Kernel Regression Estimation With Time Series Errors
Author(s): Jeffrey D. Hart
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 53, No. 1 (1991), pp. 173-187
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2345733
Accessed: 28-08-2018 14:52 UTC

J. R. Statist. Soc. B (1991)
53, *No.* 1, *pp.* 173–187

Kernel Regression Estimation with Time Series Errors

By JEFFREY D. HART†

Texas A&M University, College Station, USA

[Received September 1987. Final revision September 1989]

SUMMARY

Suppose that data $y_1, \ldots, y_n$ are observed according to the model $y_i = f((i-\frac{1}{2})/n) + \epsilon_i$, $i = 1, \ldots, n$, where f is a smooth function and $\{\epsilon_1, \ldots, \epsilon_n\}$ is a sample from a zero-mean, covariance stationary time series. The problem of objectively choosing the bandwidth of a kernel estimate for f is addressed. It is shown both theoretically and by simulation that cross-validation produces extremely rough kernel estimates when the data are sufficiently positively correlated. This makes it inadvisable to use residuals from a cross-validated kernel estimate as a means of estimating the covariance function of the errors. Alternative methods of estimating the covariance function are proposed. In a simulation study, incorporating these estimated covariances into a risk estimation procedure leads to more efficient smoothing of positively correlated data.

Keywords: AUTOREGRESSIVE PROCESS; BANDWIDTH SELECTION; MEAN AVERAGE-SQUARED ERROR; MEAN-SQUARED ERROR; SPECTRUM

1. INTRODUCTION

A vast literature now exists on using kernel-type smoothers to estimate regression functions nonparametrically. Practically all this literature is based on the assumption that the observed data are uncorrelated. In essence, this assumption implies that any observed trends, whether long or short term, are either deterministic in nature or simply anomalous chance occurrences. Such an implication is clearly undesirable. There are many settings, such as in time series analysis, where it is reasonable to model slowly varying trends deterministically, but to explain any other regular behaviour in the data by means of a correlation model. Furthermore, kernel estimators are sometimes used to smooth data which result from processing uncorrelated data. Such processing can introduce correlation into the data to be smoothed. An example of the latter phenomenon occurs in the estimation of heteroscedasticity in regression (see Müller and Stadtmüller (1987)).

In what follows it is assumed that univariate data $y_1, \ldots, y_n$ are observed, and that

$$y_i = f\left(\frac{i-\frac{1}{2}}{n}\right) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where f is a smooth function defined on [0, 1] and $\{\epsilon_1, \ldots, \epsilon_n\}$ is a sample from a zero-mean, covariance stationary process. Subsequent theoretical results do not depend in any crucial way on the fact that the design in model (1.1) is evenly spaced. The Gasser and Müller (1984) kernel estimator of the function f is defined by

†*Address for correspondence*: Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.

 0035-9246/91/53173 $2.00

$$\hat{f}_h(x) = h^{-1} \sum_{i=1}^{n} y_i \int_{(i-1)/n}^{i/n} K\left(\frac{x-u}{h}\right) du, \qquad x \in (0, 1), \tag{1.2}$$

where $h > 0$ and K is a function (often a density) whose properties will be specified later. The main objectives of this paper are twofold. First, it will be shown that positively correlated data have a disastrous effect on cross-validatory methods of choosing the bandwidth h of $\hat{f}_h$. When n is large, cross-validation has a strong tendency to choose a kernel estimate that virtually interpolates the data. The surprising aspect of this result is that the correlation can be quite weak and still destroy the usual optimality property of cross-validation. For example, a first-order moving average error structure with a correlation of 0.25 is sufficient to yield the interpolation result. Our theorem in this regard provides theoretical backing for the empirical findings of Diggle (1985) and Hart and Wehrly (1986). The second objective of the paper is to describe and analyse a way of improving smoothing methods when the data are correlated. The essence of the new method is that it allows some of the smoothness in the data to be due to correlation. The lowest frequency components in the data are filtered out by differencing, and the covariance function of the error process is estimated from the periodogram of the differenced data. The estimated covariance function is then incorporated into a risk estimation criterion for choosing the bandwidth of $\hat{f}_h$. This new methodology is illustrated using both simulated data and the time series composed of the Beveridge wheat price index from the year 1500 to 1869. (See, for example, Anderson (1971).)

Before continuing, it is worth mentioning some very recent work concerned with smoothing correlated data. Diggle and Hutchinson (1989) propose a method for choosing the smoothing parameters of cubic spline regression estimators in the presence of correlated errors. Also, on completing the first version of this paper, the author learned of unpublished manuscripts by Chiu (Rice University, USA) and by Altman (Cornell University, USA) that draw conclusions similar to those in this paper.

The rest of the paper will proceed as follows. In the next section a large sample approximation to the mean-squared error (MSE) of $\hat{f}_h(x)$ is given. This result shows in a simple way how the efficiency of kernel estimators is affected by serial correlation. Section 3 is devoted to the result mentioned earlier concerning the failure of cross-validation with positively correlated data. Finally, the new methodology and numerical results are presented in, respectively, Sections 4 and 5.

2. MEAN-SQUARED ERROR PROPERTIES OF KERNEL ESTIMATORS

For convenience, here we present some assumptions and notation concerning the kernel K. The assumptions will be implicit throughout the remainder of the paper.

(a) K vanishes outside $(-1, 1)$;
(b) K is symmetric about zero;
(c) $\int_{-1}^{1} K(y)\, dy = 1$;
(d) K is Lipschitz continuous.

We also define two functionals of K:

(a) $J_K = \int_{-1}^{1} K^2(y)\,dy$;
(b) $\sigma_K^2 = \int_{-1}^{1} y^2 K(y)\,dy$.

To demonstrate in a simple way how correlation affects kernel estimators, it is of interest to investigate the behaviour of $E(\hat{f}_h(x) - f(x))^2$ when n is large. In doing so it will be assumed that equation (1.1) holds and that $\{\epsilon_1, \ldots, \epsilon_n\}$ is covariance stationary, i.e.

$$\operatorname{cov}(\epsilon_i, \epsilon_{i+k}) = c_n(k), \qquad |k| = 0, 1, \ldots. \tag{2.1}$$

To study the asymptotic MSE of $\hat{f}_h(x)$, it is still necessary to specify the nature of c_n as $n \to \infty$. There are two main cases of interest. Let $\{k_n(u)\}$ be a sequence of positive integers such that $\lim_{n\to\infty} k_n(u)/n = u \in [0, 1]$. Then, for large n, $c_n(k_n(u))$ is the covariance between data values whose design points differ by about u. Now, the two main cases are (with $\sigma^2 > 0$)

(a) $$\lim_{n\to\infty} c_n(k_n(u)) = \sigma^2 \rho(u) \tag{2.2}$$

(b) $$\lim_{n\to\infty} c_n(k_n(u)) = \sigma^2 I_{\{0\}}(u) \tag{2.3}$$

for each $u \in [0, 1]$, where ρ is a real-valued characteristic function.

