Maximum likelihood estimation of models for residual covariance in spatial regression

By K. V. MARDIA and R. J. MARSHALL

Department of Statistics, University of Leeds, U.K.

Summary

We describe the maximum likelihood method for fitting the linear model when residuals are correlated and when the covariance among the residuals is determined by a parametric model containing unknown parameters. Observations are assumed to be Gaussian. We give conditions which ensure consistency and asymptotic normality of the estimators. Our main concern is with the analysis of spatial data and in this context we describe some simulation experiments to assess the small sample behaviour of estimators. We also discuss an application of the spectral approximation to the likelihood for processes on a lattice.

Some key words: Asymptotic normality; Consistency; Gaussian process; Kriging; Lattice process; Likelihood; Simulation; Spatial covariance.

1. Introduction

In regression analysis the assumption that residuals are uncorrelated is sometimes untenable. This is particularly so in the analysis of spatial data when correlation may exist between neighbouring entities. The nature of the covariance among residuals will usually not be known precisely, but it is often possible to adopt a simple parametric model to describe it. One then has a set of covariance parameters as well as the regression coefficients to estimate. In most applications estimation of the regression coefficients will be of primary importance; however, our interest in this problem arose in the estimation of models of spatial covariance where the nature of the covariance is of particular interest in itself (Mardia, 1980). The approach we adopt, through maximum likelihood, is akin to the earlier studies of Cliff & Ord (1973, Chapter 5), and more recently Cook & Pocock (1983).

In § 2 we discuss how the likelihood function may be numerically maximized, giving suitable formulae for the derivatives and information matrix. In § 3 we give conditions which ensure that, with dependent observations, estimators are consistent and uniformly asymptotically normal. The relevance of these conditions for the analysis of spatial data is discussed in §§ 4 and 5. The results of some Monte Carlo simulation experiments for the estimation of spatial covariance is presented in § 5. These provide a useful insight into the nature of bias for small samples. In § 6 we discuss the computationally straightforward large sample spectral approximation to the likelihood for processes observed on a lattice and we give some illustrative examples.

2. Maximum likelihood

2·1. *The likelihood for a Gaussian process*

We consider a real valued Gaussian random process $\{Y(t);\ t \in T\}$ where T is an index set. For example $T = Z^d$ describes a d-dimensional lattice process and $T = R^d$ is a continuous parameter process. Or T could be a collection of spatial entities such as regions or counties. We suppose that, for all $t \in T$, $E\{Y(t)\} = x(t)'\beta$, where $x(t) = \{x_1(t), \ldots, x_q(t)\}'$ is a $q \times 1$ vector of nonrandom regressors and $\beta \in B$ is a parameter vector, B being an open subset of R^q. Also let the covariance be defined by a parametric model $\operatorname{cov}\{Y(t), Y(s)\} = \sigma(t, s; \theta)$, for all $t, s \in T$, where $\theta \in \Theta$ is a $p \times 1$ parameter vector, Θ being an open subset of R^p. We assume that $\sigma(t, s; \theta)$ is twice differentiable with respect to θ at all points on $T^2 \times \Theta$, and that it is positive-definite in the sense that for every finite subset $T_n = \{t_1, \ldots, t_n\}$ of T the covariance matrix $V_n = \{\sigma(t_i, t_j; \theta)\}$ is positive-definite.

Suppose that $Y(t)$ is observed at each point to give the sample vector $Y_n = \{Y(t_1), \ldots, Y(t_n)\}'$. We denote the combined $(q+p) \times 1$ parameter vector by $\phi = (\beta', \theta')'$. The log likelihood for ϕ is, if we ignore a constant,

$$L_n(\phi; Y_n) = -\tfrac{1}{2}\log|V_n| - \tfrac{1}{2}(Y_n - X_n\beta)' V_n^{-1}(Y_n - X_n\beta), \qquad (2{\cdot}1)$$

where X_n is an $n \times q$ regressor matrix with jth column $x_j = \{x_j(t_1), \ldots, x_j(t_n)\}'$. We assume X_n to be rank q. By maximizing (2·1) the maximum likelihood estimates of β and θ, denoted $\hat{\beta}_n, \hat{\theta}_n$, can be obtained. There are a number of ways this can be done as we indicate in §2·2. First, however, we give some relevant formulae. For notational convenience, we shall henceforth omit the subscript n on X_n, V_n and Y_n. The derivative vector of L_n, say $L_n^{(1)}$, can be written $L_n^{(1)} = (L_\beta', L_\theta')'$, where $L_\beta = -X'V^{-1}X\beta + X'V^{-1}Y$ and the ith element of L_θ is

$$(L_\theta)_i = -\tfrac{1}{2}\operatorname{tr}(V^{-1}V_i) - \tfrac{1}{2}W'V^iW \quad (i = 1, \ldots, p), \qquad (2{\cdot}2)$$

where $V_i = \partial V/\partial\theta_i$, $V^i = \partial V^{-1}/\partial\theta_i = -V^{-1}V_iV^{-1}$ and $W = Y - X\beta$. The second derivative matrix of L_n can be written

$$L_n^{(2)} = \begin{bmatrix} L_{\beta\beta} & L_{\beta\theta} \\ L_{\beta\theta}' & L_{\theta\theta} \end{bmatrix}, \qquad (2{\cdot}3)$$

where $L_{\beta\beta} = -X'V^{-1}X$, $L_{\beta\theta}$ has ith column $-X'V^iX\beta + X'V^iY$ for $i = 1, \ldots, p$, and $L_{\theta\theta}$ has (i,j)th term

$$-\tfrac{1}{2}\{\operatorname{tr}(V^{-1}V_{ij} + V^iV_j) + W'V^{ij}W\}, \qquad (2{\cdot}4)$$

where

$$V_{ij} = \partial^2 V/\partial\theta_i\partial\theta_j, \quad V^{ij} = \partial^2 V^{-1}/\partial\theta_i\partial\theta_j.$$

