Journal of Multivariate Analysis **68**, 1–25 (1999)
Article ID jmva.1998.1774, available online at http://www.idealibrary.com on IDEAL®

Multivariate Density Estimation with General Flat-Top Kernels of Infinite Order*

Dimitris N. Politis

University of California at San Diego

and

Joseph P. Romano

Stanford University

Received September 20, 1995; revised March 12, 1998

The problem of nonparametric estimation of a multivariate density function is addressed. In particular, a general class of estimators with favorable asymptotic performance (bias, variance, rate of convergence) is proposed. The proposed estimators are characterized by the flatness near the origin of the Fourier transform of the kernel and are actually shown to be exactly $\sqrt{N}$-consistent provided the density is sufficiently smooth. © 1999 Academic Press

AMS 1991 subject classifications: primary: 62G07; secondary: 62H12.

Key words and phrases: bias reduction; Fourier transform; kernel; mean squared error; nonparametric density estimation; rate of convergence; smoothing.

1. INTRODUCTION

Suppose $X_1, ..., X_N$ are independent,[1] identically distributed random vectors taking values in R^d, and possessing an absolutely continuous distribution function F with corresponding probability density function f. The density f is assumed to be bounded, continuous, and smooth to some extent that will be quantified later; f is otherwise unknown and should be estimated using the data. In particular, it will be assumed that the characteristic function $\phi(s) = \int_{R^d} e^{i(s \cdot x)} f(x)\, dx$ tends to zero sufficiently fast as $\|s\|_p \to \infty$; here $s = (s_1, ..., s_d)$, $x = (x_1, ..., x_d) \in R^d$, $(s \cdot x) = \sum_k s_k x_k$ is the

* This research was partially supported by NSF grants DMS 94-04329, DMS 94-03826, and DMS 97-03964.

[1] The assumption of independence is not crucial here. The arguments presented in the paper apply equally well if the observations are stationary and weakly dependent, where weak dependence can be quantified through the use of mixing coefficients; see, for example, Györfi *et al.* (1989).

0047-259X/99 $30.00

inner product between s and x, and $\|\cdot\|_p$ is the l_p norm, i.e., $\|s\|_p = (\sum_k |s_k|^p)^1/p$, if $1 \leqslant p \leqslant \infty$, and $\|s\|_\infty = \max_k |s_k|$.

The nonparametric kernel smoothed estimator of $f(x)$, for some $x \in R^d$, is given by (cf., for example, Rosenblatt (1991) or Scott (1992))

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \Lambda(x - X_i) = \frac{1}{(2\pi)^d} \int_{R^d} \lambda(s)\, \phi_N(s)\, e^{-i(s \cdot x)}\, ds, \tag{1}$$

where $\Lambda(\cdot)$ is the smoothing kernel satisfying[2] $\int \Lambda(x)\, dx = 1$, $\phi_N(s) = 1/N \sum_{k=1}^{N} e^{i(s \cdot X_k)}$ is the sample characteristic function, and $\lambda(s) = \int \Lambda(x)\, e^{i(s \cdot x)}\, dx$ is the Fourier transform of the kernel. In general, $\Lambda(\cdot)$ and $\lambda(\cdot)$ both depend on a positive "bandwidth" parameter h; in particular, it will be assumed that $\Lambda(x) = h^{-d} \Omega(x/h)$, and $\lambda(s) = \omega(hs)$, where $\Omega(\cdot)$ and $\omega(\cdot)$ are some fixed (not depending on h) bounded functions, satisfying $\omega(s) = \int \Omega(x)\, e^{i(s \cdot x)}\, dx$; the bandwidth h will in general depend on N but it will not be explicitly denoted.

It is well known (cf. Rosenblatt (1991, p. 7)) that in this case

$$E\hat{f}(x) = \int \Omega(v)\, f(x - hv)\, dv, \tag{2}$$

and

$$Var(\hat{f}(x)) = \frac{1}{h^d N} \left[\int \Omega^2(v)\, f(x - hv)\, dv - h^d \left(\int \Omega(v)\, f(x - hv)\, dv \right)^2 \right]. \tag{3}$$

If f is continuous at x, and $f(x) > 0$, and if $h \to 0$, as $N \to \infty$, but with $h^d N \to \infty$, equation (3) becomes

$$Var(\hat{f}(x)) = \frac{1}{h^d N} f(x) \int \Omega^2(x)\, dx + O(1/N). \tag{4}$$

If the bandwidth h is a fixed constant as $N \to \infty$, then it is immediate from (3) that

$$Var(\hat{f}(x)) = \frac{1}{N} C_{f, \Omega}(x, h), \tag{5}$$

where $C_{f, \Omega}(x, h)$ is a bounded function depending on f and Ω.

If Ω has finite moments up to qth order, and moments of order up to $q - 1$ equal to zero, then q is called the "order" of the kernel Ω. If the

[2] In case it is not otherwise noted, integrals will be over the whole of R^d.

density f has r bounded continuous derivatives,[3] it then follows (cf. for example, Rosenblatt (1991)) that

$$Bias(\hat{f}(x)) = E\hat{f}(x) - f(x) = c_{f,\Omega}(x)\, h^k + o(h^k), \tag{6}$$

where $k = \min(q, r)$, and $c_f(x)$ is a bounded function depending on Ω, on f, and on f's derivatives. This idea of choosing a kernel of order q in order to get the $Bias(\hat{f}(x))$ to be $O(h^k)$ dates back to Parzen (1962) and Bartlett (1963); see also Cacoullos (1966) for the multivariate case. Some more recent references on "higher-order" kernels include the following: Devroye (1987), Gasser *et al.* (1985), Granovsky and Müller (1991), Jones (1995), Jones and Foster (1993), Marron (1994), Marron and Wand (1992), Müller (1988), Nadaraya (1989), Silverman(1986), and Scott (1992).

Note that the asymptotic order of the bias is limited by the order of the kernel if the true density is very smooth, i.e., if r is large. To avoid this limitation, one can define a "superkernel" as a kernel whose order can be any positive integer; Devroye (1992) contains a detailed analysis of superkernels in the univariate case. Thus, if f has r bounded continuous derivatives, a superkernel will result in an estimator with bias of order $O(h^r)$, no matter how large r may be; so, we might say that a superkernel is a kernel with "infinite order".

Note that the $O(h^r)$ order for the bias, and the corresponding rate of $O(N^{-2r/(2r+d)})$ for the Mean Squared Error of $\hat{f}$, have been shown to be optimal, i.e., they are the smallest achievable with kernel estimators if the density f is constrained to have exactly r bounded and continuous derivatives. If the characteristic function $\phi(s)$ decreases exponentially fast with increasing $\|s\|$, or if $\phi(s)$ vanishes outside a compact set, then the smallest achievable orders for the Mean Squared Error of $\hat{f}$ are $O(\log N/N)$ and $O(1/N)$ respectively. These important lower bounds on the accuracy of kernel estimators are due to Watson and Leadbetter (1963); see also Wahba (1975).

However, it might be more appropriate to say that a kernel has "infinite order" if it results in an estimator with bias of order $O(h^r)$ no matter how large r may be *regardless of whether the kernel has finite moments*. It seems that the finite-moment assumption for Ω is just a technical one, and that existence of the Lebesgue integrals used to calculate the moments is *not* necessarily required in order that a kernel has favorable bias performance; rather, it seems that if the integrals defining the moments of Ω have a Cauchy principal value of zero then the favorable bias performance follows,

[3] Existence and boundedness of derivatives up to order r includes existence and boundedness of mixed derivatives of total order r; cf. Rosenblatt (1991, p. 8).

and this is in turn ensured by setting ω to be constant over an open neighborhood of the origin.

A preliminary report on a specific type of such infinite order kernel in the univariate case (that corresponds to an ω of "trapezoidal" shape) was given in Politis and Romano (1993); in the present paper a general family of multivariate kernels of infinite order is presented, and the favorable properties of the resulting estimators are quantified. As elaborated above, the proposed kernels are characterized by the fact that their Fourier transforms are "flat" over an open neighborhood of the origin. In particular, for the class of ultra-smooth densities whose characteristic functions are supported on a compact set, the proposed kernel estimators are shown to actually be $\sqrt{N}$-consistent.

The organization of the remainder of the paper is as follows: Section 2 contains the necessary definitions and statements of our main results on the performance of the proposed kernel estimators; Section 3 contains some practical comments and simulation results; all technical proofs are placed in Section 4.

2. A GENERAL FAMILY OF FLAT-TOP SMOOTHING KERNELS OF INFINITE ORDER

Let c and p be constants satisfying $1 \leqslant c \leqslant \infty$, $1 \leqslant p \leqslant \infty$, and define

$$\lambda_c(s) = \begin{cases} 1 & \text{if} \quad \|s\|_p \leqslant 1/h \\ g_\lambda(s, h) & \text{if} \quad 1/h < \|s\|_p \leqslant c/h \\ 0 & \text{if} \quad \|s\|_p > c/h. \end{cases} \tag{7}$$

Here $g_\lambda(s, h)$ is some properly chosen continuous, real-valued function satisfying $g_\lambda(s, h) = g_\lambda(-s, h)$, $g_\lambda(s, 1) = g_\lambda(s/h, h)$, and $|g_\lambda(s, h)| \leqslant 1$, for any s, with $g_\lambda(s, h) = 1$, if $\|s\|_p = 1/h$, and $g_\lambda(s, h) = 0$, if $\|s\|_p = c/h$. We will also assume that $\int_S |g_\lambda(s, h)|^2 \, ds < \infty$, where $S = \{s\colon 1/h < \|s\|_p \leqslant c/h\}$; the latter assumption guarantees that $\int \lambda_c^2(s) \, ds < \infty$ which will be necessary in order to have kernel estimators with finite variance (see our Remark 2 in what follows).

