PREAMBLE

Knowledge comes, but wisdom lingers.

Tennyson, **Locksley Hall**

In the ideal anthology dealing with contemporary spatial statistics, one would like to include a paper by Besag, by Mardia, by Ord, and by Ripley; I am glad that the present volume is able to contain a contribution by at least three of these four statisticians. In terms of Mardia's submission, more than just a knowledge of spatial statistical modelling is imparted. The benefits of his accumulated wisdom are presented, with an emphasis on his spatial work during the past decade. The purpose of this paper is to outline maximum likelihood estimation methods for spatial linear models, with the presentation being enhanced by the inclusion of numerous abstract examples and data set analyses. Much guidance is offered here, for the researcher who is looking for a comprehensive treatment of spatial statistical modelling. Once more computational issues are raised, both by Mardia and by his discussant. Paelinck emphasizes the points raised by Mardia, and agrees with Mardia's contention that major problems meriting subsequent research attention include model identifiability, anisotropy, and second-order conditions.

The Editor

Maximum Likelihood Estimation for Spatial Models

Kanti V. Mardia*

Department of Statistics, University of Leeds, Leeds, LS2 9JT, UK

Overview: The paper gives maximum likelihood (ML) estimation methods for spatial linear models in three forms: Direct Representation (DR), Conditional Autoregressive (CAR) Models and Simultaneous Autoregressive (SAR) Models for the Gaussian Case. We also discuss the computational aspects of the methods. The problem of asymptotic bias is considered for the DR. For the intrinsic random field we obtain the maximum likelihood estimators (MLEs) and indicate a relationship with marginal likelihood. We give the exact MLEs for CAR models on a circle and a torus together with some properties. It is indicated how the same approach applies to the SAR models. For the stationary random field, we discuss the Whittle approximation. We also consider the MLE for intrinsic CAR. We then describe the estimation of a nugget parameter and its asymptotic distribution. We indicate the extension of the method to the multivariate case, block data, missing values in lattice, designs under spatial correlation, and the such. Finally, a general discussion is given.

1. Introduction

Spatial models have become increasingly used for image analysis (see, Mardia, 1989 for references). Previous applications were more in agriculture, forestry, ecology, geography, geology, to name a few disciplines. We mainly study ML estimation for three forms of spatial linear models: Direct Representation (DR), Conditional Auto–regressive (CAR) and Simultaneous Auto–regressive (SAR) when the random field is Gaussian. Finite range covariance schemes lead to a sparse covariance matrix Σ of the random field in DR, whereas CAR and SAR lead to sparse Σ^{-1}. However, the models are interrelated; see Section 2.

Section 3 gives the MLEs and their problems for DR. Section 4 gives the MLEs for the intrinsic model. Section 5 gives CAR models with some comments on the SAR case. Section 6 discusses the problem for a nugget parameter or errors in variable model. Section 7 considers some other uses and extensions such as block data, missing values in a lattice and the use of these methods in experimental design. The last section gives a discussion.

* The author is grateful to Dan Griffith for inviting him to participate in this Symposium. He is also grateful to John Kent, Tim Hainsworth and Alan Watkins for their helpful comments. This work is supported by NSF grant DMS-8803207 to Professor Watson, Princeton University.

2. The model and its three forms

2.1. The spatial linear model

Let $\{X(\mathbf{t})\}$ be a stochastic process where $\mathbf{t}$ represents a point in d–dimensional space where we write $T = R^d$ for the Euclidean space and $T = Z^d$ for points on a regular lattice. Suppose the process is sampled at points $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$ to give the sample vector

$$\mathbf{X} = \{X(\mathbf{t}_1), X(\mathbf{t}_2), \ldots, X(\mathbf{t}_n)\}'.$$

Suppose that $\{X(\mathbf{t})\}$ is Gaussian and that

$$\mu(\mathbf{t}) = E\{X(\mathbf{t})\} = \mathbf{f}(\mathbf{t})'\boldsymbol{\beta}, \tag{2.1.1}$$

where $\boldsymbol{\beta}$ is a q–by–1 parameter vector and $\mathbf{f}(\mathbf{t})$ is a vector of known functions, possibly monomials, which may describe a trend. If there is no confusion we will write

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)' \text{ and } \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'.$$

Suppose that

$$\{X(\mathbf{t}) - \mu(\mathbf{t})\}$$

is second–order stationary with

$$\mathrm{Cov}\{X(\mathbf{t}),\, X(\mathbf{t}+\mathbf{h})\} = \sigma(\mathbf{h}; \boldsymbol{\theta}), \tag{2.1.2}$$

where $\sigma(\cdot;\, \boldsymbol{\theta})$ is a positive definite function of $\mathbf{h}$, assumed known apart for a p–by–1 vector of parameters $\boldsymbol{\theta}$. [We will discuss some restrictions on $\sigma(\cdot;\, \boldsymbol{\theta})$ for the asymptotic theory in Section 3.] Thus from (2.1.1),

$$E(\mathbf{X}) = \mathbf{F}\boldsymbol{\beta}, \tag{2.1.3}$$

where $\mathbf{F} = \{\mathbf{f}(\mathbf{t}_1), \mathbf{f}(\mathbf{t}_2), \ldots, \mathbf{f}(\mathbf{t}_n)\}'$. Let the covariance matrix of $\mathbf{X}$ be $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ with

$$\sigma_{ij} = (\boldsymbol{\Sigma})_{ij} = \{\sigma(\mathbf{t}_i - \mathbf{t}_j;\, \boldsymbol{\theta})\}.$$

Thus our model is of the form

$$\text{observation} = \text{deterministic trend} + \text{stochastic fluctuation}$$

where trend measures the long–term variation whereas the stochastic fluctuation measures the short–term or the local variation.

2.2. The direct representation

We can specify $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ by modelling $\sigma(\mathbf{h}; \boldsymbol{\theta})$ directly. With this formulation, we will call the spatial linear model the Direct Representation (DR).

Consider the following example. Table 1 gives a topographic data set consisting of 52 points from Davis (1973, pp. 313-314) with $T = R^2$. Figure 1 plots the data, and looking at the values closely, there is some indication of trend. At a smaller scale, we will expect local variation. One way to model the local variation, is to take $\sigma(\mathbf{h}; \boldsymbol{\theta})$ with finite *range*

TABLE 1
GEOGRAPHIC COORDINATES
AND ELEVATIONS OF CONTROL POINTS
FOR EXAMPLE SURVEYING PROBLEM

E–W Coordinate $\mathbf{t}_1$	N–S Coordinate $\mathbf{t}_2$	Elevation $X(\mathbf{t}_1, \mathbf{t}_2)$	E–W Coordinate $\mathbf{t}_1$	N–S Coordinate $\mathbf{t}_2$	Elevation $X(\mathbf{t}_1, \mathbf{t}_2)$
0.3	6.1	870	5.2	3.2	805
1.4	6.2	793	6.3	3.4	840
2.4	6.1	755	0.3	2.4	890
3.6	6.2	690	2.0	2.7	820
5.7	6.2	800	3.8	2.3	873
1.6	5.2	800	6.3	2.2	875
2.9	5.1	730	0.6	1.7	873
3.4	5.3	728	1.5	1.8	865
3.4	5.7	710	2.1	1.8	841
4.8	5.6	780	2.1	1.1	862
5.3	5.0	804	3.1	1.1	908
6.2	5.2	855	4.5	1.8	855
0.2	4.3	830	5.5	1.7	850
0.9	4.2	813	5.7	1.0	882
2.3	4.8	762	6.2	1.0	910
2.5	4.5	765	0.4	0.5	940
3.0	4.5	740	1.4	0.6	915
3.5	4.5	765	1.4	0.1	890
4.1	4.6	760	2.1	0.7	880
4.9	4.2	790	2.3	0.3	870
6.3	4.3	820	3.1	0.0	880
0.9	3.2	855	4.1	0.8	960
1.7	3.8	812	5.4	0.4	890
2.4	3.8	773	6.0	0.1	860
3.7	3.5	812	5.7	3.0	830
4.5	3.2	827	3.6	6.0	705

Elevation is measured in feet above sea level. Coordinates are expressed in 50–feet units measured from an arbitrary origin located in the southwest corner, with $\mathbf{t}_1$ being the East–West coordinate and $\mathbf{t}_2$ being the North–South coordinate (from Davis, 1973).

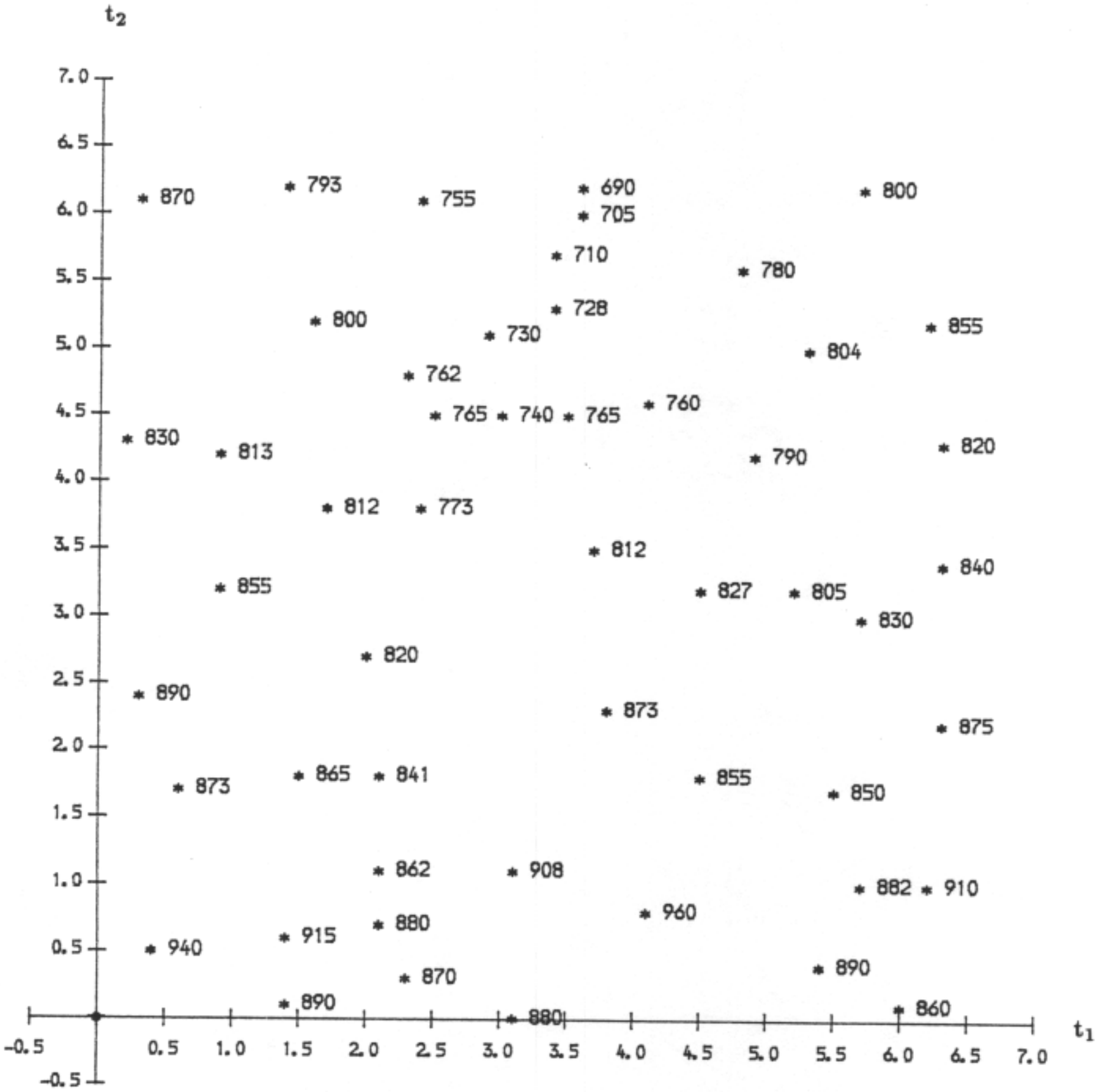
α so that $\sigma(\mathbf{h}; \boldsymbol{\theta}) = 0$ for $|\mathbf{h}| > \alpha$. For example, we could use the power scheme where the covariance function is given by (Mardia and Watkins, 1989)

$$\sigma(\mathbf{h}; \boldsymbol{\theta}) = \sigma^2(1 - \alpha^{-1}|\mathbf{h}|)^4, \quad |\mathbf{h}| < \alpha; \quad = 0 \text{ otherwise} \tag{2.2.1}$$

where $\boldsymbol{\theta} = (\sigma^2, \alpha)'$. We note that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is sparse if α is less than the maximum distance between the points. Further, the sites can be irregularly distributed.

Figure 1.

Spatial distribution of elevations (Davis, 1973).



2.3. The conditional autoregressive model

Another way to model the spatial linear model is to specify $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ through a conditional autoregression (CAR) model. We will mainly concentrate here on the case when all the n sites are on a regular lattice. On an infinite regular lattice Z^d, the CAR model is defined by (Besag, 1974)

$$E(X_i \,|\, \text{rest}) = \mu_i + \sum_{j \neq i} \phi_{ij}(\boldsymbol{\theta})(x_j - \mu_j), \tag{2.3.1}$$

$$\text{Var}(X_i|\text{rest}) = \tau^2, \tag{2.3.2}$$

where 'rest' denotes all points $j \in Z^d, j \neq i$, and $\boldsymbol{\Phi}(\boldsymbol{\theta}) = [\phi_{ij}(\boldsymbol{\theta})]$ is a symmetric matrix with $\phi_{ii}(\boldsymbol{\theta}) = 0$, and $\phi_{ij}(\boldsymbol{\theta})$ a function such that the process $\{X(\mathbf{t})\}$ is covariance stationary. In particular, we take

$$\begin{aligned} \phi_{ij}(\boldsymbol{\theta}) &= \theta_{i-j} \text{ for } i - j \in N, \\ &= 0 \text{ otherwise} \end{aligned} \tag{2.3.3}$$

where N denotes a finite symmetric neighbourhood of the origin and $\theta_{i-j} = \theta_{j-i}$. For example, for the first–order neighbourhood in 2–dimensions,

$$N = \{(-1, 0), (1, 0), (0, 1), (0, -1)\}, \tag{2.3.4}$$

and there are two parameters in $\boldsymbol{\Phi}(\boldsymbol{\theta})$, θ_{10} and θ_{01} which we will write as θ_1 and θ_2 respectively. Thus for $\mu_i = 0$; the CAR is defined by

$$E(X_i \,|\, \text{rest}) = \sum_{j \in N} \theta_{i-j} x_j, \; \text{Var}(X_i \,|\, \text{rest}) = \tau^2. \tag{2.3.5}$$

Sometimes it will be convenient to write the parameters as

$$\phi_h = \tau^{-2}, \; \mathbf{h} = \mathbf{0}; \; \phi_h = \tau^{-2}\theta_h, \; \mathbf{h} \neq \mathbf{0}. \tag{2.3.6}$$

The simplest example of (2.3.5) is when $\theta_{i-j} = \theta$ so that

$$\boldsymbol{\Phi}(\boldsymbol{\theta}) = \theta \mathbf{W}, \; E(X_i \,|\, \text{rest}) = \theta \sum_{j \in N} x_{i+j}, \tag{2.3.7}$$

where $(\mathbf{W})_{ij} = 1$ if $i - j \in N$; $= 0$ otherwise. $\mathbf{W}$ is called the adjacency matrix. We will call this the basic CAR model.

If $\mathbf{X}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$E(X_i \,|\, \text{rest}) = \mu_i + \sum_{j \neq i} (\sigma^{ij}/\sigma^{ii})(x_j - \mu_j), \text{ and } \text{Var}(X_i \,|\, \text{rest}) = 1/\sigma^{ii}.$$

On identifying these two expressions with (2.3.1) and (2.3.2) respectively, we obtain

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = \tau^{-2}[\mathbf{I} - \boldsymbol{\Phi}(\boldsymbol{\theta})]. \tag{2.3.8}$$

It should be noted that the infinite matrices $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ have to be absolutely convergent and positive definite. $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is certainly p.d. if $|1 - \sum_{h \in N} \theta_h \cos(\boldsymbol{\omega}'\mathbf{h})| < 1$. This holds if

$$\sum_{h \in N} |\theta_h| < 1 \tag{2.3.9}$$

[since $|\cos(\boldsymbol{\omega}'\mathbf{h})| \leq 1$] and therefore in the basic model $|\theta| < 1/\nu$ where ν is the number of neighbours.

In general, it should be noted that for the stationary CAR,

$$\sigma(\mathbf{h};\boldsymbol{\theta}) = (2\pi)^{-d} \int_{(-\pi,\pi)^d} [\exp(i\boldsymbol{\omega}'\mathbf{h})]/[\sum_{s \in N_0} \phi_s \cos(\boldsymbol{\omega}'\mathbf{s})] d\boldsymbol{\omega}, \tag{2.3.10}$$

where the set N_0 is N with the origin included. Hence the class of stationary process includes the class of CARs. We can use $\sigma(\mathbf{h};\boldsymbol{\theta})$ in the DR but here $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is complicated. However, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is simple and sparse.

So far we have discussed the CAR on infinite lattice Z^2 but in practice, our sites are on a finite lattice D. Let $C = Z^2 - D$. We first obtain the inverse covariance matrix of $\{X(\mathbf{t})\}, t \in D$.

For the infinite lattice, we can write the inverse covariance matrix (2.3.8) of $\mathbf{X}$ as

$$\tau^2 \boldsymbol{\Sigma}_\infty^{-1} = \mathbf{I}_\infty - \boldsymbol{\Phi}_\infty \equiv \begin{pmatrix} \mathbf{I}_D - \boldsymbol{\Phi}_D & \mathbf{B} \\ \mathbf{B}' & \mathbf{I}_c - \boldsymbol{\Phi}_c \end{pmatrix},$$

where $\mathbf{I}_D$ and $\mathbf{I}_c$ are the identity matrix, $\boldsymbol{\Phi}_D$ matrix for the finite lattice D and so forth. Then Künsch (1983) has shown that the inverse covariance matrix for the finite lattice D is

$$\boldsymbol{\Sigma}_D^{-1} = \mathbf{I}_D - \boldsymbol{\Phi}_D - \boldsymbol{\Gamma}_D,$$

where

$$\boldsymbol{\Gamma}_D = \mathbf{B}_D (\mathbf{I}_c - \boldsymbol{\Phi}_c)^{-1} \mathbf{B}'.$$

provided all matrices converge absolutely. Thus (2.3.8) is not valid for D unless $\boldsymbol{\Gamma}_D = \mathbf{0}$. However, the exact $\boldsymbol{\Sigma}_D^{-1}$ can be obtained through $\boldsymbol{\Sigma}_D$ whose elements are obtained from (2.3.10).

In fact, $\boldsymbol{\Sigma}_D(\boldsymbol{\theta})$ is the covariance function of the marginal distribution of $\{X(\mathbf{t})\}$ on $\mathbf{t} \in D$, and therefore the process is stationary on $\mathbf{t} \in D$ with $\sigma(\mathbf{h};\boldsymbol{\theta})$ given by (2.3.10). We will call this process an M–CAR. However, $\boldsymbol{\Sigma}_D(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_D(\boldsymbol{\theta})^{-1}$ are both complicated, unlike $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ given by (2.3.8) with $\boldsymbol{\Phi}(\boldsymbol{\theta})$ defined at (2.3.3). We can achieve some simplicity by making boundary adjustments in the following two ways:

(i) T–CAR. Wrap the CAR on a torus.

(ii) C–CAR. For $\mu(\mathbf{t}) = 0$, use the conditional distribution of $\{X(\mathbf{t})\}$ on D given $X(\mathbf{t}) = 0$, $\mathbf{t} \notin D$, *i. e.* use the free boundaries.

Under the C–CAR, the CAR representation (2.3.5) is preserved but we do not have stationarity. Under the T–CAR, we have stationarity as well as the CAR representation but the periodic boundaries are not realistic.

Using the C–CAR can lead to serious bias in estimation for large n. The main reason is that the boundary for $d = 2$ is of order $n^{1/2}$ and the effect of neglecting it can be of order higher than n^{-1}. For some practical examples see Guyon (1982). Martin (1987) has highlighted some confusion in this area. The above result (2.3.8) requires basic knowledge of conditional distributions for the multivariate normal case, especially Theorems 3.2.3 and 3.2.4 in Mardia, Kent and Bibby (1989).

Suppose the sites are irregularly distributed; then the extension of the CAR model is not straightforward, in general (see, Besag, 1975). An interesting particular case is for $\boldsymbol{\Phi}(\boldsymbol{\theta})$ given by (2.3.7), where $W_{ij} = 1$ if i and j are nearest neighbours, and zero otherwise. Also for the regularized process, equation (2.3.7) can be used, where now

$W_{ij} = 0$ if areas are not contiguous, and

$\propto$ a monotonic function of the length of the common boundary otherwise.

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{W}$, with $\lambda_1 < \cdots < \lambda_n$, then $\boldsymbol{\Sigma}$ is positive definite if $0 \leq \theta < 1/\lambda_n$.

