ESAIM: PS January 2009, Vol. 13, p. 1–14
DOI: 10.1051/ps:2007055

ESAIM: Probability and Statistics
www.esaim-ps.org

ASYMPTOTIC UNBIASED DENSITY ESTIMATORS

Nicolas W. Hengartner[1] and Éric Matzner-Løber[2]

Abstract. This paper introduces a computationally tractable density estimator that has the same asymptotic variance as the classical Nadaraya-Watson density estimator but whose asymptotic bias is zero. We achieve this result using a two stage estimator that applies a multiplicative bias correction to an oversmooth pilot estimator. Simulations show that our asymptotic results are available for samples as low as $n = 50$, where we see an improvement of as much as 20% over the traditionnal estimator.

Mathematics Subject Classification. 62G07, 62G20.

Received March 29, 2007.

Introduction

Under classical regularity conditions, the kernel density estimators from an i.i.d. n-sample from a distribution F with twice continuously differentiable densities f satisfy a central limit theorem with bandwidths of order $h_n = cn^{-1/5}$

$$\sqrt{nh_n}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{1}{h_n}K\left(\frac{X_j - x}{h_n}\right) - f(x)\right) \Longrightarrow \mathcal{N}(b(x), \sigma^2(x)), \tag{1}$$

with asymptotic bias and variance

$$b(x) = \frac{c^{5/2}}{2}f''(x)\int u^2K(u)\mathrm{d}u \quad \text{and} \quad \sigma^2(x) = f(x)\left(\int K^2(u)\mathrm{d}u\right),$$

respectively. See for example the book of Wand and Jones [19]. This theorem is used to justify the practice of drawing pointwise 2σ confidence intervals about the estimated density. While suggestive, these intervals do not account for the asymptotic bias $b(x)$. When the bias is non-zero, the resulting asymptotic coverage probability is less than the claimed 95%. This motivates our interest in density estimators whose bias is zero.

Estimators having zero asymptotic bias are not trivial. Rosenblatt [16] showed that if the considered class of densities is rich enough (which is typically the case in classical nonparametric density estimation), there does not exist unbiased density estimators for any finite sample n. Nevertheless, it is possible that asymptotically, the scaled bias $\sqrt{nh_n}\left(\mathbb{E}\left[\widehat{f}_n(x)\right] - f(x)\right)$ converges to zero. Examples of the latter are kernel estimators with

Keywords and phrases. Nonparametric density estimation, kernel smoother, asymptotic normality, bias reduction, confidence intervals.

[1] Stochastics Group, Los Alamos National Laboratory, NM 87545, USA.

[2] UMR 6625, IRMAR, Université Rennes 2, 35043 France; eml@uhb.fr.

Article published by EDP Sciences

small bandwidth sequence $h_n = o(n^{-1/5})$. But because in such instances, $n^{-2/5}\sqrt{nh_n} \longrightarrow \infty$, these estimators do not converge at the optimal rate and the resulting confidence intervals will be too wide.

Many methods, starting with the variable bandwidth estimators of Abrahamson [1], have attempted to reduce the asymptotic bias. The strategy of these estimators is to use a larger bandwidth in low density areas and smaller bandwidths in high density areas. The result is that for four times differentiable densities, the bias is of order $O(h^4)$ instead of the usual $O(h^2)$. The variance, while larger, remains of order $O((nh)^{-1})$. The interested reader about variable bandwidth kernel estimators is referred to Abramson [2], Hall and Marron [7], Hall [6], Jones [10], Jones *et al.* [12].

Higher order kernels achieve a similar reduction in bias under the same assumption and were studied by Granovsky and Müller [5], Marron and Wand [14], and Berlinet [4]. Other methods estimate adjustments to the estimated density to reduce the bias. For example, McKay [13] considers additive bias corrections to the density while Terell and Scott [18] apply the additive bias correction to the logdensity. Jones *et al.* [11], extending an idea of Nielson and Linton [15], estimate directly a multiplicative adjustment to the density.

However, the bandwidth sequence $h_n = cn^{-1/5}$ that is optimal for twice differentiable densities, significantly undersmooths densities assumed to be four times differentiable, an assumption made throughout the bias reduction literature. As a result, the resulting density estimators converge at a sub-optimal rate given the smoothness assumptions.

Finally, Hjort and Glad [8] and Hjort and Jones [9] have studied a multiplicative adjustment to a parametric family of densities. These estimators are unbiased when the true density belongs to the parametric family, and has smaller bias than the kernel estimator in a neighborhood of the parametric family while maintaining the same variance. Their result is important because they show how to decrease, or even eliminate, the bias without altering the smoothness assumptions made on the true density, changing the convergence rate, or even increasing the variance of the estimator.

This paper considers simple kernel based density estimators that are both rate-optimal and have, for every fixed density f, zero asymptotic bias. The major contribution is the introduction of a two stage estimator that corrects the bias without affecting the asymptotic variance, and this, without having to assume additional smoothness on the density. Pointwise confidence intervals based on these estimators have asymptotically the claimed coverage probability and their width converge to zero at the optimal rate.

