WILEY

Using Difference-Based Methods for Inference in Nonparametric Regression with Time Series Errors

Author(s): Peter Hall and Ingrid Van Keilegom

Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 65, No. 2 (2003), pp. 443-456
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/3647514
Accessed: 31-10-2017 11:55 UTC

*J. R. Statist. Soc.* B (2003)
**65**, *Part* 2, *pp.* 443–456

# Using difference-based methods for inference in nonparametric regression with time series errors

Peter Hall

*Australian National University, Canberra, Australia*

and Ingrid Van Keilegom

*Australian National University, Canberra, Australia, and Université catholique de Louvain, Louvain-la-Neuve, Belgium*

[Received February 2002. Revised September 2002]

**Summary.** 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. This new approach produces a bandwidth selector that is equivalent, to both first and second orders, to that given by the full time series cross-validation algorithm. Other applications of difference-based methods are to variance estimation and construction of confidence bands in nonparametric regression.

*Keywords*: Autoregression; Bandwidth; Covariance; Cross-validation; Kernel methods; Linear time series; Local linear regression; Time series cross-validation

## 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 that 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 estimator variance is

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

where $\gamma(j)$ denotes the error covariance at lag $j$, $n$ is sample size, $h$ is the bandwidth, $f(x)$ is the density of design at the point $x$ of estimation and $\kappa$ is a constant depending only on the types

*Address for correspondence*: Ingrid Van Keilegom, Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.
E-mail: vankeilegom@stat.ucl.ac.be

1369–7412/03/65443

of estimator and kernel. See, for example, Opsomer *et al.* (2001). We shall show how to use our difference-based approach to estimate the zero spectrum

$$\tau^2 \equiv \gamma(0) + 2 \sum_{j \geqslant 1} \gamma(j),$$

at expression (1.1). The covariances (and autoregressive parameters) are also needed 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 the choice of bandwidth. 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 $h$. In contrast, the autoregressive parameters need to be calculated for every candidate for the bandwidth that minimizes 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) technique, and in particular they 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 *et al.* (1986), Müller and Stadtmüller (1987), Müller (1988) and Hall *et al.* (1990). Müller and Stadtmüller (1988) suggested difference-based methods for estimating correlation in the setting of regression models with $m$-dependent errors. Techniques based on differencing have an extensive history in time series; see for example Anderson (1971), page 66.

In the context of estimating variance under conditions of heteroscedasticity, Fan and Yao (1998) argued that some *ad hoc* difference-based methods can be inefficient. Although this is certainly true, it is also the case that difference-based methods that have better performance in the case of large samples can have an unduly large mean-squared error for small samples. See, for example, Seifert *et al.* (1993). Considerations such as these, and the attractive simplicity of difference-based methods, can argue in favour of them despite theoretical losses of efficiency. The techniques suggested in the present paper are efficient in the case of Gaussian errors.

Härdle *et al.* (1997) have reviewed nonparametric approaches to time series analysis. Methodology for nonparametric regression with correlated errors, including different approaches to the choice of bandwidth, has been surveyed and discussed by Opsomer *et al.* (2001). This paper 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 *et al.* (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 choosing bandwidths in the case of long-range dependent errors; see Hermann *et al.* (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 the estimation of the covariance function. The work in the present paper will address particularly the instance of nonparametric inference by using local linear methods, and there the time series setting has recently been treated by Opsomer (1996), Anh *et al.* (1999), Masry and Mielniczuk (1999), Cai and Masry (2000), Cai and Ould-Said (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 $1 \leqslant i \leqslant n$, are generated by the model $Y_i = g(X_i)+\varepsilon_i$. Here $g$ is a smooth function, and we suppose that

the $X_i$s form an increasing sequence on the interval $\mathcal{I} = [0, 1]$ and are
either equally spaced there or are ordered values of independent and identically
distributed random variables whose density has support equal to $\mathcal{I}$. (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 $\varepsilon_1, \ldots, \varepsilon_n$ form a segment of a doubly infinite stationary linear time series with zero mean. Specifically, we ask that, for some $p \geqslant 1$,

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

where $\{Z_i : -\infty < i < \infty\}$ are independent and identically distributed random variables with finite variance $\sigma^2$ 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 $|z| \leqslant 1$. Since the latter series has only a finite number of non-zero terms then an equivalent assumption is that,

for some $\eta > 0$, the equation $1 - \Sigma_j \phi_j z^j = 0$ has no complex roots
satisfying $|z| \leqslant 1 + \eta$. (2.3)

This condition implies that the process $\{\varepsilon_i\}$ 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 between three or more $Y_i$s may also be used, but the additional complication does not produce asymptotic improvements in performance.

