

RESEARCH ARTICLE

Using small bias nonparametric density estimators for confidence interval estimation

Marco Di Marzio^a and Charles C. Taylor^{b*}

^a*DMQTE, Università di Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy;* ^b*Dept. of Statistics, University of Leeds, Leeds LS2 9JT, UK*

(.....)

Confidence intervals for densities built on the basis of standard nonparametric theory are doomed to have poor coverage rates due to bias. Studies on coverage improvement exist, but reasonably behaved interval estimators are needed. We explore the use of small bias kernel-based methods to construct confidence intervals, in particular using a geometric density estimator that seems better suited for this purpose.

Keywords: Bootstrap; Coverage rate; Geometric density estimators; Higher-order bias estimators; U-statistic.

AMS Subject Classification: Primary 62G07; secondary 62E20.

1. Introduction

Nonparametric density estimation is plagued by a bias problem, and much effort has been devoted to obtain modified estimators with a smaller bias. [11] perform an extensive, MISE based simulation study where many of these small bias, kernel-based estimators are compared. Their final advice favours the use of the standard kernel method in many situations.

Confidence intervals for nonparametric density estimates typically have poor coverage rates as a result of the bias problem. Bootstrap methods do not provide a remedy because the bootstrap expectation of a linear nonparametric estimator is the estimate itself. Hall [5] accurately treats bootstrap confidence intervals for kernel density estimation, and concludes that undersmoothing is preferable to explicit bias estimation. After observing that Hall's undersmoothing deteriorates the variance estimate, and consequently is unable to guarantee the promised coverage, [2] uses empirical likelihood to avoid this reported flaw. To date, it seems that off-the-shelf methods for confidence interval estimation of densities are still needed. In addition, the above studies do not give rules for practical bandwidth selection, and little account is taken of the expected width.

In this paper we explore the feasibility of confidence intervals on the basis of small bias density estimators. Apart from [5], who studies how undersmoothing of higher-order kernel estimators influences the coverage, this strategy has not been fully explored. A reason could be that often many small bias estimators neither

*Corresponding author. Email: c.c.taylor@leeds.ac.uk

have small theoretical minimum MISEs (as pointed out by [11]), nor possess efficient data-driven bandwidth selection. Another reason could be that these methods produce estimates that are not densities. But notice that here the coverage is our main target, therefore the performance of an estimator relies primarily on integrated squared bias; much less on MISE. In addition, since our final goal is a confidence interval, the fact that the estimate does not integrate to one is of secondary importance. However, the non-negativity constraint – violated by higher-order kernel estimators – remains relevant when estimating in the tails.

We focus on a couple of density estimators which implement the same bias reduction idea, one via a multiplicative structure, and the other one via an additive structure. Our simulation study compares other known reduced bias estimators for which it is straightforward to obtain a bandwidth selector. Having said that there is an edge for our methods, it seems that all the bias-reduction estimators tried give reasonable performance for confidence interval estimation, even for small samples.

In Section 2 we present the estimators. A number of different interpretations are available for them. Unifying views suggest that they rely on the same *twicing* principle or that they have a bootstrap nature. Other different interpretations are possible for the multiplicative version. In Section 3 we obtain asymptotic (integrated) mean squared errors. We also formulate normal-based bandwidth selectors, also for a number of estimators included in the paper of [11]. In Section 4 we provide theory to motivate confidence interval estimation, such as asymptotic normality, variance estimation and a Chi-square method to approximate bootstrap distributions. Section 5 contains a simulation study. Finally, some concluding remarks are given in Section 6. A few preliminaries follow.

Given a random sample X_1, \dots, X_n from an unknown density f of a continuous r.v. X , the usual kernel density estimate of f at x is

$$\hat{f}(x; h) := \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (1)$$

the function K , measurable and integrating to 1, is the kernel, the positive real number h is the bandwidth. If f has at least $p > 1$ derivatives in a neighbourhood of x , a Taylor series expansion gives

$$\mathbb{E}[\hat{f}(x; h)] - f(x) = \sum_{k=1}^{p-1} h^k \frac{(-1)^k}{k!} f^{(k)}(x) \mu_k(K) + O(h^p),$$

where $\mu_k(K) := \int u^k K(u) du$. If K is a density such that $\mu_1(K) = 0$ – as in the standard case – then the bias is $O(h^2)$. If an estimator has bias of order $O(h^p)$ with $p > 2$, we refer to it as a small (or higher-order) bias estimator. K is said to have order p if $\mu_k(K) = 0, 0 < k < p$, and $0 < \mu_p(K) < +\infty$.

2. The estimators

A multiplicative estimator of f at x is

$$\hat{f}_M(x; h) := \frac{\hat{f}^2(x; h)}{\hat{f}(x; 2^{1/2}h)};$$

it was originally motivated by the observation that the expectation of the smoothed bootstrap normal-kernel estimator is simply $\hat{f}(x; 2^{1/2}h)$. Hence, a standard bootstrap bias correction approach (see [3], pg. 103) leads to the additive estimator

$$\hat{f}_A(x; h) := 2\hat{f}(x; h) - \hat{f}(x; 2^{1/2}h)$$

or the multiplicative estimator \hat{f}_M . \hat{f}_A amounts to (1) equipped with the fourth order kernel $2K - K * K$, i.e. the “twicing” kernel proposed in fixed design regression by [14]. Although \hat{f}_A is simpler to analyze, we prefer \hat{f}_M because it cannot take negative values.

