MULTIVARIATE DENSITY ESTIMATION WITH GENERAL
FLAT-TOP KERNELS OF INFINITE ORDER

BY

DIMITRIS N. POLITIS and JOSEPH P. ROMANO

TECHNICAL REPORT NO. 471
MARCH 1995

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS92-04864

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Multivariate Density Estimation with General Flat-top Kernels of Infinite Order

By

Dimitris N. Politis and Joseph P. Romano

Technical Report No. 471

March 1995

Prepared Under the Auspices

Of

National Science Foundation Grant DMS92-04864

Department of Statistics

Stanford University

Stanford, California
Multivariate Density Estimation with General
Flat-top Kernels of Infinite Order *

Dimitris N. Politis                  Joseph P. Romano
Department of Statistics            Department of Statistics
Purdue University                   Stanford University
W. Lafayette, IN 47907              Stanford, CA 94305

Abstract

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 esti-
mators 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.

Keywords and phrases: Bias reduction, Fourier transform, kernel, mean squared
error, nonparametric density estimation, rate of convergence, smoothing.

*Research partially supported by NSF grants DMS 94-04329 and DMS 94-03826.
1 Introduction

Suppose \(X_1, \ldots, X_N\) 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 \(s = (s_1, \ldots, s_d)\), \(x = (x_1, \ldots, x_d) \in R^d\), \((s \cdot x) = \sum_k s_k x_k\) is the 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 \(1 \leq p < \infty\), 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))

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

where \(\Lambda(\cdot)\) is the smoothing kernel satisfying\(^2\) \(\int \Lambda(x) dx = 1\); \(\phi_N(s) = \frac{1}{N} \sum_{k=1}^{N} e^{i(s \cdot X_k)}\) 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

\[
E \hat{f}(x) = \int \Omega(v)f(x - hv)dv,  
\]

\(^1\)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).

\(^2\)In case it is not otherwise noted, integrals will be over the whole of \(R^d\).
and

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

(3)

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

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

(4)

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

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

(5)

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

If $\Omega$ has finite moments up to $q$th order, and moments of order up to $q-1$ equal to zero, then $q$ is called the ‘order’ of the kernel $\Omega$. If the density $f$ has $r$ bounded continuous derivatives\(^3\), it then follows (cf., for example, Rosenblatt (1991)) that

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

(6)

where $k = \min(q, r)$, and $c_{f}(x)$ is a bounded function depending on $\Omega$, on $f$, and on $f$’s derivatives. This idea of choosing a kernel of order $q$ in order to get the $\text{Bias}(\hat{f}(x))$ 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, Müller, and Mammitzsch (1985), Granovsky and Müller (1991), 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

\(^3\)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).
order $O(h^r)$, no matter how large $r$ may be; so, we might say that a superkernel is a kernel with 'infinite order'. The advantage in using such a kernel is that the practitioner does not have to select a new kernel to use with each new incoming data set; the same kernel (with different choice of bandwidth) will 'adapt' to the smoothness of the unknown density $f$ in achieving equation (6) for any degree of smoothness $r$. The $O(h^r)$ order for the bias (and the corresponding rate of $O(N^{-2r/(2r+d)})$ for the Mean Squared Error of $\hat{f}$) has been shown to be optimal, i.e., it is 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 $\Omega$ 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 $\Omega$ have a Cauchy principal value of zero then the favorable bias performance follows, and this is in turn ensured by setting $\omega$ 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 $\omega$ 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 further discussion; 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 $1 \leq c \leq \infty$, $1 \leq p \leq \infty$, and define

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

(7)

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

If $c = 1$, the drop from the value 1 to the value 0 is done in a discontinuous fashion, and no function $g_{\lambda}$ is needed. On the other hand, the case $c = \infty$ covers the situation where a compact support for $\lambda_c$ is not desired. In essence, $g_{\lambda}$ interpolates between the value 1 for $||s||_p \leq c/h$, and the value 0 for $||s||_p > 1/h$. Perhaps the most ‘natural’ way for doing 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 $g_{\lambda}$ can be found in Section 3.3.

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

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

(8)

i.e., by the 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 $\Lambda_c$ is real-valued because of the symmetry of $\lambda_c$, i.e., $\lambda_c(s) = \lambda_c(-s)$.