Given an integer $j \geqslant 1$, 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 that $g$ has a bounded derivative,

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

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

$$(D_jY)_i = \{D_j\, g(X)\}_i + (D_j\varepsilon)_i, \tag{2.5}$$

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

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

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

where $j \geqslant 1$ in the second definition and $m_1 \leqslant m_2$ are subsidiary smoothing parameters. Put $\hat{\gamma}(-j) = \hat{\gamma}(j)$.

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

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

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

$$\max_{0 \leqslant j \leqslant n} \left|\hat{\gamma}(j) - \gamma(j)\right| = O_p(n^{-1/2}). \tag{2.8}$$

Given $\hat{\gamma}(j)$, for $1 \leqslant j \leqslant p$, we construct estimators of $\phi_j$ by using the Yule–Walker equations. Specifically, since

$$\gamma(k) = \sum_{j \geqslant 1} \phi_j\, \gamma(k-j) \qquad \text{for } k \geqslant 1,$$

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 $\gamma$ by $\hat{\gamma}$ in the definition of $A$, to obtain the estimator $\hat{A}$, and putting $\hat{\gamma}_j = \hat{\gamma}(j)$, we define

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

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

$$\max_{1 \leqslant j \leqslant p} \left|\hat{\phi}_j - \phi_j\right| = O_p(n^{-1/2}). \tag{2.9}$$

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

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

then

$$(\hat{\phi}_1, \ldots, \hat{\phi}_p)^{\mathrm{T}} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}} + n^{-1} \sum_{i=1}^{n} \varepsilon_i A^{-1}(Z_{i+1}, \ldots, Z_{i+p})^{\mathrm{T}} + o_p(n^{-1/2}). \tag{2.11}$$

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 equation (2.11) (or, in the case of Hart's estimators, the respective version of equation (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 $\{\varepsilon_i\}$ 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 expression (2.6) may be employed for arbitrarily large values of $j$. However, these particular estimators do not exploit the autoregressive structure of the process $\varepsilon_i$. This can result in the covariance sequence $\{\hat{\gamma}(j_1 - j_2) : -\infty < j_1, j_2 < \infty\}$ not being positive definite, and in $\hat{\gamma}(j)$ being unnecessarily highly variable for large $j$. An alternative approach, which is available once the estimators $\hat{\phi}_j$ have been constructed, is to estimate $\gamma(j)$ by $\tilde{\gamma}(j)$, constructed as follows. Define $\hat{\psi}_1, \hat{\psi}_2, \ldots$ by

$$1 + \sum_{j \geqslant 1} \hat{\psi}_j z^j = \left(1 - \sum_{1 \leqslant j \leqslant p} \hat{\phi}_j z^j\right)^{-1}.$$

and let $\hat{\psi}_0 = 1$. Let $\bar{\gamma}(j)$ denote the coefficient of $z^j$, or equivalently of $z^{-j}$, in an expansion of

$$\left(1 - \sum_{j=1}^{p} \hat{\phi}_j z^j\right)^{-1} \left(1 - \sum_{j=1}^{p} \hat{\phi}_j z^{-j}\right)^{-1},$$

i.e.

$$\bar{\gamma}(j) = \sum_{k,l \geqslant 0} \sum \hat{\psi}_l \hat{\psi}_k.$$

Put $\hat{\sigma}^2 = \hat{\gamma}(0)/\bar{\gamma}(0)$, an estimator of $\sigma^2 = \text{var}(Z_i)$. Then define $\tilde{\gamma}(j) = \hat{\sigma}^2\, \bar{\gamma}(j)$ for $j \geqslant 1$. Hence, an estimator of the zero spectrum

$$\tau^2 \equiv \gamma(0) + 2 \sum_{j \geqslant 1} \gamma(j),$$

appearing at expression (1.1), is given by

$$\hat{\tau}^2 = \hat{\sigma}^2 \left(1 - \sum_{1 \leqslant j \leqslant p} \hat{\phi}_j\right)^{-2}.$$