Using $V^{ij} = V^{-1}(V_iV^{-1}V_j + V_jV^{-1}V_i - V_{ij})V^{-1}$ we obtain the information matrix

$$B_n = -E(L_n^{(2)}) = \operatorname{diag}(B_\beta, B_\theta), \qquad (2{\cdot}5)$$

where $B_\beta = X'V^{-1}X$ and the (i, j)th element of B_θ is $\tfrac{1}{2}t_{ij}$ with

$$t_{ij} = \operatorname{tr}(V^{-1}V_iV^{-1}V_j) = \operatorname{tr}(VV^iVV^j).$$

2·2. *Computational aspects*

One procedure to maximize L_n is by scoring using the updating equation $\phi_1 = \phi_0 + B_n^{-1}L_n^{(1)}$. In view of (2·5) this amounts to computing β_0, for a given θ_0, using

$$\beta_0 = (X'V^{-1}X)^{-1}X'V^{-1}Y \qquad (2{\cdot}6)$$

and then updating θ using

$$\theta_1 = \theta_0 + B_\theta^{-1} L_\theta, \tag{2·7}$$

where L_θ is evaluated from (2·2). Then β is updated using (2·6) and so on. This method is straightforward and we have found that it is quite efficient. It can be made more robust by incorporating a Levenberg–Marquardt parameter $\delta > 0$, that is by replacing B_θ by $B_\theta + \delta\,\mathrm{diag}\,(B_\theta)$ in (2·7) to ensure an increase in the likelihood at each step. There are, of course, other ways this maximization could be achieved. One could take $-L_{\theta\theta}$ as the Hessian in (2·7) but the matrix multiplication in its evaluation, using (2·4) is prohibitive. A quasi-Newton method of maximization could also be used and it may be that such an approach will be more efficient and reliable, though scoring has an advantage that it gives an estimate of B_n^{-1}. In a limited number of trials, we have found that a quasi-Newton method has given no significant reduction in the number of function evaluations required.

Another procedure, suggested by Cochrane & Orcutt (1949) and Cliff & Ord (1973, Chapter 5), is to maximize L_n with respect to θ for a given β_0 and then to update β using (2·6). This idea can be introduced here by allowing the updating of θ, using (2·7), to continue until an approximate conditional maximum is found, after which β is updated using (2·6). In our experience this is not so efficient as a one-step alternation between (2·6) and (2·7); the extra iterations in finding the conditional maximum do not seem to speed the overall rate of convergence.

The residuals after ordinary least squares estimation of β may, in some cases, be used to suggest an initial θ. If, for example, $Y(t)$ is a covariance stationary process, empirical covariances can be computed; see, for example, Ripley (1981, p. 57). These can be used to indicate θ initially. Unless there are *a priori* grounds for a particular covariance model a tentative model identification can sometimes be made at this stage.

The maximum likelihood procedure is computationally restrictive since V^{-1} has to be evaluated. Unless V is structured in a way that can be exploited, computational time is proportional to n^3 and the approach is impractical for large n. We have satisfactorily implemented the procedure for n up to 150 on the Leeds University Amdahl V7 computer. If n is prohibitively large one can partition T_n into subsets, obtain the maximum likelihood estimates for each and pool the results. There may be scope for work along these lines. Schagan (1980) has considered using subsets of size two.

When T_n is a d-dimensional rectangular lattice and $Y(t)$ a covariance stationary process it is possible to approximate the likelihood; we outline this procedure in §6. However, for $d = 1$ exact inversion of V, which is a Toeplitz matrix, is rapidly achieved using an algorithm by Trench (1964) which fully exploits the persymmetry property of the inverse. In two dimensions it is also possible to set up an algorithm for the inversion of V: suppose points of an $M \times N$ lattice are numbered column by column and let $V = V_N^M$. Then $V_N^M = ((C_{i-j}))$, where C_k is an $M \times M$ covariance matrix between columns separated by lag k. Hence, we can write

$$V_N^M = \begin{bmatrix} C_0 & R_{N-1} \\ R'_{N-1} & V_{N-1}^M \end{bmatrix}, \tag{2·8}$$

where $R_{N-1} = (C_1, \ldots, C_{N-1})$. If we use (2·8), it is straightforward to obtain a recursive formula for $(V_N^M)^{-1}$ in terms of $(V_{N-1}^M)^{-1}$, which requires the inversion of an $M \times M$ matrix at each iteration, and computation time is thus proportional to $N^2 M^3$.

3. CONSISTENCY AND ASYMPTOTIC NORMALITY

Since Y is a single observation from $N_n(X\beta, V)$, it is not obvious that $\hat{\theta}_n$ and $\hat{\beta}_n$ are consistent and asymptotically normal. Sweeting (1980) has given a general result concerning weak consistency and uniform asymptotic normality of maximum likelihood estimators which is applicable here. We require continuity, growth and convergence of the information $L_n^{(2)}$. Our interest centres on conditions on V and X, given T_n. First the continuity requirement is satisfied by the assumption that $\sigma(.,.;\theta)$ be twice differentiable on Θ with continuous second derivatives. Secondly, growth in information is guaranteed if the smallest latent root of B_n tends to ∞ as $n \to \infty$, or equivalently using (2·5), if, as $n \to \infty$,

$$\lim B_\theta^{-1} = 0, \quad \lim B_\beta^{-1} = 0. \tag{3·1}$$

Thirdly, convergence of $L_n^{(2)}$ is ensured if $B_n^{-\frac{1}{2}} L_n^{(2)} B_n^{-\frac{1}{2}}$ converges in probability to a unit matrix. From (2·3) and since $L_{\beta\beta}$ is nonrandom this requires that $A_1 = B_\theta^{-\frac{1}{2}} L_{\theta\theta} B_\theta^{-\frac{1}{2}}$ and $A_2 = B_\beta^{-\frac{1}{2}} L_{\beta\theta} B_\theta^{-\frac{1}{2}}$ converge in probability to a unit and zero matrix respectively. Since mean square convergence implies convergence in probability, it is sufficient that