The proposed kernel smoothed estimators of $f$ are given by

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

(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 $\Lambda_c$ with the empirical distribution, the form of $\Lambda_c$ must be calculated. In general, a closed-form expression for $\Lambda_c$ might not be available, but $\Lambda_c$ 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) it is immediate that $\text{Var}(\hat{f}_c(x)) = O(\frac{1}{Nh^d N})$, 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 $\phi$ that are in the same spirit as the conditions in Watson and Leadbetter (1963).

**Condition C₁:** For some $p \in [1, \infty]$, there is an $r > 0$, such that $\int ||s||_p^r |\phi(s)| < \infty$

**Condition C₂:** For some $p \in [1, \infty]$, there are positive constants $B$ and $K$ such that $|\phi(s)| \leq Be^{-K||s||_p}$.

**Condition C₃:** For some $p \in [1, \infty]$, there is a positive constant $B$ such that $|\phi(s)| = 0$, if $||s||_p \geq B$.

Conditions $C₁$ to $C₃$ can be interpreted as different conditions on the smoothness of the spectral density $f(w)$; 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 $C₂$ holds, then Condition $C₁$ holds as well, and if Condition $C₃$ holds, then Conditions $C₁$ and $C₂$ hold as well. Also note that if Condition $C₁$ holds, then $f$ must necessarily have $[r]$ bounded, continuous derivatives, where $[\cdot]$ is
the positive part; cf. Katzenelson (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 it is satisfied in many physical and biomedical applications (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 $\Lambda_e$ 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

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

Now let $x$ be some point in $\mathbb{R}^d$ 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 $\text{Bias}(\hat{f}_c(x))$ 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 $r$th derivative $f^{(r)}$ satisfies a uniform Lipschitz condition of order $\alpha > 1/2$; cf. Katzenelson, 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 $c < \infty$, $\omega$ has compact support.
**Theorem 2** Assume that \( h \to 0 \), as \( N \to \infty \), but with \( h^d N \to \infty \); under Condition C2, it follows that \( \sup_{x \in \mathbb{R}^d} |\text{Bias}(\hat{f}_c(x))| = O(h^{1-d} e^{-K/h}) \). If we let \( h \sim A / \log N \), as \( N \to \infty \), where \( A \) is a constant such that \( A < 2K \), it follows that

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

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

**Theorem 3** Assume Condition C3 and that, as \( N \to \infty \), \( h \) is some constant small enough such that \( h \leq B^{-1} \); it follows that

\[
\sup_{x \in \mathbb{R}^d} |\text{Bias}(\hat{f}_c(x))| = 0.
\]

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

**Remark 3.** The special case where \( c = 1 \), i.e., when the drop of \( \lambda_c \) 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.
3 Discussion and practical comments

3.1 Computational aspects and remarks

Assuming \( c < \infty \), and choosing \( g_\lambda \) to be linear, actually results into a compact expression for \( \lambda_c \), namely

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

(10)

where \((x)^+ = \max(x, 0)\) is the positive part function. A closed-form expression for

\[
\Lambda_c^{LIN}(x) = (2\pi)^{-d} \int \lambda_c^{LIN}(s)e^{-i(x \cdot s)}ds
\]

in the special case \( d = 1 \) is given by

\[
\Lambda_c^{LIN}(x) = \begin{cases} 
\frac{h \sin^2(\pi cx/h) - \sin^2(\pi x/h)}{\pi^2 x^2 (c-1)} & \text{if } c > 1 \\
\frac{1}{\pi^2} \sin(2\pi x/h) & \text{if } c = 1;
\end{cases}
\]

(11)

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

Figures 1 and 2 around here

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 \( \Lambda_c \) is rotation-invariant as well. Hence, to evaluate \( \Lambda_c(x) \) for any \( x \in \mathbb{R}^d \), it suffices to evaluate it for \( x = (x_1, 0, 0, \ldots, 0) \), with \( x_1 \) spanning \( \mathbb{R} \), and then rotate the resulting graph. But \( \Lambda_c(x_1, 0, 0, \ldots, 0) \) can be obtained by a univariate Fourier transform as

\[
\Lambda_c(x_1, 0, 0, \ldots, 0) = (2\pi)^{-1} \int \mu(s_1)e^{-is_1 x_1}ds_1,
\]

where

\[
\mu(s_1) = (2\pi)^{-d+1} \int \cdots \int \lambda_c(s_1, s_2, \ldots, s_d)ds_2ds_3 \cdots ds_d
\]
is the ‘marginal’ of the function $\lambda_c(s) = \lambda_c(s_1, s_2, \ldots, s_d)$.

