon) = O\left(n^{\frac{(8+2/\gamma)}{5}} \exp\left(-Cn^{4/5-\delta}\epsilon^2\right)\right),$$

where C is positive constant.

If $\{X_t\}$ is a strongly mixing stationary process satisfying the conditions on f and α in Theorem 4, the range of bandwidths H_n can be widened to $\left[c_1(\log(n)/n)^{1/5+\delta}, c_2(\log(n)/n)^{1/5-\delta}\right]$, where $0 < \delta < 1/5$ and $(\log(n)/n)^{1/5-\delta}\log(n)(\log\log n) = o(1)$ as $n \to \infty$. In that case,

$$P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon) = O\left(\left(\frac{n}{\log(n)}\right)^{\frac{(8+2/\gamma)}{5}} \exp\left(-C(n/\log(n))^{3/5-2\delta}\epsilon^2\right)\right), \quad \text{where } C > 0.$$

4. The moment assumption $\int |x|^\gamma f(x)dx < \infty$ in Theorems 2 and 4 depends on the tail of f. For heavy-tailed distributions this condition holds for smaller values of γ. So for the heavy-tailed densities, $P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon)$ seems to converge to zero at a slower rate. However for any positive value of γ, $\|\hat{f}_{n,\hat{h}} - f\| \to 0$ completely as $n \to \infty$, under i.i.d. assumption (see Corollary 2.1). There appears to be no effect of the tail of f on the results related to pointwise convergence of $f_{n,\hat{h}}$.

5. Under the extra assumption that the population variance is finite, one can replace the sample interquartile range by sample standard deviation in the boundary points of H_n (we are thankful to the reviewer for raising this point). In that case

$$P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon) \le 2P(|sd - sd^*| > \eta/2) + P\left(\sup_{h \in I_n'} \|\hat{f}_{n,h} - f\| > \epsilon\right),$$

where sd, sd^* are the sample and the population standard deviations respectively, $0 < \eta < sd^*/2$ and

$$I_n' = \left[0.002sd^* n^{-1/5},\ 2000sd^* n^{-1/5}\right].$$

Under the conditions stated in Theorems 2 and 4 and the extra assumption that $\int x^2 f(x)dx < \infty$ $P(\sup_{h \in I_n'} \|\hat{f}_{n,h} - f\| > \epsilon)$ converges to zero at similar rate as $P(\sup_{h \in I_n} \|\hat{f}_{n,h} - f\| > \epsilon)$ and $P(|sd - sd^*| > sd^*/2) = o(1)$ as $n \to \infty$, under i.i.d. as well as strongly mixing dependence assumptions. Therefore even if *IQR* is replaced by sd in H_n, $\|\hat{f}_{n,\hat{h}} - f\|$ converges in probability to zero under i.i.d. as well as strongly mixing dependence assumptions, provided $\int x^2 f(x)dx < \infty$.

If *IQR* is replaced by sd in the boundary points of H_n, the convergence rate of $P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon)$ depends on the convergence rate of $P(|sd - sd^*| > \eta/2)$. The later essentially depends on the convergence rate of $P(|m_2 - \mu_2| > \delta)$, where m_2, μ_2 are the sample and population 2nd moments respectively and $\delta > 0$. The best possible rate of $P(|m_2 - \mu_2| > \delta)$ is usually obtained by using a Bernstein type inequality, where the random variables are assumed to be bounded. One can also use the well known inequality $P(|m_2 - \mu_2| > \delta) \le E(m_2 - \mu_2)^2/\delta^2$. Then there is no need to assume that the random variables are bounded. But the 4th population moment is assumed to be finite, and the upper bound obtained is somewhat crude (as even under i.i.d. assumption $E(m_2 - \mu_2)^2 = O(1/n)$). So to obtain the exponential convergence rate of $P(|sd - sd^*| > \eta/2)$ we need more restrictive assumptions on f than what is used in Theorems 2 and 4 (viz. Assumption 2).

Hence if sd is used, instead of *IQR*, more conditions are needed to obtain the similar rate of convergence of $P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon)$. Moreover the role of *IQR* or sd in c_1 and c_2 is only to ensure that $\hat{h}$ is scale invariant. So there seems to be hardly any extra benefit of using sd instead of *IQR* in c_1, c_2.

3. An outline of proofs

In this final section we provide an overview of the main idea used in the proofs of all Theorems 1–4 and some important lemmas. The details are available with the authors.

3.1. Proof of theorems

Proof of Theorems 1 and 2. Let us first discuss the proofs of Theorems 1 and 2. In these theorems we assume that $X_1, \ldots, X_n$ are i.i.d. random variables with density f satisfying Assumption 2. We note that for any random bandwidth $\hat{h}$

$$r_{1n} = P(|\hat{f}_{n,\hat{h}}(x) - f(x)| > \epsilon) \le P\left(\hat{h} \notin I_n\right) + P\left(|\hat{f}_{n,\hat{h}}(x) - f(x)| > \epsilon,\ \hat{h} \in I_n\right) \tag{3.1}$$

and

$$r_{2n} = P(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon) \le P\left(\hat{h} \notin I_n\right) + P\left(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon,\ \hat{h} \in I_n\right) \tag{3.2}$$

where $I_n = \left[\frac{a}{n^{1/5}}, \frac{b}{n^{1/5}}\right]$, where $a = \frac{IQR^* - \eta/2}{2 \times 10^3}$ and $b = \frac{(IQR^* + \eta/2)10^3}{2}$. η is a positive constant as in Assumption 2, and let $\eta < IQR^*$. The definition (1.1) implies that $\hat{h} \in H_n$. Therefore it is easy to verify that

$$P\left(\hat{h} \notin I_n\right) \le 2P(|Q_1 - Q_1^*| > \eta/4) + 2P(|Q_3 - Q_3^*| > \eta/4). \tag{3.3}$$

Also we see that

$$P\left(|\hat{f}_{n,\hat{h}}(x) - f(x)| > \epsilon,\ \hat{h} \in I_n\right) \le P\left(\sup_{h \in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon\right) \tag{3.4}$$

$$P\left(\|\hat{f}_{n,\hat{h}} - f\| > \epsilon,\ \hat{h} \in I_n\right) \le P\left(\sup_{h \in I_n} \|\hat{f}_{n,h} - f\| > \epsilon\right). \tag{3.5}$$

The inequalities (3.1) to (3.5) imply that the convergence rates of r_{1n} and r_{2n} depend on the rate at which $P(|Q_i - Q_i^*| > \eta/4)$, $i = 1, 3$, $P(\sup_{h \in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon)$ and $P(\sup_{h \in I_n} \|\hat{f}_{n,h} - f\| > \epsilon)$ converge to zero, as n is increased.

Under i.i.d. assumption and Assumption 2, using Theorem 2.3.2 in page 74 in Serfling (1980), we see that $P(|Q_i - Q_i^*| > \eta/4)$, $i = 1, 3$, converges to zero at an exponential rate. The rates of convergence of $P(\sup_{h\in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon)$ and $P(\sup_{h\in I_n} \|\hat{f}_{n,h} - f\| > \epsilon)$ are obtained in Lemmas 1 and 2 respectively.

Therefore under the i.i.d. assumption, Assumption 2 and the stated conditions in Lemmas 1 and 2 we get the convergence rates of r_{1n} and r_{2n} as a direct consequence of these lemmas. This completes the proofs of Theorems 1–2. □

Proofs of Theorems 3 and 4. To prove Theorems 3 and 4 we assume that $\{X_t\}$ is a strongly mixing stationary process with marginal density f satisfying Assumption 2 and the mixing coefficient α satisfies $\alpha(n) \leq \exp(-2cn)$, for some $c > 0$. In this case we define $I_n = \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$. The choice of a and b remains as it is.

The inequalities (3.1) to (3.5) do not depend on any dependence assumption, and they hold in the presence of the strongly mixing type dependence as well. Therefore again we see that the convergence rates of r_{1n}, r_{2n} depend on the rates at which $P(|Q_i - Q_i^*| > \eta/4)$, $i = 1, 3$, $P(\sup_{h\in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon)$ and $P(\sup_{h\in I_n} \|\hat{f}_{n,h} - f\| > \epsilon)$ converge to zero as $n \to \infty$ under strongly mixing dependence assumption.

Under strongly mixing dependence assumption the convergence rates of $P(|Q_i - Q_i^*| > \eta/4)$, $i = 1, 3$ are obtained in Lemma 4. Under the same dependence assumption, Assumption 2 and some extra assumptions on f the convergence rates of $P(\sup_{h\in I_n} |\hat{f}_{n,h}(x) - f(x)| > \epsilon)$ and $P(\sup_{h\in I_n} \|\hat{f}_{n,h} - f\| > \epsilon)$ are obtained in Lemmas 3 and 6, respectively. Therefore the proofs of the Theorems 3 and 4 follow from the inequalities (3.1) to (3.5) and Lemmas 3, 4 and 6. □

3.2. Proof of lemmas

Let us now discuss the proofs of Lemmas 1–3 and 6 which are used in the proofs of Theorems 1–4. Lemma 5 are used to prove Lemma 6. The proofs of Lemmas 1–3 and 6 depend on one inequality stated and proved below. The proof Lemma 5 is discussed briefly.

Given $\epsilon > 0$. Under the Assumption 1 on the kernel K, one can partition the interval I_n into $k(n)$ non-overlapping sub-intervals $\{I_{ni}, i = 1, \ldots, k(n)\}$ each of length δ_n such that

$$\sup_{h\in I_{ni}} \|\hat{f}_{n,h} - \hat{f}_{n,h_i}\| < \epsilon/2, \quad i = 1, \ldots, k(n), \tag{3.6}$$

where h_i is a boundary point of the sub-interval I_{ni}, $i = 1, \ldots, k(n)$.

Proof. Let $I_n = \left[\frac{a}{n^{1/5}}, \frac{b}{n^{1/5}}\right]$ and $g_z(h) = \frac{1}{h}K(z/h)$. Under Assumption 1, $\|K\|$, $\sup_{-\infty<z<\infty} |zK^{(1)}(z)|$ are finite numbers. Therefore under this assumption

$$\frac{d}{dh}g_z(h) \leq \frac{1}{h^2}\left\{\|K\| + \sup_{-\infty<z<\infty} |zK^{(1)}(z)|\right\} \leq Cn^{2/5}, \ \forall h \in I_n \text{ and } -\infty < z < \infty.$$

Therefore, $|g_z(h) - g_z(h_i)| \leq (h - h_i)\sup_{h\in I_n} \left|\frac{d}{dh}g_z(h)\right| \leq Cn^{2/5}(h - h_i)$, $-\infty < z < \infty$.

Given $\epsilon > 0$, let $\delta_n = \frac{\epsilon}{2C}n^{-2/5}$. Then for $h \in I_{ni}$, $i = 1, \ldots, k(n)$,

$$\begin{aligned} |\hat{f}_{n,h}(x) - \hat{f}_{n,h_i}(x)| &\leq \frac{1}{n}\sum_{j=1}^{n}\left|\frac{1}{h}K\left(\frac{x - X_j}{h}\right) - \frac{1}{h_i}K\left(\frac{x - X_j}{h_i}\right)\right| \\ &\leq \frac{1}{n}\sum_{j=1}^{n}\left|g_{x-X_j}(h) - g_{x-X_j}(h_i)\right| < \epsilon/2, \quad \forall x. \end{aligned}$$

$$\Rightarrow \|\hat{f}_{n,h} - \hat{f}_{n,h_i}\| < \epsilon/2, \quad i = 1, \ldots, k(n).$$

The right side of the above inequality is free of h and x. Therefore (3.6) follows from the above inequality. □

We note that in the above mentioned proof no dependence assumption is used. Moreover, if $I_n = \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$, (3.6) is proved by using $\delta_n = \frac{\epsilon}{2C}\left(\frac{\log(n)}{n}\right)^{2/5}$. Therefore (3.6) continues to hold even for a density estimate based on a strongly mixing stationary process, with appropriate modification in I_n and δ_n.

Proof the Lemmas 1 and 2. Given $\epsilon > 0$, we partition the interval $I_n = \left[\frac{a}{n^{1/5}}, \frac{b}{n^{1/5}}\right]$ into $k(n)$ non-overlapping sub-intervals $\{I_{ni}, i = 1, \ldots, k(n)\}$ each of length $\delta_n = \frac{\epsilon}{2C}n^{-2/5}$, where $C = \|K\| + \sup_{-\infty<z<\infty} |zK^{(1)}(z)|$. Clearly $k(n)$ is a multiple of n

The following inequalities are trivial:

$$P\left(\sup_{h\in I_n}|\hat{f}_{n,h}(x)-f(x)|>\epsilon\right)\le P\left(\max_{1\le i\le k(n)}\left\{|\hat{f}_{n,h_i}(x)-f(x)|+\sup_{h\in I_{ni}}|\hat{f}_{n,h}(x)-\hat{f}_{n,h_i}(x)|\right\}>\epsilon\right),$$

$$P\left(\sup_{h\in I_n}\|\hat{f}_{n,h}-f\|>\epsilon\right)\le P\left(\max_{1\le i\le k(n)}\left\{\|\hat{f}_{n,h_i}-f\|+\sup_{h\in I_{ni}}\|\hat{f}_{n,h}-\hat{f}_{n,h_i}\|\right\}>\epsilon\right).$$

Using (3.6) on the right side of the above inequalities we get the following inequalities:

$$P\left(\sup_{h\in I_n}|\hat{f}_{n,h}(x)-f(x)|>\epsilon\right)\le\sum_{i=1}^{k(n)}P\left(|\hat{f}_{n,h_i}(x)-f(x)|>\epsilon/2\right)$$

$$P\left(\sup_{h\in I_n}\|\hat{f}_{n,h}-f\|>\epsilon\right)\le\sum_{i=1}^{k(n)}P\left(\|\hat{f}_{n,h_i}-f\|>\epsilon/2\right).$$

Each $h_i\in I_n$. Hence h_i is a multiple of $n^{-1/5}$, $i=1,\ldots,k(n)$. Therefore under the assumption $\|f^{(1)}\|<\infty$, $\exists N_\epsilon>1$ such that

$$\|E[\hat{f}_{n,h_i}]-f\|<\epsilon/4,\quad\forall n>N_\epsilon,\ i=1,2,\ldots,k(n).$$

Hence for $n>N_\epsilon$,

$$P\left(\sup_{h\in I_n}|\hat{f}_{n,h}(x)-f(x)|>\epsilon\right)\le\sum_{i=1}^{k(n)}P\left(|\hat{f}_{n,h_i}(x)-E[\hat{f}_{n,h_i}(x)]|>\epsilon/4\right)\tag{3.7}$$

$$P\left(\sup_{h\in I_n}\|\hat{f}_{n,h}-f\|>\epsilon\right)\le\sum_{i=1}^{k(n)}P\left(\|\hat{f}_{n,h_i}-E[\hat{f}_{n,h_i}]\|>\epsilon/4\right).\tag{3.8}$$

Assumption 1 covers the conditions on K in Condition 6 in Theorem 3.1.5, in page 183 in Rao (1983). Using inequality (27) in page 184 in Rao (1983) we get the following inequality:

$$P\left(|\hat{f}_{n,h_i}(x)-E[\hat{f}_{n,h_i}(x)]|>\epsilon/4\right)\le 2\exp(-C_1nh_i\epsilon^2/2)\le 2\exp(-Cn^{4/5}\epsilon^2),$$

where C_1, C are positive constants free of x. Substituting the above inequality on the right side of (3.7), and using the fact that $k(n)$ is a multiple of $n^{1/5}$ we get Lemma 1. □

To prove Lemma 2, we see that $nh_i/\log(n)\to\infty$, $i=1,\ldots,k(n)$, as $n\to\infty$. So each h_i satisfies the Condition 10 (on the bandwidth) in page 185 of Rao (1983). Assumption 1 covers the Condition 9 (on the kernel) in page 185 of Rao (1983). Further assumptions on f stated in Lemma 2 also cover all the conditions stated in Theorem 3.1.7 in Rao (1983) (pages 184–185).

