The Annals of Statistics
1994, Vol. 22, No. 4, 2115–2134

ON THE NONPARAMETRIC ESTIMATION OF COVARIANCE FUNCTIONS

BY PETER HALL, NICHOLAS I. FISHER AND BRANKA HOFFMANN

CSIRO and Australian National University, CSIRO, and CSIRO

We describe kernel methods for estimating the covariance function of a stationary stochastic process, and show how to ensure that the estimator has the positive semidefiniteness property. From a practical viewpoint, our method is significant because it does not demand a parametric model for covariance. From a technical angle, our results exhibit a striking departure from those in more familiar cases of kernel estimation. For example, in the context of covariance estimation, kernel estimators can have the same convergence rates as maximum likelihood estimators, and can have exceptionally fast convergence rates when employed to estimate variance.

1. Introduction. Covariance or variogram estimation is a fundamental problem in inference for stationary stochastic processes, having wide-ranging applications in areas such as ore reserve estimation and hydrosciences [e.g., Matheron (1971), Journel and Huijbregts (1978) and Christakos (1984)]. In order to preserve the property of positive semidefiniteness enjoyed by a true covariance function, statisticians commonly resort to fitting parametric models. However, such an approach can be inadequate, particularly in view of the difficulty of testing goodness of fit with dependent data [e.g., Armstrong and Diamond (1984) and Christakos (1984)]. In the present paper we propose nonparametric estimators of covariance. Our basic estimators are constructed using kernel methods. We suggest modifications of those estimators, which enjoy the positive semidefiniteness property but retain the flexibility of kernel methods.

Recently, Shapiro and Botha (1991) suggested a method based on constrained curve fitting through point estimates of the covariance function. Shapiro and Botha's method does not produce a smooth covariance function which is positive semidefinite in the continuum, although it does have that property on a discrete set. We argue that our approach is simpler, in that it does not require numerical optimization under nonlinear constraints. Furthermore, its theoretical properties are significantly more transparent. This is important, because it is unclear what convergence rates are enjoyed by Shapiro and Botha's estimators, and what methods of construction are necessary to achieve them. The reader is referred to Cressie [(1991), Section 2.6, page 90ff], for a detailed account of variogram estimation.

Sampson and Guttorp (1992) have recently proposed a nonparametric esti-

Received September 1992; revised November 1993.
AMS 1991 *subject classifications*. Primary 62G05; secondary 62G10.
Key words and phrases. Convergence rate, correlation, covariance, positive semidefinite, stochastic process, kernel, variance, variogram.

mator for variance in the case of spatial data. Their approach is very different from ours, being based on a covariance representation due to Schoenberg (1938). It would be applicable to one dimension, although the convergence rates of their technique, for either one or two dimensions, remain to be determined. Related work on estimation of second-order properties of spatial point processes may be found in Diggle, Gates and Stibbard (1987) and Berman and Diggle (1989).

Our methods and main theoretical results may be described as follows. Assume that the stationary process X is observed at "time" points $t_1, \ldots, t_n$. The aim is to estimate $\rho(t) = \operatorname{cov}\{X(s), X(s+t)\}$, for general t and without structural or parametric assumptions about ρ. The t_i's are not necessarily evenly spaced, and may, for example, represent values of independent random variables. Put $\overline{X} = n^{-1}\sum X(t_i)$, $\widehat{X}_{ij} = \{X(t_i) - \overline{X}\}\{X(t_j) - \overline{X}\}$ and $t_{ij} = t_i - t_j$. Let K denote a kernel function, which we take to be a symmetric probability density. Let h represent bandwidth, and define

$$\widehat{\rho}(t) = \left[\sum_i \sum_j \widehat{X}_{ij} K\{(t - t_{ij})/h\}\right] \left[\sum_i \sum_j K\{(t - t_{ij})/h\}\right]^{-1}, \tag{1.1}$$

where the two double series may either include or exclude the diagonal terms corresponding to $i = j$. Then $\widehat{\rho}$ is a kernel estimator of ρ.

Of course, $\widehat{\rho}$ is not necessarily itself a covariance function, since it typically lacks the positive semidefiniteness property,

$$\int \int \rho(s - t) w(s) w(t)\, ds\, dt \geq 0 \quad \text{for all integrable functions } w.$$

By Bochner's theorem, this is equivalent to nonnegativity of the Fourier transform of ρ (i.e., of the spectrum),

$$\rho^\dagger(\theta) \geq 0 \quad \text{for all } \theta,$$

where

$$\rho^\dagger(\theta) = \int_{-\infty}^{\infty} \rho(t) e^{i\theta t}\, dt = 2 \int_0^{\infty} \rho(t) \cos(\theta t)\, dt.$$

(Throughout, we use $g^\dagger$ to denote the Fourier transform of a function g.) If the time points t_i are regularly spaced on a grid, then one way of estimating ρ is via Fourier-inversion of the periodogram. This would always produce a covariance function estimate satisfying the positive definiteness property *on the grid points*. However, the problem then arises of smoothing the estimator in such a way that it is positive semidefinite in the continuum. Furthermore, this method is not available when the t_i's are not regularly spaced. We suggest instead the following approach. First, compute the Fourier transform, $\widehat{\rho}^\dagger$, of $\widehat{\rho}$, perhaps after truncation of the latter to ensure integrability. Next, render $\widehat{\rho}^\dagger$ nonnegative, for example by deleting any negative lobes and perhaps doing a little additional smoothing. Let the resulting function be represented by $\widetilde{\rho}^\dagger$.

Finally, Fourier-invert $\widetilde{\rho}^{\dagger}$ to obtain a new function $\widetilde{\rho}$ which, by construction, is guaranteed to enjoy the positive semidefiniteness property.

In our technical analysis of this procedure, we first develop a theoretical model generating the time points t_i. To ensure statistical consistency in estimation of ρ, it is usually necessary for the interval between the smallest and largest t_i's to expand as sample size, n, increases. Indeed, for many parametric problems the amount of information in the sample about ρ is roughly proportional to any interquantile range of the collection of t_i's; see Remark 3.9. On the other hand, if the range increases at the same rate as n, or at a faster rate, then it is often not possible to ensure that a large number of differences $t_i - t_j$ exist with the property $t_i - t_j \simeq t$, for any given t. The latter property is essential for consistent estimation of ρ in a nonparametric, structure-free context. For example, it fails if $t_i \equiv i$ (where the range of the t_i's increases like n); here it is not possible to get a consistent nonparametric estimator of ρ.

For these reasons we assume that the interval to which the t_i's are confined increases like λ, and that $\lambda = \lambda(n)$ diverges so slowly that $\lambda/n \to 0$. Then in many parametric settings the amount of information about ρ in the collection $\{X(t_1), \ldots, X(t_n)\}$ increases roughly in proportion to λ, and the convergence rate of maximum likelihood estimators of ρ is typically $\lambda^{-1/2}$. We claim that our *nonparametric* estimator $\widehat{\rho}$ also enjoys this convergence rate, provided the bandwidth h is chosen appropriately. Furthermore, the bias of an optimally constructed version of $\widehat{\rho}$ is typically of smaller order than the error about the mean. Thus, the effect of smoothing is much less apparent here than in more classical curve estimation problems, where the convergence rate is typically slower than in parametric contexts and where the optimal estimator usually has bias and error-about-the-mean of the same order of magnitude.

Another point of departure from the classical case is that there is a range of orders of magnitude of bandwidth for which asymptotic optimality is attained. (In the case of curve estimation using independent and identically distributed data, the order of magnitude of h, and even the constant multiplying that order, are uniquely determined by the requirement of asymptotic optimality.) This feature is particularly fortuitous, since the construction of a data-driven rule for bandwidth selection is difficult even in the case of the basic kernel estimator $\widehat{\rho}$, let alone for the transformed version $\widetilde{\rho}$.