The estimator \hat{f}_M is already known, in the sense that [9] cite it as an example within a family of special cases of a more general technique. This family has been referred to by them as “generalized jackknifing on the log scale”, and \hat{f}_M is thus an example of a multiplicative form, akin to that of [15]. Although this is a multiplicative estimator, we note that it is distinct from that of [10]. As it will be seen below, \hat{f}_M has a smaller bias than \hat{f} , but at the price that it does not integrate to one. To make the estimator well defined, we require the various denominators to be strictly positive everywhere in the support. This appears a good reason for using Gaussian kernels.

3. Bandwidth selection

The natural L_2 risk measure for a generic estimator $\hat{f}(x)$ is $\text{MSE}[\hat{f}(x)] := \mathbb{E}[\{f(x) - \hat{f}(x)\}^2]$. But, in view of a more general usage, we consider the following global version of it

$$\text{MISE}[\hat{f}] := \mathbb{E} \left[\int (f - \hat{f})^2 \right].$$

In particular, for a kernel-type estimator $\hat{f}(\cdot; h)$, h is selected in order to minimize an estimate of an asymptotic version of $\text{MISE}[\hat{f}]$. Selectors of this kind are usually indicated as h_{AMISE} . We now give the asymptotic MSEs for our estimators.

Theorem 3.1: *Let X_1, \dots, X_n be a random sample from a density f of a continuous univariate r.v. X . Given the estimators $\hat{f}_M(x; h)$ and $\hat{f}_A(x; h)$, both equipped with kernel K , assume that*

- (1) f is bounded and continuous at x ; moreover $f^{iv}(x)$ exists and is finite;
- (2) the bandwidth h depends on n ; in particular $\lim_{n \rightarrow \infty} nh = \infty$, and $\lim_{n \rightarrow \infty} h = 0$;
- (3) K is a Gaussian density;

Then, at a point x in the support of f we have

$$\mathbb{E}[\hat{f}_M(x; h)] = f(x) + \frac{h^4}{4} \left\{ \frac{f''^2(x)}{f(x)} - f^{iv}(x) \right\} + O\{(nh)^{-1} + h^6\},$$

$$\text{VAR}[\hat{f}_M(x; h)] = 0.72 \frac{f(x)}{nh\pi^{1/2}} + O\{(nh)^{-2}\},$$

$$\text{MSE}[\hat{f}_M(x; h)] = \frac{h^8}{16} \left\{ \frac{f''^2(x)}{f(x)} - f^{iv}(x) \right\}^2 + 0.72 \frac{f(x)}{nh\pi^{1/2}} + O\{(nh)^{-2} + h^{10}\}$$

Table 1. Coefficients of $h_{\text{AMISE}} = c\hat{\sigma}n^{-1/9}$ for various small bias estimators: \hat{f}_M and \hat{f}_A are given in Section 2; \hat{f}_{FO} is the fourth-order kernel estimator; \hat{f}_{JF} is an estimator (explicitly given by (4) in [11]) of [9]; \hat{f}_{JLN} is that of [10]; \hat{f}_{HR} is an estimator from [7], and \hat{f}_{TS} indicates the variable bandwidth estimator of [16].

Estimator	\hat{f}_M	\hat{f}_A	\hat{f}_{FO}	\hat{f}_{JF}	\hat{f}_{JLN}	\hat{f}_{HR}	\hat{f}_{TS}
c	0.8928	0.9126	1.0834	0.9055	0.9642	0.8617	0.8124

and

$$\begin{aligned} \mathbb{E}[\hat{f}_A(x; h)] &= f(x) + \frac{h^4}{4} f^{iv}(x) + O\{h^6\}, \\ \text{VAR}[\hat{f}_A(x; h)] &= 0.72 \frac{f(x)}{nh\pi^{1/2}} + O\{(nh)^{-2}\}, \\ \text{MSE}[\hat{f}_A(x; h)] &= \frac{h^8}{16} f^{iv}(x)^2 + 0.72 \frac{f(x)}{nh\pi^{1/2}} + O\{(nh)^{-2} + h^{10}\} \end{aligned}$$

Both $\text{MSE}[\hat{f}_M(x; h)]$ and $\text{MSE}[\hat{f}_A(x; h)]$ fit into the general form of MSE expressions for small bias estimators in [11].

3.1. Normal Reference Bandwidth Selection

A very simple bandwidth selector for the usual \hat{f} is the *normal scale rule*. It results from a normal population assumption. This gives $h_{\text{NS}} = 1.06\hat{\sigma}n^{-1/5}$. We now give similar rules for many small bias estimators. [11] give the approximated AMSE of many $O(h^4)$ -bias estimators. All of these, together with the corresponding results for \hat{f}_M and \hat{f}_A , can be integrated under the normal assumption, then optimized over h . This leads to bandwidth selectors of the form $h_{\text{AMISE}} = c\hat{\sigma}n^{-1/9}$. The coefficients c are summarized in Table 1.