2.4. The simultaneous autoregressive model

For simplicity let us assume the finite lattice case. We have

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\psi}(\boldsymbol{\theta})(\mathbf{X} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

Thus with $\boldsymbol{\psi} \equiv \boldsymbol{\psi}(\boldsymbol{\theta})$,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = \sigma^{-2}(\mathbf{I} - \boldsymbol{\psi}')(\mathbf{I} - \boldsymbol{\psi}), \text{ or } \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2(\mathbf{I} - \boldsymbol{\psi})^{-1}(\mathbf{I} - \boldsymbol{\psi}')^{-1}, \tag{2.4.1}$$

where the defining property requires $|\mathbf{I} - \boldsymbol{\psi}|$ to be non–singular. As in the CAR case, we can take a particular case as $\boldsymbol{\psi} = \theta\mathbf{W}$ so that $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ is sparse. Here $\mathbf{W}$ need not be a symmetric matrix. For almost all values of θ, the class of the SARs is included in the class of the CARs on the infinite lattice. However, note that $\boldsymbol{\psi}$ is not uniquely determined by a given $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, unlike the CAR. We will not discuss the estimation problems for the SAR in detail.

2.5. Particular cases

We now give these three representations for the Geometric Scheme, where the correlation function is

$$\rho(\mathbf{h}; \lambda, \nu) = \lambda^{|h_1|}\nu^{|h_2|}. \tag{2.5.1}$$

We have its SAR and CAR representations on the infinite lattice as

SAR: $X_{ij} = \lambda X_{i-1,j} + \nu X_{i,j-1} - \lambda\nu X_{i-1,j-1} + \varepsilon_{ij}$,

CAR: $E(X_{ij}|\cdot) = \alpha(x_{i-1,j} + x_{i+1,j}) + \beta(x_{i,j-1} + x_{i,j+1}) - \alpha\beta(x_{i-1,j-1} + x_{i-1,j+1} + x_{i+1,j-1} + x_{i+1,j+1})$,

Var $(X_{ij}|\cdot) = \sigma^2/(1+\lambda^2)(1+\nu^2)$,

where $\alpha = \lambda/(1+\lambda^2)$ and $\beta = \nu/(1+\nu^2)$. From $E(X_{ij}|\cdot)$, it follows that the neighbourhood is of the second order.

For the first–order neighbourhood for the CAR in Z^2, we have from (2.3.10)

$$\sigma(h_1, h_2) = \tau^2(2\pi)^{-2} \int_{(-\pi,\pi)^2} \frac{\cos(\omega_1 h_1)\cos(\omega_2 h_2)}{1 - 2\theta_1\cos(\omega_1) - 2\theta_2\cos(\omega_2)}\, d\omega_1 d\omega_2. \tag{2.5.2}$$

Besag (1981) shows that for $\theta_1 + \theta_2 > 0.48$ and $(h_1, h_2) \neq (0,0)$,

$$\sigma(h_1, h_2) \simeq \tau^2 \{2\pi(\theta_1\theta_2)^{1/2}\}^{-1} K_0\left((1 - 2\theta_1 - 2\theta_2)^{1/2}[\frac{h_1^2}{\theta_1^2} + \frac{h_2^2}{\theta_2^2}]^{1/2}\right),$$

where $K_0(\cdot)$ is the modified Bessel function of the second kind and order zero. In particular, for $\theta_1 = \theta_2 = \theta$, $\sigma(h_1, h_2)$ is closely approximated by a monotonic decreasing function of $|\mathbf{h}| = (h_1^2 + h_2^2)^{1/2}$, so that it is almost an isotropic scheme. Further, there is very slow decay of $\rho(\mathbf{h})$ with increasing $|\mathbf{h}|$ whenever $\rho(1,0)$ is moderately high; *e. g.*, if $\rho(1,0) = 0.85$, then $|\mathbf{h}|$ must exceed 2000 before $\rho(\mathbf{h}) < 0.1$. Note that the Geometric Scheme is highly anisotropic and cannot display the type of slow decay for the first-order CAR scheme considered here.

3. ML estimation for DR

3.1. ML equations

In this Section, we follow Mardia and Marshall (1984). Since from Section 2.1 $\mathbf{X}$ is multivariate normal, the log-likelihood function of $\mathbf{X}$ with the parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\ell = \ell(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})'[\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta}). \tag{3.1.1}$$

On differentiating (3.1.1) with respect to $\boldsymbol{\beta}$, with the help of $\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$ we get

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{X} - (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})\boldsymbol{\beta}. \tag{3.1.2}$$

Hence, the MLE of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{F})^{-1}\mathbf{F}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}, \tag{3.1.3}$$

where $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$, $\hat{\boldsymbol{\theta}}$ being the MLE of $\boldsymbol{\theta}$. To differentiate (3.1.1) with respect to $\boldsymbol{\theta}$, we first note the following two results:

$$\frac{\partial \log|\boldsymbol{\Sigma}|}{\partial \theta_i} = \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i), \quad \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \theta_i} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1},$$

where $\boldsymbol{\Sigma}_i = \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}$. Hence differentiating with respect to θ_i, with the help of these two results, we have

$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i) + \frac{1}{2}\mathbf{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\mathbf{w}, \quad i = 1, \ldots, p, \tag{3.1.4}$$

where $\mathbf{w} = \mathbf{X} - \mathbf{F}\boldsymbol{\beta}$. Thus the p equations for $\hat{\boldsymbol{\theta}}$ are

$$\hat{\mathbf{w}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{w}} = \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_i), i = 1, \ldots, p, \tag{3.1.5}$$

with $\hat{\mathbf{w}} = \mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$. We note that the equations hold even when $\{X(\mathbf{t})\}$ is not covariance stationary.

One does not see an analytical solution to the ML equations (3.1.3) and (3.1.5). However, for a nested scheme such that

$$\sigma(\mathbf{h};\boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i \sigma_i(|\mathbf{h}|) \tag{3.1.6}$$

some progress can be made, since $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is of the form $\Sigma \mathbf{K}_i \theta_i$, where the matrices $\mathbf{K}_i$ are fixed. These are realistic models in the Analysis of Variance (see, Hocking, 1984), but not so realistic in Spatial Statistics. However, we will just give one example for its mathematical contents, namely the variance component scheme. Suppose the process is stationary, with

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \theta_1 \mathbf{I}_n + \theta_2 (\mathbf{I}_r \otimes \mathbf{E}_s), \tag{3.1.7}$$

where $n = rs$, and $\mathbf{E}_s$ is an s-by-s matrix of 1s. Then $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and the MLEs of θ_1 and θ_2 are obtained from the solution to

$$r(s-1)\lambda_0^{-1} + (r-1)\lambda_1^{-1} + \lambda_2^{-1} = q_0\lambda_0^{-2} + q_1\lambda_1^{-2},\ s(r-1)\lambda_1^{-1} + n\lambda_2^{-1} = sq_1\lambda_1^{-2},$$

where

$$q_i = \mathbf{X}'\mathbf{H}\mathbf{A}_i\mathbf{H}\mathbf{X}\ (i = 0, 1),\ \lambda_0 = \theta_1,\ \lambda_1 = \theta_1 + s\theta_2,\ \lambda_2 = \theta_1 + n\theta_2, \tag{3.1.8}$$

and

$$\mathbf{H}_n = \mathbf{I}_n - n^{-1}\mathbf{E}_n,\ \mathbf{A}_0 = \mathbf{I}_r \otimes \mathbf{H}_s,\ \text{and}\ \mathbf{A}_1 = s^{-1}\mathbf{H}_r \otimes \mathbf{E}_s.$$

The ML equations can be given in closed form since here $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ can be obtained explicitly.

Also, some progress can be made for a finite lattice with the doubly geometric scheme given by (2.5.1). To obtain numerical solutions in general, we give the standard solving method involving the information matrix, which we now derive.

3.2. The information matrix and asymptotic normality

On differentiating (3.1.4) with respect to θ_j, we find that

$$2\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} = -\mathrm{tr}(\mathbf{R}_{ij} - \mathbf{S}_{ij}) - \mathbf{w}'(\mathbf{S}_{ij} + \mathbf{S}_{ji} - \mathbf{R}_{ij})\boldsymbol{\Sigma}^{-1}\mathbf{w}, \tag{3.2.1}$$

where

$$\mathbf{R}_{ij} = \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}^{-1},\ \mathbf{S}_{ij} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j, \boldsymbol{\Sigma}_{ij} = \frac{\partial^2\boldsymbol{\Sigma}}{\partial\theta_i\partial\theta_j} = \frac{\partial\boldsymbol{\Sigma}_i}{\partial\theta_j}. \tag{3.2.2}$$

Since $E(\mathbf{w}'\mathbf{X}\mathbf{w}) = \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{X})$, after taking the expectation of (3.2.1) we obtain

$$E[-\frac{\partial^2\ell}{\partial\theta_i\partial\theta_j}] = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j) = a_{ij},\ \text{say.} \tag{3.2.3}$$

We also have, on differentiating (3.1.2) with respect to θ_j, $\frac{\partial^2\ell}{\partial\beta\,\partial\theta_j} = -\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\mathbf{w}$. Since $E(\mathbf{w}) = \mathbf{0}$, we find that

$$E[\frac{\partial^2\ell}{\partial\beta\,\partial\theta_j}] = 0.$$

Furthermore, from (3.1.2) we also get $\frac{\partial^2 \ell}{\partial \beta^2} = -\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}$. Hence the information matrix for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\mathbf{B}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix} = \text{diag}(\mathbf{B}_\beta, \mathbf{B}_\theta), \text{ say,} \tag{3.2.4}$$

where $\mathbf{A} = (a_{ij})$ is defined by (3.2.3).

Under certain regularity conditions, including differentiability of $\sigma(\mathbf{h}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, it can be shown that (Mardia and Marshall, 1984)

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\theta}} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{pmatrix}, \begin{pmatrix} (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1} \end{pmatrix}\right]. \tag{3.2.5}$$

Hence $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are asymptotically independent. Further, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}$, and of $\hat{\boldsymbol{\theta}}$ is $\mathbf{A}^{-1}$.

Note that for asymptotic normality we require continuity, growth and convergence of the observed information matrix. Mardia and Marshall (1984) assume, among other conditions, that the sample set grows (*e. g.*, $|\mathbf{t}_i - \mathbf{t}_j| \geq c > 0$) in such a way that the sampling domain increases in extent as n increases. The sufficient conditions for the asymptotic normality and weak consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ are given in Mardia and Marshall (1984). One of the conditions is that $\sigma(\cdot; \boldsymbol{\theta})$ is twice differentiable, with continuous second derivatives.

Note that the above condition is not satisfied for the spherical scheme with range parameter, *e. g.* at $|\mathbf{h}| = \alpha$. Moreover,

$$\sigma(\mathbf{h}; \alpha) = 1 - \frac{3}{2}\frac{|\mathbf{h}|}{\alpha} + \frac{1}{2}\frac{|\mathbf{h}|^3}{\alpha^3}, \quad |\mathbf{h}| < \alpha; \ = 0 \text{ otherwise,}$$

is not twice differentiable (see Mardia and Watkins, 1989). Note that this scheme is commonly used in geostatistics (Matheron, 1971); but, in particular, the asymptotic standard error (SE) written from the information matrix may not be valid for this scheme.

The problem of estimation in a bounded region $D \subset R^d$ with sampling increasingly dense in D was recognized but was excluded from their discussion. Subsequently Stein (1987, 1988) has investigated such problems effectively.

Next consider the scale parameter. Note that we could write $\boldsymbol{\theta}' = (\theta_1, \boldsymbol{\theta}_2')$, where $\theta_1 = \sigma^2$ is such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2\mathbf{P}(\boldsymbol{\theta}_2)$ such that $P(\boldsymbol{\theta}_2)$ represents a correlation matrix. Then we have

$$\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) = \mathbf{P}(\boldsymbol{\theta}_2) \ (i = 1); \ = \sigma^2\mathbf{P}_i(\boldsymbol{\theta}_2) \ (i \neq 1),$$

where $\mathbf{P}_i(\boldsymbol{\theta}_2) = \frac{\partial \mathbf{P}(\boldsymbol{\theta}_2)}{\partial \theta_i}$ $(i \neq 1)$. Hence, the ML equations from (3.1.3) and (3.1.4) are

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\hat{\mathbf{P}}^{-1}\mathbf{F})^{-1}\mathbf{F}'\hat{\mathbf{P}}^{-1}\mathbf{X}, \tag{3.2.6}$$

$$\hat{\sigma}^2 = [(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}})'\hat{\mathbf{P}}^{-1}(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}})]/n, \tag{3.2.7}$$

and

$$\hat{\sigma}^2 \text{tr}(\hat{\mathbf{P}}^{-1}\hat{\mathbf{P}}_i) = \mathbf{w}'\hat{\mathbf{P}}^{-1}\hat{\mathbf{P}}_i\hat{\mathbf{P}}^{-1}\mathbf{w}.$$

Further, the information matrix has elements

$$a_{11} = (n/2\sigma^4),$$
$$a_{1i} = \frac{1}{2}\,\mathrm{tr}(\mathbf{P}^{-1}\mathbf{P}_i)/\sigma^2 \ (i \neq 1)$$
$$a_{ij} = \frac{1}{2}\,\mathrm{tr}(\mathbf{P}^{-1}\mathbf{P}_i\mathbf{P}^{-1}\mathbf{P}_j) \ (i, j \neq 1). \tag{3.2.8}$$

However, mathematically it is simpler to consider $\boldsymbol{\theta}$ itself.

3.3. Computational aspects

3.3.1 The scoring method.

From (3.2.4), the method implies updating $(\boldsymbol{\beta}_{k+1}, \boldsymbol{\theta}_{k+1})$ at stage $(k+1)$ using

$$\begin{pmatrix}\boldsymbol{\beta}_{k+1}\\ \boldsymbol{\theta}_{k+1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_{k}\\ \boldsymbol{\theta}_{k}\end{pmatrix} + \mathrm{diag}(\mathbf{B}_{\beta_k}^{-1}, \mathbf{B}_{\theta_k}^{-1}) \begin{pmatrix}\ell_{\beta k}\\ \ell_{\theta k}\end{pmatrix},$$

where ℓ_{β_k}, and $\ell_{\theta k}$ are the derivative vectors of the log–likelihood ℓ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively at $\boldsymbol{\beta} = \boldsymbol{\beta}_k, \boldsymbol{\theta} = \boldsymbol{\theta}_k$. This implies equivalence with the use of

$$\boldsymbol{\beta}_k = [\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\mathbf{F}]^{-1}\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\mathbf{X},$$

and

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{B}_{\theta_k}^{-1}\ell_{\theta_k}, \tag{3.3.1}$$

where

$$(\ell_{\theta_k})_i = -\frac{1}{2}\,\mathrm{tr}\{[\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_k)] - (\mathbf{X} - \mathbf{F}\boldsymbol{\beta}_k)'\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_k)\boldsymbol{\Sigma}(\boldsymbol{\theta}_k)^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta}_k)\}.$$

For example, one can start with $\boldsymbol{\theta}$ based on a graphical method and then update $\boldsymbol{\beta}$, or start with $\boldsymbol{\beta}$ as the least squares solution. However, for large n the numerical problem is formidable, and various approximation methods or modifications are put forward (see, for example, Mardia and Marshall, 1984). Also, the likelihood may be multimodal for small samples, causing the scoring method to lead to any stationary point (see Warnes and Ripley, 1987; Mardia and Watkins, 1989). In the above procedure, we also can use the observed Fisher information, since the second derivatives of the likelihood are known (from Section 3.2). Kitanidis and Lane (1985) fully discuss the Gauss–Newton methods (for a general discussion of this topic also see Kitanidis, 1987). However, using the observed information implies calculation of several second derivatives.

A computationally simpler method is given by Vecchia (1988). We can approximate the likelihood function by a partial likelihood function of the form

$$L_m(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|\{X_{im}\}),$$

where $\{X_{im}\}$ is an array consisting of MIN$(i-1, m)$ observations from among $X_1, \ldots, X_{i-1}$ that are closest to X_i, as measured by the distances $|\mathbf{t}_i - \mathbf{t}_j|$. As m approaches n, $L_m(\mathbf{X})$ approaches the likelihood function; but, $L_m(\mathbf{X})$ is very easy to compute for small m. Vecchia (1988) gives an iterative procedure whereby estimates based on $L_1(\mathbf{X})$ are used as initial values of estimates based upon $L_2(\mathbf{X})$, and so on. A statistic is computed at each step of the iterative procedure in order to assess the convergence of the iterative estimates.

3.3.2 Profile likelihood.

One method to check that a solution to the ML equations is really the global maximum is to plot the profile likelihood when the correlation parameters are ≤ 2. Substituting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ from (3.2.6) and (3.2.7) into the log–likelihood (3.1.1), we get the *profile likelihood*

$$\ell_p(\mathbf{X};\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{P}(\theta)| - \frac{1}{2}\log\{[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\mathbf{P}(\boldsymbol{\theta})^{-1}[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]\}, \quad (3.3.2)$$

where

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{F}]^{-1}\mathbf{F}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}.$$

If $\hat{\boldsymbol{\theta}}$ maximizes $\ell_p(\mathbf{X};\theta)$, then $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\hat{\sigma}^2(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ also maximize this likelihood. Usually $\boldsymbol{\theta}$ is a scalar, especially for the stationary case, so that it is relatively easy to plot $\ell_p(\mathbf{X};\boldsymbol{\theta})$ rather than $\ell(\mathbf{X};\boldsymbol{\theta},\sigma^2)$ (see Mardia and Watkins, 1989; cf. Warnes and Ripley, 1987). Also, we then can obtain $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\boldsymbol{\theta}})$ from (3.2.6) and (3.2.7), respectively.

3.4. Analysing topographic data

We now consider the analysis of the topographic data of Table 1. Instead of plotting the sample covariance function, we plot the semi–variogram given by

$$2g(\mathbf{h}) = \Sigma(X_i - X_{i+h})^2/n,$$

where

$$(t_i, t_{i+h}) \in D, \quad \mathrm{t}_1 = |\mathbf{h}|\cos(\theta), \ \mathrm{t}_2 = |\mathbf{h}|\sin(\theta).$$

Figure 2 shows the semi–variogram in four directions, namely $\theta = 0°$, $45°$, $90°$, and $135°$, measured along the t_1 axis in the clockwise direction. Note that if there is a trend then it can be shown that the population semi–variogram, $2\gamma(\mathbf{h}) = E(X_t - X_{t+h})^2$, is proportional to $|\mathbf{h}|^2$ for small $\mathbf{h}$. Thus there is an indication of trend. Also, since the semi–variograms are different for different directions, the data tabulated in Table 1 are not isotropic.

For simplicity, we fit the stationary model with mean $\boldsymbol{\beta}$ and the power covariance function given by (2.2.1). Having the parameters as range α and variance σ^2, we find that [with the asymptotic SE in brackets obtained from (3.2.4)]

$$\hat{\alpha} = 18.6(6.4), \ \hat{\sigma}^2 = 3103.4(1147.7), \ \hat{\beta} = 860.9(33.8),$$

$$\mathrm{CORR}(\hat{\alpha}, \hat{\sigma}^2) \simeq 0.85, \ \log(L) = -244.3.$$

Thus, $\hat{\alpha}$ and $\hat{\sigma}^2$ have large variance and are highly correlated.

The profile log–likelihood for α from (3.3.2) is shown in Figure 3, which has a unique mode at $\hat{\alpha} = 18.6$. Figure 4 shows the contour obtained from the ML predictor (Mardia and Marshall, 1984), which is given by

$$\hat{X}(\mathrm{t}) = \{\mathbf{f}(\mathrm{t})\}'\hat{\boldsymbol{\beta}} + \boldsymbol{\sigma}_X'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}). \quad (3.4.1)$$

Here, we have $\mathbf{f}(\mathrm{t}) = 1$ and $\mathbf{F} = 1$. The contour is similar to Davis (1973, p. 322, Figure 6.9), except that there are slight differences near the edges. The contour indicates that the

Figure 2.

Semi–variogram of topographic data in four directions with fitted power covariance scheme for the stationary case.

