Austral. J. Statist. **39**(3), 1997, 313–324

NONPARAMETRIC AUTOCOVARIANCE FUNCTION ESTIMATION

R.J. HYNDMAN[1] AND M.P. WAND[2]

Monash University and University of NSW

Summary

Nonparametric estimators of autocovariance functions for non-stationary time series are developed. The estimators are based on straightforward nonparametric mean function estimation ideas and allow use of any linear smoother (e.g. smoothing spline, local polynomial). The paper studies the properties of the estimators, and illustrates their usefulness through application to some meteorological and seismic time series.

Key words: Bandwidth; correlated errors; kernel smoothing; local polynomial; nonparametric regression; non-stationary model; time series.

1. Introduction

Many time series exhibit non-stationarity in mean, variance and autocorrelation. An example of this is the series of daily maximum temperatures in Melbourne, Australia over the ten year period 1981–1990. Figure 1 shows each day's maximum plotted against the previous day's maximum temperature. The data have been split into six subsets based on the month of observation. These plots show the changing autocorrelation of the time series over the year with lowest autocorrelation in summer.

Of course, the month divisions used in Figure 1 are artificial boundaries. We assume the autocovariance and autocorrelation functions vary smoothly with time (or with some other exogenous variable). In this paper we consider the problem of estimating autocorrelation where the autocorrelation is not constant over the time series. The results derived are extensions of ideas in Ruppert *et al.* (1997) who estimate the variance function using linear smoothers assuming the data are uncorrelated. We derive estimates of autocovariance using linear smoothers and hence obtain estimates of autocorrelation.

Received February 1996; revised January 1997; accepted March 1997.

[1]Dept of Mathematics and Statistics, Monash University, Clayton, VIC 3168.
email: Rob.Hyndman@sci.monash.edu.au

[2]AGSM, University of New South Wales, Sydney, NSW 2052.

Acknowledgments. The authors thank Professor Rainer Dalhaus, Dr Alun Pope and an anonymous referee for helpful comments, and Dr Greg Houseman for the Kobe earthquake data. The data for the Kobe earthquake were obtained from the Data Management Centre, Washington University. Part of this work was carried out while the second author was a visitor to l'Institut de Statistique, Université Catholique de Louvain, Belgium, funded by a grant from l'Institut de Mathématique Pure et Appliquée of the same university. The first author was supported in part by an Australian Research Council grant.

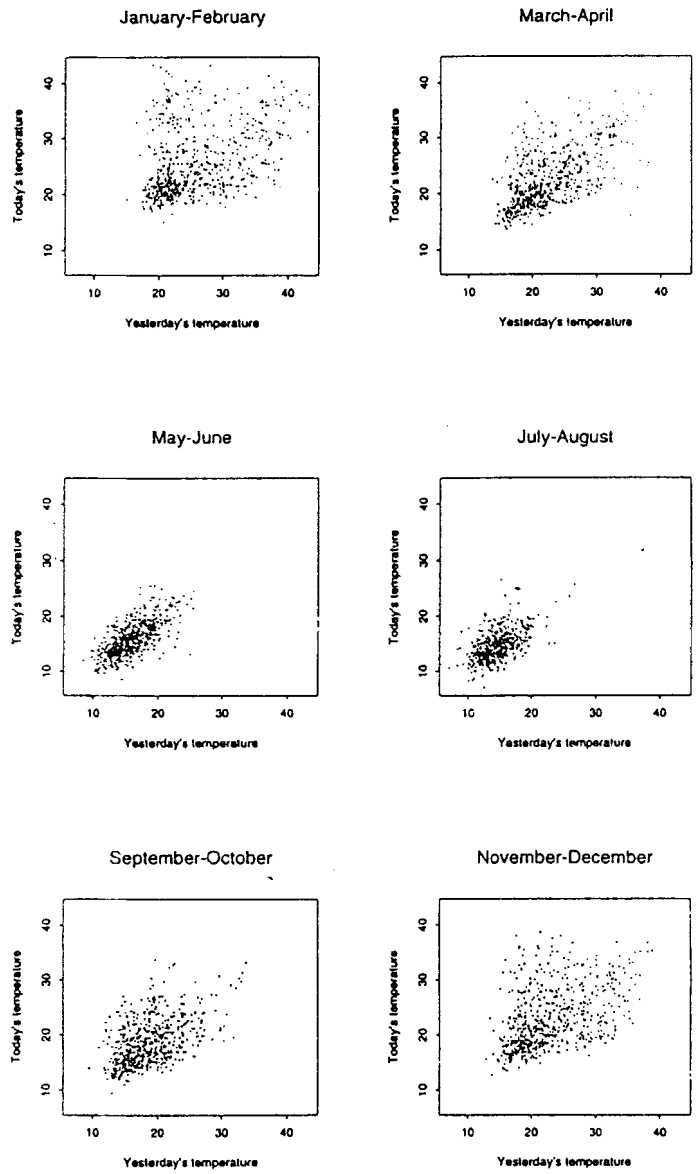

Fig. 1. Lag one scatterplots for the daily maximum temperatures. Note the changing first order autocorrelation throughout the year.