The definition of $\hat{\gamma}(0)$ implies that that quantity is positive. Moreover, since $\bar{\gamma}(0)$ equals the error variance in any autoregression (defined conditionally on the data) that has coefficients $\hat{\phi}_j$ (for $1 \leqslant j \leqslant p$) 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 $\{\tilde{\gamma}(j_1 - j_2) : -\infty < j_1, j_2 < \infty\}$ 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

$\hat{\phi}_j$, conditionally on the data. It may be proved that, provided that conditions (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 $\tilde{\gamma}(j)$, observe that we may write

$$\varepsilon_i = \sum_{k\geqslant 0} \psi_k Z_{i-k}$$

where $\psi_0 = 1$ and $\psi_1, \psi_2, \ldots$ are defined by

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

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

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

which in turn equals $\sigma^2$ multiplied by the coefficient of $z^j$ in an expansion of

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

This argument of course fails if the equation $\Sigma_j \phi_j z^j = 0$ has a root inside or on the unit circle. Moreover, for small samples the estimators $\tilde{\gamma}(j)$ may be unreliable if the equation has a root outside but close to the unit circle.

### 2.4. Curve estimators

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

$$\hat{g}(x) = \sum_{1\leqslant k\leqslant n} w_k(x) Y_k \Big/ \sum_{1\leqslant k\leqslant n} w_k(x), \tag{2.12}$$

where

$$w_k(x) = K\Big(\frac{x - X_k}{h}\Big)\{s_2(x) - (x - X_k)\, s_1(x)\},$$

$$s_l(x) = (nh)^{-1} \sum_{k=1}^{n} K\Big(\frac{x - X_k}{h}\Big)(x - X_k)^l,$$

$K$ is a kernel function and $h$ is a bandwidth.

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

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

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

$$v_k(x) = L\Big(\frac{x - X_k}{h_1}\Big)\{r_2(x) - (x - X_k)\, r_1(x)\},$$

$$r_l(x) = (nh_1)^{-1} \sum_k^{(x)} L\Big(\frac{x - X_k}{h_1}\Big)(x - X_k)^l,$$

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

In contrast with 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 $\Sigma_k^{(x)} v_k(x)(x - X_k) = 0$.

### *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 $\tilde{g}_i = \tilde{g}(X_i)$,

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

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

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

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

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

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

where

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

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

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

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 $h_{\mathrm{opt}}$ that minimizes MISE for $\hat{g}$, defined at equation (2.12):

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

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 $\hat{g}$.

An advantage that Hart's (1994) approach has over our own is that it does not require the quantities $m_1$ and $m_2$ to be selected by the user. However, as we shall show in Section 3, our method is rather insensitive to the choice of these quantities. Nevertheless, if $\gamma(0)^{-1} \int (g'')^2$ were large then the choice of $m_1$ and $m_2$ would be more of an issue.

### 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 minimize $\mathrm{CV}_2(\phi_1,\ldots,\phi_p,h)$ 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

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

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

$E(|\varepsilon_i|^c)<\infty$ 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. (2.16)

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

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

(Assuming expression (2.16), $h_0$ is asymptotic to a constant multiple of $n^{-1/5}$, and so equation (2.17) implies that expression (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. $n^{1/10}(\hat{h}_j-h_0)/h_0 \to N(0,\lambda)$ 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 Appendix B that these are equivalent to the estimators $\hat{\phi}_1,\ldots,\hat{\phi}_p$ suggested in Section 2.2, in that equation (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 summarize 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 $\tau^2=\sigma^2(1-\Sigma_{1\leqslant j\leqslant p}\,\phi_j)^{-2}$. We followed the precedent set by Hart (1994) and generated data from the model $Y_i=g(x_i)+\varepsilon_i$ for $i=1,\ldots,n$, where $x_i=(i-\frac{1}{2})/n$,