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 $\lambda_c$, followed by a discrete Fourier transform; cf., for example, Silverman (1986, p. 61). In that sense, exact knowledge of the form of $\Lambda_c$ is not needed; see also our Remark 2 after Theorem 1. However, for illustration purposes, we now construct an explicit $(\Lambda_c, \lambda_c)$ 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, $1 \leq c \leq \infty$, and $h > 0$, and define

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

(12)
and

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

(13)

it is easy to check that $\lambda_c^{PROD}$ and $\Lambda_c^{PROD}$ are related to each other by a Fourier transform, and that

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

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

Figures 3 and 4 around here

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 Figures 2 and 4 the kernels $\Lambda_2^{LIN}$ and $\Lambda_2^{PROD}$ exhibit negative 'sidelobes' beside the main prominent 'lobe' around the origin which is positive.

Nevertheless, the nonnegativity is not a serious issue; there is a natural fix-up, namely using the modified estimator\footnote{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)).} $\hat{f}_c^+(x) = \max(\hat{f}_c(x), 0)$. The estimator $\hat{f}_c^+(x)$ is not only nonnegative, but is more accurate as well, in the sense that $MSE(\hat{f}_c^+(x)) \leq MSE(\hat{f}_c(x))$, for all $x$; this fact follows from the obvious inequality $|\hat{f}_c^+(x) - f(x)| \leq |\hat{f}_c(x) - f(x)|$.

In addition, note that if $f(x) > 0$, an application of Chebychev's inequality shows that $\text{Prob}(\hat{f}_c(x) = \hat{f}_c^+(x)) \to 1$ 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 $\sqrt{h} N \hat{f}_c(x)$, or $\sqrt{h} N \hat{f}_c(x)$, degenerates to a point mass at zero.

### 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 $\lambda_c$ 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 $\mathbb{R}^d$, 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 $\lambda_c$ 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 rate 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 $j$-th 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 $\{s : |\phi_N(s)| = \text{const.}\}$) 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) for more details on use of transformations. Finally, 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 $g_\lambda$

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

Nevertheless, properly choosing $c$ and the shape of the function $g_\lambda$ will definitely have a practical impact. In terms of choosing the shape of $\lambda_0$ or of $\omega$, i.e., choosing $c$ and $g_\lambda$, Devroye (1992, p. 2053) writes: "The recommendation is to take (our $\omega$) rectangular with two smooth tails added on so as to make the tails of (our $\Omega$) small. The size of these tails has to be determined from nonasymptotic considerations, perhaps via some data-based rule."

Making the tails of $\Omega$ small has a twofold advantage:\footnote{It should be stressed however that by different choices of $c$ and $g_\lambda$ we can not change the asymptotic orders of bias and variance of the resulting estimators; that is why the actual shape of $\lambda_0$ is immaterial in our asymptotic Theorems 1–3, as long as $\lambda_0$ is flat near the origin, and has finite Euclidean norm. By choosing the value of $c$ and the shape of the function $g_\lambda$ properly, we can only influence the proportionality constants in the large-sample bias and variance of the estimators.} (a) reducing the bias of the resulting estimator by reducing the 'leakage' through the many small peaks in the — typically wavy — tails of $\Omega$, 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 $\lambda_1$ and $\Lambda_1$ that are plotted in Figures 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 $\Lambda_1$ is much bigger than those in either $\Lambda_2^{LIN}$ or $\Lambda_2^{PROD}$ (see Figures 2 and 4). As a matter of fact, to really witness the tails of $\Lambda_1$ become negligible in magnitude, we have to look at $\Lambda_1(x)$ over a wider region of the $(x_1, x_2)$ plane; see Figure 7.
Figures 5, 6 and 7 around here

