Journal of Nonparametric Statistics
Vol. 22, No. 2, February 2010, 219–235

Taylor & Francis
Taylor & Francis Group

Bias reduction in kernel density estimation via Lipschitz condition

Kairat Mynbaev[a] and Carlos Martins-Filho[b,c]*

[a] *International School of Economics, Kazakh British Technical University, Tolebi 59, Almaty 050010, Kazakhstan;* [b] *Department of Economics, University of Colorado, Boulder, CO 80309-0256, USA;* [c] *IFPRI, 2033 K Street NW, Washington, DC 20006-1002, USA*

(*Received 13 September 2008; final version received 13 August 2009*)

In this paper we propose a new nonparametric kernel-based estimator for a density function f which achieves bias reduction relative to the classical Rosenblatt–Parzen estimator. Contrary to some existing estimators that provide for bias reduction, our estimator has a full asymptotic characterisation including uniform consistency and asymptotic normality. In addition, we show that bias reduction can be achieved without the disadvantage of potential negativity of the estimated density – a deficiency that results from using higher order kernels. Our results are based on imposing global Lipschitz conditions on f and defining a novel corresponding kernel. A Monte Carlo study is provided to illustrate the estimator's finite sample performance.

Keywords: bias reduction; kernel density estimation; Lipschitz conditions

AMS 2000 Classification codes: 62G07, 62G20

1. Introduction

Let f denote the density associated with a real random variable X and let $\{X_j\}_{j=1}^n$ be a random sample of size n of X. We call a kernel any function K on $\Re$ such that

$$\int_{-\infty}^{+\infty} K(t)\,\mathrm{d}t = 1. \tag{1}$$

The Rosenblatt–Parzen estimator for the density f evaluated at $x \in \Re$ is given by $f_{\mathrm{R}}(x) = (1/n)\sum_{j=1}^n (1/h_n)K((X_j - x)/h_n)$, where $0 < h_n$ is a bandwidth sequence such that $h_n \to 0$ as $n \to \infty$. Let $B(f_{\mathrm{R}}(x)) = E(f_{\mathrm{R}}(x)) - f(x)$ denote the bias of $f_{\mathrm{R}}(x)$ at x. It is well known (Parzen 1962; Pagan and Ullah 1999; Fan and Yao 2003) that if f has its rth derivative bounded and continuous at x an interior point in the support of f and the kernel is of order r, that is, K satisfies $\int_{-\infty}^{+\infty} K(t)t^j\,\mathrm{d}t = 0$ for $j = 1, \ldots, r-1$ then $\mathrm{Bias}(f_{\mathrm{R}}(x)) = O(h_n^r)$. Bias reduction through higher order kernels (Granovsky and Müller 1991; Jones and Foster 1993) can be inconvenient in that for $r > 2$, K can no longer be nonnegative everywhere and therefore $f_{\mathrm{R}}(x)$ may

*Corresponding author. Emails: carlos.martins@colorado.edu, c.martins-filho@cgiar.org

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250903266058
http://www.informaworld.com

be negative. There exist other approaches to bias reduction in density estimation (Jones, Linton, and Nielsen 1995; DiMarzio and Taylor 2004) but the asymptotic properties of these estimators have not been fully developed.

In this paper, we propose a new nonparametric kernel-based density estimator for which reduction in the order of the bias, relative to the Rosenblatt–Parzen estimator, is attained by imposing global Lipschitz conditions on f. The use of our estimator and higher order Lipschitz conditions seems desirable for the following reasons: (a) in a sense to be made precise in Section 2, r-times differentiability of f is stronger than r-times Lipschitz smoothness; (b) we provide a full asymptotic characterisation of our estimator, including results on its uniform consistency, asymptotic normality and convergence rates. We emphasise that this is the main theoretical advantage of our estimator. Its rates of convergence are true for all bandwidths and sample sizes. By contrast, rates of convergence for higher order kernels and local polynomial estimators are valid only asymptotically; (c) our estimator is nonnegative, given a suitable choice of the seed kernel. In fact, the Cauchy kernel assures nonnegativity of the estimator (Section 2.2).

The rest of the paper is organised as follows. Section 2 provides a brief discussion of Lipschitz conditions, discusses the properties of the new kernels we propose and defines our estimator. In Section 3, the main asymptotic properties of our estimator are obtained. Section 4 contains a small Monte Carlo study that gives some evidence on the small sample performance of our estimator relative to the Rosenblatt–Parzen and local quadratic estimators. Sections 5 provides a conclusion and gives directions for future work.

2. Lipschitz conditions, associated kernels and a new nonparametric density estimator

2.1. *Lipschitz conditions*

The properties of nonparametric density estimators are traditionally obtained by assumptions on the smoothness of the underlying density. Smoothness can be regulated by finite differences, which can be defined as forward, backward or centred. The corresponding examples of finite first-order differences for a function $f(x)$ are $f(x+h) - f(x)$, $f(x) - f(x-h)$ and $f(x+h) - f(x-h)$, where $h \in \Re$. Here, we focus on centered even-order differences because the resulting kernels are symmetric. Let $C_{2k}^l = (2k)!/(2k-l)!l!, l = 0, \dots, 2k, k \in \{1, 2, \dots\}$ be the binomial coefficients, $c_{k,s} = (-1)^{s+k} C_{2k}^{s+k}, s = -k, \dots, k$ and

$$\Delta_h^{2k} f(x) = \sum_{s=-k}^{k} c_{k,s} f(x + sh), \quad h \in \Re. \tag{2}$$

We say that a function $f : \Re \to \Re$ satisfies the Lipschitz condition of order $2k$ if for any $x \in \Re$ there exist $H(x) > 0$ and $\varepsilon(x) > 0$ such that $|\Delta_h^{2k} f(x)| \le H(x) h^{2k}$ for all h such that $|h| \le \varepsilon(x)$. The following theorem shows that $H(x)$ and $\varepsilon(x)$ can be obtained for the Gaussian and Cauchy densities.

Theorem 1 (a) *Let* $f(x) = \mathrm{e}^{-(1/2)x^2}/(2\pi)^{1/2}$, *then for any small* $\varepsilon \in (0,1)$ *there exists a constant* $c_\varepsilon > 0$ *such that*

$$|\Delta_h^{2k} f(x)| \le c_\varepsilon \mathrm{e}^{-(1-\varepsilon)x^2/2} h^{2k} \quad \textit{for } |h| \le \varepsilon(1 + |x|). \tag{3}$$

(b) *Let* $f(x) = (\pi(1 + x^2))^{-1}$, *then there exist* $\varepsilon \in (0,1)$ *and a constant* $c > 0$ *such that*

$$|\Delta_h^{2k} f(x)| \le c h^{2k} f^{k+1}(x) \quad \textit{for} \quad |h| \le \varepsilon(1 + |x|). \tag{4}$$

Proof (a) We prove the statement for $f(t) = e^{-(1/2)t^2}$. For any twice differentiable function f one has $f(y) = f(x) + f^{(1)}(x)(y-x) + \int_x^y (y-t) f^{(2)}(t)\,\mathrm{d}t$, hence for $h > 0$

$$\begin{aligned} |\Delta_h^2 f(x)| &= |f(x-h) - 2f(x) + f(x+h)| \\ &= \left|\int_x^{x+h} (x+h-t) f^{(2)}(t)\,\mathrm{d}t + \int_x^{x-h} (x-h-t) f^{(2)}(t)\,\mathrm{d}t\right| \\ &= \left|\int_x^{x+h} (x+h-t) f^{(2)}(t)\,\mathrm{d}t + \int_{x-h}^{x} (t-x+h) f^{(2)}(t)\,\mathrm{d}t\right| \\ &\le \sup_{|x-t|\le h} |f^{(2)}(t)| \left(\int_x^{x+h} (x+h-t)\,\mathrm{d}t + \int_{x-h}^{x} (t-x+h)\,\mathrm{d}t\right) = h^2 \sup_{|x-t|\le h} |f^{(2)}(t)|. \end{aligned} \tag{5}$$

The case for $h < 0$ leads straightforwardly to the same bound. We now prove that

$$\Delta_h^2 \left(\Delta_h^{2(k-1)} f(x)\right) = \Delta_h^{2k} f(x). \tag{6}$$

Observe that the left-hand side of Equation (6) can be written as

$$\begin{aligned} \Delta_h^2 \left(\Delta_h^{2(k-1)} f(x)\right) &= \sum_{s=-k+1}^{k-1} (-1)^{s+k-1} C_{2(k-1)}^{s+k-1} f(x+sh-h) \\ &\quad - 2 \sum_{s=-k+1}^{k-1} (-1)^{s+k-1} C_{2(k-1)}^{s+k-1} f(x+sh) \\ &\quad + \sum_{s=-k+1}^{k-1} (-1)^{s+k-1} C_{2(k-1)}^{s+k-1} f(x+sh+h) \\ &= \sum_{s=-k}^{k-1} (-1)^{s+k} C_{2(k-1)}^{s+k} f(x+sh) + 2 \sum_{s=-k+1}^{k-1} (-1)^{s+k} C_{2(k-1)}^{s+k-1} f(x+sh) \\ &\quad + \sum_{s=-k+2}^{k} (-1)^{s+k} C_{2(k-1)}^{s+k-2} f(x+sh) \\ &= C_{2(k-1)}^{0} f(x-kh) - \left(C_{2(k-1)}^{1} + 2C_{2(k-1)}^{0}\right) f(x+(-k+1)h) \\ &\quad + \sum_{-k+2}^{k-2} (-1)^{s+k} \left(C_{2(k-1)}^{s+k} + 2C_{2(k-1)}^{s+k-1} + C_{2(k-1)}^{s+k-2}\right) f(x+sh) \\ &\quad - \left(2C_{2(k-1)}^{2k-2} + C_{2(k-1)}^{2k-3}\right) f(x+(k-1)h) + C_{2(k-1)}^{2k-2} f(x+kh). \end{aligned}$$

