The Annals of Statistics
1995, Vol. 23, No. 6, 2241–2263

ON BANDWIDTH CHOICE FOR DENSITY ESTIMATION WITH DEPENDENT DATA

BY PETER HALL, SOUMENDRA NATH LAHIRI AND YOUNG K. TRUONG

Australian National University

We address the empirical bandwidth choice problem in cases where the range of dependence may be virtually arbitrarily long. Assuming that the observed data derive from an unknown function of a Gaussian process, it is argued that, unlike more traditional contexts of statistical inference, in density estimation there is no clear role for the classical distinction between short- and long-range dependence. Indeed, the "boundaries" that separate different modes of behaviour for optimal bandwidths and mean squared errors are determined more by kernel order than by traditional notions of strength of dependence, for example, by whether or not the sum of the covariances converges. We provide surprising evidence that, even for some strongly dependent data sequences, the asymptotically optimal bandwidth for independent data is a good choice. A plug-in empirical bandwidth selector based on this observation is suggested. We determine the properties of this choice for a wide range of different strengths of dependence. Properties of cross-validation are also addressed.

1. Introduction. There is a growing body of work on density estimation under conditions that allow quite general strengths of dependence. In particular, Castellana and Leadbetter (1986) and Castellana (1989) considered kernel estimators under conditions that encompass very-long-range dependence in a Gaussian process, and addressed first-order properties of mean integrated squared error. Achieving optimal performance in this setting, even approximately, requires an empirical rule for selecting bandwidth. Hart (1984) treated this problem for the Fourier integral estimator of a density. However, while that may be viewed as a kernel estimator, its kernel does not satisfy the conditions that are typically imposed (in particular, it is not Riemann integrable) and the simulation study described by Hart indicates that the Fourier integral estimator may not perform particularly well. More recently, Hart and Vieu (1990) have developed a leave-many-out approach to cross-validation in more traditional kernel density estimation and have shown that it can produce first-order optimality provided that the strength of dependence is not too great.

Received April 1994; revised December 1994.
AMS 1991 *subject classifications.* Primary 62G07; secondary 62M10.
Key words and phrases. Bandwidth choice, cross-validation, density estimation, Gaussian process, integrated squared error, kernel methods, long-range dependence, mean integrated squared error, plug-in rule, short-range dependence, window width.

The results of Hart and Vieu (1990) are of considerable importance, not least because they represent the first detailed account of empirical bandwidth choice in kernel density estimation for dependent data. However, it is important to appreciate their limitations, in order to see just how much of the bandwidth choice problem remains unsolved. The regularity conditions imposed by Hart and Vieu demand exceptionally short-range dependence of the underlying process; for example, in the special case of a Gaussian process their assumptions appear to require that the covariance $\gamma(i)$ between observations lagged i apart decrease like $i^{-\alpha}$ for some α that is at least several hundred in size. While this indicates that geometrically weak dependence is not essential in order for empirical bandwidth choice to be viable, it does raise the question of what happens when dependence is stronger, in particular if it is so long-range that the sum of the covariances diverges.

The present paper takes up this issue directly. Our aim is to provide detail about both theoretically optimal bandwidth choice and its empirical approximation, under conditions of virtually arbitrarily long-range dependence. There seems little hope of addressing such a complex problem without making rather specific assumptions about the nature of dependence of the underlying stochastic process. We would argue that, while the generality of the work of Hart and Vieu (1990) in the context of short-range dependence is strongly enhanced by their emphasis on mixing processes, that approach does restrict the range of different levels of dependence that can be considered in detail. In our analysis we assume that the underlying stationary stochastic process is an unknown function of a Gaussian process. While structurally more restrictive than the mixing assumption of Hart and Vieu (1990), our condition allows the strength of dependence to be determined entirely by a single sequence of numbers, the covariances $\gamma(i)$. The "function of a Gaussian process" model has been used before in other problems (e.g., the analysis of fractal properties of a stochastic process in continuous time).

Assuming the model just described, our main conclusions are as follows. [It is convenient here for us to describe our results in the case where $\gamma(i) \sim ci^{-\alpha}$ for constants $c \neq 0$ and $\alpha > 0$, although the analysis that we shall present later in the paper does not require such stringent assumptions.] When $\alpha > \frac{4}{5}$ the deterministic bandwidth h_0 that minimizes MISE agrees even at *second* order with its counterpart h^* in the case of independent data. When $\frac{2}{5} < \alpha < \infty$, the bandwidth that produces the overall minimum still agrees to first order with h^*. This result fails when $\alpha \leq \frac{2}{5}$, but even there h^* produces first-order minimization of MISE, since, whenever $\alpha < \frac{4}{5}$, adjusting the bandwidth in the vicinity of the optimum has an effect only on second- and higher-order terms. More generally, if the kernel is of order $r \geq 2$, then the "boundaries" at $\frac{2}{5}$ and $\frac{4}{5}$ change to $r/(2r + 1)$ and $2r/(2r + 1)$, respectively. (For the sake of simplicity and brevity we shall not provide details of those cases.)

In Section 3 we develop dependent-data versions of the root-n consistent plug-in rule considered by Hall, Sheather, Jones and Marron (1991). It is shown in Section 4 that in the context of data that are functions of a Gaussian process, least-squares cross-validation can produce asymptotic

first-order optimality under conditions of very-long-range dependence. However, it can be much more variable than the plug-in rule, as our simulation study in Section 5 indicates.

Work on density estimation under dependence dates from the seminal contributions of Roussas (1969) and Rosenblatt (1970), who considered estimation based on Markov chain data. Two of the major issues treated have been consistency and rates of convergence, in various metrics [see, e.g., Nguyen (1979), Yakowitz (1985), Hart (1987), Roussas and Ioannides (1987), Roussas (1988), Tran (1989, 1990a), Hall and Hart (1990), Meloche (1990), Roussas (1991) and Yu (1993)]. Asymptotic distribution theory has been developed by several authors [see, e.g., Castellana and Leadbetter (1986), Yakowitz (1989), Roussas (1990a) and Tran (1990b)]. Recursive or sequential density estimation with dependent data has been treated extensively, and we cite here only two representative contributions, by Györfi and Masry (1990) and Roussas (1990b). The prevalence of data that exhibit long-range dependence, and the variety of statistical methods that are available for their analysis, have been well documented by Beran (1992). The considerable practical importance of long-range dependence has been appreciated since at least Hurst's (1951) seminal observation of long-range dependence in hydrological data.

By way of notation, we write $\{X_i\}$ for a stationary sequence of random variables, whose marginal density f we desire to estimate. Our main kernel function K is taken to be a symmetric probability density, and so satisfies the usual second-order conditions:

$$\int x^i K(x)\,dx = \begin{cases} 1, & \text{if } i = 0, \\ 0, & \text{if } i = 1, \\ \mu_2 \neq 0, & \text{if } i = 2. \end{cases}$$

Our estimator of f is defined by

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}.$$

In our analysis of its properties we shall economize our use of space by allowing relatively generous smoothness assumptions.

2. First- and second-order properties of the "optimal" bandwidth for dependent data. We begin by discussing known results in the case of independent data. There, for a twice-differentiable density, mean integrated squared error (MISE) may be Taylor-expanded as

$$M(h) = \int E(\hat{f} - f)^2 = (nh)^{-1}R(K) + h^4 \tfrac{1}{4}\mu_2^2 I_2 + o\{(nh)^{-1} + h^4\}, \tag{2.1}$$

with $R(K) = \int K^2$, $\mu_j = \int z^j K(z)\,dz$ and $I_j = \int (f^{(j)})^2$. [See, e.g., Silverman (1986), Section 3.3.] This formula indicates that the bandwidth h^* that is optimal in the sense of minimizing $M(h)$ satisfies

$$h^* \sim (J_1/n)^{1/5}, \tag{2.2}$$

where $J_1 = R(K)/(\mu_2^2 I_2)$. We say that (2.2) describes first-order properties of the optimal bandwidth.

To obtain detail about higher-order properties of h^* we require a refined version of (2.1). Under the assumption that f has at least six derivatives this is given by

$$M(h) = N(h) + O(n^{-1}h^2 + h^8), \tag{2.3}$$

where

$$N(h) = (nh)^{-1}R(K) - n^{-1}I_0 + h^4 \tfrac{1}{4}\mu_2^2 I_2 - h^6 \tfrac{1}{24}\mu_2\,\mu_4 I_3. \tag{2.4}$$

Result (2.3) implies that the optimal bandwidth for independent data satisfies

$$h^* = (J_1/n)^{1/5} + J_2(J_1/n)^{3/5} + O(n^{-4/5}), \tag{2.5}$$

where $J_2 = \mu_4 I_3/(20\mu_2 I_2)$. The extra term $J_2(J_1/n)^{3/5}$ appearing in (2.5), compared with (2.2), represents second-order terms in an asymptotic approximation to the optimal bandwidth. Third- and higher-order terms are represented by the remainder, $O(n^{-4/5})$.