Therefore, under the conditions stated in Lemma 2, using inequality (49) in the proof of Theorem 3.1.7 in Rao (1983) (pages 184 and 188),

$$P\left(\|\hat{f}_{n,h_i}-E(\hat{f}_{n,h_i})\|>\epsilon/4\right)\le\exp(-C_2nh_i)+2(1+2a_n/b_n)\exp(-C_1nh_i),$$

where $C_2=c_2\epsilon$, $a_n=\frac{c_3}{(\epsilon h_i)^{1/\gamma}}$, $b_n=\epsilon h_i^2c_4$, $C_1=\epsilon^2/(c_5+c_6\epsilon)$. c_i, $i=2,\ldots,6$, are positive constants. Clearly a_n/b_n is a multiple of $h_i^{-(2+1/\gamma)}$ and each $an^{-1/5}\le h_i$, for $i=1,\ldots,k(n)$. Therefore we have the following equation:

$$P\left(\|\hat{f}_{n,h_i}-E(\hat{f}_{n,h_i})\|>\epsilon/4\right)=O\left(n^{\frac{(2+1/\gamma)}{5}}\exp(-C\epsilon^2n^{4/5})\right),$$

where C is a positive constant. Substituting the above inequality on the right side of (3.8), and using the fact that $k(n)$ is a multiple of $n^{1/5}$ we get Lemma 2. □

Proof of the Lemmas 3 and 6. We partition $I_n=\left[a\left(\frac{\log(n)}{n}\right)^{1/5},b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$ into $k(n)$ non-overlapping sub-intervals $\{I_{ni},i=1,\ldots k(n)\}$ each of length $\delta_n=\frac{\epsilon}{2C}\left(\frac{\log(n)}{n}\right)^{2/5}$. Clearly $k(n)$ is a multiple of $\left(\frac{n}{\log(n)}\right)^{1/5}$. We note that the inequalities (3.7) and (3.8) are obtained without any dependence assumption, and that they continue to hold for the kernel estimators based on a strongly mixing stationary process with density f. Hence to prove Lemma 3 we have to obtain the convergence rate of $P\left(|\hat{f}_{n,h_i}(x)-E[\hat{f}_{n,h_i}(x)]|>\epsilon/4\right)$ under strongly mixing dependence assumption, where $h_i\in I_n$, $i=1,\ldots,k(n)$. The arguments are as follows.

Let $Y_{nj} = K\left(\frac{x-X_j}{h_i}\right) - E\left[K\left(\frac{x-X_j}{h_i}\right)\right]$, $j = 1, \ldots, n$, $n \in \mathbb{N}$. Clearly $\{Y_{nj}, j = 1, \ldots, n,\ n \in \mathbb{N}\}$ is a triangular array of ... zero random variables and $|Y_{nj}| \le 2K_1$, $\forall j,\ n$, and

$$P\left(|\hat{f}_{n,h_i}(x) - E(\hat{f}_{n,h_i}(x))| > \epsilon/2\right) = P\left(\left|\sum_{j=1}^{n} Y_{n,j}\right| > nh_i\epsilon/2\right).$$

Merlevède et al. (2009) obtained the Bernstein type inequality for strongly mixing bounded random variables, where ... mixing coefficient converges to zero at an exponential rate.

We note that if $\{X_n\}_{n\in\mathbb{N}}$ is a stationary strongly mixing process with mixing coefficient α, each row of the triangular ... $\{Y_{nj}, j = 1, \ldots, n,\ n \in \mathbb{N}\}$ represents a strongly mixing stationary sequence of mean zero bounded random varia... with a sequence of mixing coefficients bounded above by $\{\alpha(n)\}$. Under the stated condition on α, the sequence of mi... coefficients of $\{Y_{nj}, j = 1, \ldots, n,\ n \in \mathbb{N}\}$ (say $\{\alpha'\}$) also satisfies the stated condition, viz. $\alpha'(n) \le \exp(-2cn)$. Now ... Theorem 1 in page 3 of Merlevède et al. (2009) we get the following inequality. For $n \ge 4$ and $\epsilon > 0$

$$\begin{aligned} P\left(|\hat{f}_{n,h_i}(x) - E(\hat{f}_{n,h_i}(x))| > \epsilon/4\right) &= P\left(\left|\sum_{j=1}^{n} Y_{n,j}\right| > nh_i\epsilon/4\right) \\ &\le \exp\left(-\frac{C'nh_i^2\epsilon^2}{4K_1^2 + 2K_1\epsilon h_i \log(n)(\log\log n)}\right), \end{aligned}$$

where C' is positive constant free of x. Recall that $a\left(\frac{\log(n)}{n}\right)^{1/5} \le h_i \le b\left(\frac{\log(n)}{n}\right)^{1/5}$, $i = 1, \ldots, k(n)$. Therefore $\exists N$ ... and $C > 0$ such that, $\frac{C'nh_i^2\epsilon^2}{4K_1^2+2K_1\epsilon h_i \log(n)(\log\log n)} \ge C\left(\frac{n}{\log(n)}\right)^{\frac{3}{5}}\epsilon^2$, $\forall n > N_\epsilon$.

Therefore as $n \to \infty$

$$P\left(|\hat{f}_{n,h_i}(x) - E(\hat{f}_{n,h_i}(x))| > \epsilon/4\right) = O\left(\exp\left(-C\left(\frac{n}{\log(n)}\right)^{\frac{3}{5}}\epsilon^2\right)\right).$$

Substituting (3.9) on the right hand side of inequality (3.7), and using the fact that $k(n)$ is a multiple of $(n/\log(n))^{1/5}$ ... Lemma 3. □

To prove Lemma 6 we recall that inequality (3.8) continues to hold under the strong mixing type dependence assum... with

$$I_n = \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right].$$

From (3.8) we see that to prove Lemma 6 we have to obtain the rate at which $P\left(\|\hat{f}_{n,h_i} - E[\hat{f}_{n,h_i}]\| > \epsilon/4\right)$, where $h_i \in$ ... $1, \ldots, k(n)$ and $k(n)$ is multiple of $(n/\log(n))^{1/5}$. Therefore we see that Lemma 6 follows from inequality (3.8), Lem... and the fact that $k(n)$ is multiple of $(n/\log(n))^{1/5}$. □

Proof of Lemma 5.

$$P(\|\hat{f}_{n,h} - E(\hat{f}_{n,h})\| > \epsilon) \le P\left(\sup_{|y|\le a_n} |\hat{f}_{n,h}(y) - E(\hat{f}_{n,h}(y))| > \epsilon\right) + P\left(\sup_{|y|>a_n} |\hat{f}_{n,h}(y) - E(\hat{f}_{n,h}(y))| > \epsilon\right)$$

where $a_n = (16K_1K_4/\epsilon h)^{1/\gamma}$ and $K_4 = 2^\gamma \int |x|^\gamma f(x)dx$. Let $b_n = \epsilon h^2/8C_1$. We cover the interval $[-a_n, a_n]$ by ... $[2a_n/b_n] + 1$ compact intervals, say $J_{n,1}, \ldots, J_{n,k_n'}$. $[x]$ denotes the largest integer $\le x$. Now repeating the argumen... page 186 in Rao (1983) we can show that

$$P\left(\sup_{|y|\le a_n} |\hat{f}_{n,h}(y) - E(\hat{f}_{n,h}(y))| > \epsilon\right) \le \sum_{i=1}^{k_n'} P\left(|\hat{f}_{n,h}(y_i) - E(\hat{f}_{n,h}(y_i))| > \epsilon/2\right),$$

where $y_i \in J_{n,i}$, $i = 1, \ldots, k_n'$. Since $h \in \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$, using (3.9) we get

$$\begin{aligned} P\left(\sup_{|y|\le a_n} |\hat{f}_{n,h}(y) - E(\hat{f}_{n,h}(y))| > \epsilon\right) &= O\left(k_n' \exp\left(-C\left(\frac{n}{\log(n)}\right)^{\frac{3}{5}}\epsilon^2\right)\right) \\ &= O\left((n/\log(n))^{(2+1/\gamma)} \exp\left(-C\left(\frac{n}{\log(n)}\right)^{\frac{3}{5}}\epsilon^2\right)\right). \end{aligned}$$

Now repeating the arguments in the page 187 of Rao (1983) (see the proof of Theorem 3.1.7) and the Bernstein type inequality by Merlev'ede et al. (2009) for strongly mixing processes we show that for $h \in \left[a\left(\frac{\log(n)}{n}\right)^{1/5}, b\left(\frac{\log(n)}{n}\right)^{1/5}\right]$, there exists N_ϵ such that for $n > N_\epsilon$

$$P\left(\sup_{|y|>a_n} |\hat{f}_{n,h}(y) - E(\hat{f}_{n,h}(y))| > \epsilon\right) \leq P\left(\sup_{|y|>a_n} |\hat{f}_{n,h}(y)| > \epsilon/2\right) = O\left(\exp\left(-C\left(\frac{n}{\log(n)}\right)^{\frac{1}{5}}\epsilon^2\right)\right). \tag{3.12}$$

Lemma 5 follows from (3.10)–(3.12). □

Acknowledgments

The authors are thankful to Prof. J.S. Marron for his detailed remarks on several issues related to density estimation. We are thankful to the esteemed reviewer for the detailed review and the kind suggestions which lead to significant improvement of the manuscript.

The first author research supported by UGC minor research project F. No. 39-938/2010 (SR).

References

Bose, A., Dutta, S., 2013. Density estimation using bootstrap bandwidth selector. Statist. Probab. Lett. 83, 245–256.
Bowman, A.W., 1984. An alternative method of cross-validation for smoothing of density estimates. Biometrika 71, 353–360.
Cao, R., Cuevas, A., Gonzlez-Manteiga, W., 1994. A comparative study of several smoothing methods in density estimation. Comput. Statist. Data Anal. 17, 153–176.
Devroye, L., 1989. The double kernel method in density estimation. Ann. Inst. Henri Poincare 25, 533–580.
Dutta, S., 2014. Local smoothing using the bootstrap. Comm. Statist. Simulation Comput. 43 (2), 378–389.
Einmahl, U., Mason, D., 2005. Uniform in bandwidth consistency of kernel-type function estimators. Ann. Statist. 33, 1380–1403.
Habbema, J.D.F., Hermans, J., Van Der Broek, K., 1974. A stepwise discrimination analysis program using density estimation. In: Bruckmann, G. (Ed.), Compstat: Proceedings in Computational Statistics, pp. 101–110.
Hall, P., Marron, J.S., Park, B.U., 1992. Smoothed cross-validation. Probab. Theory Related Fields 92, 1–20.
Hazelton, M., 1996. Bandwidth selection for local density estimators. J. Scand. Statist. 23, 221–232.
Hazelton, M., 1999. An optimal local bandwidth selector for kernel density estimation. J. Statist. Plann. Inference 77, 37–50.
Jones, M.C., Marron, J.S., Park, B.U., 1991. A simple root n bandwidth selector. Ann. Statist. 19, 1919–1932.
Krieger, A.M., Pickands, J., 1981. Weak convergence and efficient estimation at a point. Ann. Statist. 9, 1066–1076.
Lardjane, S., 2007. Nonparametric density estimation for nonmixing approximable stochastic processes. Stat. Inference Stoch. Process. 10, 209–221.
Merlev'ede, F., Peligrad, M., Rio, E., 2009. Bernstein inequality and moderate deviations under strong mixing conditions. IMS Collect. High Dimens. Probab. V, 273–292.
Mielniczuk, J., 1990. Remark concerning data-dependent bandwidth choice in density estimation. Statist. Probab. Lett. 9, 27–33.
Park, B.U., Marron, J.S., 1990. Comparison of data-driven bandwidth selectors. J. Amer. Statist. Assoc. 85, 66–72.
Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Statist. 33, 1065–1076.
Rao, B.L.S.P., 1983. Nonparametric Functional Estimation. Academic Press, New York.
Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832–835.
Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. J. Amer. Statist. Assoc. 82, 1131–1146.
Serfling, R., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.
Stone, C.J., 1984. An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist. 12, 1285–1297.
Talagrand, M., 1994. Sharper bounds for Gaussian and empirical processes. Ann. Probab. 22, 28–76.
Wang, X., Hu, S., Yang, W., 2011. The Bahadur representation for sample quantiles under strongly mixing sequence. J. Statist. Plann. Inference 141, 662–666.
Wied, D., Weißbach, R., 2010. Consistency of the kernel density estimator—a survey. Statist. Papers 53, 1–21.

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com

Local Smoothing for Kernel Distribution Fu Estimation

Santanu Dutta

To cite this article: Santanu Dutta (2015) Local Smoothing for Kernel Distributio Estimation, Communications in Statistics - Simulation and Computation, 44:4, 2 10.1080/03610918.2013.795591

To link to this article: http://dx.doi.org/10.1080/03610918.2013.795591

Accepted author version posted online: 02 Jun 2014.
Published online: 02 Jun 2014.

Communications in Statistics—Simulation and Computation®, 44: 878–891, 2015

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2013.795591

Taylor & Francis

Local Smoothing for Kernel Distribution Function Estimation

SANTANU DUTTA

Mathematical Sciences Department, Tezpur University, Napaam, INDIA

The problem of bandwidth selection for kernel-based estimation of the distribution function (cdf) at a given point is considered. With appropriate bandwidth, a kernel-based estimator (kdf) is known to outperform the empirical distribution function. However, such a bandwidth is unknown in practice. In pointwise estimation, the appropriate bandwidth depends on the point where the function is estimated. The existing smoothing methods use one common bandwidth to estimate the cdf. The accuracy of the resulting estimates varies substantially depending on the cdf and the point where it is estimated. We propose to select bandwidth by minimizing a bootstrap estimator of the MSE of the kdf. The resulting estimator performs reliably, irrespective of where the cdf is estimated. It is shown to be consistent under i.i.d. as well as strongly mixing dependence assumption. Two applications of the proposed estimator are shown in finance and seismology. We report a dataset on the S & P Nifty index values.

Keywords Bootstrap; Distribution function estimation; Kernel-based estimator.

Mathematics Subject Classification 62G05

1. Introduction

Let $X_1, \ldots, X_n$ be n continuous random variables with common distribution function F and density f. We consider the problem of estimating F (and also the survival function or the probability of exceedance) at a given design point x_0. Distribution function estimation finds application in survival analysis (see, for instance, Swanepoel and Graan, 2005; Liu and Yang, 2008). The estimation of F or the survival function appears as a natural problem in several other contexts as well. For example, in climatological studies, for a high value c, the relevance of knowing the probability of occurrence of a wind speed bigger than c is obvious. Similar estimation problems also arise in finance and seismology. We give two such examples in this article. Del Río and Estévez-Pérez (2012) contained a detailed literature review on the applications of estimation of F and related functions, such as the probability of exceedance.

A simple nonparametric estimator of F is the empirical distribution function (we call it F_n). The asymptotic properties of F_n are well-known (e.g., see Serfling, 1980). However, there are some compelling reasons for considering a kernel-based distribution function

Received October 23, 2012; Accepted April 8, 2013
Address correspondence to Santanu Dutta, Mathematical Sciences Department, Tezpur University, Napaam, 784028, India; E-mail: tezpur1976@gmail.com
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

estimator (we call it $\hat{F}_n$). For instance, under i.i.d. assumption, Reiss (1981) proved that $F_n(x)$ is asymptotically deficient, with respect to $\hat{F}_n(x)$. The relative deficiency of the empirical estimator in comparison to the kernel-based estimator has also been established by Falk (1983). Moreover, simulations in Azzalini (1981) reveal that the MSE of $\hat{F}_n(x)$ can be much lower than that of $F_n(x)$. Another simple reason for considering a kernel-based estimate of a continuous cdf is that the resulting estimator is also a continuous distribution function.

Kernel-based methods have been widely popular in the context of nonparametric functional estimation (see, for example, Rao, 1983). For an absolutely continuous distribution, a kernel-based estimator of $F(x_0)$ is obtained by integrating a kernel density estimator up to x_0. A kernel-based estimator crucially depends on the bandwidth (say h) which controls the smoothness. The optimum h for estimating F is known to be of the order $n^{-1/3}$ (see, for instance, Azzalini, 1981; Jin and Shao, 1999). The main problem in kernel-based estimation is to specify a suitable bandwidth based on the data. Sarda (1993) proposed a "leave-one-out" method, and Altman and Leger (1995) suggested a simple plug-in normal reference bandwidth selector. Bowman et al. (1998) proposed a cross-validation method. Polanski and Baker (2000) developed an iterative method for calculating the optimal plug-in bandwidth. Del Río and Estévez-Pérez (2012) have developed a package kerdiest for implementing these bandwidth selectors using the statistical software R. Swanepoel and Graan (2005) introduced a new approach based on non-parametric transformation of the data and discussed the problem of bandwidth selection for their estimator.

The above mentioned bandwidth selectors aim to minimize some "global" or overall measure of discrepancy between the estimator and F. Consequently, the resulting estimators use the same amount of smoothing to estimate F at any point. However from Azzalini (1981) we see that while estimating $F(x_0)$ the appropriate h can vary widely, depending on whether x_0 is close to the boundary or in the tail region. Therefore, while estimating F pointwise, the existing global bandwidth selectors can perform poorly for certain choice of x_0 (we observe this in the simulation study). The problem of local bandwidth selection for pointwise estimation of F has received far less attention. Azzalini (1981) proposed to use $h = 0.5\sigma n^{-1/3}$ and $h = 1.3\sigma n^{-1/3}$ in the kernel cdf estimator for x_0 close to the boundary and in the long tail of f, respectively, where σ is the standard deviation. These seem to be the only available random bandwidths which are proposed based on the location of x_0. But the resulting estimates are drastically poor in comparison to the empirical estimate, for F equal to the cdf of the Gamma(1/2) distribution and x_0 equal to the 5th percentile. In fact, in the context of pointwise cdf estimation, we have not come across a bandwidth selector which automatically adjusts the amount of smoothing with change in x_0. So there seems to be a lot of scope for development of algorithms for data-based bandwidth selection for estimating $F(x_0)$ using kernel. The bootstrap method has been quite successful in the context of smoothing a kernel density estimator (see Dutta 2012a, Bose and Dutta 2013, and references therein, among most recent). Surprisingly no bootstrap based bandwidth selector seems to have been developed for estimation of F. We propose a bandwidth selector based on the minimization of a bootstrap estimate of the MSE of $\hat{F}_n(x_0)$.

