Biometrika (2003), **90,** 4, *pp.* 831–844

Printed in Great Britain

Nonparametric estimation of large covariance matrices of longitudinal data

BY WEI BIAO WU

Department of Statistics, The University of Chicago, Chicago, Illinois 60637, U.S.A.

wbwu@galton.uchicago.edu

AND MOHSEN POURAHMADI

Division of Statistics, Northern Illinois University, DeKalb, Illinois 60115, U.S.A.

pourahm@math.niu.edu

SUMMARY

Estimation of an unstructured covariance matrix is difficult because of its positive-definiteness constraint. This obstacle is removed by regressing each variable on its predecessors, so that estimation of a covariance matrix is shown to be equivalent to that of estimating a sequence of varying-coefficient and varying-order regression models. Our framework is similar to the use of increasing-order autoregressive models in approximating the covariance matrix or the spectrum of a stationary time series. As an illustration, we adopt Fan & Zhang's (2000) two-step estimation of functional linear models and propose nonparametric estimators of covariance matrices which are guaranteed to be positive definite. For parsimony a suitable order for the sequence of (auto)regression models is found using penalised likelihood criteria like AIC and BIC. Some asymptotic results for the local polynomial estimators of components of a covariance matrix are established. Two longitudinal datasets are analysed to illustrate the methodology. A simulation study reveals the advantage of the nonparametric covariance estimator over the sample covariance matrix for large covariance matrices.

Some key words: Cholesky decomposition; Covariance estimation; Local polynomial regression; Longitudinal study; Order selection; Varying-coefficient regression.

1. INTRODUCTION

The sample covariance matrix provides an unbiased estimator for the population covariance matrix and is generally positive definite. Various improved estimators inspired by decision-theoretic considerations and constructed by shrinking its eigenvalues (Lin & Perlman, 1985) remain unstable because of the high dimensionality of a covariance matrix, particularly when many variables are measured on a few subjects (Krzanowski et al., 1995). In the longitudinal-data literature (Diggle et al., 1994) it is a common practice to pick a stationary covariance matrix with few parameters from a menu provided by popular software packages. Of course, such estimators of the covariance matrix could have considerable bias when the selected structure is far from the truth. A Bayesian way of resolving this difficulty is to place priors on a covariance matrix so as to 'shrink' or 'smooth' it toward some structures (Daniels & Kass, 2001). This can provide robustness to misspecifi-

cation of the structure and offer stability over assuming no structure. The commonly used inverse Wishart prior offers only one parameter for controlling the amount of shrinkage or smoothing, while priors of Daniels & Kass (2001), based on decomposing a matrix into eigenvalues and Givens angles or into the variances and correlations, offer two. More flexible priors with many parameters to control shrinkage can be introduced using an unconstrained and statistically meaningful reparameterisation of the covariance matrix (Daniels & Pourahmadi, 2002).

To strike a balance between variability and bias of covariance estimators it is prudent to restrict attention to covariance structures suggested by the data. To this end, nonparametric estimators of covariance structures are useful either as a guide to the formulation of parametric models or as the basis for formal inference without imposing parametric assumptions. However, most nonparametric estimators of covariance matrices are developed either for stationary processes or without heeding the positive-definiteness constraint; see Glasbey (1988), Shapiro & Botha (1991), Sampson & Guttorp (1992), Hall et al. (1994) and Hall & Patil (1994). Diggle & Verbyla (1998) introduced a nonparametric estimator for the covariance structure of longitudinal data without assuming stationarity, but their estimator, based on kernel weighted local linear regression smoothing of sample variogram ordinates and of squared residuals, is not guaranteed to be positive definite.

In this follow-up paper to Pourahmadi (1999, 2000) and Daniels & Pourahmadi (2002), we provide nonparametric, positive-definite estimators of covariance matrices without assuming stationarity. The key idea is that the covariance matrix $\Sigma = (\sigma_{ij})$ of a zero-mean random vector $Y = (y_1, \ldots, y_m)'$ can be diagonalised by a lower triangular matrix constructed from the regression coefficients when y_t is regressed on its predecessors $y_1, \ldots, y_{t-1}$. More precisely, for $t = 2, \ldots, m$, we have

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \varepsilon_t, \quad T\Sigma T' = D, \tag{1}$$

where T and D are unique, T is a unit lower triangular matrix having ones on its diagonal and $-\phi_{tj}$ at its (t, j)th element for $j < t$, and D is diagonal with $\sigma_t^2 = \mathrm{var}(\varepsilon_t)$ as its diagonal entries (Pourahmadi, 1999, 2000). The decomposition (1) converts the constraint entries of Σ into two groups of unconstrained 'regression' and 'variance' parameters given by $\{\phi_{tj}, t = 2, \ldots, m;\ j = 1, \ldots, t-1\}$ and $\{\log \sigma_1^2, \ldots, \log \sigma_m^2\}$, respectively. Conceptually, this approach reduces the difficult and nonintuitive task of modelling a covariance matrix to the more familiar task of modelling $m - 1$ regression problems. The more general autoregressive and moving average, ARMA, models and Akaike's idea of spectral density estimation by fitting autoregressive models of successively increasing orders (Berk, 1974) are excellent instances of modelling or approximating stationary covariance matrices using regression-like models (Pourahmadi, 2001, § 2.6.3).

We rely on nonparametric methods such as local polynomial estimators (Fan & Gijbels, 1996) to smooth the subdiagonals of T. If we denote estimators of the components of Σ in (1) by $\hat{T}$ and $\hat{D}$, an estimator of Σ given by $\hat{\Sigma} = \hat{T}^{-1}\hat{D}\hat{T}^{-1'}$ is guaranteed to be positive-definite. Our idea of smoothing along the subdiagonals of T is motivated by the similarity of the regressions in (1) to varying-coefficients autogressions (Subba Rao, 1970; Kitagawa & Gersch, 1985; Dahlhaus, 1997):

$$\sum_{j=0}^{p} f_{j,m}(t/m) y_{t-j} = \sigma_m(t/m)\varepsilon_t \quad (t = 0, 1, \ldots), \tag{2}$$

where $f_{0,m}(u) \equiv 1$, $f_{j,m}(u)$ $(1 \leqslant j \leqslant p)$ and $\sigma_m(u)$ are continuous functions on $[0, 1]$ and $\{\varepsilon_t\}$

is a sequence of independent random variables each with mean zero and variance one. This analogy suggests taking the nonredundant entries of T and D as realisations of some smooth functions:

$$\phi_{t,t-j} = f_{j,m}(t/m), \quad \sigma_t = \sigma_m(t/m). \tag{3}$$

Thus, as in nonparametric regression, (3) provides an intuitively appealing framework whereby it appears that one is observing $f_{j,m}(.)$ and $\sigma_m(.)$ on a finer grid as m gets large.