We claim that (2.5) describes properties of h^* to a slightly greater order of accuracy than may be achieved by empirical approximation, and so further refinements are not really of practical interest. To appreciate this point, observe that even if f is known up to a single, estimable parameter, the quantity I_2 (and hence J_1) cannot be approximated to a greater degree of accuracy than order $n^{-1/2}$. Such an order is achievable in a nonparametric context if f has four derivatives and satisfies a Lipschitz condition of order $\frac{1}{4}$ or more on the fourth derivative [see, e.g., Bickel and Ritov (1988)]. Therefore the smallest order of error that could be associated with an estimate of h^* is $n^{-1/2}n^{-1/5} = n^{-7/10}$. The latter is larger than the remainder term in (2.5), and so that formula provides as much accuracy as is needed for an empirical study of bandwidth choice.

The discussion above addressed only the case of independent data. However, a striking conclusion to be drawn from the theoretical analysis that we shall give shortly is that (2.5) is also valid under quite general conditions of dependence. Only for processes with particularly long-range dependence, for example, Gaussian processes where the sum of the *squared* covariances does not converge, is this not true. Thus, the presence of dependence does not necessarily make a great deal of difference to the bandwidth choice problem. This contrasts strikingly with the situation in the closely related problem of nonparametric regression, where even first-order properties of bandwidth choice are altered by a very small amount of dependence. For example, the analogues of (2.2) for independent and m-dependent data are quite different, with J_1 having different forms in the two cases. [See, e.g., Hart (1987, 1991)].

In stating our first result for this section we impose conditions directly on the joint probability densities of bivariate distributions in the sequence $\{X_i\}$. Let f_i denote the density of (X_j, X_{i+j}), and put $g_i(x_1, x_2) = f_i(x_1, x_2) - f(x_1)f(x_2)$. We ask that K be bounded and compactly supported, that each g_i

have two derivatives of all types, that f have six bounded, integrable derivatives and that, for some $\varepsilon > 0$,

$$\begin{aligned} &\sum_{i=1}^{\infty} \int |g_i(x, x)|\, dx < \infty, \qquad \int \sup_{|z| \le \varepsilon} |f^{(6)}(x + z)|\, dx < \infty, \\ &\int \sup_{\substack{j_1 + j_2 = 2 \\ |z_1|, |z_2| \le \varepsilon}} |g_i^{(j_1, j_2)}(x + z_1, x + z_2) - g_i^{(j_1, j_2)}(x, x)|\, dx < \infty. \end{aligned} \tag{2.6}$$

Collectively we call these conditions (C_1). [The implications of (2.6) are discussed two paragraphs below the proof of Theorem 2.1.] Generalize the definition of N, at (2.4), to the case of a dependent sequence:

$$\begin{aligned} N(h) = {} & (nh)^{-1} R(K) \\ & + n^{-1} \left\{ 2 \sum_{i=1}^{n-1} (1 - n^{-1} i) \int g_i(x, x)\, dx - I_0 \right\} \\ & + h^4 \tfrac{1}{4} \mu_2^2 I_2 - h^6 \tfrac{1}{24} \mu_2 \mu_4 I_3. \end{aligned} \tag{2.7}$$

The assumption that f have six or more derivatives is common in settings such as this, where high-order approximations to bandwidths are treated. Indeed, estimating I_3 at a root-n consistent rate demands $6\frac{1}{4}$ derivatives of f [see Bickel and Ritov (1988)].

THEOREM 2.1. *For dependent data satisfying conditions* (C_1), *and employing the general definition of* N *at* (2.7), *the expansion at* (2.3) *remains valid.*

PROOF. Set $K_h(x) = h^{-1} K(x/h)$, and observe that

$$\begin{aligned} n \operatorname{var} \hat{f}(x) = {} & \operatorname{var} K_h(x - X_1) \\ & + 2 \sum_{i=1}^{n-1} (1 - n^{-1} i) \operatorname{cov}\{K_h(x - X_1), K_h(x - X_{1+i})\}, \end{aligned}$$

$$\begin{aligned} & \operatorname{cov}\{K_h(x - X_1), K_h(x - X_{1+i})\} \\ & \quad = \iint K(y_1) K(y_2) g_i(x - hy_1, x - hy_2)\, dy_1\, dy_2 \\ & \quad = g_i(x, x) + h^2 r(x), \end{aligned} \tag{2.8}$$

where, if K is compactly supported on $(-c, c)$,

$$|r(x)| \le 4c^2 (\sup K^2) \sup_{|z_1|, |z_2| \le ch} \sum_{\substack{0 \le j_1, j_2 \le 2 \\ j_1 + j_2 = 2}} |g_i^{(j_1, j_2)}(x + z_1, x + z_2) - g_i^{(j_1, j_2)}(x, x)|.$$

Therefore,

$$\begin{aligned} s &\equiv \int \sum_{i=1}^{n-1} (1 - n^{-1} i) \operatorname{cov}\{K_h(x - X_1), K_h(x - X_{1+i})\}\, dx \\ &= \sum_{i=1}^{n-1} (1 - n^{-1} i) \int g_i(x, x)\, dx + O(h^2). \end{aligned} \tag{2.9}$$

Similarly,

$$\begin{aligned}\int \operatorname{var} K_h(x - X_1)\,dx &= h^{-1} - \int \left\{\int K(y) f(x - hy)\,dy\right\}^2 dx \\ &= h^{-1} - \int f^2 + O(h^2),\end{aligned} \tag{2.10}$$

and so

$$\int \operatorname{var} \hat{f} = (nh)^{-1} + n^{-1}\left\{2\sum_{i=1}^{n-1}(1 - n^{-1}i)\int g_i(x, x)\,dx - I_0\right\} + O(n^{-1}h^2).$$

A similar but simpler argument may be used to show that

$$\int \left(E\hat{f} - f\right)^2 = h^4 \tfrac{1}{4}\mu_2^2 I_2 - h^6 \tfrac{1}{24}\mu_2\,\mu_4 I_3 + O(h^6). \tag{2.11}$$

The theorem follows from these two formulae. □

Note particularly that the definition of N at (2.7) differs from that at (2.4) only through the addition of terms that do not depend on h. Therefore, minimizing $N(h)$ over h is not affected by the revised definition. This fact and Theorem 2.1 together ensure that formula (2.5) holds true even in the dependent case; no changes are needed. Therefore, provided (2.6) holds, the optimal bandwidth is identical to its counterpart in the case of independent data, up to but not including terms of third order.

To discuss the implications of this result, and also conditions (2.6), let us assume for the sake of simplicity that $\{X_i\}$ is a stationary Gaussian sequence with zero mean, unit variance and covariance function $\gamma(i) = E(X_{i+j}X_j)$, for all integers $j \geq 0$. This is the type of sequence employed in most simulation studies of the problem of density estimation under dependence [see, e.g., Hart and Vieu (1990)]. Asking that the dependence of this process decay with increasing lag along the sequence is equivalent to demanding that $\gamma(i) \to 0$ as $i \to \infty$, and so we make this assumption. If in addition γ is ultimately monotone, then condition (2.6) holds if and only if

$$\sum_{i=1}^{\infty} |\gamma(i)| < \infty. \tag{2.12}$$

Indeed, if (2.12) is violated and γ is ultimately monotone, then (2.6) fails and the infinite series in the definition of N at (2.7) diverges, so that N is not well defined. Of course, (2.12) is the usual definition of "short-range dependence"; for a process with ultimately monotone covariance, (2.12) is necessary and sufficient for the sample mean to have variance of order n^{-1}.

Nevertheless, it is possible for the process $\{X_i\}$ to exhibit long-range dependence yet the optimal bandwidth agree to first order, and even to second order, with that given by (2.5) in the case of independent data. To investigate this in greater detail, we assume that $X_i = a(Y_i)$, $i \geq 1$, where $\{Y_i\}$ is a stationary Gaussian process with zero mean and unit variance, and

the function $a = b^{-1}$ (meaning that b is the inverse of a) is monotone and seven times differentiable. Additionally we ask that b satisfy $E\{a(Y)\} = 0$, $E\{a(Y)^2\} = 1$,

$$|b^{(1)}(x)|^{-1} \le B_1 \exp\{\varepsilon_1 b(x)^2\} \quad \text{and} \quad \sum_{j=1}^{7} |b^{(j)}(x)| \le B_2 \exp\{\varepsilon_2 b(x)^2\},$$

for some $B_1, B_2(\varepsilon_2) > 0$, some $0 < \varepsilon_1 \le \frac{1}{2}\{1 + \sup|\gamma(i)|\}^{-1}$ and all $\varepsilon_2 > 0$. (This is sufficient to ensure that the joint density f_i has six integrable derivatives of all types.) Collectively we call these conditions (C_2).

Set

$$B_0 = \int (2\pi)^{-1} b(x)^2 \{b(x)^2 - 3\} b^{(1)}(x)^2 \exp\{-b(x)^2\}\, dx.$$

[In the case of Gaussian data, where $b(x) \equiv x$, we have $B_0 = -3/(8\pi^{1/2})$.] Redefine

$$\begin{aligned} N(h) = {} & (nh)^{-1} R(K) + n^{-1}\left\{2 \sum_{i=1}^{n-1} (1 - n^{-1}i) \int g_i(x, x)\, dx - I_0\right\} \\ & + B_0 n^{-1} h^2 \mu_2 \sum_{i=1}^{n-1} (1 - n^{-1}i)\gamma(i) \\ & + h^4 \tfrac{1}{4} \mu_2^2 I_2 - h^6 \tfrac{1}{24} \mu_2\, \mu_4 I_3. \end{aligned} \tag{2.13}$$

This definition of N differs from that at (2.7) only by inclusion of the term involving B_0. In the context of (2.3), this term belongs to the remainder, but should condition (2.12) fail then it will be of larger size than $n^{-1}h^2$ and so should be included in N. Then terms of the next smallest order go into the remainder, as the following result shows.