Noting that $C_{2(k-1)}^{1} + 2C_{2(k-1)}^{0} = C_{2k}^{1}$, $2C_{2(k-1)}^{2k-2} + C_{2(k-1)}^{2k-3} = C_{2k}^{2k-1}$ and $C_{2(k-1)}^{s+k} + 2C_{2(k-1)}^{s+k-1} + C_{2(k-1)}^{s+k-2} = C_{2k}^{s+k}$ proves Equation (6). Using Equations (5) and (6) we have,

$$|\Delta_h^{2k} f(x)| \le h^2 \sup_{|x-t|\le|h|} |\Delta_h^{2(k-1)} f^{(2)}(t)| \le \cdots \le h^{2k} \sup_{|x-t|\le k|h|} |f^{(2k)}(t)|. \tag{7}$$

If $f(t) = \mathrm{e}^{-t^2/2}$, then $f^{(2k)}(t) = P_{2k}(t) f(t)$ where P_{2k} is a polynomial of degree $2k$. We can bound the polynomial by the exponential function, so that for any $\varepsilon \in (0,1)$ there exists a constant $c_\varepsilon > 0$

such that

$$|f^{(2k)}(t)| \le c_\varepsilon \mathrm{e}^{-(1-\varepsilon)t^2/2}. \tag{8}$$

Let $|h| \le \varepsilon(1+|x|)$ and consider two cases. First, suppose that $|x| \ge 1$. Then, $|h| \le 2\varepsilon|x|$, so that $|x-t| \le k|h|$ implies $|t| = |x+t-x| \ge |x| - |t-x| \ge |x| - 2\varepsilon k|x|$. Assuming that $2\varepsilon k < 1$, from Equation (8) we have

$$\sup_{|x-t|\le k|h|} |f^{(2k)}(t)| \le c_\varepsilon \mathrm{e}^{-(1-\varepsilon)(1-2\varepsilon k)^2x^2/2} \quad \text{if } |h| \le \varepsilon(1+|x|). \tag{9}$$

Second, suppose that $|x| < 1$. Since the function on the right-hand side of Equation (8) is bounded from above by c_ε for any t and the function $\mathrm{e}^{-(1-\varepsilon)x^2/2}$ is bounded away from zero for $|x| < 1$,

$$\sup_{|x-t|\le k|h|} |f^{(2k)}(t)| \le c_\varepsilon \le \tilde{c}_\varepsilon \mathrm{e}^{-(1-\varepsilon)x^2/2} \quad \text{if } |h| \le \varepsilon(1+|x|).$$

The last inequality together with Equation (9) and (7) proves Equation (3).

(b) We prove the statement for $f(t) = (1+t^2)^{-1}$. By induction it is easy to show that, for any natural n, $f^{(n)}(t) = P_n(t)f^{n+1}(t)$ where P_n is a polynomial of order n. Indeed, $f^{(1)}(t) = -2t(1+t^2)^{-2} = P_1(t)f^2(t)$. Suppose the formula is true for some $n > 1$, then

$$\begin{aligned} f^{(n+1)}(t) &= P_n^{(1)}(t)f^{n+1}(t) + P_n(t)(n+1)f^n(t)f^{(1)}(t) \\ &= [P_n^{(1)}(t)(1+t^2) - 2(n+1)tP_n(t)]f^{n+2}(t) = P_{n+1}(t)f^{n+2}(t). \end{aligned}$$

Since $|P_{2k}(t)| = \left|\sum_{j=0}^{2k} a_j t^j\right| \le \sum_{j=0}^{2k} |a_j|(1+t^2)^{j/2} \le c(1+t^2)^k$ by Equation (7) it follows that

$$|\Delta_h^{2k} f(x)| \le h^{2k} \sup_{|x-t|\le k|h|} |f^{(2k)}(t)| \le ch^{2k} \sup_{|x-t|\le k|h|} f^{k+1}(t). \tag{10}$$

Let $|h| \le \varepsilon(1+|x|)$ where $\varepsilon = 1/(4k)$ and suppose $|x| \ge 1$. As above, we have $|t| \ge |x|(1-2\varepsilon k) = |x|/2$. Then, $f(t) \le 4/(4+x^2) \le f(x)$ and Equation (4) follows from Equation (10). Now, suppose $|x| \le 1$, then $2f(x) \ge 1$. Since $f(t) \le 1$ we have from Equation (10) that $|\Delta_h^{2k} f(x)| \le ch^{2k} \le ch^{2k} f^{k+1}(x)2^{k+1}$, which completes the proof. ■

We note that Equation (7) shows that boundedness of $f^{(2k)}(x)$ implies a Lipschitz condition of order $2k$. A full description of the relationships between smoothness requirements in terms of derivatives and Lipschitz conditions can be found in Besov, Il'in, and Nikol'skiĭ (1978). We now turn to the definition of a family of kernels that will be used in constructing the new estimator we propose.

2.2. *Kernels and the proposed estimator*

For a kernel K and natural number k we define the set $\{M_k(x)\}_{k=1,2,3,\ldots}$ where

$$M_k(x) = -\frac{1}{c_{k,0}} \sum_{|s|=1}^{k} \frac{c_{k,s}}{|s|} K\left(\frac{x}{s}\right). \tag{11}$$

In this context we call K a seed kernel for M_k. The main impetus for the definition of $M_k(x)$ is that it allows us to express the bias of our proposed estimator in terms of higher order finite differences of the density f (Theorem 3). Let $\lambda_{k,s} = (-1)^{s+1}(k!)^2/(k+s)!(k-s)!, s = 1, \ldots, k$ and since $-(c_{k,s}/c_{k,0}) = -(c_{k,-s}/c_{k,0}) = \lambda_{k,s}, s = 1, \ldots, k$, Equation (11) can also be written as

$M_k(x) = \sum_{s=1}^{k}(\lambda_{k,s}/s)(K(x/s) + K(-(x/s)))$. It follows by construction that M_k is symmetric, that is $M_k(x) = M_k(-x), x \in \Re$. Since the coefficients $c_{k,s}$ satisfy $\sum_{|s|=0}^{k} c_{k,s} = (1-1)^{2k} = 0$, we have

$$-\frac{1}{c_{k,0}}\sum_{|s|=1}^{k} c_{k,s} = 1 \quad \text{or} \quad \sum_{s=1}^{k}\lambda_{k,s} = \frac{1}{2}. \tag{12}$$

It is therefore the case that Equations (1) and (12) imply that

$$\int_{-\infty}^{+\infty} M_k(x)\,\mathrm{d}x = \sum_{s=1}^{k}\frac{\lambda_{k,s}}{s}\left(\int_{-\infty}^{+\infty} K\left(\frac{x}{s}\right)\mathrm{d}x + \int_{-\infty}^{+\infty} K\left(-\frac{x}{s}\right)\mathrm{d}x\right) = 1,$$

which establishes that every $M_k(x)$ is a kernel for all k. The following theorem gives some properties of the family $\{M_k(x)\}_{k=1,2,\ldots}$ based on the seed kernel K.

Theorem 2 *Let* $G(x) = K(x) + K(-x)$ *and* $M_\infty(x) = \sum_{s=1}^{\infty}((-1)^{s+1}/s)G(x/s)$. *Suppose that the derivative* $K^{(1)}$ *exists and is bounded in some neighbourhood* $(-\delta, \delta)$ *of the origin. Then, we have*:

(a) *the series* $M_\infty(x)$ *absolutely converges at any* $x \neq 0$. *At* $x = 0$ *it converges conditionally to* $M_\infty(0) = 2K(0)\ln 2$,

(b) *Suppose, additionally, that* K *is bounded and continuous in* $\Re$ *and denote*

$$\|G\|_\infty = \sup_{x\in\Re}|G(x)| \quad \textit{and} \quad \|G^{(1)}\|_{\infty,\delta} = \sup_{x\in(-\delta,\delta)}|G^{(1)}(x)|.$$

For all $k > m \geq [|x|/\delta + 1]$ *(integer part) one has the estimate of the rate of convergence*

$$|M_k(x) - M_\infty(x)| \leq ||\lambda_{k,m-1}| - 1| \parallel G \parallel_\infty \sum_{s=1}^{m-1}\frac{1}{s} + 2 \parallel G \parallel_\infty \frac{1}{m} + \left(2\max\{\parallel G^{(1)} \parallel_{\infty,\delta} |x|, \parallel G \parallel_\infty\} + \parallel G^{(1)} \parallel_{\infty,\delta} |x|\right)\sum_{s=m}^{\infty}\frac{1}{s^2} \tag{13}$$

which implies locally uniform convergence of M_k *to* M_∞ *and continuity of* M_∞.