In § 2, we clarify the close relationship between estimation of covariance matrices and estimation of functional linear models. Section 3 provides a detailed discussion of smoothing and selecting the orders of the regression models. The methodology is illustrated in § 4 using two longitudinal datasets. In § 5, simulation studies are presented to compare performance of sample covariance matrices with smoothed estimators. Asymptotic bounds for the bias and variance of smoothed regression coefficients are given in § 6. Our focus here is on methodology; more technical results about appropriate definitions of AIC and BIC, their asymptotic properties and consistency of the ensuing covariance-matrix estimator are currently under study.

2. Modelling covariances via regressions

2·1. *Regression representation of a random vector*

For a time-ordered random vector $Y = (y_1, \ldots, y_m)'$ with zero mean and positive definite covariance matrix Σ, let $\hat{y}_t$ be the linear least-squares predictor of y_t based on its predecessors $y_{t-1}, \ldots, y_1$ and let $\varepsilon_t = y_t - \hat{y}_t$ be its prediction error with variance $\sigma_t^2 = \mathrm{var}(\varepsilon_t)$. Then, for certain scalars ϕ_{tj}, we have (1) where the ϕ_{tj}'s and the variances σ_t^2 are computed from Σ and have statistical meaning as regression coefficients and prediction variances (Pourahmadi, 1999). Let $D = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ and let T be the unit lower triangular matrix with $-\phi_{tj}$ as its (t, j)th entry. Then the regression models in (1), for $t = 1, \ldots, m$, written in matrix form lead to the desired factorisation such that T and D are the components of the modified Cholesky decomposition of Σ.

When Σ is unstructured, the nonredundant entries of T and $\log D = \mathrm{diag}(\log \sigma_1^2, \ldots, \log \sigma_m^2)$ are unconstrained, and the $m(m+1)/2$ constrained and hard-to-model parameters of Σ can be replaced by the same number of unconstrained parameters of the unique pair $(T, \log D)$. The new parameters ϕ_{tj} and σ_t^2 are referred to as the generalised autoregressive parameters and the innovation variances of Σ. Thus, by analogy with the generalised linear models for a mean vector, the link function suggested by the factorisation in (1) is the function $g(.)$ defined on the set of positive-definite matrices by

$$g(\Sigma) = \log D + T + T' - 2I. \tag{4}$$

Since the entries of $g(\Sigma)$ are unconstrained and statistically meaningful, the dimension of the parameter space can be reduced considerably by using covariates (Pourahmadi, 1999, 2000; Daniels & Pourahmadi, 2002) and by modelling these entries parametrically, nonparametrically or in a Bayesian way. An attractive feature of this reparameterisation is that, regardless of the modelling approach, the estimated covariance matrix is guaranteed to be positive definite.

The procedures of Fan & Zhang (2000) and Huang et al. (2002) for estimating varying-coefficients regression models for the mean of longitudinal data can be exploited in the estimation of the covariance matrix of such data. We confine our attention to the balanced-data case. Let $\{t_j, j = 1, \ldots, m\}$ be the common times over which the response y_{ij} is

measured on the ith subject and let $X_{ij} = (X_{ij1}, \ldots, X_{ijd})'$ be the d covariates measured at time t_j. This gives rise to balanced longitudinal data of the form

$$(t_j, X_{ij}, y_{ij}) \quad (i = 1, \ldots, n;\ j = 1, \ldots, m).$$

A simple model capable of accommodating the objectives of a longitudinal study is the functional linear model

$$y_i(t_j) = X_i(t_j)'\beta(t_j) + \varepsilon_i(t_j), \tag{5}$$

where $y_i(t_j) = y_{ij}$, $X_i(t_j) = X_{ij}$, $\beta(t_j) = (\beta_1(t_j), \ldots, \beta_d(t_j))'$ and $\{\varepsilon_i(t)\}$ is a zero-mean stochastic process with covariance function

$$\sigma_{s,t} = \mathrm{cov}\{\varepsilon_i(s), \varepsilon_i(t)\}. \tag{6}$$

The two-step procedure for estimating $\beta(.)$ in (5) is as follows (Fan & Zhang, 2000).

Step 1. Use the data at t_j and the ordinary least-squares method to fit the linear model (5) and to obtain the raw estimates $b(t_j) = (b_1(t_j), \ldots, b_d(t_j))'$ for $\beta(t_j) = (\beta_1(t_j), \ldots, \beta_d(t_j))'$.

Step 2. Apply a smoothing technique to the data $\{(t_j, b_r(t_j)), j = 1, \ldots, m\}$ to obtain a smooth estimate for the coefficient functions $\beta_r(.)$ $(r = 1, \ldots, d)$.

For our application of the above procedure to nonparametric estimation of covariance matrices via decomposition (1), it is instructive to concatenate the d-vectors $b(t_j)$, $\beta(t_j)$, for $j = 1, \ldots, m$, as the $d \times m$ matrices.

$$B = (b(t_1), \ldots, b(t_m)), \quad \mathscr{B} = (\beta(t_1), \ldots, \beta(t_m)), \tag{7}$$

and to view Step 2 as smoothing the rows of the matrix B.

2·2. *Least-squares estimators of a covariance matrix*

In this section we introduce a raw estimator $(\hat{T}, \hat{D})$ of a covariance matrix based on least-squares estimators of the regression coefficients and residual variances in (1).