An additional feature of our estimator $\widehat{\rho}(t)$ is that in the special case $t = 0$, it often converges at the particularly fast rate of $o(\lambda^{-1/2})$. Of course, $\rho(0) = \sigma^2 = \mathrm{var}\{X(s)\}$, and so we are claiming that the kernel estimator $\widehat{\rho}(0)$ can converge at a faster rate than the more common variance estimator

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \{X(t_i) - \overline{X}\}^2,$$

which typically satisfies a central limit theorem of the form $\lambda^{1/2}(\widehat{\sigma}^2 - \sigma^2) \to N(0, \tau^2)$ in distribution, where $0 < \tau^2 < \infty$. The reason for this "superefficiency" property of $\widehat{\rho}(0)$ is that the latter estimator uses not just information in the diagonal terms $\widehat{X}_{ii} = \{X(t_i) - \overline{X}\}^2$, which form the sole basis for $\widehat{\sigma}^2$, but

also information in the off-diagonal terms X_{ij} (for $i \neq j$). Since there are a lot more off-diagonal terms than diagonal terms [$n(n-1)$ rather than n], then there is potential for improving performance by including the off-diagonal quantities.

Section 2 briefly describes numerical properties of our estimator. Section 3 outlines theoretical results which imply the convergence rates discussed above, and also discusses other aspects of the general problem. Outlines of proofs are given in Section 4. A practical application of the method will appear elsewhere.

2. Numerical results. We conducted a simulation study to evaluate the performance of the proposed method for a variety of covariance functions. The results for one such function are presented in this section; they are typical of those obtained in the larger study. For all the results presented here, we constructed the estimator $\widehat{\rho}(t)$, $t > 0$, using the definition at (1.1) and with diagonal terms included. This choice is recommended by the theory in Section 3. To calculate the Fourier transform $\widehat{\rho}^{\dagger}$, we truncated $\widehat{\rho}$ at a point $T_1 > 0$, and then brought the curve down linearly to 0 at a point T_2, say. An example of this truncation and linear smoothing is illustrated in panel (b) of Figure 2. Software was written that would allow the truncation and smoothing to be performed interactively.

The truncation and smoothing produces a function that we may write as

$$\widehat{\rho}_1(t) = \begin{cases} \widehat{\rho}(t), & 0 \le t \le T_1, \\ \widehat{\rho}(T_1)(T_2 - t)(T_2 - T)^{-1}, & T_1 < t \le T_2, \\ 0, & t > T_2. \end{cases}$$

In a slight abuse of notation, put

$$\widehat{\rho}^{\dagger}(\theta) = 2 \int_0^{\infty} \widehat{\rho}_1(t) \cos(\theta t)\, dt,$$

and define

$$\widehat{\theta} = \inf\{\theta > 0 \colon \widehat{\rho}^{\dagger}(\theta) \ge 0\},$$

$$\widetilde{\rho}(t) = (2\pi)^{-1} \int_{-\widehat{\theta}}^{\widehat{\theta}} \widehat{\rho}^{\dagger}(\theta)\, d\theta.$$

Then $\widetilde{\rho}$ is our final estimator of ρ.

We applied this technique to data generated from the stationary Gaussian process having zero mean and covariance function

$$\rho(t) = t^{-1} \sin t.$$

Note particularly that ρ is not compactly supported. A quartic function $0.9375(1 - x^2)^2$ was used for the kernel $K(x)$. So as to simplify and abbreviate this section, we take $h = 0.01$ throughout, and define $t_i = \lambda u_i$, where the u_i's are generated from the uniform distribution on the interval $(0, 1)$, $\lambda = 20, 100$ and $1 \le i \le n = 50, 250$.

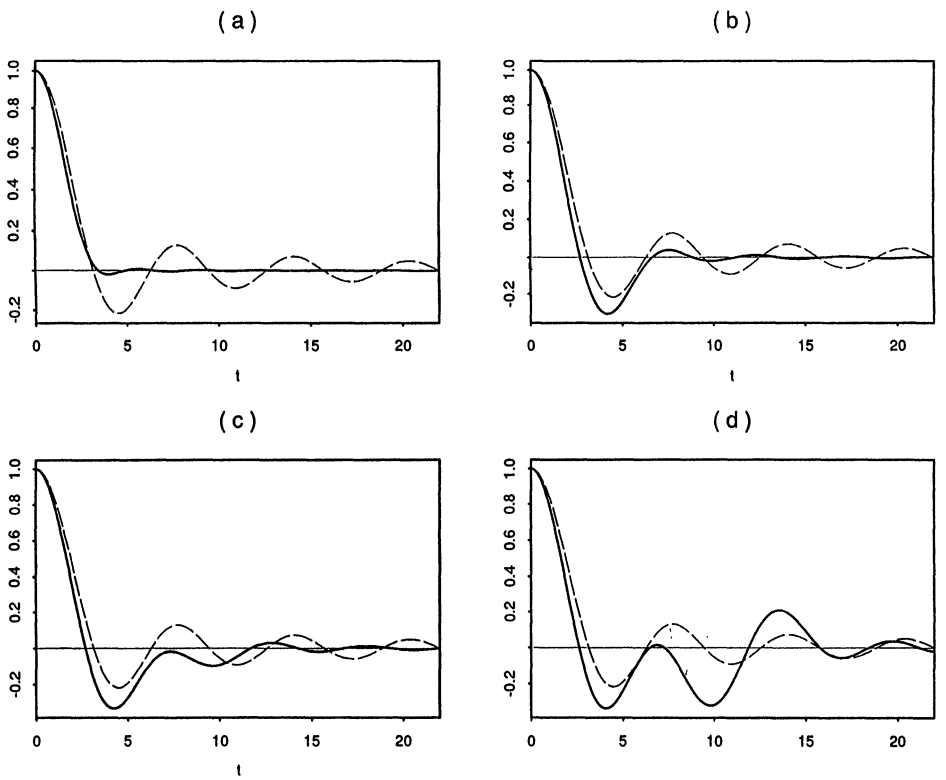

FIG. 1. *Estimator $\tilde{\rho}$ for four different trunctions (T_1, T_2). The same data set is used throughout.* (a) $T_1 = 2.4185$, $T_2 = 2.9641$; (b) $T_1 = 5.0553$, $T_2 = 5.5402$; (c) $T_1 = 9.2984$, $T_2 = 9.4499$; (d) $T_1 = 14.2385$, $T_2 = 14.4203$.

Figure 1 illustrates typical traces of the estimator $\tilde{\rho}$ for four different truncations. Note the way in which the fidelity of $\tilde{\rho}$ to the true ρ at first improves, and then deteriorates, as the truncation point is increased. Figure 2 illustrates four steps in the construction of the second panel in Figure 1: first, the data set; second, the estimator $\hat{\rho}$, with the truncation marked; third, the function $\hat{\rho}^{\dagger}$, with $\hat{\theta}$ marked; and finally, the estimator $\tilde{\rho}$ (shown here as an estimate of the correlation function, for purposes of comparison with the true function). Note particularly that imposing the constraint of positive definiteness produces a substantially smoother estimator—compare panels (b) and (d) of Figure 2.