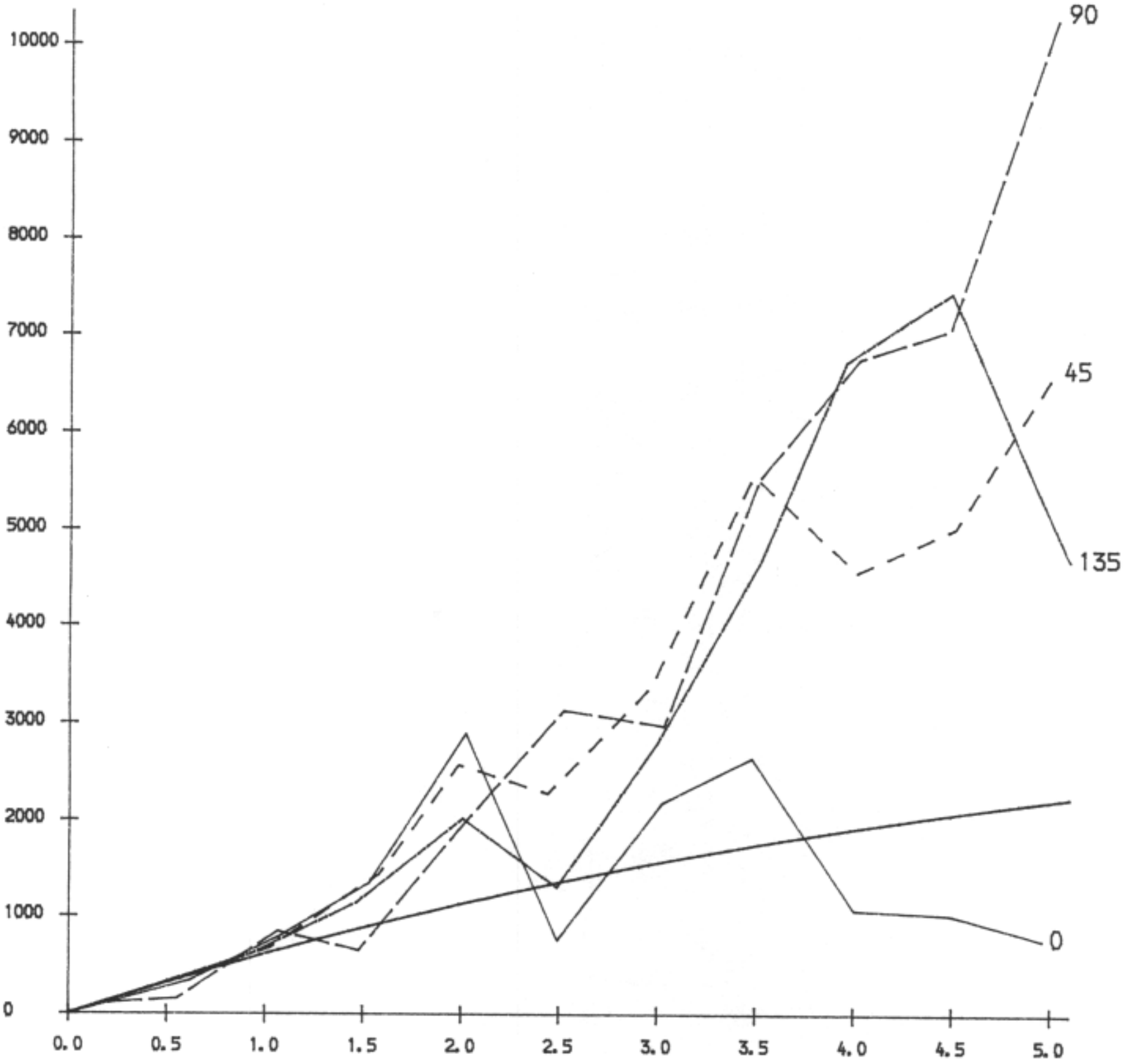


Figure 3.

Profile likelihood under stationary model for the topographic data, with power scheme with range α

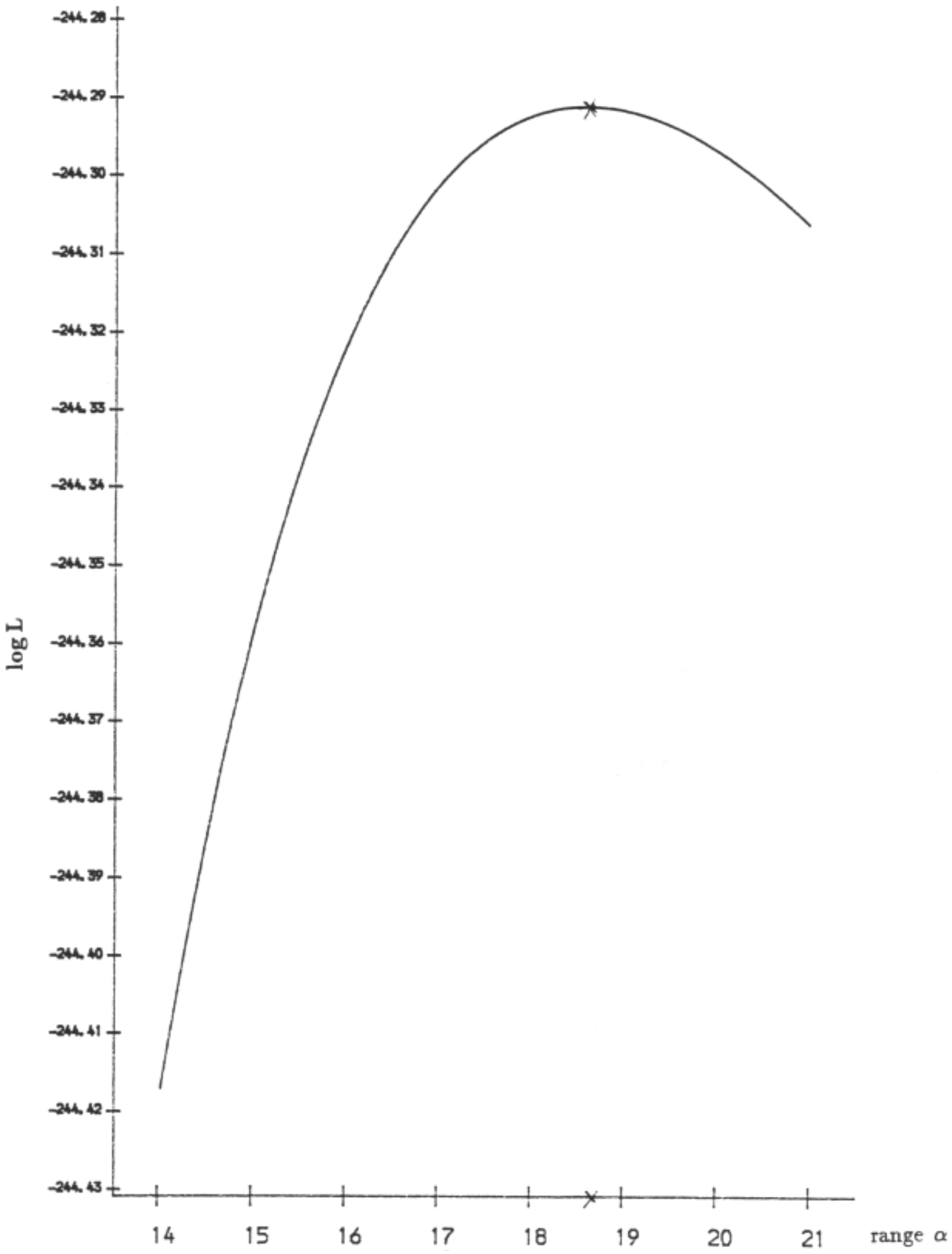
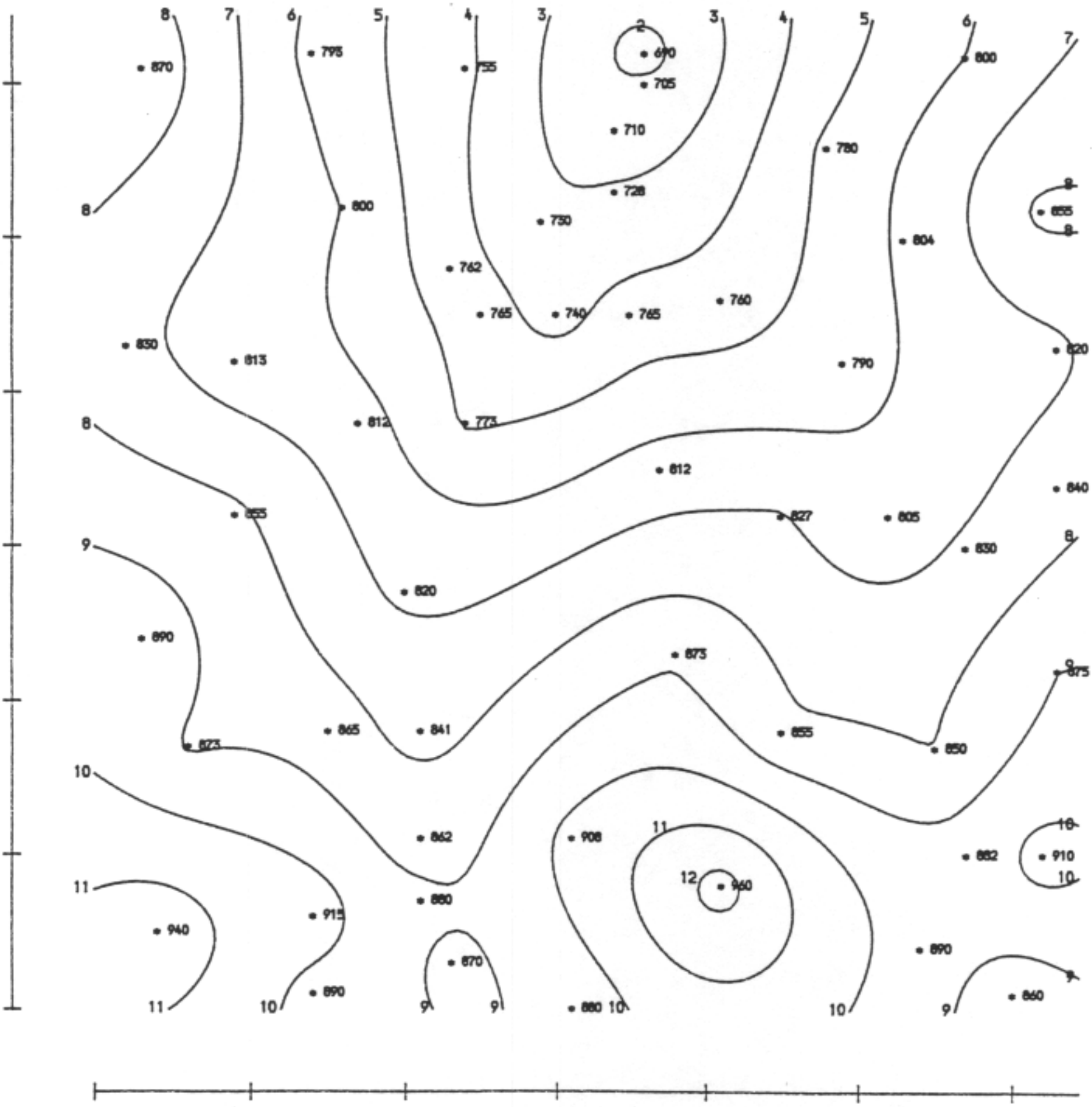


Figure 4.

Contours from the maximum likelihood predictor with stationary model and the long–range correlation. [Contours: $1 = 675(25), \ldots, 13 = 975$.]



data have a basin-shape. Note that the non-differentiability at $\mathbf{h} = \mathbf{0}$ of $\sigma(\mathbf{h}; \boldsymbol{\theta})$ produces spikes when the predictor passes through the data points.

If we fit a quadratic trend, then we find that the MLEs, with asymptotic SE in the brackets, are

$$\hat{\alpha} = 5.2(1.6),\ \hat{\sigma}^2 = 812(225.9),\ \hat{\boldsymbol{\beta}}' = (960.12, -50.38, -19.85, 6.88, 0.28, -0.2), \qquad (3.4.2)$$

which are the coefficients of 1, t_1, t_2, ${t_1}^2$, t_1t_2, and ${t_2}^2$, respectively. Their SEs are (30.2, 13.8, 13.1, 1.2, 1.6, 1.8), and $\mathrm{CORR}(\hat{\alpha}, \hat{\sigma}^2) = 0.71$. The asymptotic correlation matrix of the parameter estimates $\hat{\boldsymbol{\beta}}$ is

$$\begin{pmatrix} +1.000 & -0.708 & -0.655 & +0.429 & +0.614 & +0.350 \\ & +1.000 & +0.253 & -0.888 & -0.446 & +0.081 \\ & & +1.000 & -0.099 & -0.454 & +0.872 \\ & & & +1.000 & +0.078 & +0.077 \\ & & & & +1.000 & +0.060 \\ & & & & & +1.000 \end{pmatrix}$$

The correlations between the regression coefficients of ${t_1}^2$, t_1t_2, and ${t_2}^2$ are very small. The maximum log-likelihood is $\ell = -236.45$. The least squares estimates are

$$\hat{\boldsymbol{\beta}}' = (997.1, -51.9, -30.1, 7.3, 0.4, 0.8),\ \hat{\sigma}^2 = 784.0,$$

which are very similar to those reported in (3.4.2). Figure 5 shows the profile likelihood with quadratic trend, which is also unimodal. Figure 6 shows the ML predictor with the fitted quadratic trend. This is now more similar to the Davis contour plots, even at the edges. However, there is hardly any significant difference between the contour plots for constant trend (Figure 4) and the contour plots for quadratic trend (Figure 6). The stationary case depicts the trend as a long-range correlation, whereas in Case 2 the range is small, depicting small scale variations. Thus there is this a non-identifiability problem in modelling.

However note that the value of the Akaike Criterion $2(-\ell +$ the number of parameters) for the first case is 495, whereas for the second case it is 489, hence indicating that the trend model is better. Figure 7 shows the semi-variogram plots together with the fitted power covariance scheme after removing the trend. On comparison with Figure 2, we again note that the trend model is better. Therefore, both of these factors lead us to recommend the trend model.

We note that for the power scheme, the parameter α does not reflect the true range, unlike for the spherical scheme (see, Section 3.2). A qualitative feel for the observed range can be gained by identifying the values of $\frac{\partial \gamma(h)}{\partial h}$ at $h = 0+$ for the two schemes, which leads to using the true range for the power scheme of $3\alpha/8$ in place of α.

Numerically, one can optimize the profile likelihood first with respect to α through the profile likelihood and then obtain the other estimates from (3.2.6) and (3.2.7), as we have done here.

Figure 5.

Profile for the topographic data with quadratic trend.

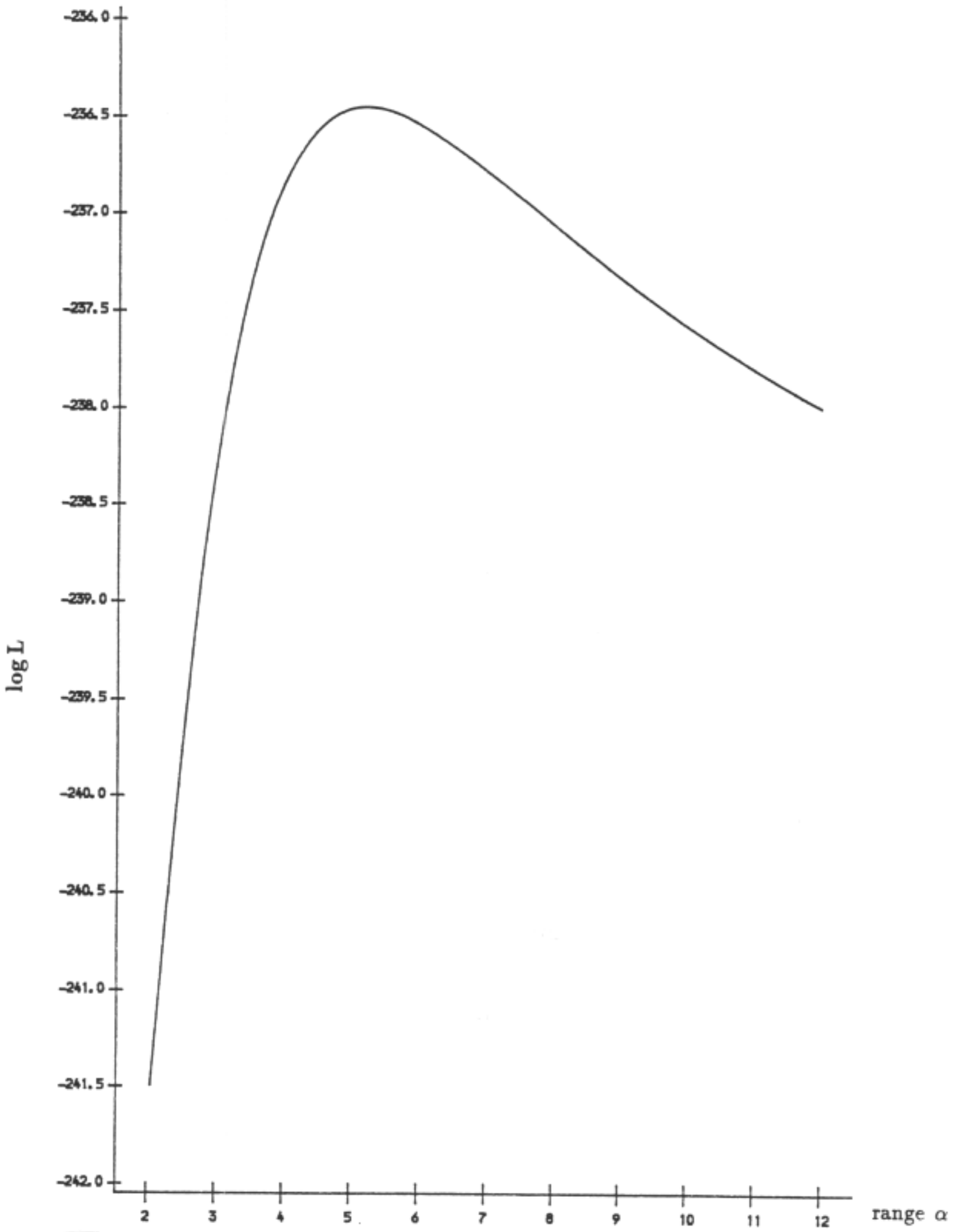


Figure 6.

Contours from the maximum likelihood predictor with quadratic trend for topographic data. [Contours: $1 = 675(25), \ldots, 13 = 975$.]

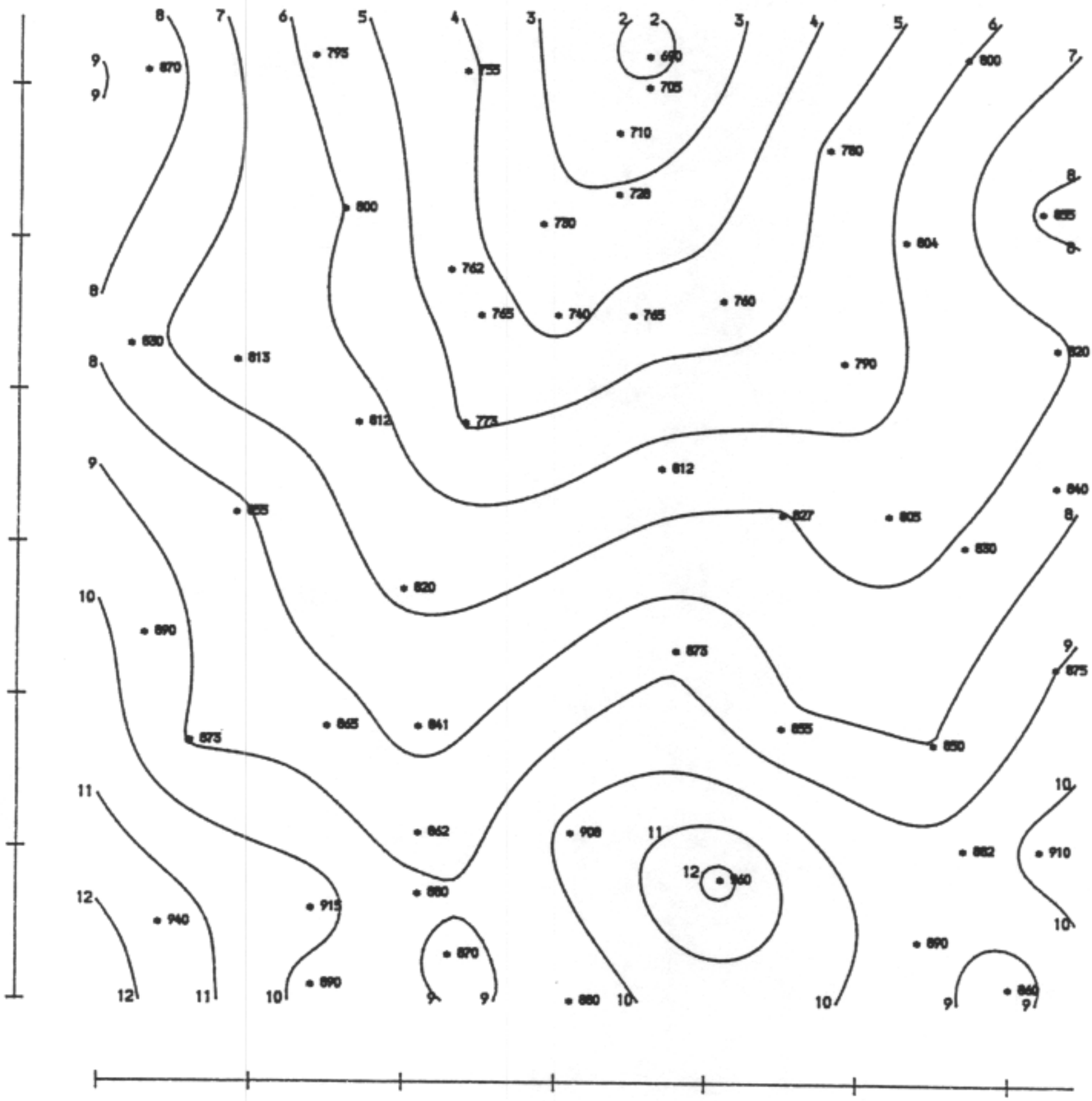
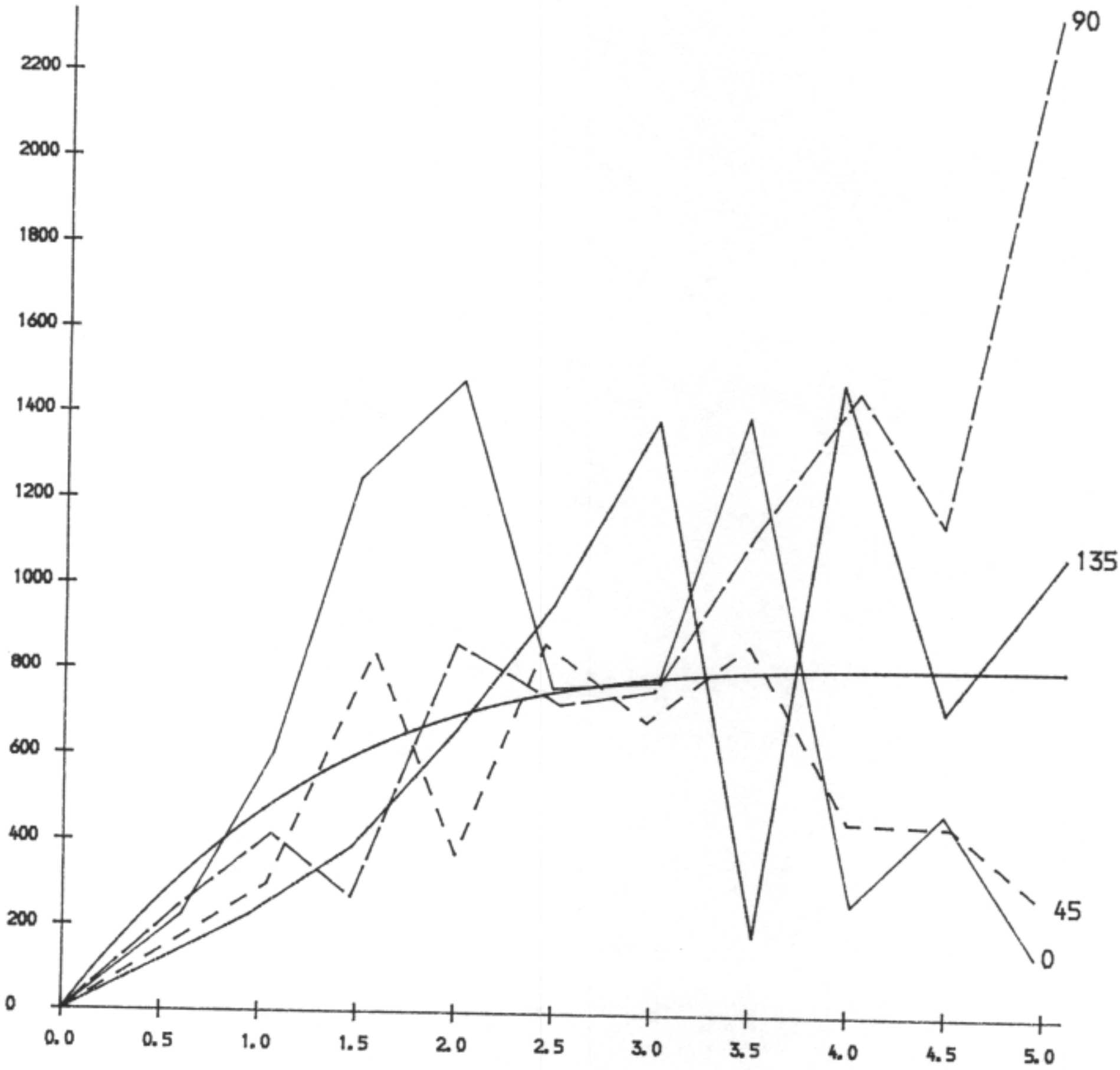