(c) *Let* G *be differentiable everywhere and fix* $x > 0$. *If* $f_x(\lambda) = (1/\lambda)G(x/\lambda)$ *has a negative derivative* $(\mathrm{d}f_x/\mathrm{d}\lambda)(\lambda)$ *for all* $\lambda \geq 1$, *then* $(k/(k+1))G(x) > M_k(x) > 0$ *for all* k. *Consequently, when* $M_k(x) \to M_\infty(x)$ *we have* $0 \leq M_\infty(x) \leq G(x)$.

(d) *If* G *is infinitely differentiable, then so is* M_∞.

Proof (a) The statement about conditional convergence at $x = 0$ follows from $G(0) = 2K(0)$ and $\ln 2 = \sum_{s=1}^{\infty}(-1)^{s+1}/s$. Now, fix $x \neq 0$. For all large s, we have $[-x/s, x/s] \subset (-\delta, \delta)$ and by the mean value theorem there exists $\theta_s \in [-x/s, x/s]$ such that $G(x/s) = K^{(1)}(\theta_s)2x/s$. This implies absolute convergence $|\sum_{s=m}^{\infty}((-1)^{s+1}/s)G(x/s)| \leq c\sum_{s=m}^{\infty} 1/s^2$.

(b) We start by establishing two properties of the coefficients $\lambda_{k,s}$. Since $C_{2k}^{k} \geq C_{2k}^{k+1} \geq \cdots \geq C_{2k}^{2k} = 1$, one has

$$1 \geq |\lambda_{k,1}| \geq |\lambda_{k,2}| \geq \cdots \geq |\lambda_{k,k}| = \frac{1}{C_{2k}^{k}}. \tag{14}$$

Furthermore, from $(-1)^{s+1}\lambda_{k,s} = ((k-s+1)\cdots k)/((k+1)\cdots(k+s)) = ((1-(s-1))/k)\cdots(1-(1/k))1/((1+(1/k))\cdots(1+(s/k)))$ we see that for any fixed s

$$(-1)^{s+1}\lambda_{k,s} \uparrow 1 \quad \text{as } k \to \infty. \tag{15}$$

To prove convergence $M_k \to M_\infty$, we take arbitrary $1 < m < k < \infty$ and split M_k and M_∞ as

$$M_k(x) = \left(\sum_{s=1}^{m-1} + \sum_{s=m}^{k}\right) \frac{\lambda_{k,s}}{s} G\left(\frac{x}{s}\right) = S_{k,m} + R_{k,m},$$

$$M_\infty(x) = \left(\sum_{s=1}^{m-1} + \sum_{s=m}^{\infty}\right) \frac{(-1)^{s+1}}{s} G\left(\frac{x}{s}\right) = S_{\infty,m} + R_{\infty,m}.$$

Let $x \geq 0$ and take, without loss of generality, $m \geq [x/\delta + 1]$ in $R_{\infty,m}$ so that $\delta > x/m$. Rearrange

$$\sum_{s=m}^{\infty} \frac{(-1)^{s+1}}{s} G\left(\frac{x}{s}\right) = \sum_{s=0}^{\infty} \frac{1}{m+2s}\left(G\left(\frac{x}{m+2s}\right) - G\left(\frac{x}{m+2s+1}\right)\right)$$
$$+ \sum_{s=0}^{\infty} G\left(\frac{x}{m+2s+1}\right)\left(\frac{1}{m+2s} - \frac{1}{m+2s+1}\right).$$

For each s in the first sum, there exists a point $\theta_s \in [x/(m+2s+1), x/(m+2s)]$ such that

$$G\left(\frac{x}{m+2s}\right) - G\left(\frac{x}{m+2s+1}\right) = G^{(1)}(\theta_s)\frac{m+2s}{m+2s+1}.$$

The last two equations imply that

$$\begin{aligned}
|R_{\infty,m}| &= \left|\sum_{s=m}^{\infty} \frac{(-1)^{s+1}}{s} G\left(\frac{x}{s}\right)\right| \\
&\leq \sum_{s=0}^{\infty}\left(\frac{\| G^{(1)} \|_{\infty,\delta}\, x}{(m+2s)(m+2s+1)} + \frac{\| G \|_{\infty}}{(m+2s)(m+2s+1)}\right) \\
&\leq 2\max\{\| G^{(1)} \|_{\infty,\delta}\, x, \| G \|_{\infty}\} \sum_{s=0}^{\infty} \frac{1}{(m+2s)(m+2s+1)} \\
&\leq 2\max\{\| G^{(1)} \|_{\infty,\delta}\, x, \| G \|_{\infty}\} \sum_{s=m}^{\infty} \frac{1}{s^2}.
\end{aligned} \tag{16}$$

Note that Equations (14) and (15) imply that

$$\begin{aligned}
|S_{k,m} - S_{\infty,m}| &\leq \sum_{s=1}^{m-1} \frac{|\lambda_{k,s} - (-1)^{s+1}|}{s} G\left(\frac{x}{s}\right) \\
&\leq |\lambda_{k,m-1} - (-1)^{m}| \|G\|_\infty \sum_{s=1}^{m-1} \frac{1}{s} \to 0 \quad \text{as } k \to \infty.
\end{aligned} \tag{17}$$

For s between m and k there are points $\tau_s \in [0, x/s]$ such that $G(x/s) = G(0) + G^{(1)}(\tau_s)x/s$. Thus,

$$R_{k,m} = G(0)\sum_{s=m}^{k}\frac{\lambda_{k,s}}{s} + x\sum_{s=m}^{k}\frac{\lambda_{k,s}}{s^2}G'(\tau_s).$$

Because of Equation (14) $|\sum_{s=m}^{k}(\lambda_{k,s}/s^2)G^{(1)}(\tau_s)| \le \|G^{(1)}\|_{\infty,\delta}\sum_{s=m}^{\infty}1/s^2$. In the series $\sum_{s=m}^{k}\lambda_{k,s}/s$ the terms have alternating signs and monotonically declining absolute values. By the Leibniz theorem $|\sum_{s=m}^{k}\lambda_{k,s}/s| \le |\lambda_{k,m}|/m \le 1/m$. Therefore

$$|R_{k,m}| \le \frac{|G(0)|}{m} + x\|G'\|_{\infty,\delta}\sum_{s=m}^{\infty}\frac{1}{s^2}. \tag{18}$$

Combining Equations (16)–(18) yields Equation (13). Also, Equations (13) and (14) show that one can choose first a large m and then a large k to make the expression on the right-hand side of Equation (13) arbitrarily small. Finally, M_∞ is continuous as a locally uniform limit of continuous functions.

(c) Pairing the terms in M_k gives

$$\begin{aligned} M_k(x) &= \sum_{l=0}^{[k/2]-1}\left[\frac{\lambda_{k,2l+1}}{2l+1}G\left(\frac{x}{2l+1}\right) + \frac{\lambda_{k,2l+2}}{2l+2}G\left(\frac{x}{2l+2}\right)\right] + R_k \\ &= \sum_{l=0}^{[k/2]-1}[\lambda_{k,2l+1}f_x(2l+1) + \lambda_{k,2l+2}f_x(2l+2)] + R_k, \end{aligned}$$

where $\lambda_{k,2l+1}$ are all positive and $R_k = 0$, if k is even, and $R_k = (\lambda_{k,k}/k)G(x/k)$, if k is odd. Further, by the assumed negativity of $\mathrm{d}f_x(\lambda)/\mathrm{d}\lambda$ one has $f_x(2l+1) > f_x(2l+2)$ for all $l \ge 0$, so that

$$\begin{aligned} M_k(x) &= \sum_{l=0}^{[k/2]-1}\lambda_{k,2l+1}\left[f_x(2l+1) - \frac{1-(2l+1)/k}{1+(2l+2)/k}f_x(2l+2)\right] + R_k \\ &> \sum_{l=0}^{[k/2]-1}\lambda_{k,2l+1}f_x(2l+2)\left(1 - \frac{1-(2l+1)/k}{1+(2l+2)/k}\right) + R_k > R_k \ge 0. \end{aligned}$$

Similarly, $M_k(x) = (k/(k+1))G(x) + \sum_{l=1}^{[(k-1)/2]}[\lambda_{k,2l}f_x(2l) + \lambda_{k,2l+1}f_x(2l+1)] + R_k$ where all $\lambda_{k,2l}$ are negative, $R_k = 0$, if k is odd, and $R_k = (\lambda_{k,k}/k)G(x/k)$, if k is even. Hence,

$$\begin{aligned} M_k(x) &< \frac{k}{k+1}G(x) + \sum_{l=1}^{[(k-1)/2]}\lambda_{k,2l}f_x(2l+1)\left(1 - \frac{1-(2l/k)}{1+(2l+1)/k}\right) + R_k \\ &< \frac{k}{k+1}G(x) + R_k \le \frac{k}{k+1}G(x). \end{aligned}$$

