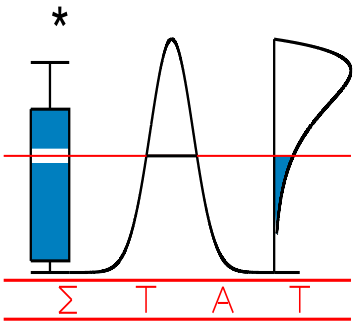
T E C H N I C A L

R E P O R T

0203

USING DIFFERENCE-BASED METHODS FOR INFERENCE IN NONPARAMETRIC REGRESSION WITH TIME-SERIES ERRORS

HALL, P. and I. VAN KEILEGOM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

USING DIFFERENCE-BASED METHODS FOR INFERENCE IN NONPARAMETRIC REGRESSION WITH TIME-SERIES ERRORS

Peter Hall[1] and Ingrid Van Keilegom[1,2]

ABSTRACT. We show that difference-based methods can be used to construct simple and explicit estimators of error covariance and autoregressive parameters in nonparametric regression with time-series errors. When the error process is Gaussian our estimators are efficient, but they are available well beyond the Gaussian case. As an illustration of their usefulness we show that difference-based estimators can be used to produce a simplified version of time-series cross-validation (TSCV). This new approach produces a bandwidth selector that is equivalent, to both first and second orders, to that given by the full TSCV algorithm. Other applications of difference-based methods are to variance estimation and construction of confidence bands in nonparametric regression.

SHORT TITLE. Difference-based methods.

KEYWORDS. Autoregression, bandwidth, covariance, cross-validation, kernel methods, linear time-series, local linear regression, time-series cross-validation.

[1]Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2]Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

The research of Van Keilegom was supported by "Projet d'Actions de Recherche Concertées", No. 98/03–217, and by an IAP research network from the Belgian government.

1. INTRODUCTION

We suggest difference-based methods for estimating error covariance and autoregressive parameters in nonparametric regression with time-series errors. Our approach is related to difference-based techniques for inference in nonparametric regression with independent errors, and uses the Yule-Walker equations to link autoregressive structure to covariance. In the case of Gaussian errors the estimators it produces are efficient, but it is available well beyond the Gaussian case.

Estimators of error covariance structure are needed to solve a range of problems connected to nonparametric regression. For example, they are required for estimating the variance of estimators of the regression function in nonparametric regression. The asymptotic formula for variance is

$$\left\{\gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)\right\}\kappa\Big/ nh\, f(x)\,, \tag{1.1}$$

where $\gamma(j)$ denotes the error covariance at lag j, n is sample size, h equals bandwidth, $f(x)$ is the density of design at the point x of estimation, and κ is a constant depending only on estimator and kernel types. See, for example, Opsomer, Wang and Yang (2001). We shall show how to use our difference-based approach to estimate the zero-spectrum $\tau^2 \equiv \gamma(0) + 2\sum_{j\geq 1}\gamma(j)$, at (1.1). The covariances (and autoregressive parameters) are needed too when using bootstrap methods to construct confidence bands for the regression mean.

A third application is simplifying Hart's (1994) time-series cross-validation (TSCV) method for bandwidth choice. An important step in TSCV is simultaneously fitting the bandwidth and autoregressive parameters. As it turns out, the parameters can be expressed as a function of the bandwidth, which implies that the cross-validation criterion effectively depends only on h. On the other hand, the autoregressive parameters need to be calculated for every candidate for the bandwidth that minimises the cross-validation criterion. Using difference-based methods, however, the autoregressive parameters can be estimated simply and directly, and the estimators do not rely on a bandwidth. Consequently, our form of Hart's cross-validation criterion depends on the bandwidth only via parameter estimators of the regression function, and not via the autoregressive parameters.

Moreover, the autoregressive parameter estimators that result from the new procedure are asymptotically equivalent to those produced by Hart's (1994) tech-

nique, and in particular have identical asymptotic distributions; and our empirical bandwidth selector is equivalent, to both first and second orders, to Hart's (1994) selector, and again has the same limiting distribution. Thus, by applying difference-based methods we can reduce the computational complexity of TSCV without appreciably affecting the final numerical result.

Difference-based methods for estimating variance in nonparametric regression with independent errors have been discussed by Rice (1984), Gasser, Sroka and Jennen-Steinmetz (1986), Müller and Stadtmüller (1987, 1989), Müller (1988, p. 99ff) and Hall, Kay and Titterington (1990). Techniques based on differencing have an extensive history in time series; see for example Anderson (1971, p. 66).

Härdle, Lütkepohl and Chen (1997) have reviewed nonparametric approaches to time-series analysis. Methodology for nonparametric regression with correlated errors, including different approaches to bandwidth choice, has been surveyed and discussed by Opsomer, Wang and Yang (2001). This article includes an account of the work of Hart (1991, 1994), who showed that when the errors are positively correlated, conventional cross-validation approaches, without an attempt to model the structure of the error process, tend to produce bandwidths that seriously undersmooth. See also Hart and Yi (1998).

Work surveyed by Opsomer, Wang and Yang (2001) includes that of Chu and Marron (1991), who gave a detailed theoretical description of second-order properties of conventional cross-validation in a time-series setting. Ray and Tsay (1997) suggested a plug-in rule for bandwidth choice in the case of long-range dependent errors; see Hermann, Gasser and Kneip (1992) for the short-range case. Yao and Tong (1998) discussed first-order properties of bandwidth selectors in a particularly general context. Hyndman and Wand (1997) suggested nonparametric approaches to estimation of the covariance function. The work in the present paper will address particularly the instance of nonparametric inference using local linear methods, and there the time-series setting has recently been treated by Opsomer (1996), Anh, Wolff, Gao and Tieng (1999), Masry and Mielniczuk (1999), Cai and Masry (2000), Cai and Elias (2001) and Francisco-Fernandez and Vilar-Fernandez (2001).

2. METHODOLOGY AND MAIN PROPERTIES

2.1. Model. Assume that the observed data (X_i, Y_i), for $1 \leq i \leq n$, are generated

by the model $Y_i = g(X_i) + \epsilon_i$. Here g is a smooth function, and we suppose that

$$\text{the } X_i\text{'s form an increasing sequence on the interval } \mathcal{I} = [0,1] \text{ and are either equally spaced there or are ordered values of independent and identically distributed random variables whose density has support equal to } \mathcal{I}. \tag{2.1}$$

The stochastic design case is useful in problems where data are recorded at "regular" but not-quite-equally-spaced times. For example, this is often the case when nightly observations are made of a star's radiative intensity.

It will be assumed that the errors $\epsilon_1, \ldots, \epsilon_n$ form a segment of a doubly infinite stationary linear time series with zero mean. Specifically, we ask that for some $p \geq 1$,

$$\epsilon_i = \sum_{j=1}^{p} \phi_j \, \epsilon_{i-j} + Z_i \,, \tag{2.2}$$

where $\{Z_i : -\infty < i < \infty\}$ are independent and identically distributed random variables with finite variance σ^2 and zero mean, independent also of the X_i's, and the constants $\phi_1, \ldots, \phi_p$ are such that the equation $1 - \sum_j \phi_j \, z^j = 0$ has no complex roots satisfying $|z| \leq 1$. Since the latter series has only a finite number of nonzero terms then an equivalent assumption is:

$$\begin{aligned} &\text{for some } \eta > 0, \text{ the equation } 1 - \textstyle\sum_j \phi_j \, z^j = 0 \\ &\text{has no complex roots satisfying } |z| \leq 1 + \eta \,. \end{aligned} \tag{2.3}$$

This condition implies that the process $\{\epsilon_i\}$ is causal.