Figure 7.

Semi–variogram after removing the quadratic drift from least squares and fitted power variogram.



3.5. Bias in the estimators

Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ be the MLE of $(\boldsymbol{\beta}, \boldsymbol{\theta})$, then under some mild conditions it can be shown that (Watkins and Mardia, 1989)

$$\begin{aligned} E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= o(n^{-1}), \qquad \text{whereas} \\ E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= (\mathbf{B}_{\theta}^{-1})\mathbf{C}_{\theta} + o(n^{-1}), \end{aligned} \tag{3.5.1}$$

where $(\mathbf{C}_{\boldsymbol{\theta}})_i = (1/2)\mathrm{tr}(\mathbf{B}_{\beta}^{-1}\mathbf{B}_{\beta i}) + (1/2)\mathrm{tr}(\mathbf{B}_{\theta}^{-1}\mathbf{M}_i)$, with $\mathbf{B}_{\beta i} = \frac{\partial \mathbf{B}_{\beta}}{\partial \theta_i}$ and

$$(\mathbf{M}_i)_{jk} = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{ik}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{jk}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}), \tag{3.5.2}$$

for $i, j, k = 1, 2, \ldots, p$, and $(\mathbf{B}_{\beta}, \mathbf{B}_{\theta})$ are defined by (3.2.4). The bias is typically of order $1/n$. We now consider some particular cases.

Case 1: If $\boldsymbol{\theta}$ is known *a priori*, then $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, so that there is no bias.

Case 2: If $q = 0$, $p = 1$, then $\mathbf{M}_1 = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}_1\ \boldsymbol{\Sigma}^1)$ and $\mathbf{B}_{\theta} = (1/2)\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1)$, where $\boldsymbol{\Sigma}^1 = \frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \boldsymbol{\theta}_1}$, so that the bias in (3.5.1) becomes

$$[\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1)]^{-2}\mathrm{tr}(\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}^1). \tag{3.5.3}$$

Case 3: For $p = 1$ with $\boldsymbol{\Sigma} = \theta_1 \mathbf{P}$, we have $\boldsymbol{\Sigma}_1 = \mathbf{P}$ and $\boldsymbol{\Sigma}_{11} = 0$ so that $\mathbf{M}_1 = 0$. But

$$\mathbf{B}_{\beta} = \theta_1^{-1}(\mathbf{F}'\mathbf{P}^{-1}\mathbf{F}) \text{ and } \mathbf{B}_{\beta 1} = -\theta_1^{-2}(\mathbf{F}'\mathbf{P}^{-1}\mathbf{F}),$$

so that $\mathrm{tr}(\mathbf{B}_{\beta}^{-1}\mathbf{B}_{\beta 1}) = -q\theta_1^{-1}$. Also $\mathbf{B}_{\theta} = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_1) = \frac{1}{2}n\theta_1^{-2}$. Hence the bias is

$$-q\theta_1/n,$$

as can be expected, considering the independent and identically distributed (i.i.d) case. There is no bias in $\hat{\boldsymbol{\beta}}$, since it does not depend on θ_1.

Case 4: Consider $q = 0$, $t \in Z^1$, $\sigma(\mathbf{h}) = \lambda^{|\mathbf{h}|}$, and $|\mathbf{h}| \in Z^1$. Then the bias is zero for $\hat{\lambda}$; but, if we take

$$\sigma(\mathbf{h}) = (1 - \lambda^2)^{-1}\lambda^{|\mathbf{h}|},$$

then the bias is $-2\lambda/n$.

This difference in behaviour of the bias can be explained. In the first case, $\sigma(\mathbf{h})$ parameterizes only the correlation structure, whereas in the second case we are taking it as the variance component of the process also. This work can be extended to the doubly geometric scheme given by (2.5.1).

4. Intrinsic models

4.1. Introduction

Consider the simple random walk along a line. Suppose that the steps ε_i are independently distributed as $N(0, \sigma^2), i = 1, 2, \ldots$. Let X_i be the distance covered after the ith step. Then

$$X_i = \varepsilon_1 + \cdots + \varepsilon_i,$$

and $\mathrm{Var}(X_i) = i\sigma^2 \to \infty$ as $i \to \infty$, so the process is non-stationary. However, the process $\{X_{i+h} - X_i\}$ is stationary with $\mathrm{Var}(X_{i+h} - X_i) = |\mathrm{h}|\sigma^2$. Thus, the semi-variogram is defined for all i. The process, which is increment stationary, will be called an intrinsic process of order 0, or the intrinsic random function (IRF) of order 0. In general, we define the intrinsic process of order k below, following Matheron (1971). It should be noted that every stationary process is intrinsic, while the converse is not true. Also, note that the class of variogram schemes for the intrinsic case is larger; compare on the line the semi-variogram

$$\gamma(\mathrm{h}; \theta) = |\mathrm{h}|^\theta, \quad 0 < \theta < 2, \tag{4.1.1}$$

with the covariance function

$$\sigma(\mathrm{h}; \theta) = 1 - |\mathrm{h}|^\theta, \ 0 < \theta < 1. \tag{4.1.2}$$

4.2. Estimation

Let us take $\mathbf{X}$ (n-by-1) as data at the points $\mathbf{t}_i, i = 1, \ldots, n$, from $N_n(\mathbf{0}, \boldsymbol{\Sigma})$, $d = 2$, say. Let

$$\mathbf{H} = \mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}', \tag{4.2.1}$$

where the matrix $\mathbf{T}$ (n-by-p) has its p-columns, $p = (k+1)(k+2)/2$, as

$$\mathbf{t}_{ij} = (\mathrm{t}_1[1]^i \mathrm{t}_1[2]^j, \mathrm{t}_2[1]^i \mathrm{t}_2[2]^j, \ldots, \mathrm{t}_n[1]^i \mathrm{t}_n[2]^j)', \ 0 \le i + j \le k,$$

and $(\mathrm{t}_i[1], \mathrm{t}_i[2])$ denotes the coordinates of the ith site, $i = 1, 2, \ldots, n$. For example, $t_{00} = (1, 1, \ldots, 1)$, and $t_{10} = (\mathrm{t}_1[1], \mathrm{t}_2[2], \ldots, \mathrm{t}_n[1])$. Thus here $\mathbf{f}(\cdot)$ is specified as the full polynomial of degree k. It should be noted that $\mathbf{H}$ is a singular and idempotent matrix of rank $n - p$. Hence

$$\mathbf{Y} = \mathbf{HX} \tag{4.2.2}$$

defines the increment of the order k. Note that

$$Y_i = b_0 - b_{11}t_{1i} - b_{12}t_{2i} - \cdots - b_{kk}t_{1i}^k - \cdots,$$

where $\mathbf{b} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}$.

If $E(\mathbf{X}) = \mathbf{T}\boldsymbol{\beta}$, then $\mathbf{b}$ is the least squares estimator of $\boldsymbol{\beta}$. For IRF-0, a constant is filtered out and we work on $X_i - \bar{X}, i = 1, 2, \ldots, n$, where $\bar{X}$ is the mean of $X_1, X_2, \ldots, X_n$. Note that for IRF-k, polynomials of degree $\le 2k$ will be filtered out. Let $\sigma_k(\mathbf{h})$ be the generalized covariance function for the IRF-k process, defined such that $\mathbf{H\Sigma H}$ is positive definite for all choices of sites and all n. Thus $\{\sigma_k(\cdot) + \mathbf{P}(\cdot) : \mathbf{P}(\cdot)\text{is a polynomial of degree } \le 2k\}$ forms an equivalence class. From (4.2.2) we have $\mathrm{Cov}(\mathbf{HX}) = \mathbf{H\Sigma H} = \mathbf{A}$, say, which is a singular matrix of rank $n - p$. We have

$$\mathbf{HAH} = \mathbf{A}. \tag{4.2.3}$$

Let $\mathbf{A}^-$ be the symmetric generalized inverse of $\mathbf{A}$, such that

$$\mathbf{H}\mathbf{A}^-\mathbf{H} = \mathbf{A}^-. \tag{4.2.4}$$

We will give a construction of $\mathbf{A}^-$ below. We first give an important result for a single parameter $\boldsymbol{\theta}$ in $\sigma(\mathbf{h};\boldsymbol{\theta})$; this result can be extended to the multi-parameter case.

Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the covariance matrix for the above case. Let $\lambda_i(\mathbf{A})$ be non-zero eigenvalues of $\mathbf{A}$. If $L(\boldsymbol{\theta})$ is the likelihood, then

$$-2\log_e[L(\boldsymbol{\theta})] = \text{constant} + \Pi\lambda_i(\mathbf{A}) + \mathbf{X}'\mathbf{A}^-\mathbf{X}, \tag{4.2.5}$$

where $\lambda_i(\mathbf{A})$ are non-zero eigenvalues of $\mathbf{A}$. Further, its derivative with respect to $\boldsymbol{\theta}$ is

$$\text{tr}(\mathbf{A}^-\mathbf{A}_\theta) + \mathbf{X}'\mathbf{A}^-\mathbf{A}_\theta\mathbf{A}^-\mathbf{X}, \tag{4.2.6}$$

where $\mathbf{A}_\theta = \frac{\partial \mathbf{A}}{\partial \boldsymbol{\theta}}$. A proof is as follows. We can find a matrix $\mathbf{D}$ $[(n-p)\text{-by-}n]$ such that

$$\mathbf{D}\mathbf{D}' = \mathbf{I}_{n-p} \quad \text{and} \quad \mathbf{D}'\mathbf{D} = \mathbf{H}, \tag{4.2.7}$$

by constructing an orthogonal matrix $\mathbf{C}$ such that

$$\mathbf{C} = \begin{pmatrix} (\mathbf{T}'\mathbf{T})^{-\frac{1}{2}}\mathbf{T} \\ \mathbf{D} \end{pmatrix}.$$

Set

$$\mathbf{Y} = \mathbf{D}\mathbf{X}; \text{ then} \tag{4.2.8}$$

$$\text{Cov}(\mathbf{Y}) = \mathbf{D}\mathbf{A}\mathbf{D}' = \mathbf{B}, \text{ say,} \tag{4.2.9}$$

which is a non-singular matrix of order $(n-p)$-by-$(n-p)$, and $E(\mathbf{Y}) = \mathbf{0}$. Note that (4.2.8) implies

$$\mathbf{X} = \mathbf{D}'\mathbf{Y} + \text{polynomial of degree k.}$$

Furthermore, from (4.2.3) and (4.2.7) we get

$$\mathbf{A} = \mathbf{D}'\mathbf{B}\mathbf{D} \text{ and } \mathbf{A}^- = \mathbf{D}'\mathbf{B}^{-1}\mathbf{D}, \tag{4.2.10}$$

as $\mathbf{A}^-$ obviously satisfies (4.2.4). Since $\mathbf{B}$ is non-singular, $-2\log_e[L(\boldsymbol{\theta})]$ is

$$\text{constant} + \log(|\mathbf{B}|) + \mathbf{Y}'\mathbf{B}^{-1}\mathbf{Y}, \tag{4.2.11}$$

and its derivative, as given in Section 3, is

$$\text{tr}(\mathbf{B}^{-1}\mathbf{B}_{\boldsymbol{\theta}}) + \mathbf{Y}'\mathbf{B}^{-1}\mathbf{B}_{\boldsymbol{\theta}}\mathbf{B}^{-1}\mathbf{Y}. \tag{4.2.12}$$

Using (4.2.8) and (4.2.10) with (4.2.11) we immediately obtain (4.2.5). From (4.2.7) we can write (4.2.12)

$$\text{tr}(\mathbf{B}^{-1}\mathbf{D}\mathbf{D}'\mathbf{B}_{\boldsymbol{\theta}}\mathbf{D}\mathbf{D}') + \mathbf{Y}'\mathbf{D}\mathbf{D}'\mathbf{B}^{-1}\mathbf{D}\mathbf{D}'\mathbf{B}^{-1}\mathbf{D}\mathbf{D}'\mathbf{Y},$$

which, using (4.2.8) and (4.2.10), leads to (4.2.6).

In a similar fashion, higher derivatives including the information matrix can be written simply by replacing $\mathbf{\Sigma}^{-1}$ by $\mathbf{A}^{-}$ in previous results. It should be noted that we could have used the algebraically "independent" variables out of $\mathbf{Y}$ from (4.2.2), and then use our previous results (see Kitanidis, 1983); but this approach destroys the symmetry achieved in (4.2.5).

Another approach is to assume that $\mathbf{\Sigma}^{-1}$ exists. On exploiting the equivalence of $\sigma_k(\mathbf{h})$, we always can obtain $\mathbf{\Sigma}$ so that $\mathbf{\Sigma}^{-1}$ exists. Then we have explicitly for the likelihood equation

$$\begin{aligned} \mathbf{A}^{-} &= \mathbf{G}'\mathbf{\Sigma}^{-1}\mathbf{G}, \text{ and} \\ \Pi\lambda_i(\mathbf{A}) &= |\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{T}||\mathbf{\Sigma}|/|\mathbf{T}'\mathbf{T}|, \text{ where} \\ \mathbf{G} &= \mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{\Sigma}^{-1}. \end{aligned} \tag{4.2.13}$$

Thus the p.d.f. is proportional to

$$|\mathbf{T}'\mathbf{T}|^{\frac{1}{2}}[|\mathbf{\Sigma}||\mathbf{T}'\mathbf{\Sigma}^{-1}\mathbf{T}|]^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{X}'\mathbf{G}'\mathbf{\Sigma}^{-1}\mathbf{G}\mathbf{X}\}, \tag{4.2.14}$$

where $\mathbf{G}$ is defined by (4.2.13). Note that $\mathbf{G}$ depends on $\mathbf{\Sigma}$, unlike in our preceding work. Nevertheless, $\mathbf{GH} = \mathbf{H}$ and $\mathbf{HG}' = \mathbf{H}$, so that $\mathbf{HA}^{-}\mathbf{H} = \mathbf{A}^{-}$ is satisfied.

The function

$$\sigma_k(\mathbf{h}) = \sum_{p=0}^{k}(-1)^{p+1}|\mathbf{h}|^{2p+1}\frac{a_p}{[(2p+1)!]}\frac{\Gamma(d/2)p!}{\sqrt{\pi}\Gamma\{p+[(d+1)/2]\}}, \tag{4.2.15}$$

defined a covariance function for the IRF-k (Delfiner, 1976) provided the a_ps satisfy

$$\sum_{p=0}^{k} a_p t^{k-p} \geq 0 \text{ for any } t > 0.$$

Thus we have, for example,

$$\sigma_0(\mathbf{h}) = -\theta|\mathbf{h}|,\ \theta > 0, \quad \text{and} \quad \sigma_1(\mathbf{h}) = \theta|\mathbf{h}|^3,\ \theta > 0,$$

defining valid intrinsic covariance functions.

4.3. Regression and the IRF-k

Let

$$X(\mathbf{t}) = \boldsymbol{\beta}_1'\mathbf{f}_1(\mathbf{t}) + \boldsymbol{\beta}_2'\mathbf{f}_2(\mathbf{t}) + \varepsilon(\mathbf{t}). \tag{4.3.1}$$

Suppose that $\boldsymbol{\beta}_1'\mathbf{f}_1(\mathbf{t})$ is a full polynomial of degree k, and $\boldsymbol{\beta}_2'\mathbf{f}_2(\mathbf{t})$ is a polynomial with no terms of degree less than k. Then under an IRF-k model for $\boldsymbol{\varepsilon}(\mathbf{t})$ we have

$$\hat{\boldsymbol{\beta}}_2' = (\mathbf{F}_2'\mathbf{\Sigma}^{-}\mathbf{F}_2)^{-1}\mathbf{F}_2'\mathbf{\Sigma}^{-}\mathbf{X}, \tag{4.3.2}$$

where $\mathbf{F}_2' = [\mathbf{f}_2(\mathbf{t}_1), \mathbf{f}_2(\mathbf{t}_2), \ldots, \mathbf{f}_2(\mathbf{t}_n)]$ since the quadratic form of importance is

$$(\mathbf{X} - \mathbf{F}_2\boldsymbol{\beta}_2)'\mathbf{\Sigma}^{-}(\mathbf{X} - \mathbf{F}_2\boldsymbol{\beta}_2)$$

4.4. Marginal likelihood and IRF-k

Let $\sigma(\mathbf{h};\boldsymbol{\theta})$ be the covariance function. Consider two models.

Model 1:

$$\mathbf{X} \sim N[\boldsymbol{\beta}\mathbf{T},\ \boldsymbol{\Sigma}(\boldsymbol{\theta})];$$

Model 2:

treat $\sigma(\mathbf{h};\boldsymbol{\theta})$ as a covariance function for an IRF-k.

We show that the marginal likelihood under Model 1 with nuisance parameter $\boldsymbol{\beta}$ is the same as the likelihood under Model 2. We have

$$\hat{\boldsymbol{\beta}} = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{T}'\mathbf{X}.$$

Let $\mathbf{Y} = \mathbf{R}\mathbf{X}$, where $\mathbf{R}$ is a singular matrix,

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix},$$

$$\mathbf{R}_1 = \mathbf{H}, \text{ and } \mathbf{R}_2 = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\Sigma}^{-1},$$

where $\mathbf{H}$ is given at (4.2.1). Now $\mathbf{Y} \sim N(\mathbf{R}\mathbf{T}\boldsymbol{\beta},\ \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}')$, with

$$\mathbf{R}\mathbf{T}\boldsymbol{\beta} = (0, \boldsymbol{\beta}')', \text{ and } \mathbf{R}\boldsymbol{\Sigma}\mathbf{R}' = \text{block diag}\,[\mathbf{R}_1\boldsymbol{\Sigma}\mathbf{R}_1', (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}].$$

Let $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$. Note that $\mathbf{Y}_2 = \hat{\boldsymbol{\beta}}$. Further, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent, and

$$\mathbf{Y}_1 \sim N(\mathbf{0},\ \mathbf{R}_1\boldsymbol{\Sigma}\mathbf{R}_1') \text{ and } \mathbf{Y}_2 \sim N[\boldsymbol{\beta},\ (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}]. \tag{4.4.1}$$

Hence the marginal likelihood is the p.d.f. of $\mathbf{Y}_1$, which is precisely the likelihood of the IRF given before. We should note that the marginal likelihood principle is expounded in Kalbfleisch and Sprott (1970); Patterson and Thompson (1975) and Harville (1977) proposed such a procedure in the context of variance components modelling. Tunnicliffe–Wilson (1989) has given the use of this principle for time series analysis.