1. Asymptotic unbiased density estimators

1.1. Higher order kernel estimators

We start this section by giving an example of a density estimate that converges at the optimal rate and for which the asymptotic bias $\sqrt{nh_n}\left(\mathbb{E}\left[\widehat{f}_n(x)\right] - f(x)\right) \to 0$. Let $X_1, \ldots, X_n$, be an n-sample from an unknown distribution F with twice differentiable density f. Consider the kernel density estimator

$$\widehat{f}(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(x - X_j),$$

where $K_h(u) = (1/h)K(u/h)$. While it is usual to match the order of the kernel with the number of assumed derivatives of the density f, one can obtain asymptotically unbiased density estimators by using a higher order kernel. The usual bias-variance calculations for twice differentiable densities yield

$$\mathbb{E}[\widehat{f}(x) - f(x)] = \int K(u)\left\{f(x - hu) - f(x)\right\}\mathrm{d}u = \frac{h^2}{2}\int u^2 K(u) f''(\xi_u)\mathrm{d}u \tag{2}$$

for some $\xi_u \in (x, x - hu)$, and

$$\mathrm{Var}(\widehat{f}(x)) = \left\{\frac{f(x)}{nh}\int K^2(u)\mathrm{d}u - \frac{f(x)^2}{n}\right\}(1 + o(h)).$$

If the density is twice continuously differentiable and $K(\cdot)$ compactly supported, an application of the dominated convergence theorem shows that the bias is $2^{-1}h^2 f''(x)\int u^2K(u)\mathrm{d}u(1+o(1))$ for second order kernels, and $h^2o(1)$ for higher order kernels. Further, this estimator is asymptotically Normal with mean zero and variance $f(x)\int K(u)^2\mathrm{d}u$.

Higher order kernels present an alternative to reducing the bandwidth with aim to get asymptotically unbiased density estimates. However Proposition 5.1 in the Appendix shows that the variance of the density estimator increases with the order of the kernels. Further, the estimator can take on negative values. These shortcomings are overcome in the next section.

1.2. **Two-step kernel estimator**

This section introduces a multiplicative adjustment to a pilot kernel estimator that reduces the bias without, asymptotically, changing the variance. Our two-stage density estimator is computed as follows: First compute a pilot kernel estimator for the density

$$\widetilde{f}(x) = \frac{1}{n}\sum_{j=1}^{n} K_{h_0}(X_j - x),$$

where $K_h(u) = (1/h)K(u/h)$. Second, estimate the ratio $\alpha(x) = f(x)/\widetilde{f}(x)$ by

$$\widehat{\alpha}(x) = \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\frac{1}{\widetilde{f}(X_j)}. \tag{3}$$

Multiplying the pilot estimator $\widetilde{f}(x)$ by $\widehat{\alpha}(x)$ produces

$$\widehat{f}(x) \quad = \quad \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\frac{\widetilde{f}(x)}{\widetilde{f}(X_j)}. \tag{4}$$

Remark that in the limit as h_0 goes to infinity, we have $\lim_{h_0\to\infty}\widetilde{f}(x)/\widetilde{f}(X_j) = 1$, so that estimator (4) becomes the usual Nadaraya-Watson kernel density estimator.

The idea of multiplicative bias reduction is not new. Hjort and Glad [8] showed that if $\widetilde{f} = f_{\hat{\theta}}$ is an estimated density from the parametric family $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$, then $\hat{f}$ has no bias if the true density belongs to $\mathcal{F}_\Theta$, and has smaller bias than the kernel density estimator for all densities in a *neighborhood* of $\mathcal{F}_\Theta$. Because the pilot estimator converges to the true density, we expect the bias to converge to zero. If the pilot estimator oversmooths, *i.e.* $n^{1/5-\varepsilon}h_0 \to \infty$, we expect the contribution to the variance of the estimator (4) from the pilot estimator to be of smaller magnitude than the contribution to the variance from multiplicative adjustment $\widehat{\alpha}(x)$. Assuming the density has four continuous derivatives, Jones *et al.* [11] proposed the estimator (4) with $h_0 = h_1 = c\cdot n^{-1/9}$ and showed that the bias is of order $O(h^4)$, instead of $O(h^2)$.

Our estimator also reduces the bias, but without requiring additional smoothness assumptions on the density nor increasing the asymptotic variance. Our results and proofs are completely different from the above referenced work, and rest on a representation theorem that we prove under the following assumptions:

K1. The kernel $K(\cdot)$ is a bounded symmetric probability density function.

K2. The kernel $K(\cdot)$ vanishes outside the interval $[-1,1]$ and, for any $0<\varepsilon<1$,

$$\inf\{K(u) : u \in [-1+\varepsilon, 1-\varepsilon]\} > 0.$$

K3. The squared L_2-norm $\|K\|_2^2 = \int K^2(u)\mathrm{d}u$ and second moment $\sigma_K^2 = \int u^2K(u)\mathrm{d}u$ are finite.

Theorem 1.1. *Assume that the kernel $K(\cdot)$ satisfies conditions (K1)–(K3). Denote by $\bar{f}(x) = \mathbb{E}[\widetilde{f}(x)]$ the expected value of the pilot estimator. If the bandwidths h_0, h_1 satisfy*

$$h_\ell \to 0, \quad nh_\ell / \log n \to \infty \quad \textit{for} \quad \ell = 0, 1,$$

then

$$\widehat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K_{h_1}(X_j - x) \frac{\bar{f}(x)}{\bar{f}(X_j)} + O_p\left(\sqrt{\frac{\log n}{nh_0}}\right).$$

Remark 1. Under assumption (K1), the expected Nadaraya-Watson density estimator $\bar{f}$ is a genuine density.

Remark 2. The ratio $\bar{f}(x)/\bar{f}(X_j)$ is equal or bigger than zero by assumption K1 and bounded away from infinity.

Section 3 is devoted to proving this theorem. Armed with this representation theorem, we can apply the results of Hjort and Glad [8], replacing $f_o(x)$ (respectively $f_o(X_j)$) by $\bar{f}(x)$ (resp. $\bar{f}(X_j)$). For completeness, we state the results of Hjort and Glad [8] in the next lemma.