Consider a sample of size n from a normal population with mean zero and covariance matrix Σ and set $Y_i = (y_{i,1}, \ldots, y_{i,m})'$ $(i = 1, \ldots, n)$, where $n > m$. A reinterpretation of (1), for each fixed $t = 2, \ldots, m$, as a varying-coefficient regression model of order $t-1$ allows us to obtain the least-squares estimators of $\phi_t = (\phi_{t1}, \ldots, \phi_{t,t-1})'$ and σ_t^2 by regressing $y_{i,t}$ on $(y_{i,t-1}, \ldots, y_{i,1})'$:

$$y_{t,i} = \sum_{k=1}^{t-1} \phi_{tk}\, y_{k,i} + \varepsilon_{t,i} \quad (i = 1, 2, \ldots, n). \tag{8}$$

In vector-matrix form this gives

$$\begin{bmatrix} y_{t,1} \\ \vdots \\ y_{t,n} \end{bmatrix} = \begin{bmatrix} y_{t-1,1} & \cdots & y_{1,1} \\ \vdots & & \vdots \\ y_{t-1,n} & \cdots & y_{1,n} \end{bmatrix} \begin{bmatrix} \phi_{t1} \\ \vdots \\ \phi_{t,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t,1} \\ \vdots \\ \varepsilon_{t,n} \end{bmatrix},$$

or

$$Y_t = X_t \phi_t + \varepsilon_t, \tag{9}$$

where Y_t is $n \times 1$, X_t is $n \times (t-1)$ and $\phi_t = (\phi_{t1}, \ldots, \phi_{t,t-1})'$ is the $(t-1) \times 1$ vector of nonredundant entries on the tth row of the matrix T and ε_t is the vector of errors with

$$E(\varepsilon_t) = 0, \quad \mathrm{cov}(\varepsilon_t) = \sigma_t^2 I_n. \tag{10}$$

Using ordinary least-squares theory, from (9) we obtain the estimators $(\hat{T}, \hat{D})$ with

$$\hat{\phi}_t = (X_t' X_t)^{-1} X_t' Y_t, \quad \hat{\sigma}_t^2 = \frac{\hat{e}_t' \hat{e}_t}{n-(t-1)}, \tag{11}$$

where $\hat{e}_t = (I_n - P_t) Y_t$ and $P_t = X_t (X_t' X_t)^{-1} X_t'$. Moreover, we have

$$E(\hat{\phi}_t | \mathscr{D}_t) = \phi_t, \quad \operatorname{cov}(\hat{\phi}_t | \mathscr{D}_t) = \sigma_t^2 (X_t' X_t)^{-1}, \tag{12}$$

where henceforth $\mathscr{D}_t = \{y_{i,s};\, s \leqslant t-1, i = 1, \ldots, n\}$. From (11)–(12) and normal linear model theory it follows that, for $t \geqslant 2$,

$$(\hat{\phi}_t | \mathscr{D}_t) \sim N\{\phi_t, \sigma_t^2 (X_t' X_t)^{-1}\}, \tag{13}$$

independently of

$$(\hat{\sigma}_t^2 | \mathscr{D}_t) \sim \frac{\sigma_t^2 \chi_{n-t+1}^2}{n-t+1}, \tag{14}$$

and consequently, for $j = 1, \ldots, t-1$,

$$\left(\frac{\hat{\sigma}_t^2 s_t^{jj}}{n-t+1} \right)^{-\frac{1}{2}} (\hat{\phi}_{tj} - \phi_{tj}) | \mathscr{D}_t \sim t_{n-t+1}, \tag{15}$$

where s_t^{jj} is the jth diagonal entry of $(X_t' X_t)^{-1}$. The use of (15) in testing whether or not certain entries of T are zero is crucial for simplifying the structure of T and is closely related to the use of a (partial) correlogram in identifying time series models and the rapidly growing area of Gaussian graphical models, Gabriel's (1962) antedependence and the emerging class of time-varying autoregression models (Macchiavelli & Arnold, 1994; Kitagawa & Gersch, 1996, p. 149).

3. Nonparametric estimators of a covariance matrix

3·1. *Preamble*

Historically, most nonparametric estimators of a covariance matrix proceed by estimating its individual entries (Glasbey, 1988; Sampson & Guttorp, 1992). However, recently Diggle & Verbyla (1998) have proposed a nonparametric estimator of the covariance matrix by smoothing separately its variance function and the variogram:

$$\sigma^2(t) = \operatorname{var}(y_t), \quad \gamma(s,t) = \frac{1}{2} E\{y_s - y_t - E(y_s - y_t)\}^2. \tag{16}$$

These two smooth components are then combined using

$$\sigma_{s,t} = \begin{cases} \frac{1}{2}\{\sigma^2(s) + \sigma^2(t)\} - \gamma(s,t) & (s \neq t), \\ \sigma^2(t) & (s = t), \end{cases} \tag{17}$$

to provide a smooth estimator of the covariance matrix or function which may not be positive-definite. Our approach is similar, but instead of using the additive decomposition (17) we rely on the multiplicative Cholesky decomposition (1), which guarantees the positive-definiteness of the estimated covariance matrix obtained by smoothing $(\hat{T}, \hat{D})$.

3·2. *Smoothing $\hat{T}$ along its subdiagonals*

The analogy between smoothing $\hat{T}$ and the mean function in Fan & Zhang (2000) can be seen through the similarities of the matrices B in (7) and $\hat{T}$. While for mean estimation

the d rows of B are smoothed, for the covariance estimation there are the options of smoothing the rows, columns and subdiagonals of $\hat{T}$ viewed as the observed values of some smooth univariate functions; see (2) and (3). There is also the possibility of smoothing the lower half of $\hat{T}$ viewed as the observed values of a smooth bivariate function. For many applications, it is more relevant to smooth $\hat{T}$ along its subdiagonals, since its jth subdiagonal entries stand for lag-j regression coefficients over time and relate to time-varying autoregression models. Furthermore, since the raw estimators of the last subdiagonals of T are unreliable and shorter, one needs to choose the number of subdiagonals to be smoothed; see § 3·3.