THEOREM 2.2. *Under conditions* (C_2), *and employing the definition of* N *at* (2.13), *the expansion at* (2.3) *remains valid with a slightly altered form of remainder*:

$$M(h) = N(h) + O\left\{n^{-1}h^2 \sum_{i=1}^{n-1} \gamma(i)^2 + n^{-1}h^3 \sum_{i=1}^{n-1} |\gamma(i)|\right\}. \tag{2.14}$$

PROOF. Set $\rho_i = \gamma(i)$, $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ and

$$\phi_i(x_1, x_2) = (2\pi)^{-1}(1 - \rho_i^2)^{-1/2} \exp\left\{-\tfrac{1}{2}(1 - \rho_i^2)^{-1}(x_1^2 + x_2^2 - 2\rho_i x_1 x_2)\right\}.$$

In this notation,

$$g_i(x_1, x_2) = \phi_i\{b(x_1), b(x_2)\} b'(x_1) b'(x_2) - \prod_{i=1}^{2} [\phi\{b(x_i)\} b'(x_i)].$$

Note particularly that

$$\phi_i(x_1, x_2) \le (2\pi)^{-1}(1 - \rho_i^2)^{-1/2} \exp\left\{-\tfrac{1}{2}(1 + |\rho_i|)^{-1}(x_1^2 + x_2^2)\right\}.$$

This inequality motivates the condition on $|b'(x)|^{-1}$, part of conditions (C_2). Noting the conditions on b, we may show from (2.8) by Taylor expansion that, in place of (2.9) and without making any assumptions about γ other than $\gamma(i) \to 0$ as $i \to \infty$, we have

$$s = \sum_{i=1}^{n-1} (1 - n^{-1}i) \int \{g_i(x, x) + h^2 \mu_2 g_i^{(2,0)}(x, x)\}\, dx + O\Big\{h^3 \sum_{i=1}^{n} |\gamma(i)|\Big\}.$$

Observe too that as $i \to \infty$ we have, for all $\varepsilon > 0$,

$$g_i^{(2,0)}(x, x) = \gamma(i)(2\pi)^{-1} b(x)^2 \{b(x)^2 - 3\} b'(x)^2 \exp\{-b(x)^2\} + O\big[\gamma(i)^2 \exp\{-(1 - \varepsilon)x^2\}\big],$$

whence

$$\begin{aligned} s = {} & \sum_{i=1}^{n-1} (1 - n^{-1}i) \int \{g_i(x, x) + h^2 \mu_2 g_i^{(2,0)}(x, x)\}\, dx \\ & + B_1 h^2 \mu_2 \sum_{i=1}^{n-1} (1 - n^{-1}i)\gamma(i) \\ & + O\Big\{h^3 \sum_{i=1}^{n} |\gamma(i)| + h^2 \sum_{i=1}^{n} \gamma(i)^2\Big\}. \end{aligned} \tag{2.15}$$

Result (2.14) follows from (2.10), (2.11) and (2.15). □

Some of the implications of Theorem 2.2 are quite unexpected. For example, consider the case where $\gamma(i)/(ci^{-\alpha}) \to 1$ as $i \to \infty$, with $c \neq 0$ and $\alpha > 0$. Then, strikingly, h_0 is always of order $n^{-1/5}$ when $\frac{2}{5} < \alpha < \infty$. When $\frac{4}{5} < \alpha < \infty$, only third- and higher-order terms in an expansion of h_0 are affected by dependence (we remind the reader that $\alpha \leq 1$ corresponds to long-range dependence); when $\frac{2}{5} < \alpha \leq \frac{4}{5}$, only second- and higher-order terms are affected by dependence; and when $0 < \alpha \leq \frac{2}{5}$, first-order terms are affected. In fact, without assuming $\gamma(i) \sim ci^{-\alpha}$, it may be shown that if $\gamma(i) = o(i^{-2/5})$, then (2.2) holds, so that the optimal bandwidth appropriate for independent data is still first-order optimal under strong long-range dependence.

In more detail, it may be shown from (2.14) that, when $\gamma(i) \sim ci^{-\alpha}$,

$$h_0 = \begin{cases} (J_1/n)^{1/5} + J_2(J_1/n)^{4/5} + o(n^{-3/5}), & \text{if } \frac{4}{5} < \alpha < \infty, \\ (J_1/n)^{1/5} + c_1 n^{(1/5)-\alpha} + o(n^{(1/5)-\alpha}), & \text{if } \frac{2}{5} < \alpha \leq \frac{4}{5}, \\ c_2 n^{(\alpha-1)/3} + o(n^{(\alpha-1)/3}), & \text{if } 0 < \alpha \leq \frac{2}{5} \text{ and } C_0 > 0, \\ c_2 n^{-\alpha/2} + o(n^{-\alpha/2}), & \text{if } 0 < \alpha \leq \frac{2}{5} \text{ and } C_0 < 0. \end{cases} \tag{2.16}$$

In these formulae, c_1 and c_2 denote nonzero constants, with $c_2 > 0$.

Observe too that, by (2.13) and (2.16), minimum mean integrated squared error is given by

$$M(h_0) \sim \begin{cases} c_3 n^{-4/5}, & \text{if } \frac{4}{5} < \alpha < \infty, \\ c_4 n^{-\alpha}, & \text{if } 0 < \alpha \le \frac{4}{5}, \end{cases} \tag{2.17}$$

where $c_3 = \frac{5}{4}\{R(K)^4 \mu_2^2 I_2\}^{1/5}$ has exactly the same formula that it would under independence, and $c_4 > 0$. Note particularly that the optimal convergence rate for independent data, $M(h^*) \sim$ const. $n^{-4/5}$, is maintained for $\alpha > \frac{4}{5}$, which includes many cases of long-range dependence. When $0 < \alpha \le \frac{4}{5}$, the slower mean square convergence rate of $n^{-\alpha}$ is obtainable for a wide range of bandwidths, indeed for h's of any size strictly between $n^{-(1-\alpha)}$ and $n^{-\alpha/4}$. When $\alpha < \frac{4}{5}$ the particular optimal choice recommended by (2.16) serves only to optimize second-order terms in an expansion of mean integrated squared error. This is quite unlike the case for short-range dependence, and serves to explain the surprising results at (2.16).

We conclude by briefly discussing the regularity conditions (C_2). They ask that the distribution F have at least seven derivatives and possess a finite moment of sufficiently high polynomial order. The smoothness part is clear from the assumption that $b^{(7)}$ exists. To appreciate the moment part, suppose for the sake of simplicity that $1 - F(x) \sim cx^{-\beta}$ as $x \to \infty$, where C, $\beta > 0$; and set $f = F'$. It may be shown that $|b^{(1)}(x)|^{-1} \sim b(x)\{1 - F(x)\}/f(x)$ and that $b(x) \sim [-2\log\{1 - F(x)\}]^{1/2}$, whence it follows that conditions (C_2) hold in respect of the upper tail if and only if $\beta > 1 + \sup|\gamma(i)|$. Likewise, the exponent in the lower tail should also exceed $1 + \sup|\gamma(i)|$.

3. Properties of a plug-in empirical bandwidth selector.

3.1. *Summary and main results.* In the present section we study the effect that different ranges of dependence have on the ability of a plug-in empirical bandwidth selector to approximate high-order terms in a formula for the optimal bandwidth, for example, as evidenced by the first two cases of (2.16). Our starting point is the classical bandwidth formula for independent data, (2.5). This is of course not strictly valid for many types of dependent data, although we know from (2.16) that it is at least approximately correct for data that exhibit only moderately long-range dependence.

Our plug-in rule is as follows. Let $\hat{I}_j$ be an estimator of I_j for $j = 1$ and 2; details will be given in Section 3.2. In the formulae $J_1 = R(K)/(\mu_2^2 I_2)$ and $J_2 = \mu_4 I_3/(20\mu_2 I_2)$, replace the unknowns I_j by these quantities, thereby obtaining estimators $\hat{J}_j$. [Of course, $R(K)$ and μ_j are known constants, depending only on the kernel K.] Then our empirical bandwidth selector is

$$\hat{h} = \left(\hat{J}_1/n\right)^{1/5} + \hat{J}_2\left(\hat{J}_1/n\right)^{3/5}; \tag{3.1}$$

compare with (2.5).