(d) If $u_n^{(1)}(x)$ are continuous, then convergence of a series $\sum u_n(x)$ in addition to uniform convergence of the series of derivatives $\sum u_n^{(1)}(x)$ are sufficient for $(\sum u_n(x))^{(1)} = \sum u_n^{(1)}(x)$. Since $G^{(1)}$ is locally bounded, $\sum_{s=1}^{\infty}(-1)^{s+1}s^{-2}G^{(1)}(x/s)$ converges locally uniformly. Therefore, M_∞ is differentiable and $M_\infty^{(1)}(x) = \sum_{s=1}^{\infty}(-1)^{s+1}/s^2G^{(1)}(x/s)$. Uniform convergence implies also continuity of $M_\infty^{(1)}$. This type of argument applies to all higher order derivatives. ■

We note that $(\mathrm{d}f_x/\mathrm{d}\lambda)(\lambda) < 0$ for $\lambda \geq 1$ if, and only if, $G(x/\lambda) + G'(x/\lambda)(x/\lambda) > 0$ for $\lambda \geq 1$. For the Gaussian and Cauchy densities this is true if $x < 1$. It is worth pointing out that the negativity of the derivative in (c) is only a sufficient condition for $M_k > 0$ for all k.[1]

We are now ready to define a new family of alternative estimators which are similar to the Rosenblatt–Parzen estimator with the exception that K is replaced by M_k. Hence, we put for $k = 1, 2, \ldots$

$$\hat{f}_k(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h_n}M_k\left(\frac{X_j - x}{h_n}\right) = \frac{1}{n}\sum_{j=1}^{n} w_j,$$

where $w_j = (1/h_n)M_k((X_j - x)/h_n)$. Given the independent and identically distributed (IID) assumption (maintained everywhere), we have

$$E(\hat{f}_k(x)) = \frac{1}{n}\sum_{j=1}^{n} E(w_j) = E(w_1), \tag{19}$$

and

$$V(\hat{f}_k(x)) = \frac{1}{n^2}\sum_{j=1}^{n} V(w_j) = \frac{1}{n}V(w_1) = \frac{1}{n}(E(w_1^2) - (E(w_1))^2). \tag{20}$$

The next theorem reveals the main idea underlying our definition of the family $\{M_k\}_{k=1,2,\ldots}$.

Theorem 3 *For any $h_n > 0$ $B(\hat{f}_k(x)) = -(1/c_{k,0})\int_{-\infty}^{+\infty} K(t)\Delta_{h_n t}^{2k} f(x)\mathrm{d}t$.*

Proof From Equation (19), we have $E(\hat{f}_k(x)) = Ew_1 = (1/h_n)\int_{-\infty}^{+\infty} M_k((t - x)/h_n)f(t)\mathrm{d}t = \int_{-\infty}^{+\infty} M_k(t)f(x + h_n t)\mathrm{d}t$. Substitution of Equation (11) and change of variables give

$$E(\hat{f}_k(x)) = -\frac{1}{c_{k,0}}\sum_{|s|=1}^{k} c_{k,s}\int_{-\infty}^{+\infty} K(t)f(x + sh_n t)\mathrm{d}t. \tag{21}$$

Hence, from Equations (2) and (1) we get

$$\begin{aligned} B(\hat{f}_k(x)) &= -\frac{1}{c_{k,0}}\int_{-\infty}^{+\infty} K(t)\sum_{|s|=1}^{k} c_{k,s}f(x + sh_n t)\mathrm{d}t - f(x)\int_{-\infty}^{+\infty} K(t)\mathrm{d}t \\ &= -\frac{1}{c_{k,0}}\int_{-\infty}^{+\infty} K(t)\sum_{|s|=0}^{k} c_{k,s}f(x + sh_n t)\mathrm{d}t = -\frac{1}{c_{k,0}}\int_{-\infty}^{+\infty} K(t)\Delta_{h_n t}^{2k} f(x)\mathrm{d}t. \end{aligned} \tag{22}$$

■

3. Asymptotic properties

In this section, we give an asymptotic characterisation of the estimator we propose. We start by providing conditions under which the estimator is asymptotically (uniformly) unbiased. We note that Theorems 4 and 5 are general and do not rely on specific properties of the family of kernels $\{M_k\}_{k=1,2,\ldots}$.

Theorem 4 *Given a kernel K satisfying* (1) *and a random sample $\{X_j\}_{j=1}^{n}$ we have,*

(*a*) *If $f(x)$ is bounded and continuous in $\Re$ then* $\lim_{n\to\infty} B(\hat{f}_k(x)) = 0$ *for all $x \in \Re$.*
(*b*) *If $f(x)$ is bounded and uniformly continuous in $\Re$ then* $\lim_{n\to\infty}\sup_{x\in R} |\mathrm{Bias}(\hat{f}_k(x))| = 0$.

Proof (a) From Equations (21) and (1), boundedness and continuity of $f(x)$ we have by the dominated convergence theorem $E(\hat{f}_k(x)) \to -(1/c_{k,0})\sum_{|s|=1}^{k} c_{k,s} f(x)$. The desired property follows from Equation (12). (b) Using Equations (21), (12) and (1), we get $B(\hat{f}_k(x)) = -(1/c_{k,0})\sum_{|s|=1}^{k} c_{k,s}\int_{-\infty}^{+\infty} K(t)[f(x+sh_nt)-f(x)]\mathrm{d}t$. Hence, for any $\delta > 0$

$$\begin{aligned}|B(\hat{f}_k(x))| &\leq c\sum_{|s|=1}^{k}\left[\int_{|sh_nt|\leq\delta}|K(t)[f(x+sh_nt)-f(x)]|\mathrm{d}t\right.\\ &\quad\left.+\int_{|sh_nt|>\delta}|K(t)[f(x+sh_nt)-f(x)]|\mathrm{d}t\right]\\ &\leq c\sum_{|s|=1}^{k}\left[\sup_{|y|\leq\delta,\ x\in\Re}|f(x+y)-f(x)|\int_{\Re}|K(t)|\mathrm{d}t + 2\sup_{x\in\Re}|f(x)|\int_{|sh_nt|>\delta}|K(t)|\mathrm{d}t\right].\end{aligned}$$

To make the right-hand side expression small, we can choose first a small δ and then a small h_n. ■

We state the next theorem without proof since it follows closely the proof of Theorem 2.8 in Pagan and Ullah (1999) with their kernel K replaced by our kernel M_k.

THEOREM 5 *If the characteristic function ϕ_K of K is integrable and $nh_n^2 \to \infty$, then*

$$\lim_{n\to\infty} E\left(\sup_{x\in\Re}|\hat{f}_k(x) - E(\hat{f}_k(x))|\right) = 0.$$

Note that if the conditions from Theorems 4 (b) and 5 are combined, we can write

$$E\left(\sup_{x\in\Re}|\hat{f}_k(x)-f(x)|\right) \leq E\left(\sup_{x\in\Re}|\hat{f}_k(x)-E(\hat{f}_k(x))|\right) + \sup_{x\in\Re}|B(\hat{f}_k(x))| \to 0,$$

establishing by the use of Markov's inequality that $\hat{f}_k(x)$ is uniformly consistent. In the next theorem, we provide the order of decay for the bias and variance of our estimator.

THEOREM 6 *Suppose that* (*a*) *$f(x)$ is bounded and continuous*, (*b*) *there exist functions $H_{2k}(x) > 0$ and $\varepsilon_{2k}(x) > 0$ such that*

$$|\Delta_h^{2k} f(x)| \leq H_{2k}(x)h^{2k} \quad \textit{for all } |h| \leq \varepsilon_{2k}(x) \tag{23}$$

and (*c*) $\int_{-\infty}^{\infty}|K(t)|t^{2k}\mathrm{d}t < \infty$. *Then, for all $x \in \Re$ and $0 < h_n \leq \varepsilon_{2k}(x)$*

$$|B(\hat{f}_k(x))| \leq ch_n^{2k}(H_{2k}(x) + \varepsilon_{2k}^{-2k}(x)), \tag{24}$$

where the constant c does not depend on x or h_n. Suppose additionally that (*d*) *K is bounded, the set $\{t : |K(t)| > 1\}$ is bounded and there exist functions $H_2(x) > 0$ and $\varepsilon_2(x) > 0$ such that*

$$|\Delta_h^2 f(x)| \leq H_2(x)h^2 \quad \textit{for all } |h| \leq \varepsilon_2(x). \tag{25}$$

Then, for all $x \in \Re$ and $0 < h_n \leq \min\{\varepsilon_{2k}(x), \varepsilon_2(x)\}$

$$V(\hat{f}_k(x)) = \frac{1}{nh_n}\left\{f(x)\int_{-\infty}^{\infty}M_k^2(t)\mathrm{d}t + R_2(x,h_n) - h_n[f(x) + R_{2k}(x,h_n)]^2\right\}, \tag{26}$$

where the residuals satisfy

$$|R_2(x,h_n)| \leq c_1h_n^2(H_2(x)+\varepsilon_2^{-2}(x)), \quad |R_{2k}(x,h_n)| \leq c_2h_n^{2k}(H_{2k}(x)+\varepsilon_{2k}^{-2k}(x)) \tag{27}$$

with constants c_1 and c_2 independent of x and h_n.