Let T_j and $\hat{T}_j$ $(j = 1, \ldots, m-1)$ stand for the vector of entries on the jth subdiagonals of T and $\hat{T}$, respectively. Note that, if the data or Σ were nearly stationary, one would expect the entries of $\hat{T}_j$ and T_j to be nearly the same. Thus, as a departure from stationarity it is natural to assume that the entries of T_j are realisations of a smooth function $f_{j,m}(.)$ on $[0, 1]$; that is $\phi_{t,t-j} = f_{j,m}(t/m)$, as in (3). By direct analogy with nonparametric regression (Fan & Gijbels, 1996, p. 28), this framework makes it possible to develop an asymptotic theory of estimation as if we observed $f_{j,m}(.)$ on a finer grid for larger m. This type of smoothness is also in line with the notions of local stationarity studied in the literature of time series analysis and signal processing (Dahlhaus, 1997; Kitagawa & Gersch, 1996). A similar smoothness assumption can be made about the logarithms of the innovation variances or the diagonal entries of $g(\Sigma)$ in (4). In the formulae that follow, it is convenient to think of the $j = 0$ case as corresponding to the main diagonal of $g(\Sigma)$.

Since most smoothing techniques are linear in the response, our typical smoother is of the form

$$\bar{\phi}_{t,t-j} = \sum_{k=j+1}^{m} w_j(k, t)\hat{\phi}_{k,k-j} \quad (j = 0, 1, \ldots, \lfloor m^{1/3} \rfloor), \tag{18}$$

where the weights $w_j(k, t)$ are determined by the smoothing technique used. It is of some theoretical interest to note that the rows of $\hat{\phi}_t$ of $\hat{T}$ are uncorrelated. Thus

$$E(\bar{\phi}_{t,t-j}) = \sum_{k=j+1}^{m} w_j(k, t)\phi_{k,k-j}, \tag{19}$$

and $\mathrm{var}(\bar{\phi}_{t,t-j})$ for local polynomial estimators is given in (25). For local polynomial estimators in this paper, we used the automatic bandwidth selector introduced in Ruppert et al. (1995). The choice of $m^{1/3}$ is guided by its role in establishing consistency of the autoregressive spectral density estimator of a stationary time series (Berk, 1974).

Since we need a smooth $f_{j,m}$ for each $j = 1, \ldots, \lfloor m^{1/3} \rfloor$, the computational complexity is $O(m^{1/3}C_m)$, where C_m is the computational complexity for local polynomial fitting of m observations; compare (23). Fast algorithms for local polynomial fitting exist. One can achieve $C_m = O(m \log m)$ by employing the Fast Fourier Transform (Fan & Marron, 1994). For more information about computational issues see Fan & Gijbels (1996, § 3.6) and references therein.

3·3. *Order selection*

In view of the apparent equivalence between modelling a covariance matrix and a sequence of regressions, parsimonious modelling of the former suggests that a bound be placed on the increasing orders of the regression models in (1). It is instructive to think of the number of subdiagonals of T whose elements are not identically 0 as the maximum order d in some varying-order AR model. The corresponding covariance matrix is then

related to the antedependence models of order d introduced by Gabriel (1962) and Macchiavelli & Arnold (1994), and the class of time-varying parameter autoregressive models of Subba Rao (1970) and Kitagawa & Gersch (1985).

In this section, we rely on penalised likelihood criteria such as AIC and BIC for selecting d, the number of nonzero subdiagonals of T, and adopt the following two-step method. In the first step, for $d = 0, 1, \ldots, \lfloor m^{1/3} \rfloor$, we estimate the raw $\hat{\phi}_{t,t-j}$ from the data as outlined in § 2·2 and smooth them using local linear regression with an automatic bandwidth selector (Ruppert et al., 1995) to obtain $\bar{\phi}_{t,t-j}$ in (18). The estimated prediction variance $\bar{\sigma}_t^2$ for a model of order d in (1) is

$$\bar{\sigma}_t^2(d) = \begin{cases} n^{-1} \sum_{i=1}^n \{y_{t,i} - (\bar{\phi}_{t,t-1} y_{t-1,i} + \ldots + \bar{\phi}_{t,t-d} y_{t-d,i})\}^2 & (d \geqslant 1), \\ n^{-1} \sum_{i=1}^n (y_{t,i} - \bar{y}_i)^2 & (d = 0). \end{cases} \tag{20}$$

Then, viewing d as a candidate for the order of (1), we may choose it to minimise an AIC-type criterion,

$$\text{AIC}(d) = n \sum_{t=1}^{m} \log \bar{\sigma}_t^2(d) + 2d \quad (d = 0, 1, \ldots, \lfloor m^{1/3} \rfloor). \tag{21}$$

The idea behind using (21) is that we do not view $\phi_{t,t-j}$, for $t = 2, \ldots, \lfloor m^{1/3} \rfloor$ and $j = 1, \ldots, t-1$, as distinct individual parameters, but rather as realisations of $\lfloor m^{1/3} \rfloor$ distinct smooth functions. Consequently, the penalty levied by AIC for estimating each subdiagonal is 2, which is twice the number of smoothing parameters required for estimating $f_{j,m}(.)$. The AIC usually overestimates the order, and this drawback is corrected by increasing the penalty in (21) from 2 to $\log n$ which amounts to using the BIC. Further research is needed into the appropriate way of counting parameters when smoothing $\hat{T}$. The topic is closely related to the notion of equivalent degrees of freedom in the theory of nonparametric curve estimation; see Hastie & Tibshirani (1990, Ch. 2) and Dahlhaus (1997, p. 23).

It might be argued that smoothing the subdiagonals of T before selecting the order of (1) could have impact on the end result. To assess the merit of such an argument, one may first select d based on the raw estimates, by minimising

$$n \sum_{t=1}^{m} \log \hat{\sigma}_t^2(d) + 2(m - d/2)(d+1) \quad (d = 0, 1, \ldots, \lfloor m^{1/3} \rfloor), \tag{22}$$

where $\hat{\sigma}_t^2(d) = n^{-1} \sum_{i=1}^n \{y_{t,i} - (\hat{\phi}_{t,t-1} y_{t-1,i} + \ldots + \hat{\phi}_{t,t-d} y_{t-d,i})\}^2$ and $(m - d/2)(d+1)$ is the total number of parameters with d subdiagonals. Then, after selecting the optimal d, we smooth the corresponding subdiagonals of T using a local linear regression with automatic bandwidth selector as before. Our limited simulations show that the two methods produce similar results. However, to compare them more formally, one needs to rely on suitable loss functions introduced in § 5 for assessing various covariance estimators.