In the first of these cases, the observed data are a sample from a continuous process on [0, 1]. In this case kernel estimators will not be consistent for f as $n \to \infty$. (See Hart and Wehrly (1986).) It still makes perfect sense to use kernel estimates to smooth the sample paths of continuous processes. Doing so amounts simply to passing the data through a linear filter, which has been a persistent theme in time series analysis.

The following theorem indicates the asymptotic behaviour of the MSE of $\hat{f}_h(x)$ when the covariance function $c_n(k)$ satisfies

$$c_n(k) = c(k) \qquad n \geqslant 1, \qquad |k| = 0, 1, \ldots. \tag{2.4}$$

This covariance structure is a special case of equation (2.3) and is essentially the same model used by Härdle and Tuan (1986) in their study of M-smoothing of time series. The proof of theorem 1 is given in Hart (1987) and is thus omitted here.

Theorem 1. Let equation (1.1) hold and assume that f has two continuous derivatives on [0, 1]. Assume also that the covariance model (2.4) holds, where $c(0) < \infty$ and $\Sigma_{k=1}^{\infty} k|c(k)| < \infty$. Now, as $n \to \infty$ and $h \to 0$, the following statement holds uniformly in $x \in (h, 1-h)$:

$$E(\hat{f}_h(x) - f(x))^2 = \frac{1}{nh}\left\{c(0) + 2\sum_{j=1}^{\infty} c(j)\right\} J_K + \frac{h^4 \sigma_K^4}{4}\{f''(x)\}^2 + o((nh)^{-1} + h^4).$$

A practical situation where model (2.4) arises may be described as follows. Suppose that the data z_i, $i = 1, \ldots, n$, follow a regression model with independent errors. If the data available for analysis are a smoothed version of the z_i, namely

$$y_i = \sum_{j=i-m}^{i+m} z_j/(2m+1), \qquad i = m+1, \ldots, n-m,$$

then the covariance structure of the y_i will satisfy model (2.4) as long as m is a constant (independent of n). Such data smoothing operations are quite common in chemical

applications. (See, for example, the interpolation scheme in Brereton (1987), p. 179.) In other situations, such as kriging (see, for example, Stein (1987)), correlation is an intrinsic property of the observations and not merely a by-product of processing the data. In these situations, the correlation is usually taken to be (independently of n) a function of the distance in space or time between data values, and hence model (2.2) is more reasonable than model (2.4). Technically, theorem 1 does not describe what happens under model (2.2). However, in practice the sample size is always finite and the only relevant question then is how good the approximation provided by an asymptotic result is. Roughly, the MSE approximation in theorem 1 is good when n is large and k_n/n is sufficiently small, where k_n is such that $|c_n(k)/c_n(0)| \approx 0$ for $k \geqslant k_n$. These conditions will often be met even when the data are a sample from a continuous process.

We close this section with the following remarks.

(a) If $c(j) = 0$ for $j = 1, 2, \ldots$, then theorem 1 gives the standard result for the MSE of $\hat{f}_h(x)$ (see Gasser and Müller (1984)). The asymptotically optimum bandwidth is (for $f''(x) \neq 0$)

$$h_n = \left[B_K \left\{ c(0) + 2 \sum_{j=1}^{\infty} c(j) \right\} \Big/ \{f''(x)\}^2 \right]^{1/5} n^{-1/5},$$

where $B_K = J_K/\sigma_K^4$. We see that h_n is at least as big as the bandwidth for uncorrelated data if $c(j) \geqslant 0$ for all $j \geqslant 1$. The quantity $c(0) + 2\Sigma_{j=1}^{\infty} c(j)$ is proportional to $S_\epsilon(0)$, where $S_\epsilon(\omega)$ is the spectrum of the error process.

(b) We can remedy boundary effects using the boundary kernels of Gasser and Müller (1979). The correlation in the data is not important in this regard since boundary effects are a bias phenomenon.

(c) Suppose that boundary kernels are used so that the bias of $\hat{f}_h(x)$ is $O(h^2)$, uniformly in $x \in [0, 1]$. Then the mean integrated squared error (MISE) of $\hat{f}_h$ is, under the conditions of theorem 1,

$$(nh)^{-1} \left\{ c(0) + 2 \sum_{j=1}^{\infty} c(j) \right\} J_K + \frac{\sigma_K^4 h^4}{4} \int_0^1 \{f''(x)\}^2 \, dx + o((nh)^{-1} + h^4).$$

If follows from this expression that the Epanechnikov kernel $K(y) = 0.75(1 - y^2) I_{(-1, 1)}(y)$ is asymptotically MISE optimum among non-negative kernels, just as in the uncorrelated data case.

(d) Hart (1987) gives an asymptotic approximation for the MISE under a special case of equation (2.3) in which $c_n(k)/c_n(0) \to 1$ as $n \to \infty$. In addition to allowing for stronger dependence than in equation (2.4), the approximation allows for unequally spaced design points.

3. FAILURE OF CROSS-VALIDATION WHEN DATA ARE POSITIVELY CORRELATED

Diggle and Hutchinson (1989) and Hart and Wehrly (1986) provide examples in which traditional automatic smoothing methods, such as cross-validation, perform

poorly with correlated data. Further evidence of this will be seen in the simulation study of the next section. The purpose of this section is to show in a precise way when cross-validation can be expected to yield poor estimates of f.