3. Asymptotic theory. We assume that $X = X(t)$ is a stationary stochastic process, observed at "time" points $t_1, \ldots, t_n$. If the t_i's are confined to a fixed interval, then we cannot necessarily estimate characteristics of the process consistently. For example, consistent estimation of the mean, $\mu = E\{X(t)\}$, demands that the process be observed over an increasingly wide range of time points. To model this situation, we assume that $t_i = \lambda u_i$, where λ is an increasingly large

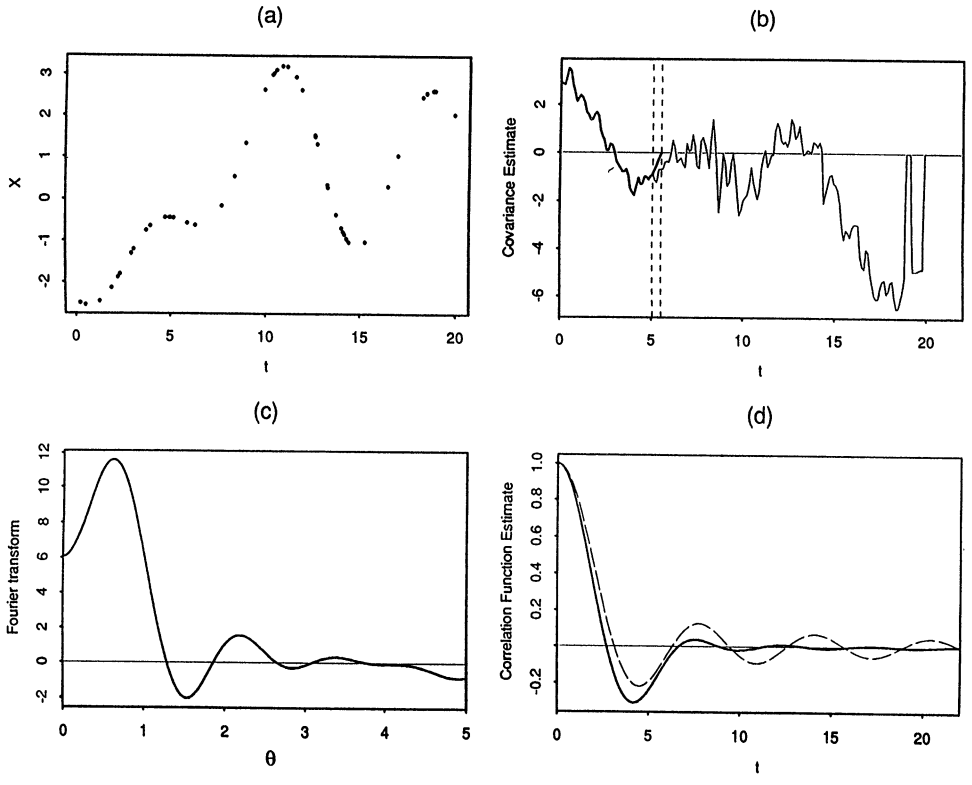

FIG. 2. *Four steps in the construction of Figure* 1(*b*). (*a*) *Data set* ($n = 50$, $\lambda = 20$); (*b*) *estimator* ρ; (*c*) *Fourier transform* $\hat{\rho}^{\dagger}$; (*d*) *final estimator* $\tilde{\rho}$ [*shown as* $\tilde{\rho}(t)/\tilde{\rho}(0)$; *dashed line is true* $\rho(t)/\rho(0)$].

positive number [i.e., $\lambda = \lambda(n) \to \infty$ as $n \to \infty$], and $u_1, \ldots, u_n$ represent observed values of independent random variables $U_1, \ldots, U_n$, all having the same distributions (not depending on n) and all independent of the process X. Alternatively, we may take $t_1, \ldots, t_n$ to be regularly spaced on an interval of width λ. Strictly speaking, we should write t_i as t_{ni}, to indicate that the time points are chosen differently for each n.

We impose the following regularity conditions. The densities f_1, f_2, f_3 of $U_1 - U_2, (U_1 - U_2, U_3 - U_2), (U_1 - U_2, U_3 - U_4, U_2 - U_4)$, respectively, are bounded and continuous, and are such that none of $f_1(0), f_2(0,0), f_3(0,0,0)$ is 0; $\int |\rho| < \infty$; the function

$$g(u,v,w) = E\Big[\{X(u+w) - \mu\}\{X(v) - \mu\}\{X(w) - \mu\}\{X(0) - \mu\}\Big] - \rho(u)\rho(v)$$

satisfies

$$\sup_{u,v} \int |g(u,v,w)|\, dw < \infty;$$

the kernel K is a bounded, piecewise continuous, compactly supported, symmet-

ric probability density with $K(0) \neq 0$; and for some $\eta > 0$, $n^{\eta}h \to 0$, $n^{1-\eta}h \to \infty$, $n^{2-\eta}h^3\lambda^{-1} \to \infty$, $n^{-\eta}\lambda \to \infty$, $n^{1-\eta}\lambda^{-1} \to \infty$; ρ has two bounded, continuous derivatives in a neighbourhood of t.

If the t_i's are equally spaced in an interval of length λ, then the functions f_1, f_2, f_3 should be taken to have the form they would if $U_1, \ldots, U_4$ we uniformly distributed on the interval $(0, 1)$.

The condition $\sup_{u,v} \int |g|\, dw < \infty$ holds if X is a smooth, polynomial function of a Gaussian process with integrable covariance. In particular, it is valid if $X = p(Y)$ where p is a polynomial and Y is a Gaussian process with covariance γ satisfying $\int |\gamma| < \infty$. If X is itself Gaussian, then $g(u, v, w) = \rho(u + w - v)\rho(w) + \rho(u + w)\rho(v - w)$, whence it follows that

$$\sup_{u,v} \int |g(u, v, w)| \leq 2\rho(0) \int |\rho(w)|\, dw.$$

Define

$$p_0 = \tfrac{1}{4}\rho''(t)^2, \qquad p_1 = \tfrac{1}{4}c^2\{cf_1(0) + K(0)\}^{-2} f(0)^2 \rho''(t)^2,$$

$$p = \begin{cases} p_0, & \text{if } t \neq 0 \text{ or if } t = 0 \text{ and } nh\lambda^{-1} \to \infty, \\ p_1, & \text{if } t = 0 \text{ and } nh\lambda^{-1} \to c,\ 0 < c < \infty, \end{cases}$$

$$q_0 = f_1(0)^{-2} f_3(0, 0, 0) \int g(t, t, u)\, du,$$

$$q_1 = c\{cf_1(0) + K(0)\}^{-2}\{cf_3(0, 0, 0) + 2K(0) f_2(0, 0)\} \int g(0, 0, u)\, du,$$

$$q = \begin{cases} q_0, & \text{if } t \neq 0 \text{ or if } t = 0 \text{ and } nh\lambda^{-1} \to \infty, \\ q_1, & \text{if } t = 0 \text{ and } nh\lambda^{-1} \to c,\ 0 < c < \infty. \end{cases}$$

Under the above regularity conditions, the sample mean $\overline{X}$ converges to the population mean μ at rate $\lambda^{-1/2}$. Indeed, $E(\overline{X}) = \mu$ and

$$\operatorname{var}(\overline{X}) = \lambda^{-1} f_1(0) \int \rho + o(\lambda^{-1}).$$

[See Step (iii) in the Appendix for a proof, and Remark 3.7 below for further details about estimation of μ.] Our next theorem shows that $\widehat{\rho}(t)$ converges to $\rho(t)$ at least as fast as $\overline{X}$ converges to μ, provided the bandwidth is chosen correctly.

THEOREM 3.1. *Assume the conditions stated above. Suppose first that the estimator $\widehat{\rho}(t)$ is defined with the $i = j$ terms included. If $t \neq 0$, or if $t = 0$ and $nh\lambda^{-1} \to c$ where $0 < c \leq \infty$, then*

$$E\{\widehat{\rho}(t) - \rho(t)\}^2 = h^4 p + \lambda^{-1} q + o(h^4 + \lambda^{-1}). \tag{3.1}$$

If $t = 0$ and $nh\lambda^{-1} \to 0$, then

$$E\{\widehat{\rho}(t) - \rho(t)\}^2 = o(h^4 + \lambda^{-1}). \tag{3.2}$$

Next, suppose that $\hat{\rho}(t)$ is defined with the $i = j$ terms excluded. Then, for both $t = 0$ and $t \neq 0$,

$$E\{\hat{\rho}(t) - \rho(t)\}^2 = h^4 p_0 + \lambda^{-1} p_0 + o(h^4 + \lambda^{-1}). \tag{3.3}$$

The remarks below elucidate consequences of the theorem. In particular, Remark 3.1 demonstrates that in many circumstances our estimator $\hat{\rho}$ achieves a convergence rate of $\lambda^{-1/2}$ (in L^1) under nonparametric assumptions on ρ; and Remark 3.9 shows that this rate is optimal in many problems, even when ρ is known parametrically. Note particularly that in many circumstances, optimal choice of bandwidth produces an estimator which is asymptotically unbiased. This is quite unusual in curve estimation problems, where asymptotic bias is typically of the same size as error about the mean.