The closest comparable work known to us is that of Dahlhaus (1996) who proposes a kernel smoother of autocovariance when the mean is zero. Our estimators are more general, allowing any linear smoother (e.g. local polynomial, smoothing spline) and a varying mean function. Our estimates of local autocorrelation may also be used to estimate a locally stationary ARMA model as discussed in Dahlhaus (1997).

In Section 2 we formulate a general class of nonparametric autocovariance

function estimators, derive their bias and obtain a bias-adjusted estimator. Section 3 investigates the definiteness properties of the local covariance matrices obtained from the autocovariance functions. The theoretical properties of the simplest autocovariance function estimator are described in Section 4, and Section 5 briefly discusses bandwidth selection. Section 6 contains some illustrations of the methodology.

2. Formulation

2.1. General Linear Smoothers

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of random pairs that are assumed to satisfy the non-stationary, correlated errors model:

$$Y_i = m(X_i) + \varepsilon_i,$$

where $\{\varepsilon_i\}$ is a ν-dependent sequence, i.e. $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $|i - j| > \nu$. We also use the vector notation

$$\boldsymbol{Y} = \boldsymbol{m} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^T$, $\boldsymbol{X} = [X_1, \ldots, X_n]^T$, $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^T$ and $\boldsymbol{m} = \text{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = [m(X_1), \ldots, m(X_n)]^T$. We call $\boldsymbol{m}$ the *mean function* and

$$\boldsymbol{\gamma}_j = [\text{cov}(\varepsilon_1, \varepsilon_{1-j}), \ldots, \text{cov}(\varepsilon_n, \varepsilon_{n-j})]^T$$

the *autocovariance function at lag* j. Many regression data sets with correlated errors, including those on which our examples are based, are such that the X_is are deterministic and equally-spaced. However, since our results hold for general, possibly random, X_is , we work at this level of generality.

Denote the error covariance matrix by

$$V = \text{cov}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \text{diag}\ \boldsymbol{\gamma}_0 + \sum_{i=1}^{\nu} \{(\text{diag}\ \boldsymbol{\gamma}_i) B_i + B_i^T (\text{diag}\ \boldsymbol{\gamma}_i)\},$$

where B_i denotes the backshift matrix of order i. By this we mean that if $\boldsymbol{c} = [c_1, \ldots, c_n]^T$ is an arbitrary n-vector then

$$(B_i \boldsymbol{c})_j = c_{j-i}$$

(with the convention that $c_i = 0$ when i is zero or negative). It is easily seen that B_i is a sub-diagonal matrix with 1s on the ith sub-diagonal. For example, if $n = 4$ then

$$B_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Also, $B_0 = I$, the identity matrix. A further piece of matrix notation that we make considerable use of is $A \odot B$ for the element-wise, or Hadamard, product of two equally-sized matrices A and B. Finally, we let $\text{Diag}(A)$ denote the column vector containing the diagonal entries of a square matrix A.

Consider the class of estimates of $\boldsymbol{m}$, $\widehat{\boldsymbol{m}} = [\widehat{m}(X_1), \ldots, \widehat{m}(X_n)]^T$, of the form

$$\widehat{\boldsymbol{m}} = S\boldsymbol{Y}$$

where S is an $n \times n$ matrix depending on $\boldsymbol{X}$ with the property that $S\mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ denotes a vector of ones. We usually refer to $\widehat{\boldsymbol{m}}$ as a linear smoother of the data and to S as the smoother matrix. Examples of linear smoothers include ordinary least squares regression, moving averages, smoothing splines, regression splines and local polynomials (see e.g. Härdle, 1990; Hastie & Tibshirani, 1990).

Suppose that S_1 is the smoother matrix corresponding to an initial smoothing of the data and put

$$\boldsymbol{r} = (I - S_1)\boldsymbol{Y},$$

the vector of residuals. Then, noting that

$$\gamma_j = \mathrm{E}\{\boldsymbol{\varepsilon} \odot (B_j\boldsymbol{\varepsilon})\},$$

a plausible estimate of γ_j is

$$\hat{\gamma}_j = S_2\{\boldsymbol{r} \odot (B_j\boldsymbol{r})\}, \tag{1}$$

where S_2 is another smoother matrix.