We argued in Section 2 that for independent data, since $\hat{I}_j$ may be chosen to be root-n consistent for I_j, the rule at (3.1) can produce a bandwidth rule

that approximates h^* up to an error of order $n^{-1/2}$, in relative terms. The absolute error is of order $n^{-1/5}n^{-1/2} = n^{-7/10}$. In Section 3.2 we shall show that for dependent data that may be modelled as a symmetric function of a Gaussian process, as was assumed in Theorem 2.2, the estimator $\hat{I}_j$ can be chosen so that

$$|\hat{I}_j - I_j| = O_p\left[\left\{n^{-1}\sum_{i=1}^{n}\gamma(i)^2\right\}^{1/2}\right]. \tag{3.2}$$

See (3.8) below. It follows that the absolute difference between our empirical bandwidth selector and the deterministic choice, $h' = (J_1/n)^{1/5} + J_2(J_1/n)^{3/5}$, is given by

$$|\hat{h} - h'| = O_p\left[n^{-7/10}\left\{\sum_{i=1}^{n}\gamma(i)^2\right\}^{1/2}\right]. \tag{3.3}$$

Of course, in the context of short-range dependence, and also for long-range dependence with $\Sigma\gamma(i)^2 < \infty$, we have $|\hat{h} - h'| = O_p(n^{-7/10})$.

To appreciate the implications of (3.3) we temporarily assume, for the sake of simplicity, that $\gamma(i) \sim ci^{-\alpha}$ for some $\alpha > 0$ and $c \neq 0$. Then by (3.3) and (2.16), $|\hat{h} - h_0| = o_p(n^{-3/5})$ if $\frac{4}{5} < \alpha < \infty$, $|\hat{h} - h_0 + c_1 n^{(1/5)-\alpha}| = o_p(n^{(1/5)-\alpha})$ if $\frac{2}{5} < \alpha \leq \frac{4}{5}$, and $\hat{h}$ is of size $n^{-1/5}$ if $0 < \alpha \leq \frac{2}{5}$. Thus, the empirical bandwidth selector $\hat{h}$ defined at (3.1) is second-order accurate for the optimal bandwidth when α is in the range $\frac{4}{5} < \alpha < \infty$; agrees with the optimal bandwidth up to but not including second-order terms if $\frac{2}{5} < \alpha < \frac{4}{5}$; and is in the range strictly between $n^{-(1-\alpha)}$ and $n^{-\alpha/4}$, which minimizes mean integrated squared error to first order, if $0 < \alpha \leq \frac{2}{5}$. The first case is of particular interest, since $\frac{4}{5} < \alpha \leq 1$ includes long-range dependence. Thus, the admittedly crudely defined bandwidth selector introduced at (3.1) can produce second-order accuracy even under conditions of long-range dependence.

Nevertheless, it is not immediately clear that a good approximation to the bandwidth that minimizes M automatically produces approximate minimization of integrated squared error, $\Delta = \Delta(h) = \int(\hat{f} - f)^2$, even to first order. To clarify these issues we shall derive a bound to the expected value of integrated squared error when the bandwidth is given by (3.1). Of course, if h is nonrandom, then $E\Delta(h) = M(h)$. However, we are interested in properties of $E\Delta(h)$ when h is stochastic. In this context it is advisable to restrict the range of values that $\hat{h}$ can assume, to prevent it from being excessively small or large. So we define $\tilde{h}$ to equal $\hat{h}$ if $\eta n^{-1/5} \leq \hat{h} \leq \eta^{-1}n^{-1/5}$, and to equal $\xi n^{-1/5}$ otherwise, where η is a very small positive constant [smaller than $\min(J_1^{1/5}, J_1^{-1/5})$] and $\xi > 0$ is an arbitrary element of $[\eta, \eta^{-1}]$. We shall prove in Section 3.2 that if the process $\{X_i\}$ is a symmetric transformation of a Gaussian process (in particular, if $\{X_i\}$ is itself Gaussian), then, for a positive constant B,

$$|E\Delta(\tilde{h}) - M(h_0)| \leq B\left\{n^{-1}\sum_{i=1}^{n}\gamma(i)^2 + n^{-7/5}\sum_{i=1}^{n}|\gamma(i)|\right\}. \tag{3.4}$$

To clearly appreciate the significance of this result, consider as before the special case where $\gamma(i) \sim ci^{-\alpha}$, with $\alpha > 0$ and $c \neq 0$. Then, by (3.4),

$$|E\Delta(\tilde{h}) - M(h_0)| = O(\delta_n), \tag{3.5}$$

where

$$\delta_n = \begin{cases} n^{-1}, & \text{if } \frac{3}{5} \leq \alpha < \infty, \\ n^{-(2/5)-\alpha}, & \text{if } \frac{2}{5} \leq \alpha < \frac{3}{5}, \\ n^{-2\alpha}, & \text{if } 0 < \alpha < \frac{2}{5}. \end{cases}$$

This bound is best possible in at least the case $\frac{3}{5} \leq \alpha < \infty$.

Recall from (2.17) that $M(h_0)$ is of size $n^{-4/5}$ if $\alpha > \frac{4}{5}$, and $n^{-\alpha}$ if $\alpha \leq \frac{4}{5}$. Result (3.5) implies that the relative error $\rho = |E\Delta(\tilde{h})M(h_0)^{-1} - 1|$ is of order $n^{-1/5}$ if $\alpha > \frac{4}{5}$; $n^{\alpha-1}$ if $\frac{3}{5} \leq \alpha < \frac{4}{5}$; $n^{-2/5}$ if $\frac{2}{5} \leq \alpha < \frac{3}{5}$; and $n^{-\alpha}$ if $0 < \alpha < \frac{2}{5}$. Therefore the relative error can actually be smaller for very-long-range dependent data than it is for independent data, despite the fact that the bandwidth selection rule is founded on the assumption of independence.

When the data are obtained by an asymmetric transformation of a Gaussian process, formulae (3.3) and (3.4) remain valid provided that $\sum_{i=1}^{n} \gamma(i)^2$ is replaced by $\sum_{i=1}^{n} |\gamma(i)|$. In particular, when $\gamma(i) \sim ci^{-\alpha}$ we have $|\hat{h} - h_0'| = O_p(n^{-1/2})$ if $\alpha > 1$; $O_p\{n^{-1/2}(\log n)^{1/2}\}$ if $\alpha = 1$; and $O_p(n^{-1/5-\alpha/2})$ if $0 < \alpha < 1$. It follows as before that $\hat{h}$ is second-order accurate for h_0 when $\alpha > \frac{4}{5}$, first-order accurate when $\frac{2}{5} < \alpha < \frac{4}{5}$, and in the range that provides first-order minimization of mean integrated squared error if $0 < \alpha \leq \frac{2}{5}$. This means that the relative error, ρ, converges to zero when $\frac{4}{5} < \alpha < \infty$, but not necessarily otherwise. The results that we shall derive in Section 3.2 will show that if $0 < \alpha \leq \frac{4}{5}$, then expected integrated squared error $E\Delta(\tilde{h})$ is of the same order, $n^{-\alpha}$, as the minimum mean integrated squared error $M(h_0)$.

In the context of our use of the MISE-optimal bandwidth as the basis for our analysis, and our use of expected integrated squared error as a measure of performance, we should mention the controversy over whether integrated squared error or mean integrated squared error is the most appropriate benchmark for assessing the accuracy of a density estimator. It is not appropriate to join that debate in the context of density estimation with dependent data, since the issues affecting performance there are not yet widely appreciated. In particular, under very-long-range dependence the ratio $\Delta(h_0)/M(h_0)$ does not converge to 1 in probability; see the discussion in Section 4. However, readers interested in acquainting themselves with the discussion for independent data will find work of Scott (1988), Mammen (1990) and Jones (1991) to be illuminating.

We should also address a simpler plug-in bandwidth formula, obtained from (3.1) by deleting the second term:

$$\breve{h} = \left(\hat{J}_1/n\right)^{1/5}. \tag{3.6}$$

Properties of $\breve{h}$ are readily obtained from those of $\hat{h}$. Indeed, if we define $\breve{h}$ to equal $\breve{h}$ if $\eta n^{-1/5} \leq \breve{h} \leq \eta^{-1} n^{-1/5}$ and to equal $\xi n^{-1/5}$ otherwise (by analogy

with our earlier definition of $\tilde{h}$), then instead of (3.3) and (3.4) we have, for data that are a symmetric transform of a Gaussian process,

$$|\breve{h} - h'| = O_p\left[n^{-7/10}\left\{\sum_{i=1}^{n}\gamma(i)^2\right\}^{1/2} + n^{-3/5}\right], \tag{3.7}$$

and

$$|E\Delta(\breve{h}) - M(h_0)| \le B\left[n^{-1}\sum_{i=1}^{n}\gamma(i)^2 + n^{-9/10}\left\{\sum_{i=1}^{n}\gamma(i)^2\right\}^{1/2} + n^{-7/5}\sum_{i=1}^{n}|\gamma(i)|\right].$$

Because we have omitted the second-order term from our empirical bandwidth formula, we cannot expect $\breve{h}$ to be second-order correct even for independent data. Nevertheless, supposing for the sake of simplicity that $\gamma(i) \sim ci^{-\alpha}$, $\breve{h}$ is first-order correct in the sense that $E\{\Delta(\breve{h})\} \sim M(h_0)$ for each $\alpha > 0$.