If $c = 1$, the drop from the value 1 to the value 0 is done in a discontinuous fashion, and no function g_λ is needed. On the other hand, the case $c = \infty$ covers the situation where a compact support for λ_c is not desired. In essence, g_λ interpolates between the value 1 for $\|s\|_p \leqslant c/h$, and the value 0 for $\|s\|_p > 1/h$. Perhaps the most "natural" way to do the interpolation would be to do it in a linear fashion provided, of course, that $c < \infty$; more details on the subject of choosing the value of c and the shape of the function g_λ can be found in Section 3.3.

Having picked a g_λ function, we now define a family of kernels $\{\Lambda_c(\cdot), c \in [1, \infty]\}$ by

$$\Lambda_c(x) = \frac{1}{(2\pi)^d} \int \lambda_c(s)\, e^{-i(s \cdot x)}\, ds, \tag{8}$$

i.e., by the (inverse) Fourier transform of $\lambda_c(s)$; note that the corresponding $\Omega(\cdot)$ and $\omega(\cdot)$ functions can be obtained by setting $h = 1$ in the definitions (7) and (8), and that Λ_c is real-valued because of the symmetry of λ_c, i.e., $\lambda_c(s) = \lambda_c(-s)$.

The proposed kernel smoothed estimators of f are given by

$$\hat{f}_c(x) = \frac{1}{N} \sum_{i=1}^{N} \Lambda_c(x - X_i) = \frac{1}{(2\pi)^d} \int \lambda_c(s)\, \phi_N(s)\, e^{-i(s \cdot x)}\, ds, \tag{9}$$

for some choice of $c \in [1, \infty]$. The estimator $\hat{f}_c$ can be computed using either of the two expressions appearing in (9). To compute $\hat{f}_c$ using the standard expression involving the convolution of Λ_c with the empirical distribution, the form of Λ_c must be calculated. In general, a closed-form expression for Λ_c might not be available, but Λ_c, can be calculated numerically over a grid of points (call it G), and consequently $\hat{f}_c(x)$ will be computed only for $x \in G$; see Section 3.1 for more details on computational aspects.

Note that by equations (4) and (5) and since, by construction, $\int \lambda_c^2(s)\, ds < \infty$, it is immediate that $Var(\hat{f}_c(x)) = O(1/h^d N)$, as $N \to \infty$, whether h is a fixed constant, or if $h \to 0$ but with $h^d N \to \infty$. Therefore, the order of magnitude of the Mean Squared Error (MSE) of $\hat{f}_c$ will hinge on the order of magnitude of the bias. We will now proceed to investigate the MSE performance of $\hat{f}_c$ under a variety of different smoothness conditions on f; for this purpose, we formulate three different conditions based on the rate of decay of the characteristic function ϕ that are in the same spirit as the conditions in Watson and Leadbetter (1963).

Condition C_1. For some $p \in [1, \infty]$, there is an $r > 0$, such that $\int \|s\|_p^r\, |\phi(s)| < \infty$

Condition C_2. For some $p \in [1, \infty]$, there are positive constants B and K such that $|\phi(s)| \leqslant Be^{-K\|s\|_p}$.

Condition C_3. For some $p \in [1, \infty]$, there is a positive constant B such that $|\phi(s)| = 0$, if $\|s\|_p \geqslant B$.

Conditions C_1 to C_3 can be interpreted as different conditions on the smoothness of the density $f(x)$; cf. Katznelson (1968), Butzer and Nessel

(1971), Stein and Weiss (1971), and the references therein. Note that they are given in increasing order of strength, i.e., if Condition C_2 holds, then Condition C_1 holds as well, and if Condition C_3 holds, then Conditions C_1 and C_2 hold as well. Also note that if Condition C_1 holds, then f must necessarily have $[r]$ bounded, continuous derivatives, where $[\cdot]$ is the positive part; cf. Katznelson (1968, p. 123). Obviously, if Condition C_2 holds, then f has bounded, continuous derivatives of *any* order; although this very high degree of smoothness for f seems like a very strong assumption, it turns out that "in many applications in the physical and biomedical sciences it can be safely assumed that the function has this high degree of smoothness" (cf. Müller (1988, p. 73)).

The following sequence of theorems quantifies the performance of the proposed family of flat-top estimators. Note that the constant p to be used in connection with the kernel Λ_c is the *same* p that appears in Conditions C_1 to C_3 (as invoked by the theorems).

THEOREM 1. *Assume that $h \to 0$, as $N \to \infty$, but with $h^d N \to \infty$; under Condition C_1, it follows that*

$$\sup_{x \in R^d} |Bias(\hat{f}_c(x))| = o(h^r).$$

Now let x be some point in R^d such that $f(x) > 0$; then by letting $h \sim AN^{-1/(2r+d)}$, for some constant $A > 0$, the asymptotic order of the Mean Squared Error of $\hat{f}_c$ is given by $MSE(\hat{f}_c(x)) = O(N^{-2r/(2r+d)})$.

Remark 1. That the $Bias(\hat{f}_c(x))$ turns out to be $o(h^r)$, rather than $O(h^r)$, should not be surprising as it was mentioned that Condition C_1 is stronger than assuming f has r bounded and continuous derivatives; however, it is not much stronger. For example, in the case $d = 1$, Condition C_1 is seen to be satisfied if it is assumed that f has r absolutely integrable derivatives, and the the rth derivative $f^{(r)}$ satisfies a uniform Lipschitz condition of order $\alpha > 1/2$; cf. Katznelson (1968, p. 32).

Remark 2. The asymptotic variance of $\hat{f}(x)$ can be calculated from equation (4). However, to compute $\int \Omega^2(x)\,dx$, it is easier to use the isometric properties of the Fourier transform, i.e., Parseval's theorem, and compute $(2\pi)^{-d} \int \omega^2(s)\,ds$ instead, especially since, if $c < \infty$, ω has compact support.

THEOREM 2. *Assume that $h \to 0$, as $N \to \infty$, but with $h^d N \to \infty$; under Condition C_2, it follows that* $\sup_{x \in R^d} |Bias(\hat{f}_c(x))| = O(h^{1-d} e^{-K/h})$. *If we let*

$h \sim A/\log N$, as $N \to \infty$, where A is a constant such that $A < 2K$, it follows that

$$\sup_{x \in R^d} |Bias(\hat{f}_c(x))| = O\left(\frac{(\log N)^{d-1}}{N^{K/A}}\right) = o\left(\frac{1}{\sqrt{N}}\right).$$

Now let x be some point in R^d such that $f(x) > 0$; the choice $h \sim A/\log N$ implies that $MSE(\hat{f}_c(x)) = O((\log N)^d/N)$.

THEOREM 3. *Assume Condition C_3 and that, as $N \to \infty$, h is some constant small enough such that $h \leqslant B^{-1}$; it follows that*

$$\sup_{x \in R^d} |Bias(\hat{f}_c(x))| = 0.$$

Now let x be some point in R^d such that $f(x) > 0$; it follows that $MSE(\hat{f}_c(x)) = O(1/N)$.

Remark 3. The special case where $c = 1$, i.e., when the drop of λ_c from the value 1 to the value 0 is done discontinuously, has been considered by many authors in the literature, e.g., Parzen (1962). Thus, considering the estimator $\hat{f}_1$, Davis (1977) proved analogs of our Theorems 1 to 3 for $d = 1$, while Ibragimov and Hasminksii (1982) have proved an analog of our Theorem 3 in the general d case. Nevertheless, the choice $c = 1$ is *not* recommendable in practice; our next Section addresses this issue, as well as other practical concerns.

Remark 4. By the formal analogy between probability spectral density estimation (see, e.g., Rosenblatt (1991)) it should not be surprising that flat-top kernels might be applicable in a context of nonparametric spectral density estimation. In Politis and Romano (1996), kernels belonging to a subset of the family of flat-top kernels are employed for the purpose of spectral density estimation using data consisting of a realization of a homogeneous random field.

Remark 5. A rather surprising observation is that smoothing with flat-top kernels does not seem to be plagued by the "curse of dimensionality" in case the underlying density is ultra-smooth, possessing derivatives of all orders. For example, in Theorem 2 under Condition C_2, the MSE of estimation achieved by flat-top kernel smoothing is of order $O(\log^d N/N)$, i.e., depending only slightly on the dimension d, while in Theorem 3 under Condition C_3, the MSE of estimation becomes exactly $O(1/N)$, i.e., not depending on d *at all*.