For the special case of stationary data where $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \gamma(i-j)$ and S_2 equal to the smoother matrix that has all elements equal to n^{-1} we obtain

$$\hat{\gamma}_j = \mathbf{1}n^{-1} \sum_{i=j+1}^{n} r_i r_{i-j}.$$

This is the usual estimator of autocovariance for a stationary series (e.g. Brockwell & Davis, 1991 p.220).

2.2. Bias Adjustments

In ordinary least squares regression it is usual to adjust the least squares estimate of the error variance to make it unbiased. This involves changing the divisor of the residual sum of squares from n to $n-k$, where k is the number of parameters being estimated. In this section we show that analogous adjustments can be made to $\hat{\gamma}_j$.

Suppose that the bias of the initial smoother S_1 can be ignored. Then the conditional expectation of this estimator can be shown to be (see Theorem 1),

$$\mathrm{E}(\hat{\gamma}_j \mid \boldsymbol{X}) = S_2\,\mathrm{Diag}\,\{B_j(I - S_1)V(I - S_1)^T\}.$$

When the errors are stationary, i.e. when

$$V = \gamma(0)I + \sum_{i=1}^{\nu} \gamma(i)(B_i + B_i^T),$$

we obtain

$$\mathrm{E}(\hat{\gamma}_j \mid \boldsymbol{X}) = \sum_{i=0}^{\nu} \gamma(i) R_{ij} \tag{2}$$

where

$$R_{ij} = \begin{cases} S_2 \,\mathrm{Diag}\,\{B_j(I - S_1)(B_i + B_i^T)(I - S_1)^T\} & \text{if } i \geq 1, \\ S_2 \,\mathrm{Diag}\,\{B_j(I - S_1)(I - S_1)^T\} & \text{if } i = 0. \end{cases}$$

Therefore, the estimator is not conditionally unbiased when the errors are stationary and when $S_1\boldsymbol{Y}$ is unbiased. One could use (2) to form an adjustment of $\{\hat{\gamma}_0, \ldots, \hat{\gamma}_\nu\}$ that does have this unbiasedness property; however the resulting set of estimators is somewhat complicated. A simpler adjustment is one that uses the fact that, for most common smoothers the entries of R_{ij} $(i \neq j)$, are of a lower order of magnitude than the corresponding entries of R_{jj}. This can be shown by noting that in the case of low smoothing, $SB_i \simeq B_i$, while for higher smoothing SB_i is approximately equal to a matrix whose columns are either $n^{-1}\mathbf{1}$ or zero. The results for these cases then follow from the fact that $B_jB_i = I$ if $j = i$ and the zero matrix otherwise. With a bit more work, the same result can be established for intermediate amounts of smoothing. This gives rise to the approximate bias-corrected estimator

$$\tilde{\gamma}_j = \frac{S_2\{\boldsymbol{r} \odot (B_j\boldsymbol{r})\}}{R_{jj}},$$

where vector division is taken to be element-wise.

Note, in particular, that

$$\tilde{\gamma}_0 = \frac{S_2(\boldsymbol{r} \odot \boldsymbol{r})}{S_2 \,\mathrm{Diag}\,\{(I - S_1)(I - S_1)^T\}} = \frac{S_2(\boldsymbol{r} \odot \boldsymbol{r})}{\mathbf{1} + S_2 \,\mathrm{Diag}\,(S_1S_1^T - 2S_1)},$$

which is the unbiased (ignoring the bias in $S_1\boldsymbol{Y}$) estimator of the conditional variance function derived by Ruppert *et al.* (1997). Thus, $\tilde{\gamma}_j$ could be used as a bias adjusted alternative to $\hat{\gamma}_j$.

2.3. Local Polynomial Smoothers

The class of linear smoothers that we use in the examples are those commonly referred to as local polynomial smoothers (see e.g. Wand & Jones, 1995

pp.116ff). The (i,j) entry of the pth degree local polynomial smoother matrix, $S_{p,h}$, is

$$(S_{p,h})_{ij} = \boldsymbol{e}_1^T \{X_p(X_i)^T W_h(X_i) X_p(X_i)\}^{-1} X_p(X_i)^T W_h(X_i) \boldsymbol{e}_j \qquad (3)$$

where $\boldsymbol{e}_i$ is the column vector with 1 in the ith position and zeros elsewhere,

$$X_p(x) = \begin{bmatrix} 1 & X_1 - x & \dots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \dots & (X_n - x)^p \end{bmatrix} \quad \text{and} \quad W_h(x) = \operatorname*{diag}_{1 \le i \le n} K\Big(\frac{X_i - x}{h}\Big).$$

Typically K is a smooth bell-shaped function such as the standard normal density, called the kernel, and $h = h(x)$ is a scaling parameter, usually referred to as the bandwidth at the point x.