4. Confidence Interval Estimation

Denote as G an element of $\{A, M\}$. To construct a $100(1 - \alpha)\%$ confidence interval, we consider normal (\mathcal{I}_N) and bootstrap percentile (\mathcal{I}_B) methods:

$$\begin{aligned} \mathcal{I}_N &:= \left(\hat{f}_G - z_{\alpha/2} \text{V}\hat{\text{A}}\text{R}[\hat{f}_G]^{1/2}, \hat{f}_G + z_{\alpha/2} \text{V}\hat{\text{A}}\text{R}[\hat{f}_G]^{1/2} \right), \\ \mathcal{I}_B &:= (F_G^{*-1}(\alpha/2), F_G^{*-1}(1 - \alpha/2)), \end{aligned}$$

where $F_G^{*-1}(u) := \inf\{x : F_G^*(x) \geq u\}$, with $F_G^*(x)$ the bootstrap distribution of $\hat{f}_G(x)$, and $\text{V}\hat{\text{A}}\text{R}[\hat{f}_G]$ derived below.

The theoretical motivation for using a normal based confidence interval lies in the following

Theorem 4.1: *Given a random sample X_1, \dots, X_n , taken from a density f of a continuous univariate r.v. X , then at x in the support of f we have*

$$(nh)^{1/2} \{ \hat{f}_G(x; h) - \mathbb{E}[\hat{f}_G(x; h)] \} \xrightarrow{L} N \left(0, \frac{0.72f(x)}{\pi^{1/2}} \right)$$

if,

- in the case $G = A$, the assumptions of Theorem 3.1 hold;

- in the case $G = M$, the assumptions of Theorem 3.1 hold, and, in addition $E[|\mathcal{K}(X_i, X_j)|^2] < \infty$ $1 \leq i, j \leq n$, where

$$\begin{aligned} \mathcal{K}(X_i, X_j) := & \frac{1}{2} \{K_h(x - X_i)K_h(x - X_j) - K_{2^{1/2}h}(x - X_i) \\ & + K_h(x - X_j)K_h(x - X_i) - K_{2^{1/2}h}(x - X_j)\}. \end{aligned}$$

Finally, if $nh^5 \rightarrow 0$, the convergence holds true also with $f(x)$ in place of $E[\hat{f}_G(x; h)]$.

Proof: See Appendix D. □

An alternative to the Normal-based confidence interval, which also avoids the resampling process, follows. By analogy with the estimation of a spectral density [17], approximate $F_G^*(x)$ by a scaled χ^2 distribution:

$$\int_0^x a(u) \chi_{b(u)}^2(u) du$$

with $a(u), b(u)$ chosen to match the mean and variance of $f_G^*(u)$:

$$a(x) := \frac{\text{VAR}_*[f_G^*(x)]}{2E_*[f_G^*(x)]}, \quad b(x) := \frac{2\{E_*[f_G^*(x)]\}^2}{\text{VAR}_*[f_G^*(x)]}.$$

where E_* and VAR_* are taken with respect to the bootstrap distribution. This leads to a third method denoted as \mathcal{I}_{χ^2} .

To obtain an estimator of the variance of $\hat{f}_M(x; h)$, we express it by the expansion in Lemma A of the Appendix – this approximation includes all of the terms of order $O\{(nh)^{-2}\}$, having a residual of order $O\{(nh)^{-3}\}$ – then we replace $f(x)$ by $\hat{f}(x; h)$. We obtain

$$\begin{aligned} \hat{\text{V}}\text{AR}[\hat{f}_M(x; h)] \simeq & \left[\frac{E[\hat{f}^2(x; h)]}{E[\hat{f}(x; 2^{1/2}h)]} \right]^2 \left[\frac{\text{VAR}[\hat{f}^2(x; h)]}{E[\hat{f}^2(x; h)]^2} + \frac{\text{VAR}[\hat{f}(x; 2^{1/2}h)]}{E[\hat{f}(x; 2^{1/2}h)]^2} \right. \\ & \left. - \frac{2\text{cov}[\hat{f}^2(x; h), \hat{f}(x; 2^{1/2}h)]}{E[\hat{f}^2(x; h)]E[\hat{f}(x; 2^{1/2}h)]} \right] + O\{(nh)^{-3}\}, \end{aligned} \quad (2)$$

The estimator $\hat{\text{V}}\text{AR}[\hat{f}_A(x; h)]$ is obtainable by similar calculations. In Lemma E the explicit expressions for the estimators involved in the above formula are reported. We notice that $\hat{\text{V}}\text{AR}[\hat{f}_M(x; h)]$ is made of ratios with the same bias order both in the numerator and in the denominator. So the bias of $\hat{\text{V}}\text{AR}[\hat{f}_M(x; h)]$ is strongly reduced just like the bias of $\hat{f}_M(x; h)$. We note that, in our simulations with small sample sizes, this estimate of the variance is occasionally negative far in the tails of the distributions. In this case bootstrap intervals can be used.

5. Simulations

We illustrate the above using the results of a simulation study where the smoothing degree is data-driven, and therefore we hope that our results are of some practical relevance. Of course, we use normal-based selectors, which are well suited only in the presence of unimodal populations. Nevertheless, also a couple of bimodal populations – for which only local bandwidths seem adequate – are included in

Table 2. Coverages and average widths for various 95% confidence interval estimators at $x = 0, 0.75, 1.5$ of a standard normal, and a bimodal density. Methods are: bootstrap percentile; normal approximation; Chi-square approximation. Averages over 100 000 simulations.