3. DISCUSSION AND PRACTICAL COMMENTS

3.1. *Computational Aspects and Remarks*

Assuming $1 < c < \infty$, and choosing $g_\lambda(s, h)$ to be linear in its first argument, actually results into a compact expression for λ_c, namely,

$$\lambda_c^{LIN}(s) = \frac{c}{c-1}\left(1 - \frac{h}{c}\,\|s\|_p\right)^+ - \frac{1}{c-1}\,(1 - h\,\|s\|_p)^+, \tag{10}$$

where $(x)^+ = \max(x, 0)$ is the positive part function. A closed-form expression for $\Lambda_c^{LIN}(x) = (2\pi)^{-d} \int \lambda_c^{LIN}(s)\, e^{-i(s \cdot x)}\, ds$ in the special case $d = 1$ is given by

$$\Lambda_c^{LIN}(x) = \begin{cases} \dfrac{h}{2\pi}\, \dfrac{\sin^2(\pi c x/h) - \sin^2(\pi x/h)}{\pi^2 x^2 (c-1)} & \text{if} \quad c > 1 \\ \dfrac{1}{2\pi}\, \dfrac{\sin(2\pi x/h)}{\pi x} & \text{if} \quad c = 1; \end{cases} \tag{11}$$

it is apparent that in the case $c > 1$, Λ_c^{LIN} is just a linear combination of Fejér kernels, whereas if $c = 1$, Λ_c^{LIN} reduces to the Dirichlet kernel. In the general case where $d > 1$, Λ_c^{LIN} depends on p and may be difficult to evaluate analytically; see Fig. 1 and 2 for graphs of λ_c^{LIN} and Λ_c^{LIN} for $d = 2$, $p = 2$, $c = 2$, and $h = 0.067$, where Λ_c^{LIN} has been computed numerically using a two-dimensional discrete Fourier transform.

In the Euclidean norm case ($p = 2$), computations can be aided by the observation that, since $\lambda_c(s)$ depends on s only through $\|s\|_2$, its functional form is rotation-invariant; consequently, $\Lambda_c(x)$ depends on x only through $\|x\|_2$, and the functional form of Λ_c is rotation-invariant as well. Hence, to evaluate $\Lambda_c(x)$ for any $x \in R^d$, it suffices to evaluate it for $x = (x_1, 0, 0, ..., 0)$, with x_1 spanning R, and then rotate the resulting graph. But $\Lambda_c(x_1, 0, 0, ..., 0)$ can be obtained by a *univariate* (inverse) Fourier transform as $\Lambda_c(x_1, 0, 0, ..., 0) = (2\pi)^{-l} \int \mu(s_1)\, e^{-is_1 x_1}\, ds_1$, where

$$\mu(s_1) = (2\pi)^{-d+1} \iint \cdots \int \lambda_c(s_1, s_2, ..., s_d)\, ds_2\, ds_3 \cdots ds_d$$

is the "marginal" of the function $\lambda_c(s) = \lambda_c(s_1, s_2, ..., s_d)$.

It should be pointed out that the computation of $\hat{f}_c$ can actually be accomplished faster by using the rightmost expression of (9), i.e., multiplication ("tapering") of the empirical characteristic function by λ_c, followed by a discrete Fourier transform; cf. for example, Silverman (1986, p. 61). In that sense, exact knowledge of the form of Λ_c is not needed; see also our Remark 2 after Theorem 1. However, for illustration purposes, we

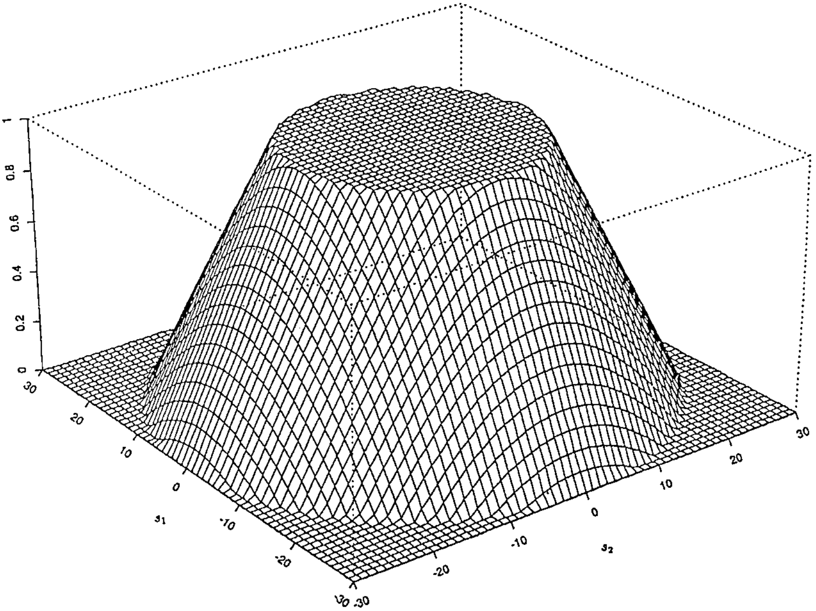

FIG. 1. The Fourier transform of Λ_c^{LIN}, i.e., $\lambda_c^{LIN}(s)$, as a function of $s=(s_1, s_2)$, for $d=2$, $p=2$, $c=2$, and $h=0.067$.

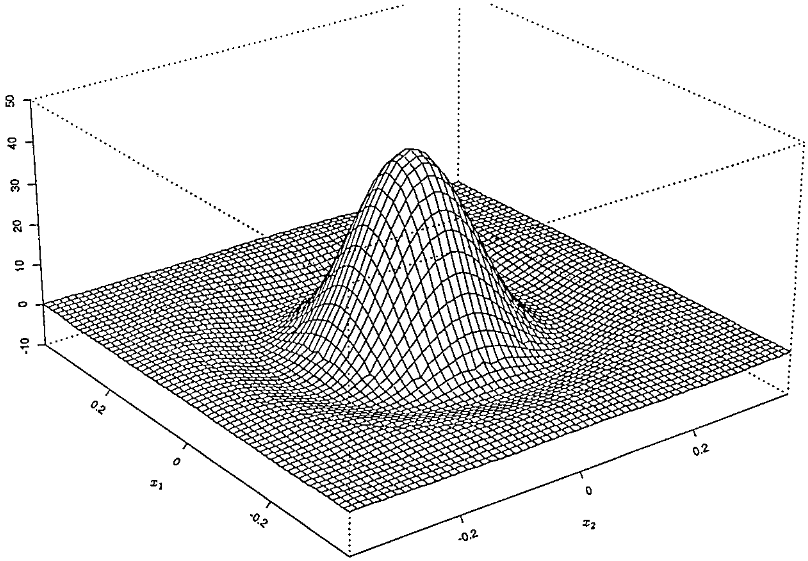

FIG. 2. The kernel $\Lambda_c^{LIN}(x)$, as a function of $x=(x_1, x_2)$, for $d=2$, $p=2$, $c=2$, and $h=0.067$.

now construct an explicit (Λ_c, λ_c) pair by taking *products* of the univariate kernel given in (11); see Müller (1988) or Scott (1992) for more details on the product method of constructing multivariate kernels. So let d be any positive integer, $1 < c < \infty$, and $h > 0$, and define

$$\Lambda_c^{PROD}(x) = \left(\frac{h}{2\pi}\right)^d \prod_{j=1}^{d} \frac{\sin^2(\pi c x_j/h) - \sin^2(\pi x_j/h)}{\pi^2 x_j^2 (c-1)}, \tag{12}$$

and

$$\lambda_c^{PROD}(s) = \left(\frac{1}{c-1}\right)^d \prod_{j=1}^{d} ((c - h\,|s_j|)^+ - (1 - h\,|s_j|)^+); \tag{13}$$

it is easy to check that Λ_c^{PROD} and λ_c^{PROD} are related to each other by a Fourier transform, and that

$$\lambda_c^{PROD}(s) = \begin{cases} 1 & \text{if} \quad \|s\|_\infty \leqslant 1/h \\ 0 & \text{if} \quad \|s\|_\infty > c/h. \end{cases}$$

The functions λ_c^{PROD} and Λ_c^{PROD} are plotted in Fig. 3 and 4 in the case $d = 2$, $p = \infty$, $c = 2$, and $h = 0.067$.

It is well-known in the literature (see, for example, Müller (1988) or Scott (1992)) that kernel density estimators corresponding to kernels of order bigger than two are not necessarily nonnegative functions; it goes without saying that the same applies for our estimators $\hat{f}_c$ that are obtained using kernels of "infinite order". To appreciate why, observe that in Fig. 2 and 4 the kernels Λ_2^{LIN} and Λ_2^{PROD} exhibit negative "sidelobes" beside the main prominent "lobe" around the origin which is positive.

Nevertheless, the nonnegativity is not a serious issue as there is a natural fix-up, namely using the modified estimator[4] $\hat{f}_c^+(x) = \max(\hat{f}_c(x), 0)$; see also Gajek (1986) and Hall and Murison (1992). Note that the estimator $\hat{f}_c^+(x)$ is not only nonnegative, but is more accurate as well, in the sense that $MSE(\hat{f}_c^+(x)) \leqslant MSE(\hat{f}_c(x))$, for all x; this fact follows from the obvious inequality $|\hat{f}_c^+(x) - f(x)| \leqslant |\hat{f}_c(x) - f(x)|$. In addition, if $f(x) > 0$, an application of Chebychev's inequality shows that $Prob\{\hat{f}_c(x) = \hat{f}_c^+(x)\} \to 1$ under the assumptions of any of our Theorems 1 to 3; on the other hand, if $f(x) = 0$, then the large-sample distribution of either $\sqrt{h^d N}\,\hat{f}_c^+(x)$, or $\sqrt{h^d N}\,\hat{f}_c(x)$, degenerates to a point mass at zero.