3. Definiteness Properties

Note that neither $\hat{\gamma}_j$ nor $\tilde{\gamma}_j$ leads to estimated covariance matrices that are guaranteed to be positive definite.

When the covariances are constant in time, they are usually estimated by (1) with S_2 equal to the $n \times n$ matrix with all entries equal to n^{-1}. This guarantees that the estimated autocovariance matrix

$$\hat{\Gamma} = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \dots & \hat{\gamma}(n-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \dots & \hat{\gamma}(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(n-1) & \hat{\gamma}(n-2) & \dots & \hat{\gamma}(0) \end{bmatrix},$$

where

$$\hat{\gamma}(j) = \frac{1}{n} \sum_{i=j+1}^{n} r_i r_{i-j},$$

is non-negative definite (e.g. Brockwell & Davis, 1991 p.no?). The non-negative definiteness property of $\hat{\Gamma}$ is sometimes desirable in time series analysis as it ensures that estimates of the spectral density are non-negative and that Yule-Walker estimates of autoregressive processes are causal. Therefore, unbiased estimators of autocorrelation are not often used (see Jenkins & Watts, 1968 Chap.5).

In the non-stationary setting, we have an analogous Toeplitz matrix for each time t, with first row

$$[\gamma_{0,t}, \gamma_{1,t}, \dots, \gamma_{n-1,t}]$$

for $t = 1, \dots, n$ where $\gamma_{j,t}$ is the tth element of γ_j. Replacing $\gamma_{j,t}$ by the tth element of one of the estimators $\hat{\gamma}_j$ or $\tilde{\gamma}_j$ leads to a matrix, $\hat{\Gamma}_t$, which

is not guaranteed to be non-negative definite. This is easily demonstrated in the case of the simplest estimator (1) by considering the trivial case where $\boldsymbol{r} = [0.5, -0.25, -0.25]^T$ and S_2 is constructed using a zero degree local polynomial smoother with kernel $K(u) = \frac{3}{4}(1-u^2)_+$ and bandwidth 1.01.

Then

$$S_2 = \begin{bmatrix} 0.9807 & 0.0193 & 0.0000 \\ 0.0190 & 0.9621 & 0.0190 \\ 0.0000 & 0.0193 & 0.9807 \end{bmatrix}, \qquad \hat{\Gamma}_2 = \begin{bmatrix} 0.0661 & -0.1191 & -0.0024 \\ -0.1191 & 0.0661 & -0.1191 \\ -0.0024 & -0.1191 & 0.0661 \end{bmatrix}$$

and

$$[1 \quad 1 \quad 1]\hat{\Gamma}_2[1 \quad 1 \quad 1]^T = -0.2841,$$

showing that $\hat{\Gamma}_t$ is not always positive definite. Therefore, it is apparent that positive definiteness needs to be sacrificed for simple linear smoother-based estimates of autocovariance functions.

4. Statistical Properties

We now present expressions for the conditional bias and covariance of $\hat{\gamma}_j$. For simplicity and ease of interpretation the covariance results are restricted to the situation where the errors are normally distributed. The Appendix, which contains the derivation of the main result in this section, explains how one may remove this restriction — but at the expense of some rather complicated algebraic expressions.

Theorem 1. *Let $\boldsymbol{b} = (S_1 - I)\boldsymbol{m}$ denote the bias vector of the smoother S_1. Then*

$$\mathrm{E}(\hat{\gamma}_j - \gamma_j \mid \boldsymbol{X}) = (S_2 - I)\gamma_j + S_2[\boldsymbol{b}_1 \odot (B_j \boldsymbol{b}_1) + \mathrm{Diag}\,\{B_j(S_1 V S_1^T - 2S_1 V)\}]. \quad (4)$$