4. Data analysis

4·1. *Bodyweights of cows*

With our first dataset we study the impact of modelling the mean on the raw estimates of ϕ_{tj}. The data consist of bodyweights of $n = 27$ cows, measured at $m = 23$ common and unequally-spaced times over a period of about 22 months, ranging between 122 and 781 days. Following Diggle et al. (1994, pp. 100–6) we use a time-scale which runs from 0 to

66, each unit representing 10 days, and use a log transformation of the bodyweights as the response variable to stabilise the variance over time. The animals were allocated to treatments in a 2×2 factorial design. The two factors were presence/absence of iron dosing and infection by an organism.

The profile plot of the data (Diggle et al., 1994, p. 103) identifies an animal with abnormally low weight gain throughout the experiment. We conduct analysis with and without this animal, to assess qualitatively the impact of outliers on our procedure. The empirical regressograms of ordinary least-squares residuals from two saturated models for the mean responses are given in Figs 1(a), (b) and Figs 1(c), (d), respectively. The first regressogram is computed assuming the four treatments have identical mean profiles while the second allows them to be different for the four treatments. Comparing the two regressograms reveals the potential impact of mean-model misspecifications on formulating models for covariances and their estimators. The apparent huge difference in the range of variations of the ϕ_{ij}'s in Figs 1(a) and (c) shows just how different the ϕ_{ij}'s are under the two different mean models. However, a closer scrutiny of the second sample covariance matrix reveals that this 23×23 matrix of rank $27 - 4 = 23$ is ill-conditioned in the sense that its largest and smallest eigenvalue pair is $(1{\cdot}67 \times 10^{-1}, 2{\cdot}21 \times 10^{-9})$. Its counterpart for the first covariance matrix with rank $27 - 1 = 26$ is $(2{\cdot}81 \times 10^{-1}, 1{\cdot}08 \times 10^{-5})$. Excluding the last row of $\hat{T}$ reduces the range of the ϕ_{ij}'s to $(-2, 6)$ which is more reasonable than the original range of $(-20, 50)$.

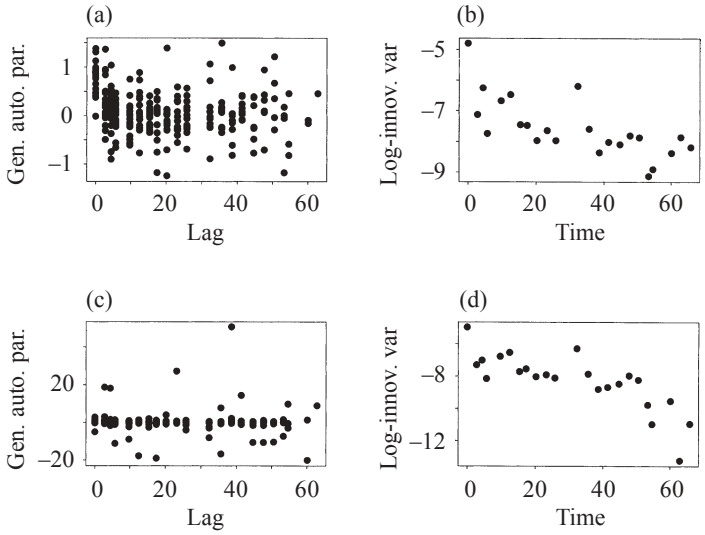

Fig. 1. Sample regressograms for the log-bodyweights of cows: (a) sample generalised autoregressive parameters for residuals from a common saturated mean, (b) sample log-innovation variances, (c) sample generalised autoregressive parameters for residuals from treatment-specific saturated mean; (d) sample log-innovation variances.

Since the ϕ_{ij}'s have different Student-t distributions and are heterogeneous, their standardised values are more appropriate for testing their significance; see (15). Figure 2 provides the standardised analogues of Figs 1(a) and (c), where now clearly the ranges of their values are comparable. As a rule of thumb, when n is large one could declare values of the standardised ϕ_{tj} falling within $(-2, 2)$ as nonsignificant at $\alpha = 0{\cdot}05$, and replace

these values by zero to obtain a parsimonious, sparse $\hat{T}$ and hence Σ. This approach to model selection for the covariance matrix is very much in line with time series model selection based on the sample correlogram and partial correlogram.

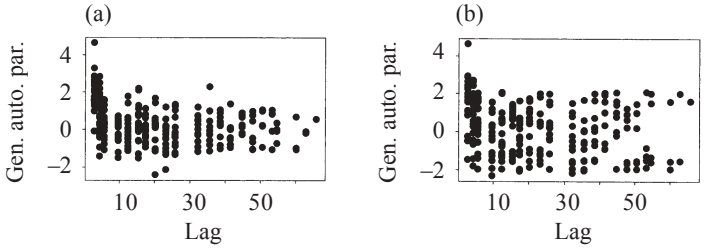

Fig. 2. Standardised sample regressograms for the log-bodyweights of cows: (a) standardised sample generalised autoregressive parameters for residuals from a common saturated mean, (b) standardised sample generalised autoregressive parameters for residuals from treatment-specific saturated mean.

Removing the animal with abnormally low weight gain leaves us with $n' = 26$ subjects and $m = 23$ repeated measures on each. When the mean profiles for the four treatments are allowed to be different and saturated, the 23×23 sample covariance matrix of the centred data is singular with rank $26 - 4 = 22$. In this case, the regression formulae (11) fail to hold for $t = 23$. The same is true of the calculations for the modified Cholesky decomposition which fail when the last row and diagonal entry of $\hat{T}$ and $\hat{D}$, respectively, are computed. Figure 3 is the analogue of Fig. 2 for the data with the lower-weight animal purged. It suggests that removing the lower-weight animal does not have much effect on the estimation of covariance other than reducing its rank. The plots analogous to Fig. 1 are also similar to those plots.