2.2. Direct estimation of $\phi_1, \ldots, \phi_p$. The method that we suggest is based on pairwise differences of the Y_i's. Differences among three or more Y_i's may also be used, but the additional complication does not produce asymptotic improvements in performance.

Given an integer $j \geq 1$, define the difference operator D_j by $(D_j \, Y)_i = Y_i - Y_{i-j}$. If the function g is smooth on $\mathcal{I}$, and if either the X_i's are equally spaced on $\mathcal{I}$, or the distribution that generated them has a density that is bounded away from zero on $\mathcal{I}$, then the differences $\{D_j \, g(X)\}_i$ of the sequence of values $g(X_i)$ are generally small in size.

Indeed, to dwell on technicalities for a moment, we note that, assuming g has a bounded derivative,

$$\max_{1 \leq i \leq n} \big|\{D_j \, g(X)\}_i\big| = O\big\{n^{-1} \, \max(j, \log n)\big\} \tag{2.4}$$

with probability 1 as $n \to \infty$. (The remainder is $O(jn^{-1})$ if the design points are equally spaced. In the case of stochastic design the $\log n$ factor arises through properties of maximal spacings of order statistics.) Therefore, the influence of differences of the $g(X_i)$'s on values of the process of differences of the Y_i's, i.e. on

$$(D_j\, Y)_i = \{D_j\, g(X)\}_i + (D_j\, \epsilon)_i\,, \tag{2.5}$$

is in most instances negligible. Hence, $D_j\, Y$ is essentially just the sequence of differences of the error sequence.

Let $\gamma(j) = \mathrm{cov}(\epsilon_i, \epsilon_{i-j})$. Properties noted in the previous paragraph, and the identity $\frac{1}{2}\, E\{(D_j\, \epsilon)_i\}^2 = \gamma(0) - \gamma(j)$, motivate the following estimators of $\gamma(0)$ and $\gamma(j)$:

$$\begin{aligned} \widehat{\gamma}(0) &= \frac{1}{m_2 - m_1 + 1} \sum_{m=m_1}^{m_2} \frac{1}{2\,(n-m)} \sum_{i=m+1}^{n} \{(D_m Y)_i\}^2\,, \\ \widehat{\gamma}(j) &= \widehat{\gamma}(0) - \frac{1}{2\,(n-j)} \sum_{i=j+1}^{n} \{(D_j\, Y)_i\}^2\,, \end{aligned} \tag{2.6}$$

where $j \ge 1$ in the second definition. Put $\widehat{\gamma}(-j) = \widehat{\gamma}(j)$.

In the definition of $\widehat{\gamma}(0)$ we are relying on the fact that covariances among the errors ϵ_i decay exponentially quickly with increasing lag. Indeed, suppose the error distribution has finite fourth moment and that

$$m_1 \le m_2\,, \quad m_1/\log n \to \infty \quad \text{and} \quad m_2 = O\big(n^{1/2}\big)\,. \tag{2.7}$$

It may be proved from (2.4) and (2.5) that, under (2.7),

$$\max_{0 \le j \le n} \big|\widehat{\gamma}(j) - \gamma(j)\big| = O_p\big(n^{-1/2}\big)\,. \tag{2.8}$$

Given $\widehat{\gamma}(j)$, for $1 \le j \le p$, we construct estimators of ϕ_j using the Yule-Walker equations. Specifically, since $\gamma(k) = \sum_{j \ge 1} \phi_j\, \gamma(k-j)$ for $k \ge 1$, then, taking A to be the $p \times p$ matrix having $\gamma(j_1 - j_2)$ as its (j_1, j_2)th element, and putting $\gamma_j = \gamma(j)$, we have $(\phi_1, \ldots, \phi_p)^{\mathrm{T}} = A^{-1}(\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$. Hence, replacing γ by $\widehat{\gamma}$ in the definition of A, to obtain the estimator $\widehat{A}$; and putting $\widehat{\gamma}_j = \widehat{\gamma}(j)$; we define

$$(\hat{\phi}_1, \ldots, \hat{\phi}_p)^{\mathrm{T}} = \widehat{A}^{-1}(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p)^{\mathrm{T}}\,.$$

It follows directly from this definition, and (2.8), that

$$\max_{0\le j\le n} \bigl|\hat\phi_j - \phi_j\bigr| = O_p\bigl(n^{-1/2}\bigr)\,. \tag{2.9}$$

Moreover, as we shall show in section 4.1, if the error distribution has finite fourth moment and (2.7) is strengthened to

$$m_2 - m_1 \to \infty\,, \quad m_1/\log n \to \infty \quad \text{and} \quad m_2 = o\bigl(n^{1/2}\bigr)\,, \tag{2.10}$$

then

$$(\hat\phi_1, \ldots, \hat\phi_p)^{\mathrm{T}} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}} + n^{-1} \sum_{i=1}^{n} \epsilon_i\, A^{-1}\, (Z_{i+1}, \ldots, Z_{i+p})^{\mathrm{T}} + o_p\bigl(n^{-1/2}\bigr)\,. \tag{2.11}$$

We shall note in section 2.6 that the more complex, implicitly-defined estimators suggested by Hart (1994) admit the same expansion, and so are equivalent to our estimators $\hat\phi_j$ up to terms that are of smaller order than $n^{-1/2}$.

It follows from (2.11) (or, in the case of Hart's estimators, the respective version of (2.11)) that the estimator of $(\phi_1, \ldots, \phi_p)$ is asymptotically normally distributed with covariance matrix $n^{-1}\sigma^2 A^{-1}$. When the error process $\{\epsilon_i\}$ is Gaussian and directly observed, the covariance matrix of the maximum likelihood estimator of $(\phi_1, \ldots, \phi_p)$ is also equal to $n^{-1}\sigma^2 A^{-1}$. Therefore, in the more general context of nonparametric regression with Gaussian autoregressive errors, where the regression mean has a bounded derivative, our estimators of the regression coefficients, and also the estimators of Hart (1994), are efficient. Likewise, the corresponding estimators of covariance are efficient.

2.3. Estimation of $\gamma(j)$. In principle, the estimators of $\gamma(j)$ suggested at (2.6) may be employed for arbitrarily large values of j. However, these particular estimators do not exploit the autoregressive structure of the process ϵ_i. This can result in the covariance sequence $\{\widehat\gamma(j_1 - j_2) : -\infty < j_1, j_2 < \infty\}$ not being positive definite, and in $\widehat\gamma(j)$ being unnecessarily highly variable for large j. An alternative approach, available once the estimators $\hat\phi_j$ have been constructed, is to estimate $\gamma(j)$ by $\widetilde\gamma(j)$, defined as follows. Let $\bar\gamma(j)$ denote the coefficient of z^j, or equivalently of z^{-j}, in an expansion of

$$\Bigl(1 - \sum_{j=1}^{p} \hat\phi_j\, z^j\Bigr)^{-1} \Bigl(1 - \sum_{j=1}^{p} \hat\phi_j\, z^{-j}\Bigr)^{-1}\,,$$

and put $\hat\sigma^2 = \widehat\gamma(0)/\bar\gamma(0)$, an estimator of $\sigma^2 = \mathrm{var}(Z_i)$. Then define $\widetilde\gamma(j) = \hat\sigma^2\,\bar\gamma(j)$ for $j \ge 1$. Hence, an estimator of the zero-spectrum $\tau^2 \equiv \gamma(0) + 2\sum_{j\ge1}\gamma(j)$, appearing at (1.1), is given by $\hat\tau^2 = \hat\sigma^2\,\bigl(1 - \sum_{1\le j\le p}\hat\phi_j\bigr)^{-2}$.