Another aspect of our work is to investigate the consistency of the estimators using the proposed bandwidth and the bandwidths in Azzalini (1981) in the presence of dependence. Several well-known datasets seem to exhibit substantial dependence. For example, we find that waiting times between eruptions and the durations of the eruptions for the Old Faithful geyser in Yellowstone National Park Certain datasets exhibit autoregressive dependence (see Fig. 1). Most of the existing bandwidth selectors for distribution function estimators assume the data to be realizations of i.i.d. random variables. The proposed estimator is shown

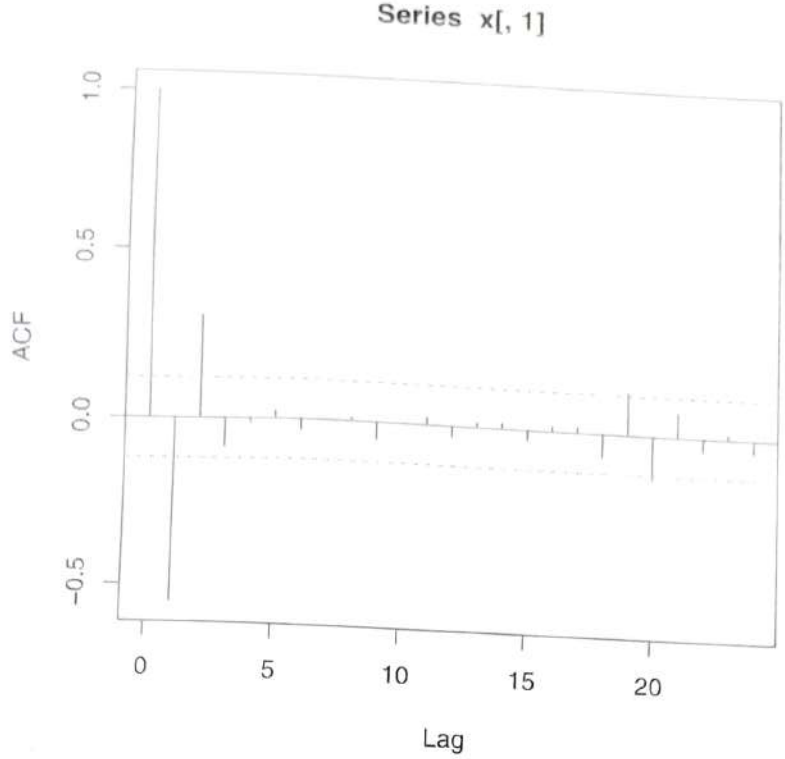


Figure 1. The auto correlation function plots for the eruption duration and waiting time values of Old Faithful data.

to work consistently under strongly mixing as well as i.i.d. assumption. We also study the performance of the Azzalini's (1981) estimators under similar dependence assumption.

In Section 2, we provide the definitions and the details of our proposal. The exact MSE of a kernel estimator using a random bandwidth is hard to obtain. However given F, one can approximate the MSE of a statistic by Monte-Carlo (MC) simulation. In Section 3, we compare the MC estimates of the MSE of a number of nonparametric estimators of $F(x_0)$, for different test distributions. We analyze two real datasets as well. One dataset consists of magnitudes of 1000 Fiji earthquakes, and the other dataset contains the annual (log) return vales of the S& P CNX Nifty, an index of the National Stock Exchange (NSE) in India, for 18 financial years from 1994–95 to 2011–12. While the data on Fiji quakes are well-known and available in standard software packages such as R, the data on the Nifty index values on the first and last trading days of each financial year (starting in April and ending in March) are collected from the NSE Web site. These data are reported in our Table 3. We show that the problem of estimation of the probability of exceedance arise naturally in the analysis of both these datasets, and apply the proposed methodology.

A simulation experiment is restricted to comparisons based on a finite number of test distributions. So in Section 4, we prove the consistency of the estimator of $F(x_0)$ using the proposed bandwidth selector under i.i.d. as well as strongly mixing dependence assumptions. We also show that the estimators using Azzalini's (1981) bandwidths are also consistent under strongly mixing dependence. These results seem to be new. In general, the simulations in Section 3 suggest the following observations.

Table 1

Monte Carlo estimates of n.MSE of the empr estimator and $\hat{F}_n(x_0, h)$ for h equal to the different random bandwidths and the optimal bandwidth in Azzalini (1981), for $n = 100$, x_0 equal to the 5th or 95th percentile

Density	x_0	$\hat{h}$	Azza[1]	Azza[2]	empr	AL	PB	CV	h_{opt}
N(0,1)	5th percentile	0.042	0.044	0.043	0.0475	0.032	0.044	0.164	0.044
	95th percentile	0.042	0.042	0.039	0.0475	0.049	0.064	0.206	0.042
Gamma(1/2)	5th percentile	0.043	1.972	8.912	0.0475	0.685	0.697	0.243	0.045
	95th percentile	0.044	0.046	0.045	0.0475	0.030	0.030	0.056	0.044
Gamma(1)	5th percentile	0.044	0.035	0.718	0.0475	0.081	0.173	0.262	0.168
	95th percentile	0.045	0.046	0.045	0.0475	0.041	0.040	0.065	0.042
Gamma(2)	5th percentile	0.044	0.039	0.092	0.0475	0.045	0.099	0.115	0.039
	95th percentile	0.043	0.046	0.044	0.0475	0.058	0.058	0.110	0.043
Gamma(5)	5th percentile	0.043	0.043	0.051	0.0475	0.038	0.067	0.078	0.044
	95th percentile	0.044	0.046	0.043	0.0475	0.034	0.030	0.063	0.042
Beta(1/2,1/2)	5th percentile	0.043	0.023	0.304	0.0475	0.059	0.139	0.231	0.041
	95th percentile	0.044	0.022	0.285	0.0475	0.051	0.126	0.231	0.038
Beta(1,1)	5th percentile	0.033	0.031	0.039	0.0475	0.037	0.072	0.128	NA
	95th percentile	0.038	0.039	0.029	0.0475	0.021	0.052	0.102	NA
Beta(1,4)	5th percentile	0.042	0.03	0.192	0.0475	0.042	0.146	0.795	0.359
	95th percentile	0.041	0.045	0.043	0.0475	0.045	0.043	0.054	0.042
Beta(2,5)	5th percentile	0.044	0.043	0.047	0.0475	0.034	0.070	0.260	0.042
	95th percentile	0.045	0.045	0.043	0.0475	0.045	0.045	0.073	0.043
Beta(2,10)	5th percentile	0.039	0.042	0.047	0.0475	0.055	0.106	0.540	0.036
	95th percentile	0.041	0.045	0.042	0.0475	0.043	0.043	0.072	0.043
Beta(5,5)	5th percentile	0.043	0.044	0.045	0.0475	0.043	0.063	0.145	0.040
	95th percentile	0.044	0.044	0.045	0.0475	0.034	0.043	0.134	0.042
AR(1), $\phi = 0.2$	5th percentile	0.058	0.061	0.058	0.065	0.058	0.064	0.175	0.046
	95th percentile	0.055	0.057	0.054	0.060	0.066	0.075	0.187	0.051
AR(1), $\phi = 0.5$	5th percentile	0.073	0.074	0.070	0.078	0.425	0.383	0.163	0.079
	95th percentile	0.079	0.080	0.078	0.083	0.374	0.302	0.183	0.08
AR(1), $\phi = 0.8$	5th percentile	0.189	0.189	0.186	0.193	0.611	0.681	0.326	0.194
	95th percentile	0.209	0.209	0.208	0.211	0.567	0.671	0.276	0.212

While the accuracy of the estimators using the existing random bandwidths can vary substantially from one example to another, the proposed estimator seems to perform reliably in a wide variety of examples. In the presence of autoregressive dependence, the accuracy of all the estimators deteriorate with increase in the extent of auto-correlation. But the proposed estimator continues to outperform the empirical estimator and compares well with the other kernel-based estimators even in the presence of substantial autocorrelation. The global bandwidths proposed by Altman and Leger (1995), Bowman et al. (1998), and Polanski and Baker (2000) perform reasonably while estimating F in the inter-quartile range. But these global bandwidth selectors do not appear to be suitable for estimation of $F(x_0)$, for x_0 in the tail region or close to the boundary of the support. In such cases, the proposed estimator performs more reliably than the other estimators. The existing estimators may perform poorly if the density is not bounded in a neighborhood of the estimation point. The proposed estimator can be used safely even in a such case.

2. Definitions and the Proposal

Let $X_1, \ldots, X_n$ be n identically distributed random variables with common distribution function F. The empirical distribution function F_n is defined as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x - X_i)$, where $I(x - X_i) = 1$ for $X_i \leq x$ and zero otherwise.

Table 2

Monte Carlo estimates of n.MSE of the empr estimator and $\hat{F}_n(x_0, h)$ for h equal to the different random bandwidths and the optimal bandwidth in Azzalini (1981), for $n = 100$. x_0 is the first, second, or the third quartile

Density	x_0	$\hat{h}$	Azza[1]	Azza[2]	empr	AL	PB	CV	h_{opt}
N(0,1)	25th percentile	0.169	0.177	0.167	0.1875	0.177	0.174	0.21	0.161
	50th percentile	0.221	0.227	0.220	0.25	0.228	0.212	0.24	0.161
	95th percentile	0.184	0.189	0.183	0.1875	0.183	0.184	0.19	0.173
Gamma(1/2)	25th percentile	0.187	0.67	0.25	0.1875	0.193	0.219	0.239	0.158
	50th percentile	0.238	0.239	0.46	0.25	0.236	0.237	0.313	0.234
	75th percentile	0.187	0.190	0.196	0.1875	0.194	0.220	0.189	0.172
Gamma(1)	25th percentile	0.178	0.175	0.181	0.1875	0.156	0.162	0.174	0.145
	50th percentile	0.253	0.252	0.267	0.25	0.251	0.254	0.261	0.228
	75th percentile	0.171	0.173	0.170	0.1875	0.181	0.178	0.191	0.177
Gamma(2)	25th percentile	0.164	0.170	0.152	0.1875	0.169	0.162	0.171	1.54
	50th percentile	0.228	0.233	0.225	0.25	0.248	0.236	0.235	0.216
	75th percentile	0.181	0.189	0.191	0.1875	0.182	0.183	0.185	0.174
Gamma(5)	25th percentile	0.177	0.180	0.173	0.1875	0.168	0.158	0.179	0.153
	50th percentile	0.259	0.252	0.445	0.25	0.248	0.236	0.31	0.21
	75th percentile	0.175	0.174	0.177	0.1875	0.171	0.173	0.191	0.158
Beta(1/2,1/2)	25th percentile	0.190	0.194	0.187	0.1875	0.191	0.186	0.21	0.170
	50th percentile	0.259	0.263	0.260	0.25	0.248	0.236	0.251	3×10^{-20}
	75th percentile	0.157	0.161	0.154	0.1875	0.184	0.178	0.183	0.166
Beta(1,1)	25th percentile	0.176	0.179	0.174	0.1875	0.176	0.170	0.178	NA
	50th percentile	0.249	0.256	0.249	0.25	0.24	0.234	0.241	NA
	75th percentile	0.180	0.185	0.179	0.1875	0.160	0.152	0.179	NA
Beta(1,4)	25th percentile	0.155	0.163	0.154	0.1875	0.042	0.146	0.178	0.161
	50th percentile	0.273	0.282	0.269	0.25	0.259	0.254	0.253	0.216
	75th percentile	0.183	0.187	0.184	0.1875	0.185	0.185	0.183	0.168
Beta(2,5)	25th percentile	0.180	0.186	0.178	0.1875	0.169	0.160	0.171	0.150
	50th percentile	0.243	0.250	0.239	0.25	0.236	0.226	0.241	0.213
	75th percentile	0.181	0.186	0.181	0.1875	0.184	0.178	0.185	0.174
Beta(2,10)	25th percentile	0.173	0.179	0.171	0.1875	0.172	0.160	0.173	0.155
	50th percentile	0.216	0.222	0.217	0.25	0.276	0.268	0.31	0.21
	75th percentile	0.201	0.203	0.201	0.1875	0.185	0.180	0.188	0.169
Beta(5,5)	25th percentile	0.166	0.167	0.164	0.1875	0.157	0.149	0.161	0.148
	50th percentile	0.24	0.246	0.236	0.25	0.262	0.245	0.258	4×10^{-29}
	75th percentile	0.175	0.177	0.171	0.1875	0.148	0.143	0.178	0.167
AR(1), $\phi = 0.2$	25th percentile	0.198	0.219	0.208	0.225	0.201	0.197	0.191	0.189
	50th percentile	0.272	0.305	0.288	0.312	0.242	0.234	0.239	NA
	75th percentile	0.221	0.229	0.216	0.238	0.199	0.198	0.189	0.190
AR(1), $\phi = 0.5$	25th percentile	0.399	0.401	0.391	0.411	0.358	0.369	0.455	0.37
	50th percentile	0.593	0.607	0.585	0.625	0.566	0.513	0.593	NA
	75th percentile	0.398	0.406	0.388	0.418	0.366	0.356	0.532	0.346
AR(1), $\phi = 0.8$	25th percentile	1.06	1.08	1.061	1.095	1.058	0.991	1.04	0.97
	50th percentile	1.569	1.590	1.558	1.596	1.466	1.413	1.493	NA
	75th percentile	1.051	1.061	1.042	1.073	1.052	0.977	1.811	0.951

A kernel-based estimator $\hat{F}_n(x_0)$ of $F(x)$ is defined as

$$\hat{F}_n(x) \equiv \hat{F}_n(x, h) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{h}\right)$$

Table 3

Nifty Index values on first and last trading days of 18 financial years, from 1994–95 to 2011–2012. Source: www.nseindia.com

Year	Date	Index closing value
1994–95	4 April 1994	1182.33
	31 March 1995	990.24
1995–96	3 April 1995	1005.37
	29 March 1996	985.3
1996–97	1 April 1996	994.8
	31 March 1997	968.3
1997–98	1 April 1997	970.15
	31 March 1998	1116.9
1998–99	1 April 1998	1150.1
	31 March 1999	1078.05
1999–00	1 April 1999	1063.45
	31 March 2000	1528.45
2000–01	3 April 2000	1534.75
	30 March 2001	1148.2
2001–02	2 April 2001	1138.1
	28 March 2002	1129.55
2002–03	1 April 2002	1138.95
	31 March 2003	978.2
2003–04	1 April 2003	984.3
	31 March 2004	1771.9
2004–05	1 April 2004	1819.65
	31 March 2005	2035.65
2005–06	1 April 2005	2067.65
	31 March 2006	3402.55
2006–07	3 April 2006	3473.3
	30 March 2007	3821.55
2007–08	2 April 2007	3633.6
	31 March 2008	4734.5
2008–09	1 April 2008	4739.55
	31 March 2009	3020.95
2009–10	1 April 2009	3060.35
	31 March 2010	5249.1
2010–11	1 April 2010	5290.5
	31 March 2011	5833.75
2011–12	1 April 2011	5826.05
	30 March 2012	5295.55

where K is a distribution function with density k, and $h \equiv h_n$ is a positive sequence satisfying $h = o(1)$, $nh \to \infty$ as $n \to \infty$. Swanepoel and Graan (2005) contains a rich literature review on the asymptotic properties of $\hat{F}_n$.

It is easy to see that the MSE of $\hat{F}_n(x_0)$ equals

$$\begin{aligned} \mathrm{MSE}(h) &= \frac{1}{n}\mathrm{Var}\left[K\left\{(x_0 - X_1)/h\right\}\right] + \left[E\left\{K\left(\frac{x_0 - X_1}{h}\right)\right\} - F(x_0)\right]^2 \\ &= r_{1n} + r_{2n}, \text{ where} \end{aligned} \tag{2.1}$$

$$r_{1n} = \frac{1}{n}\left[\int K^2((x_0-u)/h)\,dF(u) - \left\{\int K((x_0-u)/h)\,dF(u)\right\}^2\right]$$

$$r_{2n} = \left[\int K((x_0-u)/h)\,dF(u) - F(x_0)\right]^2.$$

So the MSE(h) is a functional of the distribution function of X_1, viz. F. In practice, F is not known.

Given $X_1, \ldots, X_n$, a bootstrap approximation to MSE (h) can be defined by replacing the unknown F by the empirical distribution function F_n, in the right-hand side of (2.1). We define a bootstrap MSE estimator MSE$^*(h)$ as follows:

$$\text{MSE}^*(h) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n} K^2\{(x_0 - X_i)/h\} - \{\hat{F}_n(x_0, h)\}^2\right] + \left[\hat{F}_n(x_0, h) - F_n(x_0)\right]^2 \quad (2.2)$$

In terms of re-sampling the above bootstrap proposal can be interpreted as a re-sampling scheme where each re-sample is generated by simple random sampling (with replacement) from the empirical distribution. In practice, no re-sampling is required to implement our proposal.