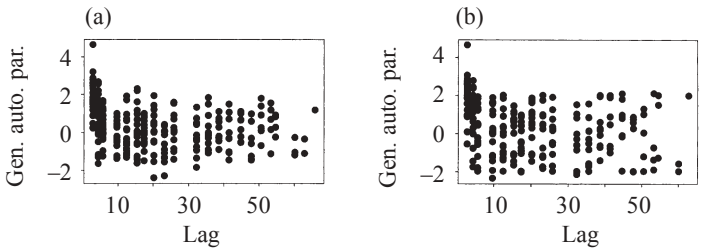

Fig. 3. Standardised sample regressograms for the log-bodyweights of cows after removing the lower-weight cow: (a) standardised sample generalised autoregressive parameters for residuals from a common saturated mean, (b) standardised sample generalised autoregressive parameters for residuals from treatment-specific saturated mean

4·2. *Cattle data*

This dataset has been studied extensively in the literature of longitudinal data analysis (Kenward, 1987; Pourahmadi, 2000; Pan & MacKenzie, 2003) to gauge the performance of new methods. Here we present a nonparametric estimator of the covariance matrix of the data for 30 animals with the treatment A. The animals were weighed eleven times over a 133-day period. Here, the data consist of $n = 30$ subjects each with $m = 11$ observations. For this balanced dataset, the ϕ_{tj} and σ_t^2 can be computed either by applying the factoris-

ation (1) to the sample covariance matrix S of the data as in Pourahmadi (1999) or by relying on the regression estimates developed in (12)–(15). Both AIC and BIC suggest smoothing the first two subdiagonals of $\hat{T}$ and replacing the rest by zero, so that $d = 2$. The correlation matrix corresponding to this nonparametric estimate $\hat{\Sigma}$ is given in Table 1. It compares favourably to the sample correlation matrix of the data and the estimated correlation matrix in Pourahmadi (1999, Table 3) obtained by fitting cubic polynomials to the ϕ_{tj} and $\log \sigma_t^2$. This lends support to the usefulness of smoothing $\hat{T}$ for m as small as 11. Our simulation study sheds further light on this important point.

Table 1: *Cattle data. Sample correlations for weights, above the main diagonal, fitted correlations, below the main diagonal, obtained by smoothing the first two subdiagonals of $\hat{T}$ and setting others to zero. Diagonal of $\hat{D}$ is not smoothed*

Time point	Time point										
	1	2	3	4	5	6	7	8	9	10	11
1	1·00	0·82	0·76	0·65	0·62	0·57	0·50	0·51	0·49	0·45	0·45
2	0·81	1·00	0·90	0·83	0·79	0·72	0·60	0·64	0·57	0·55	0·52
3	0·75	0·91	1·00	0·92	0·87	0·83	0·73	0·75	0·68	0·67	0·64
4	0·70	0·85	0·92	1·00	0·94	0·90	0·81	0·82	0·74	0·70	0·68
5	0·66	0·81	0·88	0·93	1·00	0·94	0·86	0·88	0·82	0·77	0·74
6	0·63	0·77	0·84	0·89	0·93	1·00	0·92	0·93	0·89	0·84	0·80
7	0·60	0·73	0·79	0·84	0·88	0·92	1·00	0·92	0·92	0·86	0·84
8	0·58	0·71	0·77	0·82	0·86	0·90	0·94	1·00	0·96	0·93	0·91
9	0·57	0·70	0·76	0·81	0·85	0·89	0·93	0·96	1·00	0·96	0·95
10	0·55	0·67	0·73	0·77	0·81	0·85	0·90	0·93	0·95	1·00	0·98
11	0·55	0·66	0·72	0·77	0·81	0·84	0·89	0·92	0·94	0·98	1·00

5. SIMULATIONS

We compare the performance of the sample covariance matrix S to that of a nonparametric estimator $\hat{\Sigma}$, obtained from (1) by smoothing the first few subdiagonals of $\hat{T}$ and replacing the rest by zero. The results confirm our intuition that, for large m and smooth covariance matrices, $\hat{\Sigma}$ should perform better than S for a variety of reasonable loss functions and sample sizes.

We consider the following three covariance matrices with varying level of smoothness as indicated by the smoothness of the functions associated with their (T, D) components.

Case 1: $\phi_{t,t-j} \equiv 0$, $\sigma_t^2 \equiv 1$, corresponding to the identity covariance matrix.

Case 2: $\phi_{t,t-j} = 2(t/m)^2 - 0{\cdot}5$ and $\phi_{t,t-j} \equiv 0$ $(j \geqslant 2)$, $\sigma_t = \log(t/10 + 2)$, corresponding to a varying coefficient AR(1).

Case 3: $\phi_{t,t-j} = m^{-2} \min(t + j, t^{1{\cdot}5}) e^{-j/4}$, $\sigma_t = \log(t/10 + 2)$.

For each case, we simulate n independent and identically distributed $N(0, \Sigma)$ random m-vectors and compute their sample covariance matrix S. The modified Cholesky decomposition of S is used to obtain $(\hat{T}, \hat{D})$, and these components are then smoothed, as outlined in § 3, to give a nonparametric estimate $\hat{\Sigma}$ of Σ. The above scheme is repeated $N = 100$ times, and for a given loss function $L(.,.)$ the values of $L(\Sigma, S)$, $L(\Sigma, \hat{\Sigma})$ and the corresponding risks are computed. The maximum possible order is chosen to be $\lfloor m^{1/3} \rfloor$, and we use the loss functions

$$L_1(A, B) = \operatorname{tr}(A^{-1}B) - \log|A^{-1}B| - m, \quad L_2(A, B) = \operatorname{tr}(A^{-1}B - I)^2.$$

The corresponding risk functions are defined by $E_\Sigma\{L_i(A, B)\}$ $(i = 1, 2)$. A smooth estimator $\hat{\Sigma}$ is considered to be better than the sample covariance matrix S if its risk function is smaller. For more information on simulation-based comparison of covariance estimators see Lin & Perlman (1985). The automatic bandwidth selector of Ruppert et al. (1995) is used in the simulation.

The results for L_1, presented in Table 2, show that, based on estimated risks, the smoothed estimators outperform the sample covariance matrix for every combination of (Σ, n, m). Surprisingly, this holds even for the smaller 10×10 covariances matrices where nonparametric estimation of a curve based on ten points does not sound prudent at all. The corresponding results for L_2 led to the same conclusions. However, closer examination of the individual losses revealed that, among the $N = 100$ simulation runs, in a few cases S had a smaller L_2 loss than $\hat{\Sigma}$ for $m = 10$.