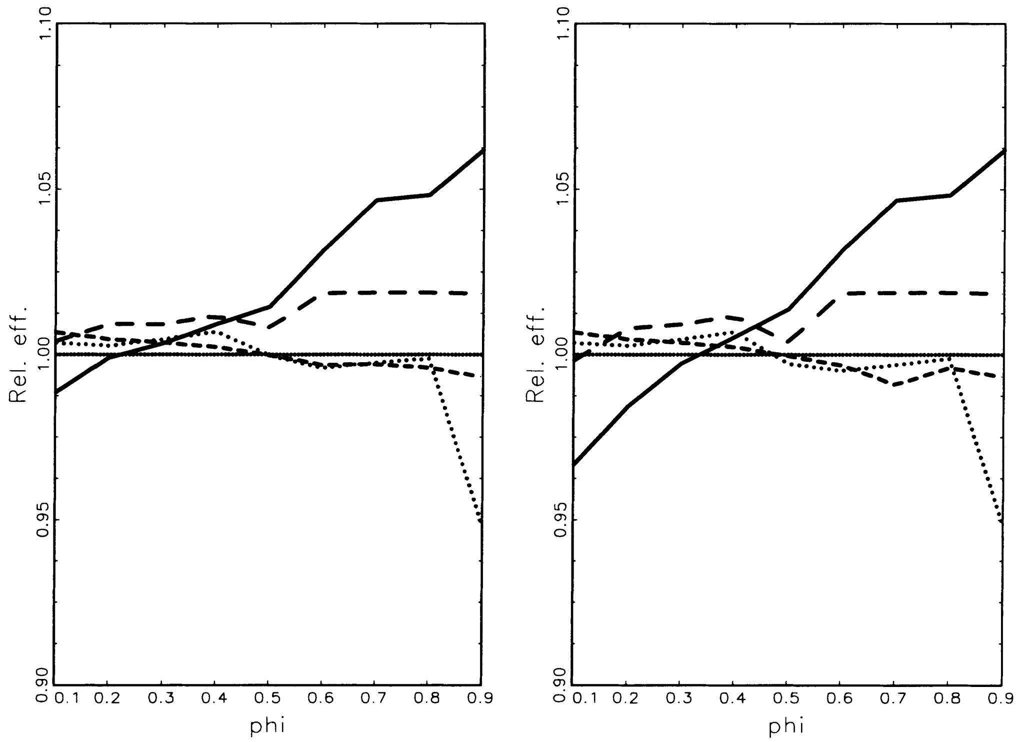
$$g(x)=10+\frac{25}{2}x^3\left(1-\frac{1}{2}x\right)^3, \tag{3.1}$$

and $\varepsilon_i=\phi\varepsilon_{i-1}+Z_i$, 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 $|x|\leqslant 1$, and took $K=L$. In this case, $R(K,L)=0.56$.

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 equation (2.12), $\hat{h}_{j,\mathrm{opt}}=R(K,L)\hat{h}_j$, $h=\hat{h}_1$ minimized $\mathrm{CV}_1(h)$ and $h=\hat{h}_2$ (along with $\phi$) minimized $\mathrm{CV}_2(\phi,h)$. Both $\hat{h}_1$ and $\hat{h}_2$ were obtained from a grid of 45 bandwidths in the interval $[0.06,0.94]$. Fig. 1 shows, for a range of values of $n$ and $\phi$, values



(a)

(b)

**Fig. 1.** Efficiency of $\hat{h}_{1,\mathrm{opt}}$ relative to $\hat{h}_{2,\mathrm{opt}}$ (the efficiencies are in terms of ratios of mean-squared errors, with $\hat{h}_{1,\mathrm{opt}}$ having a lower mean-squared error if the efficiency exceeds 1): ———, $n = 100$; — —, $n = 200$; ·······, $n = 400$; -------, $n = 600$

of the ratio of averages of ASE($\hat{h}_{j,\mathrm{opt}}$) over 200 samples. A ratio larger than 1 indicates higher efficiency of the new procedure. In Fig. 1(a), $m_1$ is the value that, for fixed $n$ and $\phi$, gives the greatest efficiency, whereas in Fig. 1(b) $m_1 = n^{0.4}$ for all settings. For 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 $m_2$.