The definition of $\widehat\gamma(0)$ implies that that quantity is positive. Moreover, since $\bar\gamma(0)$ equals the error variance in any autoregression (defined conditional on the data) that has coefficients $\hat\phi_j$ (for $1 \le j \le p$) and independent disturbances with unit variance, then $\bar\gamma(0) > 0$. These properties imply that $\hat\sigma^2$ is indeed positive. Moreover, the covariance sequence $\{\widetilde\gamma(j_1 - j_2) : -\infty < j_1, j_2 < \infty\}$ is guaranteed to be positive definite, since it is the covariance sequence for any autoregression in which the disturbance variance is $\hat\sigma^2$ and the autoregressive coefficients are $\hat\phi_j$, conditional on the data. It may be proved that, provided (2.3) and (2.7) hold and the error distribution has finite fourth moment, $\hat\tau^2 = \tau^2 + O_p(n^{-1/2})$.

To appreciate the origins of the estimator $\widetilde\gamma(j)$, observe that we may write $\epsilon_i = \sum_{k\ge0}\psi_k\,Z_{i-k}$ where $\psi_0 = 1$ and $\psi_1, \psi_2, \ldots$ are defined by

$$1 + \sum_{j=1}^{\infty}\psi_j\,z^j = \biggl(1 - \sum_{j=1}^{p}\phi_j\,z^j\biggr)^{-1}.$$

Thus, the covariance $\gamma(j)$ of the error process $\{\epsilon_i\}$ is given by

$$\gamma(j) = \sigma^2\sum_{k=0}^{j}\psi_k\,\psi_{k-j}\,,$$

which in turn equals σ^2 multiplied by the coefficient of z^j in an expansion of

$$\biggl(1 - \sum_{j=1}^{p}\phi_j\,z^j\biggr)^{-1}\biggl(1 - \sum_{j=1}^{p}\phi_j\,z^{-j}\biggr)^{-1}.$$

2.4. Curve estimators. A conventional, two-sided, local-linear smoother estimates $g(x)$ by

$$\hat g(x) = \frac{\sum_{1\le k\le n} w_k(x)\,Y_k}{\sum_{1\le k\le n} w_k(x)}\,, \tag{2.12}$$

where

$$w_k(x) = K\Bigl(\frac{x - X_k}{h}\Bigr)\,\bigl\{s_2(x) - (x - X_k)\,s_1(x)\bigr\}\,,$$
$$s_\ell(x) = (nh)^{-1}\sum_{k=1}^{n}K\Bigl(\frac{x - X_k}{h}\Bigr)\,(x - X_k)^\ell\,,$$

K is a kernel function and h is a bandwidth.

Initially, however, we construct a one-sided version of $\tilde{g}(x)$; it is

$$\tilde{g}(x) = \frac{\sum_k^{(x)} v_k(x)\, Y_k}{\sum_k^{(x)} v_k(x)}\,, \tag{2.13}$$

where $\sum_k^{(x)}$ denotes summation over k such that $X_k < x$,

$$v_k(x) = L\Big(\frac{x - X_k}{h_1}\Big)\,\big\{r_2(x) - (x - X_k)\, r_1(x)\big\}\,,$$
$$r_\ell(x) = (nh_1)^{-1} \sum_k^{(x)} L\Big(\frac{x - X_k}{h_1}\Big)\,(x - X_k)^\ell\,,$$

and L and h_1 are potentially a new kernel and new bandwidth.

In contrast to the case of local-constant methods treated by Hart (1994), in the setting of local linear techniques the kernel L does not need to be one-sided. Indeed, if L is a conventional kernel, in particular a symmetric probability density, then $E(\tilde{g}) - g = O(h^2)$ rather than $O(h)$. This follows from the property that $\sum_k^{(x)} v_k(x)\,(x - X_k) = 0$.

2.5. Cross-validation criterion. Define $\hat{\phi}_1, \ldots, \hat{\phi}_p$ as suggested in section 2.2, and construct the estimator $\tilde{g}$ as suggested in section 2.4. Put $\tilde{g}_i = \tilde{g}(X_i)$,

$$\widehat{Y}_i(h) = \tilde{g}_i + \sum_{j=1}^{p} \hat{\phi}_j\,(Y_{i-j} - \tilde{g}_{i-j})\,; \tag{2.14}$$

the dependence on h of course comes about through the fact that $\tilde{g}_i$ is a function of that quantity. Put

$$\mathrm{CV}_1(h) = \frac{1}{n - \ell + 1} \sum_{i=\ell}^{n} \{Y_i - \widehat{Y}_i(h)\}^2\,,$$

where ℓ denotes the least integer such that, for the range of values h considered, computation of $\tilde{g}_{\ell-p}$ does not require design points further to the left than X_1. Our empirical bandwidth selector is $h = \hat{h}_1$, the minimiser of $\mathrm{CV}_1(h)$.

An alternative approach, suggested by Hart (1994), is to use the bandwidth $h = \hat{h}_2$ that, along with $\phi_1, \ldots, \phi_p$, minimises

$$\mathrm{CV}_2(\phi_1, \ldots, \phi_p, h) = \frac{1}{n - \ell + 1} \sum_{i=\ell}^{n} \big\{Y_i - \widehat{Y}_i(\phi_1, \ldots, \phi_p, h)\big\}^2\,,$$

where

$$\widehat{Y}_i(\phi_1, \ldots, \phi_p, h) = \tilde{g}_i + \sum_{j=1}^{p} \phi_j\,(Y_{i-j} - \tilde{g}_{i-j})\,.$$

The rationale behind either method is that $\hat{h}_j$ should offer an accurate approximation to the bandwidth h_0 that minimises mean integrated squared error (MISE) for $\tilde{g}$, defined at (2.13):

$$\mathrm{MISE}_{\tilde{g}}(h) = \int_{\mathcal{I}} E\big\{\tilde{g}(x) - g(x)\big\}^2 f(x)\, dx\,,$$

where f denotes the design density. (Therefore, $f \equiv 1$ if the X_i's are equally spaced on the unit interval $\mathcal{I}$.)

As Hart (1994) observed, h_0 is asymptotically proportional to the bandwidth h_{opt} that minimises MISE for $\hat{g}$, defined at (2.12):

$$\mathrm{MISE}_{\hat{g}}(h) = \int_{\mathcal{I}} E\big\{\hat{g}(x) - g(x)\big\}^2 f(x)\, dx\,,$$

in the sense that $h_{\mathrm{opt}}/h_0 \to R(K, L)$ as $n \to \infty$, where $R(K, L)$ depends only on K and L. Therefore, $\hat{h}_j\, R(K, L)$ is a practicable bandwidth selector for use when estimating g by $\hat{g}$. Note that when $L = K$, $R(K, L) = 2^{-1/5}$ regardless of the choice of K. Hence, choosing L equal to K offers the additional advantage that $R(K, L)$ does not need to be worked out.