Lemma 1.2 (Hjort and Glad)**.** *Assume the kernel $K(\cdot)$ satisfies conditions (K1) and (K2). Given any fixed density $f_o(x)$ for which the ratio $r(x) = f(x)/f_o(x)$ is twice continuously differentiable, the estimator*

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K_h\left(X_j - x\right) \frac{f_o(x)}{f_o(X_j)}$$

has bias

$$\mathbb{E}\left[\widehat{f}(x)\right] - f(x) = \frac{\sigma_K^2}{2} f_o(x) r''(x) h^2 (1 + o(1))$$

and variance

$$\operatorname{Var}\left(\widehat{f}(x)\right) = \left(\frac{f(x)\|K\|_2^2}{nh} - \frac{f(x)^2}{n}\right)(1 + o(h)).$$

Remark 3. If $h = c \cdot n^{-1/5}$, then the Lindeberg conditions are readily verified and $\widehat{f}(x)$ satisfies the central limit theorem

$$\sqrt{nh}\left(\widehat{f}(x) - f(x)\right) \Longrightarrow \mathcal{N}(b_o(x), \sigma^2(x))$$

with

$$b_o(x) = c^{5/2} \frac{\sigma_K^2}{2} f_o(x) r''(x) \quad \text{and} \quad \sigma^2(x) = f(x)\|K\|_2^2.$$

Remark 4. The variance of $\hat{f}(x)$ is the same as the variance of the kernel density estimator while the bias of $\widehat{f}(x)$ is smaller whenever $r''(x) < f''(x)$.

If the bandwidth h_0 of the pilot estimator converges to zero such that $\log n/nh_0 = o(\min(h_1^4, 1/nh_1))$, then $n^{-1}\sum_{j=1}^{n} K_{h_1}(X_j - x)\left(\bar{f}(x)/\bar{f}(X_j)\right)$ is the dominant term of $\widehat{f}(x)$. When f is twice continuously differentiable, $\bar{f}(x), \bar{f}'(x)$ and $\bar{f}''(x)$ converge to $f(x), f'(x)$ and $f''(x)$, respectively. Therefore the bias of $\widehat{f}(x)$, which is proportional to $\left(f(x)/\bar{f}(x)\right)''$, converges to zero. Without further assumptions on the smoothness of the true underlying density, we can not quantify how fast the asymptotic bias converges to zero. If we are willing to

assume that the density f has three continuous derivatives, then the asymptotic bias is of order $O(h_1)$, and if f has four continuous derivatives, then the asymptotic bias will be of order $O(h_1^2)$. Smoothness beyond four derivatives does not change the order of the asymptotic bias.

We are now ready to state the main theorem of this paper.

Theorem 1.3. *Assume the smoothing kernel $K(\cdot)$ satisfies conditions (K1)-(K3). Let $h_1 = c \cdot n^{-1/5}$ and $h_0 = c \cdot n^{-\alpha}$ for $0 < \alpha < 1/5$. If the density f is twice continuously differentiable, then the two step density estimator (4) satisfies the central limit theorem*

$$n^{2/5}\left(\widehat{f}(x) - f(x)\right) \Longrightarrow \mathcal{N}(0, \sigma^2(x)),$$

with

$$\sigma^2(x) = f(x)\|K\|_2^2.$$

Remark 5. Kernel density estimators with bandwidths of order $O(n^{-\alpha})$ for $0 < \alpha < 1/5$ are oversmoothing the true density, and as a result, they have biases that are of larger order of magnitude than their standard deviations. Thus, we conclude that the multiplicative adjustment performs a bias reduction on the pilot estimator.

Remark 6. The asymptotic variances of the two step estimator and the usual kernel density estimator are the same. However, for finite samples, the two step kernel smoother can have a slightly larger variance depending on choice of the bandwidth h_0.

Corollary 1.4. *Assume that the smoothing kernel $K(\cdot)$ satisfies assumptions (K1)–(K3). Let $z_{1-\alpha/2}$ be the $1-\alpha/2$ quantile of a standard Normal and set $\nu^2 = \int K^2(u)\mathrm{d}u$. If f is twice continuously differentiable, then for every fixed x*

$$\lim_{n\to\infty} \mathbb{P}_f\left[\widehat{f}(x) - \frac{z_{1-\alpha/2}\cdot\nu}{n^{2/5}}\sqrt{\widehat{f}(x)} \le f \le \widehat{f}(x) + \frac{z_{1-\alpha/2}\cdot\nu}{n^{2/5}}\sqrt{\widehat{f}(x)}\right] = 1-\alpha.$$

2. Extensions

2.1. Higher order kernel smoothers

Higher order kernels take advantage of higher order derivatives of the density to reduce the bias by orders of magnitude. This also holds for the estimator considered in Lemma 1.2, that is, if the ratio $r(x) = f(x)/f_{\mathrm{o}}(x)$ is s-times continuously differentiable and $K^{[s]}(\cdot)$ is a kernel of order s, then the estimator

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}K^{[s]}\left(\frac{X_j - x}{h}\right)\frac{f_{\mathrm{o}}(x)}{f_{\mathrm{o}}(X_j)}$$

has bias

$$\mathbb{E}\left[\widehat{f}(x)\right] - f(x) = \frac{f_{\mathrm{o}}(x)}{s!}\left\{\int u^s K^{[s]}(u)\mathrm{d}u\right\} r^{(s)}(x)h^s(1+o(1))$$

and variance

$$\mathrm{Var}\left(\widehat{f}(x)\right) = \frac{f(x)\int K^{[s]}(u)^2\mathrm{d}u}{nh} - \frac{f(x)}{n}.$$