If, in addition, the ε_i are normally distributed then

$$\begin{aligned} \mathrm{cov}(\hat{\gamma}_j \mid \boldsymbol{X}) = S_2[&\{(S_1 - I)V(S_1 - I)^T\} \odot \{B_j(S_1 - I)V(S_1 - I)^T B_j^T\} \\ &+ \{B_j(S_1 - I)V(S_1 - I)^T\} \odot \{B_j(S_1 - I)V(S_1 - I)^T\}^T \\ &+ (\boldsymbol{b}_1 \boldsymbol{b}_1^T) \odot \{B_j(S_1 - I)V(S_1 - I)^T B_j^T\} \\ &+ (\boldsymbol{b}_1 \boldsymbol{b}_1^T B_j^T) \odot \{B_j(S_1 - I)V(S_1 - I)^T\} \\ &+ (B_j \boldsymbol{b}_1 \boldsymbol{b}_1^T) \odot \{B_j(S_1 - I)V(S_1 - I)^T\}^T \\ &+ \{(S_1 - I)V(S_1 - I)^T\} \odot (B_j \boldsymbol{b}_1 \boldsymbol{b}_1^T B_j^T)]S_2^T. \end{aligned}$$

One can use Theorem 1 to gain a better understanding of the properties of $\hat{\gamma}_j$ by comparing them with those of

$$\tilde{\gamma}_j = S_2\{\boldsymbol{\varepsilon} \odot (B_j \boldsymbol{\varepsilon})\}.$$

Notice that $\bar{\gamma}_j$ is the estimate of γ_j based on the true, but unobservable, errors rather than the residuals from S_1. It is easy to see that

$$\mathrm{E}(\bar{\gamma}_j - \gamma_j \mid \boldsymbol{X}) = (S_2 - I)\gamma_j,$$

and, using Lemma 1 of the Appendix,

$$\mathrm{cov}(\bar{\gamma}_j \mid \boldsymbol{X}) = S_2 \,\mathrm{cov}\{\boldsymbol{\varepsilon} \odot (B_j\boldsymbol{\varepsilon})\} S_2^T = S_2\{V \odot (B_j V B_j^T) + (B_j V) \odot (V B_j^T)\} S_2^T.$$

Now suppose that the initial smoother S_1 possesses the properties

$$(S_1 - I)\boldsymbol{m} \xrightarrow{P} \mathbf{0} \qquad \text{and} \qquad S_1 V \xrightarrow{P} 0$$

as $n \to \infty$, where convergence of a matrix to a limit is taken to be element-wise. Provided $\nu = o(n)$, these conditions hold for most common smoothers including the local polynomial smoothers described in Section 2. Then it follows from Theorem 1 that

$$\begin{aligned} \mathrm{E}(\hat{\gamma}_j - \gamma_j \mid \boldsymbol{X}) &= \mathrm{E}(\bar{\gamma}_j - \gamma_j \mid \boldsymbol{X}) + o_P\{\mathrm{E}(\bar{\gamma}_j - \gamma_j \mid \boldsymbol{X})\} \\ \text{and} \quad \mathrm{cov}(\hat{\gamma}_j \mid \boldsymbol{X}) &= \mathrm{cov}(\bar{\gamma}_j \mid \boldsymbol{X}) + o_P\{\mathrm{cov}(\bar{\gamma}_j \mid \boldsymbol{X})\} \end{aligned}$$

for large n. Thus, the initial smoother S_1 has only a second-order asymptotic effect on the bias and variance of $\hat{\gamma}_j$, and asymptotically, $\hat{\gamma}_j$ performs as well as it would if the errors were observable. This phenomenon has been observed in variance function estimation by Hall & Carroll (1989) and Ruppert *et al.* (1997).

5. Bandwidth Selection

In the case of local polynomial smoothing, the bandwidths h_1 and h_2, corresponding to S_1 and S_2, need to be specified. In the previous section we argued that S_1, and therefore h_1, has only a second-order effect on the performance of $\hat{\gamma}_j$; however, this is an asymptotic result and, in practice, its choice will have some effect on the estimator. One could use Theorem 1 to develop rules for choosing h_1 and h_2 jointly. However, as remarked in Ruppert *et al.* (1997), a simpler strategy, that still leads to asymptotic optimality when the degrees of the local polynomials of each smoother are equal, is to choose h_1 using a bandwidth selector that is optimal for estimation of m. Then apply the same rule to choose the bandwidth h_2 for smoothing the $\boldsymbol{r} \odot (B_j\boldsymbol{r})$ vector. However, the difference between the settings considered here and by Ruppert *et al.* (1997) is that the errors are correlated — so bandwidths selectors that take correlations into account would need to be used to achieve asymptotic optimality. Proposals for bandwidth selection in the presence of correlated errors, that could be extended to local polynomials and possibly to non-stationarity, include those of Altman (1990), Chu & Marron (1991), Hart (1991, 1994) and Herrmann *et al.* (1992).