REMARK 3.1. In formulae (3.1) and (3.3), the first term represents the contribution from squared bias and the second denotes the contribution from variance. By choosing bandwidth h so that $h^4 = o(\lambda^{-1})$, we may ensure that the bias contribution is asymptotically negligible. The estimator $\hat{\rho}(t)$ is then asymptotically unbiased, with

$$E\{\hat{\rho}(t) - \rho(t)\}^2 \sim \operatorname{var}\{\hat{\rho}(t)\} \sim \lambda^{-1} q.$$

Such a choice of λ is compatible with the conditions of the theorem. Indeed, if $n^{-\eta}\lambda \to \infty$ and $n^{1-\eta}\lambda^{-7/8} \to \infty$ for some $\eta > 0$, then we may choose $h = h(n)$ such that for some $\eta' \to 0$, we have $n^{\eta'} h \to 0$, $n^{1-\eta'} h \to \infty$, $n^{2-\eta'} h^3 \lambda^{-1} \to \infty$ and $\lambda h^4 \to 0$.

REMARK 3.2. In formula (3.2), the exact convergence rate (when $nh\lambda^{-1} \to 0$) is

$$n^2 h^6 \lambda^{-2} + nh\lambda^{-2} + n^{-1} = (nh\lambda^{-1})^2 h^4 + (nh\lambda^{-1})\lambda^{-1} + n^{-1} = o(h^4 + \lambda^{-1}).$$

If $n^{-1/2-\eta}\lambda \to \infty$ and $n^{1-\eta}\lambda^{-1} \to \infty$ for some $\eta > 0$, then we may choose $h = h(n)$ such that for some $\eta' > 0$, we have $n^{\eta'} h \to 0$, $n^{1-\eta'} h \to \infty$, $n^{2-\eta'} h^3 \lambda^{-1} \to 0$, $nh\lambda^{-1} \to 0$ and $\lambda h^4 \to 0$. Therefore, when λ diverges to ∞ at a rate between $n^{1/2}$ and n, it is possible to choose h such that $\hat{\rho}$ is "superefficient" at $t = 0$, in the sense that $E\{\hat{\rho}(0) - \rho(0)\}^2 = o(\lambda^{-1})$. That is, $\hat{\rho}(0)$ converges to $\rho(0)$ at rate $o(\lambda^{-1/2})$.

REMARK 3.3. An alternative estimator of $\rho(0) = \operatorname{var}\{X(t)\}$ is

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \{X(t_i) - \bar{X}\}^2.$$

This estimator converges only at rate $\lambda^{-1/2}$, and does not have the potential for the speed of convergence to be improved to $o(\lambda^{-1/2})$. It is of particular interest that, by including diagonal terms and doing an appropriate amount of smoothing, we may construct an estimator which improves on $\hat{\sigma}^2$.

REMARK 3.4. Asymptotic normality of $\widehat{\rho}$ may be proved under a variety of regularity conditions. For example, if $X = p(Y)$ where p is a polynomial and Y is a stationary Gaussian process whose covariance function γ satisfies $|\gamma(t)| \leq C_1 \exp(-C_2|t|)$ for some $C_1, C_2 > 0$ and all t, then the method of moments may be used to prove that $\widehat{\rho} - E\,\widehat{\rho}$ is asymptotically normal $N(0, \operatorname{var}\widehat{\rho})$.

REMARK 3.5. The condition $n^{2-\eta}h^3\lambda^{-1} \to \infty$, introduced prior to the statement of the theorem, may seem unusual. However, it is close to being necessary, and indeed the theorem may fail if $n^2h^3\lambda^{-1}$ is bounded. This point is elucidated in Step (ii) of our proof of the theorem.

REMARK 3.6. An alternative approach is to estimate the variogram,

$$\gamma(t) = 2\{\rho(0) - \rho(t)\},$$

instead of the correlation function. Since

$$Y_{ij} = \{X(t_i) - X(t_j)\}^2$$

is unbiased for $\gamma(t_i - t_j)$, then a kernel estimator of the variogram is given by

$$\widehat{\gamma}(t) = \left[\sum_i \sum_j Y_{ij} K\{(t - t_{ij})/h\}\right]\left[\sum_i \sum_j K\{(t - t_{ij})/h\}\right]^{-1},$$

where the $i = j$ terms may be either included or excluded. Asymptotic theory may be developed for $\widehat{\gamma}$, and is very similar to that for estimation of $\rho(t)$. In particular, when the $i = j$ terms are included in the estimator $\widehat{\gamma}(t)$, "superefficient" estimation of $\gamma(0) = 0$ is possible, although, of course, estimating 0 is not of such interest as estimating $\rho(0)$! Furthermore, asymptotically unbiassed estimators of $\gamma(t)$, with variance of size λ^{-1}, may be developed in the case $t \neq 0$.

REMARK 3.7. If the process X is Gaussian, then, although $\overline{X}$ is seldom equal to the maximum likelihood estimator of μ, $\overline{X}$ is sometimes asymptotically optimal, in the sense that the amount of information about μ in the sample $\mathcal{X} = \{X(t_1), \ldots, X(t_n)\}$ is asymptotic to $(\operatorname{var}\overline{X})^{-1}$. For example, take $t_i = \lambda_i/n$, suppose ρ is known, and define $X_i = X(t_i)$, $\sigma_{ij} = \rho(t_i - t_j)$, $(\sigma_{ij}^{(-1)}) = (\sigma_{ij})^{-1}$. Then the maximum likelihood estimator of μ is

$$\widehat{\mu} = \left(\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}^{(-1)}\right)^{-1} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}^{(-1)} X_i.$$

The inverse of the variance of $\widehat{\mu}$, and the Fisher information in $\mathcal{X}$ about μ, are both equal to $\Sigma\Sigma\sigma_{ij}^{(-1)}$. Now, the circulant version of the $n \times n$ Toeplitz matrix (σ_{ij}) has inverse (τ_{ij}), say, and the sum of the elements of (τ_{ij}) precisely equals

n^2 multiplied by the inverse of the sum of the elements of (σ_{ij}). Under the hypothesis that $\lambda/n \to 0$ and ρ vanishes on a compact interval, we may prove that $\Sigma\Sigma\sigma_{ij}^{(-1)} \sim \Sigma\Sigma\tau_{ij}$. (The argument is based on noting that the amount of information in $\mathcal{X}$ lies between the amounts of information in two other stochastic sequences, each of which has a circulant variance matrix and whose respective lengths are $n \pm O(\lambda)$.] Hence, since $f(0) = 1$ in this case,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}^{(-1)} \sim n^2\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\right)^{-1},$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}\rho(t_i - t_j) \sim n^2\lambda^{-1}f(0)\int\rho.$$

Therefore, $(\operatorname{var}\widehat{\mu})^{-1}$ and (information in $\mathcal{X}$ about μ) are both asymptotic to

$$n^2\left\{n^2\lambda^{-1}f(0)\int\rho\right\}^{-1} = \lambda\left\{f(0)\int\rho\right\}^{-1} \sim (\operatorname{var}\overline{X})^{-1}.$$

REMARK 3.8. If the process X is Gaussian and the covariance function ρ is known parametrically, then, in principle, the exact form of ρ can be determined precisely (i.e., without error) from observation of X over any finite interval, no matter how short. This is one of the paradoxes of inference about the spectrum of a stationary Gaussian process; see, for example, Grenander (1950) and Yaglom (1963). It follows that even if ρ is known only up to smoothness conditions, as hypothesized in our theorem, ρ can be estimated with extraordinary accuracy (at a faster rate than $\lambda^{-1/2}$) from the data $\mathcal{X} = \{X(t_1), \ldots, X(t_n)\}$. In this sense, the convergence rate described in the theorem falls short of optimality in the case where X is Gaussian.