Thus, we can extend Theorem 1.3 to this setting to show that the estimator $\widehat{f}(x)$

$$\widehat{f}(x) = \left\{ \frac{1}{n} \sum_{j=1}^{n} K_{h_1}^{[s]}(X_j - x) \frac{1}{\widetilde{f}(X_j)} \right\} \times \widetilde{f}(x)$$

is asymptotic Normal with mean zero and variance $f(x) \int K^{[s]}(u)^2 \mathrm{d}u$, provided that $h_0 = n^{-\alpha}$ for some $0 < \alpha < 1/(1+2s)$. While the true density needs to be s times differentiable, we do not need to use a higher order kernel for the pilot estimator.

2.2. **The multivariate case**

Multivariate kernel density estimators in R^p are of the form

$$\widehat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K_{\mathbf{H}}(X_j - x)$$

where $K_H(u) = (1/\det \mathbf{H})\, K(\mathbf{H}^{-1}x)$ and $\mathbf{H}$ is a $p \times p$ invertible bandwidth matrix. For fixed multivariate densities $f_{\mathrm{o}}(x)$, the estimator

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} K_{\mathbf{H}}(X_j - x) \frac{f_{\mathrm{o}}(x)}{f_{\mathrm{o}}(X_j)},$$

with $\mathbf{H} = \mathrm{diag}(h)$, has bias

$$\mathbb{E}\left[\hat{f}(x)\right] - f(x) = \frac{h^2 f_{\mathrm{o}}(x)}{2} \sum_{k=1}^{p} \frac{\partial^2}{\partial x_k^2} \left(\frac{f(x)}{f_{\mathrm{o}}(x)} \right) \int u_k^2 K(u) \mathrm{d}u \, (1 + o(1))$$

and variance

$$\mathrm{Var}\left(\hat{f}(x)\right) = \frac{f(x) \int K(u)^2 \mathrm{d}u}{nh} (1 + o(1)).$$

Therefore if the bandwidth h_0 of the pilot estimator is of order $O(n^{-\alpha})$; with $0 < \alpha < 1/(4+p)$, estimators of the form (4) will have a Normal distribution with mean zero and variance $f(x) \int K(u)^2 \mathrm{d}u$.

3. Discussion

From the theory, we know that as soon as h_0 is bigger than h_1, the estimator (4) has smaller bias than the classical kernel estimate. We conduct a simulation study in order to compare the finite sample performance of our estimator with the classical one and to see how its performance depends on our choice for the bandwidth h_0 of the pilot estimator. In our simulations, we restricted the bandwidth of the multiplicative adjustment h_1 to lie in a grid $\mathcal{H}$ and set the bandwidth of the pilot estimator $h_0 = ch_1$. Our theory discusses the case where $c > 1$, but in order to explore the sensitivity of our two stage estimator on h_0, we also considered some values for $c < 1$. The bandwidth selection problem was avoided by doing a "best possible performance" assuming that the true density f is known and for each h_0 we obtain $\hat{h}_1$ such as

$$\hat{h}_1 \quad = \quad \min_{h_1 \in \mathcal{H}} \max_{x \in \mathcal{G}} |\hat{f}(x) - f(x)|$$

where the maximum on x is taken on a grid $\mathcal{G}$ of 100 points equispaced in the range of X_i. Of course the true bandwidth is unknown in practice but the simulations were done to verify that our estimator does well for finite

TABLE 1. Median over 100 replications of $\|\hat{f}(x) - f(x)\|_\infty$ divide by $\|\hat{f}_{NW}(x) - f(x)\|_\infty$ where $\hat{f}_{NW}(x)$ is Nadaraya-Watson estimator.

Density	n	$h_0 = 0.5\hat{h}_1$	$h_0 = 0.75\hat{h}_1$	$h_0 = \hat{h}_1$	$h_0 = 1.1\hat{h}_1$	$h_0 = 1.5\hat{h}_1$	$h_0 = 2\hat{h}_1$	$h_0 = 10\hat{h}_1$
Laplace	50	0.983	0.950	0.932	0.937	0.939	0.948	0.993
	100	0.956	0.957	0.956	0.951	0.961	0.960	0.993
	200	0.957	0.941	0.940	0.941	0.939	0.933	0.990
	500	0.941	0.936	0.933	0.929	0.941	0.947	0.988
Gamma3	50	1.400	1.111	1.048	1.036	1.007	0.985	0.982
	100	1.543	1.375	1.304	1.270	1.226	1.186	0.993
	200	1.348	1.187	1.122	1.095	1.031	0.981	0.976
	500	1.181	0.968	0.875	0.849	0.808	0.817	0.987
Gaussian	50	1.058	0.929	0.869	0.863	0.879	0.895	0.989
	100	1.048	0.919	0.871	0.863	0.869	0.880	0.986
	200	0.976	0.858	0.827	0.821	0.828	0.859	0.982
	500	0.906	0.829	0.800	0.802	0.811	0.833	0.974
Bimodal Gaussian	50	1.155	0.946	0.928	0.928	0.947	0.983	1
	100	0.984	0.867	0.851	0.854	0.877	0.885	1
	200	1.062	0.981	0.968	0.965	0.952	0.952	1.007
	500	1.016	0.878	0.853	0.855	0.878	0.917	0.981

sample sizes. We tried $n = 50$, $n = 100$, $n = 200$ and $n = 500$ for different densities such as Laplace, Gamma, Gaussian, Gaussian bimodal. The number of replication is 100. In each case, we calculate for each replication the sup norm over the grid $\mathcal{G}$

$$\|\hat{f}(x) - f(x)\|_\infty,$$

and then calculate the median over 100 replications.