$$\lim E(\| A_1 - I \|^2) = 0, \quad \lim E(\| A_2 \|^2) = 0,$$

where $\| . \|$ denotes Euclidean matrix norm. Now,

$$E(\| A_1 - I \|^2) = E\{\operatorname{tr}(B_\theta^{-1} L_{\theta\theta} B_\theta^{-1} L_{\theta\theta} - I)\} = \Sigma_{ij} \operatorname{cov}\{(B_\theta^{-1} L_{\theta\theta})_{ij}, (B_\theta^{-1} L_{\theta\theta})_{jl}\}, \tag{3·2}$$

since $E(B_\theta^{-1} L_{\theta\theta}) = I$, so that we require, as $n \to \infty$,

$$\lim \sum_{i,j,k,l=1}^{p} b_\theta^{kl} b_\theta^{lj} \operatorname{tr}(V V^{kj} V V^{li}) = 0, \tag{3·3}$$

where $b_\theta^{kl} = (B_\theta^{-1})_{kl}$ and the trace in (3·3) is $\operatorname{cov}\{(L_{\theta\theta})_{kj}, (L_{\theta\theta})_{li}\}$. Similarly for $\lim E(\| A_2 \|^2) = 0$ we require, as $n \to \infty$,

$$\lim \sum_{i,k=1}^{p} \sum_{j,l=1}^{q} b_\theta^{ki} b_\beta^{lj} (x_j' V^k V V^i x_l) = 0, \tag{3·4}$$

where the bracketed term is $\operatorname{cov}\{(L_{\beta\theta})_{kj}, (L_{\beta\theta})_{li}\}$. We have therefore established the following theorem.

THEOREM 1. *The conditions* (3·1), (3·3), (3·4) *are sufficient for the asymptotic normality and weak consistency of* $\hat{\phi}_n$; *that is* $\hat{\phi}_n \sim N(\phi, B_n^{-1})$.

The conditions of Theorem 1 can sometimes be checked directly, for example, if V is diagonal and $Y(t_i)$ $(i = 1, \ldots, n)$ are heteroscedastic. However, the following theorem gives simpler sufficient conditions which we demonstrate in §4 to be pertinent to spatial sampling.

Let $\lambda_1 \leqslant \ldots \leqslant \lambda_n$ be the eigenvalues of V and let those of V_i and V_{ij} be λ_k^i and λ_k^{ij} for $k = 1, \ldots, n$ respectively, with $|\lambda_1^i| \leqslant \ldots \leqslant |\lambda_n^i|$ and $|\lambda_1^{ij}| \leqslant \ldots \leqslant |\lambda_n^{ij}|$ for $i, j = 1, \ldots, p$.

THEOREM 2. *Suppose that as* $n \to \infty$:

(i) $\lim \lambda_n = C < \infty$, $\lim |\lambda_n^i| = C_i < \infty$, $\lim |\lambda_n^{ij}| = C_{ij} < \infty$ *for all* $i, j = 1, \ldots, p$;

(ii) $\| V_i \|^{-2} = O(n^{-\frac{1}{2}-\delta})$, *for some* $\delta > 0$ *for* $i = 1, \ldots, p$;

(iii) *for all* $i, j = 1, \ldots, p$, $a_{ij} = \lim\{t_{ij}/(t_{ii}t_{jj})^{\frac{1}{2}}\}$ *exists, where* $t_{ij} = \operatorname{tr}(V^{-1} V_i V^{-1} V_j)$ *and* $A = ((a_{ij}))$ *is a nonsingular matrix;*

(iv) $\lim (X'X)^{-1} = 0$.

Then the conditions of Theorem 1 *hold.*

Proof. Consider the first part of (3·1). We have $V^{-1}-\lambda_n^{-1} I > 0$ so that $X'V^{-1}X > \lambda_n^{-1} X'X$. Inversion gives $(X'V^{-1}X)^{-1} < \lambda_n(X'X)^{-1}$, so that, by (i) and (iv), $\lim (X'V^{-1}X)^{-1} = 0$.

Next, we have $B_\theta^{-1} = 2D^{-1}A_n^{-1}D^{-1}$, where $A_n = ((t_{ij}/(t_{ii}t_{jj})^{\frac{1}{2}}))$ and

$$D = \operatorname{diag}(t_{11}^{\frac{1}{2}}, \ldots, t_{pp}^{\frac{1}{2}}).$$

Hence $b_\theta^{ij} = 2t_{ii}^{-\frac{1}{2}} t_{jj}^{-\frac{1}{2}} a_n^{ij}$, where a_n^{ij} is the (i,j) element of A_n^{-1}. Now

$$t_{ii} = \operatorname{tr}(V^{-1}V_i V^{-1} V_i) \geqslant \lambda_n^{-2} \| V_i \|^2.$$

Hence, by (i), (ii) and (iii), $b_\theta^{ij} = O(n^{-\frac{1}{2}-\delta})$, so that $\lim (B_\theta^{-1}) = 0$.

Thirdly, let Q_1 denote the sum in (3·3). Since $V_n > 0$ we can write

$$\operatorname{tr}(V V^{kj} V V^{li}) = \operatorname{tr}(S_{kj}S_{li}),$$

where $S_{kj} = V^{\frac{1}{2}} V^{kj} V^{\frac{1}{2}}$. Hence

$$|Q_1| \leqslant \Sigma |b_\theta^{ki}| |b_\theta^{lj}| \{\operatorname{tr}(S_{kj}^2)\operatorname{tr}(S_{li}^2)\}^{\frac{1}{2}}. \tag{3·5}$$

Now $\operatorname{tr}(S_{kj}^2) \leqslant n \| S_{kj} \|_s^2$, where $\| . \|_s$ denotes spectral norm. Also, since

$$V^{kj} = V^{-1}(V_k V^{-1} V_j + V_j V^{-1} V_k - V_{kj}) V^{-1},$$

we have that

$$\| S_{kj} \|_s \leqslant 2 \| V_k \|_s \| V_j \|_s \| V^{-\frac{1}{2}} \|_s^4 + \| V^{-\frac{1}{2}} \|_s^2 \| V_{kj} \|_s,$$

which is bounded above as $n \to \infty$ using (i). Thus, since, as shown above, $b^{kj} = O(n^{-\frac{1}{2}-\delta})$, each term in (3·5) is $O(n^{-2\delta})$, so that (3·5) converges to zero.