The reason $h = 0.05$ was used in connection with $\Lambda_1$ in Figures 6 and 7, (as opposed to $h = 0.067$ that was used for $\Lambda_2^{LIN}$ and $\Lambda_2^{PROD}$ in Figures 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 1, with these choices of $h$, the variance integrals $\int \lambda^2(s)ds = h^{-d} \int \omega^2(s)ds$ 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 $\Lambda_1$ will result to more biased estimators than either $\Lambda_2^{LIN}$ or $\Lambda_2^{PROD}$ 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, $\Lambda_1$ will result to an estimator with bigger variance than either $\Lambda_2^{LIN}$ or $\Lambda_2^{PROD}$.

<table>
<thead>
<tr>
<th></th>
<th>$\int \lambda^2(s)ds$</th>
<th>$\int \omega^2(s)ds$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>0.360</td>
<td>0.0016</td>
</tr>
<tr>
<td>Figure 3</td>
<td>0.445</td>
<td>0.0020</td>
</tr>
<tr>
<td>Figure 5</td>
<td>0.467</td>
<td>0.0243</td>
</tr>
</tbody>
</table>

Table 1. Entries of the first column are the variance integrals $\int \lambda^2(s)ds = h^{-d} \int \omega^2(s)ds$ that equal the asymptotic variance of $\sqrt{N}f(x)/\sqrt{f(x)}$ for the three functions shown in Figures 1, 3, and 5; entries of the second column are the variance constants $\int \omega^2(s)ds$ for the three functions shown in Figures 1, 3, and 5.
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 \( g_\lambda \) corresponding to either \( \Lambda^\text{LIN}_c \) or \( \Lambda^\text{PROD}_c \) will give good results; see also our discussion in Section 3.2 where the choices of \( p = 2 \) and \( p = \infty \) that correspond to \( \Lambda^\text{LIN}_c \) and \( \Lambda^\text{PROD}_c \) come up rather naturally. As evidenced by the variances presented in Table 1, \( \Lambda^\text{LIN}_2 \) might be somewhat preferable to \( \Lambda^\text{PROD}_2 \), 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 \( g_\lambda(s) \) will turn out to be \( c = \infty \), and a \( g_\lambda(s) \) that decays to zero fast enough as \( s \to \infty \), but that it is not necessarily nonnegative for all values of \( s \); rather \( g_\lambda(s) \) 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. 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 \( \Lambda^\text{LIN}_c \) or \( \Lambda^\text{PROD}_c \) (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 bandwidth \( h \). 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 $h = \text{const.} \leq 1/B$ 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.

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 $\hat{B}$. 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 $\phi_N(s) \rightarrow \phi(s)$ as $N \rightarrow \infty$) would be 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 $C_1$–$C_3$.

Acknowledgement. Many thanks are due to Prof. George Kyriazis of the University of Cyprus for many helpful discussions.
4 Technical proofs.

Proof of Theorem 1. Let \( x \) be any point in \( \mathbb{R}^d \) and note that

\[
\text{Bias}(\hat{f}_c(x)) = E\hat{f}_c(x) - f(x)
\]

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

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

\hspace{1cm} (14)

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

Now note that

\[
|\text{Bias}(\hat{f}_c(x))| \leq \frac{2}{(2\pi)^d} \int_{||s||_p > 1/h} ||s||_p^p |\phi(s)|ds
\]

\[
= \frac{2}{(2\pi)^d} \int_{||s||_p > 1/h} ||s||_p^p |\phi(s)|ds \leq h^{-r} \frac{2}{(2\pi)^d} \int_{||s||_p > 1/h} ||s||_p^p |\phi(s)|ds = o(h^{-r}),
\]

where it was used that, since \( |g(s)| \leq 1 \), \( |\lambda_c(s) - 1| \leq 2 \). The reason the little \( o(\cdot) \) arises in the above is the following: note that \( \int ||s||_p^p |\phi(s)|ds = \int_{||s||_p > 1/h} ||s||_p^p |\phi(s)|ds + \int_{||s||_p \leq 1/h} ||s||_p^p |\phi(s)|ds \); as \( h \to 0 \), we have \( \int_{||s||_p \leq 1/h} ||s||_p^p |\phi(s)|ds \to 0 ||s||_p^p |\phi(s)|ds \to 0 \), which is finite by Condition \( C_1 \), and thus it follows that \( \int_{||s||_p > 1/h} ||s||_p^p |\phi(s)|ds \to 0 \).