[4] Strictly speaking, the modified estimator should read $\hat{f}_c^+(x) = \max(\hat{f}_c(x), 0)/\int \max(\hat{f}_c(y), 0)\,dy$, so that the estimator integrates to one; nevertheless, this renormalization is an asymptotically negligible adjustment because under appropriate conditions $\int \max(\hat{f}_c(y), 0)\,dy \to 1$ in probability (cf. Nadaraya (1989)).

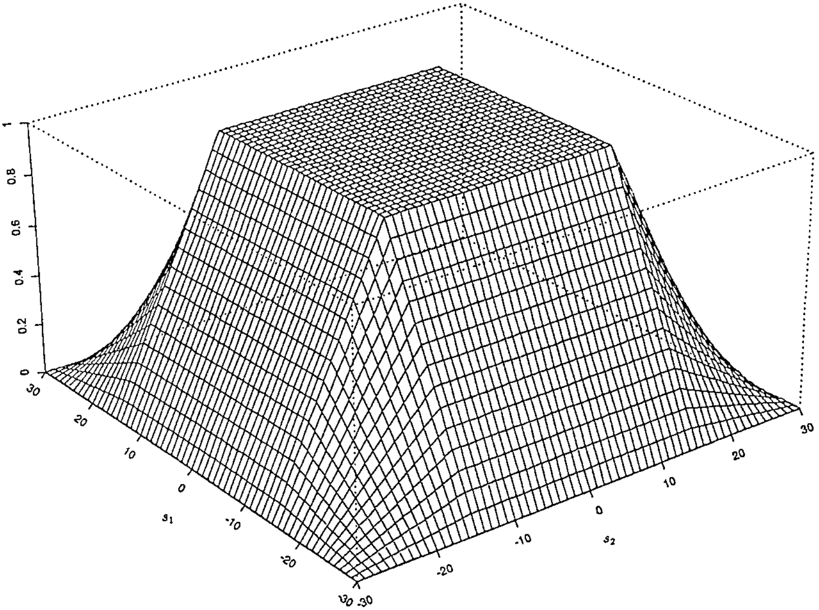

FIG. 3. The Fourier transform of Λ_c^{PROD}, i.e., $\lambda_c^{PROD}(s)$, as a function of $s=(s_1, s_2)$, for $d=2$, $p=\infty$, $c=2$, and $h=0.067$.

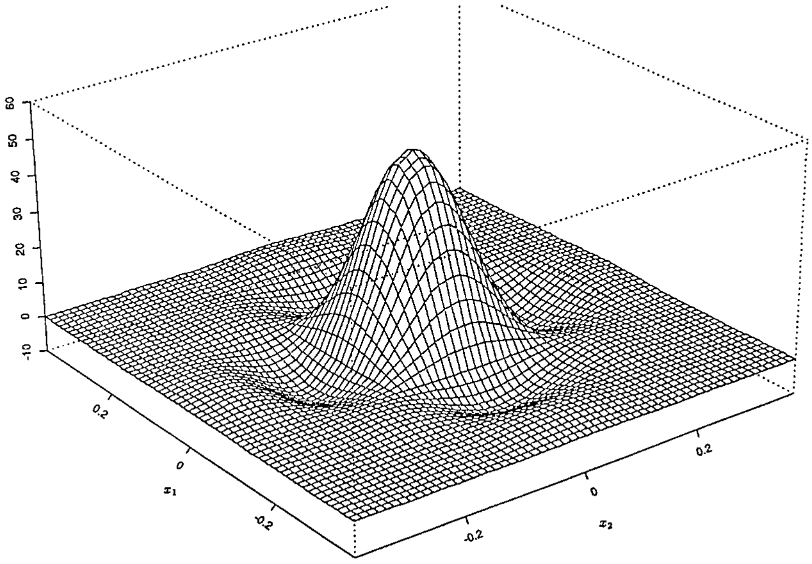

FIG. 4. The kernel $\Lambda_c^{PROD}(x)$, as a function of $x=(x_1, x_2)$, for $d=2$, $p=\infty$, $c=2$, and $h=0.067$.

3.2. *Choosing the Value of p and Transformations*

The implicit assumption in our Theorems 1 to 3 was that the value of p used in λ_c and the subsequent computation of the estimator $\hat{f}_c$ (or $\hat{f}_c^+$) was the *same* as the value of p appearing in the invoked Conditions C_1 to C_3. Note, however, that if one of Conditions C_1 to C_3 holds for some $p \in [1, \infty]$, then, by the equivalence of l_p norms for R^d, that same Condition would hold for *any* $p \in [1, \infty]$, perhaps with a change in the constants B and K. In that sense, the matching of the values of p in λ_c with that of the invoked Condition C_1, C_2, or C_3 is *not* required for the asymptotic arguments to go through, and Theorems 1 to 3 are true even without the matching.

Nevertheless, it makes good sense to have this matching occur (even approximately) as it *would* make a difference in practice. The reason it would be beneficial can be attributed to this possible change in the constants B and K that influence the proportionality constants in calculating the bias of $\hat{f}_c$. While the asymptotic order of the bias remains unchanged, the proportionality constant can be reduced by this matching of the values of p; see, for example, the proof of Theorem 2.

A practical way to ensure that this approximate matching occurs is described next. Once $|\phi_N(s)|$ is calculated, it can be plotted as a diagnostic tool, in analogy to correlogram plots in the spectral analysis of time series (cf. Priestley (1981)). Since s is in general multi-dimensional, "slices" of $|\phi_N(s)|$ can be plotted, i.e., varying only one or two of the coordinates of s at a time; alternatively, we can vary s subject to a linear constraint of the type $Ms = m$, where M is a $(d-k)$ by d matrix (and k is 1 or 2), and m is a $(d-k)$ dimensional vector. By so doing, one can get a rough estimate of the different rates of decay of $|\phi_N(s)|$ along all directions, and certainly along the d principal directions. Note that the rates of decay of $|\phi_N(s)|$ can be influenced by scaling the X data. Thus, a first step is to employ a diagonal transformation D to come up with transformed data $Y_i = DX_i$, $i = 1, ..., N$; here $D = diag(D_1, ..., D_d)$ should be chosen such that D_j^{-1} equals an estimate of scale (say, sample standard deviation) of the jth coordinate of the X data. In conjunction with the new Y data, using $p = \infty$ seems like a reasonable choice.

Ideally however, we would want the "level" curves of $|\phi_N(s)|$ (i.e., the sets of the type $\{s\colon |\phi_N(s)| = const.\}$) to be shaped like an l_p unit ball. If the "level" curves of the sample characteristic function of the Y data are not shaped like l_p balls, another linear (not diagonal) transformation can be employed in an effort to achieve approximately equal rate of decay of the sample characteristic function in *all* directions (and not just the d principal ones); cf. Scott (1992, p. 153) and Wand and Jones (1993) for more details on use of transformations and more general bandwidth parameterizations.

Note that the value $p = 2$ can be used in conjunction with kernel estimation of the probability density of the transformed data where the sample characteristic function has equal rate of decay in all directions.

3.3. *Choosing the Value of c and the Shape of the Function* g_λ

It is quite interesting that the actual value of c and the actual shape of the function g_λ do not enter at all in our asymptotic Theorems 1–3; this observation agrees with the findings of Devroye (1992) who considered infinite-order kernels in the univariate case ($d = 1$).

Nevertheless, properly choosing c and the shape of the function g_λ will definitely have a practical impact. In terms of choosing the shape of λ_c or of ω, i.e., choosing c and g_λ, Devroye (1992, p. 2053) writes: "The recommendation is to take (our ω) rectangular with two smooth tails added on so as to make the tails of (our Ω) small. The size of these tails has to be determined from nonasymptotic considerations, perhaps via some data-based rule."

Making the tails of Ω small has a twofold advantage;[5] (a) reducing the bias of the resulting estimator by reducing the "leakage" through the many small peaks in the (typically wavy) tails of Ω, and (b) reducing the variance of the resulting estimator which is approximately proportional to $\int \Omega^2(x)\,dx$. Therefore, comparison between different kernels can be accomplished by inspecting the relative magnitude (and sign) of the "sidelobes" as compared to the main "lobe" around the origin.

In particular, the choice $c = 1$ which was considered by Davis (1977) and Ibragimov and Hasminksii (1982) is *not* recommendable in practice. To see this, consider the functions λ_1 and Λ_1 that are plotted in Figs. 5 and 6 in the case $d = 2$, $p = \infty$, and $h = 0.05$. It is apparent that the magnitude of the wavy "sidelobes" of Λ_1 is much bigger than those in either Λ_2^{LIN} or Λ_2^{PROD} (see Figs. 2 and 4). As a matter of fact, to really witness the tails of Λ_1 become negligible in magnitude, we have to look at $\Lambda_1(x)$ over a wider region of the (x_1, x_2) plane; see Fig. 7.

The reason $h = 0.05$ was used in connection with Λ_1, in Figs. 6 and 7 (as opposed to $h = 0.067$ that was used for Λ_2^{LIN} and Λ_2^{PROD} in Figs. 2 and 4) was the effort to compare kernels that yield estimators with approximately equal variance. As can be seen from the first column of Table I, with these choices of h, the variance integrals $\int \lambda_c^2(s)\,ds = h^{-d} \int \omega^2(s)\,ds$ (that equal

[5] It should be stressed however that by different choices of c and g_λ we can *not* change the asymptotic orders of bias and variance of the resulting estimators; that is why the actual shape of λ_c is immaterial in our asymptotic Theorems 1–3, as long as λ_c is flat near the origin, and has finite Euclidean norm. By choosing the value of c and the shape of the function g_λ properly, we can only influence the proportionality constants in the large-sample bias and variance of the estimators.