		observed coverage rates (average width)					
		$N(0, 1)$			$0.5N(0, 1) + 0.5N(3, 1)$		
	x	0	0.75	1.5	0	0.75	1.5
n = 50	\mathcal{I}_B	93.3 (0.222)	94.3 (0.204)	94.1 (0.151)	81.9 (0.122)	94.4 (0.109)	65.6 (0.093)
	\mathcal{I}_N	91.6 (0.202)	94.0 (0.193)	93.9 (0.152)	74.1 (0.107)	95.2 (0.107)	81.6 (0.107)
	\mathcal{I}_{χ^2}	92.4 (0.202)	94.4 (0.192)	93.2 (0.150)	78.0 (0.106)	94.8 (0.107)	76.0 (0.107)
n = 100	\mathcal{I}_B	93.6 (0.163)	94.7 (0.151)	94.2 (0.112)	77.7 (0.090)	94.6 (0.081)	52.3 (0.070)
	\mathcal{I}_N	92.3 (0.153)	94.3 (0.145)	94.7 (0.114)	69.8 (0.081)	95.0 (0.081)	71.0 (0.080)
	\mathcal{I}_{χ^2}	92.9 (0.153)	94.5 (0.145)	93.8 (0.114)	73.0 (0.081)	94.6 (0.081)	66.0 (0.080)

our study. This is in order to check how the performance deteriorates in such scenarios. Curiously, to the best of our knowledge, data driven smoothing is new both for higher order kernels and kernel-based confidence intervals. In particular, higher order estimators have been compared on the basis of their best possible MISEs, while in the only two empirical studies existing on confidence intervals based on kernel density estimators (see [5] and [2]) the coverage rates are reported at predetermined smoothing levels.

In what follows kernels are Gaussian; the bandwidths are given by the normal-based plug-in rules specified in Table 1; the confidence levels are $1 - \alpha = 0.95$, and, finally, the number of bootstrap samples is 1000.

5.1. Interval estimation

As a first case study, we use the setup of [5]: estimate the standard normal, and a symmetric, bimodal, normal mixture at $x = 0, 0.75, 1.5$; use $n = 50$ and $n = 100$. Also [2] estimates the standard normal density at 0 with $n = 50$.

Our results – contained in Table 2 – are averages over 100 000 simulations. It can be seen that our coverage favorably compares with those of [5] and [2]. In particular, \mathcal{I}_N works also for these small sized samples, yet its theoretical motivation has an asymptotic nature. It is noticeable that \mathcal{I}_{χ^2} also performs well. For the bimodal density, the coverage at $x = 1.5$, which is a local minimum, is, as expected, poor. But it is still superior to most of the performances seen in [5].

The second case study is more general: we estimate models #1 (Gaussian), #2 (Skewed Unimodal) and #6 (Bimodal) of [12] in $[-3, 3]$; moreover a Student t with five degrees of freedom in $[-4, 4]$. Sample sizes are 50 and 500. As a motivation for this choice consider that we have included the main three unimodal models, i.e. symmetric, skewed and heavy tailed, to obtain a more general conclusion on the matter. We compute bootstrap percentile confidence intervals based on various estimators. Concerning our choice of estimators, consider what follows. It is possible to divide small bias methods into two categories, depending on their output: “positive” estimators and “negative” ones. Now, from the extensive comparative study provided by [11], it results that excellent candidates to represent these categories are, respectively, the proposal of [10] (\hat{f}_{JLN}), and the fourth-order kernel estimator in Section 2.1 of [11] (\hat{f}_{FO}). As a benchmark, also \hat{f} is included.

We adopt the following performance indices:

$$P := \int p(x)f(x)dx,$$

Table 3. Integrated performance measures $\bar{P}, \bar{W}, \bar{O}$ (10 000 simulations) for a variety of bootstrap 95% percentile confidence intervals (indicated by the corresponding point estimator symbol).

		Performance Measures: $100\bar{P}, \bar{W}, 100\bar{O}$							
		#1	#2	#6	$t(5)$				
n = 50	\hat{f}	89.4, 0.167, 0.94	86.7, 0.214, 1.90	81.0, 0.140, 2.00	84.0, 0.140, 1.56				
	\hat{f}_M	93.3, 0.191, 0.28	92.4, 0.243, 0.55	87.0, 0.162, 1.26	91.2, 0.159, 0.41				
	\hat{f}_{JLN}	90.3, 0.167, 0.71	87.2, 0.215, 1.69	74.8, 0.138, 2.86	85.7, 0.141, 1.26				
	\hat{f}_{FO}	92.8, 0.183, 0.36	91.5, 0.233, 0.78	84.3, 0.153, 1.63	89.6, 0.152, 0.71				
n = 500	\hat{f}	90.3, 0.0722, 0.35	85.5, 0.0904, 0.93	73.6, 0.0610, 1.30	81.3, 0.0578, 0.83				
	\hat{f}_M	94.0, 0.0713, 0.06	91.4, 0.0891, 0.31	72.6, 0.0607, 1.30	90.1, 0.0579, 0.26				
	\hat{f}_{JLN}	92.5, 0.0643, 0.15	86.1, 0.0801, 0.79	54.3, 0.0536, 2.26	82.6, 0.0516, 0.70				
	\hat{f}_{FO}	93.5, 0.0684, 0.09	89.5, 0.0853, 0.50	66.9, 0.0579, 1.62	85.6, 0.0547, 0.52				

$$W := \int w(x)f(x)dx,$$

the expectations of the coverage (p) and width (w). Strictly, narrower intervals are of importance only when the desired coverage is attained, so the trade-off we have used is

$$O := \int |1 - \alpha - p(x)|w(x)f(x)dx.$$

Table 3 gives the results for each of the measures P, W, O calculated on 10 000 samples. It can be seen from Table 3 that small bias methods give much better coverage than \hat{f} , recalling that the bandwidth is always automatically selected. The results for \hat{f}_A (not shown) were quite similar to those of \hat{f}_M , but not quite as good. Overall, it seems that \hat{f}_M is well behaved for the unimodal case. In order to investigate why \hat{f}_M seems to outperform \hat{f}_{JLN} and \hat{f}_{FO} , we now consider a more typical analysis of performance in point estimation.