Our recommendation is that the polynomial of degree greater than k and $\boldsymbol{\theta}$ be estimated first by the marginal likelihood. (The polynomial of degree k is not modelled.) The estimate of $\boldsymbol{\theta}$ then is used to estimate the full polynomial of degree k.

5. Estimation for the CAR model

5.1. The general case

We now consider the CAR model of Section 2.3. The key point is that $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ is given by (2.3.5). Consequently we could write the MLE for $\boldsymbol{\theta}$ given by (3.1.5) in relation to $\boldsymbol{\Sigma}^i = \frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\theta_i}$ using $\boldsymbol{\Sigma}_i = -\boldsymbol{\Sigma}\boldsymbol{\Sigma}^i\boldsymbol{\Sigma}$. Thus the ML equation for $\boldsymbol{\theta}_i$ given by (3.1.5) becomes simply

$$\hat{\mathbf{w}}'\hat{\boldsymbol{\Sigma}}^i\hat{\mathbf{w}} = \text{tr}(\hat{\boldsymbol{\Sigma}}^i\hat{\boldsymbol{\Sigma}}). \tag{5.1.1}$$

Also note for the matrix $\mathbf{A}$ in the information matrix (3.2.4) we can use that

$$2a_{ij} = \text{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^i\boldsymbol{\Sigma}\boldsymbol{\Sigma}^j).$$

We now consider a particular case of Section 2.3 with

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} = (\mathbf{I} - \theta\mathbf{W})/\tau^2, \tag{5.1.2}$$

where $\mathbf{W}$ is a given matrix and $\boldsymbol{\theta}' = (\theta, \tau^2)$ are the parameters. Implicitly, we are neglecting the boundary effects (see Section 2.3). It should be noted that the log–likelihood becomes

$$\text{constant} - \frac{1}{2}\log|\mathbf{I} - \theta\mathbf{W}| - \frac{n}{2}\log(\tau^2) - (\mathbf{X} - \mathbf{F}\boldsymbol{\beta})'(\mathbf{I} - \theta\mathbf{W})(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})/(2\tau^2) \tag{5.1.3}$$

Further, from (3.2.6)–(3.2.8), we obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{F})^{-1}(\mathbf{F}'\mathbf{X} - \hat{\theta}\mathbf{F}'\mathbf{W}'\hat{\mathbf{w}}), \tag{5.1.4}$$

$$\hat{\tau}^2 = \hat{\mathbf{w}}'(\mathbf{I} - \hat{\theta}\mathbf{W})\hat{\mathbf{w}}/n, \tag{5.1.5}$$

and

$$\hat{\mathbf{w}}'\mathbf{W}\hat{\mathbf{w}} = \hat{\tau}^2\text{tr}[\mathbf{W}(\mathbf{I} - \hat{\theta}\mathbf{W})^{-1}], \tag{5.1.6}$$

where $\hat{\mathbf{w}} = \mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}$. In Design of Experiments, $\mathbf{F}'\mathbf{F}$ is of a simple form so that (5.1.4) can be simplified further (see Section 7). Some alternative iterative procedure can be suggested, *e. g.* for given $\hat{\boldsymbol{\beta}}$ (or from the least square estimation), we can obtain $\hat{\theta}$ from (5.1.4) which when substituted into (5.1.5) leads to $\hat{\tau}^2$. Also note that the profile likelihood can be simplified through (3.3.2).

5.2. Properties

5.2.1. Unimodality.

For $\mu = 0$, we can write the likelihood for the basic CAR, *i. e.* with $\boldsymbol{\Sigma} = \mathbf{I} - \theta\mathbf{W}$, as

$$\psi(\theta_1, \theta_2)\exp(\theta_1 T_1 + \theta_2 T_2), \tag{5.2.1}$$

where $\theta_1 = -\frac{1}{2}\tau^2$, $\theta_2 = -\frac{1}{2}\theta\tau^{-2}$, $T_1 = \mathbf{x}'\mathbf{x}$, and $T_2 = \mathbf{x}'\mathbf{W}\mathbf{x}$.

Hence the density (5.2.1) belongs to the canonical exponential family, and therefore various well–known results for this family apply. In particular, except when the sufficient statistics T_1 and T_2 are on the boundary of (θ_1, θ_2), the MLEs of θ_1 and θ_2 exist and are unique. Hence the likelihood will be well–behaved (*e. g.* the log–likelihood will be concave) and unimodal. The exceptional cases are when x_i equals a constant or $\mathbf{x}$ is an eigenvector of $\mathbf{W}$. In the latter case, T_2 is a constant. For an alternative treatment, see Ripley (1988).

For the uniqueness of the MLE for the general CAR on a lattice, see Künsch (1981); for $d = 2$, the proof of concavity of the log–likelihood is simpler.

5.2.2. Asymptotic normality.

The asymptotic normality of the MLEs for the Gaussian CAR follows from Section 3.2. Further, Künsch (1983) proves the following result. Let $f(x)$ be the spectral density, and let, for a fixed subset, $\mathbf{C}^\star$ be the vector of the sample covariance functions with the divisor as the number of terms in the product. Then under certain regularity conditions, we have

$$(2n+1)^{d/2}[\mathbf{C}^\star - E(\mathbf{C}^\star)] \sim N[\mathbf{0}, 2(2\pi)^{-d}\boldsymbol{\Sigma}^\star],$$

where

$$E(\mathbf{C}^*)_h = (2\pi)^{-d} \int \cos(h\omega) f(\omega) d\omega, \text{ and}$$

$$(\mathbf{\Sigma}^*)_{g,h} = \int \cos(g\omega)\cos(h\omega)[f(\omega)]^2 d\omega.$$

We have only considered the stationary case but for the trend case, similar behaviour is expected. For small θ, the CAR and the SAR are similar and therefore the MLEs are expected to behave the same way (see, Cliff and Ord, 1981; Griffith, 1988).

5.3. Other estimators

Another way of estimating the parameters is to use pseudo–likelihood which maximises the conditional probabilities

$$\prod f(\mathbf{x}_i \,|\, \mathbf{x}_j, j \neq i).$$

This leads, after some algebra, to the ordinary least squares estimation. Another approach is to use coding methods, where some sites are coded in a region so that no two encoded sites are to be neighbours of each other. Then the coded variables, given the rest of the sites, are mutually independent, and again we can use the least squares estimation. A simple example is for the first–order CAR on a line, where even sites are encoded given the odd sites, and vice–versa. We then can use the mean from the two estimates as the final estimator (also see Plackett, 1960).

Besag and Moran (1975) have emphasized that the coding technique always provides unbiased estimators and exact tests of significance. Also Besag (1975) has shown that under certain regularity conditions, the pseudo–likelihood estimators are consistent. In general, though, these estimators will not be as efficient as the MLE.

Besag (1977a) has given examples where the pseudo–likelihood estimators are more efficient than the coding estimators. For our discussion, consider the basic CAR given by (2.3.7). The pseudo–likelihood estimators $\{\theta^*, \tau^{*2}\}$ of $\{\theta, \tau^2\}$ can be obtained. It is found that

$$\text{Var}(\theta^*) \sim 2\theta^2[(1-\nu\rho_1)^2/n]\nu\rho_1^2, \quad \text{Var}(\tau^*) \sim \tau^2(1+\nu\theta^2)/n,$$

where ν is the number of neighbours and ρ_1 is the correlation between variates at the neighbouring sites. Further, if f denotes the fraction of sites used in estimation (*i. e.*, coded) then the coding estimators $\tilde{\theta}$ and $\tilde{\tau}^2$ of θ and τ^2 have

$$\text{Var}(\tilde{\theta}) = \theta(1-\nu\theta\rho_1)/(fn\nu\rho_1), \quad \text{Var}(\tilde{\tau}) = 2\tau^2/(fn).$$

Hence the pseudo–likelihood estimators are more efficient than the coding estimators.

5.4. Estimation for the T–CAR

5.4.1. The circle case.

Let x_t be on a circle, $t = 0, 1, \ldots, n-1$, with x_{n-1} being "next to" x_0. Initially let us assume that $\mu = 0$ so that the population covariance function is

$$\sigma_h = E\{x_t \; x_{(t+h)\text{mod}(n)}\},$$

whereas the sample covariance function at lag h is

$$C(h) = n^{-1} \sum_{t=0}^{n-1} x_t \; x_{(t+h) \bmod (n)}.$$

The periodogram of x_t is defined by

$$I(\omega) = \sum_{h=0}^{n-1} \exp(i\omega h) C(h). \tag{5.4.1}$$

Define from $\sigma_h, h = 0, 1, \ldots, n-1$, a circulant matrix $\mathbf{\Sigma}$ with

$$\sigma_{st} = \sigma_{(s-t) \bmod (n)} = \sigma_{(t-s) \bmod (n)}.$$

Then $\mathbf{\Sigma}$ has eigenvalues

$$\lambda_j = \sum_{h=0}^{n-1} \exp\{-2\pi i j h/n\} \sigma_h, \;\; j = 0, 1, \ldots, n-1, \tag{5.4.2}$$

and the eigenvectors $\mathbf{w}_j (n \times 1)$, say, $j = 0, 1, \ldots, n-1$, where

$$\mathbf{w}_j = [(1/\sqrt{n}) \exp\{2\pi j h/n\}, \;\; h = 0, 1, \ldots, n-1]'.$$

Thus we have a spectral decomposition of $\mathbf{\Sigma}, \mathbf{\Sigma} = \mathbf{W \Lambda W^*}$, where $\mathbf{W^*}$ is the complex conjugate transpose of $\mathbf{W}$, has columns $\mathbf{w}_j$, and where $\mathbf{\Lambda} = \text{diag}(\lambda_j)$. Hence it can be shown that the log–likelihood function simplifies to

$$\text{constant} + \frac{1}{2} \sum_{j=0}^{n-1} \log \lambda_j^{-1} - \frac{n}{2} \sum_{j=0}^{n-1} I(2\pi j/n)/\lambda_j. \tag{5.4.3}$$

From (5.4.2) we have the Fourier expansion of σ_h as

$$\Sigma \lambda_j \cos(2\pi j h/n) = \sigma_h. \tag{5.4.4}$$

Using $\mathbf{\Lambda^{-1}} = \mathbf{W \Sigma^{-1} W^*}$, we can express λ_j^{-1} in the form

$$\lambda_j^{-1} = \sum_{h=0}^{n-1} \phi_h \cos(2\pi j h/n), \tag{5.4.5}$$

where ϕ_h from (5.4.3) must satisfy

$$\sigma_h = (1/n) \sum_{j=0}^{n-1} \cos(2\pi j h/n) / \sum_{h=0}^{n-1} \phi_h \cos(2\pi j h/n). \tag{5.4.6}$$

Note that $\phi_h = \phi_{n-h}(= \phi_{-h})$; otherwise ϕ_h are arbitrary parameters with $\lambda_j^{-1} > 0$. If we substitute $I(\cdot)$ and λ_j^{-1} from (5.4.1) and (5.4.5), respectively, into (5.4.3), we obtain

$$n\sum_{j=0}^{n-1}\lambda_j^{-1}I(2\pi j/n) = \sum_{h=0}^{n-1}\phi_h C(h).$$

Hence the log–likelihood from (5.4.3) becomes, except for a constant,

$$\frac{1}{2}\sum_{j=0}^{n-1}\log(\lambda_j^{-1}) - \frac{1}{2}\sum_{h=0}^{n-1}\phi_h C(h). \tag{5.4.7}$$

Now from (5.4.5), $\frac{\partial\lambda_j}{\partial\phi_h} = \cos(2\pi jh/n)$, so that on differentiating (5.4.7) with respect to ϕ_h we obtain the ML equation

$$\Sigma\lambda_j\cos(2\pi jh/n) = C(h).$$

The left–hand side is simply $n\sigma_h$ from (5.4.4), and thus we obtain

$$\hat{\sigma}_h = C(h). \tag{5.4.8}$$

That the MLEs of ϕ_h coincide with the moment estimators of covariance function ϕ_h is an important result. Usually, the number of ϕ_hs is limited by N.

In fact, we could have started with a Gaussian CAR on a circle, with

$$E(X_t|\text{rest}) = \sum_{s\neq t}\theta_t \quad x_{(t+s)\bmod(n)}, \quad \text{Var}(X_t|\text{rest}) = \tau^2. \tag{5.4.9}$$

Then the MLEs of θ_h from (5.4.6) and (5.4.8) are the solutions to

$$C(h) = (1/n)\sum_{j=0}^{n-1}[\cos(2\pi jh/n)]/[\sum_{h=0}^{n-1}\phi_h\cos(2\pi jh/n)], \tag{5.4.10}$$

where $\phi_0 = \tau^{-2}$ and $\phi_h = -\tau^{-2}\theta$, $h \neq 0$. Analogous to (5.4.3), we can approximate the log–likelihood for large n by

$$\text{constant} + \frac{1}{2}\{\int \log[\Sigma\phi_s\cos(s\omega)]^{-1}d\omega - \Sigma\phi_h C(h)\}, \tag{5.4.11}$$

which leads to the approximate ML equation (5.4.8), since

$$C(h) = (2\pi)^{-d}\int[\cos(h\omega)][\sum_s \hat{\phi}_s\cos(s\omega)]^{-1}d\omega, \tag{5.4.12}$$

where the right–hand side is precisely $\hat{\sigma}(h)$. Further, we also can express (5.4.11) as

$$\text{constant} + (n/2)\int \log[f(\omega)]^{-1}d\omega - (n/2)\int[I(\omega)/f(w)]d\omega, \tag{5.4.13}$$

where $f(w) = [\Sigma\phi_s\cos(s\omega)]^{-1}$. The importance of this result will become apparent later on, when we discuss the Whittle approximation.

5.4.2. The torus case.

Consider the discrete torus

$$T = \prod_{\ell=1}^{d} \{0, 1, \ldots, n[\ell] - 1\} \subset Z^d,$$

with opposite faces identified. Write $s = t \bmod(T)$ if $s[\ell] - t[\ell]$ is an integer multiple of $n[\ell]$, for each $\ell = 1, 2, \ldots, d$. Then the preceding section carries over to the torus case directly. For example, the sums over $s = 0, 1, \ldots, n-1$ are now over $s \in T$. The eigenvectors $\mathbf{w}_j, j \in T$, have entries

$$\frac{1}{\sqrt{|T|}} \exp\{2\pi i \sum_{\ell} j[\ell]h[\ell]/n[\ell]\}, \; h \in D.$$

Let σ_h, $h \in T$, be a covariance function; then the covariance matrix $\boldsymbol{\Sigma} = (\boldsymbol{\sigma}_{st})$ has entries $\boldsymbol{\sigma}_{st} = \boldsymbol{\sigma}_{s-t}$. The matrix $\boldsymbol{\Sigma}$ is called a block circulant matrix. We have the eigenvalues

$$\lambda_j = \Sigma_h \exp(-2\pi i \mathbf{j}'\mathbf{h}/n)\sigma_h. \tag{5.4.14}$$

To represent $\boldsymbol{\Sigma}$ it is necessary to arrange the elements of $\sigma_h, \mathbf{h} \in T$, in some order, *e. g.* lexicographic, though the results do not depend upon the order chosen. Again we have a spectral decomposition of $\boldsymbol{\Sigma}$, and with

$$I(\boldsymbol{\omega}) = \sum_{h \in T} \exp(i\boldsymbol{\omega}'\mathbf{h})C(\mathbf{h}), C(\mathbf{h}) = (1/n) \sum_{t, t+h \in T} x_t \; x_{(t+h) \bmod n}. \tag{5.4.15}$$

We now can write the log–likelihood, which is the same as (5.4.3) but with the above modification. In addition, for the Gaussian CAR with the representation similar to (5.4.9),

$$\lambda_j^{-1} = \sum_h \phi_h \cos(2\pi \mathbf{j}'\mathbf{h}/n),$$

and it is found that the MLEs of σ_h are given by

$$\hat{\sigma}_h = C(\mathbf{h}), \tag{5.4.16}$$

where σ_h satisfy an equation similar to (2.3.10). If there are restrictions on ϕ_h (other than $\phi_h = \phi_{-h}$, $\mathbf{h} \neq \mathbf{0}$), then we can modify the ML equations (see below). We also should note that if μ is the population mean, then from $\hat{\mu} = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{X}/\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}$ we get

$$\hat{\mu} = \bar{x},$$

as $\mathbf{1}$ is an eigenvector of $\boldsymbol{\Sigma}$. Also note that the ML equation and estimate also can be written from the general case, since

$$\boldsymbol{\Sigma} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^{\star} \text{ and } \boldsymbol{\Sigma}^{-1} = \mathbf{W}\boldsymbol{\Lambda}^{-1}\mathbf{W}^{\star}, \text{ with}$$
$$|\boldsymbol{\Sigma}| = \prod \lambda_i \text{ and } \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} = (\mathbf{W}^{\star}\mathbf{X})'\boldsymbol{\Lambda}^{-1}(\mathbf{W}\mathbf{X})^{\cdot}$$

The key point is that *only* λ_i depends on the parameters. Thus, for example, we can write for the Fisher information matrix for $\boldsymbol{\theta}$ (or ϕ_h), from (3.2.4),

$$(\mathbf{B}_\theta)_{ij} = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j) = \frac{1}{2}\sum_{r=1}^{n}\frac{1}{\lambda_r^2}\frac{\partial\lambda_r}{\partial\theta_i}\frac{\partial\lambda_r}{\partial\theta_j}. \tag{5.4.17}$$

We also note that we can wrap (unnaturally) on the torus if there is a trend (see Mardia and Marshall, 1984).

We now consider the Gaussian first–order T–CAR with $d = 2$ given by

$$E(X_{rs}\,|\,\text{rest}) = \mu + \theta(x_{r-1,s} + x_{r+1,s} + x_{r,s-1} + x_{r,s+1} - 4\mu),$$

where X_{rs} is the (r,s)-th observation on a torus, and $|\theta| < 1/4$. We have $\hat{\mu} = \bar{x}$.

Let $\mathrm{Var}(X_{rs}|\text{rest}) = \tau$, and let us write

$$f(\theta) = (2\pi)^{-2}\int_{(-\pi,\pi)^2}\{1 - 2\theta[\cos(x) + \cos(y)]\}^{-1}dx\,dy.$$

The ML equations for $\hat{\tau}$ and $\hat{\theta}$ are $\hat{\sigma}(0,0) = C(0,0)$, and $\hat{\sigma}(1,0)+\hat{\sigma}(0,1) = C(0,1)+C(1,0)$. For the circle, the ML equations for $\hat{\tau}$ and $\hat{\theta}$ are

$$\hat{\tau} = C(0,0) - 2\hat{\theta}[C(1,0) + C(0,1)], \text{ and } \{C(0,0) - 2\hat{\theta}[C(1,0) + C(0,1)]\}f(\hat{\theta}) = C(0,0).$$

The reason is that for the T–CAR we have $\sigma(0,0) - 2\theta[\sigma(0,1) + \sigma(1,0)] = \tau$.

We also can simplify the information matrix of (μ, τ, θ) from (5.4.14) and (5.4.17) (see Besag and Moran, 1975) for the simplified expressions and their numerical calculations.

5.5. The Whittle approximation

Let us consider a lattice for the stationary case with zero mean. Consider data (X_t) available in a region $D \subset Z^d$, where D is typically rectangular. Let $\{\sigma_h, \mathbf{h} \in Z^d\}$ be a covariance function with spectral density $f(\boldsymbol{\omega})$, $\boldsymbol{\omega} \in (-\pi,\pi)^d$. Then, motivated by the torus case [see equation (5.4.13) for the circle case], we consider a spectral approximation to the log–likelihood ℓ given by

$$2\ell = \text{constant} + n\int_{(-\pi,\pi)^d}\log[f(\boldsymbol{\omega})^{-1}]d\boldsymbol{\omega} - n\int_{(-\pi,\pi)^d}[I(\boldsymbol{\omega})/f(\boldsymbol{\omega})]d\boldsymbol{\omega}, \tag{5.5.1}$$

where

$$I(\boldsymbol{\omega}) = \sum_h \exp(i\mathbf{h}'\boldsymbol{\omega})C(\mathbf{h}),\ f(\boldsymbol{\omega}) = (2\pi)^{-d}\sum\exp(i\mathbf{h}'\boldsymbol{\omega})\sigma_h, \tag{5.5.2}$$

and $C(\mathbf{h})$ is some estimate of the covariance at lag h. We now outline some estimates.