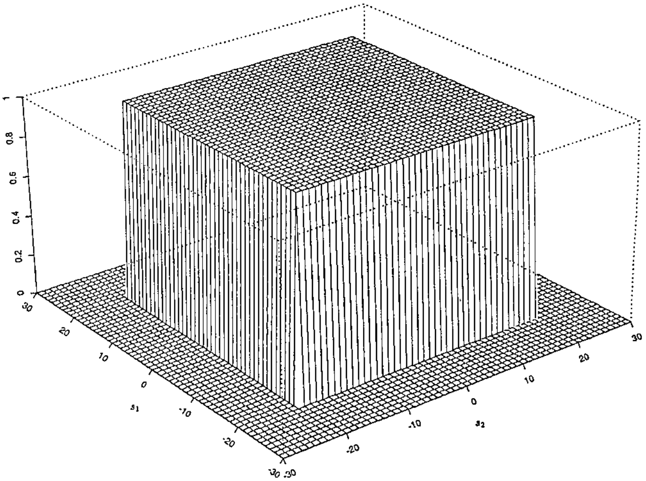

FIG. 5. The Fourier transform of Λ_1, i.e., $\lambda_1(s)$, as a function of $s=(s_1, s_2)$, for $d=2$, $p=\infty$, and $h=0.05$.

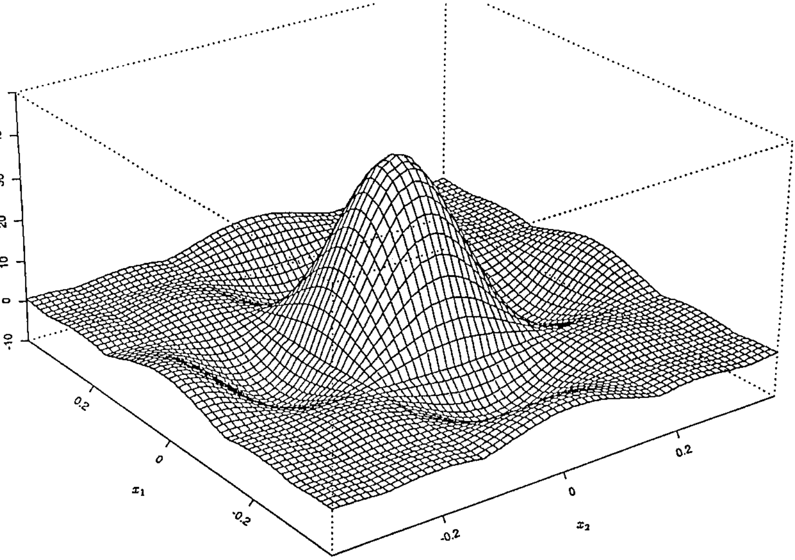

FIG. 6. The kernel $\Lambda_1(x)$, as a function of $x=(x_l, x_2)$, for $d=2$, $p=\infty$, and $h=0.05$.

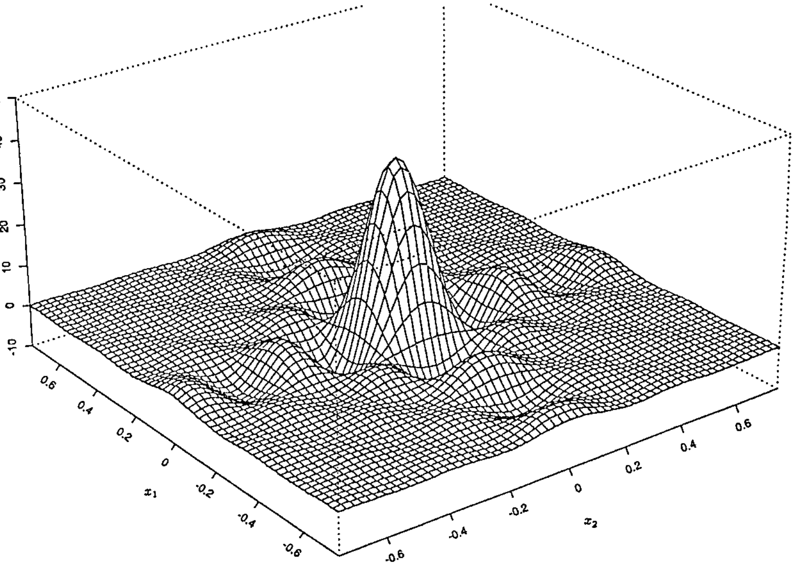

FIG. 7. Same as Fig. 6, i.e., $d=2$, $p=\infty$, and $h=0.05$, but here $\Lambda_1(x)$ is shown over a wider region of the (x_1, x_2) plane.

the asymptotic variance of $\sqrt{N}\,\hat{f}(x)/\sqrt{f(x)})$ are about equal for the three kernels. So, in other words, choosing the h bandwidths so that we achieve similar variances we empirically verify that Λ_1 will result in more biased estimators than either Λ_2^{LIN} or Λ_2^{PROD} because of the more pronounced "sidelobes". Alternatively, suppose that the *same* bandwidth was used for all three kernels. Then, as can be seen from the second column of Table 1, Λ_1 will result in an estimator with bigger variance than either Λ_2^{LIN} or Λ_2^{PROD}.

In short, $c=1$ is a bad choice. Our empirically-based recommendations at this point suggest that using $c=2$, or c in the neighborhood of 2 (say $c \in [1.5, 3]$), and using the g_λ corresponding to either Λ_c^{LIN} or Λ_c^{PROD} will

TABLE I

Entries are the Variance Integrals $\int \lambda_c^2(s)\,ds$ and the Variance Constants $\int \omega^2(s)\,ds$ for the Three Functions Shown in Fig. 1, 3, and 5

	$\int \lambda_c^2(s)\,ds$	$\int \omega^2(s)\,ds$
Figure 1:	0.360	0.0016
Figure 3:	0.445	0.0020
Figure 5:	0.467	0.0243

give good results; see also our discussion in Section 3.2 where the choices of $p=2$ and $p=\infty$ that correspond to Λ_c^{LIN} and Λ_c^{PROD} come up rather naturally. As evidenced by the variances presented in Table 1, Λ_2^{LIN} might be somewhat preferable to Λ_2^{PROD}, but it is also a bit harder to work with because it is not given in closed form. We conjecture that the "optimal" (with respect to some reasonable criterion, say exact MSE of the resulting estimators) choices of c and $g_\lambda(s, h)$ will turn out to be $c=\infty$, but with a very carefully constructed g_λ function that decays to zero fast enough as $s \to \infty$, but that is not necessarily nonnegative for all values of s; rather, $g_\lambda(s, h)$ will have small negative (and positive) "sidelobes" for s large, in much the same way as the kernel $\Omega(x)$ has to go negative for some x-regions to achieve optimality—see Devroye (1992) for more discussion. Nevertheless, this extra fine-tuning of kernel choice will not be very significant in practice—unless the sample size N is really huge, and higher-order refinements acquire importance; using either Λ_c^{LIN} or Λ_c^{PROD} (with c in the neighborhood of 2) will probably be as good for all practical purposes.

3.4. *Choosing the Bandwidth h*

Last, but not in any means least in terms of practical importance, is the choice of bandwidtd h. Müller (1988, p. 61) writes "... the behavior of kernel estimates with kernels of higher order is less sensitive towards a suboptimal choice of bandwidth." Consequently, our kernels of infinite order should also share this robustness property. Nevertheless, to take full advantage of the smoothness of the underlying true probability density using our infinite order kernels one should be prepared to use really large bandwidths if deemed necessary.

As a matter of course, our Theorems 1–3 give expressions for the optimal bandwidth (optimal with respect to minimization of the asymptotic order of the resulting MSE), i.e., $h \sim AN^{-1/(2r+d)}$, $h \sim A/\log N$, and $h = const. \leqslant 1/B$, respectively, where the constants A and B are described in Theorems 1–3. However, this is not entirely satisfactory from a practical point of view since it is assumed we know which of Conditions C_1-C_3 holds true (and we know r and B) which is not given in any real data-analytic situation. Rather, the degree of smoothness of the true probability density should also be gauged from the available data at hand; one way of doing this is looking at a plot of $|\phi_N(s)|$ vs. s as discussed in Section 3.2. Since smoothness of the probability density function is a property of the tails of $\phi(\cdot)$, the apparent decay of $\phi_N(s)$ for large s may give useful information on the smoothness of $f(\cdot)$.

Although more work is needed in order to settle the problem of optimal bandwidth choice, we now give a practical recommendation based on our Theorem 3 in conjunction with a diagnostic plot of $|\phi_N(s)|$ as discussed in Section 3.2. Suppose that the empirical plot of $|\phi_N(s)|$ reveals that $|\phi_N(s)|$

is of negligible magnitude for $\|s\|_p$ bigger than some number $\hat{B}$, and that $|\phi_N(s)|$ is nonnegligible if $\|s\|_p \leqslant \hat{B}$. Then, $\hat{B}$ can be considered as an estimate of the constant B appearing in Condition C_3, and we should be advised to choose $h = 1/\hat{B}$. Note that even if the weaker Conditions C_1 or C_2 hold instead of Condition C_3, still $|\phi(s)|$ (and therefore $|\phi_N(s)|$ as well, since $\phi_N(s) \to \phi(s)$ as $N \to \infty$) would be practically negligible for big enough $\|s\|_p$; hence, the above simple diagnostic procedure should give reasonable choices for the bandwidth h under any of our assumed smoothness Conditions C_1–C_3.