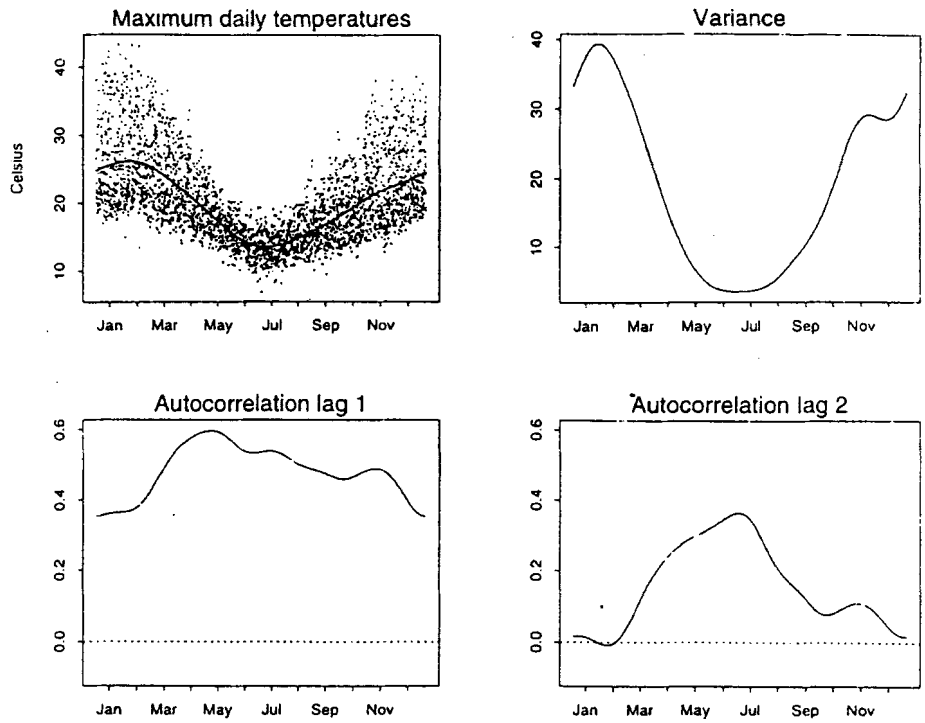

Fig. 2. Ten years of daily maximum temperatures for Melbourne plotted against day of observation (top left); — estimated smoothing mean. The other plots show estimates of the variance and the first two autocorrelation functions for these data.

However, these extensions have not yet been performed so a practical recommendation for automatic bandwidth selection cannot be made at this stage. In our examples the bandwidths were chosen subjectively.

6. Examples

6.1. Melbourne Maximum Daily Temperatures

We now return to the maximum daily temperature data described in Section 1. Figure 2 shows each day's maximum temperature plotted against the day of the year; the temperatures for the three occurrences of 29 February have been removed to simplify the plots and computations. The changing mean is shown as a solid line. The X variable here is defined as the day of observation, i.e. the numbers from 1 to 365, repeated 10 times over the period in which the data were observed.

Figure 2 also shows the changing variance and autocorrelation over time for these data. The linear smoother used was a local linear regression (e.g. Wand & Jones, 1995 pp.116ff) with kernel

$$K(u/h) = \tfrac{3}{4}\left[1 - \frac{1}{h^2}\{\min(|u|, 365 - |u|)\}^2\right]_+ \qquad (|u| < 365),$$