Finally taking (3·4) we have

$$\Sigma_{k,i}\Sigma_{l,j} b_\theta^{ki} b_\beta^{lj} x_j' V^k V V^i x_l = \Sigma_{l,j} b_\beta^{lj} x_j' H x_l = \operatorname{tr}\{(X'V^{-1}X)^{-1}(X'HX)\}, \tag{3·6}$$

where $H = \Sigma_{k,i} b^{ki} V^k V V^i$. Now $\operatorname{tr}\{(X'V^{-1}X)^{-1}(X'HX)\}$ is bounded above by q times the spectral radius of $V_n H$, that is by $q \| V_n H \|_s$. Also,

$$\| VH \|_s^2 = \| V^{-\frac{1}{2}}(\Sigma b_\theta^{ki} V_k V^{-1} V_i) V^{-\frac{1}{2}} \|_s \leqslant \Sigma |b_\theta^{kl}| \| V^{-\frac{1}{2}} \|_s^4 \| V_k \|_s \| V_l \|_s$$

if we use $V^i = -V^{-1} V_i V^{-1}$. Hence (3·6) converges to zero, since $b_\theta^{kl} = O(n^{-\frac{1}{2}-\delta})$ and using (i). This completes the proof.

4. Application of Theorem 1 for sampling on R^d

Consider sampling a continuous parameter spatial process $\{Y(t), t \in R^d\}$ to estimate a spatial covariance and a possible spatial trend. We assume that the sample set, T_n, is predetermined and nonrandom with the restriction $\| t_r - t_s \| \geqslant \gamma > 0$, for all pairs $r, s = 1, \ldots, n$, to ensure that the sampling domain increases in extent as n increases, although how it increases remains, for the moment, unspecified. We shall not discuss the problem of estimation in a bounded region $C \subset R^d$, with sampling increasingly dense in C.

We are interested in establishing easily applicable conditions on σ to ensure that Theorem 2 holds. Condition (iv) does not involve σ and we shall assume that X is suitably chosen for it to hold, thus confining our comments to (i)–(iii) which involve only

σ and T_n. Essentially (iii) ensures that B_θ^{-1} is nonsingular in the limit, that is that elements of $\hat{\theta}_n$ are not asymptotically linearly dependent, whilst (ii) guarantees that information on θ accrues given that V is bounded in the sense of (i). One important case is when $Y(t)$ is a covariance stationary process in R^d with $\sigma(t, t+h;\ \theta) = \sigma(h;\ \theta)$, and when T_n represents an $n_1 \times \ldots \times n_d$ regular lattice, not necessarily rectangular, with fixed spacings. Let σ_k, for $k \in Z^d$, denote the covariance for lag $k = (k_1, \ldots, k_d)$ of the lattice. From matrix norm properties, the spectral radius of V, that is λ_n, is no larger than the row sum norm of V. Now, as $n_1, \ldots, n_d \rightarrow \infty$, the row sum norm of V is the sum of $|\sigma_k|$ over $k \in Z^d$. Hence, in condition (i), $\lim \lambda_n < \infty$ is ensured if σ_k is absolutely summable over Z^d. Similarly the remaining parts of (i) hold if $\sigma_{k,i} = \partial\sigma_k/\partial\theta_i$, and $\sigma_{k,ij} = \partial^2\sigma_k/\partial\theta_i\partial\theta_j$ are absolutely summable over Z^d. Consider now condition (ii). In view of the structure of V we have

$$\| V_i \|^2 = \Sigma_{k_1}(n_1-k_1)\,\Sigma_{k_2}(n_2-k_2)\ldots\Sigma_{k_d}(n_d-k_d)\,\sigma^2_{k,i},$$

so that, if $\sigma^2_{k,i}$ is summable, repeated application of Kronecker's lemma gives $\lim n^{-1} \| V_i \|^2 = \Sigma\,\sigma^2_{k,i}$. Hence if $\sigma^2_{k,i}$ is summable, which is implied by the summability of $|\sigma_{k,i}|$, condition (ii) holds with $\delta = \frac{1}{2}$. In view of this discussion we have the following theorem.

THEOREM 3. *For a covariance stationary Gaussian process sampled on a regular lattice, and subject to conditions* (iii) *and* (iv) *of Theorem* 2, $\hat{\phi}_n$ *is consistent and asymptotically normal if* σ_k, $\sigma_{k,i}$ *and* $\sigma_{k,ij}$ *are absolutely summable for all* $i, j = 1, \ldots, p$.

Consider now the more general case in which $Y(t)$ is not necessarily covariance stationary and sample points may be irregularly spaced, subject to $\| t_r - t_s \| \geqslant \gamma > 0$. Arguing as above, we require that rows of V, V_i and V_{ij} be absolutely summable to ensure (i) of Theorem 2. However, while this requirement may hold, that of condition (ii) could fail. Suppose, for example, that θ is a range or cut-off point so that $\sigma(t_r, t_s;\ \theta) = 0$ for $\| t_r - t_s \| > \theta$, and that $\sigma(t_r, t_r;\ \theta)$ does not depend on θ, for all $r, s = 1, \ldots, n$. Then if sampling becomes increasingly widely spaced, that is, if there exists an $N < \infty$ such that, for $n > N$, $\| t_n - t_r \| > \theta$ for $r = 1, \ldots, n-1$, then $\| V_i \|^2$ is finite as $n \rightarrow \infty$.

Often, in computing empirical covariances for second-order stationary processes, there is an apparent discontinuity at the origin, suggesting superimposed noise, or, for a geostatistical interpretation, suggesting small-scale discontinuities in $Y(t)$, the so-called nuggett effect (Matheron, 1971; Ripley, 1981, p. 50). Whatever the interpretation, a better representation of the covariance can sometimes be obtained by writing $\theta = (\theta^*, \sigma_2^2)$, $\sigma_2^2 > 0$, and $V_n = V_n^* + \sigma_2^2 I_n$, where V_n^* contains unknown parameters θ^*. If Theorem 2 holds for θ^* it evidently also holds for θ, except in the degenerate case $V_n^* = \theta^* I_n$ for which (iii) fails. We discuss further aspects to the estimation of σ_2^2 in §5.