Proof Condition (c) implies for any $N > 0$

$$\int_{|t|>N} |K(t)|\mathrm{d}t \le \int_{|t|>N} |K(t)||\frac{t}{N}|^{2k}\mathrm{d}t \le N^{-2k}\int_{-\infty}^{\infty} |K(t)|t^{2k}\mathrm{d}t. \tag{28}$$

Using Equation (22) and conditions (a) and (b) we have

$$\begin{aligned}|B(\hat{f}_k(x))| &\le c_1\left(\int_{|h_nt|\le\varepsilon_{2k}(x)} + \int_{|h_nt|>\varepsilon_{2k}(x)}\right)|K(t)\Delta_{h_nt}^{2k}f(x)|\mathrm{d}t\\ &\le c_2\left[H_{2k}(x)\int_{|h_nt|\le\varepsilon_{2k}(x)}|K(t)|(h_nt)^{2k}\mathrm{d}t + \sup_{x\in\Re}|f(x)|\int_{|h_nt|>\varepsilon_{2k}(x)}|K(t)|\mathrm{d}t\right].\end{aligned}$$

It remains to apply Equation (28) and condition (c) to obtain Equation (24).

Now we proceed with the derivation of Equation (26). According to Equation (20), we need to evaluate $E(w_1^2)$ and $(Ew_1)^2$. By Equations (19) and (24),

$$E(w_1) = E(\hat{f}_k(x)) = f(x) + R_{2k}(x, h_n) \quad \text{where } R_{2k} \text{ satisfies Equation (27).} \tag{29}$$

Now, $E(w_1^2) = (1/h_n^2)\int M_k^2((t-x)/h_n)f(t)\mathrm{d}t = (1/h_n)\int M_k^2(t)f(x+h_nt)\mathrm{d}t$ and by symmetry of M_k we have

$$\begin{aligned}\int M_k^2(t)f(x+h_nt)\mathrm{d}t - f(x)\int M_k^2(t)\mathrm{d}t &= \left(\int_0^\infty + \int_{-\infty}^0\right)M_k^2(t)f(x+h_nt)\mathrm{d}t\\ &\quad - 2\int_0^\infty M_k^2(t)f(x)\mathrm{d}t\\ &= \int_0^\infty M_k^2(t)\Delta_{h_nt}^2 f(x)\mathrm{d}t.\end{aligned}$$

Using Equation (25) the same way we applied Equation (23) to obtain Equation (24), we get

$$\int M_k^2(t)f(x+h_nt)\mathrm{d}t = f(x)\int M_k^2(t)\mathrm{d}t + R_2(x, h_n), \tag{30}$$

where the residual $R_2(x, h_n)$ satisfies Equation (27). In this argument, we used the fact that $\int_{-\infty}^{\infty} K^2(t)t^2\mathrm{d}t = (\int_{t:|K(t)|>1} + \int_{t:|K(t)|<1})K^2(t)t^2\mathrm{d}t \le c\,l(\{t : |K(t)| > 1\}) + \int_{-\infty}^{\infty} t^2|K(t)|\mathrm{d}t < \infty$, where $l(\{t : |K(t)| > 1\})$ denotes the measure of the set $\{t : |K(t)| > 1\}$. As a result $\int_{-\infty}^{\infty} M_k^2(t)t^2\mathrm{d}t < \infty$. Note that Equation (26) is a consequence of Equation (20) and Equations (29) and (30). ■

We note that the order of the bias for our estimator is similar to that attained by a Rosenblatt–Parzen estimator constructed with a kernel of order $2k$ for $k = 1, 2, \ldots$. The advantage of our estimator in this case results from the fact that it can be constructed to be nonnegative and, as observed from Theorem 1, boundedness of $f^{(2k)}$ implies a Lipschitz condition of order $2k$. In addition, if x is fixed and $f(x) \ne 0$ then Equation (26) can be (for small h_n) simplified to

$$V(\hat{f}_k(x)) = \frac{1}{nh_n}\left\{f(x)\int_{-\infty}^{\infty} M_k^2(t)\mathrm{d}t + f(x)O(h_n)\right\} \tag{31}$$

which is of order similar to that of a Rosenblatt–Parzen estimator.

It is also instructive to compare the results in Theorem 6 with those obtained for the nonparametric density estimator $f_J(x) = f_{\mathrm{R}}(x)(1/nh_n)\sum_{j=1}^n(1/f_{\mathrm{R}}(X_j))K((X_j - x)/h_n)$ proposed by

Jones et al. (1995). The fact that $f_{\mathrm{R}}(X_j)$ appears in the denominator creates theoretical difficulties for the analysis of the bias of $f_J(x)$. In particular, the expressions for the bias obtained by Jones et al. (1995) ignore terms of order $O((nh_n)^{-1})$ and $o(h_n^4)$, and as a result the expression for the bias is valid only asymptotically. Unlike their expressions, our results hold for all bandwidths h_n. The same comments apply to the variance of $f_J(x)$.

Certain seed kernels may not satisfy condition (c) in Theorem 6. One example is the Cauchy kernel which has been considered above. In the next theorem, we show that the Cauchy kernel can produce undesirable results when attempting to reduce bias.

THEOREM 7 *Let K be a Cauchy seed kernel and, for a given k, let H_{2k} and ε_{2k} be Lipschitz parameters as implied by Theorem* 1 - (*b*) : $H_{2k}(x) = cK^{k+1}(x)$, $\varepsilon_{2k}(x) = \varepsilon(1 + |x|)$. *Denote $q_0 = (2k+1)/2$, take any $q > q_0$ and let $p = q/(q-1), \alpha = (2k/q) - (1/p)$. Then, there exists a small $h_0 > 0$ such that*

$$|B(\hat{f}_k(x))| \le c(H_{2k}(x)\varepsilon_{2k}(x)^{(2k+1)/p}|h|^{\alpha} + |h|\varepsilon_{2k}(x)^{-1}) \quad \textit{for } |h| \le h_0. \tag{32}$$

Since $\alpha < 1$ can be made arbitrarily close to 1 *by selecting q close to q_0 we have $|B(\hat{f}_k(x))| = O(h_n^{\alpha})$ irrespective of the choice of k.*

Proof We have $(1/p) + (1/q) = 1$ and by Hölder's inequality

$$\begin{aligned}\int_{|ht|\le\varepsilon_{2k}(x)} |K(t)\Delta_{ht}^{2k} f(x)|\mathrm{d}t &= \int_{|ht|\le\varepsilon_{2k}(x)} K(t)|\Delta_{ht}^{2k} f(x)|^{(1/p)+(1/q)}\mathrm{d}t \\ &\le \left(\int_{|ht|\le\varepsilon_{2k}(x)} |\Delta_{ht}^{2k} f(x)|\mathrm{d}t\right)^{1/p} \left(\int_{|ht|\le\varepsilon_{2k}(x)} K(t)^q|\Delta_{ht}^{2k} f(x)|\mathrm{d}t\right)^{1/q}.\end{aligned} \tag{33}$$

Applying Equation (23) we can bind the right-hand expression by

$$(H_{2k}(x)|h|^{2k})^{1/p}\left(\int_{|ht|\le\varepsilon_{2k}(x)} t^{2k}\mathrm{d}t\right)^{1/p} (H_{2k}(x)|h|^{2k})^{1/q}\left(\int_{|ht|\le\varepsilon_{2k}(x)} K(t)^q t^{2k}\mathrm{d}t\right)^{1/q}. \tag{34}$$

Here,

$$\int_{|t|\le\varepsilon_{2k}(x)/|h|} t^{2k}\mathrm{d}t = 2\int_0^{\varepsilon_{2k}(x)/|h|} t^{2k}\mathrm{d}t = c(\varepsilon_{2k}(x)/|h|)^{2k+1}. \tag{35}$$

The condition for convergence of $\int_{-\infty}^{\infty} K(t)^q t^{2k}\mathrm{d}t$ is $2q - 2k > 1$ and it is satisfied by our choice of q. Hence, Equations (33)–(35) lead to

$$\begin{aligned}\int_{|ht|\le\varepsilon_{2k}(x)} |K(t)\Delta_{ht}^{2k} f(x)|\mathrm{d}t &\le cH_{2k}(x)|h|^{2k-(2k+1)/p}(\varepsilon_{2k}(x))^{(2k+1)/p} \\ &= cH_{2k}(x)|h|^{\alpha}(\varepsilon_{2k}(x))^{(2k+1)/p}.\end{aligned} \tag{36}$$

Furthermore,

$$\int_{|ht|>\varepsilon_{2k}(x)} |K(t)\Delta_{ht}^{2k} f(x)|\mathrm{d}t \le c\sup_{x\in\Re}|f(x)|\int_{|ht|>\varepsilon_{2k}(x)} K(t)\mathrm{d}t. \tag{37}$$

Since $\varepsilon_{2k}(x) = \varepsilon(1 + |x|) \ge \varepsilon$, $K(t)$ can be estimated by $c_1 t^{-2}$ in the domain of interest for all $|h| \le h_0$ where h_0 is sufficiently small. Hence,

$$\int_{|t|>\varepsilon_{2k}(x)/|h|} K(t)\mathrm{d}t \le c_1\int_{|t|>\varepsilon_{2k}(x)/|h|} \frac{\mathrm{d}t}{t^2} = c_2\frac{|h|}{\varepsilon_{2k}(x)}. \tag{38}$$

Equations (36)–(38) prove Equation (32).