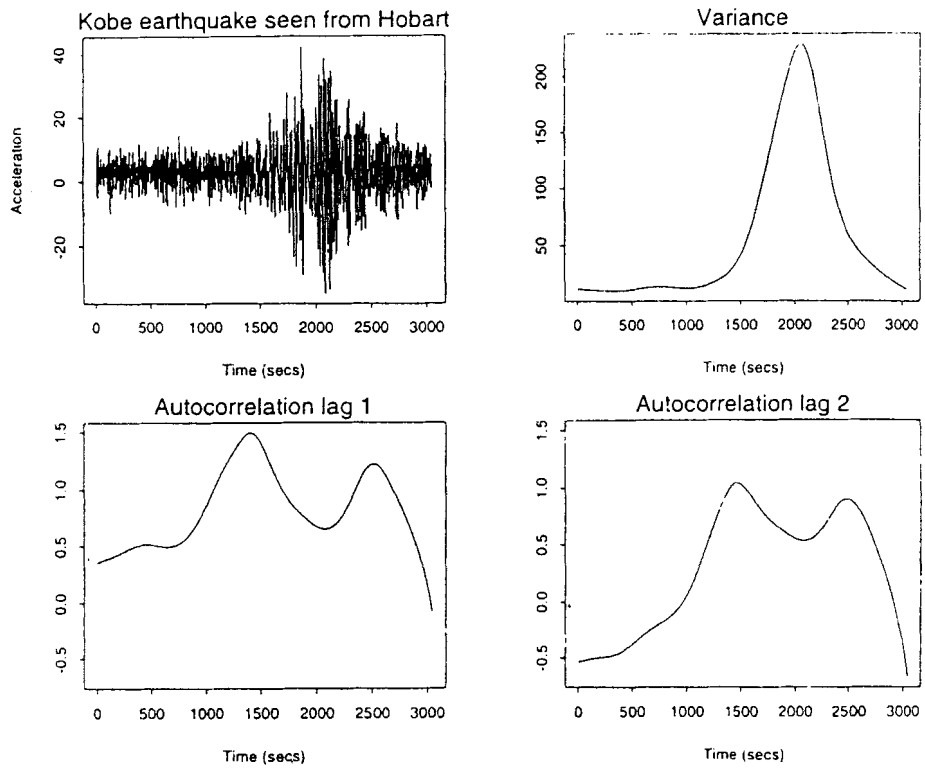

Fig. 3. A seismogram (vertical acceleration, nm/s^2) of the Kobe earthquake, recorded in Hobart, with estimates of the variance and the first two autocorrelation functions

which takes the seasonality into account. The initial smoother (S_1) to compute the mean, and the smoother (S_2) to compute the variance, each used a bandwidth of 30 days. The smoothers (S_2) that compute the autocovariance used a bandwidth of 50 days. All bandwidths were subjectively chosen and the estimates were computed using (1).

Clearly the serial correlation is higher in winter than in summer. The varying autocorrelation reflects the varying difficulty in meteorological forecasting through the year. When the temperature and other meteorological time series have higher serial correlation, they are more predictable.

6.2. Kobe Earthquake

Figure 3 shows a seismogram recorded at the University of Tasmania, Hobart, during the Kobe earthquake in Japan. The seismogram is for 16 January 1995, beginning at 20:56:51 (GMT) and continuing for 51 minutes at 1 second intervals. The particular attribute shown is the vertical acceleration in nm/s^2. The X variable used is simply the number of seconds since the start of observations on this seismograph.

The first earthquake waves occurred between 500 and 1000 seconds from the start of the record but are almost impossible to see in the time plot because of the high level of background noise. However, the autocorrelation estimates

pick up the seismological waves earlier. Interestingly, the autocorrelations do not return to their former levels after the main shockwaves, but become more negative.

The mean function $\boldsymbol{m}$ was taken to be constant and estimated as the mean of the data. The variance function was estimated using a local linear regression with kernel $K(u) = \frac{3}{4}(1-u^2)_+$ and bandwidth 150 seconds. The autocovariance estimates were calculated similarly but with bandwidths of 300 seconds. Again, all bandwidths were chosen subjectively and the estimates were computed using (1).

These estimates could be used to produce a locally stationary model of the seismogram for simulation purposes or to identify the earthquake activity earlier than is evident from the time plot.

Appendix: Proof of Theorem 1

Noting the result $\boldsymbol{a} \odot \boldsymbol{b} = \operatorname{Diag}(\boldsymbol{b}\boldsymbol{a}^T)$ for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\mathrm{E}(\hat{\gamma}_j \mid \boldsymbol{X})$ equals

$$\begin{aligned}
&S_2 \mathrm{E}[\operatorname{Diag}\{B_j(I-S_1)\boldsymbol{Y}\boldsymbol{Y}^T(I-S_1)^T\}] \\
&= S_2[\operatorname{Diag}\{B_j(S_1-I)(\boldsymbol{m}\boldsymbol{m}^T+V)(S_1-I)^T\}] \\
&= S_2[\{\boldsymbol{b}_1 \odot (B_j\boldsymbol{b}_1)\} + \operatorname{Diag}(B_jV) + \operatorname{Diag}\{B_j(S_1VS_1^T - 2S_1V)\}] \\
&= S_2\gamma_j + S_2[\{\boldsymbol{b}_1 \odot (B_j\boldsymbol{b}_1)\} + \operatorname{Diag}\{B_j(S_1VS_1^T - 2S_1V)\}],
\end{aligned}$$

from which the result for the conditional bias follows immediately.

The result for $\operatorname{cov}(\hat{\gamma}_j \mid \boldsymbol{X})$ follows directly from Lemma 1.