3.5. *Some Finite-Sample Numerical Results*

With the goal to empirically substantiate and further illustrate our heuristic recommendations on choosing c and h in Sections 3.3 and 3.4, a small finite-sample simulation study was conducted. In order to produce better graphs where our heuristics become more apparent, and to avoid having to look only at "slices" of $|\phi_N|$, we focused on the univariate case $d = 1$. Three different "true" densities were considered; the "skewed unimodal density" (#2 in Marron and Wand (1992, p. 717)), the "asymmetric bimodal density" (#8 in Marron and Wand (1992, p. 717)), and the heavy-tailed density of Student's t-distribution with 3 degrees of freedom. It can be easily checked that Condition C_2 holds true for each of the three densities considered.

We used a sample size of $N = 200$, as it seems that for "nice densities", i.e., very smooth densities without "sharp" prominent features, an N between 100 and 1000 would be sufficient in order for the asymptotic approximations to the MSE to have some validity; see, e.g., Fig. 9 in Marron and Wand (1992). All computations were performed using the statistical language S+ on a 486 IBM PC.

The smoothed estimators were computed using a discrete approximation to the RHS of Eq. (9). The function $g_\lambda(s, h)$ was chosen to be linear (in its first argument), so in effect the kernel (11) was used. To elaborate, $\phi_N(s)$ was computed for s taking values on a grid, i.e., $s = s_j = jG$, where $j = 1, 2, \ldots$; note that $|\phi_N(0)| = 1$ always. The gridsize constant G was taken approximately equal to $2\pi(0.14)$ for the first two densities, and $2\pi(0.07)$ for Student's t. Finally, the (inverse) Fourier transform of the product $\lambda_c(s)\, \phi_N(s)$ was computed (via an FFT—Fast Fourier Transform) yielding the kernel smoothed estimator $\hat{f}_c(x)$.

Figure 8 concerns the "skewed unimodal density". Note in Fig. 8a that $|\phi_N(s_j)|$ drops sharply with increasing j, and then (for large j) exhibits erratic fluctuations of undying (seemingly constant) magnitude. The plot can be interpreted to suggest that $|\phi(s)|$ should be close to zero for large s, and that the aforementioned erratic fluctuations are simply due to the (random) error in estimating a quantity that is almost zero with a sample

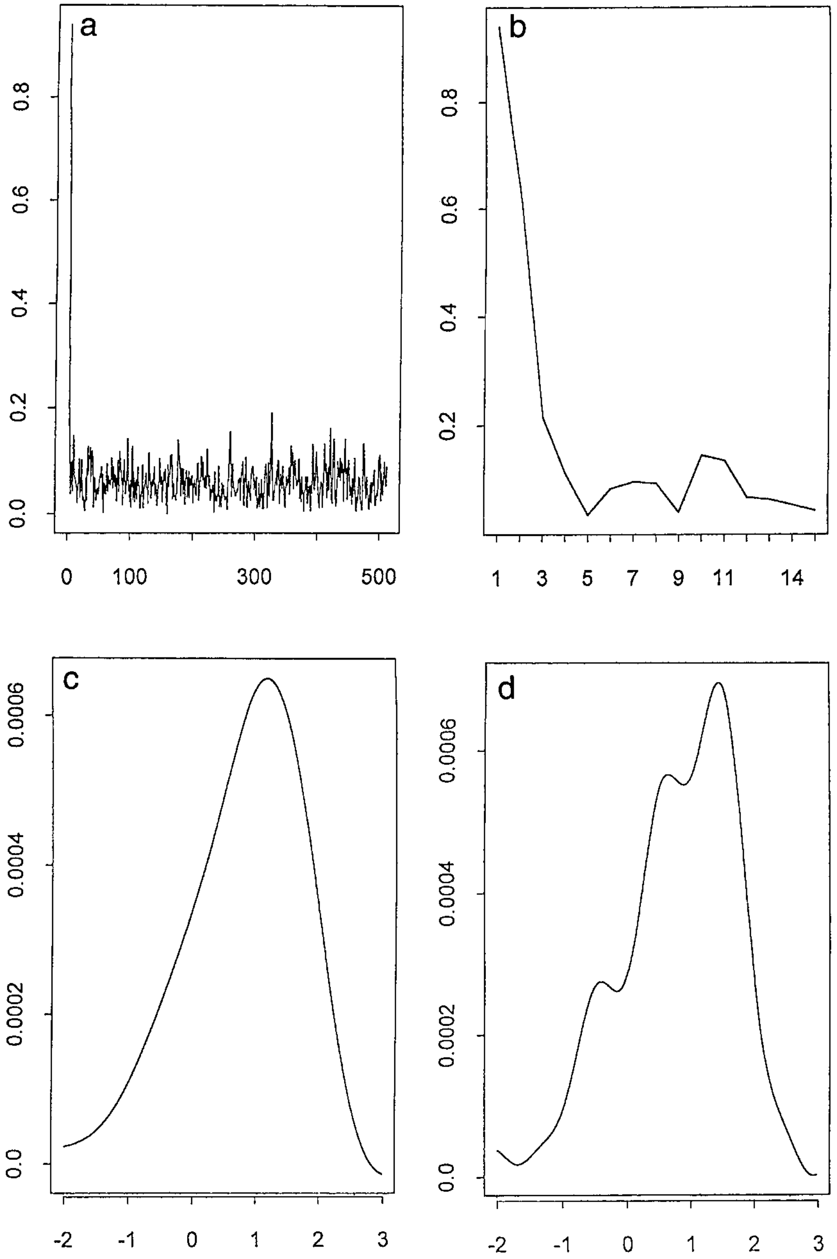

FIG. 8. "Skewed unimodal density": (a) Plot of $|\phi_N(s_j)|$ vs j for $j = 1, 2, ..., 152$. (b) Same as (a) for $j = 1, 2, ..., 15$ only. (c) Graph of estimator $\hat{f}_2$ with $h = 1/(3G)$ (optimal bandwidth). (d) Graph of estimator $\hat{f}_2$ with $h = 1/(6G)$ (undersmoothed case).

size of 200. Indeed, we know—by construction of the dataset—that this is exactly the case.

To apply our heuristic suggested in Section 3.4 for choosing the bandwidth h, we need to identify a threshold B, such that $|\phi(s)| \simeq 0$ for $|s| > B$. Figure 8b is a "magnification" of Fig. 8a for s near the origin that helps us estimate B as being 3 or 4 (times the grid-size G). Comparing the plots of density estimator $\hat{f}_2(x)$, for x spanning the range of the data, in the cases $h = 1/(3G)$ (Fig. 8c), and $h = 1/(6G)$ (Fig. 8d) confirms our heuristic of Section 3.4 for choosing h, as Fig. 8d is obviously undersmoothed. Similarly (although not shown for brevity's sake), choosing $h = 1/(4G)$ for use in computing $\hat{f}_1(x)$ was observed to give optimal smoothing results in the $c = 1$ case. Comparing the optimally smoothed estimators $\hat{f}_c(x)$ in the two cases $c = 1$ and 2, a small advantage was observed in favor of choice $c = 2$ for x near the left endpoint; notably, in either case, $\hat{f}_c(x)$ goes slightly negative for x near the right endpoint of the range.

Figure 9 concerns the "asymmetric bimodal density" and similar comments apply regarding the drop of $|\phi_N(s_j)|$ for increasing j. From Fig. 9a we again estimate the threshold B as being 3 or 4 (times the grid-size G). Note that plots of density estimator $\hat{f}_2(x)$ in the cases $h = 1/(3G)$ and $h = 1/(2G)$ again confirm our heuristic as choice $h = 1/(2G)$ leads to an obviously undersmoothed estimator, reducing the two modes to a single one. Going into further detail, we compare the plots of $\hat{f}_2(x)$ in the cases $h = 1/(3G)$ (Fig. 9b) and $h = 1/(4G)$ (Fig. 9c); it is apparent that a difference between $3G$ and $4G$ for our estimated B is of some import. In such an ambiguous situation in picking out a single value for B, we recommend using the *smaller* of the two candidates for B (i.e., the one leading to a *larger* bandwidth) as this was observed to lead to better results in either case ($c = 1$ or 2). This finding is not unexpected since, as mentioned before, with infinite-order kernels we should be prepared to use large bandwidths; see also the discussion in Devroye (1992, p. 2055).

Figure 10 concerns the Student's t density with 3 degrees of freedom, and similar findings are apparent. Figure 10b shows an optimally smoothed $\hat{f}_2(x)$ where $h = 1/(4G)$, i.e., with an implicit estimation of $\hat{B} = 4$. Comparing the plot of $\hat{f}_2(x)$ to that of $\hat{f}_1(x)$ (with either $h = 1/(4G)$ or $h = 1/(6G)$) confirms our preference of the $c = 2$ case vs $c = 1$. In particular, the graph of $\hat{f}_1(x)$ (not shown here) takes a pronounced negative dip near $x = 4$ where the data are sparse in the $h = 1/(4G)$ case; this is somehow corrected in the $h = 1/(6G)$ case showing that proper bandwidth choice is important in ensuring practical nonnegativity as well. Although it is understandable that the density estimate near the tails of a heavy-tailed distribution will be inaccurate, the $c = 1$ case gives predictably "wavy" errors that are avoided in the $c > 1$ case.