3.2. *Technical details.* As in Theorem 2.1, we suppose that $X_i = a(Y_i)$, $i \ge 1$, where $\{Y_i\}$ is a stationary Gaussian process with zero mean and unit variance, and the function $a = b^{-1}$ is monotone and satisfies $E\{a(Y)\} = 0$, $E\{a(Y)^2\} = 1$,

$$|b^{(1)}(x)|^{-1} \le B_1\exp\{\varepsilon_1 b(x)^2\} \quad \text{and} \quad \sum_{j=1}^{20}|b^{(j)}| \le B_2\exp\{\varepsilon_2 b(x)^2\},$$

for some $B_1, B_2(\varepsilon) > 0$, some $0 < \varepsilon_1 < \frac{1}{2}\{1 + \sup|\gamma(i)|\}^{-1}$ and all $\varepsilon_2 > 0$. Of course, if the process $\{X_i\}$ is itself Gaussian, then we may take $a(x) \equiv x$. We further assume that K is compactly supported and twice differentiable and that K'' is Hölder continuous; and we let K_1 and h_1 denote a new, six times boundedly differentiable, compactly supported kernel function and a new bandwidth, respectively, such that K_1 is of order at least 4 [i.e., $\int K_1 = 1$ and $\int x^i K_1(x)\,dx = 0$ for $1 \le i \le 3$]. Collectively we call these conditions (C_3).

Write $\hat{f}_1(x) = (nh_1)^{-1}\sum_{i=1}^{n}K_1\{(x - X_i)/h_1\}$ for the corresponding density estimator, and set $\theta_{1j} = \int(E\hat{f}_1)f^{(j)}\,dx$ and $\theta_{2j} = \int(E\hat{f}_1^{(j)})^2$ for $j = 2$ or 3. Our respective estimators of θ_{1j}, θ_{2j} and I_j are given by

$$\hat{\theta}_{1j} = 2\{n(n-1)h_1^{2j+1}\}^{-1}\sum\sum_{1\le i_1<i_2\le n}K_1^{(2j)}\left(\frac{X_{i_1} - X_{i_2}}{h_1}\right),$$

$$\hat{\theta}_{2j} = 2\{n(n-1)h_1^{2(j+1)}\}^{-1}\sum\sum_{1\le i_1<i_2\le n}\int K_1^{(j)}\left(\frac{x - X_{i_1}}{h_1}\right)K_1^{(j)}\left(\frac{x - X_{i_2}}{h_1}\right)dx,$$

$$\hat{I}_j = 2\hat{\theta}_{1j} - \hat{\theta}_{2j}.$$

For independent data the estimator $\hat{\theta}_{kj}$ is unbiased for θ_{kj} whenever $j \in \{2, 3\}$ and $k \in \{1, 2\}$. This is one of the motivations for our approach to

estimating I_j. In the case of dependent data there is a degree of bias, but it is of smaller order than the error about the mean, as Theorem 3.1 will show.

Let $B_1, B_2, \ldots$ denote generic positive constants. Define the following:

$$v(h) = \begin{cases} n^{-1} \sum_{i=1}^{n} \gamma(i)^2 + n^{-1}h^8 \sum_{i=1}^{n} |\gamma(i)|, & \text{if } a \text{ is symmetric,} \\ n^{-1} \sum_{i=1}^{n} |\gamma(i)|, & \text{otherwise;} \end{cases}$$

$$v_n = \begin{cases} n^{-1} \sum_{i=1}^{n} \gamma(i)^2, & \text{if } a \text{ is symmetric,} \\ n^{-1} \sum_{i=1}^{n} |\gamma(i)|, & \text{otherwise.} \end{cases}$$

Note that if $h \le B_1 n^{-1/16}$, then $v(h) \le B_2 v_n$.

Our first theorem implies that, under the function-of-a-Gaussian-process model introduced in Section 2, and provided bandwidth is chosen in the range $B_1 n^{-1/(4j+1)} \le h_1 \le B_2 n^{-1/16}$, for some B_1, $B_2 > 0$, we have, for $m = 1$ and 2,

$$E\left\{\left(\hat{I}_j - I_j\right)^{2m}\right\} \le B_3 v_n. \tag{3.8}$$

THEOREM 3.1. *Under the above conditions, and assuming that* $n^{-1/(4j+1)} \le h_1 \le 1$, *there exist positive constants* B_4, B_5 *and* B_6 *such that, for each* $j \in \{2, 3\}$, $k \in \{1, 2\}$ *and* $m \in \{1, 2\}$,

$$|E\left(\hat{\theta}_{kj}\right) - \theta_{kj}| \le B_4 n^{-1} \sum_{i=1}^{n} |\gamma(i)|, \tag{3.9}$$

$$E\left\{\left|\hat{\theta}_{kj} - E\left(\hat{\theta}_{kj}\right)\right|^{2m}\right\} \le B_5 v(h_1)^m, \tag{3.10}$$

$$|2\theta_{1j} - \theta_{2j} - I_j| \le B_6 h_1^{2(7-j)}. \tag{3.11}$$

Finally, we derive (3.4), for which purpose we use the following result. Let $\mathcal{I}$ denote the interval $[\eta n^{-1/5}, \eta^{-1} n^{-1/5}]$, where $\eta \in (0, 1)$.

THEOREM 3.2. *Assuming conditions* (C_3) *and that* $n^{-1/(4j+1)} \le h_1 \le 1$,

$$E\{\Delta'(h)^2\} = O\{h^{-2} v(h)\}, \tag{3.12}$$

and, for any $\varepsilon > 0$,

$$E\left\{\sup_{h \in \mathcal{I}} \Delta''(h)^2\right\} = O\{n^{\varepsilon} h^{-4} v(h)\}. \tag{3.13}$$

The proof of Theorem 3.1 is given in Section 6, while that of Theorem 3.2 (which is relatively straightforward) may be found in a longer version of the paper obtainable from any one of the authors.

Observe that by Taylor expansion about h', valid if K has at least two bounded derivatives,

$$E\Delta(\tilde{h}) = M(h') + E\{(\tilde{h} - h')\Delta'(h')\} + \tfrac{1}{2}E\{(\tilde{h} - h')^2\Delta''(h^\dagger)\},$$

where $h^\dagger$ lies between $\tilde{h}$ and h', and hence between $\eta n^{-1/5}$ and $\eta^{-1}n^{-1/5}$. Therefore,

$$|E\Delta(\tilde{h}) - M(h')| \le \{E(\tilde{h} - h')^2 E\Delta'(h')^2\}^{1/2} + \{E(\tilde{h} - h')^4 E\Delta''(h^\dagger)^2\}^{1/2}.$$

In view of (3.8) and the definition of $\tilde{h}$,

$$E(\tilde{h} - h')^{2m} = O\{(n^{-2/5}v_n)^m\}.$$

By (3.12) and (3.13),

$$E\Delta'(h')^2 = O(n^{2/5}v_n), \qquad E\Delta''(h^\dagger)^4 = O(n^{\varepsilon+(4/5)}v_n).$$

Furthermore, by Theorems 2.1 and 2.2,

$$\begin{aligned} |M(h_0) - M(h')| &= O\left[n^{-7/5}\sum_{i=1}^{n}|\gamma(i)| + n^{-2}\left\{\sum_{i=1}^{n}|\gamma(i)|\right\}^2\right] \\ &= O\left\{n^{-7/5}\sum_{i=1}^{n}|\gamma(i)| + v_n\right\}. \end{aligned}$$

Therefore, for each $\varepsilon > 0$,

$$\begin{aligned} |E\Delta(\tilde{h}) - M(h_0)| &= O\Big\{(n^{-2/5}v_n)^{1/2}(n^{2/5}v_n)^{1/2} + (n^{-2/5}v_n)n^{\varepsilon}(n^{4/5}v_n)^{1/2} \\ &\qquad + n^{-7/5}\sum_{i=1}^{n}|\gamma(i)| + v_n\Big\} \\ &= O\left\{v_n + n^{\varepsilon}v_n^{3/2} + n^{-7/5}\sum_{i=1}^{n}|\gamma(i)|\right\}. \end{aligned}$$

Provided that $|\gamma(i)| \le Ci^{-\delta}$ for some C, $\delta > 0$ and all sufficiently large i, the formula just above implies that $n^{\varepsilon}v_n^{3/2} = o(v_n)$ if ε is sufficiently small. This proves (3.4).

4. Properties of least-squares cross-validation. The least-squares cross-validation algorithm was introduced by Rudemo (1982) and Bowman (1984). Its theoretical properties were described by Hall (1983) and Stone (1984), and its more general features have been discussed by Silverman [(1986), page 48ff]. Hart and Vieu (1990) proposed a variant that would seem to be more appropriate for dependent data. To define it, first set

$$\hat{f}_{-i}(x) = (n_l h)^{-1}\sum_{j:\,|i-j|>l} K\{(x - X_j)/h\},$$

for $1 \le i \le n$, where $l = l_n$ is a sequence of positive integers, called the *leave-out* sequence, and n_l is such that

$$nn_l = \#\{(i,j)\colon |i-j| > l \quad \text{and} \quad 1 \le i,j \le n\}.$$

Then, following Hart and Vieu (1990), define the "leave-out l cross-validation function" by

$$CV_l(h) = \int \hat{f}^2(x)\,dx - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(X_i).$$

Minimizing $CV_l(h)$ over a suitable interval gives the cross-validation bandwidth $\hat{h}$ for estimating f using $\hat{f}$.