Assuming that equation (1.1) holds, define the cross-validation curve C_n by

$$C_n(h) = \frac{1}{n}\sum_{j=1}^{n}\left\{\hat{f}_h^j\left(\frac{j-\frac{1}{2}}{n}\right) - y_j\right\}^2 I_{(b,1-b)}\left(\frac{j-\frac{1}{2}}{n}\right), \tag{3.1}$$

where $\hat{f}_h^j(x)$ is the estimate (1.2) calculated without the jth observation and $0 < b < \frac{1}{2}$. The indicator function in equation (3.1) is used to avoid problems with boundary effects. Let $\hat{h}_n$ be the minimizer of $C_n(h)$ for $n^{-1+\delta} \leqslant h \leqslant n$ $(\delta > 0)$. Härdle *et al.* (1988) investigated the performance of $\hat{h}_n$ when the data are independent and showed that $\hat{h}_n$ is asymptotic to $\tilde{h}_n$, the minimizer of the loss function

$$L_n(h) = \frac{1}{n}\sum_{j=1}^{n}\left\{\hat{f}_h\left(\frac{j-\frac{1}{2}}{n}\right) - f\left(\frac{j-\frac{1}{2}}{n}\right)\right\}^2 I_{(b,1-b)}\left(\frac{j-\frac{1}{2}}{n}\right).$$

The following theorem concerns the behaviour of $C_n(h)$ under the covariance structure defined by model (2.4). We reiterate that such a model for the covariance does not cover all the cases of practical interest. However, the main concern here is to establish an existence result on the extent to which the traditional form of cross-validation fails when the data are correlated.

Theorem 2. Let the conditions of theorem 1 hold, and suppose that $\{\epsilon_i\}$ is fourth order stationary. Define $\kappa_4(s-t, r, 0)$ to be the fourth joint cumulant of the distribution of $(\epsilon_t, \epsilon_{t+r}, \epsilon_s, \epsilon_{s+r})$ and assume that $|\kappa_4(m, r, 0)| \leqslant \beta_m$ for all m and r, where $\Sigma_{m=-\infty}^{\infty}\beta_m < \infty$. (If the ϵ_i are Gaussian, $\kappa_4 \equiv 0$.) Suppose also that

$$J_K\left\{c(0) + 2\sum_{j=1}^{\infty} c(j)\right\} < 2K(0)\left\{c(1) + 2\sum_{j=1}^{\infty} c(j)\right\}. \tag{3.2}$$

Let $C_n(h)$ be as in equation (3.1) with $b = Bn^{-1/5}$, where B is arbitrarily large but fixed. Let S_n be a set of cardinality $[Cn^{\beta}]$ $(C, \beta > 0)$ whose elements are in the interval $[n^{-1}, Bn^{-1/5}]$. If $\hat{h}_n$ is the minimizer of $C_n(h)$ among $h \in S_n$ and if $\beta < \alpha < \frac{4}{5}$, then

$$\lim_{n\to\infty} P(\hat{h}_n < n^{-1+\alpha}) = 1.$$

For the proof of this theorem see Appendix A.

Theorem 2 shows that, when the data are sufficiently positively correlated, cross-validation will choose a kernel estimate that very nearly interpolates the data. This phenomenon is explained quite simply by considering

$$E(C_n(h)) = E\left[\frac{1}{n}\sum_{j=1}^{n}\left\{\hat{f}_h^j\left(\frac{j-\frac{1}{2}}{n}\right) - f\left(\frac{j-\frac{1}{2}}{n}\right)\right\}^2 I_{(b,1-b)}\left(\frac{j-\frac{1}{2}}{n}\right)\right] + \frac{n_b}{n}\sigma^2 - CT(h),$$

where n_b is the number of non-zero terms in the sum and

$$CT(h) = \frac{2}{n}\sum_{j=1}^{n} \operatorname{cov}\left(\hat{f}_h^j\left(\frac{j-\frac{1}{2}}{n}\right), \epsilon_j\right) I_{(b,1-b)}\left(\frac{j-\frac{1}{2}}{n}\right).$$

Now, when the ϵ_j are uncorrelated, $CT(h) = 0$ and the minimizer of $EC_n(h)$ is essentially the minimizer of the risk function $EL_n(h)$. However, when the ϵ_j are positively correlated, $CT(h)$ is positive. If the correlation is sufficiently strong for condition (3.2) to hold, then as $nh_n \to \infty$ and $nh_n^5 \to 0$

$$E(C_n(h)) \approx \sigma^2 + c/nh_n,$$

where $c < 0$. Thus, when n is sufficiently large, $EC_n(h)$ will be minimized at a value of h that is very near zero. Alternatively, theorem 2 may be explained by first noting that cross-validation is designed to yield good predictors of the observations $y_1, \ldots, y_n$. When the data are uncorrelated, the best mean-squared error predictor of y_j given $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n$ is simply $f_j = f((j-\frac{1}{2})/n)$, and so it is not surprising that cross-validation tends to pick a kernel smoother that is a good estimate of f. However, when the data are correlated, the best predictor of y_j will depend on neighbouring data as well as on f_j. For example, if the errors follow a first-order autoregressive process, the minimum MSE predictor of y_j given the other data is equal to $f_j + c_\rho(\epsilon_{j-1} + \epsilon_{j+1})$, where $c_\rho = \rho/(1+\rho^2)$. When $\rho > 0$ and is sufficiently large, this predictor is well approximated by $\hat{f}_h^j((j-\frac{1}{2})/n)$ with h very small.

To give an indication of when condition (3.2) will be met, suppose that $c(k) = \sigma^2\rho^{|k|}$, $|k| = 0, 1, \ldots$, where $|\rho| < 1$. In this case the errors have the covariance function of a first-order autoregressive process. It may be verified that, for the Epanechnikov kernel, condition (3.2) is equivalent to $\rho > 0.164$, while, for the quartic kernel $K(y) = (15/16)(1-y^2)^2 I_{(-1,1)}(y)$, condition (3.2) becomes $\rho > 0.155$. Suppose now that the errors have the covariance function of a first-order moving average process, i.e. $c(1) = \sigma^2\rho$ ($|\rho| \leqslant 0.5$) and $c(k) = 0$ for $k \geqslant 2$. In this case condition (3.2) holds for the Epanechnikov kernel if and only if $\rho > 0.227$, and for the quartic kernel if and only if $\rho > 0.170$. The latter example shows that the most innocent of correlation structures can cause cross-validation to perform very poorly.