The exponent α satisfies $\alpha = (2k/q) - 1 + (1/q) = (2k + 1/q_0)(q_0/q) - 1 = 2(q_0/q) - 1 < 1$ and can be made arbitrarily close to 1 by selecting $q > q_0$ close to q_0. ■

The Cauchy density declines at infinity too slowly, and this slow decay is inherited by our kernel M_k. As a result, the reduction in bias achieved through an increase in the Lipschitz smoothness is limited, even when that smoothness and, correspondingly, the order k of the kernel M_k is very high. We have also verified this in Monte Carlo simulations. Better estimation results have been obtained (Section 4) using the Gaussian density as a seed but in this case M_k is not necessarily nonnegative. Other seed kernels, for which M_k is nonnegative, may exist but we have failed to find one.

In many instances there is an interest in integration of bias and variance expressions over the range of the random variable X. In this case, it is necessary to investigate the convergence of integrals involving x before omitting terms of higher order in h_n. This is done in the following theorem, where we denote the mean squared error by $\mathrm{MSE}(\hat{f}_k(x)) = V(\hat{f}_k(x)) + B(\hat{f}_k(x))^2$ and the integrated MSE by $\mathrm{IMSE} = \int_{\Re} \mathrm{MSE}(\hat{f}_k(x))\mathrm{d}x$.

Theorem 8 *Let assumptions* (*a*)–(*d*) *of Theorem* 6 *be satisfied. Then*,

(1) *If* $h_n \to 0$ *and* $n \to \infty$ *in such a way that* $nh_n \to \infty$, *then* $\mathrm{MSE}(\hat{f}_k(x)) \to 0$. *If, additionally*, f, H_2, H_{2k}, ε_2^{-1} *and* ε_{2k}^{-1} *are bounded, then* $\sup_{x\in\Re} \mathrm{MSE}(\hat{f}_k(x)) \to 0$.

(2) *Suppose that* $H_{2k}, \varepsilon_{2k}^{-2k} \in L_2(\Re)$, $f, H_2, \varepsilon_2^{-2} \in L_1(\Re)$, *then* IMSE *is bounded by a function of the form* $\phi(h) = c_1/(nh) + c_2 h^{4k}$. *The optimal* h_n *resulting from the minimisation of* ϕ *is of order* $h_{\mathrm{opt}} \asymp n^{-1/(4k+1)}$.

Proof (1) The first statement follows from Equations (24) and (31). The second is an implication of Equations (24), (26) and (27).

(2) Replacing $V(\hat{f}_k(x))$ and $B(\hat{f}_k(x))$ in IMSE by their approximations (24) and (26), we get an approximation for IMSE, which we denote by

$$\mathrm{AIMSE} = \int_{\Re} \left\{ \frac{1}{nh} \left\{ f(x) \int_{\Re} M_k^2(t)\mathrm{d}t + R_2(x,h) - h[f(x) + R_{2k}(x,h)]^2 \right\} + R_{2k}^2(x,h) \right\} \mathrm{d}x.$$

Under the conditions imposed, the integrals in x are finite. $f \in L_2(\Re)$ because $f \in L_1(\Re) \cap L_\infty(\Re)$. Since all terms of higher order in h can be omitted for small h, we have $\mathrm{AIMSE} \le c_1/(nh) + c_2 h^{4k} = \phi(h)$. ■

Note that for the optimal h_n we have $nh_n \to \infty$, $nh_n^2 \to \infty$, like in the classical treatment of the Rosenblatt–Parzen estimator. By Theorem 1, for the Gaussian density all conditions of Theorem 8 are satisfied. We now establish the asymptotic normality of our estimator under suitable normalisation.

Theorem 9 *Suppose that* f *is continuous and bounded*, $f(x) > 0$, *there exist functions* $H_2(x) > 0$ *and* $\varepsilon_2(x) > 0$ *such that Equation* (25) *holds, and for some* $\delta > 0$, $\int_{\Re} |K(t)|^{2+\delta}(t)\mathrm{d}t < \infty$. *If* $nh_n \to \infty$, *then*

$$(nh_n)^{1/2}\left(\hat{f}_k(x) - E(\hat{f}_k(x))\right) \xrightarrow{\mathrm{d}} N\left(0, f(x)\int_{\Re} M_k^2(t)\mathrm{d}t\right). \tag{39}$$

If additionally,

$$nh_n^{4k+1} \longrightarrow 0, \tag{40}$$

then

$$(nh_n)^{1/2}(\hat{f}_k(x) - f(x)) \xrightarrow{\mathrm{d}} N\left(0, f(x)\int_{\Re} M_k^2(t)\mathrm{d}t\right). \tag{41}$$

Proof Normalising $\hat{f}_k(x) - E(\hat{f}_k(x))$ by its standard deviation, we obtain by Equations (19) and (20)

$$S_n \equiv \frac{\hat{f}_k(x) - E(\hat{f}_k(x))}{V(\hat{f}_k(x))^{1/2}} = \frac{1}{n}\sum_{j=1}^{n} \frac{w_j - E(w_j)}{(V(w_1)/n)^{1/2}} = \sum_{j=1}^{n} X_{nj}.$$

Here $X_{nj} = (w_j - E(w_j))/(nV(w_1))^{1/2},\ E(X_{nj}) = 0,\ V(X_{nj}) = 1/n,\ V(S_n) = 1.$ Recall that X_i are IID and therefore so are X_{nj}. Using the notation in the Lindeberg–Feller Theorem (Davidson 1994) $\mu_{nj} = 0,\ \sigma_{nj} = 1/n,\ \sigma_n = 1$ and $\max_j \sigma_{nj}/\sigma_n \to 0,\ n \to \infty$. Let F_{nj} be the distribution function of X_{nj}. All F_{nj} coincide with F_{n1} and the Lindeberg function takes the form

$$\begin{aligned}\lambda &\equiv \frac{1}{\sigma_n^2}\sum_{j=1}^{n}\int_{|x|>\varepsilon} x^2 \mathrm{d}F_{nj}(x) = n\int_{|x|>\varepsilon} x^2\mathrm{d}F_{n1}(x) \le \frac{n}{\varepsilon^\delta}\int |x|^{2+\delta}\mathrm{d}F_{n1}(x)\\ &= \frac{n}{\varepsilon^\delta}E(|X_{n1}|^{2+\delta}) = \frac{nE(|w_1 - E(w_1)|^{2+\delta})}{\varepsilon^\delta (nV(w_1))^{1+\delta/2}}.\end{aligned}$$

Here by Minkowski's and Hölder's inequality $E(|w_1 - E(w_1)|^{2+\delta}) \le 2^{2+\delta}E(|w_1|^{2+\delta})$. In addition, by a result similar to Equation (30) we have

$$\begin{aligned}E(|w_1 - E(w_1)|^{2+\delta}) &\le \left(\frac{2}{h_n}\right)^{2+\delta}\int_{\Re}\left|M_k\left(\frac{s-x}{h_n}\right)\right|^{2+\delta} f(s)\mathrm{d}s\\ &= 2\left(\frac{2}{h_n}\right)^{1+\delta}\int_{\Re}|M_k|^{2+\delta}(t)f(x+h_nt)\mathrm{d}t \asymp 2\left(\frac{2}{h_n}\right)^{1+\delta} f(x)\\ &\quad\times\int_{\Re}|M_k|^{2+\delta}(t)\mathrm{d}t.\end{aligned}$$

By Equation (31) $V(w_1) = nV(\hat{f}_k(x)) \asymp (1/h_n)f(x)\int_{\Re} M_k^2(t)\mathrm{d}t$. Consequently,

$$\lambda \le \frac{(nh_n)^{-\delta/2}2^{2+\delta}f(x)\int_{\Re}|M_k|^{2+\delta}(t)\mathrm{d}t}{\varepsilon^\delta(f(x)\int_{\Re}M_k^2(t)\mathrm{d}t)^{1+\delta/2}} = O((nh_n)^{-\delta/2}) \to 0.$$

By the Lindeberg–Feller Theorem $S_n \overset{\mathrm{d}}{\to} N(0,1)$. Since $nh_nV(\hat{f}_k(x)) \to f(x)\int_{\Re}M_k^2(t)\mathrm{d}t$, the equation $(nh_n)^{1/2}(\hat{f}_k(x) - E(\hat{f}_k(x))) = (nh_nV(\hat{f}_k(x)))^{1/2}S_n$ implies Equation (39). Finally, since $(nh_n)^{1/2}(\hat{f}_k(x) - f(x)) = (nh_n)^{1/2}(\hat{f}_k(x) - E(\hat{f}_k(x))) + (nh_n)^{1/2}(E(\hat{f}_k(x)) - f(x))$ we see that Equation (41) is true if $\lim(nh_n)^{1/2}(E(\hat{f}_k(x)) - f(x)) = 0$. By Equation (24) this follows from Equation (40). ■

4. Monte Carlo study and example

In this section, we perform a small Monte Carlo study to implement our proposed estimator and illustrate its finite sample performance. In addition, we provide an example that shows that the negativity problem of density estimators based on higher order kernels (or local polynomial estimators) can be severe while our proposed estimator is everywhere positive.