Fig. 1 shows that our approach is quite insensitive to the 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 $m_1 = n^{0.4}$ is close to optimal for all values of $n$ and $\phi$. Simulations carried out for regression functions $g$ different from the function (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 $\hat{\sigma}^2 = \hat{\gamma}(0)/\bar{\gamma}(0)$ 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 by using Hart's (1994) procedure. Fig. 2 shows graphs of MSE($\tilde{\tau}^2$)/MSE($\hat{\tau}^2$). We took $m_2 = n^{1/2}$ in all cases, and we took $m_1$ to be the optimal $m_1$ in Fig. 2(a) and to equal $n^{0.1}$ in Fig. 2(b). For most values of $\phi$ the new procedure is more efficient than Hart's procedure, with efficiency losses arising only for small $\phi$. As the sample size increases the curves become flatter, indicating that the influence of $\phi$ on efficiency diminishes. Simulations for other regression functions gave similar results.

We mention that simulations carried out for autoregressive AR(2) models give results that are similar to those in the AR(1) case. Once again, taking $m_2 = n^{0.5}$, $m_1 = n^{0.4}$ for the esti-

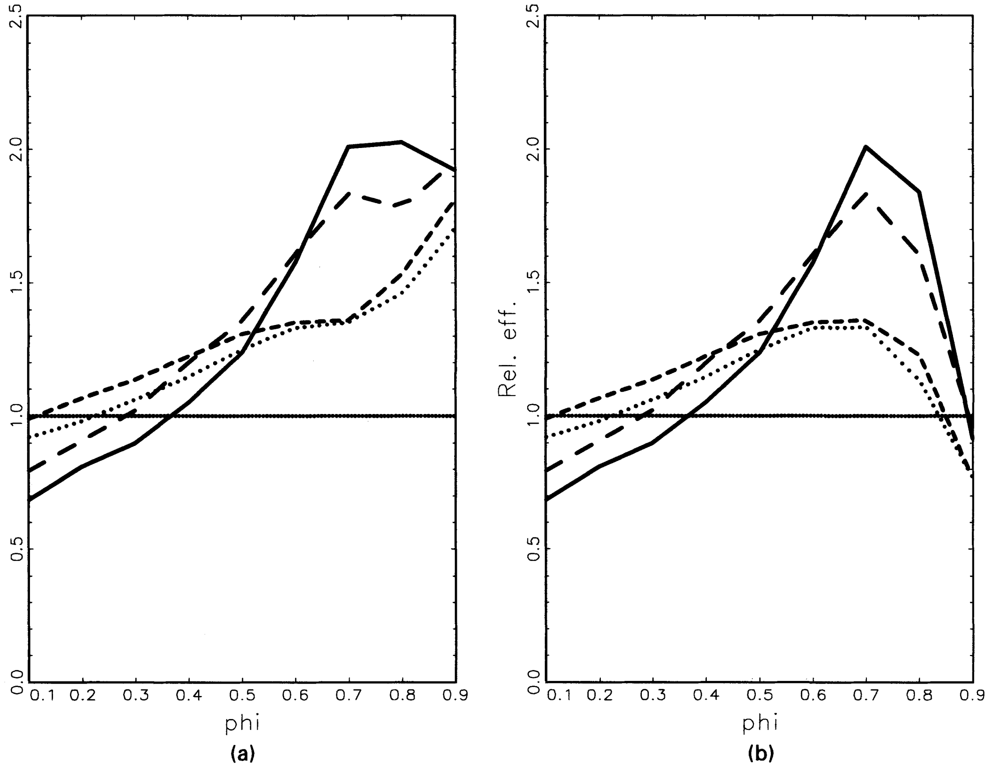


**Fig. 2.** Efficiency of $\hat{\tau}^2$ relative to $\tilde{\tau}^2$ (the efficiencies are in terms of mean-squared error ratios, with $\hat{\tau}^2$ having a lower mean-squared error if the efficiency exceeds 1): ———, $n = 100$; — —, $n = 200$; ······, $n = 400$; ------, $n = 600$

mation of the bandwidth and $m_1 = n^{0.1}$ for estimating $\tau^2$ 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 $\phi_2$ 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 $\phi_2 = -0.6$).