5.2. Point estimation

For the same models and estimators as before, we have calculated the usual L_2 integrated discrepancies. For each model 10 000 samples were drawn. Each column of Table 4 gives the ratio of MISE, integrated variance and integrated bias-squared between an element of $\{\hat{f}_M, \hat{f}_{JLN}, \hat{f}_{FO}\}$ and those of \hat{f} . As can be seen from Table 4, \hat{f}_M is the best in bias reduction, even though it is not so good at minimizing MISE. Now, note that in presence of bias (small bias estimators are still biased) we will need the variance to have a bigger magnitude than bias, in order to get the right coverage. But this is exactly the case of \hat{f}_M , on the basis of Table 4, where we can observe comparatively small biases and big variances. In conclusion, the point estimation results, if read from a confidence interval perspective, explain a certain superiority of \hat{f}_M .

5.3. On normalizing the estimators

It is well known that small bias methods produce estimates that do not integrate to 1, and/or take negative values. A large number of techniques that transform these estimates into densities have been proposed; see [4]. Nevertheless, we have preferred to not involve estimate corrections. This is simply to avoid linking the

Table 4. Integrated performance measures for a variety of estimators and models. Ratios of estimators ($\tilde{f} \in \{\hat{f}_M, \hat{f}_{JLN}, \hat{f}_{FO}\}$), for MISE, integrated variance, and integrated bias-squared, relative to those of \hat{f} , i.e. $MISE[\tilde{f}]/MISE[\hat{f}]$, $IVAR[\tilde{f}]/IVAR[\hat{f}]$, and $IBIAS[\tilde{f}]/IBIAS[\hat{f}]$. Quantities are averages over 10 000 simulations.

	#1			#2			#6			t(5)		
	MISE	IVAR	IBIAS	MISE	IVAR	IBIAS	MISE	IVAR	IBIAS	MISE	IVAR	IBIAS
n = 50												
\hat{f}_M	1.035	1.270	0.192	0.897	1.245	0.324	1.020	1.318	0.767	0.866	1.223	0.322
\hat{f}_{JLN}	0.853	0.982	0.392	0.829	0.988	0.568	1.069	0.951	1.170	0.816	1.003	0.532
\hat{f}_{FO}	0.980	1.196	0.206	0.890	1.186	0.403	1.027	1.208	0.873	0.882	1.182	0.426
n = 500												
\hat{f}_M	0.790	0.964	0.205	0.722	0.958	0.418	1.020	0.984	1.039	0.788	0.964	0.209
\hat{f}_{JLN}	0.673	0.784	0.300	0.716	0.783	0.630	1.287	0.765	1.577	0.672	0.786	0.300
\hat{f}_{FO}	0.756	0.913	0.230	0.744	0.911	0.529	1.118	0.915	1.231	0.756	0.914	0.235

performance of an estimator — both absolute and relative — to a subjective choice of the correction method. It would be a *subjective choice* exactly because the formal properties of these estimators refer to the uncorrected versions. The only fair alternative could have been to select the best correction method for each pair {estimator, model}, but this seems a long way from practical usage. However, we note that the correction subject seems problematic, for example, [4] show that the simple dividing by the integral of the estimate – inappropriate for correcting higher-order kernel methods – could even deteriorate the performance, depending on the model to estimate, and with no way to know this in advance from the data.

6. Concluding Remarks

Higher-order bias methods have been much studied in kernel density estimation, but are less used. Given that, in some cases, explicit bias correction of an ordinary kernel is essentially equivalent to using a small bias estimator [8], there seems to be justification for using such methods when the goal is confidence interval estimation rather than point estimation. In this case, it seems that the strength of any method lies mainly in its ability to reduce bias with the availability of a suitable plug-in rule for the smoothing parameter. Further work could extend these methods to nonparametric regression, which could also be incorporated in hypothesis testing, for example, in tools such as SiZer [1].

Finally, we note that our data-based smoothing parameters are chosen to minimize AMISE (under a normal assumption). However (as also pointed out by [5]) there is absolutely no reason that an adequate choice for the bandwidth which minimizes MISE will be the correct one in terms of coverage accuracy. However, our simulations suggest that these AMISE-bandwidth selectors may nevertheless provide a good trade-off between coverage and expected width in many situations. Moreover, practical selectors which “optimize” the coverage do not yet exist.

Acknowledgements

The authors are grateful to the Editor, Associate Editor and two anonymous referees for their valuable comments which led to considerable improvements in this article.