It is of interest to note that we can construct a predictor of $Y(t)$, for $t \in T - T_n$, when ϕ is unknown. Let us write

$$\operatorname{cov}\{Y(t), Y\} = \{\sigma(t, t_1;\ \theta), \ldots, \sigma(t, t_n;\ \theta)\} = \{\sigma_t(\theta)\}',$$

say. Then the conditional mean of $Y(t)$ given Y is

$$\{x(t)\}'\beta + \{\sigma_t(\theta)\}' V^{-1}(Y - X\beta).$$

This provides the standard predictor of $Y(t)$ when ϕ is known, and it becomes the well-known universal kriging predictor when β is replaced by the generalized least squares estimator, (2·6) with θ assumed known, although in practice θ is usually estimated by *ad hoc* procedures. When both β and θ are replaced by $\hat{\beta}$ and $\hat{\theta}$ respectively, we obtain its maximum likelihood estimator which may be termed a unified universal kriging predictor. Its properties will be discussed elsewhere.

5. Simulation experiments

In practice the approach suggested in §2 will be of most use for small to moderate n. It is therefore of interest to investigate the validity of the asymptotic results for n not too large. We now outline some Monte Carlo investigations of this problem for isotropic covariance stationary process on $T = R^2$. We take T_n to be an $n = N \times N$ square lattice with unit spacing. To simulate Y_n we generated $Z_n \sim N_n(0, I_n)$ and took $Y_n = X_n \beta + L_n Z_n$, where $V_n = L_n L_n'$ is a Cholesky decomposition. Our concern is primarily with the estimation of covariance parameters and in the examples to follow we estimate a constant mean, for convenience $\beta = 0$.

Example 1. Here, with $\theta = (\sigma_1^2, \alpha)'$. we took the covariance $\sigma(h; \sigma_1^2, \alpha) = \sigma_1^2 \rho(h; \alpha)$, where $\rho(h; \alpha) = 1 - 3|h|/(2\alpha) + |h|^3/(2\alpha^3)$, for $|h| \leqslant \alpha$, 0 elsewhere. This so-called spherical model is commonly used by geostatisticians (Matheron, 1971; Journel & Huijbregts, 1978) for processes on R^2 and R^3, though it arises naturally only in R^3 (Matérn, 1960). It is easy to show that the conditions of Theorem 3 hold for $1 < \alpha < \infty$. We took $\sigma_1^2 = 1$, $\alpha = 3$ and carried out in this example, and those to follow, about 300 replications for each of $N = 6$, 8 and 10 obtaining for each replication the estimate $\hat{\beta}$, $\hat{\sigma}_1^2$ and $\hat{\alpha}$. Figure 1a shows the empirical probability density function of $\hat{\sigma}_1^2$ for $N = 6$ and $N = 10$ as well as normal asymptotic distribution for $N = 10$. There is a significant negative bias for $N = 6$ and $N = 8$. For $N = 10$ there is still some disparity with the normal distribution but the bias is not serious. Figure 1b demonstrates that there is some negative bias to $\hat{\alpha}$ for $N = 6$, however, this bias is not evident for $N = 8$ or for $N = 10$ when the normal approximation is good. The correlation between $\hat{\sigma}_1^2$ and $\hat{\alpha}$ was 0·55 (0·50) for $N = 6$, and 0·51 (0·47) for $N = 10$, where the two values in each case denote the empirical and asymptotic values respectively. We omit the empirical distribution of $\hat{\beta}$ since it did not display significant bias and agreed quite closely with asymptotic theory even for $N = 6$.

In the following examples we also estimate an extra parameter σ_2^2, where σ_2^2 represents the variance of added noise. Thus we can now write $V_n = \sigma_1^2 P_\alpha + \sigma_2^2 I_n$, where P_α is a correlation matrix. In estimating σ_1^2, σ_2^2 and α it is possible that either $\hat{\sigma}_1^2 = 0$ or $\hat{\sigma}_2^2 = 0$ maximize the likelihood, and this possibility should be accounted for by imposing the constraints $\hat{\sigma}_1^2 \geqslant 0$ and $\hat{\sigma}_2^2 \geqslant 0$ in the numerical maximization. Boundary values are a possibility when $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$ for small n but the probability of their occurrence tends to zero as $n \rightarrow \infty$. If, however, $\sigma_2^2 = 0$ for example, the possibility $\hat{\sigma}_2^2 = 0$ remains as $n \rightarrow \infty$ and $\hat{\sigma}_2^2$ behaves asymptotically as a mixture of discrete and continuous distributions. Moran (1971) has shown that in the identically independently distributed case the asymptotic distribution of maximum likelihood estimators of boundary values is obtained by considering boundary occurrences as censored values of the usual asymptotic normal distribution. Whether this result carries over to our case remains open.

Example 2. With parameters and model as in Example 1 we simulated processes with added noise having $\sigma_2^2 = 0·1$. This resulted in about a 3% increase in var $(\hat{\beta})$ and a 40% increase in var $(\hat{\alpha})$ and var $(\hat{\sigma}_1^2)$, but in other respects their distributions remained similar