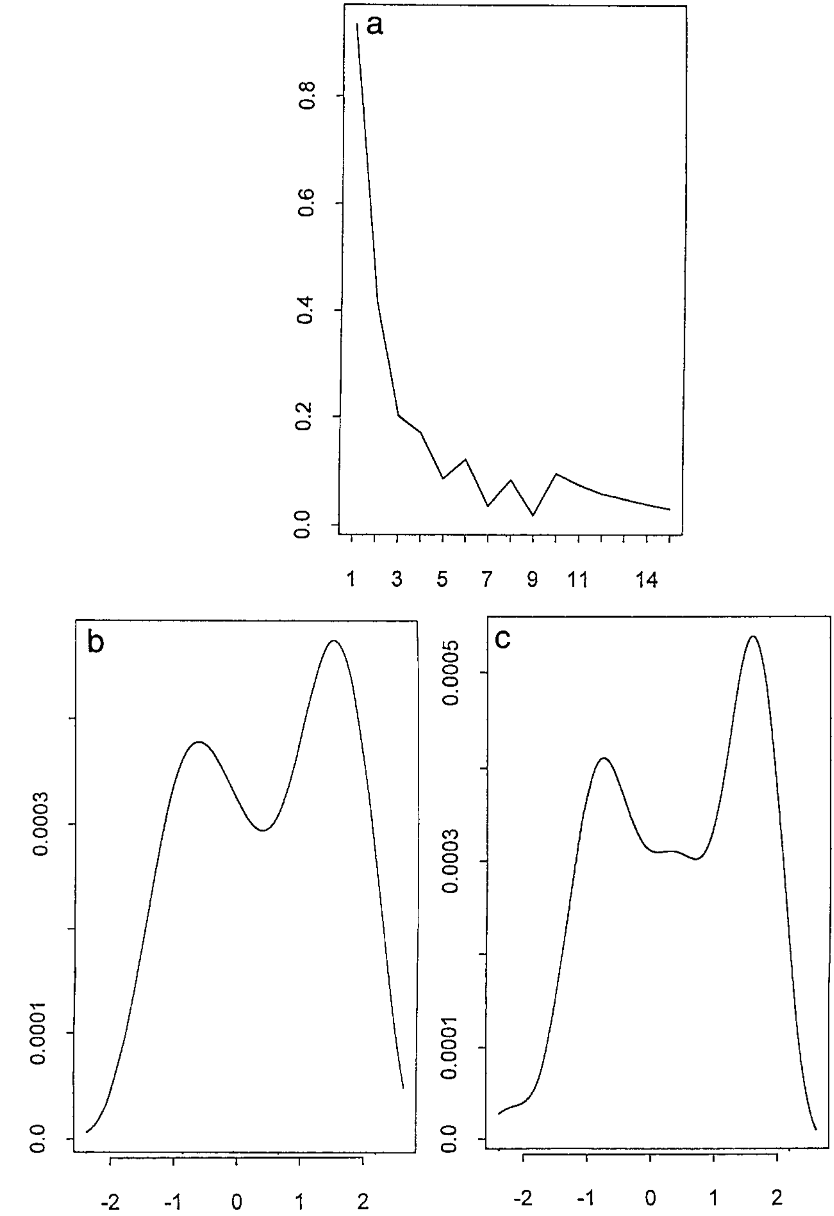

FIG. 9. "Asymmetric bimodal density": (a) Plot of $|\phi_N(s_j)|$ vs j, for $j = 1, 2, ..., 15$. (b) Graph of estimator $\hat{f}_2$ with $h = 1/(3G)$ (optimal bandwidth). (c) Graph of estimator $\hat{f}_2$ with $h = 1/(4G)$(undersmoothed case).

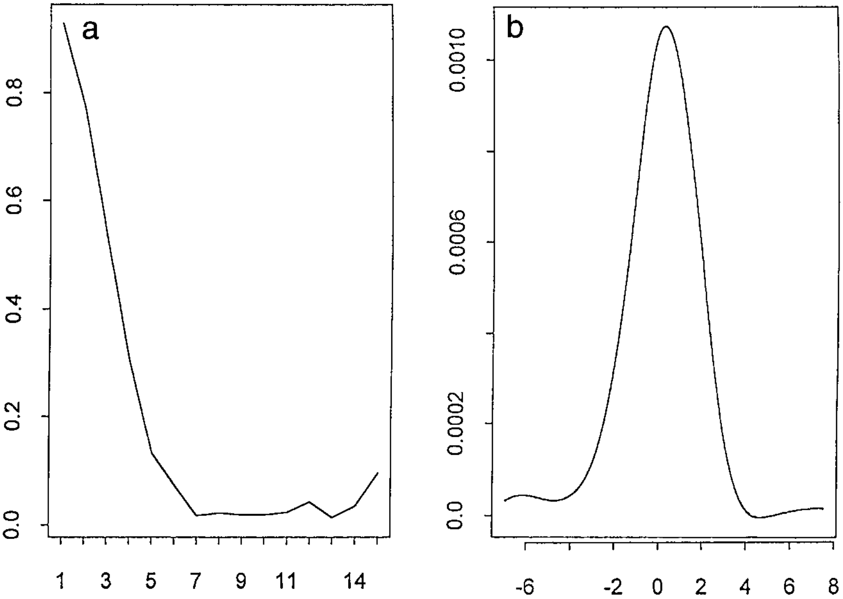

FIG. 10. Student's t-distribution with 3 degrees of freedom: (a) Plot of $|\phi_N(s_j)|$ vs. j, for $j = 1, 2, ..., 15$. (b) Graph of estimator $\hat{f}_2$ with $h = 1/(4G)$ (optimal bandwidth).

The moral is that flat-top kernels can indeed be useful in practice, although more work may be in order on the subject of optimal data-based bandwidth choice. Indeed, an interesting interplay between kernel choice (choosing c) and bandwidth choice (choosing h) has been observed in our simulations -see the discussion preceeding Table I. Our theoretical results indicate that higher (and infinite) order kernels have a practical advantage (in terms of increasing estimation accuracy) over second order kernels in case the true density is very smooth. On the contrary, if "the true density has features that make their presence felt, but can not be well recovered, the higher order kernels have no advantage over the nonnegative kernel"; cf. Marron and Wand (1992, p. 732).

4. TECHNICAL PROOFS

Proof of Theorem 1. Let $x \in R^d$; then,

$$Bias(\hat{f}_c(x) = E\hat{f}_c(x) - f(x)$$

$$= \frac{1}{(2\pi)^d} \int \lambda_c(s)\, E\phi_N(s)\, e^{-i(s \cdot x)}\, ds - \frac{1}{(2\pi)^d} \int \phi(s)\, e^{-i(s \cdot x)}\, ds$$

$$= \frac{1}{(2\pi)^d} \int (\lambda_c(s) - 1)\, \phi(s)\, e^{-i(s \cdot x)}\, ds$$

$$= \frac{1}{(2\pi)^d} \int_{\|s\|_p > 1/h} (\lambda_c(s) - 1)\, \phi(s)\, e^{-i(s \cdot x)}\, ds, \qquad (14)$$

since $\lambda_c(s) = 1$, for all s such that $\|s\|_p \leqslant 1/h$.

Now note that

$$|Bias(\hat{f}_c(x))|$$

$$\leqslant \frac{2}{(2\pi)^d} \int_{\|s\|_p > 1/h} |\phi(s)|\, ds$$

$$= \frac{2}{(2\pi)^d} \int_{\|s\|_p > 1/h} \frac{\|s\|_p^r}{\|s\|_p^r} |\phi(s)|\, ds \leqslant h^r \frac{2}{(2\pi)^d} \int_{\|s\|_p > 1/h} \|s\|_p^r\, |\phi(s)|\, ds = o(h^r),$$

where it was used that, since $|g_\lambda(s, h)| \leqslant 1$, $|\lambda_c(s) - 1| \leqslant 2$. The reason the little $o(\cdot)$ arises in the above is the following: note that

$$\int \|s\|_p^r\, |\phi(s)|\, ds = \int_{\|s\|_p > 1/h} \|s\|_p^r\, |\phi(s)|\, ds + \int_{\|s\|_p \leqslant 1/h} \|s\|_p^r\, |\phi(s)|\, ds;$$

as $h \to 0$, we have

$$\int_{\|s\|_p \leqslant 1/h} \|s\|_p^r\, |\phi(s)|\, ds \to \int \|s\|_p^r\, |\phi(s)|\, ds.$$

which is finite by Condition C_1, and thus it follows that $\int_{\|s\|_p > 1/h} \|s\|_p^r\, |\phi(s)|\, ds \to 0$.

Therefore, $Bias(\hat{f}_c(x)) = o(h^r)$, uniformly in $x \in R^d$. Finally, under Condition C_1, f is continuous at x; now if $f(x) > 0$, equation (4) holds true, and the theorem is proved. Q.E.D.