4. INCORPORATING COVARIANCE ESTIMATES INTO CHOICE OF BANDWIDTH

Correlation in the data can be accounted for in the bandwidth selection problem by using a risk estimation procedure. Define the mean average-squared error (MASE) curve by

$$M(h) = E\left[\frac{1}{n}\sum_{j=1}^{n}\left\{\hat{f}_h\left(\frac{j-\frac{1}{2}}{n}\right) - f\left(\frac{j-\frac{1}{2}}{n}\right)\right\}^2\right]$$

$$= E\left[\frac{1}{n}RSS(h)\right] - c(0) + 2n^{-1}\left\{w_0(h)\, c(0) + 2\sum_{j=1}^{n(h)} w_j(h)\, c(j)\right\}, \qquad (4.1)$$

where

$$RSS(h) = \sum_{j=1}^{n}\left\{\hat{f}_h\left(\frac{j-\frac{1}{2}}{n}\right) - y_j\right\}^2,$$

$$w_j(h) = (n-j) \int_{(j-\frac{1}{2})/nh}^{(j+\frac{1}{2})/nh} K(y)\, dy,$$

and $n(h)$ is the largest integer less than $nh+\frac{1}{2}$. (If boundary kernels are used, the definition of the $w_j(h)$ changes.) If the errors are uncorrelated, Rice (1984) proposes that $M(h)$ be estimated by

$$\hat{M}(h) = \frac{1}{n} RSS(h) - \hat{\sigma}^2 \left\{ 1 - 2 \int_{-1/2nh}^{1/2nh} K(y)\, dy \right\}, \tag{4.2}$$

where $\hat{\sigma}^2 = (2n)^{-1} \Sigma_{i=2}^{n} (y_i - y_{i-1})^2$. Rice's criterion can be modified to account for correlation if we know or have estimates of the covariances $c(k)$, $k = 0, 1, \ldots$. For a general covariance function $\tilde{c}$, define

$$\hat{M}(h; \tilde{c}) = \frac{1}{n} RSS(h) - \tilde{c}(0) + 2n^{-1} \left\{ w_0(h)\, \tilde{c}(0) + 2 \sum_{j=1}^{n(h)} w_j(h)\, \tilde{c}(j) \right\}. \tag{4.3}$$

As suggested by Hart and Wehrly (1986), it is natural to estimate $M(h)$ by $\hat{M}(h; \hat{c})$, where $\hat{c}(k)$ is some estimate of $c(k)$.

If selection of the bandwidth is to remain an objective procedure, it will be necessary to estimate $c(k)$ without having knowledge of f. Doing so is straightforward in the setting of Hart and Wehrly (1986), who consider a case where independent copies of a process are observed. If instead a single datum is recorded at each design point, estimating $c(k)$ without having a model for f becomes more difficult. By contrast, when f is in a parametric family, $c(k)$ can be well estimated without having an efficient estimate of f. Solo (1981) gives conditions on the error process under which ordinary least squares estimates of regression parameters are strongly consistent. Residuals from the ordinary least squares fit may be used to estimate the covariance function of the errors. In the Cochrane and Orcutt (1949) and Prais–Winsten procedures, this method of estimating $c(k)$ is used, and then the regression function is re-estimated using generalized least squares based on $\hat{c}(k)$. See Rao and Griliches (1969), Spitzer (1979) and Krämer (1980) for more on these two-stage and other procedures.

It is reasonable to regard ordinary cross-validation as the nonparametric analogue of parametric ordinary least squares. Theorem 2, however, is a good argument against carrying the analogy a step further by using ordinary cross-validation to obtain the initial $\hat{f}$ from which $c(k)$ is estimated. The residuals from such a rough estimate of f will be poor substitutes for the ϵ_i.

We now propose a way in which $c(k)$ can be estimated without having an initial estimate of f. The method involves differencing the data $y_1, \ldots, y_n$. Some theoretical aspects of the methodology are pursued in Hart (1989). Here we shall merely describe the method of estimating $c(k)$. Define

$$\Delta_j = y_{j+1} - 2y_j + y_{j-1}, \qquad j = 2, \ldots, n-1.$$

Since f'' is continuous,

$$\Delta_j = n^{-2} f''(x_j) + d_j, \tag{4.4}$$

where $(j-\frac{3}{2})/n \leqslant x_j \leqslant (j+\frac{1}{2})/n$ and $d_j = \epsilon_{j+1} - 2\epsilon_j + \epsilon_{j-1}$. Now, if model (2.4) holds, $\text{var}(d_j)$ is constant for all n, and so $\Delta_j \approx d_j$ when n is large. This suggests that the spectrum S_ϵ of the process $\{\epsilon_j\}$ can be consistently estimated from the data $\Delta_j, j = 2, \ldots, n-1$, since

$$S_d(\omega) = |1 - e^{i\omega}|^4 S_\epsilon(\omega), \qquad \omega \in [-\pi, \pi],$$

where S_d is the spectrum of $\{d_j\}$. Müller and Stadtmüller (1988) have used differences to estimate $c(k)$ when the errors form a moving average process. Gasser *et al.* (1986) propose the use of $\{\Delta_j\}$ to estimate the variance function in a heteroscedastic regression model.

The first step in the proposed method of estimating $c(k)$ is to compute the periodogram

$$I_\Delta(\omega) = \frac{1}{T_n} \left| \sum_{j=2}^{n-1} \Delta_j^* e^{-i\omega j} \right|^2, \qquad \omega \in [-\pi, \pi],$$

where

$$\Delta_j^* = t\left(\frac{j-\frac{1}{2}}{n}\right)\Delta_j,$$

$$T_n = 2\pi \sum_{j=2}^{n-1} t^2\left(\frac{j-\frac{1}{2}}{n}\right)$$

and t is a twice differentiable function that vanishes at zero and unity. See Hart (1989) for a discussion of the benefits of using the taper t. If a parametric model, $S(\omega; \theta)$, is assumed for $S_\epsilon(\omega)$, then, following Rice (1979), we can estimate θ by the maximizer of the approximate log-likelihood

$$\tilde{L}_n(\theta) = -\frac{2\pi}{n} \sum_{\omega_j \in \Lambda} \{\log S(\omega_j; \theta) + \tilde{I}_\epsilon(\omega_j)/S(\omega_j; \theta)\}, \tag{4.5}$$

where $\tilde{I}_\epsilon(\omega) = |1 - e^{i\omega}|^{-4} I_\Delta(\omega)$, $\omega_j = 2\pi j/n$, $\Lambda = [\delta, \pi]$ and $\delta > 0$. Owing to the one-to-one correspondence between $c(k)$ and S_ϵ, estimating θ yields an estimate of $c(k)$. Differencing the data does not (in general) completely eliminate the effect of f. Therefore, if δ is taken too small the estimate of θ will be biased because of the low frequency contribution of f to the periodogram.