It is well known that an optimum h, asymptotically minimizing the MSE(h), is a multiple of $n^{-1/3}$ (see, for instance, Azzalini, 1981; Jin and Shao, 1999). Therefore, without loss in generality, we restrict the search for an optimum h in a compact interval I_n, with the boundary points equal to some multiple of $n^{-1/3}$. Let

$$\hat{h} = \text{argmin}_{h\in I_n}\text{MSE}^*(h), \text{ where } I_n = [c_1 n^{-1/3},\ c_2 n^{-1/3}], \quad (2.3)$$

The boundary points of I_n are chosen to be scale-invariant bandwidths. We choose c_1, c_2 in such a way that for $h \in I$, $\log_{10}(h)$ varies over a broad range. From Azzalini (1981) we see that for a broad class of distributions, the multiplier of $n^{-1/3}$ in the optimal bandwidth varies between σ to 2σ, when x_0 is a point in the right tail of F. If x_0 is not in the long tail, $0.5\sigma n^{-1/3}$ is the more appropriate value of h. σ denotes the standard deviation. Motivated by these observations we use $c_1 = 0.2\hat{\sigma}$ and $c_1 = 2\hat{\sigma}$, where $\hat{\sigma}$ is the sample standard deviation.

The survival function or the risk function or the probability of exceedence, at x_0, is defined as

$$S(x_0) = 1 - F(x_0).$$

Therefore given an estimator $\hat{F}(x_0)$ of $F(x_0)$, a natural estimator $\hat{S}(x_0)$ of $S(x_0)$ equals $1-\hat{F}(x_0)$. Clearly $\hat{F}(x_0)$ and $\hat{S}(x_0)$ have the same MSE. Consequently, $\hat{S}(x_0) = 1-\hat{F}_n(x_0, \hat{h})$ is the proposed estimator of $S(x_0)$, where $\hat{h}$ is as defined in (2.3).

3. Simulation and Data Analysis

We compare the values of n times the MSE of the empirical estimator $F_n(x_0)$ with that of $\hat{F}_n(x_0, h)$ using h equal to $\hat{h}$ in (2.3), and the random bandwidths proposed by Azzalini (1981), Altman and Leger (1995), Bowman et al. (1998) and Polanski and Baker (2000) based on i.i.d. observations from 11 test distributions and for x_0 equal to the 5th, 25th, 50th, 75th and 95th percentiles. While the 5th and the 95th percentile values are in the tail region, the other percentile values are in the inter quartile range. We know that the optimal

bandwidth asymptotically minimizing the MSE of a kernel-based estimator is of the form $h_{opt} = (u \cdot 4v)^{1/3} n^{-1/3}$, where u, v are as mentioned in Azzalini (1981). In practice u, v are unknown. But in a simulation study we can compute h_{opt} explicitly. We also compare the values of n times the MSE of the proposed estimator with that of the estimator $\hat{F}_n(x_0, h_{opt})$ using the optimal bandwidth (we are thankful to the esteemed reviewer for this suggestion).

In order to study the performance of the above mentioned estimators in presence of dependence, we compare the values of n times the MSE of these estimators based on data generated by the AR(1) process described below.

$$X_t = \phi X_{t-1} + \sqrt{1-\phi^2} Z_t, \quad t = 2, 3, 4, \ldots,$$

where $\{Z_t\}_{t=1,2,\ldots}$ is an i.i.d. process. Z_1, X_1 follow $N(0, 1)$ distribution. Also $\{Z_t\}$ is independent of X_1. We consider three different values of ϕ, viz. $\phi = 0.2, 0.5, 0.8$.

The above experiment is repeated for $n = 100, 1000$. It is difficult to accommodate all the simulation results in one Table. The values of n times the MSE of all the estimators are provided in Table 1, for x_0 equal to the 5th and the 95th percentile and $n = 100$. In Table 2, we report these values for x_0 equal to the 25th, 50th and 75th percentile values. The values of n times the MSE for $n = 100$ and 1000 are similar. Under i.i.d. assumption the exact MSE of the empirical estimator $F_n(x_0)$ equals $p(1-p)/n$, where x_0 is the pth percentile. The exact MSE of a kernel estimator using a random h is hard to obtain. We approximate the same by Monte Carlo (MC) method. To compute the MC estimate of the MSE of a statistic we generate 5000 samples of a specific size from a particular model. The average of the squared deviations of the values of the statistic from the exact parameter is the MC estimate of its MSE. Under auto-regressive dependence the MSE of even the empirical estimator is approximated by MC method.

Let us introduce some notations. Let $N(0, 1)$, Gamma(l) and Beta(l_1, l_2) denote the standard normal distribution, the gamma distribution with parameter l, and the beta distribution of the 1st kind with parameters l_1, l_2. The different choices of the parameters in the gamma and beta distributions, used in the simulations, are mentioned in Table 1. In Table 1, Azza[1], Azza[2], AL, PB and CV represent the random bandwidths proposed by Azzalini (1981), Altman and Leger (1995), Polanski and Baker (2000) and the cross-validation bandwidth by Bowman et al. (1998). AL[1] $= 0.5\sigma n^{-1/3}$ and AL[2]$=1.3\sigma n^{-1/3}$, where σ is the sample standard deviation. "empr" is an abbreviation for the empirical estimator. As mentioned above, h_{opt} represents the MSE optimal bandwidth in Azzalini (1981). We note that if $f^{(1)}(x) = 0$, h_{opt} is not defined. In that case we write "NA" corresponding to the values of n times the MSE of $\hat{F}_n(x_0, h_{opt})$ in Tables 1 and 2. Throughout these simulations we use Epanechnikov kernel. From Table 1, we have the following main observations.

1. The estimator $\hat{F}_n(x_0, \hat{h})$, using $\hat{h}$ in (2.3), outperforms the empirical estimator $F_n(x_0)$ for most of the choices of F and x_0.
2. The estimators using the global bandwidth selectors, such as the AL, PB, and CV selectors, perform reasonably well in estimating $F(x_0)$ for x_0 in the interquartile (see Table 2). But these estimators seem to struggle to estimate F in the tail region or at a point close to the boundary (see Table 1). If f is not bounded in a neighborhood of x_0, these estimators seem to struggle. In particular for x_0 equal to the 5th percentile of the Gamma(1/2), Gamma(1) and Beta(1/2, 1/2) distributions, the MSE of kernel-based estimators using Azza[2], AL, PB and CV bandwidths are much larger than the MSE of the empirical estimator. In contrast, the performance of the proposed

estimator $\hat{F}_n(x_0, \hat{h})$ does not deteriorate drastically, in comparison to the other estimators, for any choice of F or x_0.

3. In the presence of auto-regressive dependence, the accuracy of all the estimators deteriorate as ϕ is increased. Even in the presence of substantial auto regressive dependence, the proposed estimator outperforms the empirical estimator and compares well with the other kernel-based estimators, especially while estimating F in the tail region or at a point close to the boundary of the support (see Table 1).
4. From Table 1, we see that the MSE of the proposed estimator compares well with the MSE of $\hat{F}_n(x_0, h_{\text{opt}})$, based on the ideal bandwidth h_{opt}, for x_0 in the tail region and also in the presence of substantial autoregressive dependence. In fact for 15 combinations of F and x_0 in Table 1, the MSE of the proposed estimator is less than or equal to that for the estimator based on h_{opt}.
5. The cross-validation method by Bowman et al. (1998) is computationally expensive and in most of the examples its MSE is larger than the MSE of the empirical estimator and the kernel-based estimators using h equal to $\hat{h}$ in (2.3) and Azza[1].

Based on the above observations, we recommended $\hat{F}_n(x_0, \hat{h})$ for estimating $F(x_0)$, especially when x_0 is a point in the tail region or close to the boundary of the support. Unlike the other kernel-based estimators, it can used safely even if f is not bounded at x_0.

3.1 *Analysis of Real Data*

1. **Fiji earthquake magnitude**. A well-known dataset consists of observations on 1000 earthquakes in Fiji since 1964. This dataset is available in the package "quakes" in the software R for statistical computing. Quakes of magnitude up to 4.9 on Richter Scale are considered as slight and are negligible (see classification of quakes in http://www.imd.gov.in/section/seismo/static/earthquake-terminology.htm). So it is of natural interest to estimate the probability of occurrence of an earthquake of magnitude exceeding 5 on Richter scale. For the Fiji data, the empirical estimate of this probability of exceedance equals 0.14.

   However, the kernel-based estimate of this probability equals 0.176 for h equal to $\hat{h}$ in (2.3). The kernel-based estimates using the other random bandwidths mentioned above are also similar. They vary in the range 0.175–0.177. So the kernel-based methods assign more probability than the empirical distribution to the event of occurrence of a quake of magnitude more than 5 in Fiji.

   Since the estimator using the proposed bandwidth $\hat{h}$ in (2.3) seems to perform reliably, especially in estimating extreme probabilities, we conclude that the chance of an earthquake of magnitude exceeding 5 in Fiji is between 17 and 18%.
2. **S & P NIFTY annual return**. The S& P CNX Nifty is a well diversified 50 stock index accounting for 22 sectors of the Indian economy. It is used for a variety of purposes such as benchmarking fund portfolios (see www.nseindia.com for details). For investors in the Indian equity market, the relevance of knowing the chance of annual return of this index exceeding some high value (say 10%) is obvious. In India, a financial year starts on 1 April and ends on the 31 March of the next year. In Table 3, we report the closing values of the Nifty index on the first trading and the last trading days for the 18 consecutive financial years from 1994–95 to 2011–12 (source: http://www.nseindia.com/products/content/equities/indices).

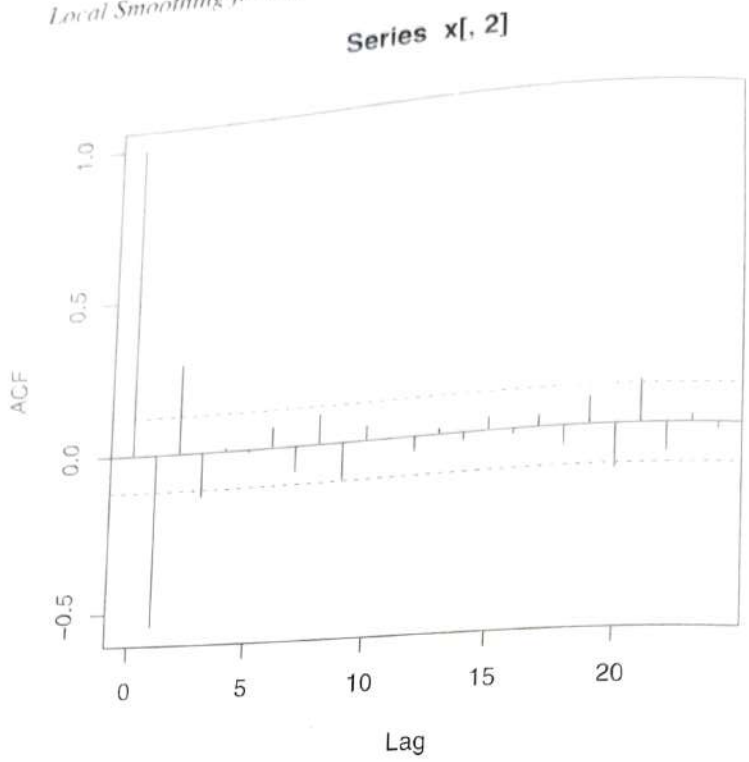


Figure 2. Plot of acf for the NIFTY annual log return values from 1994–95 to 2011–2012.

The annual log return for a financial year is equal to the logarithm of the ratio of the index values on the last and first trading days of that financial year. We calculate the annual log return values for the 18 yr. In Fig. 2, we plot the auto correlation function (acf) based on these log return values.

There seems to be substantial first-order autoregressive dependence (see Fig. 2). The empirical estimate of the probability of the annual log return exceeding 10% equals 0.278. The kernel-based estimates of this probability of exceedance using $\hat{h}$ and Azzalini's bandwidths vary between 0.291 and 0.322. From simulations we see that the proposed estimator performs slightly better than the empirical estimator under substantial auto regressive dependence. So we conclude the chance that the annual Nifty log return exceeds 10% is close to 0.30.

We also compute the chance that the log annual return is less than -0.10 , i.e., an annual loss of more than 10%. The empirical estimate of this probability is 0.111, and the kernel-based estimate (using $\hat{h}$) of the same is close to 0.12. So the chance of the annual Nifty (log) return exceeding 10% seems to be much larger than the chance of a loss (in log scale) of 10% or more.

4. Asymptotic Properties

Let $\tilde{h} \equiv h(X_1, \ldots, X_n)$ be a random bandwidth (i.e., a function of the data) and $\hat{F}(x_0, \tilde{h})$ be the corresponding kernel-based estimator of $F(x_0)$. In this section, we prove the consistency of $\hat{F}(x_0, \tilde{h})$, for $\tilde{h}$ equal to the proposed bandwidth $\hat{h}$ in (2.3), and the random bandwidths in Azzalini (1981) under i.i.d. and strongly mixing conditions. We have the following theorems.

4.1 *i.i.d. Case*

Theorem 4.1. *Let* K *be a distribution function with density* k. *Let* $X_1, \ldots, X_n$ *be i.i.d. random variables with distribution function* F *and density* f. *If* $\tilde{h} + \frac{1}{n\tilde{h}} \to 0$, *completely/almost surely/ in probability,* $\hat{F}(x_0, \tilde{h}) \to F(x_0)$, *completely/almost surely/in probability.*

The following corollary is immediate.

Corollary 4.1. *Under the conditions stated in Theorem 4.1,* $\int x^2 dF(x) < \infty$ *and for* $\tilde{h} = c\hat{\sigma} n^{-1/3}$, $\hat{F}(x_0, \tilde{h})$ *converges almost surely to* $F(x_0)$, *where* c *is a positive constant and* $\hat{\sigma}$ *is the sample standard deviation.*

$c = 0.5$ *and* $c = 1.3$ *correspond to the random bandwidths proposed by Azzalini (1981). Hence under the conditions in Theorem 4.1 and* $\int x^2 dF(x) < \infty$, *the kernel-based estimators of* $F(x_0)$ *proposed by Azzalini (1981) are strongly consistent.*

Proof of Theorem 4.1. It is easy to see that for any arbitrary $\epsilon > 0$,

$$P\left(|\hat{F}_n(x_0, \tilde{h}) - F(x_0)| > \epsilon\right) \leq P\left(J > \epsilon\right),$$

where $J = \int |f_n(x, \tilde{h}) - f(x)|dx$, $f_n(x, \tilde{h}) = \frac{1}{n\tilde{h}} \sum_{i=1}^{n} k\left((y - X_i)/\tilde{h}\right)$ and k is the density function corresponding to the distribution function K.

Theorem 1 in Devroye and Gyorfi (1985), chapter 6, p. 148, states that if $f_n(\cdot, \tilde{h})$ is a kernel density estimator using a random bandwidth $\tilde{h}$ such that $\tilde{h} + \frac{1}{n\tilde{h}} \to 0$, completely/almost surely/in probability, $J \to 0$ completely/almost surely/in probability, for any density f on the real line. Hence our Theorem 4.1 is a direct consequence of the Theorem 1 in Devroye and Gyorfi (1985), chap. 6, p. 148.
This completes the proof. □

Next we obtain an asymptotic property of the proposed random bandwidth $\hat{h}$ in (2.3).

Lemma 4.1. *Let* $X_1, \ldots, X_n$ *be i.i.d. random variables with distribution function F, satisfying* $\int x^2 dF(x) < \infty$. *Then* $\hat{h} + \frac{1}{n\hat{h}} \to 0$ *almost surely.*

Proof. From the definition of $\hat{h}$ we see that

$$c_1 n^{-1/3} \leq \hat{h} \leq c_2 n^{-1/3} \text{ and } \frac{1}{c_2 n^{2/3}} \leq \frac{1}{n\hat{h}} \leq \frac{1}{c_1 n^{2/3}},$$

where c_1, c_2 are constant multiples of the sample standard deviation. So under the stated conditions, c_1, c_2 converge (almost surely) to positive constants as $n \to \infty$.

Using the above Lemma and Theorem 4.1 we get the following Theorem.

Theorem 4.2. *Let* K *be a distribution function with density* k. *Let* $X_1, \ldots, X_n$ *be i.i.d. random variables with distribution function* F *and* $\int x^2 dF(x) < \infty$. *Then for* $h = \hat{h}$ *in (2.3)*

$$\hat{F}_n(x_0, \hat{h}) \to F(x_0) \text{ almost surely, as } n \to \infty.$$

The above theorem ensures the almost sure convergence of the proposed estimator $\hat{F}_n(x_0, \tilde{h})$ to $F(x_0)$. Next, we prove the consistency of the proposed estimator and the estimators in Azzalini (1981) under strongly mixing dependence assumption.