References

- [1] P. Chaudhuri and J.S. Marron, *SiZer for exploration of structures in curves*, J. Amer. Statist. Assoc. 94 (1999), pp. 807–823.
- [2] S.X. Chen, *Empirical likelihood confidence intervals for nonparametric density estimation*, Biometrika 83 (1996), pp. 329–341.
- [3] A.C. Davison and D.V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press, 1997.
- [4] I.K. Glad, N.L. Hjort, and N. Ushakov, *Correction of density estimators that are not densities*, Scand. J. Statist. 30 (2003), pp. 415–427.
- [5] P. Hall, *Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density*, Ann. Statist. 20 (1992), pp. 675–694.
- [6] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, Ann. Math. Statist. 19 (1948), pp. 293–325.
- [7] O. Hössjer and D. Ruppert, *Asymptotics for the transformation kernel density estimator*, Ann. Statist. 23 (1995), pp. 1198–1222.
- [8] M.C. Jones, *On higher order kernels*, J. Nonparametr. Stat. 5 (1995), pp. 215–221.
- [9] M.C. Jones and P.J. Foster, *Generalized Jackknifing and higher order kernels*, J. Nonparametr. Stat. 3 (1993), pp. 81–94.
- [10] M.C. Jones, O. Linton and J.P. Nielsen, *A simple bias reduction method for density estimation*, Biometrika 82 (1995), pp. 327–38.
- [11] M.C. Jones and D.F. Signorini, *A comparison of higher-order bias kernel density estimators*, J. Amer. Statist. Assoc. 92 (1997), pp. 1063–1073.
- [12] J.S. Marron and M.P. Wand, *Exact mean integrated squared error*, Ann. Statist. 20 (1992), pp. 712–736.
- [13] R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.
- [14] W. Stuetzle and Y. Mittal, *Some comments on the asymptotic behavior of robust smoothers*, in *Lecture Notes in Math 757: Smoothing Techniques for Curve Estimation*, T. Gasser and M. Rosenblatt, eds., Springer-Verlag, Berlin, 1979, pp. 191–195.
- [15] G.R. Terrell and D.W. Scott, *On improving convergence rates for nonnegative kernel density estimators*, Ann. Statist. 8 (1980), pp. 1160–1163.
- [16] G.R. Terrell and D.W. Scott, *Variable kernel density estimation*, Ann. Statist. 20 (1992), pp. 1236–1265.
- [17] J.W. Tukey, *The sampling theory of power spectrum estimates in Proc. Symp. on Applications of Autocorrelation Analysis to Physical Problems, NAVEXOS-P-735*, Office of Naval Research, Department of the Navy, Washington, USA, 1949, pp. 47–67.

Appendix A. Lemma A

Consider two continuous, real random variables X and Y , If both μ_X and μ_Y are non-zero, and $\text{VAR}[X/Y]$ exists finite, then

$$\text{VAR} \left[\frac{X}{Y} \right] \simeq \left(\frac{\mu_X}{\mu_Y} \right)^2 \left(\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Y^2}{\mu_Y^2} - \frac{2\sigma_{XY}}{\mu_X \mu_Y} \right),$$

provided that all the involved moments are finite.

Proof: This standard result is obtained by calculating the expectation of the second-order bivariate Taylor expansion of X/Y in a neighborhood of (μ_X, μ_Y) . \square

Appendix B. Lemma B

Consider a random sample X_1, \dots, X_n , taken from a continuous univariate density f . Let $\bar{\phi}(h) := \text{E}[K_h(x - X_1)]$ where x belongs to the support of f . Assume that the kernel K is a Gaussian density. Then

$$\text{E}[\hat{f}^2(x; h)] = \frac{1}{n} \left\{ (n-1) \bar{\phi}(h)^2 + \frac{\bar{\phi}(h/2^{1/2})}{2\pi^{1/2}h} \right\},$$

$$\text{E}[\hat{f}(x; 2^{1/2}h)] = \bar{\phi}(2^{1/2}h),$$

$$\text{VAR}[\hat{f}(x; 2^{1/2}h)] = \frac{1}{n} \left\{ \frac{\bar{\phi}(h)}{8^{1/2}\pi^{1/2}h} - \bar{\phi}(2^{1/2}h)^2 \right\},$$

$$\begin{aligned} n^4 \text{VAR}[\hat{f}^2(x; h)] &= n \left\{ \frac{\bar{\phi}(h/2)}{32^{1/2}\pi^{3/2}} h^{3/2} - \frac{\bar{\phi}(h/2^{1/2})^2}{4\pi h^2} \right\} + \frac{2n!}{(n-2)!} \left\{ \frac{\bar{\phi}(h/2^{1/2})^2}{4\pi h^2} - \bar{\phi}(h)^4 \right\} \\ &+ \frac{4n!}{(n-3)!} \left[\bar{\phi}(h)^2 \left\{ \frac{\bar{\phi}(h/2^{1/2})}{2\pi^{1/2}h} - \bar{\phi}(h)^2 \right\} + \frac{\bar{\phi}(h)}{n-3} \left\{ \frac{\bar{\phi}(h/3^{1/2})}{12^{1/2}\pi h^2} - \frac{\bar{\phi}(h/2^{1/2})}{2\pi^{1/2}h} \bar{\phi}(h) \right\} \right], \end{aligned}$$

$$\begin{aligned} \text{COV}[\hat{f}^2(x; h), \hat{f}(x; 2^{1/2}h)] &= \frac{n!}{(n-2)!} \left\{ (n-2)\bar{\phi}(h)^2\bar{\phi}(2^{1/2}h) + \frac{\bar{\phi}(h/2^{1/2})}{2\pi^{1/2}h} \bar{\phi}(2^{1/2}h) \right. \\ &+ \left. \frac{2\bar{\phi}((2/3)^{1/2}h)}{6^{1/2}\pi^{1/2}h} \bar{\phi}(h) \right\} + \frac{n}{20^{1/2}\pi h^2} \bar{\phi}((2/5)^{1/2}h) - \frac{1}{n} \left\{ (n-1)\bar{\phi}(h)^2 + \frac{\bar{\phi}(h/2^{1/2})}{2\pi^{1/2}h} \right\}. \end{aligned}$$