A somewhat different approach to estimating θ is to use non-linear regression to fit the model $|1 - e^{i\omega}|^4 S(\omega; \theta)$ to $I_\Delta(\omega_j), j = 1, \ldots, [n/2]$. An advantage of this approach is that it is not nearly so important as before to choose δ since the periodogram $I_\Delta(\omega)$ need not be multiplied by the factor $|1 - e^{i\omega}|^{-4}$ before fitting a model. In either approach to estimating θ, it is advisable to difference the data more than twice and to repeat the estimation procedure. If $\hat{\theta}$ changes little, one feels confident that the deterministic component, f, has effectively been eliminated.

Before proceeding to a numerical study, a few comments are in order concerning the assumption of a parametric model for $c(k)$. This assumption may seem inconsistent with the notion that cross-validation should provide a completely objective way of estimating the mean function. However, it is inevitable that *some* assumption has to be made about the error process; otherwise the model composed of f and $c(k)$

is not even identifiable. We can also argue that the smoothing algorithm defined by the assumption that $c(k) = c(k;\theta)$ will lead to reasonable estimates of f for a much wider class of covariance functions than just $c(k;\theta)$. Finally, when the data are correlated, any allowance for correlation is preferable to using the ordinary form of cross-validation, as evidenced by theorem 2.

5. SIMULATION STUDY AND DATA ANALYSIS

A simulation study was undertaken to investigate the methodology presented in the previous section. The function considered was $f(x) = 16x^2(1-x)^2$ for $0 \leqslant x \leqslant 1$, and the error process was taken to be first order autoregressive, i.e. $\epsilon_1 \sim N(0, \sigma^2)$ and $\epsilon_i = \rho\epsilon_{i-1} + z_i$, $i = 2, \ldots, n$, where the z_i are independently and identically distributed $N(0, \sigma^2(1-\rho^2))$. Six combinations of ρ and σ ($\rho = 0, 0.35, 0.70$; $\sigma = 0.05, 0.20$) and the two sample sizes $n = 100$ and $n = 300$ were investigated. For each combination of n, ρ and σ, 100 independent sets of data were generated. For each set of data, three estimates of the covariance function were calculated. The three estimates were of the form $\hat{c}_\delta(k) = \hat{\sigma}_\delta^2\hat{\rho}_\delta^k$, $k \geqslant 0$, where $\hat{\sigma}_\delta$ and $\hat{\rho}_\delta$ were obtained by maximizing equation (4.5) for three choices of δ. For both sample sizes the choices for δ were 0.05π, 0.10π and 0.15π. In calculating $\tilde{I}_\epsilon(\omega)$, the taper

$$t(u) = \begin{cases} 10(10u)^3 - 15(10u)^4 + 6(10u)^5, & 0 \leqslant u \leqslant 0.1, \\ 1, & 0.1 < u \leqslant 0.5, \\ t(1-u), & 0.5 < u \leqslant 1, \end{cases}$$

was used throughout the study.

Having obtained a covariance function estimate $\hat{c}$, the criterion $\hat{M}(h;\hat{c})$ was calculated for 101 equally spaced values of h in $[1/n, 1/2]$, where the kernel employed was $K(y) = 0.75(1-y^2)\, I_{(-1,1)}(y)$. Three bandwidths, each the approximate minimizer of an estimated MASE curve, were thereby obtained for each data set. In addition, the approximate minimizer of $\hat{M}(h)$, the criterion that assumes the data are uncorrelated, was determined. Finally, to see how well the various bandwidths performed, the minimizer of

$$L_n(h) = \frac{1}{n}\sum_{j=1}^{n}\left\{\hat{f}_h\left(\frac{j-\frac{1}{2}}{n}\right) - f\left(\frac{j-\frac{1}{2}}{n}\right)\right\}^2$$

over 101 equally spaced values of h in $[1/n, 1/2]$ was determined for each set of data.

The results of the simulation are summarized in Tables 1 and 2. The nomenclature used in the ensuing discussion is explained in Table 1. One remarkable aspect of the study is how poorly the procedure UC performed when the data were correlated. The performance of UC (relative to the optimum) deteriorates with

(a) increasing ρ for n and σ fixed,
(b) increasing σ for n and $\rho > 0$ and
(c) increasing n for σ and $\rho > 0$ fixed.

It is encouraging that each of the covariance-based criteria outperformed UC when the data were correlated. This suggests that using an even somewhat inefficient $\hat{c}$ in $\hat{M}(h;\hat{c})$ is preferable to using $\hat{M}(h)$ when the data are correlated. Also encouraging is

TABLE 1
Means and standard deviations of various bandwidths obtained in the simulation study †

Bandwidth type	*Means and standard deviations (× 10000) for the following values of σ and ρ:*					
	σ = 0.05			*σ = 0.2*		
	ρ			*ρ*		
	0	*0.35*	*0.70*	*0*	*0.35*	*0.70*
n = 100						
OPT	687 (107)	838 (138)	943 (218)	1309 (275)	1487 (285)	1673 (337)
UC	731 (113)	418 (99)	356 (32)	1150 (352)	358 (222)	254 (28)
C1	765 (102)	904 (136)	1014 (288)	1311 (256)	1596 (346)	1882 (744)
C2	744 (111)	871 (142)	1058 (360)	1258 (297)	1550 (393)	2118 (1112)
C3	729 (115)	855 (172)	1067 (367)	1247 (312)	1483 (437)	2116 (1136)
n = 300						
OPT	600 (87)	668 (107)	792 (122)	995 (166)	1158 (201)	1405 (265)
UC	555 (112)	161 (55)	102 (25)	955 (225)	101 (37)	83 (0)
C1	562 (90)	654 (117)	806 (178)	977 (223)	1093 (295)	1368 (453)
C2	565 (103)	633 (151)	1105 (923)	974 (252)	1038 (352)	1463 (994)
C3	543 (121)	602 (187)	1151 (1065)	962 (239)	983 (386)	1610 (1380)

†The standard deviations are in parentheses. OPT denotes the minimizer of the loss $L_n(h)$. The other bandwidths minimize $\hat{M}(h; \hat{c})$ for some $\hat{c}$. C1, C2 and C3 use the parametric method of Rice for estimating $c(k)$ with, respectively, $\delta = 0.05\pi$, $\delta = 0.10\pi$ and $\delta = 0.15\pi$. UC is the criterion which assumes that the data are uncorrelated.