4.1. *Monte Carlo study*

We implement our estimator and for comparison purposes we also include the Rosenblatt–Parzen estimator and the local quadratic estimator of Lejeune and Sarda (1992), which is given by $\hat{f}_{\mathrm{LS}}(x) = (1/nh_n)\sum_{i=1}^{n} W((X_i - x)/h_n)$, where $W(u) = (\frac{3}{2} - \frac{1}{2}u^2)K(u)$ and $K(u)$ is the Gaussian kernel. We note that $W(u)$ is a fourth-order kernel, and consequently, $\hat{f}_{LS}(x)$ can be negative as all other density estimators obtained using different higher order kernels.

We consider simulated data from five different densities. The first four were proposed in Marron and Wand (1992) and are examples of normal mixtures. They are: (1) Gaussian $(f_1(x) \equiv N(0,1))$, (2) Bimodal $(f_2(x) \equiv \frac{1}{2}N(-1, 4/9) + \frac{1}{2}N(1, 4/9))$, (3) Separated-Bimodal $(f_3(x) \equiv \frac{1}{2}N(-1.5, 1/4) + \frac{1}{2}N(1.5, 1/4))$ and (4) Trimodal $(f_4(x) \equiv (9/20)N(-6/5, 9/25) + \frac{9}{20}N(6/5, 9/25) + \frac{1}{10}N(0, 1/16))$. The fifth density is given by

$$f_5(x) = \begin{cases} \dfrac{1}{c}\exp\left(\dfrac{-(x+2)^2}{2}\right) & \text{if } x \leq -1, \\ \dfrac{1}{c}\exp\left(\dfrac{-(x-2)^2}{2}\right) & \text{if } x \geq 1\ , \\ \dfrac{1}{2c}\exp(-1/2)(x^2+1) & \text{if } -1 < x < 1, \end{cases}$$

where $c = 2F_1(1)\sqrt{2\pi} + \frac{4}{3}\exp(-1/2)$, $F_1(a) = \int_{-\infty}^{a} f_1(x)\mathrm{d}x$. It is easy to verify that $f_5^{(2)}(x)$ is not continuous for all x, but it does satisfy a Lipschitz condition of order 2 for all x.

For each of these densities 1000 samples of size $n = 200$, 400 and 600 were generated.[2] In our first set of simulations five estimators were obtained for each sample: $\hat{f}_k(x)$ for $k = 2, 4, 8$, $\hat{f}_{\mathrm{R}}(x)$ and $\hat{f}_{\mathrm{LS}}(x)$. The bandwidths for each estimator (say $\hat{f}_{\mathrm{E}}(x)$) were selected by minimising integrated squared error $I(\hat{f}_{\mathrm{E}}) = \int(\hat{f}_{\mathrm{E}}(x) - f(x))^2\mathrm{d}x$ for each simulated sample. In practice, this bandwidth is infeasible given that $f(x)$ is unknown. However, in the context of a Monte Carlo study it is desirable since estimation performance is not impacted by the noise introduced through a data-driven bandwidth selection. See Jones and Signorini (1997) for an approach that is similar to ours. Table 1 provides average absolute bias (B) and average MSE for each estimator and each density considered for $n = 200, 400$, respectively.[3]

In our second set of simulations, we consider the performance of $\hat{f}_2(x)$, $\hat{f}_{\mathrm{R}}(x)$ and $\hat{f}_{\mathrm{LS}}(x)$ based on data-driven bandwidths obtained from the minimisation of a suitably defined cross-validation

Table 1. Five estimators with optimal bandwidth (h_n). Average bias $(\times 10^3)$ (B), mean squared error $(\times 10^3)$ (MSE).

	$f_1(x)$		$f_2(x)$		$f_3(x)$		$f_4(x)$		$f_5(x)$	
Estimators	B	MSE	B	MSE	B	MSE	B	MSE	B	MSE
$n = 200$										
$\hat{f}_{\mathrm{R}}$	6.637	0.317	7.644	0.437	8.710	0.645	8.919	0.529	14.020	0.324
$\hat{f}_{LS}$	5.239	0.251	6.079	0.403	6.784	0.544	8.126	0.523	13.285	0.269
$\hat{f}_2$	5.493	0.250	6.403	0.410	7.038	0.551	8.292	0.521	13.453	0.279
$\hat{f}_4$	5.109	0.235	5.839	0.407	6.936	0.539	8.097	0.536	10.294	0.159
$\hat{f}_8$	4.936	0.216	5.744	0.403	6.999	0.557	8.045	0.547	12.316	0.231
$n = 400$										
$\hat{f}_{\mathrm{R}}$	4.975	0.184	5.959	0.271	6.674	0.393	6.839	0.344	8.700	0.128
$\hat{f}_{LS}$	3.727	0.132	4.629	0.236	4.996	0.313	5.759	0.334	7.820	0.098
$\hat{f}_2$	3.908	0.135	4.845	0.243	5.195	0.321	6.010	0.333	7.926	0.102
$\hat{f}_4$	3.762	0.134	4.348	0.225	5.225	0.308	5.638	0.329	7.618	0.127
$\hat{f}_8$	3.779	0.125	4.240	0.230	4.940	0.302	5.560	0.331	7.280	0.087

function. Thus, we define

$$h^{\mathrm{CV}} \equiv \underset{h}{\operatorname{argmin}} \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} G * G\left(\frac{X_i - X_j}{h}\right) - 2\frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} G\left(\frac{X_i - X_j}{h}\right),$$

where $G * G(u) = \int G(u-t)G(t)\mathrm{d}u$. For $\hat{f}_2(x)$, $\hat{f}_{\mathrm{R}}(x)$ and $\hat{f}_{\mathrm{LS}}(x)$, $G(u)$ is, respectively, $M_2(u)$, $K(u)$, and $W(u)$. Given that $K(u)$ is a Gaussian kernel we can easily obtain through Fourier transform methods the convolutions $W * W(u) = (1/2\sqrt{2\pi})\exp(-\frac{1}{4}u^2)(u^4/64 - 7x^2/16 + 27/16)$ and $M_2 * M_2(u) = 16/(9\sqrt{2}\sqrt{2\pi})\exp(-\frac{1}{4}u^2) - 8/(9\sqrt{5}\sqrt{2\pi})\exp(-\frac{1}{10}u^2) + 4/(3\sqrt{2\pi})\exp(-\frac{1}{16}u^2)$. Table 2 provides average absolute bias (B) and average MSE for each estimator and each density considered for $n = 200$ and 400.

We first discuss the results in Table 1. We observe the following general regularities. First, as predicted by our asymptotic results, for all densities considered, the average absolute bias and average MSE of our estimators $\hat{f}_k(x)$ for $k = 2, 4, 8$ fall as the sample size increases. Second, as suggested in Theorem 6, increases in the values of k reduce average absolute bias and MSE, but this is not verified for all experiments. Specifically, when the sample size is small ($n = 200$) bias does not fall with k for some of the densities that are more difficult to estimate, i.e. f_3 and f_5. Reductions in average MSE due to increases in k are much less pronounced. Third, density functions with larger curvature (in increasing order of curvature f_1, f_2, f_3, f_4 and f_5) are more difficult to estimate both in terms of bias and MSE for all estimators considered. Our proposed estimators ($\hat{f}_2$, $\hat{f}_4$, $\hat{f}_8$) and the local quadratic estimator ($\hat{f}_{\mathrm{LS}}$) outperform the Rosenblatt–Parzen estimator both in terms of bias and MSE. For $k = 2$, the case where the smallest bias reductions are attained, bias can be reduced by as much as 20% relative to the Rosenblatt–Parzen estimator. Additionally, the magnitude of bias reduction produced by our estimator increases with sample size. We observe that $\hat{f}_2$, the estimator we propose that is more directly comparable with the local quadratic estimator, and $\hat{f}_{\mathrm{LS}}$ perform very similarly both in terms of bias and MSE. In summary, all of the asymptotic characterisations provided in Section 3 seem to accurately predict the behaviour of our estimators in reasonably small sample sizes.