Let $D_h = \{t : (t, t+h) \in D\}$. Whittle (1954) recommended using

$$C_W(\mathbf{h}) = \frac{1}{|D|}\sum_{\mathbf{t}\in D_h} x_t x_{t+h},\ \mathbf{h} \in D;\ = 0, \mathbf{h} \notin D,$$

which essentially amounts to taking $X_t = 0$ outside D. It is noted by Guyon (1982) that this leads to bias asymptotically in estimating the MLE. He recommends using

$$C_G(\mathbf{h}) = \frac{1}{|D_h|} \sum_{\mathbf{t} \in D_h} x_t x_{t+h}, \mathbf{h} \in D; \ = 0, \mathbf{h} \notin D,$$

where $|D_h| = \prod (n_i - |h_i|)^{-1}$. While this modification removes the bias in the MLE, $I(\omega)$ now can be less than zero. Its effect on the variance remains unclear.

Dahlhaus and Künsch (1987) recommend using $C_K(\mathbf{h})$ in place of $C(\mathbf{h})$ in (5.5.2), where

$$C_K(\mathbf{h}) = \frac{\sum_{(t,\, t+h) \in D} x_t x_{t+h} w_t w_{t+h}}{\sum_{(t,\, t+h) \in D} w_t^2},$$

with

$$w_t = \prod_{i=1}^{d} u[(t_i - \frac{1}{2}) n_i],$$

and

$$u(y) = w(2y/\rho), \ 0 \le y \le \rho/2; \ = 1, \ \rho/2 \le y \le 1/2; \text{ and, } = u(1-y), \ 1/2 \le y \le 1,$$

where ρ is a smoothing parameter. The tapering pulls the data toward zero near the boundary, while keeping exactly the same value at the centre. A common taper in time series analysis is the Tukey–Hanning taper, with $w(u) = [1 - \cos(u\pi)]/2$. Dahlhaus and Künsch (1987) have shown that the bias is asymptotically negligible in estimating θ by the MLE using this approximation, and the estimator is asymptotically efficient. Also $I(\omega) > 0$.

For the M–CAR we have from (2.3.10), and $f(\omega) = 1/\sum_{s \in N_0} \phi_s \cos(\omega' \mathbf{h})$. Thus the Whittle approximation (5.5.1) leads to the moment estimates $\hat{\sigma}_h = C(\mathbf{h})$, where σ_h is given by (2.3.10). It should be noted that $C_W(\mathbf{h})$ and $C_G(\mathbf{h})$ use the free boundaries of the C–CAR, and therefore asymptotically they may experience some loss of efficiency. Again $C_K(\mathbf{h})$ is recommended.

Assuming that the data are from the infinite CAR, then the MLE based on the C–CAR and the T–CAR will lead to a biased estimator, in general, unless these are adjusted as above. This problem does not arise for the estimates by the M–CAR.

5.6. The intrinsic CAR (IRF-0)

Let $f(\omega)$ be the spectral density of a process. For intrinsic processes we have

$$\int_{(-\pi,\pi)^d} f(\boldsymbol{\omega}) d\boldsymbol{\omega} = \infty \text{ and } \int_{(-\pi,\pi)^d} |\boldsymbol{\omega}|^2 f(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty.$$

Thus for the CAR given by (2.3.5) and (2.3.9), we have

$$\sum_{h \in N} \theta_h = 1 \text{ or } \sum_{h \in N} \phi_h = 0. \tag{5.6.1}$$

Given this restriction, we can proceed as before (see, Künsch, 1987). For example, consider the approximate log–likelihood given by (5.4.11), which under restriction (5.6.1) leads to

$$\hat{\gamma}_h = g(\mathbf{h}), \tag{5.6.2}$$

where $g(\mathbf{h})$ is the sample semi–variogram and γ_h is the population semi–variogram on an infinite lattice, given by

$$\gamma_h = (2\pi)^{-d} \int [1 - \cos(\mathbf{h}'\boldsymbol{\omega})]/\{\sum_s \phi_s[1 - \cos(\mathbf{s}'\boldsymbol{\omega})]\} d\boldsymbol{\omega}. \tag{5.6.3}$$

Since in the intrinsic case the generalized covariance function is defined only up to an additive constant, equation (5.6.2) makes sense. It simply is an expression of the moment estimators

$$\sigma_{(\mathbf{0})} - \sigma_h = C(\mathbf{0}) - C(\mathbf{h})$$

from (5.4.8), when $\sigma(\mathbf{0})$ exists. Of course, this is completely analogous to (5.4.11). Further, equation (5.6.2) is exact for the T–CAR.

A particular "intrinsic" CAR of interest for the finite lattice is

$$E(X_i|\text{rest}) = \bar{x}_i \text{ and } \text{Var}(X_i|\text{rest}) = \tau^2/\nu_i, \tag{5.6.4}$$

where ν_i is the number of neighbours of i, and $\bar{x}_i$ is the mean of the neighbouring values of i. This specification suggests a neat way of defining boundary corrections for finite lattices (Besag, 1989). It has the joint density (Künsch, 1987)

$$\text{constant } \tau^{-n} \exp[-\frac{1}{2\tau^2} \sum_{i \sim j} (x_i - x_j)^2], \tag{5.6.5}$$

where the sum is over all $i \sim j$ that are neighbours. A practical example appears in Kent and Mardia (1988). Note that the density is singular, but the MLE of τ^2 is straightforward and it is clearly well–behaved. For the infinite lattice, this becomes a particular case of the basic CAR defined by (2.3.7).

5.7. Landsat data

We consider Switzer's Landsat data of three rock–types on one of four spectral bands (namely, red). The assumption of stationarity is inadequate since the mean depends upon the rock type. The simple model that x_i is a function only of rock type plus white noise also is not adequate, since the values in two regions of the same rock type differ considerably; this also may be due to some blurring and other features of the ground, such as texture and orientation, contributing to the signal. It seems that fitting an overall trend or fitting an intrinsic model may prove adequate. The data are on a 16–by–25 lattice, and have been analyzed by Künsch (1987) through IRF-0. Table 2 shows the MLE with various neighbourhood schemes, utilizing a Whittle–type approximation to the IRF.

TABLE 2
ESTIMATED PARAMETERS OF AND INTRINSIC CAR OF ORDER 0 FOR LANDSAT DATA

Number of Neighbors	$\hat{\theta}_h$ with h specified (0,1)	(0,2)	(1,−1)	(1,0)	(1,1)	(2,0)	$\hat{\tau}^2$	400L
4	0.346			0.154			7.35	931.7
8	0.385		−0.014	0.267	−0.138		6.05	905.0
12	0.442	−0.057	−0.002	0.206	−0.104	0.015	5.74	897.8

Table 2 gives estimated parameters for the first, second, and third–order neighbourhoods, based upon solving (5.6.2) numerically. It also gives $L = -(2) \times (400) \times$ log–likelihood, which has been approximated, where 400 is the number of observations. There is a clear anisotropy in the data, since $\hat{\theta}_{1,1}$ and $\hat{\theta}_{1,-1}$ are different.

One should note that the Akaike criterion,

$$L + 2 \times \text{(the number of parameters)},$$

gives values for $400L$ of

$$937.7 \quad 915.0 \quad 911.8.$$

These values indicate that it is better to use at least the second–order neighbourhoods. There is only a slight gain in using the third–order neighbourhood.

We now check how the model fits the data and how it compares with the corresponding stationary models. If we use the third–order neighbourhood, the sum of regression coefficients $\Sigma\theta_h = 0.9954$, so that we are on the boundary of the parameter space (*i. e.*, it indicates an intrinsic model). Figure 8 shows the theoretically fitted combined semi–variograms for (i) an intrinsic CAR with third–order neighbours, (ii) an ordinary CAR with third–order neighbours, and (iii) an empirical variogram. Figure 8 also shows that the agreement is good for both models, for small lags, while the intrinsic model fits better for larger lags.

These findings imply that the simple discrimination technique using a constant mean for each rock type cannot work; but, an intrinsic model with constant mean in each region is plausible. Kent and Mardia (1988) have given another analysis of the same data.

6. The errors in variable model

6.1. The direct model case

Consider

$$\text{Cov}(\mathbf{X}) = \sigma^2\mathbf{P} + \psi^2\mathbf{I}, \tag{6.1.1}$$

where $\mathbf{P}$ is a correlation matrix, or

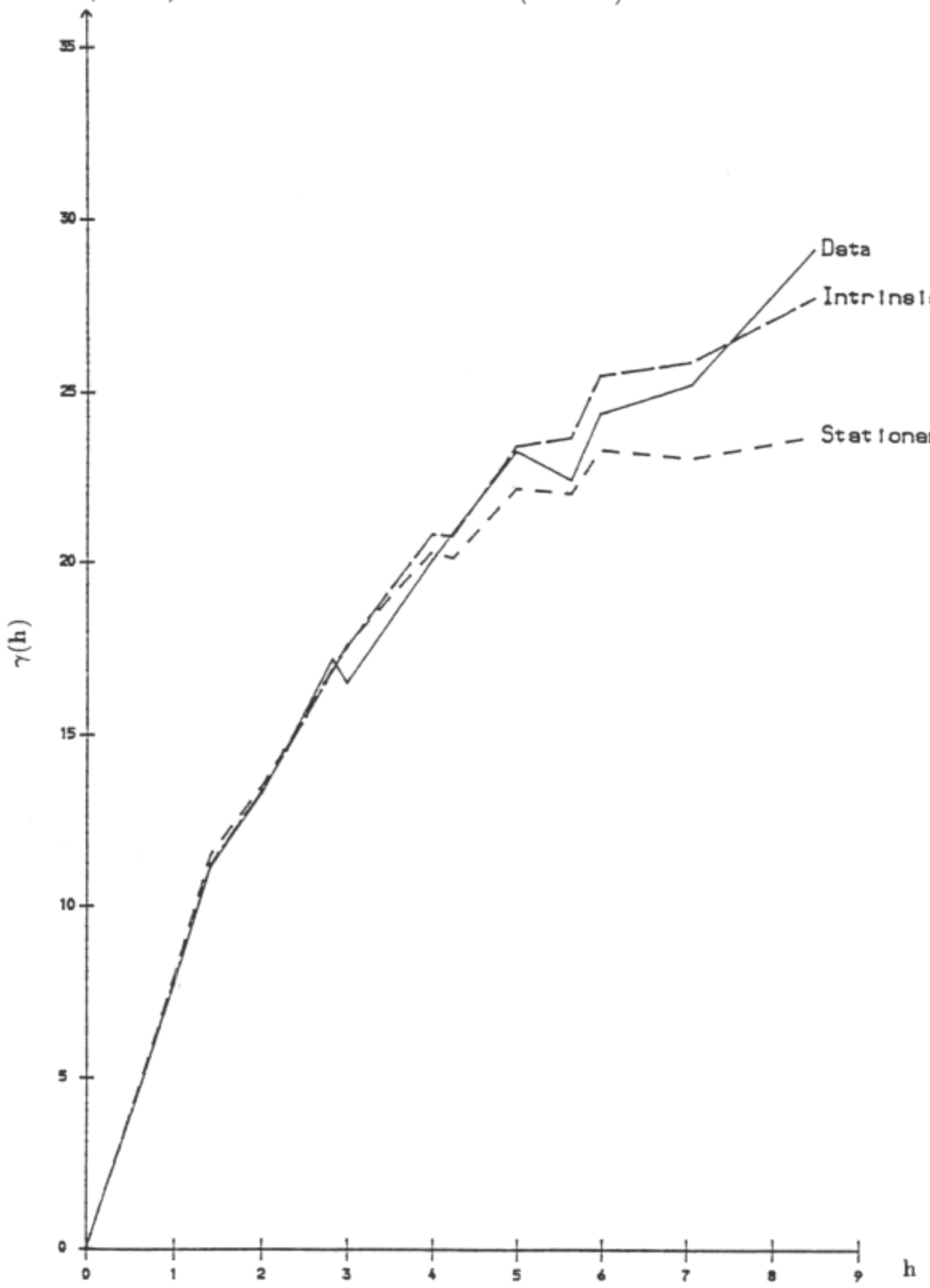
$$\text{Cov}(X_i, X_j) = \sigma^2\rho_{ij} + \psi^2\delta_{ij}, \ \rho_{ij} = \rho(\mathbf{t}_i - \mathbf{t}_j),$$

with

$$\rho(0) = 1, \text{ and } \delta_{ij} = 1 \text{ if } i = j; \ = 0 \text{ if } i \neq j.$$

Figure 8.

Semi-variogram for the Landsat data (———), fitted variograms for 3rd order stationary CAR (- - - -) and for 3rd order intrinsic CAR (— — —).



We will call ψ^2 a nugget parameter.

For the lattice case, we could write the underlying process as

$$X(\mathbf{t}) = \varepsilon(\mathbf{t}) + \eta(\mathbf{t}), \tag{6.1.2}$$

with $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2\rho(\mathbf{t}_i - \mathbf{t}_j)$, $\rho(0) = 1$, $\text{Var}[\varepsilon(\mathbf{t})] = \psi^2$, and $\text{Cov}\{\varepsilon(\mathbf{t}), \eta(\mathbf{t})\} = 0$.

Hence (6.1.2) is an errors–in–variable model, where $\varepsilon(\mathbf{t})$ is uncontrollable error and $\eta(\mathbf{t})$ is measurement error. Representation (6.1.2) does not make sense for a continuous process; but (6.1.1) is still meaningful where the nugget parameter ψ^2 represents the error, depending upon the choice of the sampling frame (*i. e.*, the sites $\mathbf{t}_i$). The point to bear in mind here is that for continuous processes, the "white noise" term $\eta(\mathbf{t})$ in (6.1.2) is not meaningful. Note that if $\rho(\mathbf{h};\theta)$ is the autocorrelation function of the process, where θ denotes the range parameter (see Section 2.2), then white noise can appear with $\theta \to 0$ or $\sigma^2 \to 0$.

We have described the model with three parameters, namely nugget, range and sill $(\psi^2 + \sigma^2)$. We can easily incorporate the trend into the model. Also, we can extend these results in order to incorporate anisotropy; Brewer and Mead (1986) have used, in addition to the nugget,

$$\sigma(\mathbf{h}) = \exp[-(\mathbf{h}'\mathbf{G}^{-1}\mathbf{h})^{\beta/2}],\ 0 \le \beta \le 2,$$

where $\mathbf{G}$ is a symmetric matrix. This scheme models geometric anisotropy. Note that $\sigma(\mathbf{h})$ can be easily seen to be a covariance function, since it can be related to the characteristic function of a multivariate distribution (see Mardia, 1986). Brewer and Mead (1986) provide an intuitively plausible method to estimate the parameters, and it remains to be seen how far the exact MLEs succeed in capturing a similar behaviour.

We now describe a computational procedure.

6.2. A profile likelihood

Extending the profile likelihood approach of Section 3.3 is worthwhile for numerical work. Let us now take the vector $\boldsymbol{\theta}$ with just three parameters, and write

$$\sigma^2,\ \delta = \psi^2/\sigma^2,\ \text{and}\ \theta,$$

where we now assume that θ is a correlation parameter in $\mathbf{P}(\theta)$. For a given θ we can use the spectral decomposition of $\mathbf{P}(\theta)$, so that

$$\mathbf{P}(\theta) = \mathbf{C}(\theta)'\boldsymbol{\Lambda}(\theta)\mathbf{C}(\theta), \tag{6.2.1}$$

where $\mathbf{C}(\theta)$ is an orthogonal matrix, and

$$\boldsymbol{\Lambda}(\theta) = \text{diag}\{\lambda_1(\theta), \lambda_2(\theta), \ldots, \lambda_n(\theta)\}.$$

As before, suppose the "trend" is $\mathbf{F}\boldsymbol{\beta}$. Then the log–likelihood from equation (3.1.1) is simply

$$\ell = -(n/2)\log(\sigma^2) - (1/2)\sum_{i=1}^{n}\log[\delta + \lambda_i(\theta)] - \sum_{i=1}^{n} u_i^2(\theta, \boldsymbol{\beta})/\{(2\sigma^2)[\delta + \lambda_i(\theta)]\}, \tag{6.2.2}$$

where

$$\mathbf{u}(\theta, \boldsymbol{\beta}) = \mathbf{C}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta}). \tag{6.2.3}$$

It can be shown, as in Mardia (1980), that

$$\hat{\sigma}^2(\delta, \theta, \boldsymbol{\beta}) = \{\sum_{i=1}^{n} u_i^2[\delta + \lambda_i(\theta)]^2\} / \sum_{i=1}^{n} [\delta + \lambda_i(\theta)]^{-2}. \tag{6.2.4}$$

The ML equations for δ and $\boldsymbol{\beta}$ do not depend on σ^2, and are

$$\hat{\boldsymbol{\beta}}(\theta, \delta) = [\mathbf{F}'\mathbf{C}(\boldsymbol{\Lambda} + \delta\mathbf{I})^{-1}\mathbf{C}\mathbf{F}]^{-1}\mathbf{C}'(\boldsymbol{\Lambda} + \delta\mathbf{I})^{-1}\mathbf{C}\mathbf{X}, \tag{6.2.5}$$

and

$$(1/n)\Big[\sum_{i=1}^{n} u_i(\theta, \hat{\boldsymbol{\beta}})^2[\hat{\delta} + \lambda_i(\theta)]^{-1}\Big]\Big[\sum_{i=1}^{n} [\hat{\delta} + \lambda_i(\theta)]^{-2}\Big] = \sum_{i=1}^{n} u_i(\theta, \hat{\boldsymbol{\beta}})^2[\hat{\delta} + \lambda_i(\theta)]^{-2}. \tag{6.2.6}$$

Now in substituting $\hat{\boldsymbol{\beta}}$ from (6.2.5), for a given θ, a solution can be obtained in terms of $\delta(\theta)$. In turn, we substitute δ into (6.2.5) to obtain $\hat{\boldsymbol{\beta}}(\theta, \hat{\delta})$, and then from (6.2.4) we get $\hat{\sigma}^2 = \sigma^2(\hat{\delta}, \theta, \hat{\boldsymbol{\beta}})$. Thus the profile likelihood with respect to θ can be obtained from (6.2.2). Hence, the ML estimate of θ can be derived, which in turn gives the MLEs of δ, $\boldsymbol{\beta}$, and σ^2. Alternatively we could work as we did in Section 3.3, so that the profile likelihood is a function of θ and δ, since δ is not eliminated. It also should be noted that for IRF-k, equation (6.2.5) does not exist. Therefore, following Section 4.2, we must first solve equations similar to (6.2.6) for δ, and then we can obtain $\hat{\sigma}^2(\theta, \delta)$ from an equation similar to (6.2.4). Next, from an equation similar to (6.2.2) we obtain the profile likelihood of θ, then $\hat{\theta}$, and finally the MLEs of σ^2 and δ.

Example. Figure 9 shows bauxite grade values in two dimensions for a region of Southern France (Marechal and Serra, 1970). The empirical, combined semi–variogram given in Figure 10 shows that there is a possible nugget. There is a peak at about $|\mathbf{h}| = 7$, which reflects the depression in the centre of the data. We fitted a power scheme with constant mean, nugget ψ^2, variance σ^2, and range α. The profile log–likelihood for α and δ is shown in Figure 11 and appears to be quite flat. Figure 10 also shows the fitted semi–variogram. It should be borne in mind that the contours are not equi–spaced. The accompanying parameter estimates are found to be

$$\hat{\mu} = 14.31, \ \hat{\delta} = 0.278, \ (\hat{\psi}^2 = 24.51), \ \hat{\sigma}^2 = 88.18, \ \hat{\alpha} = 8.12, \ \text{and} \ \log(\ell) = -119.97.$$

Figure 12 shows contours from the ML predictor that clearly captures the depression in the data.

6.3. The CAR model case

For large scale data the above computational method is not feasible. We again could use the Whittle approximation, which in turn relies on the torus approximation (Besag, 1977b). Let us write $\boldsymbol{\Sigma}_0 = \sigma^2\mathbf{P}$. Let

$$\lambda_j = 1/\sum_{h} \phi_h \cos(2\pi \mathbf{j}'\mathbf{h}/n);$$

Figure 9.

Spatial distribution of Bauxite grades (Marechal & Serra, 1970).