TABLE 2
Medians and 75th percentiles of $|\hat{h} - \hat{h}_0|$ *from the simulation study* †

Bandwidth type	*Medians and 75th percentiles (× 10000) for the following values of σ and ρ:*					
	σ = 0.05			*σ = 0.20*		
	ρ			*ρ*		
	0	*0.35*	*0.70*	*0*	*0.35*	*0.70*
n = 100						
UC	99 (198)	446 (545)	594 (743)	297 (631)	1139 (1386)	1436 (1634)
C1	99 (198)	99 (198)	198 (347)	248 (495)	396 (644)	495 (978)
C2	99 (198)	99 (235)	297 (483)	272 (495)	446 (681)	644 (1250)
C3	99 (198)	149 (248)	322 (495)	248 (545)	495 (693)	718 (1349)
n = 300						
UC	99 (199)	497 (596)	695 (783)	199 (348)	1043 (1192)	1291 (1490)
C1	99 (149)	99 (199)	149 (298)	199 (348)	298 (497)	447 (745)
C2	99 (199)	99 (199)	248 (348)	199 (385)	323 (546)	646 (894)
C3	99 (199)	149 (298)	323 (546)	199 (385)	397 (546)	844 (1130)

†The 75th percentiles are in parentheses. The quantity $\hat{h}_0$ is the minimizer of $L_n(h)$ for a given set of data while $\hat{h}$ is the minimizer of either $\hat{M}(h)$ or $\hat{M}(h; \hat{c})$. See Table 1 for an explanation of bandwidth types.

that, when the data were uncorrelated, the procedures C1, C2 and C3 performed about the same as UC. Except when ρ was 0.7, the covariance-based procedure was relatively insensitive to the choice of δ. Too large a choice for δ at $\rho = 0.7$ led to somewhat *oversmoothed* kernel estimates rather than to drastically undersmoothed estimates as in the UC method.

To illustrate the methodology presented in Section 4 further, we shall apply it to a time series consisting of the Beveridge (1921) index of wheat prices from the year 1500

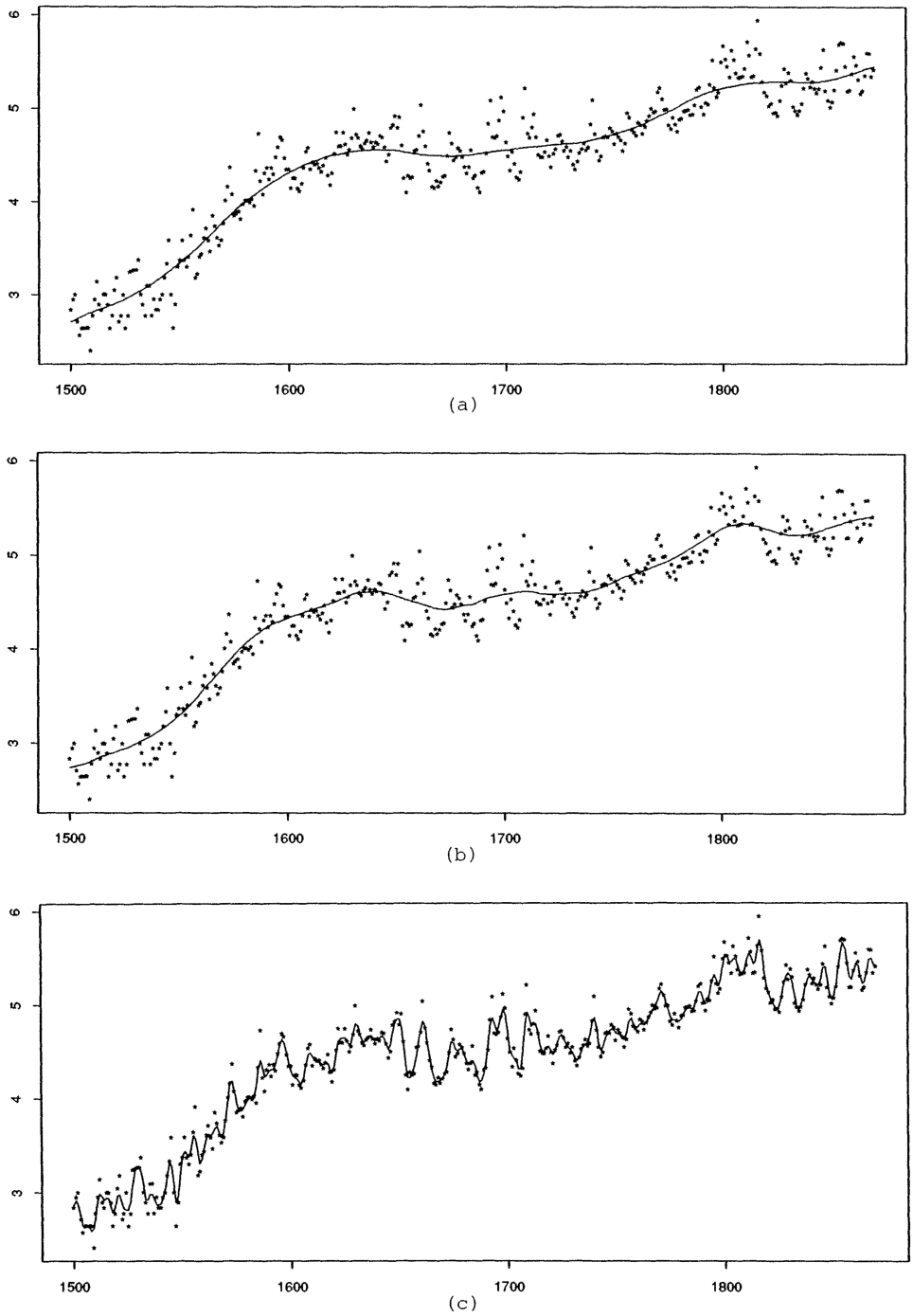

Fig. 1. Kernel estimates from the Beveridge wheat price index data: the horizontal axes are in years and each vertical axis is the natural logarithm of the price index; the bandwidths for (a) ($h = 37$) and (b) ($h = 24.1$) minimize equation (4.3) for covariance estimates based on, respectively, AR(1) and AR(2) models; the bandwidth for (c) ($h = 2.1$) is obtained on the assumption that the data are uncorrelated

to 1869 (see Anderson (1971)). These data are an annual index of prices at which wheat was sold in European markets. The data used for analysis were the natural logarithms of the Beveridge indices. This transformation was done to correct for heteroscedasticity in the original series. As can be seen from Fig. 1, the time series is clearly non-stationary. It is reasonable to account for the trend in the data by means of a kernel smoother. Having done so, we can then carefully analyse residuals to try to detect periodicities in the stationary part of the series. A boundary-modified Epanechnikov kernel was used in the analysis of the Beveridge data. (The Epanechnikov kernel was multiplied by an appropriate linear function for each point x within a bandwidth of either end of the interval [1499.5, 1869.5]. See Gasser and Müller (1979) for the details.)