In Table 2, we observe that the MSE of all estimators across all densities increases when the bandwidth is selected by cross validation for $n = 200$ and 400. This is not surprising, as additional noise is introduced in computing $\hat{f}_{\mathrm{R}}$, $\hat{f}_{\mathrm{LS}}$ and $\hat{f}_2$. Interestingly, there is not a significant change in bias between the results in Tables 1 and 2 for $n = 200$ or 400. As in Table 1 $\hat{f}_2$ and $\hat{f}_{\mathrm{LS}}$ outperform $\hat{f}_{\mathrm{R}}$ in both bias and MSE. It is worth noting that with estimated bandwidths $\hat{f}_2$ seems to outperform

Table 2. Three estimators with cross validation bandwidth (h^{CV}). Average bias ($\times 10^3$) (B), mean squared error ($\times 10^3$) (MSE).

	$f_1(x)$		$f_2(x)$		$f_3(x)$		$f_4(x)$		$f_5(x)$	
Estimators	B	MSE	B	MSE	B	MSE	B	MSE	B	MSE
$n = 200$										
$\hat{f}_{\mathrm{R}}$	6.580	0.484	7.500	0.613	8.782	0.839	9.148	0.706	14.159	0.326
$\hat{f}_{LS}$	5.214	0.693	6.188	1.308	6.918	0.977	8.932	1.861	15.276	0.360
$\hat{f}_2$	5.356	0.406	6.978	0.579	7.040	0.742	8.329	0.709	13.616	0.284
$n = 400$										
$\hat{f}_{\mathrm{R}}$	4.788	0.272	5.867	0.348	6.437	0.486	6.914	2.049	8.546	0.128
$\hat{f}_{LS}$	3.810	0.325	4.465	0.454	4.661	0.574	5.238	0.518	8.788	0.134
$\hat{f}_2$	3.682	0.215	4.724	0.321	4.955	0.443	5.997	0.410	7.870	0.102

$\hat{f}_{\text{LS}}$ in terms of MSE for all densities and for both $n = 200$ and 400. However, in terms of bias, the estimators continue to perform rather similarly, with the exception of the density f_5, where our estimator outperforms $\hat{f}_{\text{LS}}$. This might be due to the fact that f_5 satisfies an order 2 Lipschitz condition but does not have a continuous second derivative. We note that the bias of $\hat{f}_2$ was smaller relative to that $\hat{f}_{\text{LS}}$ in the case for f_5 in Table 1, but the difference was of smaller magnitude.

4.2. *Example*

We apply our estimator with $k = 3$ based on a Gaussian seed kernel and a Rosenblatt–Parzen estimator constructed with an order 6 kernel given by $W(u) = \frac{1}{8}(15 - 10u^2 + u^4)K(u)$ to a sample of 600 realisations from a Dickey–Fuller statistic.[4] Bandwidths for both density estimators are obtained via cross-validation and the estimated densities evaluated at the sample points are shown in Figure 1. The figure shows that our estimator is everywhere positive but the higher order

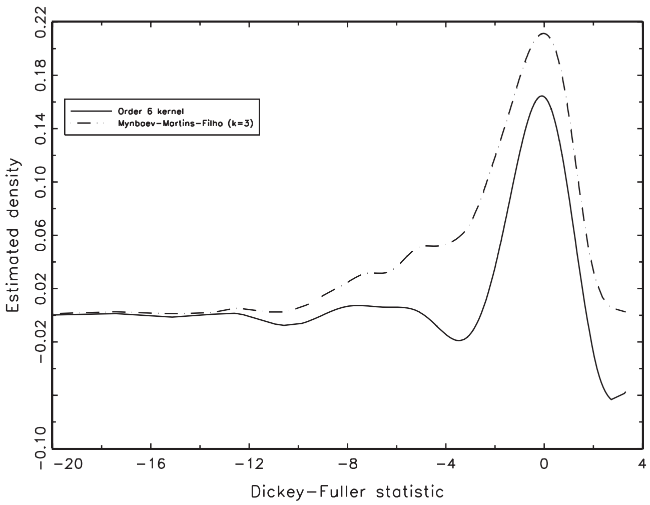

Figure 1. Estimated Dickey–Fuller density using order 6 kernel and $\hat{f}_3$.

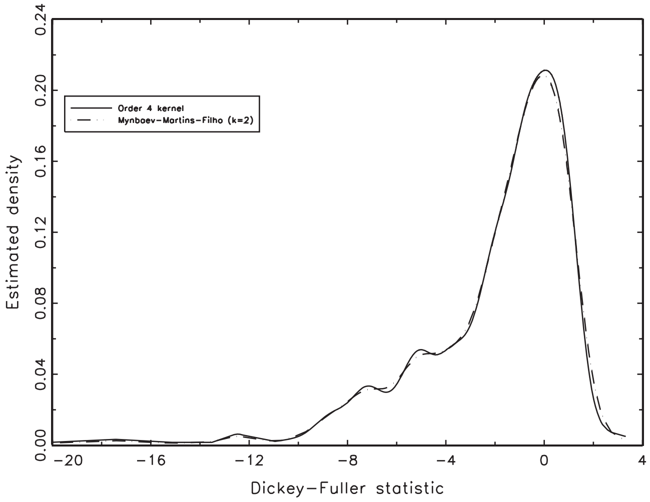

Figure 2. Estimated Dickey–Fuller density using order 4 kernel and $\hat{f}_2$.

kernel estimator is negative at a number of points in which it is evaluated. It is important to note that when the same sample of Dickey–Fuller statistics is treated with $\hat{f}_2$ $(k = 2)$ and $\hat{f}_{\mathrm{LS}}$ (order four kernel) the estimated densities are rather similar and $\hat{f}_{\mathrm{LS}}$ is everywhere positive (Figure 2).

5. Summary

In this paper, we attain reduced bias for nonparametric kernel density estimation by defining a new kernel-based estimator that explores the theory of finite differences. The main characteristic of the proposed estimator is that bias reduction may be achieved relative to the classical Rosenblatt–Parzen estimator without the disadvantage of potential negativity (depending on the seed kernel) of the estimated density – a deficiency that results from using higher order kernels to attain bias reduction. Contrary to other popular approaches for bias reduction, e.g. Jones et al. (1995) and DiMarzio and Taylor (2004) we provide a full asymptotic characterisation of our estimator. A small Monte Carlo study reveals that our estimator performs well relative to the Rosenblatt–Parzen estimator and the promised bias reduction is obtained in fairly small samples. Future work should provide seed kernels K that assure nonnegativity of M_k and are different from the Cauchy kernel.

Acknowledgements

We thank two anonymous referees for their comments. The remaining errors are the authors’ responsibility.

Notes

1. We have several examples and graphical illustrations for which $M_k > 0$ with the Cauchy seed, but we have been unable to establish this fact analytically.
2. Results for samples of size $n = 600$ are not reported but are available upon request from the authors.
3. As expected from asymptotic theory, when $n = 600$ bias and MSE for all estimators across all densities are reduced.
4. See Fuller (1976), Dickey and Fuller (1979) and Pagan and Ullah (1999).

References

Besov, O., Il’in, V., and Nikol’skiĭ, S. (1978), *Integral Representations of Functions and Imbedding Theorems*, New York: Wiley.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford: Oxford University Press.

Dickey, D.A., and Fuller, W. (1979), ‘Distribution of the Estimators for Autoregressive Time Series with a Unit Root’, *Journal of the American Statistical Association*, 74, 427–431.

DiMarzio, M., and Taylor, C.C. (2004), ‘Boosting Kernel Density Estimates: A Bias Reduction Technique?’, *Biometrika*, 91, 226–233.

Fan, J., and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer.

Fuller, W. (1976), *Introduction to Statistical Time Series*, New York: Wiley.

Granovsky, B., and Müller, H.G. (1991), Optimizing Kernel Methods: A Unifying Variational Principle, *International Statistical Review*, 59, 373–388.

Jones, M.C., and Foster, P.J. (1993), ‘Generalized Jackknifing and Higher Order Kernels’, *Journal of Nonparametric Statistics*, 3, 81–94.

Jones, M.C., Linton, O., and Nielsen, J.P. (1995), ‘A Simple Bias Reduction Method for Density Estimation’, *Biometrika*, 82, 327–338.

Jones, M.C., and Signorini, D.F. (1997), ‘A comparison of Higher-Order Bias Kernel Density Estimators’, *Journal of the American Statistical Association*, 92, 1063–1073.

Lejeune, M., and Sarda, P. (1992), ‘Smooth Estimators of Distribution and Density Functions’, *Computational Statistics & Data Analysis*, 14, 457–471.

Marron, J.S., and Wand, M.P. (1992), ‘Exact Mean Integrated Squared Error’, *Annals of Statistics*, 20, 712–736.

Pagan, A., and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge: Cambridge University Press.

Parzen, E. (1962), ‘On Estimation of a Probability Density and Mode’, *Annals of Mathematical Statistics*, 33, 1065–1076.