2.6. Theoretical properties of cross-validation. Hart (1994) argued that, if $(\tilde{\phi}_1, \ldots, \tilde{\phi}_p, \hat{h}_2)$ and h_0 denote the quantities that respectively minimise $\mathrm{CV}_2(\phi_1, \ldots, \phi_p, h)$ and $\mathrm{MISE}_{\tilde{g}}(h)$, then $\hat{h}_2/h_0 \to 1$ in probability. This result may be refined by showing, analogously to Chu and Marron (1991), that

$$n^{1/10}\,(\hat{h}_2 - h_0)/h_0 \to \mathrm{N}(0, \lambda)\,, \tag{2.15}$$

where $0 < \lambda < \infty$ and the convergence is in distribution. Assuming the model described in section 2.1, sufficient regularity conditions for (2.15) are (2.1), (2.3),

$$\begin{minipage}{0.85\textwidth}$E(|\epsilon_i|^c) < \infty$ for all $c > 0$, g has two continuous derivatives on the interval $\mathcal{I}$, h is constrained to lie in the interval $\mathcal{H} = [n^{\delta-1}, n^{-\delta}]$ for some $\delta \in (0, \frac{1}{5})$, and K is a symmetric, compactly supported probability density with a Hölder continuous derivative.\end{minipage} \tag{2.16}$$

We shall show in section 4.2 that, under the same assumptions, our bandwidth selector $\hat{h}_1$ also satisfies (2.15), and in fact

$$\hat{h}_1 - \hat{h}_2 = o_p\big(n^{-3/10}\big)\,. \tag{2.17}$$

(Assuming (2.16), h_0 is asymptotic to a constant multiple of $n^{-1/5}$, and so (2.17) implies that (2.15) holds for $\hat{h}_1$ if it applies to $\hat{h}_2$.) Therefore, $\hat{h}_1$ and $\hat{h}_2$ are equivalent to both first and second orders; they satisfy the same law of large numbers, i.e. $\hat{h}_j/h_0 \to 1$ in probability, and the same central limit theorem, i.e. $n^{1/10}(\hat{h}_j - h_0)/h_0 \to \mathrm{N}(0, \lambda)$ in distribution.

As noted above, Hart's (1994) approach produces implicitly-defined estimators $\tilde{\phi}_1, \ldots, \tilde{\phi}_p$ of $\phi_1, \ldots, \phi_p$. We shall prove in section 4.2 that these are equivalent to the estimators $\hat{\phi}_1, \ldots, \hat{\phi}_p$ suggested in section 2.2, in that (2.11) continues to hold if the left-hand side is replaced by $(\tilde{\phi}_1, \ldots, \tilde{\phi}_p)^{\mathrm{T}}$.

3. NUMERICAL PROPERTIES

We shall summarise the results of a simulation study comparing the finite-sample performance of the full TSCV procedure of Hart (1994) with our simplified approach, and assess relative performances of estimators of the zero-spectrum $\tau^2 = \sigma^2 \, (1 - \sum_{1 \le j \le p} \phi_j)^{-2}$. We followed the precedent set by Hart (1994) and generated data from the model $Y_i = g(x_i) + \epsilon_i$ for $i = 1, \ldots, n$, where $x_i = (i - \frac{1}{2})/n$,

$$g(x) = 10 + \tfrac{25}{2} \, x^3 \left(1 - \tfrac{1}{2} \, x\right)^3 , \tag{3.1}$$

$\epsilon_i = \phi \, \epsilon_{i-1} + Z_i$, and $\{Z_i : -\infty < i < \infty\}$ was an independent and identically distributed sequence of normal random variables with zero mean and variance $(\frac{1}{4})^2$. We used the biweight kernel $K(x) = (15/16) \, (1 - x^2)^2$, for $|x| \le 1$, and took $K = L$.

To compare the full TSCV procedure of Hart (1994) with our simplified procedure, we used as error criterion the average squared error (ASE):

$$\mathrm{ASE}(\hat{h}_{j,\mathrm{opt}}) = n^{-1} \sum_{i=1}^{n} \{\hat{g}(x_i) - g(x_i)\}^2 ,$$

for $j = 1, 2$, where $\hat{g}$ was as at (2.12), $\hat{h}_{j,\mathrm{opt}} = 2^{-1/5} \, \hat{h}_j$, $h = \hat{h}_1$ minimised $\mathrm{CV}_1(h)$, and $h = \hat{h}_2$ (along with ϕ) minimised $\mathrm{CV}_2(\phi, h)$. Both $\hat{h}_1$ and $\hat{h}_2$ were obtained from a grid of 45 bandwidths in the interval $[0.06, 0.94]$. Figure 1 shows, for a range of values of n and ϕ, values of the ratio of averages of $\mathrm{ASE}(\hat{h}_{j,\mathrm{opt}})$ over 200 samples. A ratio larger than 1 indicates higher efficiency of the new procedure. In the left-hand panel of Figure 1, m_1 is the value that, for fixed n and ϕ, gives greatest efficiency, whereas in the right-hand panel, m_1 equals $n^{0.1}$ for all settings. For the sake of brevity and simplicity we give here only results for $m_2 = n^{1/2}$; results that are more favourable to us are generally obtained for smaller values of m_2.

The figure shows that our approach is quite insensitive to choice of m_1, and that there is little difference between Hart's procedure and our simplified approach. The relative efficiency is everywhere between 0.95 and 1.05. In addition, the choice $m_1 = n^{0.1}$ is close to optimal for all values of n and ϕ. Simulations carried out for regression functions g different from the function at (3.1) gave similar conclusions.

Next we addressed the estimator $\hat{\tau}^2 = \hat{\sigma}^2 (1 - \hat{\phi})^{-2}$, introduced in section 2.3, where $\hat{\sigma}^2 = \widehat{\gamma}(0)/\bar{\gamma}(0)$ and $\bar{\gamma}(0) = (1 - \hat{\phi}^2)^{-1}$. We compared the mean squared error (MSE) of $\hat{\tau}^2$, obtained from 200 samples, with the MSE of $\tilde{\tau}^2 = \tilde{\sigma}^2 (1 - \tilde{\phi})^{-2}$, calculated using Hart's (1994) procedure. Figure 2 shows graphs of $\mathrm{MSE}(\tilde{\tau}^2)/\mathrm{MSE}(\hat{\tau}^2)$. As before, we took $m_2 = n^{1/2}$ in all cases, and took m_1 to be the optimal m_1 in the left-hand panel and to equal $n^{0.1}$ in the right-hand panel. For most values of ϕ the new procedure is more efficient than Hart's procedure, with efficiency losses arising only for small ϕ. Once again the choice $m_1 = n^{0.1}$ seems appropriate for most values of ϕ. As sample size increases the curves become flatter, indicating that the influence of ϕ on efficiency diminishes. Simulations for other regression functions gave similar results.