Table 1 presents the median over 100 replications of the ratio:

$$\frac{\|\hat{f}(x) - f(x)\|_\infty}{\|\hat{f}_{NW}(x) - f(x)\|_\infty},$$

where $\hat{f}_{NW}$ is the classical Nadaraya-Watson estimator. As predicted by Theorem 3, the sup norm of our estimator was smaller than the sup norm for the classical Nadaraya estimator when $c > 1$ except for the Gamma 3 density. The differences vary from 5% to 20% depending the sample size (the difference increases with the sample size) and the type of densities. Interestingly, the performance of our estimator is not very sensitive to the particular choice of the bandwidth of the pilot estimator h_0.

We proposed a new density estimator which is positive everywhere, which reduces the bias and which is simple to implement. We do not conduct any theory in order to choose the two bandwidths in a optimal way.

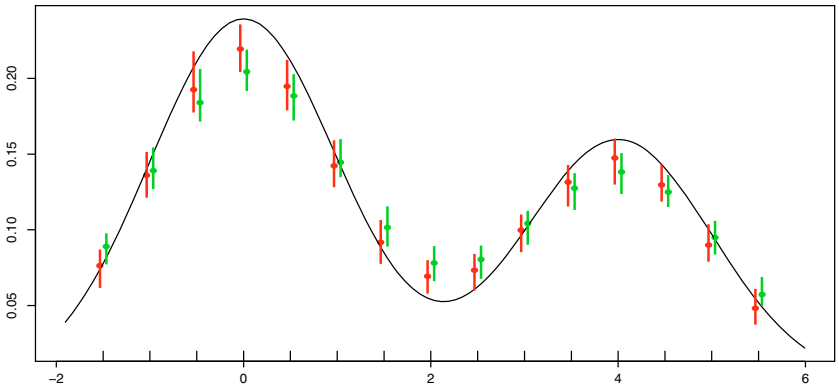

FIGURE 1. The solid line represents the true density for the bimodale Gaussian.

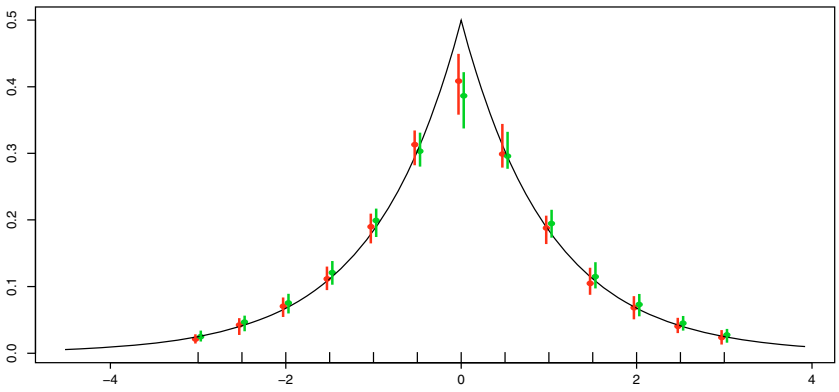

FIGURE 2. The solid line represents the true density for the Laplace.

The rule of thumb we proposed and which we try on others simulations which are not reported here, is the following:

- use any data-driven choice method existing in the literature for the bandwidth as for example the one proposed in the GNU-R library of **KernSmooth** of Wand [19] which gives $\hat{h}_1$;
- apply our estimator with $(\hat{h}_0 = 1.5\hat{h}_1, \hat{h}_1)$.

We applied that simple rule to the same data sets and densities as above and improve often Nadaraya-Watson estimator. The following pictures present, for the Laplace and Bimodal Gaussian density with $n = 100$ and 100 iterations, different estimators on a grid of points. In lines is the true density which is unknown. For every point on a fixed grid, we plot, side by side, the median over 100 replications of our estimator at that point (left side) and on the right side of that point the median over 100 replications of Nadaraya Watson estimator. We add also the interquartile interval in order to see the fluctuations of the different estimators.

On both example, our estimator reduces the bias by increasing the peak and decreasing the valley (see the circles) and the interquertile intervals look similar for both estimator as predicted by the theory.

4. Asymptotics: Proof of Theorem 1.1

Denote by $\bar{f}(x) = \mathbb{E}[\tilde{f}(x)]$ the expected value of the pilot estimator and decompose (4) as follows:

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\frac{\bar{f}(x)}{\tilde{f}(X_j)}\left(1 + \frac{\tilde{f}(x) - \bar{f}(x)}{\bar{f}(x)}\right)\left(1 - \frac{\tilde{f}(X_j) - \bar{f}(X_j)}{\tilde{f}(X_j)}\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\frac{\bar{f}(x)}{\tilde{f}(X_j)} + S_2(x) - S_3(x) - S_4(x),\end{aligned}$$

where

$$S_2(x) = \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\left(\frac{\tilde{f}(x) - \bar{f}(x)}{\tilde{f}(X_j)}\right), \tag{5}$$

$$S_3(x) = \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\frac{\bar{f}(x)}{\tilde{f}(X_j)}\left(\frac{\tilde{f}(X_j) - \bar{f}(X_j)}{\tilde{f}(X_j)}\right), \tag{6}$$

and

$$S_4(x) = \frac{1}{n}\sum_{j=1}^{n} K_{h_1}(X_j - x)\frac{\bar{f}(x)}{\tilde{f}(X_j)}\left(\frac{\tilde{f}(x) - \bar{f}(x)}{\tilde{f}(x)}\right)\left(\frac{\tilde{f}(X_j) - \bar{f}(X_j)}{\tilde{f}(X_j)}\right). \tag{7}$$