Next, we discuss the implications of Theorem 3.1 for a covariance estimator, $\widetilde{\rho}$, designed to satisfy the condition of positive definiteness. We assume that the stationary process X has all moments finite; that ρ vanishes outside a compact interval, say $(-t_0, t_0)$; that an upper bound, T, to t_0 is known; and that for some $\alpha > 1$, the ratio

$$\rho^{\dagger}(\theta)/(1 + |\theta|)^{\alpha}$$

is bounded away from 0 and ∞, uniformly in θ.

The fact that ρ is known to vanish outside $(-T, T)$ is utilized by defining

$$\widehat{\rho}^{\dagger}(\theta) = \int_{-T}^{T}\widehat{\rho}(t)\cos(\theta t)\,dt. \tag{3.4}$$

Let $\beta \geq \frac{1}{2}$, and put

$$\widehat{\theta} = \inf\Big[\big\{\theta > 0\colon \widehat{\rho}^{\dagger}(\theta) \leq 0\big\} \cup \big\{\lambda^{\beta}\big\}\Big],$$

$$\widetilde{\rho}^{\dagger}(\theta) = \begin{cases} \widehat{\rho}^{\dagger}(\theta), & \text{if } |\theta| \leq \widehat{\theta}, \\ 0, & \text{if } |\theta| > \widehat{\theta}, \end{cases} \tag{3.5}$$

$$\widetilde{\rho}(t) = (2\pi)^{-1} \int_{-\hat{\theta}}^{\hat{\theta}} \widetilde{\rho}^{\dagger}(\theta) \cos(\theta t)\, d\theta.$$

At alternative approach to defining $\widetilde{\rho}$ is to replace $\widehat{\rho}^{\dagger}$ by

$$\widetilde{\rho}^{\dagger}(\theta) = \int_{-T}^{T} \left[\sum_i \sum_j \widehat{X}_{ij} \cos(t_{ij}\theta) K\big\{(t - t_{ij})/h\big\}\right] \times \left[\sum_i \sum_j K\big\{(t - t_{ij})/h\big\}\right]^{-1} dt, \tag{3.6}$$

and then define $\widetilde{\rho}^{\dagger}$ and $\widetilde{\rho}$ as before. The validity of the theorem is not affected by this change. It is tempting to replace the definition of $\widetilde{\rho}^{\dagger}(\theta)$ by, for example, a quantity proportional to

$$\sum_i \sum_j \widehat{X}_{ij} \cos(t_{ij}\theta).$$

However, this change is really only permissible if the points t_{ij} are asymptotically uniformly distributed over $(-T, T)$. The definition of $\widetilde{\rho}^{\dagger}$ at (3.6) represents one of several different possible approaches to correcting for nonuniformity of the t_{ij}'s.

THEOREM 3.2. *Assume the conditions stated above, and that for some $\eta > 0$, we have $\lambda \to \infty$ and $n^{1-\eta}\lambda^{-7/8} \to \infty$ as $n \to \infty$. [Then we may choose $h = h(n)$ such that for some $\eta' > 0$, we have $n^{\eta'} \to 0$, $n^{1-\eta'}h \to \infty$, $n^{2-\eta'}h^3\lambda^{-1} \to \infty$ and $\lambda h^4 \to 0$; see Remark 3.1. We assume that h has these properties.] Suppose, too, that either $t_1, \ldots, t_n$ are regularly spaced on an interval of width λ or that they are generated as described in the opening paragraph of this section, with U_i having a continuous distribution with a bounded and continuous density, and that K is a bounded, continuous, compactly supported probability density with $K(0) \neq 0$. Then*

$$\sup_{t \geq 0} |\widetilde{\rho}(t) - \rho(t)| = O_p\big(\lambda^{-(1/2)+(1/2\alpha)}\big) \tag{3.7}$$

as $\lambda \to \infty$.

REMARK 3.9. The rate of convergence, that is, $\lambda^{-(1/2)+(1/2\alpha)}$, described by (3.7), is optimal in the following sense. Under appropriate mixing conditions

on the process X, and assuming that $\rho^\dagger(\theta) \sim \text{const.}|\theta|^{-\alpha}$ as $|\theta| \to \infty$, it may be proved that $\widehat{\theta}/\lambda^{1/(2\alpha)}$ has a proper, nonzero weak limit. See that part of the proof of Theorem 3.2 subsequent to (4.23). It follows from (4.21) and (4.23) of that proof that for fixed t, in particular for $t = 0$, $\lambda^{-(1/2)+(1/2\alpha)}\{\widetilde{\rho}(t) - \rho(t)\}$ has a proper weak limit.

REMARK 3.10. Closely related techniques may be employed to derive convergence rates of estimators in the case where ρ is not compactly supported. There, we may still use (3.4) to define $\widehat{\rho}^\dagger$, except that $T = T(\lambda, n)$ now diverges slowly to ∞. The overall convergence rate is now a function of the rate at which $\rho(t) \to 0$ as $|t| \to \infty$, as well as the rate at which $\rho^\dagger(\theta) \to 0$ as $|\theta| \to \infty$. For example, if ρ decreases exponentially quickly—in particular, if $|\rho(t)| \le C_1 \exp(-C_2|t|)$ for constants $C_1, C_2 > 0$—then in definition (3.4) we may take $T = C_3 \log n$ for any sufficiently large $C_3 > 0$. Minor modifications of our present proof of Theorem 3.2 show that (3.7) continues to hold, provided that the right-hand side is replaced by $O_p(\lambda^{-(1/2)+(1/2\alpha)+\varepsilon})$ for arbitrary $\varepsilon > 0$.

4. Proofs.

4.1. *Proof of Theorem* 3.1. We give the proof only in outline, confining almost all attention to the case where $i = j$ terms are included in the definition of $\widehat{\rho}(t)$. The following notation is used:

$$t_{ij} = t_i - t_j, \qquad \widehat{X}_{ij} = \{X(t_i) - \overline{X}\}\{X(t_j) - \overline{X}\},$$

$$X_{ij} = \{X(t_i) - \mu\}\{X(t_j) - \mu\},$$

$$a(t) = \sum_i \sum_j K\{(t - t_{ij})/h\}, \qquad B(t) = \sum_i \sum_j X_{ij} K\{(t - t_{ij})/h\},$$

$$b = E(B),$$

$$\widetilde{\rho}(t) = B(t)/a(t), \qquad \varepsilon_{ij} = X_{ij} - \rho(t_{ij}),$$

$$\Delta(t) = a(t)^{-1} \sum_i \sum_j \{X(t_i) - \mu\} K\{(t - t_{ij})/h\}.$$

In this notation,

$$\widehat{\rho}(t) = \widetilde{\rho}(t) - 2(\overline{X} - \mu)\Delta(t) + (\overline{X} - \mu)^2. \tag{4.1}$$

The remainder of our proof is comprised of five steps, dealing with different aspects of formula (4.1).

Step (i): $a(t)$. Since $t_{ij} = 0$, then

$$a(t) = a_1(t) + nK(t/h), \tag{4.2}$$

where

$$a_1(t) = \sum_{i \ne j} \sum K\{(t - t_{ij})/h\}.$$

If $t \neq 0$, then $K(t/h) = 0$ for all sufficiently small h.