Finally we mention that simulations carried out for AR(2) models give results similar to those in the AR(1) case. Once again, taking $m_1 = n^{0.1}$ and $m_2 = n^{0.5}$ generally produces estimators that either outperform, or perform similarly to, those of Hart (1994). The AR(2) model selected by Hart (1994) for numerical illustration is a case in point. There, taking $n = 100$, $\phi_1 = 1.0$ and varying ϕ_2 between -0.6 and -0.1, and considering the zero-spectrum estimation problem, our method had greater efficiency than Hart's for $\phi_2 > -0.45$ and less efficiency on the other side of -0.45 (and worst-case efficiency 0.95 at $\phi_2 = -0.6$). Treating the bandwidth choice problem, our method had greater efficiency than Hart's for $\phi_2 > -0.35$ and less efficiency on the other side of -0.35 (and worst-case efficiency 0.92 at $\phi_2 = -0.6$).

4. TECHNICAL ARGUMENTS

4.1. Proof of (2.11). Observe that, by (2.4) and (2.5),

$$\frac{1}{2\,(n-j)} \sum_{i=j+1}^{n} \{(D_j\, Y)_i\}^2 = \frac{1}{2\,(n-j)} \sum_{i=j+1}^{n} \{(D_j\, \epsilon)_i\}^2 + o_p\big(n^{-1/2}\big) \tag{4.1}$$

uniformly in $1 \le j \le m_2$. Furthermore,

$$\frac{1}{2\,(n-j)} \sum_{i=j+1}^{n} \{(D_j\, \epsilon)_i\}^2 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 - \frac{1}{n} \sum_{i=j+1}^{n} \epsilon_i\, \epsilon_{i-j} + o_p\big(n^{-1/2}\big) \tag{4.2}$$

uniformly in $1 \le j \le m_2$; and, provided $m_2 - m_1 \to \infty$,

$$\frac{1}{n\,(m_2 - m_1 + 1)} \sum_{j=m_1}^{m_2} \sum_{i=j+1}^{n} \epsilon_i\,\epsilon_{i-j} = o_p\big(n^{-1/2}\big)\,. \tag{4.3}$$

Together, (4.1)–(4.3) imply that if we define $\Delta_j = n^{-1}\,\sum_{1\le i\le n}\{\epsilon_i\,\epsilon_{i-j} - \gamma(j)\}$ then $\widehat{\gamma}(j) = \gamma(j) + \Delta_j + o_p(n^{-1/2})$ for $0 \le j \le n$. Hence, if Δ_A denotes the $p \times p$ matrix with $\Delta_{j_1-j_2}$ at position (j_1, j_2), if $\Delta = (\Delta_1, \ldots, \Delta_p)^{\mathrm{T}}$ and if $\gamma = (\gamma(1), \ldots, \gamma(p))^{\mathrm{T}}$, then $\widehat{A} = A + \Delta_A + o_p(n^{-1/2})$, whence

$$(\hat{\phi}_1, \ldots, \hat{\phi}_p)^{\mathrm{T}} = \widehat{A}^{-1}(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p)^{\mathrm{T}} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}} + A^{-1}\big(\Delta - \Delta_A A^{-1}\gamma\big) + o_p\big(n^{-1/2}\big)\,. \tag{4.4}$$

Now, $\Delta_A A^{-1}\gamma = \Delta_A(\phi_1, \ldots, \phi_p)^{\mathrm{T}}$, of which the jth component is

$$\begin{aligned}\sum_{k=1}^{p} \Delta_{j-k}\,\phi_k &= n^{-1} \sum_{i=1}^{n} \epsilon_i \sum_{k=1}^{p} \phi_k\,\epsilon_{i+j-k} - \gamma(j) + o_p\big(n^{-1/2}\big) \\ &= n^{-1} \sum_{i=1}^{n} \epsilon_i\,(\epsilon_{i+j} - Z_{i+j}) - \gamma(j) + o_p\big(n^{-1/2}\big)\,.\end{aligned}$$

Therefore, the jth component of $\Delta - \Delta_A A^{-1}\gamma$ is equal to $n^{-1}\,\sum_{1\le i\le n}\,\epsilon_i\,Z_{i+j} + o_p(n^{-1/2})$. Result (2.11) follows from this property and (4.4).

4.2. Proof of (2.17), and of version of (2.11) with $(\tilde{\phi}_1, \ldots, \tilde{\phi}_p)$ on left-hand side. Put $N = n - \ell + 1$, $g_i = g(X_i)$, $U_i = \tilde{g}_i - g_i - \sum_j\,\phi_j\,(\tilde{g}_{i-j} - g_{i-j})$,

$$V_i = \sum_{j=1}^{p} (\hat{\phi}_j - \phi_j)\,(\tilde{g}_{i-j} - g_{i-j})\,, \quad W_{1i} = \sum_{j=1}^{p} (\hat{\phi}_j - \phi_j)\,\epsilon_{i-j}$$

and $W_i = W_{1i} - Z_i$. In this notation, and with $\widehat{Y}_i(h)$ defined as at (2.14), $\widehat{Y}_i(h) - Y_i = U_i - V_i + W_i$, and so

$$\mathrm{CV}_1(h) = N^{-1} \sum_{i=\ell}^{n} \{Y_i - \widehat{Y}_i(h)\}^2 = S_1 + S_2 + S_3 - 2\,(T_1 - T_2 + T_3)\,,$$

where $NS_1 = \sum_i U_i^2$, $NS_2 = \sum_i V_i^2$, $NS_3 = \sum_i W_i^2$, $NT_1 = \sum_i U_i\,V_i$, $NT_2 = \sum_i U_i\,W_i$ and $NT_3 = \sum_i V_i\,W_i$. Define too $NT_4 = \sum_i U_i\,W_{1i}$, $T_5 = T_2 - T_4$, $NT_6 = \sum_i V_i\,W_{1i}$ and $T_7 = T_3 - T_6$.

Recall from (2.9) that $\max_j |\hat{\phi}_j - \phi_j| = O_p(n^{-1/2})$, and note that it may be shown that for each $\eta > 0$,

$$\sup_{h\in\mathcal{H}} \big\{(nh)^{-1/2} n^{\eta} + h^2\big\}^{-1} \max_{\ell\le i\le n} |\tilde{g}_i - g_i| = O_p(1) \tag{4.5}$$

as $n \to \infty$. (The proof of this result, and of (4.7) below, uses the "continuity argument"; see Cheng and Hall (2001), for example, for discussion of this technique.) Therefore, because $\max_{1 \le i \le n} |\epsilon_i| = O_p(n^\eta)$ for each $\eta > 0$ (since each $E(|\epsilon_i|^c) < \infty$; see (2.16)), then

$$\max_{\ell \le i \le n} |U_i| = O_p\big\{(nh)^{-1/2} n^\eta + h^2\big\}\,, \quad \max_{\ell \le i \le n} |V_i| = O_p\big(n^{\eta-1} h^{-1/2} + n^{-1/2} h^2\big)$$

and $\max_{\ell \le i \le n} |W_{1i}| = O_p(n^{\eta-(1/2)})$, uniformly in $h \in \mathcal{H}$, for each $\eta > 0$. Therefore, $S_2 = O_p(n^{\eta-2} h^{-1} + n^{-1} h^4)$, $T_1 = O_p(n^{\eta-(3/2)} h^{-1} + n^{-1/2} h^4)$ and $T_6 = O_p[\{(nh)^{-1/2} + h^2\}\, n^{\eta-1}]$ uniformly in $h \in \mathcal{H}$. It follows that

$$\mathrm{CV}_1 = S_1 + 2\,(T_2 - T_7) + R_1\,, \tag{4.6}$$

where, here and below, R_j denotes a random function of h that can be written as $R_j = R' + R''$, with R' not depending on h and, for all $\eta > 0$, $R'' = O_p[\{(nh)^{-1} + h^4\}\, n^{\eta-(1/2)}]$ uniformly in $h \in \mathcal{H}$.