Proof: The first three equations are immediate. Set $Y_i := x - X_i$, now

$$\text{VAR}[\hat{f}^2(x; h)] = \frac{1}{n^4} \text{VAR} \left[\sum \sum K_h(Y_i) K_h(Y_j) \right],$$

then $n^4 \text{VAR}[\hat{f}^2(x; h)]$ is equal to

$$\begin{aligned} n \text{VAR} [K_h(Y_1)^2] &+ 2 \frac{n!}{(n-2)!} \text{VAR} [K_h(Y_1) K_h(Y_2)] + \frac{n!}{(n-4)!} \text{COV} [K_h(Y_1) K_h(Y_2), K_h(Y_3) K_h(Y_4)] \\ &+ 4 \frac{n!}{(n-3)!} \text{COV} [K_h(Y_1) K_h(Y_2), K_h(Y_1) K_h(Y_3)] + 2 \frac{n!}{(n-3)!} \text{COV} [K_h(Y_1) K_h(Y_1), K_h(Y_2) K_h(Y_3)] \\ &+ 4 \frac{n!}{(n-2)!} \text{COV} [K_h(Y_1) K_h(Y_1), K_h(Y_1) K_h(Y_2)] + \frac{n!}{(n-2)!} \text{COV} [K_h(Y_1) K_h(Y_1), K_h(Y_2) K_h(Y_2)]. \end{aligned}$$

Assuming that the kernel is Gaussian, a little algebra leads to the result.

Concerning the mixed moment of $\text{COV}[\hat{f}^2(x; h), \hat{f}(x; 2^{1/2}h)]$, we have

$$\begin{aligned} n^3 \sum \sum \sum \mathbb{E}[K_h(Y_i) K_h(Y_j) K_{2^{1/2}h}(Y_s)] &= \frac{n! \mathbb{E}[K_h(Y_1) K_h(Y_2) K_{2^{1/2}h}(Y_3)]}{(n-3)!} \\ &+ \frac{n! \mathbb{E}[K_h(Y_1)^2 K_{2^{1/2}h}(Y_2)]}{(n-2)!} + 2 \frac{n! \mathbb{E}[K_h(Y_1) K_h(Y_2) K_{2^{1/2}h}(Y_1)]}{(n-2)!} + n \mathbb{E}[K_h(Y_1)^2 K_{2^{1/2}h}(Y_1)]. \end{aligned}$$

Assuming that the kernel is Gaussian, again a little algebra leads to the result.

□

Appendix C. Proof of Theorem 3.1

The proof is based on a “linearization” argument. First of all, we note that

$$\{ \hat{f}_M(x; h) - f(x) \} - \left\{ \frac{\hat{f}_h(x; h)^2 - f(x) \hat{f}(x; 2^{1/2}h)}{f(x)} \right\} \sim o_p \left\{ (nh)^{-1/2} \right\},$$

and so the bias is approximated by the expectation of the second term, *i.e.*

$$\frac{\mathbb{E}[\hat{f}(x; h)^2] - f(x)\mathbb{E}[\hat{f}(x; 2^{1/2}h)]}{f(x)} = \frac{h^4}{4} \left\{ \frac{f''(x)^2}{f(x)} - f^{iv}(x) \right\} + O(h^6) + O\{(nh)^{-1}\}.$$

Now let $f_n(x) := \mathbb{E}[\hat{f}(x; h)^2]/\mathbb{E}[\hat{f}(x; 2^{1/2}h)]$ and note that

$$\hat{f}_M(x; h) - f_n(x) = \left[\frac{\hat{f}(x; h)^2}{f(x)} - \frac{\hat{f}(x; 2^{1/2}h)f_n(x)}{f(x)} \right] \frac{f(x)}{\hat{f}(x; 2^{1/2}h)}.$$

Since

$$\frac{f(x)}{\hat{f}(x; 2^{1/2}h)} \xrightarrow{p} 1$$

the variance of the LHS is equal to the variance of the term in square brackets. Lemma B provides the various elements. For an approximate version of $\text{VAR}[\hat{f}_M(x)]$, consider that $\bar{\phi}(h) = f(x) + h^2 f''(x)\mu_2(K) + O(h^4)$ and replace $\bar{\phi}(h)$ with $f(x)$. Finally consider that K is a Gaussian density. To get the asymptotic moments of $\hat{f}_A(x; h)$ apply Lemma B, then approximate as above.

Appendix D. Proof of Theorem 4.1

We have

$$\hat{f}_A(x; h) = \frac{1}{nh} \sum_{i=1}^n H\left(\frac{x - X_i}{h}\right)$$

where $H(z) := 2K(z) - 1/2^{1/2}K(z/2^{1/2})$, a higher order kernel. But the Lindberg condition holds, as for the standard kernel, as follows. For any $\epsilon > 0$,

$$\begin{aligned} & h^{-1} \mathbb{E} \left[H\left(\frac{x - X_1}{h}\right)^2 I_{\{|H(\frac{x-X_1}{h}) - \mathbb{E}[H(\frac{x-X_1}{h})]| > \sqrt{nh}\epsilon\}} \right] \\ &= \int_{\{|H(y) - \mathbb{E}[H(\frac{x-X_1}{h})]| > \sqrt{nh}\epsilon\}} H(y)^2 f(x - hy) dy \end{aligned}$$

which converges to 0 if $nh \rightarrow \infty$.