4.2 *Strongly Mixing Case*

Suppose $\{X_t,\ t \in \mathbb{Z}\}$ is a $\mathbb{R}$ valued, strictly stationary process with marginal density f. Let $M_{-\infty}^{0}$ and M_n^{∞} denote σ-fields generated by $\{X_t,\ t \le 0\}$ and by $\{X_t, t \ge n\}$, respectively. Then X_t is a strong mixing process if

$$\alpha(n) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in M_n^{\infty},\ B \in M_{-\infty}^{0}\} \downarrow 0,\ \text{as } n \to \infty.$$

Let us collect some assumptions on $\alpha(n)$ to be used in the following.

Assumption 4.1. $n^3 m^{-2} (\log n)^{-1} [\alpha(m)]^{2\gamma/(2\gamma+1)} \to 0$ as $n \to \infty$ for some positive number γ and for some sequence of integers $m = m(n)$ with $m \to \infty$ and $m/n \to 0$ as $n \to \infty$.

The next Lemma is an extension of the Lemma 2 in Chapter 6 in Devroye and Gyorfi (1985) to the case of strongly mixing dependence.

Lemma 4.2. *Let* K *be any density function. Let us consider a sequence of intervals* $H_n = [h', h'']$ *such that* $h' = o(1)$ *and* $nh''m^{-2}(\log n)^{-1} \to \infty$. *Then under Assumption 1, for every* $\epsilon > 0$

$$P\left(\sup_{h \in H_n} \int |\hat{f}_n(y, h) - f(y)| dy > \epsilon\right) = o(1).$$

The proof of the above lemma is similar to the proof of the Lemma 2 in Chapter 6 in Devroye and Gyorfi (1985). The proof is available in Dutta (2012b), and can be obtained from the author.

Theorem 4.3. *Let* K *be a distribution function with density* k. *Let* $\{X_t,\ t \in \mathbb{Z}\}$ *be a strongly mixing stationary process with marginal distribution function* F *and density* f. *Suppose that Assumption 1 holds. If* $\tilde{h} + \frac{m^2 \log(n)}{n\tilde{h}} \to 0$ *in probability,* $\hat{F}(x_0, \tilde{h}) \to F(x_0)$ *in probability.*

Proof. For any $\epsilon > 0$,

$$P\left(|\hat{F}(x_0, \tilde{h}) - F(x_0)| > \epsilon\right) \le P\left(\int_{-\infty}^{\infty} \left|\hat{f}_n(y, \tilde{h}) - f(y)\right| dy > \epsilon\right). \tag{4.1}$$

Repeating the arguments in the proof of Theorem 1 in Chapter 6 of Devroye and Gyorfi (1985) p. 159, we see that for any random bandwidth $\tilde{h}$ satisfying $\tilde{h} + \frac{m^2 \log(n)}{n\tilde{h}} \to 0$ (in probability) there exists a sequence of $H_n \equiv \left[h_n', h_n''\right]$, where $h_n'' = o(1)$, $\frac{nh_n'}{m^2 \log(n)} \to \infty$ as $n \to \infty$ and

$$P\left(\int_{-\infty}^{\infty} \left|\hat{f}(y, \tilde{h}) - f(y)\right| dy > \epsilon\right) \le 2P\left(\tilde{h} + \frac{m^2 \log(n)}{n\tilde{h}} > \epsilon\right) + P\left(\sup_{h \in H_n} \int \left|\hat{f}(y, h) - f(y)\right| dy > \epsilon\right).$$

Therefore under the assumption $\tilde{h} + \frac{m^2 \log(n)}{n\tilde{h}} = o(1)$ (in probability), Theorem 4.3 is a consequence of Lemma 4.2. □

Remark 4.1. Let $\{X_t\}_{t\in Z}$ be a stationary, ergodic process such that $E|X_1| < \infty$. Birkhoff ergodic theorem ensures that the average of $\sum_{i=0}^{n-1} X_{t+i}/n$ converges almost surely to $E(X_1)$ as $n \to \infty$. A stationary strongly mixing process is a stationary ergodic process (see Rieders, 1993). Hence, if $\{X_t,\ t \in \mathbb{Z}\}$ be a strongly mixing stationary process with marginal distribution function F and $\int x^2 dF(x) < \infty$, the standard deviation of $X_1, \ldots, X_n$ converges (almost surely) to the standard deviation of the marginal distribution of X_1.

The following corollary is immediate.

Corollary 4.2. *Suppose that the conditions stated in Theorem 4.3 hold. Moreover, $\int x^2 dF(x) < \infty$ and $\tilde{h} = c\hat{\sigma} n^{-1/3}$. Then, $\hat{F}(x_0, \tilde{h})$ converges almost surely to $F(x_0)$, where c is a positive constant and $\hat{\sigma}$ is the sample standard deviation.*

The above corollary implies that estimators using the random bandwidths proposed by Azzalini (1981) remain consistent under the strongly mixing dependence assumption. Now we prove the consistency of our estimator under such dependence assumption.

Using arguments as in the proof of Lemma 4.1 and Remark 1 we have the following Lemma.

Lemma 4.3. *Let $\{X_t,\ t \in \mathbb{Z}\}$ is a strongly mixing stationary process with marginal distribution function* F *and $\int x^2 dF(x) < \infty$. Suppose that Assumption 1 holds. Then $\hat{h} + \frac{m^2 \log(n)}{n\hat{h}} \to 0$ in probability.*

Lemma 4.3 is an extension of Lemma 4.1 from i.i.d. to strongly mixing dependence assumption. A direct consequence of the above Lemma and Theorem 4.3 is the following Theorem.

Theorem 4.4. *Let $\{X_t,\ t \in \mathbb{Z}\}$ is a strongly mixing stationary process with marginal distribution function* F *and $\int x^2 dF(x) < \infty$. Suppose that Assumption 1 holds. Then for $h = \hat{h}$ in (2.3), $\hat{F}_n(x_0, \hat{h})$ converges in probability to $F(x_0)$ as $n \to \infty$.*

Theorem 4.4 ensures consistency of the estimator using $h = \hat{h}$, in (2.3), in the presence of strongly mixing dependence.

Acknowledgments

We are thankful to the esteemed reviewer for his suggestions that lead to substantial improvement of the article. The author is thankful to Pinky Dutta for help in collection of the data on the Nifty index.

Funding

This research has been supported by the UGC minor research project F. No. 39-938/2010 (SR) of the author.

References

Altman, N., Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference* 46:195–214.

Azzalini, A. (1981). Estimation of a distribution function and quantiles by a kernel method. *Biometrika* 68(1): 326–328.

Bose, A., Dutta, S. (2013). Density estimation using bootstrap bandwidth selector. *Statistics and Probability Letters* 83:245–256.

Bowman, A., Hall, P., Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika* 85:799–808.

Devroye, L., Gyorfi, L. (1985). *Nonparametric Density Estimation The L_1 View*. New York: John Wiley and Sons, Inc.

Dutta, S. (2012a). Local smoothing using the bootstrap. *Communications in Statistics-Simulation and Computation* 43:378–389.

Dutta, S. (2012b). *Cross-validation revived*. Unpublished monograph.

del Río, A. Q., Estévez-Pérez, G. (2012). Nonparametric kernel distribution function estimation with kerdiest: An R package for bandwidth choice and applications. *Journal of Statistical Software* 50(8): 1–21.

Falk, F. Y. (1983). Relative efficiency and deficiency of kernel type estimators of distribution functions. *Statist. Neerlandica* 37(2): 73–83.

Jin, Z., Shao, Y. (1999). On kernel estimation of a multivariate distribution function. *Statistics and Probability Letters* 41:163–168.

Liu, R., Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics* 20(8): 661–677.

Polanski, A., Baker, E. R. (2000). Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation* 65:63–80.

Rao, B. L. S. P. (1983). *Nonparametric Functional Estimation*. New York: Academic Press Inc.

Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics* 8(2): 116–119.

Rieders, E. (1993). The size of the averages of strongly mixing random variables. *Statistics and Probability Letters* 18:57–64.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference* 35:65–214.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Swanepoel, J. W. H., Graan, F. C. V. (2005). A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scandinavian Journal of Statistics* 32(4): 551–562.

BANDWIDTH SELECTION FOR KERNEL BASED INTERVAL ESTIMATION OF A DENSITY

SANTANU DUTTA *

Abstract

We propose a bandwidth selection method for kernel based interval estimation of a density at a design point, with an aim to minimize the coverage error. The bandwidth is chosen by minimizing a bootstrap estimate of the coverage error. The proposed algorithm seems to be the first bandwidth selector for kernel based interval estimation of a density.

1 INTRODUCTION

We consider the problem of construction of confidence interval for $f(x_0)$, where f is the unknown density generating the given data and x_0 is a given design point. A density function may be arbitrarily specified at a point x_0. This technical difficulty is overcome by assuming that f is continuous.

One of the most well known estimators of f is a kernel density estimator (KDE) defined as follows.

Let $X_1, ..., X_n$ be independent and identically distributed random variables with an unknown density $f(\cdot)$. The *kernel density estimator* of f based on the kernel $K(\cdot)$ and bandwidth $h \equiv h_n$, is defined as

$$\hat{f}_n(y) \equiv \hat{f}(y,h) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{y - X_i}{h}\Big), \tag{1.1}$$

where $h \to 0$ and $nh \to \infty$ as $n \to \infty$. The problem of data based selection of h for estimating $f(x_0)$ using $\hat{f}_n$ has been well studied. See for instance, Chan et al. (2010), Dutta (2012) among most recent.

In contrast, far less seems to be known regarding the choice of h for constructing a confidence interval for $f(x_0)$ using $\hat{f}_n(x_0)$. For instance, Chan et at. (2010) have mentioned that there seems to be no automatic method for practical interval estimation for $f(x_0)$ available in the literature. From the simulation study in Hall (1992) we see that the bandwidth which is appropriate (in terms of coverage accuracy) for confidence interval construction is not easy to determine. No data based method for selecting such an h was suggested by the author. Chen (1996) proposed empirical likelihood confidence intervals for density estimation, but again with no bandwidth selection method was provided. Fiorio (2004) discussed two programs, viz. "asciker" and "bsciker" in Stata, to compute asymptotic and bootstrap confidence intervals for kernel density estimation. However these programs assume that the search for the correct bandwidth has been performed beforehand (see page 173 in Fiorio (2004)). Therefore these algorithms cannot be used for determining the appropriate amount of smoothing for kernel interval estimation.

A kernel based confidence interval for $f(x_0)$ crucially depends on the approximations of the quantiles of the sampling distribution of $S = (\hat{f}_n(x_0) - E(f_n(x_0))/\hat{\sigma}$ and the bias $b = \hat{f}_n(x_0) - f(x_0)$, where $\hat{\sigma}$ is an estimated standard deviation of $\hat{f}_n(x_0)$. The bias b is not negligible even for a bandwidth minimizing the

*Mathematical Science Dept., Tezpur University Napaam:784028, Tezpur, Assam, INDIA, e-mail: tezpur1976@gmail.com.
Research supported by UGC minor research project F. No. 39-938/2010 (SR).

mean squared error. There are two approaches to tackle the bias b, viz. either to estimate bias explicitly, or reduce it substantially by under smoothing (see Hall (1992)). Hall (1992) showed that the under smoothing method produces confidence intervals with greater coverage accuracy than those obtained by explicit bias correction. There are several other practical advantages of the under smoothing method. For instance, the under smoothing approach no estimator of the bias is required (see Hall (1992)). Given $X_1, ..., X_n$ and some bandwidth h, a two sided under-smoothed bootstrap $1-\alpha$ confidence interval of $f(x_0)$ is defined as

$$I(1-\alpha) = (\hat{f}_n(x_0,\ h) - \hat{\sigma}(h)\hat{u}_{1-\alpha/2},\ \hat{f}_n(x_0,\ h) - \hat{\sigma}(h)\hat{u}_{\alpha/2}),$$

where $\hat{u}_\alpha$ is the αth quantile of a bootstrap approximation of the sampling distribution of S. We use following $\hat{\sigma}^2(h)$ proposed by Hall (1992)

$$\hat{\sigma}^2(h) = \frac{1}{nh}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{h}\right)^2 - h\hat{f}_n(x_0)^2\right].$$

Clearly $I(1-\alpha)$ is a function of h. In this paper we propose a method for data based choice of h which is suitable (in some sense) for $I(1-\alpha)$. The (exact) coverage probability of $I(1-\alpha)$ is defined as

$$\beta(1-\alpha) = P(f(x_0) \in I(\alpha)).$$

Hall (1992) suggested to select h with an aim to minimize the absolute value of the coverage error $CE = |\beta(1-\alpha) - 1 + \alpha|$. However $\beta(1-\alpha)$ is a function of the unknown f. So for practical data based choice of h; CE has to be estimated based on $X_1, ..., X_n$. Using classical bootstrap method we propose estimate of the CE and it is minimized (with respect to h) for data based choice of the bandwidth. Let denote the proposed data based bandwidth. The details of our proposal are given in Section 2.

The exact coverage probability $\beta(1-\alpha)(\hat{h})$, of the confidence interval using $\hat{h}$, is hard to compute. However for any given f, we can approximate the coverage probability using Monte-Carlo simulations. In simulation study, in Section 3, we compute the Monte-Carlo estimates of $\beta(1-\alpha)(\hat{h})$ for different choices of f and x_0. We also report the average width and the variance of the widths of the confidence intervals. The results are compared with the findings of of Hall (1992) and the results in Table 2 in Chan et al. (2010). The proposed two sided confidence interval using $\hat{h}$ seems to work well (in terms of coverage probability and average width) for sample size greater than or equal to 100.

2 Our proposal

Given $X_1, ..., X_n$ and h, we propose a bootstrap estimate $\beta^*(1-\alpha)$ of the coverage probability as follows

$$\beta^*(1-\alpha) \equiv \beta^*(1-\alpha)(h) = P^*(\hat{f}_n(x_0, h) \in I^*(1-\alpha)),$$

where

$$I^*(1-\alpha) = (\hat{f}_n^*(x_0, h) - \hat{\sigma}^*(h)\hat{u}^*_{1-\alpha/2},\ \hat{f}_n^*(x_0, h) - \hat{\sigma}^*(h)\hat{u}^*_{\alpha/2}).$$

Given $X_1, ..., X_n$, let $X_1^*, ..., X_n^*$ be a simple random sample drawn with replacement (srswr) from empirical distribution. As mentioned earlier $\hat{f}_n^*(x_0)$, $\hat{\sigma}^*$ are the bootstrap versions of $\hat{f}_n(x_0)$ and denotes the conditional probability, given $X_1, ..., X_n$. $\hat{u}^*_\alpha$ is a bootstrap version of the statistic $\hat{u}_\alpha$ is the αth quantile of the conditional distribution of $S^{**} = (\hat{f}_n^{**}(x_0,\ h) - \hat{f}_n^*(x_0,\ h))/\hat{\sigma}^{**}$, given $X_1^*, ..., X_n^*$ and h. $\hat{f}_n^{**}(x_0,\ h)$ and $\hat{\sigma}^{**}$ are obtained by replacing $X_1, ..., X_n$ in (1.1) and (1.2) by $X_1^{**}, ..., X_n^{**}$ which is a second stage re-sample drawn with replacement from $X_1^*, ..., X_n^*$.

$\beta^*(1-\alpha)$ is a function of the bandwidth h. We define a bootstrap estimator $\widehat{CE}$ of the coverage error as follows

$$\widehat{CE} \equiv \widehat{CE}(h) = |\beta^*(1-\alpha)(h) - (1-\alpha)|.$$

We minimize $\widehat{CE}$ with respect to h for data based bandwidth selection. The resulting random $\hat{h}$ is defined as follows

$$\hat{h} = \text{argmin}_{h \in J_n} \widehat{CE}(h), \tag{2.1}$$

where J_n is a compact interval with endpoints equal to scale invariant bandwidths, which are smaller than the bandwidth minimizing the MISE. As mentioned earlier, Hall suggested to use $h = 1.05 c \hat{\gamma} n^{-1/5}, 0 < c \leq 1$, for under smoothing (see [6]). Motivated by this proposal we use

$$J_n = [c_1 1.05 \hat{\gamma} n^{-1/5},\ c_2 1.05 \hat{\gamma} n^{-1/5}],\ 0 < c_1 < c_2 \leq 1.$$

Hall considered a wide range of values of c varying from 0.1 to 1, and showed that widely different values of c are appropriate under different circumstances (see Table 1 in page 687 in [6]). Motivated by this, we use $c_1 = 0.1$ and $c_2 = 1$. With these choices of c_1, c_2 , J_n covers all the under smoothing bandwidths considered by Hall in the simulation study in [6].

The proposed two sided under smoothed bootstrap $1-\alpha$ confidence interval of $f(x_0)$ is defined as

$$I(1-\alpha)(\hat{h}) = (\hat{f}_n(x_0,\ \hat{h}) - \hat{\sigma}(\hat{h})\hat{u}_{1-\alpha/2},\ \hat{f}_n(x_0,\ \hat{h}) - \hat{\sigma}(\hat{h})\hat{u}_{\alpha/2}). \tag{2.2}$$

2.1 Some computational details.

2.1.1 Computation of $\hat{u}_\alpha$

Given $X_1, ..., X_n$ and h, we compute $\hat{u}_\alpha$ as follows.