Note that

$$NT_4 = \sum_{j=1}^p (\hat\phi_j - \phi_j) \left(\Delta_{j0} - \sum_{k=1}^p \phi_k\, \Delta_{jk}\right), \quad NT_7 = -\sum_{j=1}^p (\hat\phi_j - \phi_j)\, \Delta_j\,,$$

where $\Delta_{jk} = \sum_{\ell \le i \le n} \epsilon_{i-j}\, (\tilde g_{i-k} - g_{i-k})$ and $\Delta_j = \sum_{\ell \le i \le n} Z_i\, (\tilde g_{i-j} - g_{i-j})$. It may be proved that for each $\eta > 0$,

$$\max_{h \in \mathcal{H}} \left[\big\{(nh)^{-1/2} + h^2\big\}\, n^{(1/2)+\eta}\right]^{-1} \left(\max_{1 \le j \le p\,,\, 0 \le k \le p} |\Delta_{jk}| + \max_{1 \le j \le p} |\Delta_j|\right) = O_p(1)\,. \tag{4.7}$$

Therefore, $|T_4| + |T_7| = O_p[\{(nh)^{-1/2} + h^2\}\, n^{\eta-1}]$ uniformly in $h \in \mathcal{H}$. Hence, by (4.6),

$$\mathrm{CV}_1 = S_1 + 2\,T_5 + R_2 = N^{-1} \sum_{i=\ell}^n (U_i - Z_i)^2 + R_3\,. \tag{4.8}$$

Let ϕ_j^0 denote the true value of ϕ_j (denoted above by simply ϕ_j), write ϕ_j for a generic value, and continue to take U_j to be the version of that quantity when ϕ_j assumes its true value. The equation $(\partial/\partial\phi_j)\, \mathrm{CV}_2(\phi_1, \ldots, \phi_p, h) = 0$ reduces to

$$\sum_{k=1}^p \big(\tilde\phi_k - \phi_k^0\big) \sum_{i=\ell}^n \Delta_{ikk}\, \Delta_{ijj} = \sum_{i=\ell}^n (U_i - Z_i)\, \Delta_{ijj}\,, \tag{4.9}$$

where $\Delta_{ijk} = \tilde{g}_{i-j} - g_{i-j} - \epsilon_{i-k}$. At the extremum, $h = \hat{h}_2 = \{1 + o_p(1)\}\, cn^{-1/5}$, where $c > 0$. It therefore follows from (4.5) and (4.7) that

$$N^{-1} \sum_{i=\ell}^{n} \Delta_{ikk}\, \Delta_{ijj} = n^{-1} \sum_{i=1}^{n} \epsilon_{i-k}\, \epsilon_{i-j} + o_p\big(n^{-1/2}\big)\,,$$
$$N^{-1} \sum_{i=\ell}^{n} (U_i - Z_i)\, \Delta_{ijj} = n^{-1} \sum_{i=1}^{n} Z_i\, \epsilon_{i-j} + o_p\big(n^{-1/2}\big)\,.$$

Result (2.11), with its left-hand side replaced by $(\tilde{\phi}_1, \ldots, \tilde{\phi}_p)^{\mathrm{T}}$, now follows via (4.9). Moreover, it holds for values of h that lie inside a small interval of which $\hat{h}_2$ is an interior point.

In particular we have proved that for each j, $\tilde{\phi}_j - \phi_j$ may be expressed as a quantity which does not depend on h and equals $O_p(n^{-1/2})$, plus a quantity which does depend on h but equals $o_p(n^{-1/2})$. We may now rework the argument leading to (4.8), with $\tilde{\phi}_j$ replacing $\hat{\phi}_j$ for $1 \le j \le p$, to show that

$$\mathrm{CV}_2(\tilde{\phi}_1, \ldots, \tilde{\phi}_p, h) = N^{-1} \sum_{i=\ell}^{n} (U_i - Z_i)^2 + R_1'\,, \tag{4.10}$$

where, here and below, R_j' denotes the sum of a term that does not depend on h and a term that equals $o_p(n^{-1})$ uniformly in $h \in [h_0\,(1-\eta), h_0\,(1+\eta)]$, for any $\eta \in (0,1)$.

Arguing as in Chu and Marron (1991) it may be proved that (a) the minimiser, $\hat{h}$, of $N^{-1} \sum_{\ell \le i \le n} (U_i - Z_i)^2$ over $h \in \mathcal{H}$, satisfies $\hat{h}/h_0 \to 1$ in probability, and (b)

$$N^{-1} \sum_{i=\ell}^{n} (U_i - Z_i)^2 = \xi_1\, h_0^2 \left\{h - h_0 \left(1 + n^{-1/10}\, \xi_2\right)\right\}^2 + R_2'\,,$$

where ξ_1 and ξ_2 denote random variables not depending on h, ξ_1 converges to a finite, strictly positive constant, and ξ_2 has an asymptotic $\mathrm{N}(0, \lambda)$ distribution with $0 < \lambda < \infty$. Combining this property with (4.8) and (4.10) we deduce that the minimisers, $\hat{h}_1$ and $\hat{h}_2$, of $\mathrm{CV}_1(h)$ and $\mathrm{CV}_2(\tilde{\phi}_1, \ldots, \tilde{\phi}_p, h)$ respectively, both satisfy $\hat{h}_j = h_0\,(1 + n^{-1/10}\xi_2) + o_p(n^{-3/10})$, for the same random variable ξ_2 in each case. This implies (2.17).

REFERENCES

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

ANH, V.V., WOLFF, R.C., GAO, J.T. AND TIENG, Q. (1999). Local linear regression with long-range dependent errors. *Austral. and New Zeal. J. Statist.* **41**, 463–479.

CAI, Z. AND ELIAS, O.-S. (2001). Local robust regression estimation for time series. Manuscript.

CAI, Z. AND MASRY, E. (2000). Nonparametric estimation of additive nonlinear ARX time series: local linear fitting and projections. *Econometric Theory* **16**, 465–501.

CHENG, M.-Y. AND HALL, P. (2001). Methods for tracking support boundaries with corners. *Ann. Statist.*, under revision.

CHU, C.-K. AND MARRON, J.S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.* **19**, 1906–1918.

FRANCISCO-FERNANDEZ, M. AND VILAR-FERNANDEZ, J.M. (2001). Local polynomial regression estimation with correlated errors. *Commun. Statist. Theor. Meth.* **30**, 1271–1293.

GASSER, T., SROKA, L. AND JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

HÄRDLE, W., LÜTKEPOHL, H. AND CHEN, R. (1997). A review of nonparametric time series analysis. *Internat. Statist. Rev.* **65**, 49–72.

HALL, P., KAY, J.W. AND TITTERINGTON, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

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

HART, J.D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *J. Roy. Statist. Soc.* Ser. B **56**, 529–542.

HART, J.D. AND YI, S. (1998). One-sided cross-validation. *J. Amer. Statist. Assoc.* **93**, 620–631.

HERRMANN, E., GASSER, T. AND KNEIP, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika* **79**, 783–795.

HYNDMAN, R.J. AND WAND, M.P. (1997). Nonparametric autocovariance function estimation. *Aust. J. Statist.* **39**, 313–324.