While these sums are local averages of $\left(\tilde{f}(X_j) - \bar{f}(X_j)\right)/\tilde{f}(X_j)$ and $\left\|\tilde{f}(X_j) - \bar{f}(X_j)\right\|_\infty$ is of order $O_p\left((\log n/nh_0)^{-1/2}\right)$, care must be taken to ensure that dividing by $\tilde{f}(X_j)$ will not unduly affect the magnitude $S_k(x)$. For densities uniformly bounded away from zero, Theorem 1.1 follows from the uniform convergence of kernel estimators. Proposition 4.1 extends the result to arbitrary densities.

Proposition 4.1. *If the kernel $K(\cdot)$ satisfies conditions (K2) then the two first moments of $S_2(x)$, $S_3(x)$ and $S_4(x)$ are of order*

$$\mathbb{E}\left[S_2(x) + S_3(x) + S_4(x)\right] = O\left(\sqrt{\frac{\log(1/h_0)}{nh_0}}\right)$$

and

$$\mathbb{E}\left[S_2(x)^2 + S_3(x)^2 + S_4(x)^2\right] = O\left(\frac{\log(1/h_0)}{nh_0}\right).$$

Remark 7. Proposition 4.1 implies that

$$S_2(x) + S_3(x) + S_4(x) = O_P\left(\sqrt{\frac{\log(1/h_0)}{nh_0}}\right).$$

Proof. By Theorem 1 of Stute [17], there exists a positive constant C_0 depending on K and f, such that the event

$$\left\{\sup_x \left|\tilde{f}(x) - \bar{f}(x)\right| \le C_0\sqrt{\frac{\log\frac{1}{h_0}}{nh_0}} \quad \text{for all but finitely many } n\right\}$$

has probability one. Thus, for large n,

$$S_2(x) \le C_0 \sqrt{\frac{\log \frac{1}{h_0}}{nh_0}} \frac{1}{n} \sum_{j=1}^{n} K_{h_1}(X_j - x) \frac{1}{\bar{f}(X_j)}, \tag{8}$$

$$S_3(x) \le C_0 \sqrt{\frac{\log \frac{1}{h_0}}{nh_0}} \frac{1}{n} \sum_{j=1}^{n} K_{h_1}(X_j - x) \frac{\bar{f}(x)}{\bar{f}(X_j)\widetilde{f}(X_j)} \tag{9}$$

$$S_4(x) \le \left(C_0 \sqrt{\frac{\log \frac{1}{h_0}}{nh_0}} \right)^2 \frac{1}{n} \sum_{j=1}^{n} K_{h_1}(X_j - x) \frac{1}{\bar{f}(X_j)\widetilde{f}(X_j)}. \tag{10}$$

Direct calculations yield

$$\mathbb{E}\left[S_2(x)\right] = O\left(\sqrt{\frac{\log \frac{1}{h_0}}{nh_0}} \right) \quad \text{and} \quad \mathbb{E}\left[S_2(x)^2\right] = O\left(\frac{\log \frac{1}{h_0}}{nh_0} \times \frac{1}{nh_1} \right).$$

To bound $S_3(x)$ and $S_4(x)$, define $N(X_1) = \#\{k \neq 1 : |X_k - X_1| \le h_0/2\}$ and $0 < c = \inf\{K(u) : |u| \le 1/2\}$, and note that

$$\tilde{f}(X_j) \ge \frac{cN(X_j) + 1}{nh_0} \ge \frac{c}{nh_0} \left(N(X_j) + 1 \right).$$

By Lemma 5.2,

$$\begin{aligned} \mathbb{E}\left[S_3(x)\right] &\le C_1 \sqrt{\frac{\log \frac{1}{h_0}}{nh_0}} nh_0 \mathbb{E}\left[\frac{K_{h_1}(X_1 - x)\bar{f}(x)}{\bar{f}(X_1)} \times \mathbb{E}\left[\frac{1}{N(X_1)+1} \middle| X_1 \right] \right] \\ &= C_1 \sqrt{nh_0 \log \frac{1}{h_0}} \mathbb{E}\left[\frac{1}{h_1} K\left(\frac{X_1 - x}{h_1} \right) \frac{\bar{f}(x)}{\bar{f}(X_1)} \frac{1}{n\bar{p}(X_1)} \right], \end{aligned}$$

where

$$\bar{p}(X_1) = \int_{X_1 - h_0/2}^{X_1 + h_0/2} f(z) \mathrm{d}z = h_0 f(X_1) \left[1 + O(h_0) \right].$$

Hence

$$\mathbb{E}\left[S_3(x)\right] \le C_3 \sqrt{\frac{\log \frac{1}{h_0}}{nh_0}},$$

and similarly,

$$\mathbb{E}\left[S_4(x)\right] \le C_4 \frac{\log \frac{1}{h_0}}{nh_0}.$$

The second moment of $S_3(x)$ is bounded by

$$\mathbb{E}\left[S_3(x)^2\right] \le C_0^2 \frac{\log \frac{1}{h_0}}{nh_0} \mathbb{E}\left[\frac{K_{h_1}(X_1 - x) K_{h_1}(X_2 - x) \bar{f}(x)^2}{\bar{f}(X_1)\widetilde{f}(X_1)\bar{f}(X_2)\widetilde{f}(X_2)} \right] \tag{11}$$

$$+ C_0^2 \frac{\log \frac{1}{h_0}}{nh_0} \frac{1}{n} \mathbb{E}\left[\left(\frac{K_{h_1}(X_1 - x)\bar{f}(x)}{\bar{f}(X_1)\widetilde{f}(X_1)} \right)^2 \right]. \tag{12}$$