Table 2: *Simulations. Risks, i.e. average losses, of the estimators S, sample, and* $\hat{\Sigma}$*, smooth, at three test matrices* Σ_i $(i = 1, 2, 3)$ *for the loss function* L_1

		Σ_1		Σ_2		Σ_3	
n	m	Sample	Smooth	Sample	Smooth	Sample	Smooth
50	10	1·225	0·246	1·209	0·274	1·171	0·274
	20	5·074	0·384	5·008	0·439	5·111	0·411
	30	12·524	0·399	12·563	0·462	12·446	0·389
	40	25·638	0·448	25·317	0·506	25·812	0·472
100	10	0·588	0·118	0·599	0·159	0·578	0·154
	20	2·264	0·178	2·254	0·213	2·325	0·206
	30	5·266	0·194	5·271	0·242	5·283	0·196
	40	9·656	0·216	9·626	0·271	9·712	0·202

As a check on the accuracy of the simulated results for the L_1 loss, the estimated risks of S in Table 2 can be compared to their exact values computed using (4.1) in Lin & Perlman (1985). For sample sizes $n = 50$ and 100, and matrices of dimensions $m = 10$, 20, 30, 40, these exact values are, for $n = 50$, 1·185, 4·929, 12·130, 24·703; and, for $n = 100$, 0·570, 2·260, 5·211, 9·605. These agree well with the estimated L_1 risks given in the first column of Table 2.

6. Asymptotic results

In this section an asymptotic error bound is derived for the mean squared error of the smoothed estimators $\bar{\phi}_{t,t-j}$. We list the following technical conditions, some of which are from Fan & Zhang (2000), and adopt their methods of proof to our setting.

Condition 1. For some $p \geqslant 0$, $f_{j,m}(.)$ and $\sigma_m(.)$ are $p + 1$ times continuously differentiable.

Condition 2. The kernel K is a bounded probability density function with a bounded support.

Condition 3. The n random vectors $y_i = (y_{k,i}, 1 \leqslant k \leqslant m)$ are independent and identically distributed for $1 \leqslant i \leqslant n$.

Condition 4. All eigenvalues of $\Omega_t = E\{(y_1, \ldots, y_t)'(y_1, \ldots, y_t)\}$, for $1 \leqslant t \leqslant m$, are larger than $\lambda_0 > 0$.

Condition 5. The diagonal entries of D in (1) are uniformly bounded: $\max_i \sigma_i \leqslant C < \infty$.

Condition 1 imposes a certain degree of smoothness on the coefficients $\phi_{t,t-j}$, where p will play the role of the order of local polynomials. Condition 2 is standard. Condition 4 imposes a uniform positive definiteness on Ω_t.

Let the raw estimator $\hat{f}_{j,m}(l/m) = \hat{\phi}_{l,l-j}$ $(l = j+1, \ldots, m)$ be the $(l-j)$th element of the vector $\hat{\phi}_l$ in (11). We briefly describe how to obtain the local polynomial estimator $\bar{f}_{j,m}(l/m) = \bar{\phi}_{l,l-j}$. Write

$$C_l(t) = (1, l-t, \ldots, (l-t)^p)' \quad (l = j+1, \ldots, m),$$

$$C(t) = (C_{j+1}(t), \ldots, C_m(t))', \quad W(t) = \mathrm{diag}\{W_{j+1}(t), \ldots, W_m(t)\},$$

where $W_l(t) = K\{(l-t)/h\}/h$ and the bandwidth h satisfies $h/m \to 0$ and $h \to \infty$. Thus the bandwidth for $\hat{f}_{j,m}$ has the form $b = h/m$ and it satisfies $b \to 0$ and $bm \to \infty$. To obtain the local pth-order polynomial estimator for $\phi_{t,t-j}$ as a function of t when j is fixed, we minimise

$$\sum_{l=j+1}^{m} \{\hat{\phi}_{l,l-j} - \beta C_l(t)\}^2 K\{(l-t)/h\}$$

over $\beta = (\beta_0, \ldots, \beta_p) \in R^{p+1}$. Then the estimator $\bar{\phi}_{t,t-j}$ has the form

$$\bar{\phi}_{t,t-j} = \sum_{l=j+1}^{m} w_{0,p+1}(l, t)\hat{\phi}_{l,l-j}, \tag{23}$$

where $w_{0,p+1}(l, t) = e'_{1,p+1}\{C'(t)W(t)C(t)\}^{-1}C_l(t)W_l(t)$ for $l = j+1, \ldots, m$.

Note that, for $t = 2, \ldots, m$, $\hat{\phi}_t - \phi_t = (X_t'X_t)^{-1}X_t'\varepsilon_t$ form martingale differences with respect to the filter generated by the past $\{\varepsilon_t\}$: $\mathscr{F}_t = \sigma(\varepsilon_1, \ldots, \varepsilon_t)$. Thus $E(\hat{\phi}_t) = \phi_t$ and $E(\hat{\phi}_{l,l-j}) = f_{j,m}(l/m)$. Therefore, under Condition 1, the bias of the smoothed estimator $\bar{\phi}_{t,t-j}$ is

$$E(\bar{\phi}_{t,t-j}) - \phi_{t,t-j} = \sum_{i=j+1}^{m} w_{0,p+1}(i, t)f_{j,m}(i/m) - f_{j,m}(t/m) = O\{(h/m)^{p+1}\}. \tag{24}$$

Next we derive the variance of $\bar{\phi}_{t,t-j}$. Since $\hat{\phi}_{i,i-j} - \phi_{i,i-j}$ $(i = j+1, \ldots, m)$ are orthogonal,

$$\mathrm{var}\,(\bar{\phi}_{t,t-j}) = \sum_{i=j+1}^{m} w_{0,p+1}^2(i, t)E(\hat{\phi}_{i,i-j} - \phi_{i,i-j})^2. \tag{25}$$

By conditioning, since $E(\varepsilon_t\varepsilon_t'|\mathscr{F}_{t-1}) = \sigma_t^2 I_n$, we have

$$\begin{aligned} E\{(\hat{\phi}_t - \phi_t)(\hat{\phi}_t - \phi_t)'\} &= E[E\{(X_t'X_t)^{-1}X_t'\varepsilon_t\varepsilon_t'X_t(X_t'X_t)^{-1}|\mathscr{F}_{t-1}\}] \\ &= E\{(X_t'X_t)^{-1}X_t'E(\varepsilon_t\varepsilon_t'|\mathscr{F}_{t-1})X_t(X_t'X_t)^{-1}\} = \sigma_t^2 E\{(X_t'X_t)^{-1}\}. \end{aligned} \tag{26}$$