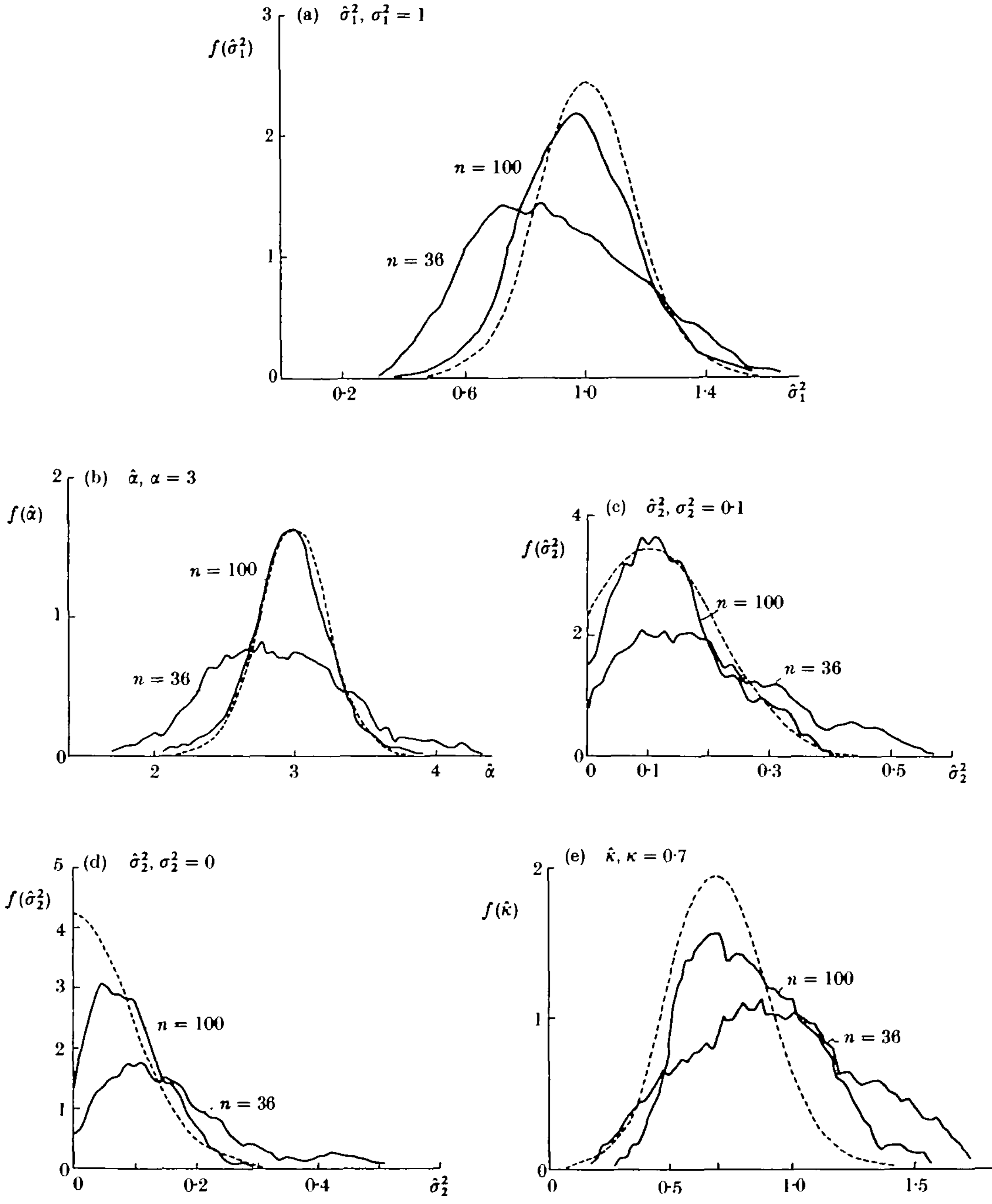

Fig. 1. Distribution of maximum likelihood estimates; solid lines, empirical; broken lines, asymptotic.

and, in particular, negative bias in $\hat{\sigma}_1^2$ and $\hat{\alpha}$ was the same order of magnitude. The estimator $\hat{\sigma}_2^2$ is of interest for its $\hat{\sigma}_2^2 = 0$ values. In Fig. 1c the continuous part of the distribution of $\hat{\sigma}_2^2$ is shown. Also we give the asymptotic distribution cut-off at $\sigma_2^2 = 0$ for $N = 10$. This seems to suggest that it is safe to treat the $\hat{\sigma}_2^2 = 0$ values as censored points of the asymptotic distribution. The corresponding probabilities of $\hat{\sigma}_2^2 = 0$, were 31%, 24% and 19% for $N = 6$, 8 and 10 respectively, and these compare favourably with the empirical frequencies 34%, 24% and 21%. The correlation between $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ is strongly negative: $-0{\cdot}55\,(-0{\cdot}75)$ for $N = 6$ and $-0{\cdot}71\,(-0{\cdot}75)$ for $N = 10$. We also note that $\operatorname{corr}(\hat{\alpha}, \hat{\sigma}_1^2) = 0{\cdot}09\,(0{\cdot}11)$ and $\operatorname{corr}(\hat{\alpha}, \hat{\sigma}_2^2) = 0{\cdot}17\,(0{\cdot}23)$ for $N = 10$.

Example 3. Here we fitted the model as in Example 2 using a true value $\sigma_2^2 = 0$. Thus we estimate a parameter, σ_2^2, which lies on the boundary of Θ. Once again negative bias was evident in $\hat{\sigma}_1^2$ and to a lesser extent in $\hat{\alpha}$. Boundary values occurred with frequencies 61%, 55% and 51% for $N = 6$, 8 and 10 respectively. Thus with increasing N the agreement with a 50% probability for $\hat{\sigma}_2^2 = 0$, derived by censoring the asymptotic normal distribution, seems quite convincing. However, Fig. 1d shows that the empirical distribution for $\hat{\sigma}_2^2 > 0$ does not follow the normal approximation particularly well near the origin.

Example 4. Whittle (1954) has shown that the isotropic covariance model,

$$\sigma(h;\, \sigma_1^2, \kappa) = \sigma_1^2 |h| \kappa K_1(|h| \kappa) \quad (h \in R^2, \kappa > 0), \tag{5·1}$$

plays a natural role on R^2. Here K_1 is a modified Bessel function of the second kind. We can check that the conditions of Theorem 3 are valid using the fact that $K_1(x) \sim (\frac{1}{2}\pi/x)^{-\frac{1}{2}} e^{-x}$. In these simulations we took model (5·1) with $\kappa = 0{\cdot}7$, $\sigma_1^2 = 1$ and added noise with $\sigma_2^2 = 0{\cdot}1$. For $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ the comments in Example 2 are pertinent although now $\operatorname{corr}(\hat{\sigma}_1^2, \hat{\sigma}_2^2) = -0{\cdot}09\,(-0{\cdot}14)$ for $N = 10$. Figure 1e shows the empirical distribution of $\hat{\kappa}$ and the normal approximation for $N = 10$. Evidently $\hat{\kappa}$ is rightward biased for both $N = 6$ and $N = 10$ and its probability density function is skewed. Note also that $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are both negatively correlated with $\hat{\kappa}$, that is $\operatorname{corr}(\hat{\sigma}_1^2, \hat{\kappa}) = -0{\cdot}61\,(-0{\cdot}52)$ and $\operatorname{corr}(\hat{\sigma}_2^2, \hat{\kappa}) = -0{\cdot}68\,(-0{\cdot}64)$ for $N = 10$.