Note that $l = 0$ corresponds to the usual cross-validation criterion for independent data. The motivation behind the definition of CV_l is that deleting l neighbouring data points reduces the dependence between the two sets of random variables $\{X_j\colon |i-j| > l\}$ and X_i that define the ith summand $\hat{f}_{-i}(X_i)$ in CV_l.

Generally, in the context of both independent and dependent data, the cross-validation criterion CV_l represents an approximation to $\Delta - \int f^2$, where $\Delta = \int(\hat{f} - f)^2$ denotes integrated squared error. Under independence, and assuming $l = 0$, this approximation is exact in the mean, in the sense that $E(CV_0) + \int f^2 \equiv E(\Delta)$. However, the equivalence is lost for dependent data, and indeed under very strong dependence it is not even valid in an approximate, relative sense, at the minimum, since the ratio

$$\frac{E\{CV_l(h_0)\} + \int f^2 - M(h_0)}{M(h_0)} \tag{4.1}$$

does not converge to zero as $n \to \infty$. [Recall from Section 2 that $M = E(\Delta)$ denotes mean integrated squared error.]

To appreciate why the ratio at (4.1) fails to converge to zero under very-long-range dependence, recall from Section 2 that in such circumstances M depends on bandwidth only to second order. First-order terms in formulae for MISE are determined entirely by the covariance structure of the process, and cross-validation fails to take adequate account of them. However, since those terms do not depend on bandwidth then this is not a fatal shortcoming, and in fact cross-validation can minimize MISE, to first order, over a wide range, as our next theorem shows.

Assume the following: the conditions of the function-of-a-Gaussian-process model in Section 2; that the transformation a is symmetric; that $\gamma(i) \sim ci^{-\alpha}$ as $i \to \infty$, for some $\alpha > 0$ and $c \ne 0$; and that $l = l(n)$ satisfies $0 \le l = o(n)$. Let $\mathscr{H}_n$ denote the interval $[n^{-c_2}, n^{-c_1}]$, where $\frac{1}{3} < c_1 < \frac{1}{5} < c_2 < 1$, and write $\tilde{h}$ for the bandwidth that minimizes CV_l over $\mathscr{H}_n$. The result below is a version of formula (3.4) for the cross-validation bandwidth $\tilde{h}$.

THEOREM 4.1. *Under the conditions above,*

$$|E\Delta(\tilde{h}) - M(h_0)| = o\{M(h_0)\}.$$

However, the ratio at (4.1) *converges to zero if and only if* $\alpha > \frac{4}{5}$.

We do not address the second-order accuracy of cross-validation, since even in the case of independent data this method does not have any second-order virtues from the viewpoint of MISE. Furthermore, cross-validation can be particularly difficult to implement with heavily dependent data and can produce bandwidths of very high variability, owing to the relative flatness of the function $CV_l(h)$; see Section 5. For these reasons we tend not to favour cross-validation as a bandwidth choice method for dependent data.

A number of variants of the theorem are possible. In particular, in the first part of the theorem the condition that $\gamma(i) \sim ci^{-\alpha}$ may be replaced by a milder one. Note that in the present form of the theorem, it is possible to choose c_1 sufficiently small and c_2 sufficiently large, within the constraints of the theorem, so that the interval $\mathscr{H}_n$ contains the bandwidth that asymptotically minimizes M. A proof of the theorem is very similar to that of the results in Section 3 and so is not given here.

It should be noted that, no matter what the value of l, the intuition which motivates cross-validation fails rather spectacularly in circumstances of very-long-range dependence. The main reason that cross-validation still works to first order, as evidenced by Theorem 4.1, is that under very-long-range dependence the choice of bandwidth is not particularly critical. However, when $\alpha \leq \frac{4}{5}$ neither CV_l nor $\Delta = \text{ISE}$ provides a good approximation to MISE. Indeed, if we define $R_n = -\int f^2 + 2n^{-1}\sum_{i=1}^n \{f(X_i) - Ef(X_i)\}$, then for $\alpha \leq \frac{4}{5}$ each of the ratios

$$\frac{E(\Delta - \text{MISE})^2}{(\text{MISE})^2},$$

$$\frac{E(CV_l - \text{MISE} - R_n)^2}{(\text{MISE})^2},$$

$$\frac{E(CV_l - \Delta - R_n)^2}{(\text{MISE})^2}$$

converges to a finite, positive number as $n \to \infty$. The limit is zero, for each ratio, when $\alpha > \frac{4}{5}$. In the boundary case where $\alpha = \frac{4}{5}$, CV_l approximates with sufficient accuracy that part of MISE that depends on h, and so Theorem 4.1 holds there.

5. Simulation. The poor performance of cross-validation, relative to plug-in rules that have been specifically constructed to enjoy high orders of accuracy, has been extensively documented in the context of independent data. [See, e.g., Park and Marron (1990) and Hall, Sheather, Jones and Marron (1991).] It is straightforward to reproduce those numerical results for short-range and moderately long-range dependent data, particularly in view of the theoretical conclusions described in Section 3. For very-long-range dependent data it is to be expected that the effect of bandwidth choice will be less apparent than in the context of independence, since the bandwidth

adjusts only high-order terms in an expansion of MISE; see Section 2. To make the numerical study more interesting in spite of these features, we chose to treat a relatively crude plug-in rule whose convergence rate to h_0 is similar to that of the cross-validated bandwidth selector, and to compare those two approaches. It turns out that the plug-in rule is still superior in the sense that it enjoys significantly less variability, but in other respects both approaches perform creditably.

Our plug-in bandwidth selector was identical to $\tilde{h}$, defined at (3.6), except that we took K_1 to be the standard Gaussian kernel. This was also our choice for K, in both plug-in and cross-validation algorithms. Numerical performance is very similar for compactly supported kernels, and indeed we set $K(x)$ equal to zero for large values of x.

A Gaussian process with a covariance structure $\gamma(i) \sim i^{-\alpha}$ for $0 < \alpha < 1$ was produced by simulating from the process $\{X_i\}$ defined by $(1 - B)^d X_i = Z_i$, where B is the backward-shift operator with $0 < d < \frac{1}{2}$, and Z_i are independent $N(0, 1)$ random variables. This is an ARIMA$(0, d, 0)$ process with fractional difference d, and it has been used to model long-memory time series [see Granger and Joyeux (1980) and Hosking (1981)]. One may show that the process $\{X_i\}$ is stationary with covariance function given by

$$\gamma(k) = E(X_i X_{i-k}) = 2\sin(d\pi)\frac{\Gamma(k + d)}{\Gamma(k + 1 - d)}\Gamma(1 - 2d) \sim ck^{-\alpha},$$

where $\alpha = 1 - 2d$. Thus, X_i has a Gaussian distribution with mean zero and variance $2\sin(d\pi)\Gamma(d)\Gamma(1 - 2d)/\Gamma(1 - d)$.

We simulated the process $\{X_i\}$ by following Haslett and Raftery (1989). One hundred independent replications were performed for each of the 24 combinations of methods (plug-in and least-squares cross-validation), sample sizes (n = 100, 200, 400) and α (= 0.2, 0.4, 0.6, 0.8). In the case of cross-validation estimates, we used l_n = 0, 5, 10, 15, 20, 25 for each replication.

The results of the simulation study are summarized in Tables 1 and 2, for n = 100 and 400. Each figure in the tables is obtained by averaging over 100 independent simulations; more complete data are available in a longer version of the paper obtainable from any one of the authors. The following features emerge from the data. First, even compared with our relatively error-prone plug-in rule the cross-validation approach produced an empirical bandwidth selector with substantially higher variance, by a factor of up to 12,

TABLE 1
Summary statistics for plug-in method

	Mean of $\hat{h}$	**SD of $\hat{h} \times 10^{-1}$**	**MISE $\times 10^{-3}$**
(n = 100, α = 0.2)	0.43453	1.09722	9.72028
(n = 100, α = 0.8)	0.38006	0.53682	2.80584
(n = 400, α = 0.2)	0.35928	0.44113	7.07293
(n = 400, α = 0.8)	0.29123	0.16861	1.01234

TABLE 2
Summary statistics for least-squares cross-validation

	l_n	Mean of $\hat{h}$	SD of $\hat{h} \times 10^{-1}$	MISE $\times 10^{-3}$
($n = 100$, $\alpha = 0.2$)	0	0.39549	1.01835	10.20120
	10	0.49195	0.93076	8.91394
	20	0.47989	1.02900	9.09588
($n = 100$, $\alpha = 0.8$)	0	0.37202	0.80889	2.96664
	10	0.48147	0.82437	2.35508
	20	0.45975	0.94874	2.47952
($n = 400$, $\alpha = 0.2$)	0	0.35480	0.69066	7.06990
	10	0.47394	0.90266	6.41401
	20	0.47538	0.93607	6.41099
($n = 400$, $\alpha = 0.8$)	0	0.30658	0.45706	0.92363
	10	0.36457	0.50652	0.83955
	20	0.36123	0.52814	0.84517

depending on circumstance. This reflects the flatness of the cross-validation function, depicted in Figure 1. The position of the minimum of that function is very prone to sampling fluctuations, and was particularly difficult to find in our numerical study. Second, varying the value of l in the cross-validation algorithm of Hart and Vieu (1990) had relatively little effect on either the mean value of $\hat{h}$ or on MISE, although it did influence variability of the algorithm. From that viewpoint, small values of l (indeed, the independent data prescription $l = 0$, even when the data exhibit very-long-range dependence) would seem to be advisable. Third, the relative insensitivity of MISE to bandwidth choice under long-range dependence, predicted by our theory in Section 2, is borne out by our results. For example, when $n = 400$ and $\alpha = 0.2$ the mean value of the cross-validated bandwidth $\hat{h}$ alters by 34%, and its variance by 86%, as l changes from 0 to 25. However, MISE alters by only 10% over that range.