Proof of Theorem 2. We will do the proof in the case $p = \infty$, the other cases $p \in [1, \infty)$ being similar; alternatively, note that if Condition C_2 is true for some $p \in [1, \infty]$, then (by the equivalence of l_p norms for R^d) it is also true for *any* $p \in [1, \infty]$, perhaps with a change in the constants B and K, therefore for $p = \infty$ as well. Let x be any point in R^d and, as in the proof of Theorem 1, note that

$$Bias(\hat{f}_c(x)) = \frac{1}{(2\pi)^d} \int_{\|s\|_\infty > 1/h} (\lambda_c(s) - 1)\, \phi(s)\, e^{-i(s \cdot x)}\, ds,$$

since $\lambda_c(s) = 1$ for $\|s\|_\infty \leqslant 1/h$.

Consider the following partition of the set $\{\|s\|_\infty > 1/h\}$, namely $\{\|s\|_\infty > 1/h\} = \bigcup_{i=1}^{d} (A_i \cup \bar{A}_i)$, where $A_i = \{s$ such that $\|s\|_\infty > 1/h$ and $s_i = \max_k |s_k|\}$, and $\bar{A}_i = \{s$ such that $\|s\|_\infty > 1/h$ and $-s_i = \max_k |s_k|\}$. Note that the A_i's and $\bar{A}_i$'s are essentially disjoint except for their boundaries, e.g., in the case where $s_1 = s_2 = \max_k |s_k|$, etc.

Therefore, we can write

$$Bias(\hat{f}_c(x)) = \int_{A_1} + \int_{A_2} + \cdots + \int_{A_n} + \int_{\bar{A}_1} + \int_{\bar{A}_2} + \cdots + \int_{\bar{A}_n}, \tag{15}$$

where for $j = 1, 2, ..., n$,

$$\int_{A_j} = \frac{1}{(2\pi)^d} \int_{s \in A_j} (\lambda_c(s) - 1)\, \phi(s)\, e^{-i(s \cdot x)}\, ds,$$

and

$$\int_{\bar{A}_j} = \frac{1}{(2\pi)^d} \int_{s \in \bar{A}_j} (\lambda_c(s) - 1)\, \phi(s)\, e^{-i(s \cdot x)}\, ds.$$

We now proceed to analyze in detail the first term, i.e., $\int_{A_1}$. Observe again that

$$\left| \int_{A_1} \right| \leqslant \frac{2}{(2\pi)^d} \int_{\|s\|_\infty > 1/h} |\phi(s)|\, ds,$$

since $|g_\lambda(s, h)| \leqslant 1$ implies $|\lambda_c(s) - 1| \leqslant 2$. But

$$\int_{\|s\|_\infty > 1/h} |\phi(s)|\, ds \leqslant \int_{1/h}^{\infty} s_1^{d-1} B e^{-K s_1}\, ds_1 = O\left(\frac{e^{-K/h}}{h^{d-1}}\right).$$

Note that to bound the multiple integral by the single integral above, the following argument was used: let $\Delta_1 = \{s\colon s_1 \in (s_1, s_1 + ds_1)\}$; the volume of the set $A_l \cap \Delta_1$ is $s_1^{d-1} ds_1$, and $|\phi(s)| \leqslant B e^{-K s_1}$, for $s \in A_1 \cap \Delta_1$, since $s_1 = \|s\|_\infty$ over A_1.

A similar analysis shows the terms $\int_{A_2}, ..., \int_{A_n}, \int_{\bar{A}_1}, ..., \int_{\bar{A}_n}$ being bounded above by $O(e^{-K/h}/h^{d-1})$ uniformly in $x \in R^d$. Hence, $|Bias(\hat{f}_c(x))| = O(e^{-K/h}/h^{d-1})$, uniformly in $x \in R^d$. Letting $h \sim A/\log N$, where A is a constant such that $A < 2K$, it follows that

$$\sup_{x \in R^d} |Bias(\hat{f}_c(x))| = O\left(\frac{(\log N)^{d-1}}{N^{K/A}}\right) = o\left(\frac{1}{\sqrt{N}}\right),$$

as required. Finally, under Condition C_2, f is continuous at x; now if $f(x) > 0$, equation (4) holds true, and the theorem is proved. Q.E.D.

Proof of Theorem 3. The proof of Theorem 3 is again based on the decomposition (15) presented in the proof of Theorem 2. We take $p = \infty$ here as well; the other cases $p \in [1, \infty)$ are similar.

Note that $h < B^{-1}$, and thus $1/h > B$. Since $|\phi(s)| = 0$, if $\|s\|_\infty > B$, it follows that $|\phi(s)| = 0$, if $\|s\|_\infty > 1/h$. Hence,

$$\sup_{x \in R^d} |Bias(\hat{f}_c(x))| = 0,$$

as stated in the theorem.

Finally, under Condition C_3, f is continuous at x; now if $f(x) > 0$, equation (5) holds true, and the theorem is proved. ▌

ACKNOWLEDGMENTS

Many thanks are due to Professor George Kyriazis of the University of Cyprus for many helpful discussions, and to the anonymous reviewers for constructive feedback.

REFERENCES

1. M. S. Barlett, Statistical estimation of density functions, *Sankhya*, *Ser. A* **25** (1963), 245–254.
2. P. Butzer and R. Nessel, "Fourier Analysis and Approximation," Academic Press, New York, 1971.
3. T. Cacoullos, Estimation of a multivariate density, *Ann. Inst. Statist. Math.* **18** (1966), 178–189.
4. K. B. Davis, Mean integrated square error properties of density estimates, *Ann. Statist.* **5** (1977), 530–535.
5. L. Devroye, "A Course in Density Estimation," Birkhäuser, Boston, 1987.
6. L. Devroye, A note on the usefulness of superkernels in density estimates, *Ann. Statist.* **20** (1992), 2037–2056.
7. L. Gajek, On improving density estimators which are not bona fide functions, *Ann. Statist.* **14** (1968), 1612–1618.
8. T. Gasser, H. G. Müller, and V. Mammitzsch, Kernels for nonparametric curve estimation, *J. Roy. Statist. Soc. B* **47** (1985), 238–252.
9. B. L. Granovsky and H. G. Müller, Optimal kernel methods: A unifying variational principle, *Internat. Statist. Rev.* **59** (1991), 373–388.
10. L. Györfi, W. Härdle, P. Sarda, and P. Vieu, "Nonparametric Curve Estimation from Time Series," Lecture Notes in Statistics No. 60, Springer-Verlag, Berlin/New York, 1989.
11. P. Hall and R. D. Murison, Correcting the negativity of high-order kernel density estimators, *J. Multivar. Anal.* **47** (1992), 103–122.
12. I. A. Ibragimov and R. Z. Hasminksii, Estimation of distribution density belonging to a class of entire functions, *Theor. Probab. Appl.* **27** (1982), 551–562.
13. M. C. Jones, On higher order kernels, *J. Nonparametr. Statist.* **5** (1995), 215–221.

14. M. C. Jones and P. J. Foster, Generalized jacknifing and higher order kernels, *J. Nonparametr. Statist.* **3** (1993), 81–94.
15. Y. Katznelson, “An Introduction to Harmonic Analysis,” Dover, New York, 1968.
16. J. S. Marron, Visual understanding of higher order kernels, *J. Comput. Graphical Statist.* **3** (1994), 447–458.
17. J. S. Marron and M. P. Wand, Exact mean integrated squared error, *Ann. Statist.* **20** (1992), 712–736.
18. H. G. Müller, “Nonparametric Regression Analysis of Longitudinal Data,” Springer-Verlag, Berlin, 1988.
19. E. A. Nadaraya, “Nonparametric Estimation of Probability Densities and Regression Curves,” Kluwer Academic, Dordrecht, 1989.
20. E. Parzen, On estimation of a probability density function and its mode, *Ann. Math. Statist.* **33** (1962), 1065–1076.
21. D. N. Politis and J. P. Romano, On a family of smoothing kernels of infinite order, *in* “Computing Science and Statistics, Proceedings of the 25th Symposium on the interface” (M. Tarter and M. Lock, Eds.), pp. 141–145, The Interface Foundation of North America, San Diego, California, April 14–17, 1993.
22. D. N. Politis and J. P. Romano, On flat-top kernel spectral density estimators for homogeneous random fields, *J. Statist. Plann. Inference* **51** (1996), 41–53.
23. M. B. Priestley, “Spectral Analysis and Time Series,” Academic Press, New York, 1981.
24. M. Rosenblatt, “Stochastic Curve Estimation,” NSF-CBMS Regional Conference Series, Vol. 3, Institute of Mathematical Statistics, Hayward, 1991.
25. D. W. Scott, “Multivariate Density Estimation: Theory, Practice, and Visualization,” Wiley, New York, 1992.
26. B. W. Silverman, “Density Estimation for Statistics and Data Analysis,” Chapman and Hall, London, 1986.
27. E. M. Stein and W. Weiss, “Introduction to Fourier Analysis on Euclidean Spaces,” Princeton Univ. Press, Princeton, NJ, 1971.
28. G. Wahba, Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation, *Ann. Statist.* **3** (1975), 15–29.
29. M. P. Wand and M. C. Jones, Comparison of smoothing parameterizations in bivariate kernel density estimation, *J. Amer. Statist. Assoc.* **88** (1993), 520–528.
30. G. S. Watson and M. R. Leadbetter, On the estimation of the probability density, I, *Ann. Math. Statist.* **33** (1963), 480–491.