Non-linear regression was used to fit AR(1) and AR(2) spectra to the periodograms of the second-, third- and fourth-differenced data. The fitted models were only trivially different for the three degrees of differencing, and hence the models obtained from the second differences were used to estimate the MASE. Two MASE estimates of the form (4.3) were calculated, one for each of the two autoregressive models. The covariance estimates were of the form $\hat{c}(k) = \tilde{\sigma}^2\rho(k;\hat{\phi})$ where $\rho(k;\phi)$ is the correlation function corresponding to an autoregression model with autoregressive parameters ϕ, $\hat{\phi}$ is the estimate of the autoregressive parameters from the differences analysis and

$$\tilde{\sigma}^2 = (2n)^{-1}\sum_{i=2}^{n}(y_i - y_{i-1})^2/\{1-\rho(1;\hat{\phi})\}.$$

This variance estimate was chosen to make the procedure more analogous to Rice's criterion in the setting of uncorrelated data.

The bandwidths minimizing the MASE estimates were 37 and 24 for, respectively, the AR(1) and AR(2) correlation models. The kernel estimates corresponding to these two bandwidths are shown in Fig. 1. The minimizer of equation (4.2) was also determined. Visually, the two correlation-based kernel estimates are little different from each other. The estimate based on the assumption of no correlation is very rough. Essentially, the latter estimate provides a good kernel interpolator of the data. This would be useful for estimating missing values in the time series. In contrast, the other two kernel estimates describe the most slowly varying part of the series.

Before concluding, a few remarks should be made about the methodology that has been proposed in this paper. The author has found this methodology to be a useful guide in arriving at an estimate of a mean function when the data are correlated. However, it should be clear that dependence in the data calls for a good deal of interaction between the analyst and the data. Under the most general assumptions on the mean function and the error process, no solely data-based procedure will be able to estimate consistently both the function and the error structure. Even when the data are independent, the results of Härdle *et al.* (1988) indicate that cross-validated bandwidths converge very slowly to the optimum bandwidth. Thus, under the best of circumstances it seems advisable to maintain a healthy scepticism about the answers obtained from automated procedures.

ACKNOWLEDGEMENTS

The advice of Ted Hannan was extremely helpful in formulating the covariance

estimates of Section 4. The author also benefited from conversations with Tom Wehrly and Daren Cline. Finally, the author is grateful to the referees and Associate Editor, whose comments led to this more expository version of the original manuscript.

This research was supported in part by Office of Naval Research contract N00014-85-K-0723.

APPENDIX A: PROOF OF THEOREM 2

Let $b_n = \log n + 1$. In what follows we assume that $n > n_0$, where $b_n < n^\alpha$ for all $n > n_0$. Then $P(\hat{h}_n < n^{-1+\alpha}) \geqslant P(\cap_{h\in\Gamma_n}\{C_n(b_n/n) < C_n(h)\})$, where $\Gamma_n = S_n \cap [n^{-1+\alpha}, Bn^{-1/5}]$. Define

$$\beta_n(h) = C_n(h) - (1/n)\sum_{j=1}^{n} \epsilon_j^2 I_{n,j} = C_n(h) - \hat{\sigma}_n^2,$$

where $I_{n,j} = I_{(Bn^{-1/5},\,1-Bn^{-1/5})}((j-\frac{1}{2})/n)$. It follows that $P(\cap_{h\in\Gamma_n}\{C_n(b_n/n) < C_n(h)\}) \geqslant P(E_n)$, where $E_n = \{b_n\beta_n(b_n/n) < -\eta + \epsilon\} \cap \{\sup_{h\in\Gamma_n} b_n|\beta_n(h)| < \eta - \epsilon\}$ and $0 < \epsilon < \eta$. We have

$$P(E_n) \geqslant P(\sup_{h\in\Gamma_n} b_n|\beta_n(h)| < \eta - \epsilon) + P(b_n\beta_n(b_n/n) < -\eta + \epsilon) - 1.$$

Now, $|\beta_n(h)| \leqslant A_n(h) + 2\hat{\sigma}_n\{A_n(h)\}^{1/2}$, where

$$A_n(h) = (1/n)\sum_{j=1}^{n}\{\hat{f}_h^j((j-\tfrac{1}{2})/n) - f((j-\tfrac{1}{2})/n)\}^2 I_{n,j}.$$

By relating $A_n(h)$ to $L_n(h)$ and using lemma 1 in Hart (1987), it is now straightforward to show that $\sup_{h\in\Gamma_n} b_n|\beta_n(h)|$ converges to zero in probability.