The expectation in the right-hand side of (11) is bounded by

$$\frac{(nh_0)^2}{c^2}\,\mathbb{E}\left[K_{h_1}(X_1-x)K_{h_1}(X_2-x)\frac{\bar{f}(x)^2}{\bar{f}(X_1)\bar{f}(X_2)}\right. \tag{13}$$

$$\left.\times\,\mathbb{E}\left[\frac{1}{(N(X_1)+1)(N(X_2)+1)}\middle| X_1,X_2\right]\right]. \tag{14}$$

Divide the (x_1,x_2) plane into the three regions: $R_1=\{(x_1,x_2):|x_1-x_2|\geq 2\}$, $R_2=\{(x_1,x_2): 1\leq|x_1-x_2|<2\}$ and $R_3=\{(x_1,x_2):|x_1-x_2|<1\}$, and set $N_0(x_1,x_2)=\#\{k>2:X_k\in R_3\}$. When $(X_1,X_2)\in R_1$, apply Lemma 5.2 to conclude that

$$\mathbb{E}\left[\frac{1}{(N(X_1)+1)(N(X_2)+1)}\middle| X_1,X_2\right]\leq\frac{C_5}{(nh_0)^2}$$

while for $(X_1,X_2)\in R_3$, bound

$$\frac{1}{(N(X_1)+1)(N(X_2)+1)}\leq\frac{2}{(N_0(X_1,X_1)+1)(N(X_1,X_2)+1)},$$

and apply Lemma 5.2 to get

$$\mathbb{E}\left[\frac{1}{(N(X_1)+1)(N(X_2)+1)}\middle| X_1,X_2\right]\leq\frac{C_6}{(nh_0)^2}.$$

Finally, for $(X_1,X_2)\in R_2$, bound

$$\frac{1}{(N(X_1)+1)(N(X_2)+1)}\leq\frac{1}{(N(X_1)-N_0(X_1,X_1)+1)(N(X_2)-N_0(X_1,X_1)+1)}$$

and apply Lemma 5.2 to conclude that

$$\mathbb{E}\left[\frac{1}{(N(X_1)+1)(N(X_2)+1)}\middle| X_1,X_2\right]\leq\frac{C_7}{(nh_0)^2}.$$

These bounds and a change of variables in (13) shows that it is of order $O(1)$, from which it follows that the right-hand side of (11) is of order $O(\log(1/h_0)/(nh_0))$.

Finally, bound (12) by

$$C_0\frac{\log\frac{1}{h_0}}{nh_0}\frac{1}{n}\mathbb{E}\left[\left(\frac{K_{h_1}(X_1-x)\bar{f}(x)}{\bar{f}(X_1)\widetilde{f}(X_1)}\right)^2\right]\leq C_8h_0\log\frac{1}{h_0}\mathbb{E}\left[\left(\frac{K_{h_1}(X_1-x)\bar{f}(x)}{\bar{f}(X_1)}\right)^2\right.$$

$$\left.\times\,\mathbb{E}\left[\frac{1}{(N(X_1)+1)(N(X_1)+2)}\middle| X_1\right]\right],$$

apply Lemma 5.2 and operate a change of variables to conclude that the latter is bounded by

$$C_9\frac{\log\frac{1}{h_0}}{nh\cdot nh_0}.$$

Hence

$$\mathbb{E}\left[S_3(x)^2\right]=O\left(\frac{\log\frac{1}{h_0}}{nh_0}\right).$$

The expectation of $S_4(x)^2$ is similarly bounded, and the conclusion of Proposition 4.1 follows.

5. Appendix

Proposition 5.1. *Assume that the smoothing kernel $K(\cdot)$ satisfyes assumptions (K1)-(K2) Denote by $K^{[r]}(\cdot)$ the hierarchy of kernels generated by $K(\cdot)$. If $s < r$, then*

$$\int K^{[s]}(u)^2 \mathrm{d}u < \int K^{[r]}(u)^2 \mathrm{d}u.$$

Remark 8. It immediately follows that within the same hierarchy, the variance of kernel density estimators is monotone increasing in the order.

Proof. Let $p_\ell(u)$ be the orthonormal polynomials of order ℓ in $L_2(K)$ and define as in Berlinet [4] the hierarchy of higher order kernels by

$$K^{[r]}(u) = \sum_{\ell=0}^{r} p_\ell(u) p_\ell(0) K(u).$$

From the ortonormality of $p_\ell(u)$ follows the conclusion

$$\int K^{[r]}(u)^2 \mathrm{d}u = \sum_{\ell=0}^{r} p_\ell(0)^2 > \sum_{\ell=0}^{s} p_\ell(0)^2 = \int K^{[s]}(u)^2 \mathrm{d}u.$$

Lemma 5.2. *Let (N_1, N_2, N_3) have a Multinomial$(n; p_1, p_2, p_3)$ distribution. Then*

$$\mathbb{E}\left[\frac{1}{N_1+1}\right] = \frac{1}{(n+1)\,p_1}\left[1-(1-p_1)^{n+1}\right] \tag{15}$$

and

$$\mathbb{E}\left[\frac{1}{(N_1+1)(N_2+1)}\right] = \frac{1-(1-p_1)^{n+2}-(1-p_2)^{n+2}+[(1-p_1)(1-p_2)]^{n+2}}{p_1 p_2 (n+1)(n+2)}. \tag{16}$$

Remark 9. The proof of (15) is also given in Barron *et al.* [3].