MASRY, E. AND MIELNICZUK, J. (1999). Local linear regression estimation for time series with long-range dependence. *Stoch. Proc. Appl.* **82**, 173–193.

MÜLLER, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Springer Lecture Series in Statistics **46**. Springer, New York.

MÜLLER, H.-G. AND STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–635.

MÜLLER, H.-G. AND STADTMÜLLER, U. (1989). Detecting dependencies in smooth regression models. *Biometrika* **75**, 639–650.

OPSOMER, J.D. (1996). Estimating an unknown function by local linear regression when the errors are correlated. *Proc. Sect. Statist. Comput.* Amer. Statist.

Assoc., Alexandria, VA, 102–108.

OPSOMER, J., WANG, Y.D AND YANG, Y.H. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16**, 134–153.

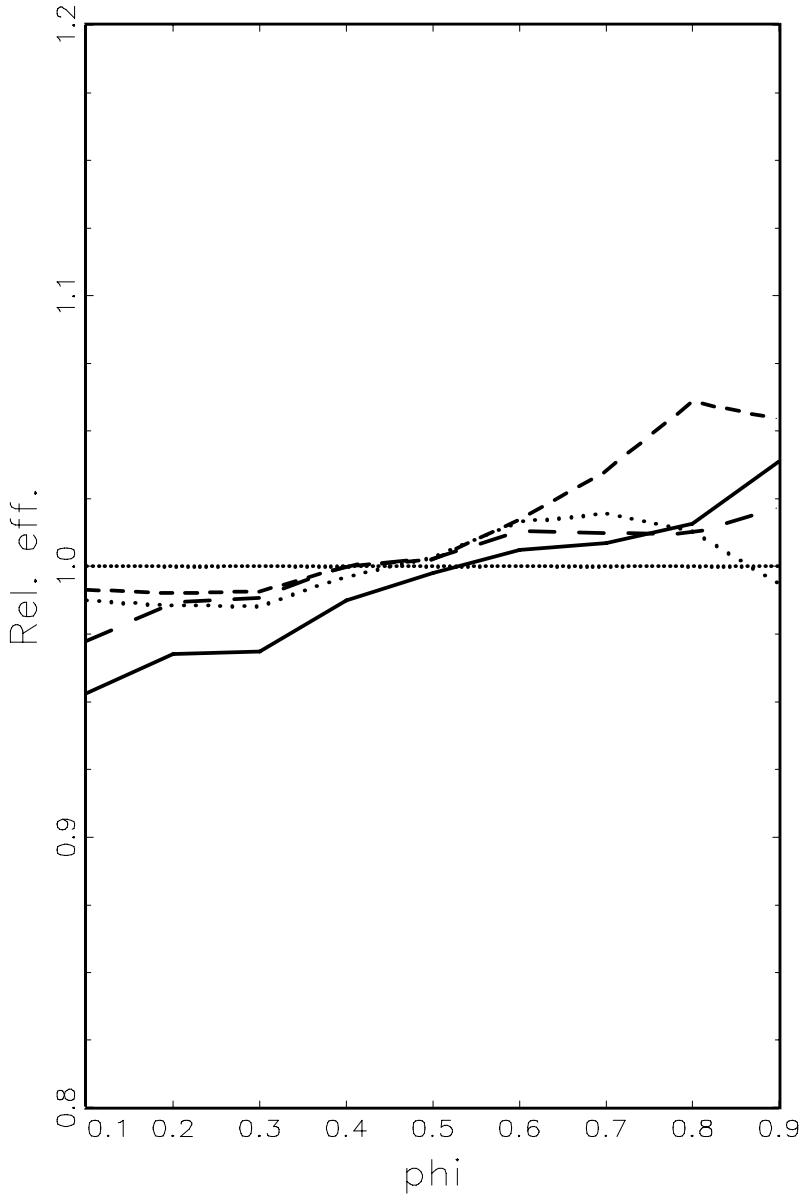
RAY, B.K. AND TSAY, R.S. (1997). Bandwidth selection for kernel regression with long-range dependent errors. *Biometrika* **84**, 791–802.

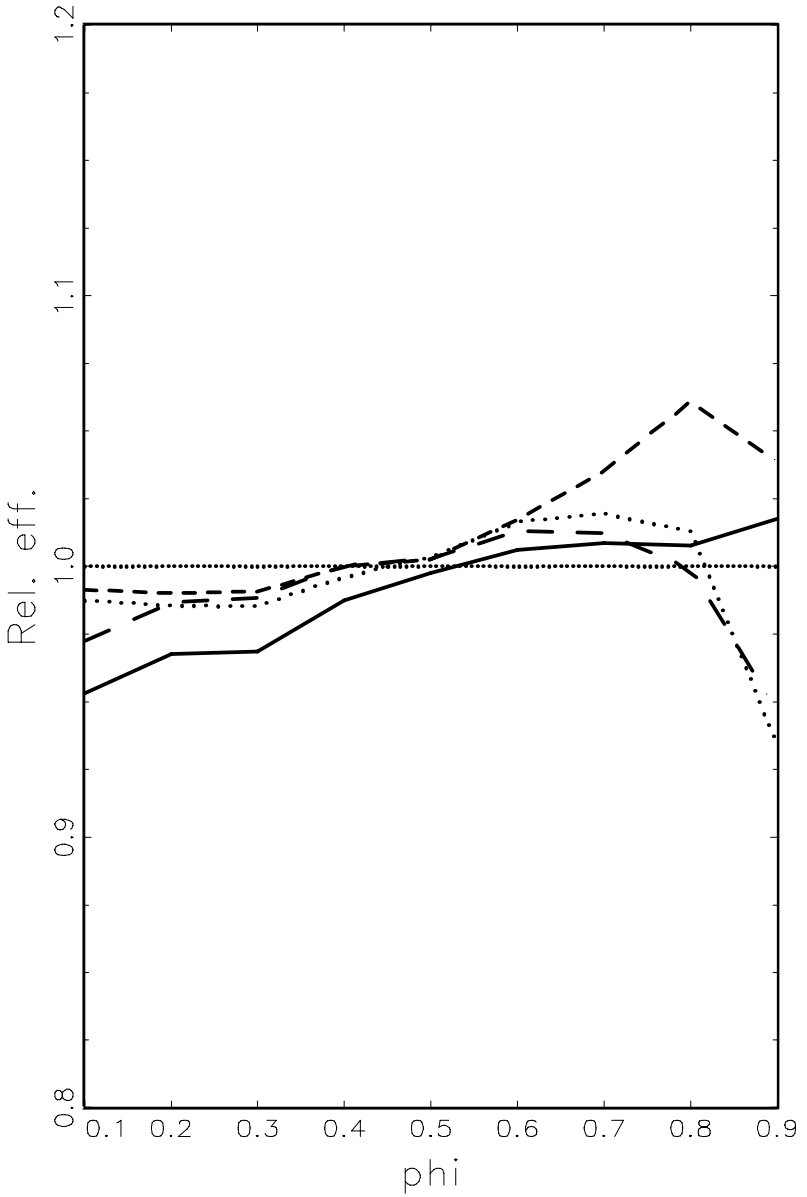
RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215–1230.