Finally we applied our method, and TSCV, to first differences of annual global surface air temperatures from 1880 to 1985. The differences are expressed in degrees Celsius and are shown in Fig. 3. The data come from Hansen and Lebedeff (1987). First we fitted an AR(1) model, using $m_1 = n^{0.4}$ and $m_2 = n^{0.5}$ in view of the results above. The bandwidths that minimized $\mathrm{CV}_1(h)$ and $\mathrm{CV}_2(h)$ were respectively $\hat{h}_1 = 39$ and $\hat{h}_2 = 32$. The two estimates of $g$, produced using bandwidths $R(K, L)\hat{h}_j$ $(j = 1, 2)$, are very close to one another and are shown in Fig. 3. The zero spectrum $\tau^2$ was estimated as 0.041 by using the new procedure with $m_1 = n^{0.1}$ and 0.033 by using TSCV. Estimates of the autoregressive coefficient $\phi_1$ were 0.414 with our method and 0.295 in the case of TSCV.

Results obtained on fitting an AR(2) model were very similar to these, with the bandwidths determined as 38 and 33, and $\tau^2$ estimated as 0.036 and 0.038, when using the respective methods. The curve estimates were virtually identical with their counterparts in Fig. 3. Estimates of

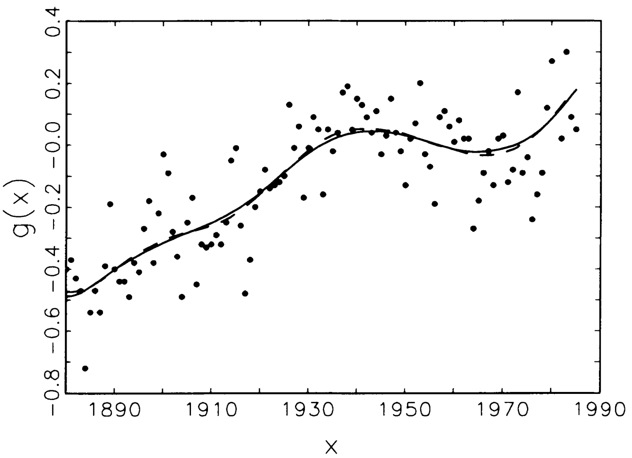


**Fig. 3.** Regression function for surface air temperature data: ———, new method; – – –, TSCV

the autoregressive coefficients $(\phi_1, \phi_2)$ were (0.442, −0.068) and (0.290, 0.040) in the respective cases.

## 4. Discussion

We have suggested methods for estimating autoregressive parameters in nonparametric regression problems, and we indicated their usefulness by applying them to choose the bandwidth. The estimators depend on secondary smoothing parameters, $m_1$ and $m_2$, and potential values of these could be indicated by a second-order theoretical analysis. This is one open problem to which our work points. Another is the application of the same ideas to estimate the variance function in heteroscedastic settings, and a third is a method for selecting the bandwidth in the latter context.

## Acknowledgements

We are grateful to three reviewers for helpful comments. The research of Van Keilegom was supported by contract ‘Projet d’actions de recherche concertées’ 98/03–217 and by the Interuniversity Attraction Pole research network P5/24 of the Belgian Government.

## Appendix A: Proof of result (2.11)

Observe that, by equations (2.4) and (2.5),

$$\frac{1}{2(n-j)} \sum_{i=j+1}^{n} \{(D_j Y)_i\}^2 = \frac{1}{2(n-j)} \sum_{i=j+1}^{n} \{(D_j \varepsilon)_i\}^2 + o_p(n^{-1/2}) \qquad \text{(A.1)}$$

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

$$\frac{1}{2(n-j)} \sum_{i=j+1}^{n} \{(D_j \varepsilon)_i\}^2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 - \frac{1}{n} \sum_{i=j+1}^{n} \varepsilon_i \varepsilon_{i-j} + o_p(n^{-1/2}) \qquad \text{(A.2)}$$