Put $W_{ij} = K\{(t - \lambda U_{ij})/h\}$,

$$\alpha_1(u) = E(W_{12} \mid U_1 = u), \qquad \alpha_2(u) = E(W_{12} \mid U_2 = u), \qquad \alpha = E(W_{12}), \tag{4.3}$$

$$D_{ij} = W_{ij} - \alpha_1(U_i) - \alpha_2(U_j) + \alpha, \qquad Z_j = \sum_{i=1}^{j-1} D_{ij}. \tag{4.4}$$

In this notation, $a_1(t) - n(n-1)\alpha$ represents an observed value of

$$\sum\sum_{i \neq j}(W_{ij} - \alpha) = 2\sum_{j=2}^{n} Z_i + (n-1)\sum_{k=1}^{2}\sum_{i=1}^{n}\{\alpha_k(U_i) - \alpha\}. \tag{4.5}$$

Now, $E(Z_j \mid U_1, \ldots, U_{j-1}) = 0$, and $Z_2, \ldots, Z_n$ are martingale differences. Hence, by Rosenthal's inequality [Hall and Heyde (1980), page 23], for integers $p \geq 1$,

$$E\left(\sum_{j=2}^{n} Z_j\right)^{2p} \leq C_{1p}\left[E\left\{\sum_{j=2}^{n} E(Z_j^2 \mid U_1, \ldots, U_{j-1})\right\}^p + \sum_{j=2}^{n} E(Z_j^{2p})\right]. \tag{4.6}$$

Here, C_{lp} denotes a constant depending only on p.

Next, observe that

$$\begin{aligned} E\left\{\sum_{j=2}^{n} E\,(Z_j^2 \mid U_1, \ldots, U_{j-1})\right\}^p &\leq n^{p-1}\sum_{j=2}^{n} E(Z_j^{2p}), \\ E\,(Z_j^{2p}) &\leq C_{1p}\left[E\left\{\sum_{j=2}^{i=1} E(D_{ij}^2 \mid U_j)\right\}^p + (j-1)E(D_{ij}^{2p})\right], \\ \sup_u E(D_{ij}^2 \mid U_j = u) &= O(h\lambda^{-1}). \end{aligned} \tag{4.7}$$

Combining the results from (4.6) down, we deduce that for large p,

$$E\left(\sum_{j=2}^{n} Z_j\right)^{2p} \leq C_{2p} n^{2p}(h\lambda^{-1})^p. \tag{4.8}$$

Furthermore,

$$|\alpha_k(u)| \leq Ch\lambda^{-1} \tag{4.9}$$

for $k = 1, 2$. Hence, by Rosenthal's inequality again,

$$E\left[\sum_{i=1}^{n}\{\alpha_k(U_k) - \alpha\}\right]^{2p} \leq C_{3p} n^p (h\lambda^{-1})^{2p}. \tag{4.10}$$

Combining (4.5), (4.8) and (4.10), we obtain

$$E\left\{\sum\sum_{i \neq j}(W_{ij} - \alpha)\right\}^{2p} \leq C_{4p}\left\{n^{2p}(h\lambda^{-1})^p + n^{3p}(h\lambda^{-1})^{2p}\right\}.$$

Therefore, by Markov's inequality, for each $\eta > 0$ and $p \geq 1$,

$$P\left\{\left|\sum\sum_{i \neq j}(W_{ij} - \alpha)\right| > \eta n^2 h\lambda^{-1}\right\} = O\left\{\left(n^2 h\lambda^{-1}\right)^{-p} + n^{-p}\right\}. \tag{4.11}$$

It now follows via the Borel–Cantelli lemma that with probability 1,

$$\left(n^2 h\lambda^{-1}\right)^{-1} \sum\sum_{i \neq j}(W_{ij} - \alpha) \to 0. \tag{4.12}$$

Now,

$$\alpha = h\lambda^{-1} \int K(u) f_1\{\lambda^{-1}(t - hu)\}\, du = h\lambda^{-1} f_1(0) + o\left(h\lambda^{-1}\right).$$

Hence by (4.12), for a sequence $t_1, t_2, \ldots$ arising with probability 1,

$$a_1(t) = n^2 h\lambda^{-1} f_1(0) + o\left(n^2 h\lambda^{-1}\right).$$

We may now deduce from (4.2) that

$$a(t) \sim \begin{cases} n^2 h\lambda^{-1} f_1(0), & \text{if } t \neq 0 \text{ or } t = 0 \text{ and } nh\lambda^{-1} \to \infty, \\ n\{cf_1(0) + K(0)\}, & \text{if } t = 0 \text{ and } nh\lambda^{-1} \to c < \infty. \end{cases} \tag{4.13}$$

Step (ii): $b(t)$. Since $t_{ii} = 0$, then

$$b(t) - a(t)\rho(t) = b_1(t) + n\{\rho(0) - \rho(t)\}K\left(t/h\right), \tag{4.14}$$

where

$$b_1(t) = \sum\sum_{i \neq j}\{\rho(t_{ij}) - \rho(t)\}K\{(t - t_{ij})/h\}.$$

Paralleling the argument in Step (i), put $W_{ij} = \{\rho(\lambda U_{ij}) - \rho(t)\}\, K\{(t - \lambda U_{ij})/h\}$, and in that notation define α_1, α_2, α, D_{ij}, Z_i by (4.3) and (4.4). The arguments given before continue to apply, except that in place of (4.7) and (4.9) we have

$$\sup_u E\left(D_{ij}^2 \mid U_j = u\right) = O\left(h^3\lambda^{-1}\right), \qquad |\alpha_k(u)| \leq Ch^3\lambda^{-1}.$$

Therefore, in place of (4.12),

$$P\left\{\left|\sum\sum_{i \neq j}(W_{ij} - \alpha)\right| > \eta n^2 h^3\lambda^{-1}\right\} = O\left\{\left(n^2 h^3\lambda^{-1}\right)^{-p} + n^{-p}\right\},$$

where by the Borel–Cantelli lemma,

$$\left(n^2 h^3\lambda^{-1}\right)^{-1} \sum\sum_{i \neq j}(W_{ij} - \alpha) \to 0,$$

with probability 1. Now,

$$\alpha = h\lambda^{-1}\int \{\rho(t-hu)-\rho(t)\}K(u)f_1\{\lambda^{-2}(t-hu)\}\,du$$
$$= \tfrac{1}{2}h^3\lambda^{-1}f_1(0)\rho''(t)+o\big(h^3\lambda^{-1}\big),$$

and so for a sequence $t_1, t_2, \ldots$ arising with probability 1,

$$b_1(t) = \tfrac{1}{2}n^2h^3\lambda^{-1}f_1(0)\rho''(t)+o\big(n^2h^3\lambda^{-1}\big).$$

Therefore, by (4.14),

$$b(t)-a(t)\rho(t) = \tfrac{1}{2}n^2h^3\lambda^{-1}f_1(0)\rho''(t)+o\big(n^2h^3\lambda^{-1}\big). \tag{4.15}$$

It is straightforward to check that, with

$$S = \big(n^2h^3\lambda^{-1}\big)^{-1/2}\sum\sum_{i\neq j}(W_{ij}-\alpha),$$

we have $\operatorname{var}(S)\to\sigma^2$ where $0<\sigma^2<\infty$, and that the fluctuations of S are of order at least 1. This demonstrates that, for sequences $t_1, t_2, \ldots$ arising with probability 1, the fluctuations of

$$\big(n^2h^3\lambda^{-1}\big)^{-1/2}\{b_1(t)-n(n-1)\alpha\}$$

are of order at least 1. Therefore, result (4.15) will fail if $n^2h^3\lambda^{-1}$ is bounded. This indicates that the condition $n^{2-\eta}h^3\lambda^{-1}\to\infty$ is close to being necessary for the theorem to hold.