We draw B_1 bootstrap re-samples. For each re-sample we compute S^*. There are B_1 values of S^* corresponding to the re-samples. Now $\hat{u}_\alpha$ is the αth sample quantile based on these B_1 values.

2.1.2 Computation of $\hat{u}^*_\alpha$

Let $X^*_1, ..., X^*_n$ be a bootstrap re sample drawn from $X_1, ..., X_n$. Based on $X^*_1, ..., X^*_n$, we compute $\hat{u}^*_\alpha$ as follows.

We generate B_2 second stage re-samples from $X^*_1, ..., X^*_n$, and compute the values of S^{**} based on the B_2 second stage re-samples. The αth sample quantile of these B_2 values of S^{**} is a Monte Carlo approximation to $\hat{u}^{**}_\alpha$.

2.1.3 Computation of $\beta^*(1-\alpha)(h)$

Given $X_1, ..., X_n$ and h, the computation of $\beta^*(1-\alpha)(h)$ involves the following steps.

(i) Generate B_1 re-samples, each of size n, by simple random sampling with replacement (srswr) from $X_1, ... X_n$, and compute $\hat{f}^*_n(x_0, h)$, $\sigma^*(h)$ for each re-sample.

(ii) From each re-sample, we further generate B_2 second stage re-samples by srswr. Using these second stage re-samples we compute $\hat{u}^*_{\alpha/2}$ and $\hat{u}^*_{1-\alpha/2}$ by the procedure mentioned above.

(iii) Using $f_n^*(x_0, h)$, $\sigma^*(h)$, $\hat{u}_{\alpha/2}^*$ and $\hat{u}_{1-\alpha/2}^*$ we compute $I^*(1-\alpha)$ for each (1st stage) re-sample. There are B_1 such intervals corresponding to the B_1 first stage re-samples.

(iv) The Monte-Carlo estimate of $\beta^*(1-\alpha)(h)$ is equal to the number of the intervals (obtained in step (iii)) containing $f_n(x_0, h)$ divided by B_1.

Remark 1. *1. As mentioned earlier $I(1-\alpha)$ is a two sided confidence interval for $E\left(\hat{f}_n(x_0)\right)$. The above mentioned algorithm essentially imitates the Mone-Carlo (MC) method of approximating the exact coverage probability of $\beta(1-\alpha)(h)$, for any given f and h. In the MC method we draw random samples from a given distribution, and for each sample we compute $I(1-\alpha)$ by the re-sampling method described earlier. The MC estimate of $\beta(1-\alpha)(h)$ is the number of the intervals containing $E\left(\hat{f}_n(x_0)\right)$ divided by the number of random samples drawn. We imitate this procedure, replacing the actual distribution by the empirical distribution.*

We note that $\hat{f}_n(x_0) = E^*\left(\hat{f}_n^*(x_0)\right)$, where E^* denotes the expectation with respect to the empirical distribution. So the bootstrap version of $I(1-\alpha)$ is a confidence interval for $\hat{f}_n(x_0)$, given $X_1, \ldots, X_n$. In our method the 1st stage re-samples, drawn from the empirical distribution, mimic the role played by the random samples drawn from the actual distribution in the MC method.

2. We use the same 1st stage re-samples and 2nd stage re-samples (obtained by re-sampling each 1st stage re-sample in step [ii] of the above algorithm) to compute $\beta^*(1-\alpha)(h)$ for different values of h, as required in a numerical minimization algorithm. This feature reduces the computational burden.

3. Given a confidence interval, Monte-Carlo approximation of its coverage probability essentially involves estimating an average of a random function using Monte-Carlo simulations. From [5] we see that much larger number of Monte-Carlo re-samples are required for approximating a bootstrap quantile estimator accurately, than the same required for approximating a bootstrap estimator of the expectation of some random function. Therefore we use different number of re-samples, viz. B_2 and B_1, to approximate the bootstrap estimators of the quantiles and the coverage probability by Monte-Carlo method.

2.2 Monte Carlo sample size for bootstrap-resampling

From [10] we see that the selection of appropriate B_1 and B_2 are not easy problems. As a rule of thumb [5] suggested that for Monete-Carlo approximation of bootstrap moment estimators the number of bootstrap re-samples should be 50 to 200. For approximating bootstrap quantile estimators the number of bootstrap re-samples should be at least 1000 (see [5]) . We use this rule of thumb, and use $B_1 = 200$, $B_2 = 1000$.

3 Simulation

Hall conducted simulations to study the effect of the choice of h on the coverage probability of an under smoothed bootstrap confidence interval $I(1-\alpha)(h)$ was examined for six combinations of f and x_0 (see [6]). The author used $h = c1.05\hat{\gamma}n^{-1/5}$, where $0 < c \leq 1$, for under smoothing the density estimator. In his simulations f equals to the $N(0,1)$ density and the $(1/2)N(0,1) + (1/2)N(3,1)$ density, and c equal to 0, 0.75 and 1.5. The notation $pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$ represents a two component mixed normal distribution, where μ_i, σ_i^2 are the mean and variance of the ith mixing component. For both these test densities, $x_0 = 0$ is the peak of the density. Hall reported the Monte Carlo estimates of the exact coverage probability $\beta(1-\alpha)(h)$, along with the average and standard deviation of the interval length. It was observed that the coverage accuracy of the confidence interval for f at the peak was less than the same at other point.

In [1], the authors considered the problem of interval estimation of $f(0)$, where f is a standard normal density. From their simulations (page 513, in [1] we see that neither the coverage error nor the length of their 95 percent interval seem to decrease as n is increased more than two times. This is perhaps due to the fact that random bandwidth proposed by Chan Lee and Peng is suitable for point estimation of f at x_0. In [7], the author pointed out that nonparametric point estimation and interval estimation are different tasks that require different degrees of smoothing.

In this section we study effect of the proposed random bandwidth $\hat{h}$ on the coverage probability and the average length of $I(1-\alpha)$, for different choices of f and x_0 and $\alpha = 0.05$. We consider the above mentioned choices of f and x_0 as in [6]. Both these densities are unimodal, with peak at $x_0 = 0$. In addition we consider two more test densities, viz. f equal to the $(1/2)N(-1,1/2) + (1/2)N(1,1/2)$ density and the gamma(2,1) density. For the $(1/2)N(-1,1/2) + (1/2)N(1,1/2)$ density there are two peaks of same height at -1 and 1, and a trough at 0. We estimate this density at x_0 equal to 0 and 1. For the gamma density peak occurs at 1. We estimate the height of the gamma density and x_0 equal to 1 and 4.474, which is the 95th percentile. To compute the Monte-Carlo estimate of the coverage probability of a confidence interval we draw m random samples of a specific size from a test distribution, and compute the confidence interval for each sample. So there are m such intervals. The Monte-Carlo estimate of the coverage probability is equal to number of intervals containing $f(x_0)$, divided by m. In Table 1 we use $c_1 = 0.1$ and $c_2 = 1$.

In Table 2 we report the Monte-Carlo estimates of the coverage probability, average length and variance of the confidence intervals using $h = c1.05\hat{\gamma}n^{-1/5}$, for different choices of c and f equal to the $(1/2)N(-1,1/2) + (1/2)N(1,1/2)$ density and the gamma(2,1) density. If the mean or the variance of the length of the confidence interval exceeds 100, we write "large".

In Table 1 we report the Monte-Carlo estimate of the coverage probability, average length and variance of the proposed confidence interval $I(1-\alpha)(\hat{h})$, in (2.2), for 10 combinations of f and x_0. We compute each estimate for $n = 50$ and $n = 100$. To compute Monte-Carlo estimate we draw $m = 300$ samples from each test density. We have the following observations.

(i) The confidence interval $I(1-\alpha)(\hat{h})$, using the proposed random bandwidth $\hat{h}$ in (2.1), seems to perform consistently. The coverage error, the mean and the variance of the interval length seem to reduce as sample size is increased for all choices of f and x_0.

(ii) From the simulation study in [6] and our Table 2, we see that the coverage probability and length of the confidence intervals using $h = c1.05\hat{\gamma}n^{-1/5}$, $0 < c \leq 1$, can vary widely depending on estimation point x_0 and c.

(iii) In contrast, the simulations in Table 1 indicate that for a given distribution the coverage accuracy of the confidence interval using $\hat{h}$ does not seem to vary drastically with the change in x_0, especially for $n = 100$. This is due to the fact that proposed bandwidth selector is a function of the estimation point x_0, and so the resulting bandwidth $\hat{h}$ automatically adjusts the amount of smoothing depending on x_0.

(iv) From the simulations in [6] we see that for f equal to the $(1/2)N(0,1) + (1/2)N(3,1)$ density and x_0 equal to the peak, the coverage probability of the under smoothed confidence interval is poor especially for $c > 0.5$ in $h = c1.05\hat{\gamma}n^{-1/5}$. From our Table 2 we see that a similar observation is also true for x_0 equal the trough between the two peaks of the $(1/2)N(-1,1/2)+(1/2)N(1,1/2)$ density. Hall pointed out that the coverage error of confidence interval for estimation f at the peak is in general higher than the same at other points, as the bias in a kernel density estimator is more pronounced at a peak. We observe that the same argument is also true for x_0 equal to a trough. Moreover from Table 2 we see that while estimating the gamma density at the peak the under smoothed confidence interval using $h = c1.05\hat{\gamma}n^{-1/5}$ performs poorly for every choice c.

Table 1: Monte Carlo estimates of $\beta(1-\alpha)(\hat{h})$ for h equal to $\hat{h}$ and $\alpha = 0.05$

Density	(x_0, n)	Coverage Probability	Interval Width average (variance)
N(0,1)	(0, 50)	0.90	0.371 (0.014)
	(0, 100)	0.96	0.151 (0.002)
	(0.75, 50)	0.91	0.381 (0.013)
	(0.75, 100)	0.958	0.239 (0.006)
	(1.5, 50)	0.88	0.221 (0.007)
	(1.5, 100)	0.935	0.143 (0.002)
(1/2)N(-1, 1/2) + (1/2)N(1, 1/2)	(0, 50)	0.90	0.229 (0.009)
	(0, 100)	0.91	0.167 (0.003)
	(1, 50)	0.90	0.384 (0.033)
	(1, 100)	0.91	0.295 (0.005)
(1/2)N(0, 1) + (1/2)N(3, 1)	(0, 50)	0.924	0.179 (0.003)
	(0, 100)	0.935	0.129 (0.001)
	(0.75, 50)	0.97	0.162 (0.002)
	(0.75, 100)	0.962	0.117 (0.001)
	(1.5, 50)	0.915	0.160 (0.012)
	(1.5, 100)	0.94	0.112 (0.001)
gamma(2,1)	(1, 50)	0.87	0.306 (0.011)
	(1, 100)	0.965	0.255 (0.004)
	(4.474,50)	0.84	0.081 (0.001)
	(4.474,100)	0.88	0.071 (0.002)

However, simulations in Table 1 suggest that the proposed confidence interval $I(1-\alpha)(\hat{h})$ perform well in estimating f at the peak as well as the trough, in terms of the coverage accuracy, especially for $n = 100$ and irrespective of f.

(v) From the simulations in [6] and our Tables 1 and 2, we see that the mean and the variance of the length proposed confidence interval compares well with the lengths of the corresponding confidence intervals using $h = c1.05\hat{\sigma}$ in [6].

Final Remarks. From the above simulation study it appears that the confidence interval $I(1-\alpha)(\hat{h})$ in (2.2) performs well for all the test densities, especially for $n = 100$. Simulations in our Table 2 suggest that if f is a density with positive support and x_0 is the peak, the under smoothed confidence interval for $f(x_0)$ using $h = c1.05\hat{\gamma}n^{-1/5}$ performs poorly for all the different choices of c mentioned in [6]. In contrast, the coverage error or the average length of $I(1-\alpha)(\hat{h})$ does not seem to vary drastically for different choices of x_0. So the proposed bandwidth selector can be recommended safely for interval estimation of $f(x_0)$ especially for large sample size.

REFERENCES

[1] Chan, N.H., Lee, T.C.M., and Peng, L. (2010). On nonparametric local inference for density estimation. Comuptational Statistics and Data Analysis 54: 509-515.

[2] Chen, S.X., 1996. Empirical likelihood confidence intervals for nonparametric density estimation. Biometrika 83: 329-341.

[3] Dutta, S. (2012). Local smoothing using the bootstrap. Communications in Statistics-Simulation and Computation. Accepted for publication.

[4] Efron, B. (1979). Bootstrap methods: Another look at the jackknife, The Annals of Statistics 7:1-26.

[5] Efron, B. and Tibshirani, R. J. (1986). Bootstrap methods for standard error, confidence intervals, and other measures of statistical accuracy. Statist. Science1:54-77.

[6] Hall, P. (1992). Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. The Annals of Statistics 20, 2 : 675-694.

[7] Horowitz, J. L. (2001). The bootstrap. In Handbook of Econometrics, ed. J. J. Heckman and E. Leamer, vol. 5, 31593228. Amsterdam: North-Holland.

[8] Fiorio, C. V. (2004). Confidence intervals for kernel density estimation. The Stata Journal 4 2: 168179.

[9] Sheather, S.J.,and Jones, M.C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. J. Roy. Statist. Soc. Ser. B. 53: 683-690.

[10] Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. Springer-Verlag, New-York.

Table 2: Monte Carlo estimates of $\beta(1-\alpha)(\hat{h})$ for $h = c \cdot 1.05 \hat{\gamma} n^{-1/5}$ for different values c

Density	c	(x_0, n)	Coverage Probability	Interval width mean (variance)
(1/2)N(-1, 1/2) + (1/2)N(1, 1/2)	0.1	(0, 50)	0.95	2435.31(large)
		(0, 100)	0.975	15.744 (large)
		(1, 50)	0.975	86.57 (large)
		(1, 100)	0.97	0.850 (0.219)
	0.2	(0, 50)	0.965	243 (large)
		(0, 100)	0.96	1.672 (19.850)
		(1, 50)	0.98	0.616 (0.022)
		(1, 100)	0.96	0.432 (0.004)
	0.3	(0, 50)	0.95	0.879 (1.656)
		(0, 100)	0.965	0.272 (0.019)
		(1, 50)	0.955	0.455 (0.0038)
		(1, 100)	0.975	0.323 (0.0013)
	0.5	(0, 50)	0.865	0.233 (0.0026)
		(0, 100)	0.87	0.156 (0.0006)
		(1, 50)	0.965	0.303 (0.001)
		(1, 100)	0.9	0.227 (0.0005)
	0.75	(0, 50)	0.415	0.161 (0.0005)
		(0, 100)	0.32	0.120 (0.0002)
		(1, 50)	0.755	0.220 (0.0003)
		(1, 100)	0.62	0.167 (0.0002)
	1	(0, 50)	0.01	0.125 (0.0003)
		(0, 100)	0.01	0.097 (0.0001)
		(1, 50)	0.285	0.172 (0.0003)
		(1, 100)	0.225	0.132 (0.0001)
Gamma(2,1)	0.1	(0, 50)	0.83	large(large)
		(0, 100)	0.75	"
		(4.474, 50)	0.975	
		(4.474, 100)	0.955	"
	0.2	(0, 50)	0.425	large(large)
		(0, 100)	0.195	"
		(4.474, 50)	0.975	"
		(4.474, 100)	0.965	"
	0.3	(0, 50)	0.125	4.019(large)
		(0, 100)	0.01	0.198 (0.024)
		(4.474, 50)	0.97	large (large)
		(4.474, 100)	0.965	"
	0.5	(0, 50)	0.01	0.199 (0.024)
		(0, 100)	0.01	0.103 (0.0004)
		(4.474, 50)	0.96	large (large)
		(4.474, 100)	0.975	"
	0.75	(0, 50)	0.0	0.123 (0.006)
		(0, 100)	0.01	0.088 (0.0001)
		(4.474, 50)	0.98	194.5 (large)
		(4.474, 100)	0.955	0.096 (0.006)

MULTIVARIATE DENSITY ESTIMATION USING AUTOMATIC L_1 SMOOTHING

SANTANU DUTTA
Mathematical Sciences Department
Tezpur University
Napaam 784028
INDIA
tezpur1976@gmail.com

Abstract

We present a new method for automatic selection of the bandwidth matrix for a multivariate kernel density estimate, under weak conditions. The existing multivariate methods for data based choice of a bandwidth matrix aim to minimize some L_2 measure of accuracy, and impose a number of assumptions on the underlying density and its derivatives. In contrast we suggest to choose the bandwidth matrix with an aim to minimize a suitable L_1 distance, and we impose no conditions on the density function at all.

We only assume that the kernel is a probability density function, and the bandwidth matrix is positive definite. Under these few assumptions, $P(\int |f_n - f| > \epsilon)$ converges to zero exponentially as sample size is increased, where f_n is the density estimate using our automatic bandwidth matrix and f is the density. This result answers the important question that "how well does a kernel density estimate, using our automatic bandwidth matrix, estimate the true density?" This question does not seem to have been answered for any other multivariate bandwidth matrix selector.