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

$$\frac{1}{n(m_2 - m_1 + 1)} \sum_{j=m_1}^{m_2} \sum_{i=j+1}^{n} \varepsilon_i \varepsilon_{i-j} = o_p(n^{-1/2}). \qquad \text{(A.3)}$$

Together, equations (A.1)–(A.3) imply that if we define $\Delta_j = n^{-1}\Sigma_{1\leqslant i\leqslant n}\,\{\varepsilon_i\varepsilon_{i-j} - \gamma(j)\}$ then $\hat\gamma(j) = \gamma(j) + \Delta_j + o_p(n^{-1/2})$ for $0 \leqslant j \leqslant n$. Hence, if $\Delta_A$ denotes the $p \times p$ matrix with $\Delta_{j_1-j_2}$ at position $(j_1, j_2)$, if $\Delta = (\Delta_1, \dots, \Delta_p)^{\mathrm{T}}$ and if $\gamma = (\gamma(1), \dots, \gamma(p))^{\mathrm{T}}$, then $\hat A = A + \Delta_A + o_p(n^{-1/2})$, whence

$$(\hat\phi_1, \dots, \hat\phi_p)^{\mathrm{T}} = \hat A^{-1}(\hat\gamma_1, \dots, \hat\gamma_p)^{\mathrm{T}} = (\phi_1, \dots, \phi_p)^{\mathrm{T}} + A^{-1}(\Delta - \Delta_A A^{-1}\gamma) + o_p(n^{-1/2}). \tag{A.4}$$

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

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

Therefore, the $j$th component of $\Delta - \Delta_A A^{-1}\gamma$ is equal to $n^{-1}\,\Sigma_{1\leqslant i\leqslant n}\,\varepsilon_i Z_{i+j} + o_p(n^{-1/2})$. Result (2.11) follows from this property and equation (A.4).

## Appendix B: Proof of result (2.17), and of version of result (2.11) with $(\tilde\phi_1, \dots, \tilde\phi_p)$ on the left-hand side

Put $N = n - l + 1$, $g_i = g(X_i)$,

$$U_i = \tilde g_i - g_i - \sum_j \phi_j(\tilde g_{i-j} - g_{i-j}),$$

$$V_i = \sum_{j=1}^{p}(\hat\phi_j - \phi_j)(\tilde g_{i-j} - g_{i-j}),$$

$$W_{1i} = \sum_{j=1}^{p}(\hat\phi_j - \phi_j)\varepsilon_{i-j}$$

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

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

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

Recall from equation (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 $\eta > 0$,

$$\sup_{h\in\mathcal{H}}\{(nh)^{-1/2}n^{\eta} + h^2\}^{-1}\max_{l\leqslant i\leqslant n}|\tilde g_i - g_i| = O_p(1) \tag{B.1}$$

as $n \to \infty$. (The proof of this result, and of result (B.3) below, uses the 'continuity argument'; see Cheng and Hall (2001), for example, for a discussion of this technique.) Therefore, because $\max_{1\leqslant i\leqslant n}|\varepsilon_i| = O_p(n^{\eta})$ for each $\eta > 0$ (since each $E(|\varepsilon_i|^c) < \infty$; see expression (2.16)), then

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

and $\max_{l\leqslant i\leqslant n}|W_{1i}| = O_p(n^{\eta-1/2})$, 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

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

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

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

where $\Delta_{jk} = \Sigma_{l \leqslant i \leqslant n} \, \varepsilon_{i-j}(\tilde{g}_{i-k} - g_{i-k})$ and $\Delta_j = \Sigma_{l \leqslant i \leqslant n} \, Z_i(\tilde{g}_{i-j} - g_{i-j})$. It may be proved that, for each $\eta > 0$,

$$\max_{h \in \mathcal{H}} [\{(nh)^{-1/2} + h^2\} n^{1/2+\eta}]^{-1} \left( \max_{1 \leqslant j \leqslant p, \, 0 \leqslant k \leqslant p} |\Delta_{jk}| + \max_{1 \leqslant j \leqslant p} |\Delta_j| \right) = O_p(1). \tag{B.3}$$