Step (iii): $\operatorname{var}(\overline{X}), \operatorname{var}\{\Delta(t)\}$. Observe that

$$n^2\operatorname{var}(\overline{X}) = \sum\sum_{i\neq j}\rho(t_{ij})+n\rho(0),$$

$$E\left\{\sum\sum_{i\neq j}\rho(\lambda U_{ij})\right\} = n(n-1)\lambda^{-1}\int\rho(u)f_1\big(\lambda^{-1}u\big)du$$
$$= n^2\lambda^{-1}f_1(0)\int\rho+o\big(n^2\lambda^{-1}\big).$$

From these results, arguing as in Steps (i) and (ii), we may deduce that for a sequence of values $t_1, t_2, \ldots$ arising with probability 1,

$$\operatorname{var}(\overline{X}) = \lambda^{-1}f_1(0)\int\rho+o\big(\lambda^{-1}\big).$$

Similarly,

$$a(t)^2\operatorname{var}\{\Delta(t)\} = \sum_{i_1}\sum_{i_2}\sum_{j_1}\sum_{j_2}\rho\big(t_{i_1i_2}\big)K\Big\{\big(t-t_{i_1j_1}\big)/h\Big\}K\Big\{\big(t-t_{i_2j_2}\big)/h\Big\}$$
$$= n^4h^2\lambda^{-3}f_1(0)\int\rho+o\big(n^4h^2\lambda^{-3}\big)$$

whence by (4.13), $\operatorname{var}\{\Delta(t)\} = (\lambda^{-1})$. Hence, for a sequence of values $t_1, t_2, \ldots$ arising with probability 1,

$$\operatorname{var}(\overline{X}) + \operatorname{var}\{\Delta(t)\} = O(\lambda^{-1}). \tag{4.16}$$

Step (iv): $\operatorname{var}\{\widetilde{\rho}(t)\}$. Observe that

$$\begin{aligned} a(t)^2 \operatorname{var}\{\widetilde{\rho}(t)\} = {} & \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} g(t_{i_1 j_1}, t_{i_2 j_2}, t_{j_1 j_2}) \\ & \times K\{(t - t_{i_1 j_1})/h\} K\{(t - t_{i_2 j_2})/h\}. \end{aligned} \tag{4.17}$$

If we replace each t_{rs} on the right-hand side by λU_{rs} and take expectations, we obtain the value

$$\begin{aligned} s = {} & n(n-1)(n-2)(n-3) E\Big(g\{\lambda(U_1 - U_2), \lambda(U_3 - U_4), \lambda(U_2 - U_4)\} \\ & \times K\big[\{t - \lambda(U_1 - U_2)\}/h\big] K\big[\{t - \lambda(U_3 - U_4)\}/h\big] \Big) \\ & + 2n(n-1)(n-2) E\Big(g\{0, \lambda(U_2 - U_3), \lambda(U_1 - U_3)\} K(t/h) \\ & \times K\big[\{t - \lambda(U_2 - U_3)\}/h\big] \Big) \\ & + 4n(n-1)(n-2) E\Big(g\{\lambda(U_1 - U_2), \lambda(U_1 - U_3), \lambda(U_2 - U_3)\} \\ & \times K\big[\{t - \lambda(U_1 - U_2)\}/h\big] \Big) \\ & + 7n(n-1) E\Big(g\{0, \lambda(U_1 - U_2), \lambda(U_1 - U_2)\} \\ & \times K(t/h) K\big[\{t - \lambda(U_1 - U_2)\}/h\big] \Big) \\ & + ng(0,0,0) K(t/h)^2. \end{aligned}$$

Let $f_3(v_1, v_2, v_3)$, $f_2(v_1, v_2)$ and $f_1(v)$ denote the joint densities of $(U_1 - U_2, U_3 - U_4, U_2 - U_4)$, $(U_2 - U_3, U_1 - U_3)$ and $U_1 - U_2$, respectively. In this notation,

$$\begin{aligned} s = {} & n^4 \{1 + O(n^{-1})\} \int g(\lambda v_1, \lambda v_2, \lambda v_3) K\{(t - \lambda v_1)/h\} K\{(t - \lambda v_2)/h\} \\ & \times f_3(v_1, v_2, v_3)\, dv_1\, dv_2\, dv_3 \\ & + 2n^3 \{1 + O(n^{-1})\} K(t/h) \int g(0, \lambda v_1, \lambda v_2) K\{(t - \lambda v_1)/h\} f_2(v_1, v_2)\, dv_1\, dv_2 \end{aligned}$$

$$
\begin{aligned}
&+ 4n^3\{1+O(n^{-1})\}\int g\{\lambda v_1, \lambda v_2, \lambda(v_2 - v_1)\}K\{(t-\lambda v_1)/h\}K\{(t-\lambda v_2)/h\} \\
&\quad \times f_3(v_1, v_2)\,dv_1\,dv_2 \\
&+ 7n^2\{1+O(n^{-1})\}K(t/h)\int g(0, \lambda v, \lambda v)K\{(t-\lambda v)/h\}f_1(v)\,dv \\
&+ ng(0,0,0)K(t/h)^2 \\
&= n^4h^2\lambda^{-3}\int g(t, t, u_3)K(u_1)K(u_2)f_3(0,0,0)\,du_1\,du_2\,du_3 \\
&+ 2n^3h\lambda^{-2}K(t/h)\int g(0, t, u_2)K(u_1)f_2(0,0)\,du_1\,du_2 + ng(0,0,0)K(t/h)^2 \\
&+ o(n^4h^2\lambda^{-3}) + O\{n^3h^2\lambda^{-2} + n^2h\lambda^{-1}K(t/h)\} \\
&\sim \begin{cases}
n^4h^2\lambda^{-3}f_3(0,0,0)\int g(t,t,u)\,du, & \text{if } t \neq 0 \text{ or } t = 0 \text{ and } nh\lambda^{-1} \to \infty \\
n^3h\lambda^{-2}\{cf_3(0,0,0)\int g(t,t,u)\,du & \\
\qquad + 2K(0)f_2(0,0)\int g(0,t,u)\,du\} & \\
\qquad + ng(0,0,0)K(0)^2, & \text{if } t = 0 \text{ and } nh\Lambda \to c,\ 0 \le c < \infty.
\end{cases}
\end{aligned}
$$

Therefore, using (4.13),

$$
(4.18)\quad a(t)^{-2}s \sim \begin{cases}
\lambda^{-1}f_1(0)^{-2}f_3(0,0,0)\int g(t,t,u)\,du, & \\
\qquad \text{if } t \neq 0 \text{ or } t = 0 \text{ and } nh\lambda^{-1} \to \infty, & \\
nh\lambda^{-2}\{cf_1(0) + K(0)\}^{-2} & \\
\quad \times \{cf_3(0,0,0)\int g(t,t,u)\,du & \\
\qquad + 2K(0)f_2(0,0)\int g(0,t,u)\,du\} & \\
\quad + n^{-1}\{cf_1(0) + K(0)\}^{-2}g(0,0,0)K(0)^2, & \\
\qquad \text{if } t = 0 \text{ and } nh\lambda^{-1} \to c,\ 0 \le c < \infty.
\end{cases}
$$

From the second case in the formula above, it follows that when $t = 0$ and $nh\lambda^{-1} \to c$ where $0 < c < \infty$, we have $a(t)^{-2}s \sim \lambda^{-1}q$; and when $t = 0$ and $nh\lambda^{-1} \to 0$, $a(t)^{-2}s = o(\lambda^{-1})$. Techniques from Steps (i) and (ii) may be used to prove that, for a sequence $t_1, t_2, \ldots$ arising with probability 1,

$$
(4.19)\quad \operatorname{var}\{\widetilde{\rho}(t)\}\begin{cases}
= \lambda^{-1}q + o(\lambda^{-1}), & \text{if } t \neq 0 \text{ and } nh\lambda^{-1} \to c,\ 0 < c \le \infty, \\
= o(\lambda^{-1}), & \text{if } t = 0 \text{ and } nh\lambda^{-1} \to 0,
\end{cases}
$$

the precise convergence rate in the latter case being $(nh\lambda^{-1})\lambda^{-1} + n^{-1}$.