Simulations and analysis of real data confirm that this new method is not merely of academic interest, but compares well with the existing sophisticated bandwidth selectors, such as the plug-in method based on 2 stage of pilot estimation (Duong and Hazelton (2003)).

Keywords and Phrases: Kernel density estimator, automatic bandwidth, L_1 distance.

AMS Subject Classification: 62G07, 62G09, 62G20.

Let $X_1, \ldots, X_n$ be n i.i.d. R^d valued ($d > 1$) random variables with joint density f. In its most gene... form, the kernel estimator $f_n \equiv f_n(.., H)$ of f is defined as

$$f_n(y) = \frac{1}{n|H|^{1/2}} \sum_{i=1}^{n} K\left((y - X_i)H^{-1/2}\right), \quad \text{where } y = (y_1, .., y_d) \in R^d \text{ and}$$

K is a $d-$variate probability density function and $H \equiv H_n$ is a $d \times d$ positive definite matrix. Wand ... Jones (1993) have compared different parameterizations of H, for bivariate kernel density estimates.

A kernel density estimate is a useful tool for exploratory data analysis, especially for bivariate data wh... it can be visualized using the familiar perspective or contour plots (see Duong and Hazelton 2005). It ... sensitive to the choice of H (for example, see Wand and Jones (1993), Simonoff (1996) and Duong ... Hazelton (2003, 2005)). Duong and Hazelton (2005) have rightly mentioned that "...the study of data-dri... methods for selecting H is not only important in its own right, but also because it sheds light on m... general multivariate kernel smoothing problems."

For univariate data, there are a number of data based bandwidth selection schemes with good theoreti... properties and strong practical performance (see Jones et al. (1996)). Research on automatic choice of H ... the asymptotic properties of a data based bandwidth matrix seems to be a relatively new area (for instan... see Sain et al. (1994), Wand and Jones (1994), Duong and Hazelton (2003, 2005), Dutta (2010) among m... recent). A number of the multivariate bandwidth selectors are designed assuming H is a diagonal matri... e.g. Sain et al. (1994) and Dutta (2010). Wand and Jones (1993), Duong and Hazelton (2003) provi... motivation for using a full bandwidth matrix H. The development of selectors for full H is rather mo... challenging than that for diagonal H. The off diagonal entries of H control the orientation of the kern... functions to the coordinate axes. The need to automatically select these off diagonal entries introduces ... problem without a univariate analogue.

A common feature among all the above mentioned bandwidth selectors is that they aim to minimize th... mean integrated squared error (MISE), a L_2 measure of overall accuracy of a density estimate. Moreove... these methods impose some conditions on f and its derivatives. Let us review some of these assumption... For instance, Duong and Hazelton (2003) obtained automatic plug-in bandwidth matrices, assuming that th... partial derivatives of f, up to order four, exist and all the second order partial derivatives are square inte... grable. Similar assumptions have been imposed by Duong and Hazelton (2005) and Dutta (2010). Anothe... common assumption in all the above mentioned bandwidth selectors is that each entry of the matrix H go... to zero as n is increased.

We propose a radically different approach of selecting H by minimizing some appropriate L_1 distance... avoiding conditions on f, K and H as far as possible. In the sequel, we only assume that K is a d vari... density and H is a $d \times d$ positive definite matrix. There are no conditions on the density function f at all.

There is a clear lack of papers about the automatic L_1 smoothing approach in density estimation (Cao ... al. (1994)). In the univariate context the double kernel method, proposed by Devroye (1989), is a propos... that aims to choose bandwidth by minimizing the L_1 distance between the given kernel density estimate an... a pilot density estimate. Devroye (1989) obtained some interesting asymptotic properties of the result... density estimate, and Cao et al. (1994) observed that that the double kernel method is a very promisi... selector in estimation of hard densities (see page 171, Cao et al. (1994)). However we have not come acro... any such method, that aims to minimize a L_1 distance, for automatic choice of the bandwidth matrix of ... multivariate kernel density estimate.

Our proposal. We propose to select H with an aim to minimize $L_B = \int_B |f_n(y) - f(y)|dy$, where B ... Borel subset of R^d. L_B is always well defined. B represents the region over which the density is estimat...

and it may be R^d or some proper subset of R^d. For $B = R^d$, $L \equiv L_{R^d}$ equals the *integrated absolute error* (IAE), which is invariant under monotone transformations of coordinate axes (see Devroye (1983), Devroye and Gyorfi (1985)).

Sometimes a choice of B may be obvious from the context of the data analysis problem itself. We present one example. The severity of an earthquake depends both on its magnitude and the depth of its focus (from earth surface). So estimating the joint density of the depth and magnitude of earthquakes, occurring in a region, seems to be of interest. Usually quakes of magnitude less than 3 are imperceptible and the largest magnitude recorded till date is 9.5. Quakes can occur anywhere between the earth surface and about 700 Km below the surface. These information on quakes have been obtained from "http://en.wikipedia.org/wiki/Earthquake" and the references therein. So for estimating the joint density of depth and magnitude, B equal to $(0, 700] \times (3, 9.5]$ seems to be a more reasonable choice than R^2.

The choice of B can be data driven as well. For instance B may be a $d-$ dimensional rectangle with endpoints equal to the sample extremes along each coordinate direction.

We note that L_B is a function of the density f, which is unknown. So we propose an estimate of L_B, call it L_B^*, which is defined as follows.

Definition 1.

$$L_B^* = \int_B |f_n(y) - g_n(y)| dy,$$

where g_n is a kernel density estimate with a $d-$ variate kernel K^0, $K^0 \neq K$, and bandwidth matrix $H_\lambda = \frac{1}{n^{2/(4+d)}} I_d$, where I_d is a $d \times d$ identity matrix.

L_B^* is in fact a smooth bootstrap estimate of L_B. Let $\hat{H}$ denote the bandwidth matrix minimizing L_B^*. L_B^* is minimized over a class of positive definite matrices. So $\hat{H}$ is always a positive definite matrix.

The resulting automatic density estimate $\hat{f}_n$ equals

$$\hat{f}_n(y) = \frac{1}{n|\hat{H}|^{1/2}} \sum_{i=1}^{n} K\left((y - X_i)\hat{H}^{-1/2}\right).$$

Remark 1. *a) The concept of minimizing L^*, for selecting H, has one advantage. Let H^* denote the bandwidth minimizing of the L^*. Then $L(H^*)$ is the integrated absolute error of $\hat{f}_n$. In contrast, if H_M denotes the minimizer of some data based estimate of the MISE M, then $M(H_M)$ is not the MISE of the density estimate using the random bandwidth matrix H_M. In fact $M(H_M)$ does not have a conceptual interpretation, except that it represents the value of M at $H = H_M$.*
(b) The condition $K^0 \neq K$ is a necessary condition. For $K^0 = K$, $H = \frac{1}{n^{2/(4+d)}} I_d$ is the minimizer of L_B^*. Consequently for $K^0 = K$, H is no longer automatic data based bandwidth matrix.

From the perspective of density estimation, the important question is that "how well does $\hat{f}_n$ estimate f?". The following Theorem provides some insight.

Theorem 1. *Let f and K be $d-$variate density functions and $H \in \mathbb{F}$, the class of all $d \times d$ positive definite matrices. Then for every $\epsilon > 0$, there exists positive constants $c, n_0 > 0$ such that*

$$P\left(\int |\hat{f}_n(y) - f(y)| dy \geq \epsilon\right) \leq 2e^{-cn\epsilon^2}, \ \forall n \geq n_0.$$

The following Corollary is immediate.

Corollary 1.1. *Under the conditions stated in Theorem 1,* $\int |\hat{f}_n(y) - f(y)|dy \to 0$, *almost surely, and* $E\int |\hat{f}_n(y) - f(y)|dy \to 0$.

It is important to note that the above Theorem and the Corollary hold for any density function f.

Most of the theoretical research on assessing the asymptotic performance of a univariate automatic bandwidth selector has been concerned with questions like "how fast does $\frac{\hat{h}}{h^*} - 1$ converge to zero?", where h^* is the minimizer of the MISE (see Loader (1999)). Similar question is raised and answered for multivariate bandwidth matrix selectors as well. For example, Duong and Hazelton (2005) have proved that $vech(\hat{H} - H_{AMISE}) = O_P(JH_{AMISE})$, where $\hat{H}, H_{AMISE}$ are bandwidth matrices minimizing the smoothed cross validation and asymptotic approximations to the MISE, J is a square matrix of ones, and $vech(H)$ is the lower triangular half of H strung out column-wise into a vector. However, Loader (1999) have argued that the real question for assessing the asymptotic performance of a bandwidth selector is "how accurately does the resulting kernel density estimator approximate the density?" Unfortunately this question seems to be unanswered for existing multivariate bandwidth matrix selectors.

The asymptotic accuracy of a density estimator, using an automatic bandwidth or bandwidth matrix, can be measured by the rate at which its integrated squared error (ISE) or the integrated squared error (IAE) goes to zero. Our Theorem 1 provides the rate of convergence of the IAE of a density estimate, using bandwidth matrix $\hat{H}$.

Theorem 1 and the subsequent corollary provide insight into the asymptotic behaviour of $\hat{f}_n$. But it is important to assess its finite sample performance, based on real and simulated data. This issue addressed in the next section. A proof of Theorem 1 is given in the appendix (section 3).

2 Simulation and analysis of real data.

Let us demonstrate our method for bivariate data. We draw samples of size 1000 from four target densities from the mixed bivariate normal family. In Table 1 we provide the formulae of the four test densities.

We compare our density estimates with the unconstrained plug-in density estimates, using 2 stage of plug-in estimation, (Duong and Hazelton (2003)). The latter is known to perform well for a wide variety of underlying densities, and is recommended by Duong, T. (2007). We use Gaussian kernel, and density estimates are produced using the "kde" function in "ks" package in R. We use $g_n(y) = \frac{1}{n\lambda^d}\sum_{i=1}^{n}\prod_{j=1}^{d} K^0\left(\frac{y_j - Y}{\lambda}\right)$, where K^0 is is the density function of uniform distribution on $[-1, 1]$ and $\lambda = \frac{1}{n^{1/6}}$.

Table 1: Parameters of 4 bivariate normal mixture distributions

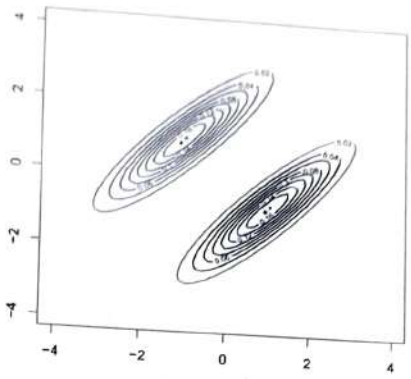
Density	$w_1.N(m_{11}, m_{12}; \sigma_{11}^2, \sigma_{12}^2, \rho_1\sigma_{11}\sigma_{12}) + ... + w_k.N(m_{k1}, m_{k2}; \sigma_{k1}^2, \sigma_{k2}^2, \rho_k\sigma_{k1}\sigma_{k2})$
A	$\frac{1}{2}N(1, -0.9; 1, 1, 0.9) + \frac{1}{2}N(-1, 0.9; 1, 1, 0.9)$
B	$\frac{4}{11}N(-2, 2; 1, 1, 0) + \frac{3}{11}N(0, 0; 0.8, 0.8, -0.72) + \frac{4}{11}N(2, -2; 1, 1, 0)$
C	$\frac{3}{7}N(-1, 0; 1, 1, 0) + \frac{3}{7}N(1, \frac{2}{\sqrt{3}}; \frac{9}{25}, \frac{49}{100}, 0) + \frac{1}{7}N(1, -\frac{2}{\sqrt{3}}; \frac{9}{25}, \frac{49}{100}, 0)$
D	$\frac{1}{2}N(1, -1; \frac{4}{9}, \frac{4}{9}, \frac{14}{45}) + \frac{1}{2}N(-1, 1; \frac{4}{9}, \frac{4}{9}, 0)$

Observations. From Figures 9-12 we see that the density estimates, using our $\hat{H}$ and the plug-in bandwidth matrix, are almost indistinguishable especially for the test densities B and C.

For test density A, the location and orientation of the two modes are same both the density estimates but the the plug-in density estimate, the lower right mode is slightly shorter than the other mode. However in our density estimate both the modes appear to be of the same length.

For the test density D, the position of the two modes in both the density estimates are same. However the shape and orientation of the upper left mode appears to be slightly different in the two density estimates. In our density estimate, the upper left mode appears to be more circular in shape, and the two modes appear to be angled away from each other. These features resemble the contour plot of the test density D. In the plug-in density estimate the modes appear to be parallel to each other

For the test densities A and D the plug-in density estimate appears to slightly smoother, but our density estimate does not seem to exhibit any spurious sampling artifacts for any of the underlying test densities.



Contour plot of test density A

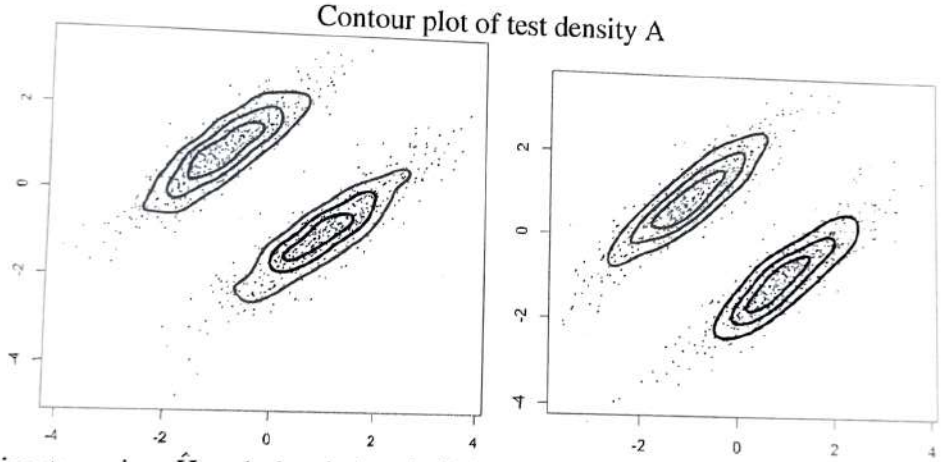
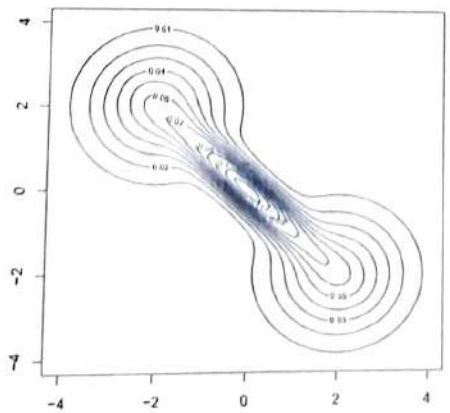


Fig 9:Density estimates using $\hat{H}$ and plug-in bandwidth, based on sample of size $n = 1000$ from test density A.