6. Proof of Theorem 3.1.

PART 1. PROOF OF (3.9). Let $g_i(x_1, x_2) = f_i(x_1, x_2) - f(x_1)f(x_2)$ be as in Section 2. Observe that

$$E\bigl(\hat{\theta}_{1j}\bigr) - \theta_{1j} = (n-1)^{-1} \sum_{i=1}^{n} (1 - n^{-1}i) \times \iint K_1(x_1) g_i^{(2j,0)}(x_2 - h_1 x_1, x_2)\, dx_1\, dx_2,$$

$$E\bigl(\hat{\theta}_{2j}\bigr) - \theta_{2j} = (n-1)^{-1} \sum_{i=1}^{n} (1 - n^{-1}i) \iiint K_1(-x_1) K\bigl(-h^{-1}x - x_2\bigr) \times g_i^{(j,j)}(h_1 x_1 - x, h x_2)\, dx\, dx_1\, dx_2.$$

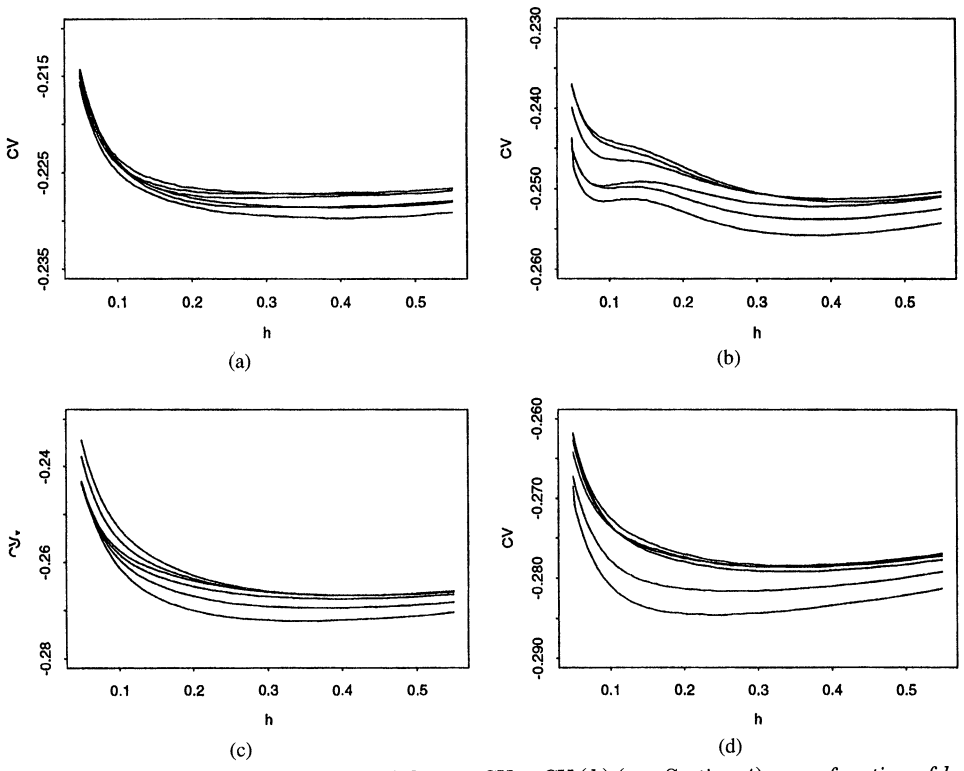

FIG. 1. *Graphs of leave-out l cross-validation, CV = $CV_1(h)$ (see Section 4), as a function of h, $0.05 \le h \le 0.55$, $l_n = 0$, 5, 10, 15, 20, 25, and α. Here the sample size is 400 and values of $CV_1(h)$ with $\alpha = 0.2$, 0.4, 0.6, 0.8 are plotted in* (a)–(d), *respectively.*

Since, for all $\varepsilon > 0$,

$$|g_i^{(2j,0)}(x_1, x_2)| + |g_i^{(j,j)}(x_1, x_2)|$$
$$\le B_1(\varepsilon)|\gamma(i)b'(x_1)b'(x_2)|\exp\Big[-\tfrac{1}{2}(1-\varepsilon)\big\{b(x_1)^2 + b(x_2)^2\big\}\Big],$$

then $|E(\hat{\theta}_{kj}) - \theta_{kj}| \le B_2 n^{-1}\Sigma|\gamma(i)|$, as had to be shown. □

PART 2. PROOF OF (3.10). We begin by describing the case $m = 1$. Observe that

$$(6.1)\quad \operatorname{var}\big(\hat{\theta}_{kj}\big) = \big\{2/n(n-1)h_1^{2j+k}\big\}^{-2}\Big(\sum^{(2)} + \sum^{(3)} + \sum^{(4)}\Big)A_{kj}(\mathbf{i}),$$

where $\Sigma^{(l)}$ denotes summation over indices $\mathbf{i} = (i_1, \ldots, i_4)$ such that $1 \le i_1 <$

$i_2 \le n$, $1 \le i_3 < i_4 \le n$ and just l of $i_1, \ldots, i_4$ are distinct; and

$$A_{1j}(\mathbf{i}) = \operatorname{cov}\left\{K_1^{(2j)}\left(\frac{X_{i_1} - X_{i_2}}{h_1}\right), K_1^{(2j)}\left(\frac{X_{i_3} - X_{i_4}}{h_1}\right)\right\}$$

$$A_{2j}(\mathbf{i}) = \iint \operatorname{cov}\left\{K_1^{(j)}\left(\frac{x - X_{i_1}}{h_1}\right) K_1^{(j)}\left(\frac{x - X_{i_2}}{h_1}\right), K_1^{(j)}\left(\frac{y - X_{i_3}}{h_1}\right) K_1^{(j)}\left(\frac{y - X_{i_4}}{h_1}\right)\right\} dx\, dy.$$

We claim that, for $l = 2$, 3 and 4,

$$\left(n^2 h_1^{2j+k}\right)^{-2} \sum{}^{(l)} |A_{kj}(\mathbf{i})| \le B_1 v(h_1), \tag{6.2}$$

whence follows (3.10). We shall prove (6.2) only in the case $l = 4$, since $l = 2$ and $l = 3$ may be treated similarly but more simply.

Let f, $f_{(i_1, i_2)}$ and $f_{(i_3, i_4)}$ denote the respective joint densities of $(X_{i_1}, \ldots, X_{i_4})$, (X_{i_1}, X_{i_2}) and (X_{i_3}, X_{i_4}), and set $g(x_1, \ldots, x_4) = f(x_1, \ldots, x_4) - f_{(i_1, i_2)}(x_1, x_2) \times f_{(i_3, i_4)}(x_3, x_4)$. In this notation,

$$\left(h_1^{2j+1}\right)^{-2} A_{1j}(\mathbf{i}) = \int \cdots \int K_1(x_1) K_1(x_3) \times g^{(2j, 0, 2j, 0)}(x_2 + h_1 x_1, x_2, x_4 + h_1 x_3, x_4)\, dx_1 \cdots dx_4.$$

Now Taylor-expand $g^{(2j, 0, 2j, 0)}(x_2 + h_1 x_1, x_2, x_4 + h_1 x_3, x_4)$ about (x_2, x_2, x_4, x_4) up to terms in h_1^8, obtaining

$$\left|\left(h_1^{2j+1}\right)^{-2} A_{1j}(\mathbf{i}) - \sum\sum_{0 \le k_1 + k_2 \le 3} c_{k_1 k_2} h_1^{2(k_1 + k_2)} \times \iint g^{(2j + 2k_1, 0, 2j + 2k_2, 0)}(x, x, y, y)\, dx\, dy\right| \tag{6.3}$$

$$\le B_2 h_1^8 \{|\gamma(i_1 - i_3)| + |\gamma(i_1 - i_4)| + |\gamma(i_2 - i_3)| + |\gamma(i_2 - i_4)|\},$$

where the $c_{k_1 k_2}$'s are constants. (For example, $c_{00} = 1$.) If the function a is symmetric, then so is b, whence it may be shown on Taylor-expanding the exponents of each of the functions $g = g^{(2j + 2k_1, 0, 2j + 2k_2, 0)}$ that the terms in $\gamma(i_1 - i_3)$, $\gamma(i_1 - i_4)$, $\gamma(i_2 - i_3)$ and $\gamma(i_2 - i_4)$ alone are all parts of quantities that integrate to zero. This proves the result that

$$\left|\iint g(x, y)\, dx\, dy\right| \le B_3\left\{\gamma(i_1 - i_3)^2 + \gamma(i_1 - i_4)^2 + \gamma(i_2 - i_3)^2 + \gamma(i_2 - i_4)^2\right\}.$$