These simulations are intended only to assess qualitatively the small-sample properties of the maximum likelihood estimators. We have not tried to quantify the bias or the departure from normality. However, the results do provide some insight into the sampling distribution for small n. One feature to emerge that does seem important is the negative bias to $\hat{\sigma}_1^2$.

6. The approximate likelihood for a rectangular lattice

For a d-dimensional rectangular lattice of $n = n_1 \times \ldots \times n_d$ points wrapped on to a torus, the log likelihood function for a second-order stationary process can.be written, if we ignore the constant, as (Besag & Moran, 1975; Besag, 1977)

$$L(Y;\, \phi) = -\tfrac{1}{2}\Sigma_{\Omega_n} \left\{ \log f(\omega;\, \theta) + \frac{I_n(\omega;\, \beta)}{f(\omega;\, \theta)} \right\}, \tag{6·1}$$

where $I_n(\omega;\, \beta)$ is the periodogram of W, that is with $W(t_i) = Y(t_i) - x_i'\beta$,

$$I_n(\omega;\, \beta) = \frac{1}{n} \left| \sum_{t \in T_n} W(t)\, e^{i\omega' t} \right|^2, \tag{6·2}$$

and $f(\omega;\, \theta)$ is the spectral density function on the torus. The summation in (6·1) is over multiples of $2\pi/n_i$, that is

$$\Omega_n = \{\omega;\ \omega = 2\pi(k_1/n_1, \ldots, k_d/n_d);\ k_i = 1, \ldots, n_i;\ i = 1, \ldots, d\},$$

and its derivation follows from the fact that $\{f(\omega);\ \omega \in \Omega_n\}$ is the set of eigenvalues of V and that V is diagonalized by a unitary matrix not dependent on θ. For a planar lattice

(6·1) provides an approximation for large $n_1, \ldots, n_d$ and, in its limiting integral form, represents Whittle's (1954) approximation; see also Guyon (1982).

For the problem to which this study is directed, $f(\omega;\, \theta)$ is not often known explicitly; however, one can use (6·1), by suitably truncating the summation

$$f(\omega;\, \theta) = \Sigma_{k \in Z^d} \sigma_k e^{i\omega' k}, \tag{6·3}$$

where σ_k for $k \in Z^d$ is lag k covariance. When a continuous parameter process, $Y(t)$ $(t \in R^d)$ with covariance $\sigma(h;\, \theta)$, is observed on a rectangular lattice with spacings $\Delta_1, \ldots, \Delta_d$, then clearly $\sigma_k = \sigma(h_k;\, \theta)$, for $h_k = (k_1 \Delta_1, \ldots, k_d \Delta_d)$. In this case an alternative computation of $f(\omega;\, \theta)$ is obtained by noting that $f(\omega;\, \theta)$ is composed of $g(\omega;\, \theta)$, the spectral density of $Y(t)$, $t \in R^d$, and its aliases. Thus if $g(\omega;\, \theta)$ is known we can calculate $f(\omega;\, \theta)$ by truncating

$$f(\omega;\, \theta) = \Sigma_{k \in Z^d} g(\omega + 2\pi\delta_k;\, \theta), \tag{6·4}$$

where $\delta_k = (k_1/\Delta_1, \ldots, k_d/\Delta_d)$. For the model (5·1), for example,

$$g(\omega;\, \theta) = \pi\sigma_1^2(|\omega|^2 + \kappa^2)^{-2}$$

(Whittle, 1954, 1963). The estimates, say $\tilde{\theta}_n$ and $\tilde{\beta}_n$, that maximize (6·1) can then be computed as in §2. Note that (6·2) can also be written in terms of the empirical covariances, C_k:

$$I_n(\omega;\, \beta) = \Sigma_{k \in T_n} C_k \cos(\omega' k), \quad C_k = \frac{1}{n} \Sigma_{t, t+k \in T_n} W(t)\, W(t+k). \tag{6·5}$$

Also the approximations for $(X'V^{-1}X)_{ij}$ and $(X'V^{-1}Y)_i$, in (2·6), are respectively

$$\Sigma_{\Omega_n} \frac{I(\omega;\, x_i, x_j)}{f(\omega;\, \theta)}, \quad \Sigma_{\Omega_n} \frac{I(\omega;\, x_i, Y_n)}{f(\omega;\, \theta)},$$

where $I(\omega;\, .\, ,\, .)$ is the coperiodogram of two vectors and need only be computed once.

There is some computational saving in using the ordinary least squares estimation of β, say b_n, and maximizing (6·1) over Θ yielding $\tilde{\theta}(b_n)$, say, thereby obviating the updating of (6·2). We could equally do this in the general case of §2 but its use may be more justified here, at least for $d = 1$, since b_n is asymptotically fully efficient (Grenander & Rosenblatt, 1957, p. 244) under commonly met conditions on f and X_n. Further, it is probable that their result generalizes for $d > 1$ but this point will not be pursued here.

We outline two examples.

The following analysis of a gravity survey of sulphide ore deposits gives a comparison of the exact maximum likelihood estimates $\hat{\theta}_n$ and $\hat{\beta}_n$ with the approximations $\tilde{\theta}_n$, $\tilde{\beta}_n$ and $\tilde{\theta}(b_n), b_n$. Data consist of 187 observations made at nodes of a square 11×17 lattice (Grant, 1957), with spacings 100 feet, which we take as one unit. There is a dome-like trend to the data which we assumed to be quadratic in t. Also following Huijbregts & Matheron (1971) we assume a spherical covariance function with a discontinuity σ_2^2 at $h = 0$. We used (6·3) to compute $f(\omega;\, \theta)$ exactly. In Table 1 are given the three sets of estimates, which are all comparable given the standard errors. However, the exact estimates required at least a hundred times more computer time, even using the algorithm (2·8).