Remark 10. It follows that

$$\mathbb{E}\left[\frac{1}{N_1+1}\right] \le \frac{1}{np_1}$$

and

$$\mathbb{E}\left[\frac{1}{(N_1+1)(N_2+1)}\right] \le \frac{1}{np_1 \cdot np_2}.$$

It follows that

$$\mathbb{E}\left[\frac{1}{(N_1+1)(N_1+2)}\right] \le \frac{1}{n^2 p_1^2}. \qquad \square$$

Proof. With the identity

$$\binom{n}{k}\frac{1}{k+1} = \binom{n+1}{k+1}\frac{1}{n+1}$$

one shows that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N+1}\right] &= \sum_{k=0}^{n}\frac{1}{k+1}\binom{n}{k}p_1^k(1-p_1)^{n-k}\\ &= \sum_{k=0}^{n}\frac{1}{n+1}\binom{n+1}{k+1}p_1^k(1-p_1)^{n-k}\\ &= \frac{1}{(n+1)\,p_1}\sum_{k=1}^{n+1}\binom{n+1}{k}p_1^k(1-p_1)^{n+1-k}\\ &= \frac{1}{(n+1)\,p_1}\left[1-(1-p_1)^{n+1}\right].\end{aligned}$$

Combining (16) and (15), the conditional expectation of N_2 given N_1 is

$$\mathbb{E}\left[\frac{1}{(N_1+1)(N_2+1)}\right]=\mathbb{E}\left[\frac{1}{N_1+1}\mathbb{E}\left[\frac{1}{N_2+1}\middle|N_1\right]\right]=\frac{1-p_1}{p_2}\mathbb{E}\left[\frac{1-\left(1-\frac{p_2}{1-p_1}\right)^{n-N_1+1}}{(N_1+1)(n-N_1+1)}\right].$$

With the identity

$$\frac{1}{(k+1)(n+1-k)}\binom{n}{k}=\frac{1}{(n+1)(n+2)}\binom{n+2}{k+1},$$

a similar derivation to (15) gives

$$\mathbb{E}\left[\frac{1}{(N_1+1)(N_2+1)}\right]=\frac{1-(1-p_1)^{n+2}-(1-p_2)^{n+2}+[(1-p_1)(1-p_2)]^{n+2}}{p_1p_2(n+1)(n+2)}.$$ □

Acknowledgements. We would wish to thank two anonymous referees whose insightful comments have improved the presentation and cohesion of our paper. This paper was written while the second author was visiting Los Alamos National Laboratory.

References

[1] I.S. Abramson, On bandwidth variation in kernel estimates – a square root law. *Ann. Statist.* **10** (1982) 1217–1223.
[2] I.S. Abramson, Adaptive density flattening-metric distortion principle for combining bias in nearest neighbor methods. *Ann. Statist.* **12** (1984) 880–886.
[3] A.R. Barron, L. Györfi and E.C. van der Meulen, Distribution Estimation Consistent in Total Variation and in Two Types of Information Divergence. *IEEE Trans. Inf. Theory* **38** (1992) 1437–1453.
[4] A. Berlinet, Hierarchies of higher order kernels. *Prob. Theory Related Fields* **94** (1993) 489–504.
[5] B.L. Granovsky and H.-G. Müller, Optimizing kernel methods: a unifying variational principle. *Ins. Statist. Rev.* **59** (1991) 373–388.
[6] P. Hall, On the bias of variable bandwidth curve estimators. *Biometrika* **77** (1990) 529–535.
[7] P. Hall and J.S. Marron, Variable window width kernel estimates of a probability density. *Prob. Theory Related Fields* **80** (1988) 37–49.
[8] N.L. Hjort and I.K. Glad, Nonparametric density estimation with a parametric start. *Ann. Statist.* **23** (1995) 882–904.
[9] N.L. Hjort and M.C. Jones, Locally parametric nonparametric density estimation. *Ann. Statist.* **24** (1996) 1619–1647.
[10] M.C. Jones, Variable kernel density estimates variable kernel density estimates. *Aust. J. Statist.* **32** (1990) 361–371. Correction **33** (1991) 119.
[11] M.C. Jones, O.B. Linton and J.P. Nielsen, A simple bias reduction method for density estimation. *Biometrika* **82** (1995) 327–38.
[12] M.C. Jones, I.J. McKay and T.-C. Hu, Variable location and scale kernel density estimation. *Inst. Statist. Math.* **46** (1994) 521–535.
[13] I. McKay, A note on bias reduction in variable kernel density estimates. *Can. J. Statist.* **21** (1993) 367–375.

[14] J.S. Marron and M.P. Wand, Exact mean integrated squared error. *Ann. Statist.* **20** (1992) 712–736.
[15] J.P. Nielson and O. Linton, A multiplicative bias reduction method for nonparametric regression. *Statist. Probab. Lett.* **19** (1994) 181–187.
[16] M. Rosenblatt, Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** (1956) 832–837.
[17] W. Stute, A law of iterated logarithm for kernel density estimators. *Ann. Probab.* **10** (1982) 414–422.
[18] G. Terrel and D. Scott, On improving convergence rates for non-negative kernel density estimators. *Ann. Statist.* **8** (1980) 1160–1163.
[19] M.P. Wand and M.C. Jones, *Kernel Smoothing*. Chapman and Hall, London (1995).