Since η can be chosen arbitrarily small, theorem 2 will be proven if we can show that $b_n\beta_n(b_n/n)$ converges in probability to a negative number. It can be shown that, as $nh \to \infty$ and $h \to 0$,

$$\beta_n(h) = \mathrm{ASE}(h) - \alpha_n(h) + o_p(nh)^{-1}, \tag{A.1}$$

where

$$\mathrm{ASE}(h) = \frac{1}{n}\sum_{j=1}^{n}\left\{\hat{f}_h\left(\frac{j-\frac{1}{2}}{n}\right) - f\left(\frac{j-\frac{1}{2}}{n}\right)\right\}^2 I_{n,j}$$

and

$$\alpha_n(h) = \frac{2}{n}\sum_{j=1}^{n}\epsilon_j\hat{f}_h^j\left(\frac{j-\frac{1}{2}}{n}\right)I_{n,j}.$$

Using equation (A.1), theorem 1 and techniques similar to those used to prove theorem 2 in Hart (1987), we have

$$b_n\,E(\beta_n(b_n/n)) \to J_K\left\{c(0) + 2\sum_{j=1}^{\infty}c(j)\right\} - 2K(0)\left\{c(1) + 2\sum_{j=1}^{\infty}c(j)\right\}.$$

Since this quantity is negative under condition (3.2), the theorem will be proven if we show

$$b_n\{\mathrm{ASE}(b_n/n) - E(\mathrm{ASE}(b_n/n))\} \xrightarrow{p} 0$$

and

$$b_n\{\alpha_n(b_n/n) - E(\alpha_n(b_n/n))\} \xrightarrow{p} 0.$$

Since the two cases are handled similarly, we only consider the second. We can write $\alpha_n(h)$ as

$$\gamma_n(h) + \frac{1}{n}\sum_{j=1}^{n} \epsilon_j Z_{j,n,h} I_{n,j}, \tag{A.2}$$

where $\gamma_n(h) = (2/n)\Sigma_{j=1}^{n} \epsilon_j \hat{f}_h((j-\frac{1}{2})/n) I_{n,j}$, $Z_{j,n,h} = (nh)^{-1}\Sigma_{k=-1}^{1} w_{kj}(n,h) y_{j+k}$, and the $w_{kj}(n,h)$ are constants bounded uniformly in k, j, n and h. We have

$$\gamma_n(h) = \frac{1}{nh}\sum_{i=-n(h)}^{n(h)} K(z_i)\left[\frac{1}{n}\sum_{j=1}^{n}\left\{\epsilon_j\epsilon_{j-i} + \epsilon_j f\left(\frac{j-i-\frac{1}{2}}{n}\right)\right\} I_{n,j}\right],$$

and hence

$$nh\{\gamma_n(h) - E(\gamma_n(h))\} = \sum_{i=-n(h)}^{n(h)} K(z_i)\left\{\hat{c}_i(i) - E(\hat{c}_i(i)) + \frac{1}{n}\sum_{j=1}^{n}\epsilon_j f\left(\frac{j-i-\frac{1}{2}}{n}\right) I_{n,j}\right\},$$

where $\hat{c}_i(i) = (1/n)\Sigma_{j=1}^{n}\epsilon_j\epsilon_{j-i} I_{n,j}$. To show that $nh\{\gamma_n(h) - E(\gamma_n(h))\}$ converges to zero in probability it is sufficient to show that

$$\sum_{i=-n(h)}^{n(h)}\left[\{\mathrm{var}(\hat{c}_i(i))\}^{1/2} + \left\{\mathrm{var}\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j f\left(\frac{j-i-\frac{1}{2}}{n}\right) I_{n,j}\right)\right\}^{1/2}\right] \tag{A.3}$$

tends to zero as $nh \to \infty$ and $h \to 0$. By the conditions imposed on $c(k)$, the two variances in expression (A.3) are bounded uniformly in i by A/n for some $A > 0$ (see Priestley (1981)), and so expression (A.3) is $O(h\sqrt{n})$. From this it is clear that $b_n\{\gamma_n(b_n/n) - E(\gamma_n(b_n/n))\}$ tends to zero in probability. The second term in expression (A.2) is handled similarly and so the proof is complete.

REFERENCES

Anderson, T. W. (1971) *The Statistical Analysis of Time Series*. New York: Wiley.

Beveridge, W. H. (1921) Weather and harvest cycles. *Econ. J.*, **31**, 429–452.

Brereton, R. G. (1987) Spectral analysis of multivariate geochemical time series. *Chemometr. Intell. Lab. Syst.*, **2**, 177–185.

Cochrane, D. and Orcutt, G. H. (1949) Application of least squares regression to relationships containing autocorrelated error terms. *J. Am. Statist. Ass.*, **44**, 32–61.

Diggle, P. J. (1985) Discussion on Some aspects of the spline smoothing approach to non-parametric regression curve fitting (by B. W. Silverman). *J. R. Statist. Soc.* B, **47**, 28–29.

Diggle, P. J. and Hutchinson, M. F. (1989) On spline smoothing with autocorrelated errors. *Aust. J. Statist.*, **31**, 166–182.

Gasser, Th. and Müller, H. G. (1979) Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (eds Th. Gasser and M. Rosenblatt), pp. 23–68. Heidelberg: Springer.

——— (1984) Nonparametric estimation of regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, **11**, 171–185.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum (with discussion)? *J. Am. Statist. Ass.*, **83**, 86–101.

Härdle, W. and Tuan, P.-D. (1986) Some theory on M-smoothing of time series. *J. Time Ser. Anal.*, **7**, 191–204.

Hart, J. D. (1987) Kernel smoothing when the observations are correlated. *Technical Report 35*. Department of Statistics, Texas A&M University, College Station.

—— (1989) Differencing as an approximate de-trending device. *Stoch. Processes Appl.*, **31**, 251–259.

Hart, J. D. and Wehrly, T. E. (1986) Kernel regression estimation using repeated measurements data. *J. Am. Statist. Ass.*, **81**, 1080–1088.

Krämer, W. (1980) Finite sample efficiency of ordinary least squares in the linear regression model with auto-correlated errors. *J. Am. Statist. Ass.*, **75**, 1005–1009.

Müller, H. G. and Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610–625.

—— (1988) Detecting dependencies in smooth regression models. *Biometrika*, **75**, 639–650.

Priestley, M. B. (1981) *Spectral Analysis and Time Series*, pp. 325–326. New York: Academic Press.

Rao, P. and Griliches, Z. (1969) Small-sample properties of several two-stage regression methods in the context of autocorrelated errors. *J. Am. Statist. Ass.*, **64**, 253–272.

Rice, J. (1979) On the estimation of the parameters of a power spectrum. *J. Multiv. Anal.*, **9**, 378–392.

—— (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215–1230.

Solo, V. (1981) Strong consistency of least squares estimators in regression with correlated disturbances. *Ann. Statist.*, **9**, 689–693.

Spitzer, J. (1979) Small-sample properties of nonlinear least squares and maximum likelihood estimators in the context of autocorrelated errors. *J. Am. Statist. Ass.*, **74**, 41–47.

Stein, M. L. (1987) Minimum norm quadratic estimation of spatial variograms. *J. Am. Statist. Ass.*, **82**, 765–772.