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

$$\mathrm{CV}_1 = S_1 + 2T_5 + R_2 = N^{-1} \sum_{i=l}^{n} (U_i - Z_i)^2 + R_3. \tag{B.4}$$

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

$$\sum_{k=1}^{p} (\tilde{\phi}_k - \phi_k^0) \sum_{i=l}^{n} \Delta_{ikk} \Delta_{ijj} = \sum_{i=l}^{n} (U_i - Z_i) \Delta_{ijj}, \tag{B.5}$$

where $\Delta_{ijk} = \tilde{g}_{i-j} - g_{i-j} - \varepsilon_{i-k}$. At the extremum, $h = \hat{h}_2 = \{1 + o_p(1)\} c n^{-1/5}$, where $c > 0$. It therefore follows from equations (B.1) and (B.3) that

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

Result (2.11), with its left-hand side replaced by $(\tilde{\phi}_1, \ldots, \tilde{\phi}_p)^{\mathrm{T}}$, now follows via equation (B.5). 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 equation (B.4), with $\tilde{\phi}_j$ replacing $\hat{\phi}_j$ for $1 \leqslant j \leqslant p$, to show that

$$\mathrm{CV}_2(\tilde{\phi}_1, \ldots, \tilde{\phi}_p, h) = N^{-1} \sum_{i=l}^{n} (U_i - Z_i)^2 + R_1', \tag{B.6}$$

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 minimizer $\hat{h}$, of $N^{-1} \, \Sigma_{l \leqslant i \leqslant n} \, (U_i - Z_i)^2$ over $h \in \mathcal{H}$, satisfies $\hat{h}/h_0 \to 1$ in probability and
(b)

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

where $\xi_1$ and $\xi_2$ denote random variables not depending on $h$, $\xi_1$ converges to a finite, strictly positive constant and $\xi_2$ has an asymptotic $N(0, \lambda)$ distribution with $0 < \lambda < \infty$. Combining this property with equations (B.4) and (B.6) we deduce that the minimizers $\hat{h}_1$ 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 $\xi_2$ in each case. This implies equation (2.17).

## References

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

Anh, V. V., Wolff, R. C., Gao, J. T. and Tieng, Q. (1999) Local linear regression with long-range dependent errors. *Aust. New Z. J. Statist.*, **41**, 463–479.

Cai, Z. and Masry, E. (2000) Nonparametric estimation of additive nonlinear ARX time series: local linear fitting and projections. *Econometr. Theory*, **16**, 465–501.

Cai, Z. and Ould-Said, E. (2001) Local robust regression estimation for time series. *Technical Report*. University of North Carolina, Charlotte.

Cheng, M.-Y. and Hall, P. (2001) Methods for tracking support boundaries with corners. Submitted to *Ann. Statist.*

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

Fan, J. and Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645–660.

Francisco-Fernandez, M. and Vilar-Fernandez, J. M. (2001) Local polynomial regression estimation with correlated errors. *Communs Statist. Theory 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.

Hall, P., Kay, J. W. and Titterington, D. M. (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.

Hansen, J. and Lebedeff, S. (1987) Global trends of measured surface air temperature. *J. Geophys. Res.* D, **92**, 13345–13372.

Härdle, W., Lütkepohl, H. and Chen, R. (1997) A review of nonparametric time series analysis. *Int. Statist. Rev.*, **65**, 49–72.

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

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

Hart, J. D. and Yi, S. (1998) One-sided cross-validation. *J. Am. Statist. Ass.*, **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. Process. Applic.*, **82**, 173–193.

Müller, H.-G. (1988) Nonparametric regression analysis of longitudinal data. *Lect. Notes Statist.*, **46**, 99 ff.

Müller, H.-G. and Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610–635.

Müller, H.-G. and Stadtmüller, U. (1988) 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. Statist. Comput. Sect. Am. Statist. Ass.*, 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.

Seifert, B., Gasser, T. and Wolf, A. (1993) Nonparametric estimation of residual variance revisited. *Biometrika*, **80**, 373–383.

Yao, Q. W. and Tong, H. W. (1998) Cross-validatory bandwidth selections for regression estimation based on dependent data. *J. Statist. Planng Inf.*, **68**, 387–415.