Step (v): *Mean squared error of* $\widetilde{\rho}(t)$. Combining (4.13), (4.15) and (4.19), we

deduce that

$$E\{\widetilde{\rho}(t)-\rho(t)\}^2 = a(t)^{-2}\{b(t)-a(t)\rho(t)\}^2 + \operatorname{var}\{\widetilde{\rho}(t)\}$$

$$\sim \begin{cases} \frac{1}{4}h^4\rho''(t)^2 + \lambda^{-1}q, & \text{if } t \neq 0 \text{ or } t = 0 \text{ and } nh\lambda^{-1} \to \infty, \\ \frac{1}{4}h^4c^2\{cf(0)+K(0)\}^{-2}f(0)^2\rho''(t)^2 + \lambda^{-1}d, & \\ \qquad\qquad \text{if } t = 0 \text{ and } nh\lambda^{-1} \to c,\ 0 < c < \infty, & \\ o(h^4+\lambda^{-1}), & \text{if } t = 0 \text{ and } nh\lambda^{-1} \to 0, \end{cases} \tag{4.20}$$

the precise convergence rate in the latter case being $(nh\lambda^{-1})^2h^4 + (nh\lambda^{-1})\lambda^{-1} + n^{-1}$.

If the diagonal terms, corresponding to $i = j$, are deleted from the definition of $\widetilde{\rho}(t)$, then, on inspection, it is clear that the first options in formulae (4.13) and (4.18) hold for both $t = 0$ and $t \neq 0$. That is,

$$a(t) \sim n^2h\lambda^{-1}f_1(0),$$

$$a(t)^{-1}s \sim \lambda^{-1}f_1(0)^{-1}f_3(0,0,0)\int g(t,t,u)\,du.$$

Therefore, the first option in (4.20) holds.

The theorem follows from (4.1) and (4.20), on noting [via Step (iii)] that only the first term on the right-hand side of (4.1) makes a nonnegligible contribution to mean squared error. □

4.2. *Proof of Theorem* 3.2. Observe that

$$\pi\{\widetilde{\rho}(t)-\rho(t)\} = \int_0^{\hat{\theta}} \{\widehat{\rho}^{\dagger}(\theta)-\rho^{\dagger}(\theta)\}\cos(\theta t)\,d\theta - \int_{\widehat{\theta}}^{\infty} \rho^{\dagger}(\theta)\cos(\theta t)\,d\theta, \tag{4.21}$$

and for any $p, q > 1$ with $p^{-1} + q^{-1} = 1$,

$$\begin{aligned} I_1(t) &\equiv \left| \int_0^{\hat{\theta}} \{\widehat{\rho}^{\dagger}(\theta)-\rho^{\dagger}(\theta)\}\cos(\theta t)\,d\theta \right| \\ &= \tfrac{1}{2}\left| \int_0^T \{\widehat{\rho}(u)-\rho(u)\}\left[(t+u)^{-1}\sin\{\widehat{\theta}(t+u)\} + (t-u)^{-1}\sin\{\widehat{\theta}(t-u)\}\right]du \right| \\ &\le \left\{ \int_0^T |\widehat{\rho}(u)-\rho(u)|^p\,du \right\}^{1/p} \\ &\quad \times \left(\int_0^T \left[|t+u|^{-q}\left|\sin\{\widehat{\theta}(t+u)\}\right|^q + |t-u|^{-q}\left|\sin\{\widehat{\theta}(t-u)\}\right|^q \right] du \right)^{1/q}. \end{aligned}$$

The arguments used to establish Theorem 3.1 may be used to prove that, for h selected as in the statement of Theorem 3.2, we have for $p \ge 1$,

$$\sup_{0 \le t \le T} E|\widehat{\rho}(t)-\rho(t)|^p = O\left(\lambda^{-p/2}\right). \tag{4.22}$$

(The case where p is an even integer is simplest to treat; other cases follow via Hölder's inequality.) Furthermore,

$$\sup_{t \geq 0} \int_0^T |t \pm u|^{-q} \left|\sin\{\widehat{\theta}(t \pm u)\}\right|^q du \leq C(q)\widehat{\theta}^{q-1}$$

Therefore,

$$\sup_{t \geq 0} I_1(t) = O_p\left(\lambda^{-1/2+\beta(q-1)}\right).$$

Since this is true for all $q > 1$, then for all $\varepsilon > 0$,

$$\sup_{t \geq 0} I_1(t) = O_p\left(\lambda^{-1/2+\varepsilon}\right). \tag{4.23}$$

Note that by (4.22),

$$\begin{aligned}\sup_{t \geq 0} |\widehat{\rho}^{\dagger}(\theta) - \rho^{\dagger}(\theta)| &\leq 2 \int_0^T |\widehat{\rho}(t) - \rho(t)|\, dt \\ &= O_p\left(\lambda^{-1/2}\right).\end{aligned}$$

Hence

$$\widehat{\rho}^{\dagger}(\theta) \geq \rho^{\dagger}(\theta) - O_p\left(\lambda^{-1/2}\right),$$

uniformly in θ. It follows that, since $\rho^{\dagger}(\theta)$ decreases like $|\theta|^{-\alpha}$ as $|\theta| \to \infty$, $\widehat{\theta}$ must be at least of size $\lambda^{1/(2\alpha)}$. Therefore,

$$\int_{\widehat{\theta}}^{\infty} \widehat{\rho}^{\dagger}(\theta)\cos(\theta t)\, d\theta = O_p\left(\widehat{\theta}^{-(\alpha-1)}\right) = O_p\left(\lambda^{-1/2+(1/2\alpha)}\right).$$

The theorem follows from this result, (4.21) and (4.23). □

Acknowledgments. The helpful comments of reviewers and an Associate Editor and the Editor have assisted in the preparation of this more succinct version of the paper.

REFERENCES

ARMSTRONG, A. G. and DIAMOND, P. (1984). Testing variograms for positive-definiteness. *Math. Geol.* **16** 407–421.

BERMAN, M. and DIGGLE, P. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *J. Roy. Statist. Soc. Ser. B* **51** 81–92.

CHRISTAKOS, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research* **20** 251–265.

CRESSIE, N. (1991). *Statistics for Spatial Data*. Wiley, New York.

DIGGLE, P. J., GATES, D. J. and STIBBARD, A. (1987). A nonparametric estimator for pairwise-interaction point processes. *Biometrika* **74** 763–770.

GRENANDER, U. (1950). Stochastic processes and statistical inference. *Arch. Math.* **1** 195–277.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application.* Academic, New York.

JOURNEL, A. G. and HUIJBREGTS, CH. J. (1978). *Mining Geostatistics.* Academic, London.

MATHERON, G. (1971). *The Theory of Regionalised Variables.* Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau #5.

SAMPSON, P. D. and GUTTORP, P. (1992). Nonparametric representation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.

SCHOENBERG, I. J. (1938). Metric spaces and completely monotone functions. *Ann. of Math.* **79** 811–841.

SHAPIRO, A. and BOTHA, J. D. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Comput. Statist. Data Anal.* **11** 87–96.

YAGLOM, A. M. (1963). On the equivalence and perpendicularity of two Gaussian probability measures in function space. In *Proceedings of the Symposium on Time Series Analysis* (M. Rosenblatt, ed.) 327–346. Wiley, New York.

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

NICHOLAS I. FISHER
CSIRO DIVISION OF MATHEMATICS
AND STATISTICS
LOCKED BAG 17
NORTH RYDE, NSW 2113
AUSTRALIA

BRANKA HOFFMANN
CSIRO DIVISION OF MATHEMATICS
AND STATISTICS
LOCKED BAG 17
NORTH RYDE, NSW 2113
AUSTRALIA