On the other hand, if a is asymmetric, then the best bound we may obtain is

$$\left|\iint g(x, y)\, dx\, dy\right| \le B_3\{|\gamma(i_1 - i_3)| + |\gamma(i_1 - i_4)| + |\gamma(i_2 - i_3)| + |\gamma(i_2 - i_4)|\}.$$

In the former case we have, by (6.3),

$$
\begin{aligned}
&\left(h_1^{2j+1}\right)^{-2}|A_{1j}(\mathbf{i})| \\
&\quad \le B_4\Big[\gamma(i_1-i_3)^2+\gamma(i_1-i_4)^2+\gamma(i_2-i_3)^2+\gamma(i_2-i_4)^2 \\
&\qquad +h_1^8\{|\gamma(i_1-i_3)|+|\gamma(i_1-i_4)| \\
&\qquad\quad +|\gamma(i_2-i_3)|+|\gamma(i_2-i_4)|\}\Big]
\end{aligned} \tag{6.4}
$$

and, in the latter,

$$
\begin{aligned}
\left(h_1^{2j+1}\right)^{-2}|A_{1j}(\mathbf{i})| \le B_5\{&|\gamma(i_1-i_3)|+|\gamma(i_1-i_4)| \\
&+|\gamma(i_2-i_3)|+|\gamma(i_2-i_4)|\}.
\end{aligned} \tag{6.5}
$$

Results (6.4) and (6.5) readily give (6.2) for $k = 1$ and $l = 4$. The case $k = 2$ and $l = 4$ is similar, since

$$
\begin{aligned}
\left(h_1^{2j+2}\right)^{-2}A_{2j}(\mathbf{i}) = \int\cdots\int &K_1(x_1)K_1(x_2)K_1(x_3)K_1(x_4) \\
&\times g^{(j,j,j,j)}(x-h_1x_1, x-h_1x_2, y-h_1x_3, y-h_1x_4) \\
&\times dx_1\cdots dx_4\,dx\,dy.
\end{aligned}
$$

This completes our proof of (3.10) when $m = 1$. The case $m = 2$ is similar, but more complex notationally. Analogously to (6.1) we may write

$$
E\left(\hat\theta_{kj}-E\hat\theta_{kj}\right)^4 = \left\{2/n(n-1)h_1^{2j+k}\right\}^{-4}\left(\sum{}^{(4)}+\cdots+\sum{}^{(8)}\right)A_{kj}(\mathbf{i}),
$$

where $\mathbf{i} = (i_1,\ldots,i_8)$, $\Sigma^{(l)}$ denotes summation over $\mathbf{i}$ such that $1 \le i_{2j-1} < i_{2j} \le n$ for $j = 1,\ldots,4$ and just l of $i_1,\ldots,i_8$ are distinct, and $A_{kj}(\mathbf{i})$ is defined by formulae similar to those in the proof for $m = 1$, involving expectation in the joint distribution of $X_{i_1},\ldots,X_{i_8}$. We shall sketch the derivation of a bound to $\Sigma^{(l)}$, which determines the overall bound. Arguing as in the earlier proof we may show that, instead of (6.4),

$$
\begin{aligned}
\left(h_1^{2j+k}\right)^{-4}|A_{kj}(\mathbf{i})| \le B_6\Big\{&\sum{}^*\gamma(i^{(1)}-i^{(2)})^2\gamma(i^{(3)}-i^{(4)})^2 \\
&+h_1^8\sum{}^*\gamma(i^{(1)}-i^{(2)})^2|\gamma(i^{(3)}-i^{(4)})| \\
&+h_1^{16}\sum{}^*|\gamma(i^{(1)}-i^{(2)})\gamma(i^{(3)}-i^{(4)})|\Big\},
\end{aligned}
$$

where Σ^* denotes summation over quadruples $i_1,\ldots,i_4$ such that the $i^{(k)}$'s come from distinct pairs (i_{2j-1}, i_{2j}), $1 \le j \le 4$. Adding over $\mathbf{i}$ we deduce that, for $l = 8$,

$$
\left(n^2h_1^{2j+k}\right)^{-4}\sum{}^{(l)}|A_{kj}(\mathbf{i})| = O\left[n^{-2}\left\{\sum_{i=1}^n\gamma(i)^2+h_1^8\sum_{i=1}^n|\gamma(i)|\right\}^2\right].
$$

This bound is also valid for $l = 2,\ldots,7$, and so (3.10) holds for $m = 1$. □

PART 3. PROOF OF (3.11). Since K_1 is of order 4 or more, and the regularity conditions imposed on b are sufficient to give f six bounded, integrable derivatives, then

$$|2\theta_{1j} - \theta_{2j} - I_j| = \int \left(E\hat{f}_1^{(j)} - f^{(j)}\right)^2 \le B_6 h_1^{2(7-j)}. \qquad \square$$

Acknowledgments. Helpful discussions with Professor J. D. Hart are gratefully acknowledged. The constructive comments of three referees and an Associate Editor, who encouraged this shortened version of the original paper, were also particularly helpful.

REFERENCES

BERAN, J. (1992). Statistical methods for data with long-range dependence (with discussion.) *Statist. Sci* **7** 404–427.

BICKEL, P. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BOWMAN, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

CASTELLANA, J. V. (1989). Integrated consistency of smoothed probability density estimators for stationary sequences. *Stochastic Process. Appl.* **33** 335–346.

CASTELLANA, J. V. and LEADBETTER, M. R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.* **21** 179–193.

GRANGER, C. W. J. and JOYEUX, R. (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Anal.* **1** 15–29.

GYÖRFI, L. and MASRY, E. (1990). The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples. *IEEE Trans. Inform. Theory* **36** 531–539.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

HALL, P. and HART, J. D. (1990). Convergence rates for density estimation for data from infinite order moving average sequences. *Probab. Theory Related Fields* **87** 253–274.

HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 263–269.

HART, J. D. (1984). Efficiency of a kernel density estimator under an autoregressive dependence model. *J. Amer. Statist. Assoc.* **79** 110–117.

HART, J. D. (1987). Kernel smoothing when the observations are correlated. Technical Report 35, Dept. Statistics, Texas A & M Univ.

HART, J. D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B* **53** 173–187.

HART, J. D. and VIEU, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* **18** 873–890.

HASLETT, J. and RAFTERY, A. E. (1989). Space-time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussions). *J. Roy. Statist. Soc. Ser. C* **38** 1–50.

HOSKING, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176.

HURST, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* **116** 770–779.

JONES, M. C. (1991). The role of ISE and MISE in density estimation. *Statist. Probab. Lett.* **12** 51–56.

MAMMEN, E. (1990). A short note on optimal bandwidth selection for kernel estimators. *Statist. Probab. Lett.* **9** 23–25.

MELOCHE, J. (1990). Asymptotic behaviour of the mean integrated squared error or kernel density estimators for dependent observations. *Canad. J. Statist.* **18** 205–211.

NGUYEN, H. T. (1979). Density estimation in a continuous-time stationary Markov process. *Ann. Statist.* **7** 341–348.

PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

ROSENBLATT, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 199–210. Cambridge Univ. Press.

ROUSSAS, G. G. (1969). Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.* **21** 73–87.

ROUSSAS, G. G. (1988). Nonparametric estimation in mixing sequences of random variables. *J. Statist. Plann. Inference* **18** 135–139.

ROUSSAS, G. G. (1990a). Asymptotic normality of the kernel estimate under dependence conditions: application to hazard rate. *J. Statist. Plann. Inference* **25** 81–104.

ROUSSAS, G. G. (1990b). Exact rates of almost sure convergence of a recursive estimate of a probability density function: application to regression and hazard rate estimation. *Journal of Nonparametric Statistics* **1** 171–195.

ROUSSAS, G. G. (1991). Kernel estimates under association: strong uniform consistency. *Statist. Probab. Lett.* **12** 393–403.

ROUSSAS, G. G. and IOANNIDES, D. (1987). Note on the uniform convergence of density estimates for mixing random variables. *Statist. Probab. Lett.* **5** 179–285.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SCOTT, D. W. (1988). Discussion of "How far are automatically chosen regression smoothing parameters from their optimum?" by W. Härdle, P. Hall and J. S. Marron. *J. Amer. Statist. Assoc.* **83** 96–98.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

TRAN, L. T. (1989). The L_1 performance of kernel density estimates under dependence. *Canad. J. Statist.* **17** 197–208.

TRAN, L. T. (1990a). Kernel density estimation under dependence. *Statist. Probab. Lett.* **10** 193–201.

TRAN, L. T. (1990b). Kernel density estimation on random fields. *J. Multivariate Anal.* **34** 37–53.

YAKOWITZ, S. (1985). Nonparametric density estimation, prediction and regression for Markov sequences. *J. Amer. Statist. Assoc.* **80** 215–221.

YAKOWITZ, S. (1989). Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivariate Anal.* **30** 124–136.

YU, B. (1993). Density estimation in the L^∞ norm for dependent data, with applications to the Gibbs sampler. *Ann. Statist.* **21** 711–735.

PETER HALL
CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 0200
AUSTRALIA

SOUMENDRA NATH LAHIRI
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 50011-0001

YOUNG K. TRUONG
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-7400