Lemma 1. *Suppose that the random vector $\boldsymbol{Y}$ has a multivariate normal distribution with mean vector $\boldsymbol{m}$ and covariance matrix V. Then for any square constant matrix A having the same number of rows as $\boldsymbol{Y}$,*

$$\begin{aligned}
&\operatorname{cov}\{(A\boldsymbol{Y}) \odot (B_sA\boldsymbol{Y})\} \\
&\quad = (AVA^T) \odot (B_sAVA^TB_s^T) + (B_sAVA^T) \odot (B_sAVA^T)^T \\
&\qquad + \{A\boldsymbol{m}(A\boldsymbol{m})^T\} \odot (B_sAVA^TB_s^T) + \{A\boldsymbol{m}(B_sA\boldsymbol{m})^T\} \odot (B_sAVA^T) \\
&\qquad + \{B_sA\boldsymbol{m}(A\boldsymbol{m})^T\} \odot (AVA^TB_s^T) + (AVA^T) \odot \{B_sA\boldsymbol{m}(B_sA\boldsymbol{m})^T\}.
\end{aligned}$$

Proof. Let a_{ij} denote the (i,j) entry of A and, following McCullagh (1987), let the generalised cumulants of $\boldsymbol{Y}$ be denoted by κ with an appropriate partitioned superscript. For example,

$$\kappa^{i,j} = \operatorname{cum}(Y_i, Y_j) = \operatorname{cov}(Y_i, Y_j) \qquad \text{and} \qquad \kappa^{i,j,k\ell} = \operatorname{cum}(Y_i, Y_j, Y_kY_\ell).$$

Then the (p,q) entry of $\operatorname{cov}\{(A\boldsymbol{Y}) \odot (B_sA\boldsymbol{Y})\}$ is

$$\operatorname{cov}\{(A\boldsymbol{Y}) \odot (B_sA\boldsymbol{Y})\}_{pq} = \sum_i \sum_j \sum_k \sum_\ell a_{pi}a_{p-s,j}a_{qk}a_{q-s,\ell}\kappa^{ij,k\ell}.$$

From McCullagh (1987 p.58) we have the identity

$$\kappa^{ij,k\ell} = \kappa^{i,j,k,\ell} + \kappa^{i}\kappa^{j,k,\ell} + \kappa^{j}\kappa^{i,k,\ell} + \kappa^{k}\kappa^{i,j,\ell} + \kappa^{\ell}\kappa^{i,j,k} + \kappa^{i,k}\kappa^{j,\ell} + \kappa^{i,\ell}\kappa^{j,k} + \kappa^{i}\kappa^{k}\kappa^{j,\ell} + \kappa^{i}\kappa^{\ell}\kappa^{j,k} + \kappa^{j}\kappa^{k}\kappa^{i,\ell} + \kappa^{j}\kappa^{\ell}\kappa^{i,k}.$$

For normal $\boldsymbol{Y}$, $\kappa^{i,j,k} = \kappa^{i,j,k,\ell} = 0$. The stated result then follows by noting that $\kappa^{i} = m_i$ and $\kappa^{i,j} = V_{ij}$.

The above proof shows how the normality assumption can be removed. For general $\boldsymbol{Y}$, the (p,q) entry of $\mathrm{cov}\{(A\boldsymbol{Y}) \odot (B_s A\boldsymbol{Y})\}$ depends on terms such as

$$\sum_i \sum_j \sum_k \sum_\ell a_{pi} a_{p-s,j} a_{qk} a_{q-s,\ell} \mathrm{E}\{(Y_i - m_i)(Y_j - m_j)(Y_k - m_k)(Y_\ell - m_\ell)\}.$$

However, it is not easy to express this term in matrix notation so we satisfy ourselves with the result for normal $\boldsymbol{Y}$.

References

ALTMAN, N.S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749–759.

BROCKWELL, P.J. & DAVIS, R.A. (1991). *Time Series: Theory and Methods.* New York: Springer-Verlag.

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

DAHLHAUS, R. (1996). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens Conference on Applied Probability and Time Series Analysis*, Vol.II, eds P.M. Robinson and M. Rosenblatt. New York: Springer Verlag.

— (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25**, 1–37.

HALL, P. & CARROLL, R.J. (1989). Variance function estimation in regression: the effect of estimating the mean. *J. Roy. Statist. Soc. Ser. B* **51**, 3–14.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* New York: Cambridge University Press.

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

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

HASTIE, T.J. & TIBSHIRANI, R.J. (1990). *Generalized Additive Models.* New York: Chapman and Hall.

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

JENKINS, G.M. & WATTS, D.M. (1968). *Spectral Analysis and its Applications.* San Francisco: Holden-Day.

McCULLAGH, P. (1987). *Tensor Methods in Statistics.* London: Chapman and Hall.

RUPPERT, D., WAND, M.P., HOLST, U. & HÖSSJER, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262–273.

WAND, M.P. & JONES, M.C. (1995). *Kernel Smoothing.* London: Chapman and Hall.