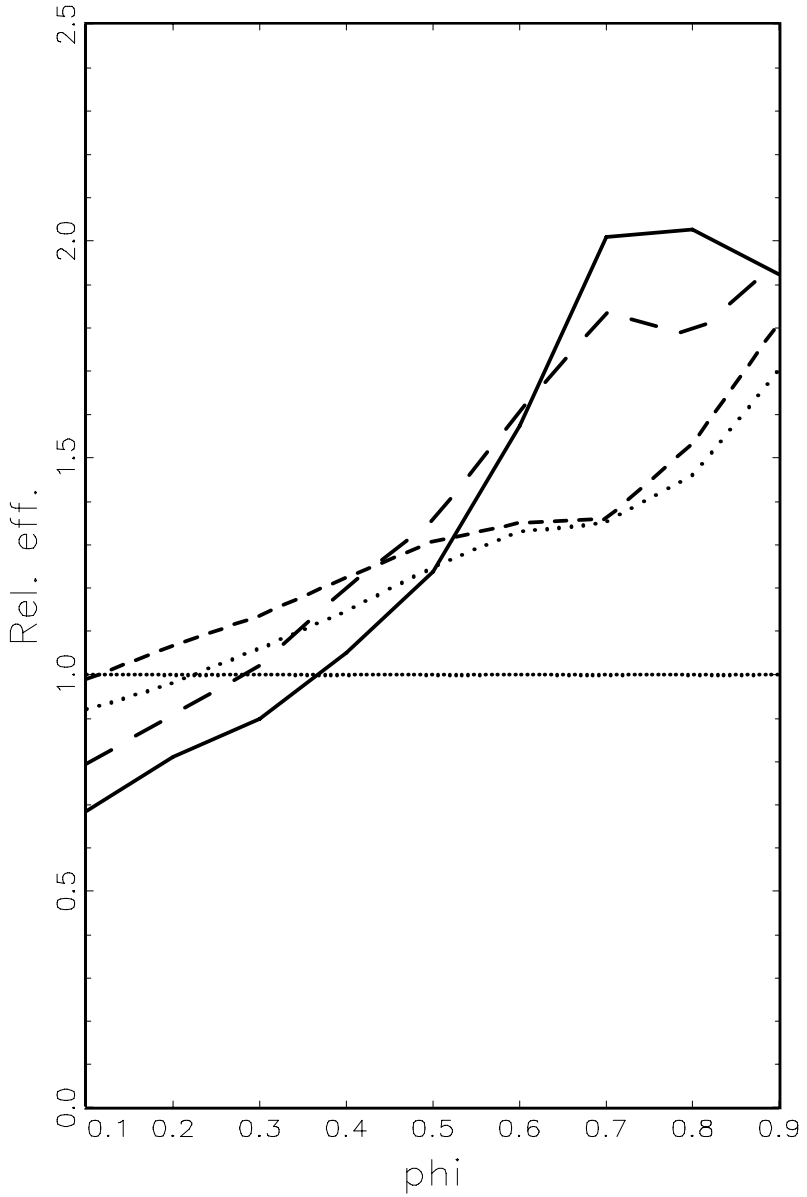
YAO, Q.W. AND TONG, H.W. (1998). Cross-validatory bandwidth selections for regression estimation based on dependent data. *J. Statist. Plann. Inf.* **68**, 387–415.

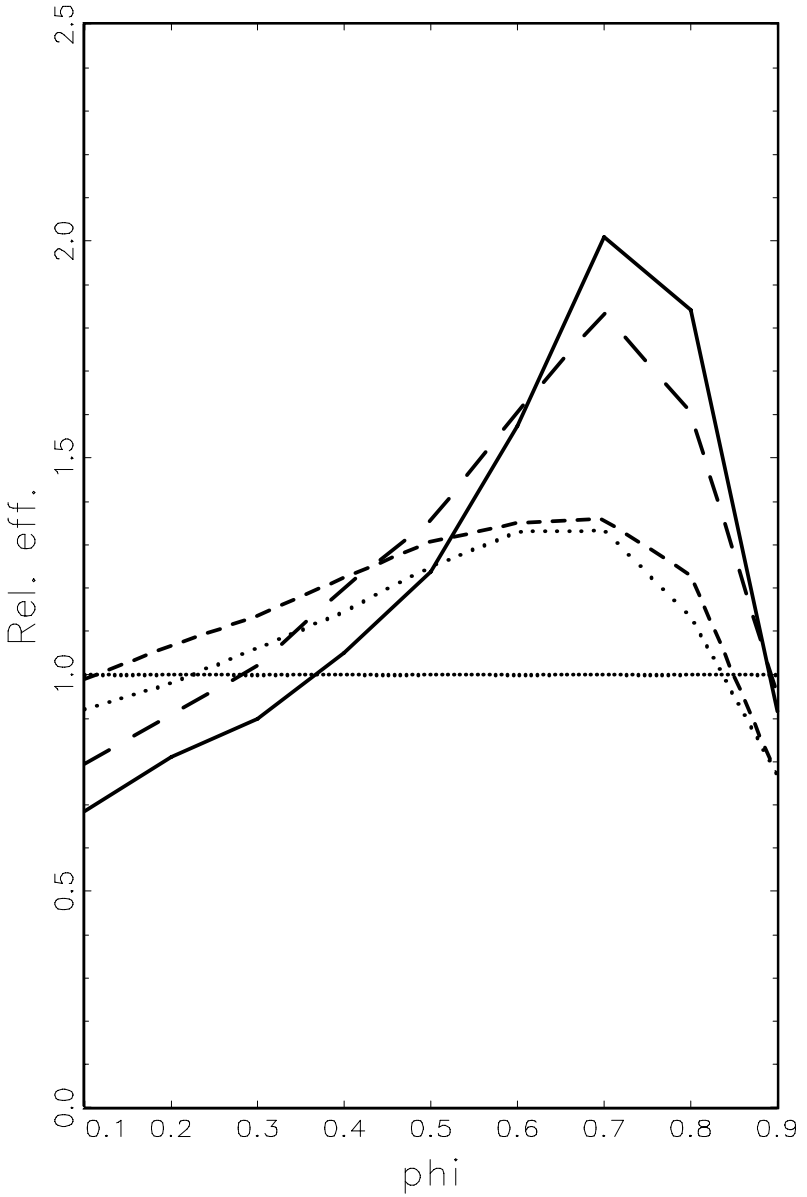
Caption for Figure 1: *Efficiency of* $\hat{h}_{1,\mathrm{opt}}$ *relative to* $\hat{h}_{2,\mathrm{opt}}$. Efficiencies are in terms of ratios of mean squared errors, with $\hat{h}_{1,\mathrm{opt}}$ having lower mean squared error if efficiency exceeds 1. The unbroken line, long-dashed line, dotted line and short-dashed line represent the efficiency curves for $n = 100$, 200, 400 and 600, respectively.

Caption for Figure 2: *Efficiency of* $\hat{\tau}^2$ *relative to* $\tilde{\tau}^2$. Efficiencies are in terms of mean squared error ratios, with $\hat{\tau}^2$ having lower mean squared error if efficiency exceeds 1. Line types for curves are as in Figure 1.


Rel. eff.
1.2
1.1
1.0
0.9
0.8
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
phi


Rel. eff.
1.2
1.1
1.0
0.9
0.8
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
phi


Rel. eff.
0.0
0.5
1.0
1.5
2.0
2.5
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
phi


Rel. eff.
0.0
0.5
1.0
1.5
2.0
2.5
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
phi