By the law of large numbers, for fixed i, $(X_i'X_i)/n \to \Omega_{i-1}$ almost surely, or $n(X_i'X_i)^{-1} \to \Omega_{i-1}^{-1}$ almost surely. By Condition 4, all eigenvalues of Ω_{i-1}^{-1} are bounded by $1/\lambda_0$. Hence, by Condition 5, all diagonal entries of $E\{(\hat{\phi}_t - \phi_t)(\hat{\phi}_t - \phi_t)'\}$ are bounded by $\sigma_t^2/(n\lambda_0) \leqslant C^2/(n\lambda_0)$. Note that the $(i-j, i-j)$th element of the matrix $\sigma_t^2 E\{(X_i'X_i)^{-1}\}$ is $E(\hat{\phi}_{i,i-j} - \phi_{i,i-j})^2$, which is also bounded by $C^2/(n\lambda_0)$. Thus, (25) and (26) yield

$$\mathrm{var}(\bar{\phi}_{t,t-j}) \leqslant \frac{C^2}{n\lambda_0} \sum_{l=j+1}^{m} w_{0,p+1}^2(l, t) = O(n^{-1})O(h^{-1}) \tag{27}$$

by (A.4) in Fan & Zhang (2000). Therefore, (27) together with (24) results in

$$E(\bar{\phi}_{t,t-j} - \phi_{t,t-j})^2 = \mathrm{var}(\bar{\phi}_{t,t-j}) + \{E(\bar{\phi}_{t,t-j}) - \phi_{t,t-j}\}^2 = O\{(nh)^{-1} + (h/m)^{2p+2}\}, \tag{28}$$

which has order $O\{(nm)^{-(2p+2)/(2p+3)}\}$ if we let $h = m(nm)^{-1/(2p+3)}$. The optimal bandwidth is therefore of the form $h = Cm(nm)^{-1/(2p+3)}$ for some $C > 0$ which may depend on the kernel K and functions $f_{j,m}(.)$. In summary, we have the following theorem.

THEOREM 1. *Assume that Conditions* 1–5 *hold and that* $m^{2+2p}/n \to \infty$. *Then the optimal bandwidth* $h = Cm(nm)^{-1/(2p+3)}$ *for some* $C > 0$ *and*

$$E(\bar{\phi}_{t,t-j} - \phi_{t,t-j})^2 = O\{(nm)^{-(2p+2)/(2p+3)}\}.$$

In particular, the local linear estimator gives the L^2 error bound $(nm)^{-4/5}$. We require that the bandwidth $h = m(nm)^{-1/(2p+3)} \to \infty$, or $m^{2p+2}/n \to \infty$. This restriction reflects the main focus of the paper, which is to model a covariance matrix with high dimensions.

In the classical setting, the errors in the local polynomial regression are assumed to be independent. There is some discussion of extensions to dependent errors; see for example Opsomer et al. (2001) and Hart (1991). Fortunately in our framework, since $\hat{\phi}_t - \phi_t = (X_t'X_t)^{-1}X_t'\varepsilon_t$ are martingale differences, the problem of bandwidth selection based on mean squared error can be reduced to the classical independent case.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the helpful and constructive comments provided by the anonymous referee.

REFERENCES

BERK, K. N. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2**, 489–502.
DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **36**, 1–37.
DANIELS, M. J. & KASS, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1173–84.
DANIELS, M. J. & POURAHMADI, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89**, 553–66.
DIGGLE, P. J., LIANG, K. Y. & ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.
DIGGLE, P. J. & VERBYLA, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54**, 401–15.
FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.
FAN, J. & MARRON, J. S. (1994). Fast implementations of nonparametric curve estimators. *J. Comp. Graph. Statist.* **3**, 35–56.
FAN, J. & ZHANG, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc.* B **62**, 303–22.
GABRIEL, K. R. (1962). Antedependence analysis of an ordered set of variables. *Ann. Math. Statist.* **33**, 201–12.
GLASBEY, C. A. (1988). Standard errors resilient to error variance misspecification. *Biometrika* **75**, 201–6.
HALL, P. & PATIL, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Prob. Theory Rel. Fields* **99**, 399–424.
HALL, P., FISHER, N. I. & HOFFMANN, B. (1994). On the nonparametric estimation of covariance functions. *Ann. Statist.* **22**, 2115–34.
HART, J. D. (1991). Kernel regression estimation with time-series errors. *J. R. Statist. Soc.* B **53**, 173–87.
HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
HUANG, J. Z., WU, C. & ZHOU, L. (2002). Varying-coefficient models and basis function approximations for analysis of repeated measurements. *Biometrika* **89**, 111–28.
KENWARD, M. G. (1987). A method for comparing profiles of repeated measurements. *Appl. Statist.* **36**, 296–308.
KITAGAWA, G. & GERSCH, W. (1985). A smoothness priors time varying AR coefficients modeling of nonstationary time series. *IEEE Trans. Auto. Contr.* **30**, 48–56.
KITAGAWA, G. & GERSCH, W. (1996). *Smoothness Priors Analysis of Time Series*. New York: Springer-Verlag.
KRZANOWSKI, W. J., JONATHAN, P., MCCARTHY, W. V. & THOMAS, M. R. (1995). Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Appl. Statist.* **44**, 101–15.

LIN, S. P. & PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis,* **6**, Ed. P. R. Krishnaiah, pp. 411–29. Amsterdam: North-Holland.

MACCHIAVELLI, R. E. & ARNOLD, S. F. (1994). Variable order antedependence models. *Commun. Statist*. A **23**, 13–22.

OPSOMER, J., WANG, Y. D. & YANG, Y. H. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16**, 134–53.

PAN, J. & MACKENZIE, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika* **90**, 239–44.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–90.

POURAHMADI, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–35.

POURAHMADI, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. New York: Wiley-Interscience.

RUPPERT, D., SHEATHER, S. J. & WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–70.

SAMPSON, P. D. & GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Assoc.* **87**, 108–19.

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

SUBBA RAO, T. (1970). The fitting of nonstationary time series models with time dependent parameters. *J. R. Statist. Soc.* B **32**, 312–22.

[*Received June* 2002. *Revised January* 2003]