Whittle (1954) suggested that (5·1) adequately fitted the correlogram of yields of orange trees planted in a 20×50 lattice (Batchelor & Reed, 1918). He obtained an empirical value $\kappa = 0{\cdot}13$ with a lag zero intercept of the correlogram at 0·554, this figure

Table 1. *Likelihood estimates for an $n = 11 \times 17$ square grid with unit spacing; approximate standard errors in brackets*

Coeff. of	$\hat{\beta}_n$	$\tilde{\beta}_n$	b_n
1	−17·053 (10·34)	−18·161 (9·33)	−28·282 (11·15)
x	1·879 (2·05)	3·284 (2·35)	3·696 (2·37)
y	6·552 (2·88)	7·600 (3·21)	9·350 (3·39)
x^2	−0·956 (0·11)	−1·060 (0·14)	−1·073 (0·14)
xy	1·076 (0·15)	1·065 (0·07)	1·091 (0·13)
y^2	−0·873 (0·25)	−1·034 (0·31)	−1·084 (0·31)
Covariance	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\theta}(b_n)$
σ_1^2	237·09 (53·14)	252·02 (56·05)	232·85 (53·35)
σ_2^2	13·31 (15·47)	4·82 (15·13)	14·51 (15·34)
α	3·73 (0·26)	3·58 (0·18)	3·50 (0·22)

being an estimate of $\sigma_1^2/(\sigma_1^2+\sigma_2^2)$ in our notation. To maximize the exact likelihood is clearly impractical for $n = 1000$ but it is interesting to compare the approximate maximum likelihood estimates with Whittle's figures. Using (6·4) to compute $f(\omega;\, \theta)$ and taking a constant mean value we obtain $\tilde{\kappa} = 0{\cdot}4721\,(0{\cdot}084)$, $\tilde{\sigma}_1^2 = 1464{\cdot}4\,(287{\cdot}8)$ and $\tilde{\sigma}_2^2 = 1248{\cdot}1\,(81{\cdot}1)$ with $\tilde{\beta} = 138{\cdot}0665\,(9{\cdot}15)$, where asymptotic standard errors are in brackets. The ordinary least squares estimator is $b_n = \bar{y} = 138{\cdot}0650$, differing only marginally from $\tilde{\beta}$, and the covariance estimates are identical. The estimate of the intercept is 0·540, agreeing well with Whittle's 0·554; however, our $\tilde{\kappa}$ indicates a faster rate of decay of correlation than is suggested by the empirical values. It is also of interest to note that if Guyon's (1982) alternative likelihood approximation is used the estimators are quite different: $\overset{*}{\kappa} = 1{\cdot}296\,(0{\cdot}073)$, $\overset{*}{\sigma}{}_1^2 = 3723{\cdot}2\,(265{\cdot}7)$ and $\overset{*}{\sigma}{}_2^2 = 0$. Guyon's correction ensures that bias is $O(n^{-1})$ by replacing C_k in (6·5) with the unbiased covariance estimator $C_k^* = nC_k/n_k$, n_k being the cardinality of $\{t;\, t, t+k \in T_n\}$. This adjustment probably results in some loss in efficiency for finite n and further C_k^* is only unbiased if the model $E(Y_n) = \beta X_n$ is correctly specified and β is known. If this is not the case neither C_k nor C_k^* are unbiased and the bias to C_k^* is n/n_k times larger than that of C_k. Thus Guyon's estimators may be more sensitive to a misspecified trend. This seems likely to explain the differences between the estimates for the data here, since there are evidently spatial variations in fertility unaccounted for by the assumption of a constant mean. This example is only illustrative and further analysis will not be pursued here.

We would like to thank the referees and Dr J. T. Kent for valuable comments on earlier drafts and the Science and Engineering Research Council for support.

References

Batchelor, L. D. & Reed, H. S. (1918). Relation of the variability of yields of fruit trees to the accuracy of field trials. *J. Agric. Res.* **22**, 245–82.

Besag, J. (1977). Errors-in-variables estimation for Gaussian lattice schemes. *J. R. Statist. Soc.* B **39**, 73–8.

Besag, J. & Moran, P. A. P. (1974). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* **62**, 555–62.

Cliff, A. D. & Ord, J. K. (1973). *Spatial Autocorrelation.* London: Pion Press.

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

Cook, D. G. & Pocock, S. J. (1983). Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics* **39**, 361 71.

Grant, F. (1957). A problem in the analysis of geophysical data. *Geophysics* **22**, 309–44.

Grenander, U. & Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series.* New York: Wiley.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**, 95-105.

Huijbregts, G. & Matheron, G. (1971). Universal Kriging. *The Canadian Institute of Mining and Metallurgy*. Special Volume **12**, 159–69.

Journel, A. G. & Huijbregts, C. J (1978). *Mining Geostatistics*. London: Academic Press.

Mardia, K. V. (1980). Some statistical inference problems in kriging. II: Theory. *Proc. 26th Int. Geol. Congress, Sciences de le Terre, Série Informatique Geologique.* 113–31.

Mátern, B. (1960). Spatial variation. *Meddelanden Fran Statens Skogsforskningsinstitut*, Band **49**, No. 5.

Matheron, G. (1971). The theory of regionalised variables and its applications. *Les Cahiers du Centre de Morphologie Mathematique.* Fasc. No. 5, Fontainebleau.

Moran, P. A. P. (1971). Maximum likelihood estimation in non-standard conditions. *Proc. Camb. Phil. Soc.* **70**, 441–50.

Ripley, B. D. (1981). *Spatial Statistics*. New York: Wiley.

Schagan, I. P. (1980). The use of stochastic processes in interpolation and approximation. *Int. J. Computer Math.* B, **8**, 63 76.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.* **8**, 1375 81.

Trench, W. F. (1964). An algorithm for the inversion of finite Toeplitz matrices. *J. Soc. Indust. Appl. Maths.* **12**, 515–22.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–49.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Int. Statist. Inst.* **40**, 974–85.

[*Received March* 1982. *Revised June* 1983]