Contour plot of test density B

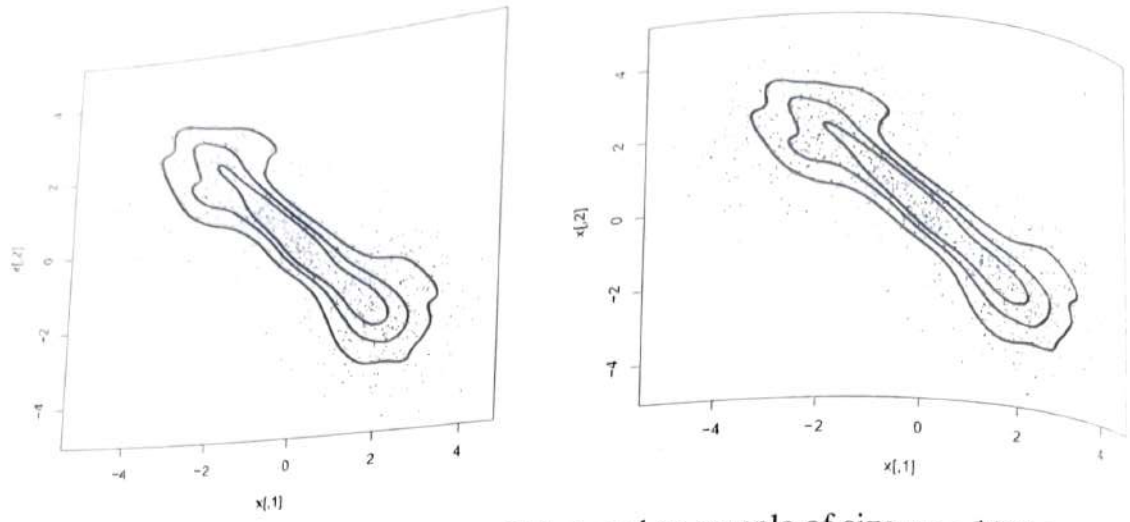
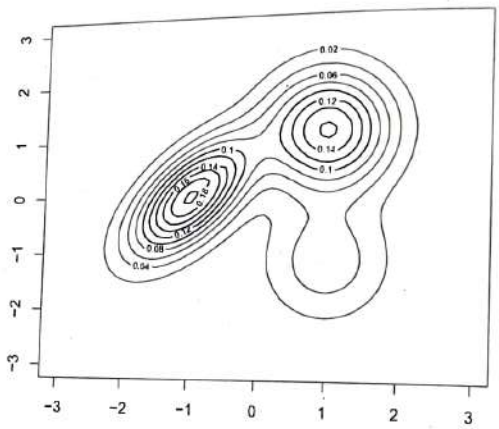


Fig 10:Density estimates using $\hat{H}$ and plug-in bandwidth, based on sample of size $n = 1000$ from test d



Contour plot of test density C

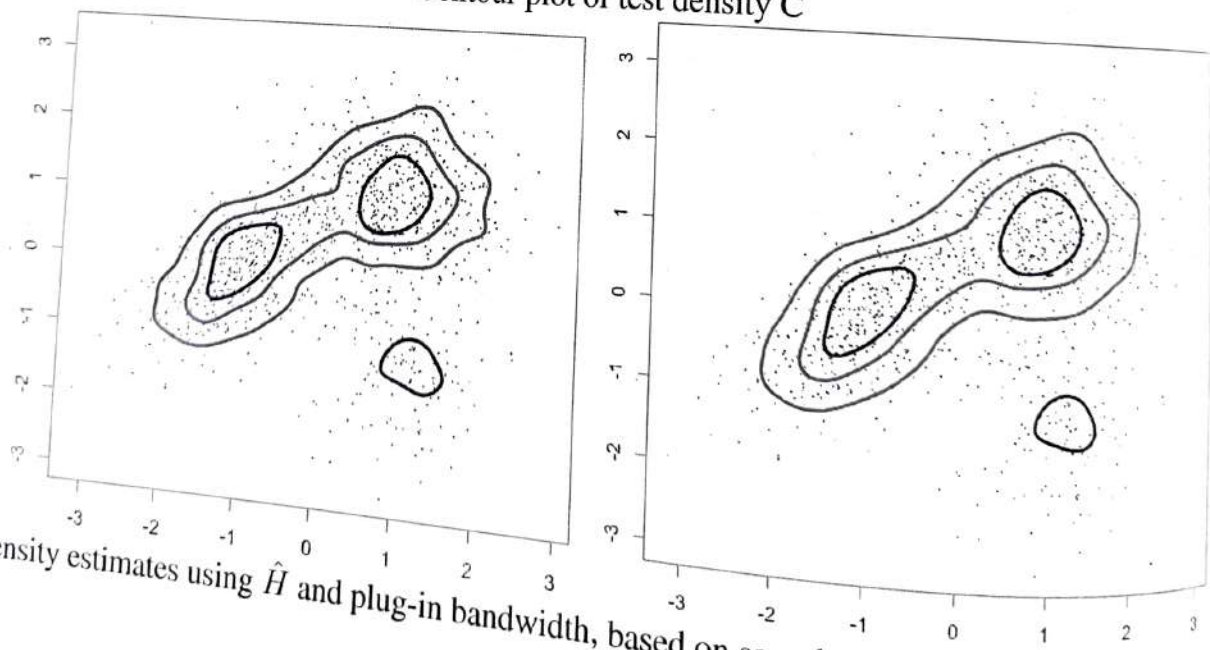
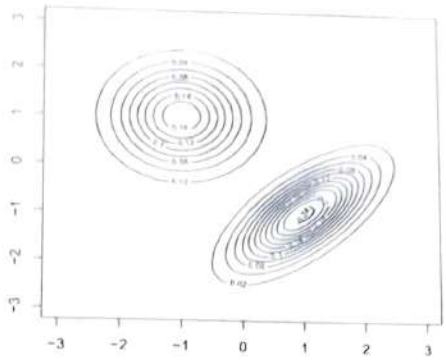
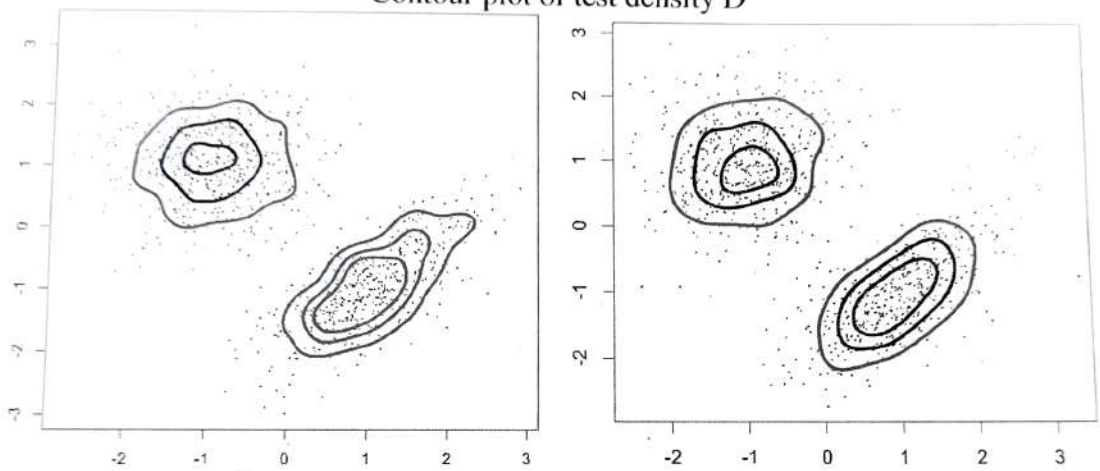


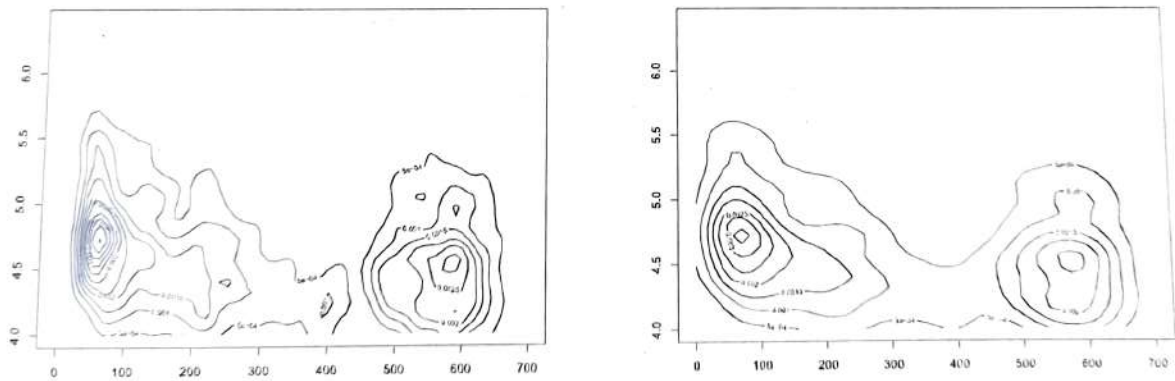
Fig 11:Density estimates using $\hat{H}$ and plug-in bandwidth, based on sample of size $n = 1000$ from test den



Contour plot of test density D



g 12:Density estimates using $\hat{H}$ and plug-in bandwidth, based on sample of size $n = 1000$ from test density D.



igure 13: Estimates of the joint density of depth and magnitude of quakes in Fiji, using $\hat{H}$ and the plug-in bandwidth matrix by Duong and Hazelton (2003)

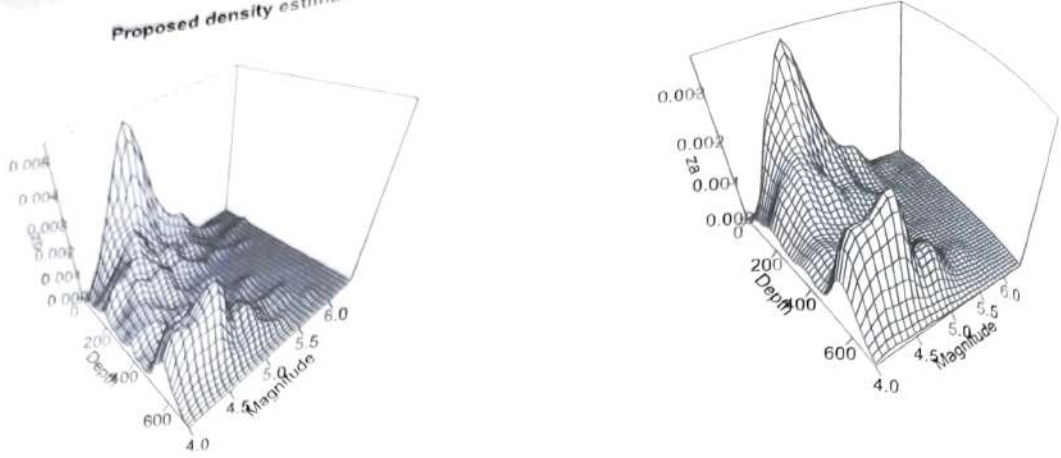


Figure 14: Estimates of the joint density of depth and magnitude of quakes in Fiji, using $\hat{H}$ and the plug-in bandwidth matrix by Duong and Hazelton (2003)

Analysis of real bivariate data. Let us consider a data set consisting of the depths and magnitudes of seismic events which occurred in a cube near Fiji since 1964. The data are available in R library. We estimate the joint density of the depth and magnitude of the quakes, using our $\hat{H}$ and the plug-in bandwidth matrix based on 2 stage of pilot estimation (Duong and Hazelton (2003)). The depth of the focus of an earthquake can vary from 0 to 700 km from earth surface, and the minimum and maximum magnitudes of the 1000 quakes equal to 4 and 6.4. So we use B equal to $[0, 700] \times [4, 6.4]$.

Conclusion. From figures 13 and 14, we see that our density estimate is slightly under smoothed in comparison to the plug-in density estimate. But both the density estimates confirm that the data is strongly bimodal. The left peak represents the quakes with (depth, magnitude) in $(0, 100) \times (4.5, 5)$, and the right peak represents quakes with (depth, magnitude) in $(550, 650) \times (4.25, 4.75)$.

Integrating the density estimates over these two regions we see that the taller left peak covers more probability than the right peak. The taller left peak indicates that a significant proportion of the Fiji quakes occurred within 100 km from earth surface, with magnitude 4.5 to 5 Richter. Being closer to the surface these quakes can cause more damage than the "deep focus" quakes occurring at 550 to 700 Km from the surface.

3 APPENDIX.

Proof of Theorem 1. Let us recall that $L(H) = \int |f_n(y) - f(y)|dy$ and $L^*(H) = \int |f_n(y) - g_n(y)|dy$ where f_n is a kernel density estimate with a $d-$ variate kernel K and bandwidth matrix H. we see almost surely,

$$\begin{aligned} |L(H) - L^*(H)| &\leq \int |g_n(y) - f(y)|dy \\ \Rightarrow ||L - L^*|| &\leq \int |g_n(y) - f(y)|dy, \quad \text{where} \end{aligned}$$

$||L - L^*|| = \sup_{H \in F} |L(H) - L^*(H)|$, where F is the class of all $d \times d$ positive definite matrices. We know that $\int |\hat{f}_n(y) - f(y)|dy = L(\hat{H})$ and $L^*(\hat{H}) = \int |\hat{f}_n(y) - g_n(y)|dy$, and recall that H^*, $\hat{H}$ are

definite matrices which minimize L and L^* respectively. Then

$$\begin{aligned} L(\hat{H}) &\leq |L(\hat{H}) - L^*(\hat{H})| + L^*(\hat{H}) \\ &\leq \|L - L^*\| + |L^*(\hat{H}) - L(H^*)| + L(H^*) \\ &\leq 2\|L - L^*\| + L(H^*) \leq 2\int |g_n(y) - f(y)|dy + L(H^*) \quad \text{using (3.1)} \end{aligned}$$

Using the above inequality, it is easy to verify that

$$\begin{aligned} P\left(\int |\hat{f}_n(y) - f(y)|dy > \epsilon\right) &= P(L(\hat{H}) > \epsilon) \\ &\leq P\left(\int |g_n(y) - f(y)|dy > \frac{\epsilon}{4}\right) + P\left(L(H^*) > \frac{\epsilon}{2}\right) \\ &\leq P\left(\int |g_n(y) - f(y)|dy > \frac{\epsilon}{4}\right) + P\left(L(H^*) > \frac{\epsilon}{2}\right) \\ &\leq P\left(\int |g_n(y) - f(y)|dy > \frac{\epsilon}{4}\right) + P\left(L\left(\frac{1}{n^{1/(4+d)}} I_{d\times d}\right) > \frac{\epsilon}{2}\right) \end{aligned} \tag{3.2}$$

$L\left(\frac{1}{n^{1/(4+d)}} I_{d\times d}\right) = \int |f_n^*(y) - f(y)|dy$, where f_n^* is a d-variate kernel density estimate with kernel K and bandwidth matrix $\frac{1}{n^{1/(4+d)}} I_{d\times d}$. We recall that g_n is also a d-variate kernel density estimate with a kernel K^0 and same bandwidth matrix.

In fact, $g_n(y) = \frac{1}{n\lambda^d}\sum_{i=1}^n K^0\left(\frac{1}{\lambda}(y - X_i)\right)$ and $f_n^*(y) = \frac{1}{n\lambda^d}\sum_{i=1}^n K\left(\frac{1}{\lambda}(y - X_i)\right)$, where $\lambda = \frac{1}{n^{1/(4+d)}}$. Clearly $\lambda = o(1)$ and $n\lambda^d \to \infty$, as n increased. Therefore using Theorem 1 (Devroye (1983)), we see that for every $\epsilon > 0$ there exists constants $r_1,\ n_1$ and $r_2,\ n_2$, such that

$$P\left(\int |g_n(y) - f(y)|dy \geq \frac{\epsilon}{4}\right) \leq e^{-r_1 n}, \quad n \geq n_1$$

$$\text{and} \quad P\left(L\left(\frac{1}{n^{1/(4+d)}} I_{d\times d}\right) > \frac{\epsilon}{2}\right) = P\left(\int |f_n^*(y) - f(y)|dy \geq \frac{\epsilon}{2}\right) \leq e^{-r_2 n}, \quad n \geq n_2.$$

Let $n_0 = \max(n_1,\ n_2)$ and $r = \max(r_1,\ r_2)$. Substituting the above inequalities in the right side of (3.2), we get

$$P(L(\hat{h}) > \epsilon) \leq 2e^{-rn}, \quad n \geq n_0.$$

This completes the proof of Theorem 1.

References

[1] Cao, R. (1993). Bootstrapping the Mean Integrated Squared Error. *Journal of Multivariate Analysis*, No. 45, 137-160.

[2] Cao, R. and Cuevas, A. and Gonzlez-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation.*Computational Statistics and Data Analysis*, 17, 153-176.

[3] Devroye, L. (1983). The Equivalence of Weak, Strong and Complete Convergence in L_1 for Kernel Density Estimates'. *The Annals of Statistics*, 11, No. 3, 896-904.

[4] Devroye, L. and Gyorfi, L. (1985). *Nonparametric Density Estimation The L_1 View*. John Wiley and Sons, Inc.

[5] Duong, T. and Hazelton M.L. (2003). Plug-in bandwidth matrices for bivariate kernel density es
Journal of Nonparametric Statistics, 15: 17-30.

[6] Duong, T. and Hazelton M.L. (2005). Cross-validation bandwidth matrices for multivariate ker
sity estimation. *Scandinavian Journal of Statistics*, 32: 485-506.

[7] Duong, T. (2007). ks:Kernel Density Estimation and Kernel Discriminant Analysis for Multivar
in R. *Journal of Statistical Software*, 21, Issue 7: 1-16.

[8] Dutta, S. (2010). Estimation of the MISE and the optimal bandwidth vector of a product kerne
estimate. *Journal of Statistical Planning and Inference*, 141, 1817-1831.

[9] Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). A brief survey of bandwidth selection fo
estimation. *J. Amer. Statist. Assoc.*, 91, No. 433, 401-407.

[10] Loader C.R. (1999). Bandwidth Selection: Classical or Plug-in? *The Annals of Statistics*, Vol
2, 415-438.

[11] Sain, S.R., Baggerly, K.A. and Scott, D.W. (1994). Cross-Validation of Multivariate Densities, J
Statist. Assoc., 89, No. 427, 807-817, Theory and Methods.

[12] Scott D.W. (1992). *Multivariate Density Estimation* Theory, Practice and Visualization, John
and Sons, Inc.

[13] Simonoff, S.J. (1996). *Smoothing Methods in Statistics*, Springer-Verlag New York, Inc.

[14] Wand, M.P. and Jones, M.C. (1993). Comparison of Smoothing Parameterizations in Bivariate
Density estimation. *J. Amer. Statist. Assoc.*, 88, No. 422, Theory and Methods.

[15] Wand, M. P. and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Comput. Sc*
97116