Therefore, \( \text{Bias}(\hat{f}_c(x)) = o(h^{-r}) \), uniformly in \( x \in \mathbb{R}^d \), as we were supposed to prove. Finally, under Condition \( C_1 \), \( f \) is continuous at \( x \); now if \( f(x) > 0 \), equation (4) holds true, and the asymptotic order of the \( \text{MSE}(\hat{f}_c(x)) \) is \( O(N^{-2r/(2r+d)}) \) as stated in the theorem. 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 \( \mathbb{R}^d \)) 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 \( \mathbb{R}^d \) and, as in the proof of Theorem 1, note that

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

Consider the following partition of the set $\{\|s\|_\infty > 1/h\}$, namely $\{\|s\|_\infty > 1/h\} = \bigcup_{i=1}^d \left(A_i \cup \bar{A}_i\right)$, where $A_i = \{s \text{ such that } \|s\|_\infty > 1/h \text{ and } s_i = \max_k |s_k|\}$, and $\bar{A}_i = \{s \text{ such that } \|s\|_\infty > 1/h \text{ and } -s_i = \max_k |s_k|\}$. 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

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

where for $j = 1, 2, \ldots, n$

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

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 that

$$
|\int_{A_j}| \leq \frac{2}{(2\pi)^d} \int_{\|s\|_\infty > 1/h} |\phi(s)| ds,
$$

since $|g_\lambda(\|s\|_\infty)| \leq 1$ implies $|\lambda_c(s) - 1| \leq 2$. But

$$
\int_{\|s\|_\infty > 1/h} |\phi(s)| ds \leq \int_{1/h}^{\infty} s_1^{d-1} Be^{-Ks_1} ds = O\left(\frac{e^{-K/h}}{h^{d-1}}\right).
$$

Note that to bound the multiple integral by the single integral above, the following argument was used: let $\Delta_1 = \{s : s_1 \in (s_1, s_1 + ds_1)\}$; the volume of the set $A_1 \cap \Delta_1$ is $s_1^{d-1} ds_1$, and $|\phi(s)| \leq Be^{-Ks_1}$, for $s \in A_1 \cap \Delta_1$, since $s_1 = \|s\|_\infty$ over $A_1$.

A similar analysis shows the terms $\int_{A_2}, \ldots, \int_{A_n}, \int_{\bar{A}_1}, \ldots, \int_{\bar{A}_n}$ being bounded above by $O\left(\frac{e^{-K/h}}{h^{d-1}}\right)$, uniformly in $x \in \mathbb{R}^d$. Hence, $|\text{Bias}(\hat{f}_c(x))| = O\left(\frac{e^{-K/h}}{\sqrt{A}}\right)$, uniformly in $x \in \mathbb{R}^d$.

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

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

Finally, under Condition $C_2$, $f$ is continuous at $x$; now if $f(x) > 0$, equation (4) holds true, and the asymptotic order of the $MSE(\hat{f}_c(x))$ is $O(\frac{\log N}{N})$ as stated in the theorem. 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,

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

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 asymptotic order of the $MSE(\hat{f}_c(x))$ is $O(1/N)$ as stated in the theorem. Q.E.D.
References


CAPTIONS FOR FIGURES.

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

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

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

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

Figure 5. The Fourier transform of $\Lambda_1$, i.e., $\lambda_1(s)$, as a function of $s = (s_1, s_2)$, for $d = 2$, $p = \infty$, and $h = 0.05$.

Figure 6. The kernel $\Lambda_1(x)$, as a function of $x = (x_1, x_2)$, for $d = 2$, $p = \infty$, and $h = 0.05$.

Figure 7. Same as Figure 6, i.e., $d = 2$, $p = \infty$, and $h = 0.05$, but here $\Lambda_1(x)$ shown over a wider region of the $(x_1, x_2)$ plane.
FIGURE 1.