Concerning \hat{f}_M , as seen in the proof of Theorem 3.1, $\hat{f}_M(x; h) - f_n(x)$ and $\hat{f}(x; h)^2/f(x) - \hat{f}(x; 2^{1/2}h)f_n(x)/f(x)$ have the same asymptotic distribution. So it is sufficient to prove that

$$n^{1/2}\{V_n - \mathbb{E}[V_n]\} \xrightarrow{L} N(0, 4\zeta_1)$$

with $V_n := \hat{f}(x; h)^2 - \hat{f}(x; 2^{1/2}h)$ and $\zeta_1 := \text{VAR}[\int \mathcal{K}(X_1, x_2)f(x_2) dx_2] > 0$.

We have

$$V_n = \frac{1}{n^2} \sum_{1 \leq i, j \leq n} \mathcal{K}(X_1, X_2),$$

Now observe that $\mathcal{K}(X_i, X_j) = \mathcal{K}(X_j, X_i)$ $1 \leq i, j \leq n$, therefore V_n is a von Mises statistic. Now define the U-statistic $V_n^* := \{n(n-1)\}^{-1} \sum \sum_{1 \leq i \neq j \leq n} \mathcal{K}(X_i, X_j)$. Provided that $E[|\mathcal{K}(X_i, X_j)|^2] < \infty$ $1 \leq i, j \leq n$, we have $n^{1/2}(V_n - V_n^*) \xrightarrow{P} 0$ (see [13], pg. 206). Now,

$$n^{1/2}\{V_n^* - E[V_n^*]\} \xrightarrow{L} N(0, 4\zeta_1)$$

provided that $\text{VAR}[\tilde{\mathcal{K}}(X_1)] > 0$, where $\tilde{\mathcal{K}}(x_1) := E[\mathcal{K}(x_1, X_2)]$ (see [6]). But observe that

$$E[\mathcal{K}(x_1, X_2)] = K_h(x_1 - x)E[K_h(X_2 - x)] - \frac{1}{2}\{K_{2^{1/2}h}(x_1 - x) + E[K_{2^{1/2}h}(X_2 - x)]\},$$

which is not degenerate.

Finally, if $nh^5 \rightarrow 0$,

$$\lim_{n \rightarrow \infty} \{E[\hat{f}_G(x; h)] - f(x)\} = \lim_{n \rightarrow \infty} (nh)^{1/2}O(h^4) = 0.$$

Appendix E. Lemma E

The expressions of the estimators contained in formula (2) are listed below

$$\hat{E}[\hat{f}^2(x; h)] = \frac{(n-1)\hat{f}^2(x; 2^{1/2}h) + \hat{f}(x; (3/2)^{1/2}h)/(2\pi^{1/2}h)}{n^2};$$

$$\hat{E}[\hat{f}(x; 2^{1/2}h)] = \hat{f}(x; 3^{1/2}h);$$

$$\hat{\text{VAR}}[\hat{f}(x; 2^{1/2}h)] = \frac{1}{n} \left\{ \frac{\hat{f}(x; 2^{1/2}h)}{8^{1/2}\pi^{1/2}h} - \hat{f}^2(x; 3^{1/2}h) \right\};$$

$$\begin{aligned} n^3 \hat{\text{VAR}}[\hat{f}^2(x)] &= 4(n-1)(n-2)\hat{f}^2(x; 2^{1/2}h) \left\{ \frac{\hat{f}(x; (3/2)^{1/2}h)}{2\pi^{1/2}h} - \hat{f}^2(x; 2^{1/2}h) \right\} \\ &+ \left\{ \frac{\hat{f}(x; (5/4)^{1/2}h)}{32^{1/2}\pi^{3/2}h^3} - \frac{\hat{f}^2(x; (3/2)^{1/2}h)}{4\pi h^2} \right\} + 2(n-1) \left\{ \frac{\hat{f}^2(x; (3/2)^{1/2}h)}{4\pi h^2} - \hat{f}^4(x; 2^{1/2}h) \right\} \\ &+ 4(n-1)\hat{f}(x; 2^{1/2}h) \left\{ \frac{\hat{f}(x; (4/3)^{1/2}h)}{12^{1/2}\pi h^2} - \frac{\hat{f}(x; 2^{1/2}h)\hat{f}(x; (3/2)^{1/2}h)}{2\pi^{1/2}h} \right\}; \end{aligned}$$

$$\begin{aligned} n^2 \hat{\text{CÔV}} \left[\hat{f}^2(x; h), \hat{f}(x; 2^{1/2}h) \right] &= \frac{2(n-1)}{6^{1/2}\pi^{1/2}h} \hat{f}(x; 2^{1/2}h)\hat{f}(x; (5/3)^{1/2}h) \\ &- \frac{\hat{f}(x; (3/2)^{1/2}h)\hat{f}(x; 3^{1/2}h)}{2\pi^{1/2}h} - 2(n-1)\hat{f}^2(x; 2^{1/2}h)\hat{f}(x; 3^{1/2}h) + \frac{\hat{f}(x; (7/5)^{1/2}h)}{20^{1/2}\pi h^2}. \end{aligned}$$