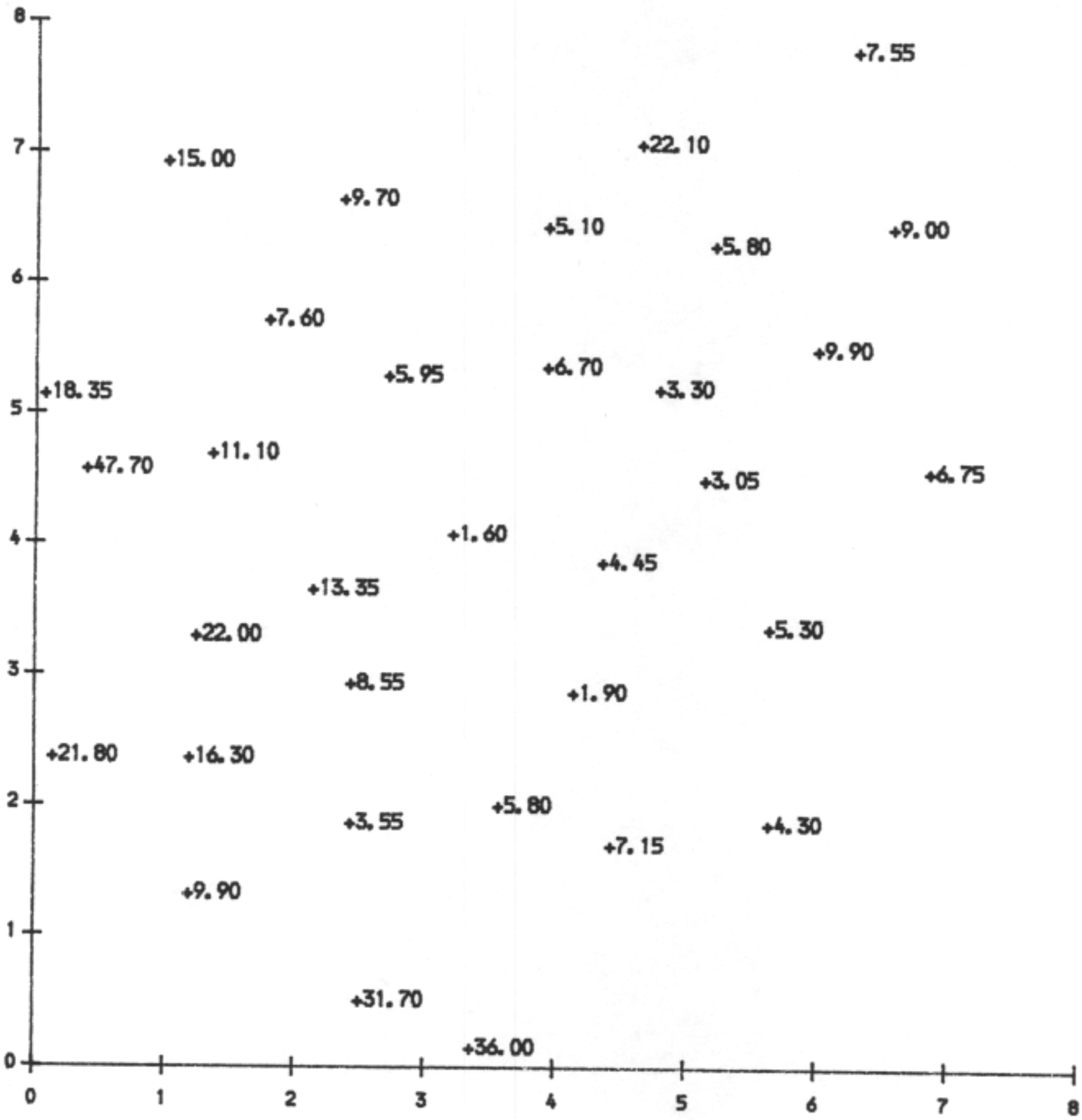


Figure 10.

A fitted isotropic variogram to the Bauxite data reflecting a nugget effect.

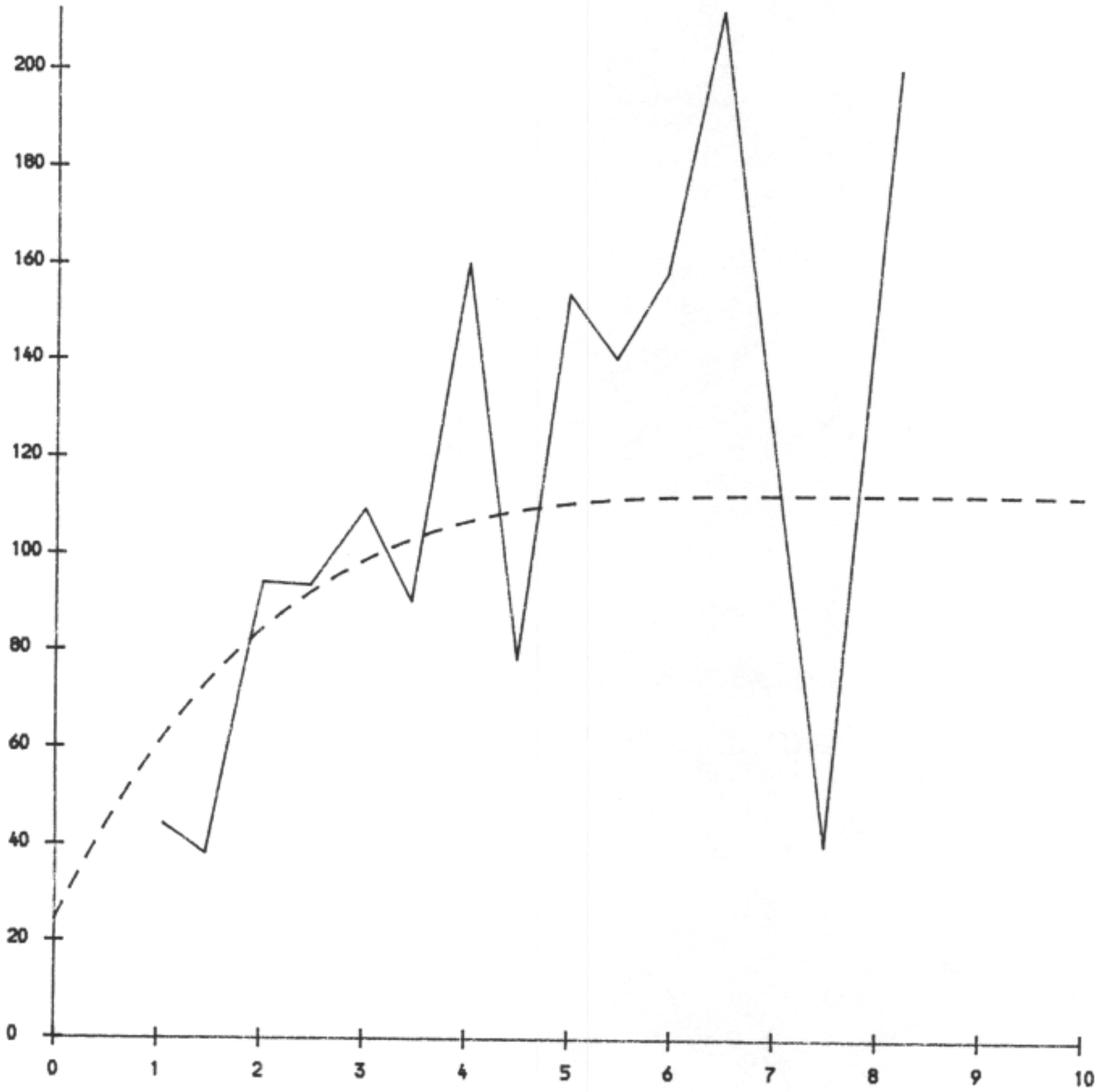


Figure 11.

Profile likelihood with nugget effect for the Bauxite data. [Contours:

$$1 = 123.0,\ 2 = -122.0(-0.5), \ldots, 5 = -120.50,\ 6 = -120.4(-0.1), \ldots, 8 = -120.20,$$

$$9 = -120.09(-0.02), \ldots, 13 = -120.01,\ 14 = -120.00(-0.01), \ldots, 17 = -119.97.]$$

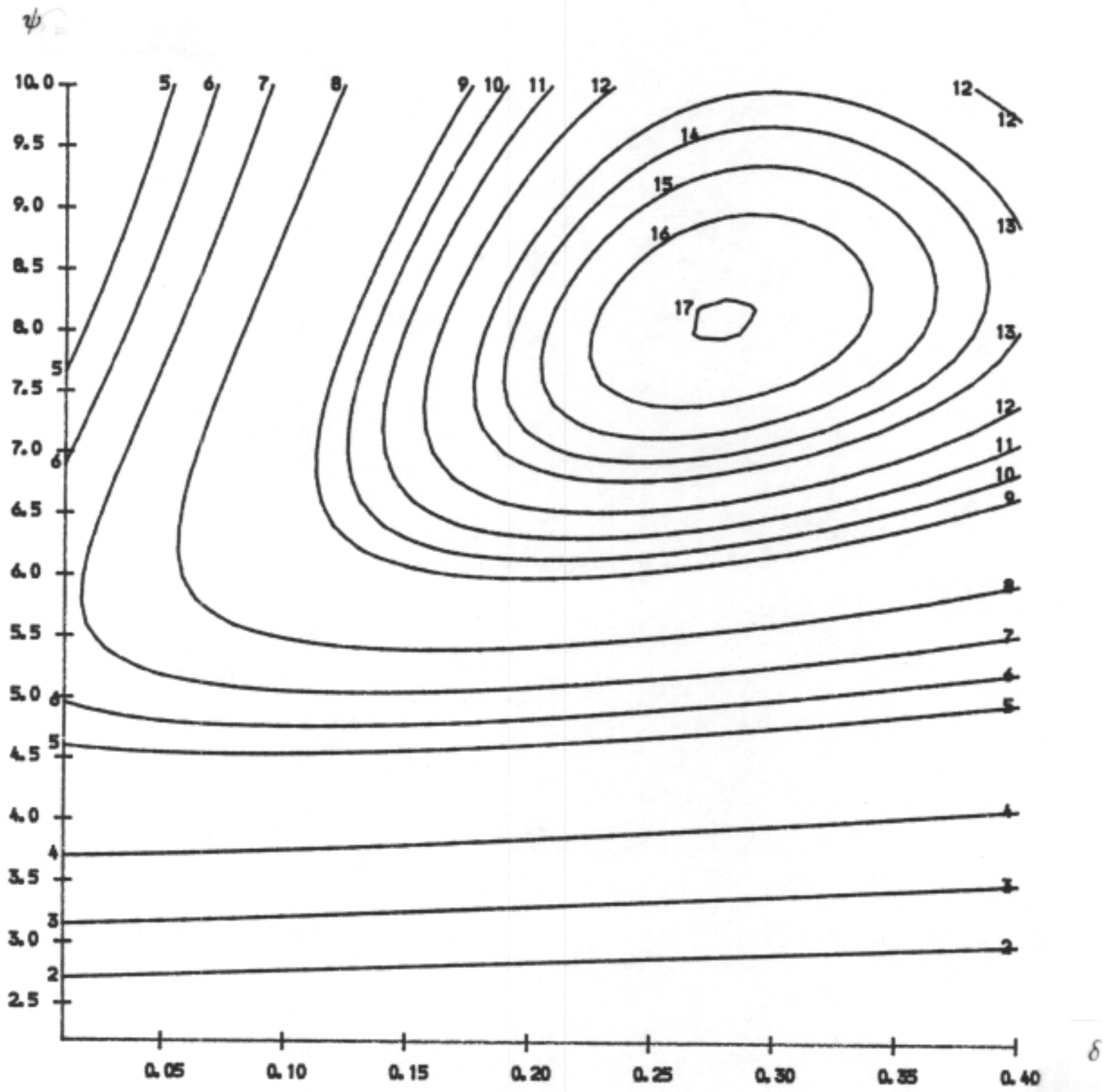
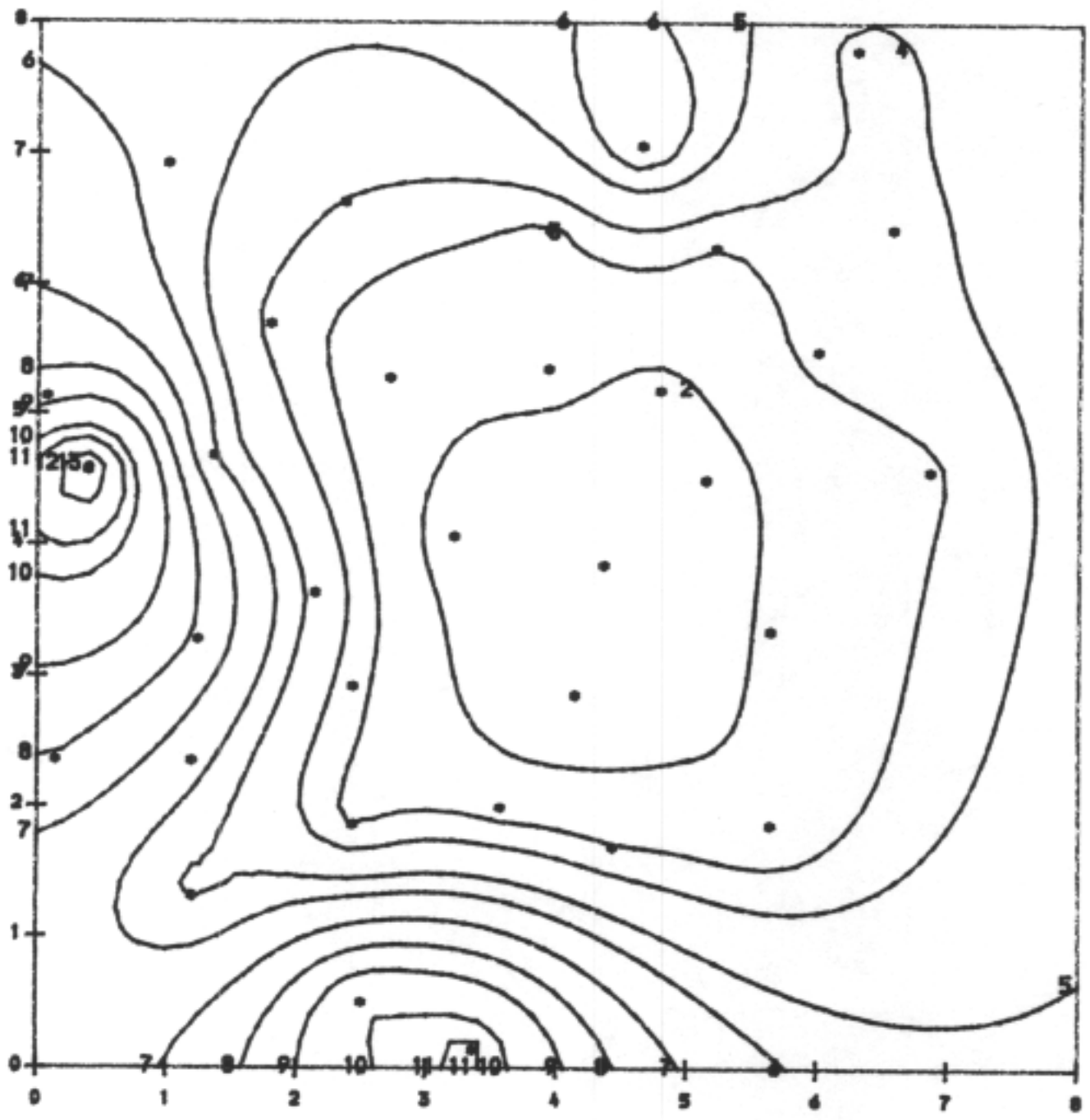


Figure 12.

Contour from the maximum likelihood predictor for the Bauxite data under stationary model with power scheme. [Contours: $1 = 3(3), \ldots, 3 = 8,\ 4 = 10,\ 5 = 13,\ 6 = 15(3), \ldots, 9 = 24,\ 10 = 28,\ 11 = 30,\ 12 = 33,\ 13 = 35(3), \ldots, 15 = 41$.]



then we can find an orthogonal matrix $\mathbf{W}$ (as in Section 5.3 for the torus case), given by

$$\boldsymbol{\Sigma}_0 = \mathbf{W}'\boldsymbol{\Lambda}\mathbf{W}, \quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where $\mathbf{W}$ does not depend on the parameters ϕ_h. For the zero mean case,

$$|\boldsymbol{\Sigma}| = \Pi(\psi^2 + \lambda_j), \ \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \Sigma(\psi^2 + \lambda_j)^{-1} I(2\pi j/n),$$

where $I(\cdot)$ is the periodogram given by (5.4.14). However, we no longer have $\hat{\sigma}_h = C(\mathbf{h})$, although, as before, the ML equations can be written from the general equations. An exception is the first–order C–CAR model (Besag, 1978). Besag and Kempton (1986) have considered a regression case on the line with nugget effect, with the error being IRF–0. In general, we could add a nugget term to (4.3.1).

Example. Mercer and Hall (1911) have given the results of a uniformity trial on the wheat plots on a 20–by–25 lattice. Following Besag (1977b), we fit the first–order Gaussian CAR model with nugget parameter, mean and variance. Let us write

$$E(X_{ij}|\text{rest}) = \mu(1 - 2\theta_1 - 2\theta_2) + \theta_1(x_{i-1,j} + x_{i+1,j}) + \theta_2(x_{i,j-1} + x_{i,j+1}),$$

and

$$\operatorname{Var}(X_{ij}|\text{rest}) = \tau^2.$$

The accompanying estimation results are

$$\hat{\mu} = 3.95, \ \hat{\tau}^2 = 0.033, \ \hat{\delta} = 2.108, \ \hat{\theta}_1 = 0.4758, \ \text{and} \ \hat{\theta}_2 = 0.0203,$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ have SEs of 0.01, and have a high negative correlation of -0.97. This latter correlation reflects the relationship $|\theta_1| + |\theta_2| < 1/2$. It follows from Section 2.5 that $\hat{\psi}^2 = 0.0696$ and $\hat{\sigma}^2 = 0.1335$. Figure 13 gives the semi–variograms of the data in four directions, with fitted semi–variogram plots that indicate a nugget effect as well as anisotropy. Also, $\hat{\theta}_1 + \hat{\theta}_2 = 0.4961$ indicates that there is a trend; the plots also indicate ripples for the semi–variograms especially around $|\mathbf{h}| = 3$. In fact, the data have a periodic trend, as is revealed by plotting the empirical correlation function, or more clearly by graphing the empirical spectral density (see McBratney and Webster, 1983).

6.4. Asymptotics

Let us write

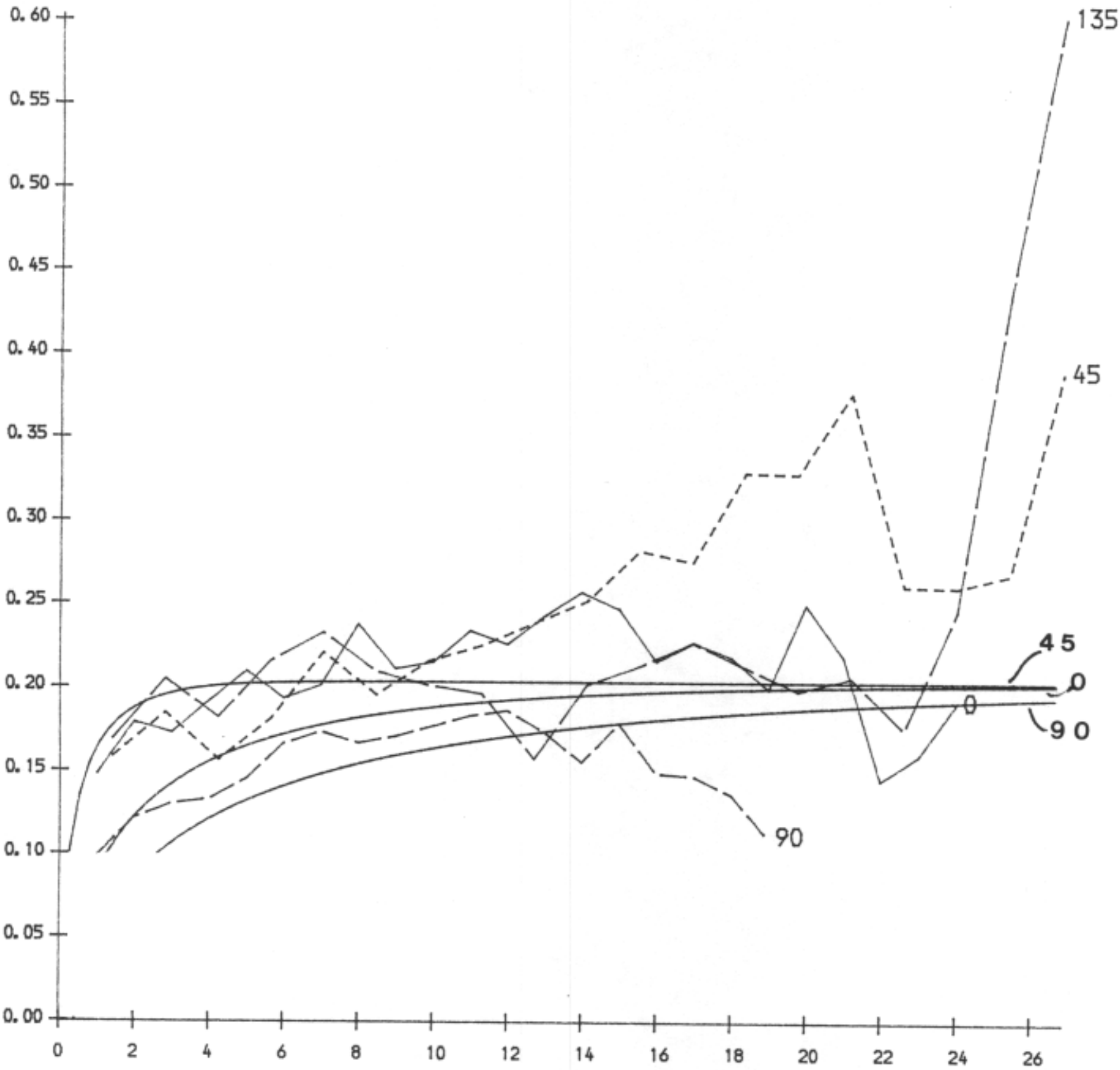
$$\sigma(\mathbf{h}; \boldsymbol{\theta}) = \sigma_1(\mathbf{h}; \boldsymbol{\theta}_2) + \theta_1\delta(\mathbf{h}), \ \theta_1 \geq 0, \tag{6.4.1}$$

where θ_1 is the nugget parameter, and $\delta(\mathbf{h}) = 1$ if $\mathbf{h} = \mathbf{0}$; $\delta(\mathbf{h}) = 0$ if $\mathbf{h} \neq \mathbf{0}$. Since $\theta_1 = 0$ lies on the boundary of the parameter space, the asymptotics are not straightforward. For such cases in the independent and identically distributed variables, see Moran (1971) and Chant (1974).

Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be positive definite for all values of $\boldsymbol{\theta}$, where we restrict $\theta_1 \geq 0$. Although $\sigma(\mathbf{h}; \boldsymbol{\theta})$ is not positive definite for all $\theta_1 < 0$, it is the case that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ will remain positive definite for $\theta_1 < 0$ but near enough to zero. Let $\hat{\theta}_1^\star$ and $\hat{\theta}_1$ be the MLEs of θ_1 for the unrestricted case and the restricted case, respectively. Following Mardia and Marshall (1984), it seems plausible that the likelihood will be quadratic in θ_1 as $h \to \infty$. Hence $\hat{\theta}_1^\star$ will be

Figure 13.

For Mercer and Hall data, the semi–variogram in the four principal directions and the semi–variogram from fitted first order CAR.



asymptotically normal, and $\hat{\theta}_1$ will be asymptotically equivalent to $\text{MAX}(0, \hat{\theta}_1^*)$. Therefore, asymptotically $\hat{\theta}_1$ will have a censored normal distribution at $\hat{\theta}_1 = 0$. Hence to test

$$H_0 : \theta_1 \geq 0 \text{ versus } H_1 : \theta_1 \text{ unrestricted},$$

it can be shown for unknown $\boldsymbol{\theta}_2$ for large n that, with probability $1/2$,

$$-2\log_e(\lambda) = \begin{cases} 0 & \text{if } \hat{\theta}_1 \leq 0, \\ \hat{\theta}_1/\text{Var}(\hat{\theta}_1) & \text{if } \hat{\theta}_1 > 0, \text{ distributed as } \chi_1^2, \end{cases}$$

where λ is the likelihood ratio statistic. Hence to test $\theta_1 = 0$, one should use

$$\hat{\theta}_1^2/\text{Var}(\hat{\theta}_1) > \chi_1^2(2\alpha), \tag{6.4.2}$$

where $\text{Var}(\hat{\theta}_1)$ is computed at $\theta_1 = 0$ as described in the subsequent discussion. (It should be noted that the level of significance gets doubled.)

Recall from (3.2.2) that the (i,j)-th element of the information matrix $\mathbf{A}$ for $\boldsymbol{\theta}$ is $a_{ij} = (1/2)\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_j)$. If $\mathbf{A}$ is partitioned for $(\theta_1, \boldsymbol{\theta})$, then

$$\text{Var}(\theta_1)_{\text{ at } \theta_1=0} \sim (a_{11} - \mathbf{a}_{21}'\mathbf{A}_{22}^{-1}\mathbf{a}_{21})_{\theta_1=0}^{-1}. \tag{6.4.3}$$

Some simplification of $\mathbf{A}$ can be achieved with

$$\boldsymbol{\Sigma} = \theta_1\mathbf{I} + \theta_2\mathbf{P}(\boldsymbol{\theta}), \tag{6.4.4}$$

since we have $\boldsymbol{\Sigma}_1 = \mathbf{I}$, $\boldsymbol{\Sigma}_2 = \mathbf{P}(\boldsymbol{\theta})$, and $\boldsymbol{\Sigma}_i = \theta_2\mathbf{P}_i(\boldsymbol{\theta})$ for $i > 2$. Hence,

$$\begin{aligned} a_{11} &= \theta_2^{-2}\text{tr}(\mathbf{P}^{-2}), \\ a_{22} &= n\theta_2^{-2}, \\ a_{12} &= \theta_2^{-1}\text{tr}(\mathbf{P}^{-1}), \text{ and} \\ a_{ij} &= \theta_2^{-1}\text{tr}(\mathbf{P}^{-1}\mathbf{P}_i\mathbf{P}^{-1}\mathbf{P}_j), \ i,j = 3,4,\ldots,n. \end{aligned} \tag{6.4.5}$$

Under (6.4.4), consider on the line $\rho(\theta)$ with the correlation function

$$\rho(h;\theta) = \theta^{|h|}, \ |\theta| < 1.$$

Under $\theta_1 = 0$, we find that

$$\begin{aligned} a_{11} &= \frac{1}{2}\theta_2^{-2}(1-\theta^2)^{-2}[n + 2(2n-3)\theta^2 + (n-2)\theta^4] \\ a_{12} &= \frac{1}{2}\theta_2^{-2}(1-\theta^2)^{-1}\{2 + (n-2)(1+\theta^2)\}, \\ a_{22} &= \frac{1}{2}n\theta_2^{-2}. \end{aligned}$$

Hence asymptotically $\text{Var}(\hat{\theta}_1)$ under $\theta_1 = 0$ from (6.4.3) is

$$\theta^2/\{n\theta_2^2(1-\theta^2)^2\}.$$

Hence, the variance is well defined. For $\theta = 0$, the variance as expected, is zero.

7. Some extensions

7.1. Multivariate spatial model

Let $\mathbf{X}$ be a matrix of $n \times m$ observations, where the i-th row $(\mathbf{X})_i$ is the m–variate observation at the i–th site, $i = 1, 2, \ldots, n$. We can write

$$E(\mathbf{X})_i = \mathbf{f}(\mathbf{t}_i, \boldsymbol{\beta}),\ \boldsymbol{\beta} \in R^q,\ \text{and}\ \mathrm{Cov}\{(\mathbf{X})_i,\ (\mathbf{X})_j\} = \rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in R^p.$$

We assume that

$$\Sigma\Sigma\alpha_i\alpha_j\rho(\mathbf{t}_i, \mathbf{t}_j; \boldsymbol{\theta})$$

is a p.d. matrix for all $\alpha_1, \alpha_2, \ldots, \alpha_s$, and that vectors $\mathbf{t}_i$ and $\mathbf{t}_j$ are contained in R^d. Writing $\mathbf{X}$ as a stacked vector, denoted by $\mathrm{Vec}(\mathbf{X})$, yields

$$\mathrm{Cov}[\mathrm{Vec}(\mathbf{X})] = \boldsymbol{\Sigma}(\boldsymbol{\theta}),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is an mn–by–mn matrix. Assuming $\mathrm{Vec}(\mathbf{X})$ to be normally distributed, then both the likelihood of $\mathbf{X}$ and its ML equations can be written as before.

Of special interest here is the linear model, where

$$E(\mathbf{X}) = \mathbf{F}\boldsymbol{\beta}.$$

The regression coefficient $\boldsymbol{\beta}$ is now a matrix (*e. g.*, for $d = 2$ with a quadratic trend, $\boldsymbol{\beta}$ is a 6–by–m matrix and $\mathbf{F}$ is an n–by–6 matrix). We should note that $\mathbf{F}$ is the same as that for the univariate case of $m = 1$. We can use a "factorized model" for the covariance matrix (Mardia, 1984), such that

$$\mathrm{Cov}[(\mathbf{X})_i, (\mathbf{X})_j] = \rho(\mathbf{i} - \mathbf{j}; \boldsymbol{\theta})\boldsymbol{\Lambda},$$

with $\rho(0) = 1$, so that $\mathrm{Var}[(\mathbf{X})_i] = \boldsymbol{\Lambda}$ for all i. Here $\boldsymbol{\Lambda}$ is an m–by–m symmetric matrix. Thus

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}.$$

Hence, the log–likelihood is simply

$$\text{constant} - (n/2)\log_e|\boldsymbol{\Lambda}| - (m/2)\log_e|\boldsymbol{\Gamma}| - (1/2)\mathrm{tr}[(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})'\boldsymbol{\Gamma}^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\beta})\boldsymbol{\Lambda}^{-1}].$$

We can now obtain the ML equation for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ for a given $\mathbf{F}$ and $\rho(\cdot)$. The profile likelihood for $\boldsymbol{\theta}$ simply maximizes

$$-(n/2)\log|\hat{\boldsymbol{\Lambda}}(\boldsymbol{\theta})| - (m/2)|\hat{\boldsymbol{\Gamma}}(\boldsymbol{\theta})|,$$

where

$$\hat{\boldsymbol{\Lambda}}(\boldsymbol{\theta}) = (1/n)[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})]'\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}[\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})],$$

and

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = [\mathbf{F}'\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\mathbf{F}]^{-1}\mathbf{F}'\boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\mathbf{X}.$$

For some further details see Mardia (1984). For $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ known, the prediction problem is called co–kriging. If $\boldsymbol{\Lambda}$ is known and $\boldsymbol{\theta}$ is a scalar, then the profile is univariate and can be easily plotted. For a multivariate CAR model, see Mardia (1988).

7.2. Regularized process

Let A_i denote the i-th plot and let $|A_i|$ be its area. We assume that univariate observations $X^{\star}_{A_i}$ are taken on these blocks. Then for the model

$$X(\mathbf{t}) = \mathbf{f}'(\mathbf{t})\boldsymbol{\beta} + \varepsilon(\mathbf{t}),$$

we have

$$E(X^{\star}_{A_i}) = \frac{1}{|A_i|}\left[\int_{A_i} \mathbf{f}'(\mathbf{t})d\mathbf{t}\right]\boldsymbol{\beta}, \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \frac{1}{(|A_i||A_j|)}\int_{A_i}\int_{A_j} \sigma(\mathbf{s}, \mathbf{t}; \boldsymbol{\theta})d\mathbf{s}\, d\mathbf{t},$$

and $\sigma(.,\ .,\ \boldsymbol{\theta})$ is the covariance function of $\boldsymbol{\theta}$.

Usually $A_i = A$. The estimator proceeds as before, but in practice one uses $X^{\star}_{A_i} = X(\mathbf{t}_i)$ if $\mathbf{t}_i$ is the centre of A_i, especially when the A_is are small. Thus we are back to the usual model.

7.3. Applications to design

Let

$$\mathbf{X} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{F}$ is a design matrix and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. Consider the n–by–r design matrix $\mathbf{F}$ for treatments, with

$$(\mathbf{F})_{ij} = 1 \text{ if the j-th treatment is applied to the i-th plot;}$$
$$= 0 \text{ otherwise,}$$

$i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, r$, so that $\mathbf{F}'\mathbf{F} = r\mathbf{I}$. Consider the basic CAR model

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I} - \theta\mathbf{N},$$

where $(\mathbf{N})_{ij} = 1$ if i and j are neighbours; $= 0$ otherwise, and θ is some unknown parameter.

We can use results reported in Section 2 to estimate θ and $\boldsymbol{\beta}$. Note that from (5.1.4)

$$\hat{\boldsymbol{\beta}} = (1/r)[\mathbf{F}'\mathbf{X} - \theta\mathbf{F}'\mathbf{N}(\mathbf{X} - \mathbf{F}\hat{\boldsymbol{\beta}})].$$

An alternative method is to use a Papadakis estimator of regression parameters $\boldsymbol{\beta}$ defined by (see Martin, 1982)

$$\hat{\boldsymbol{\beta}}_P = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'(\mathbf{X} - \hat{\beta}\mathbf{Y}),$$

where here $\hat{\boldsymbol{\beta}} = \mathbf{Y}'\mathbf{Q}\mathbf{X}/\mathbf{Y}'\mathbf{Q}\mathbf{Y}$, with $\mathbf{Y} = c\mathbf{N}\mathbf{Q}\mathbf{X}$, c = a constant depending upon the layout (*e. g.*, circle, torus, or as such), and

$$\mathbf{Q} = \mathbf{I} - \mathbf{F}'(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}.$$

For the circle, $c = 1/2$, and $\mathbf{N} = \mathbf{W} + \mathbf{W}^{-1}$, where

$$\mathbf{W} = \begin{cases} 1 & \text{along the leading upper diagonal;} \\ 1 & \text{in the lower left corner; and} \\ 0 & \text{otherwise.} \end{cases}$$

7.4. Missing values in lattice data

Let us assume that some data on a lattice are missing, but suppose that $\boldsymbol{\Sigma}^{-1}$ is known explicitly (*e. g.*, for a CAR model). Then one strategy is to estimate the missing values by ML prediction (with the auto–regression coefficient matrix explicitly evaluated or an approximation used), and obtain $\hat{\beta}$ and $\hat{\theta}$ on the lattice (see Martin, 1984). This procedure should not make any difference in the estimates of $\hat{\beta}$ and $\hat{\theta}$.

Also the loss of information through missing values on β and θ can be computed as

$$I_{\beta,\theta} \text{ whole data} - I_{\beta,\theta} \text{ observed data.}$$

There is no difficulty in interpreting this measure for a single parameter; but, for higher dimensions, a determinant must be evaluated to measure the loss. Such important cases are studied in Haining *et al.* (1989).

A simple example of this situation is when the density is (5.6.5). We can maximize the density with respect to the missing values. For example, if two neighbouring values x_i and x_j are missing, with the first–order neighbourhood in 2–dimensions, then the respective estimates of x_i and x_j are simply

$$(4\tilde{x}_i + \tilde{x}_j)/5 \text{ and } (\tilde{x}_i + 4\tilde{x}_j)/5,$$

respectively, where now $\tilde{x}_i$ is the mean of the three observed neighbours of the i-th site.

8. Discussion

We have described mainly the ML method of estimation for the spatial linear model in two forms, DR and CAR. For the CAR model, there are some interesting features. For the T–CAR, the MLEs of $\boldsymbol{\theta}$ are obtained through the moment estimator of the covariance function. Another key feature is that the matrix of the eigenvectors for the covariance matrix does not involve $\boldsymbol{\theta}$. The same comment applies to the first–order C–CAR, even when the nugget parameter is present. However, the M–CAR is the realistic model, and the MLE from the approximate ML equations, $\hat{\sigma}_h = C_k(h)$ of Section 5.5, have better properties.

We have not examined alternative estimators, such as the minimum quadratic unbiased estimators for θ (see Kitanidis, 1985; Marshall and Mardia, 1985; Stein, 1986), but these estimates are closely related to the MLEs for the intrinsic model (see Stein, 1987). Also, we have not examined any exploratory techniques for spatial data (see, for example, Cressie, 1986).

Consequently, great care should be taken in formulating a model, especially since there can be some identifiability problems. For example, when the stationary model versus a trend is used, the stationary model will tend to estimate the long–term correlations if there is a trend. Plotting the data, empirical semi–variogram surface or semi–variogram in different directions, periodogram, and so forth, can be revealing. Further, the computation of the ML estimates requires care due to possible local maxima, and whenever possible a profile likelihood should be examined.

9. References

Besag, J. E. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, **36B**: 192-236.

Besag, J. E. (1975) Statistical analysis of non–lattice data. *Statistician*, **24**: pp.79-195.

Besag, J. E. (1977a) Efficiency of pseudo–likelihood estimation for sample Gaussian fields. *Biometrika*, **64**: 616-8.

Besag, J. E. (1977b) Errors in–variables estimation for Gaussian lattice scheme. *Journal of the Royal Statistical Society*, **39B**: 73-78.

Besag, J. E. (1978) Discussion to Nearest neighbour models in the analysis of field experiments. *Journal of the Royal Statistical Society*, **40B**: 165-166.

Besag, J. E. (1981) On a system of two–dimensional recurrence equations. *Journal of the Royal Statistical Society*, **43B**: 302-309.

Besag, J. E. (1989) Towards Bayesian image analysis, to appear.

Besag, J. E. and P. A. P. Moran. (1975) On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, **62**: 555-562

Besag, J. E. and R. Kempton. (1986) Statistical analysis of field experiments using neighbouring plots. *Biometrics*, **42**: 321-351.

Brewer, A. C. and R. Mead. (1986) Continuous second order models of spatial variation with application to the efficiency of crop experiments (with discussion). *Journal of the Royal Statistical Society*, **149A**: 314-348.

Chant, D. (1974) On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika*, **61**: 291-298.

Cliff, A. D. and J. K. Ord. (1981) *Spatial Processes*. London: Pion.

Cressie, N. (1986) Kriging nonstationary data. *Journal of the American Statistical Association*, **81**: 625-634.

Dahlhaus, R., and H. R. Künsch. (1987) Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, **74**: 877-82.

Davis, J. C. (1973) *Statistics and Data Analysis in Geology*. New York: Wiley.

Delfiner, P. (1976) Linear estimation of non–stationary spatial phenomena, in *Advances in Geostatistics in the Mining Industry*, edited by M. Guarascio, M. David, and C. Huijbregts, pp. 49-68. Dordrecht: Reidel.

Griffith, D. A. (1988) *Advanced Spatial Statistics*. Dordrecht: Kluwer.

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

Haining, R., D. A. Griffith, and R. Bennett. (1989) Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data. *Communications in Statistics*, **18**: 1875-1894.

Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**: 320-340.

Hocking, R. R. (1985) *The Analysis of Linear Models*. New York: Brooks/Cole.

Kalbfleisch, D. and D. A. Sprott. (1970) Applications of likelihood methods to models in-

volving large numbers of parameters. *Journal of the Royal Statistical Society*, **32B**: 175-194.

Kent, J. T. and K. V. Mardia. (1988) Spatial classification using fuzzy membership models. *Transactions of the IEEE/PAMI*, **10**: 659-671.

Kitanidis, P. K. (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, **19**: 909-921.

Kitanidis, P. K. (1985) Minimum–variance unbiased quadratic estimation of covariances of regionalised variables. *Mathematical Geology*, **17**: 195-208.

Kitanidis, P. K. (1987) Parametric estimation of covariance of regionalised variables. *Water Resources Bulletin of the American Water Resources Association*, **23**: 557-567.

Kitanidis, P. K. and R. W. Lane. (1985) Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss–Newton methods. *Journal of Hydrology*, **79**: 53-71.

Künsch, H. R. (1981) Thermodynamics and statistical analysis of Gaussian random fields. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **98**: 407-421.

Künsch, H. R. (1983) Approximations to the maximum likelihood equations for some Gaussian random fields. *Scandinavian Journal of Statistics*, **10**: 239-246.

Künsch, H. R. (1987) Intrinsic autoregressions and related models on the two–dimensional lattice. *Biometrika*, **74**: 517-24.

Mardia, K. V. (1980) Some statistical inference problems, in Kriging II: Theory, Proceedings of the 26th International Geology Congress, pp. 113-131. Sciences de la Terre: "Advances in Automatic Processing and Mathematical Models in Geology," Series "Informatique Geologie," # 15.

Mardia, K. V. (1984) Spatial discrimination and classification maps. *Communications in Statistics*, **13**: 2181-2197.

Mardia, K. V. (1986) Discussion to the paper by Brewer and Mead. *Journal of the Royal Statistical Society*, **149A**: 341-342.

Mardia, K. V. (1988) Multi–dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, **24**: 265-284.

Mardia, K. V. (1989) Markov models and Bayesian methods in image analysis. *Journal of Applied Statistics*, **16**: 125-130.

Mardia, K. V., J. T. Kent, and J. M. Bibby. (1989) *Multivariate Analysis*, 7th printing with corrections. New York: Academic Press.

Mardia, K. V. and R. J. Marshall. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**,: 135-46.

Mardia, K. V. and A. J. Watkins. (1989) On multimodality of the likelihood for the spatial linear model. *Biometrika*, **76**: 289-295.

Marechal, A. and J. Serra. (1970) Random kriging, in Geostatistics: A Colloquium, edited by D. Merriam, pp. 91-112. New York: Plenum Press.

Marshall, R. J. and K. V. Mardia. (1985) Minimum norm quadratic estimation of components of spatial covariance. *Mathematical Geology*, **17**: 517-525.

Martin, R. J. (1982) Some aspects of experimental design and analysis when errors are

correlated. *Biometrika*, **69**: 597-612.

Martin, R. J. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communication in Statistics*, **13**: 1275-1288.

Martin, R. J. (1987) Some comments on correction techniques for boundary effects and missing value techniques. *Geographical Analysis* **19**: 273-282.

Matheron, G. (1971) The Theory of Regionalized Variables and Its Applications. Fountainbleau: Les Cahiers du Morphologie Mathematique, Fasc. No. 5.

McBratney, A. B. and R. Webster. (1981) Detection of ridge and furrow pattern by spectral analysis of crop yield. *International Statistical Review*, **49**: 45-52.

Mercer, W. B. and A. D. Hall. (1911) The experimental error of field trials. *Journal of Agricultural Science*, **4**: 107-132.

Moran, P. A. P. (1971) Maximum likelihood estimation in nonstandard conditions. *Proceedings of the Cambridge Philosophical Society*, **70**: 441-450.

Patterson, H. D. and R. Thompson. (1974) Maximum likelihood estimation of components of variance, in Proceedings of the 8th International Biometrics Conference, edited by L. Corsten and T. Postelnicu, pp. 197-208. Bucharest: Academy of the Socialist Republic of Rumania.

Plackett, R. L. (1960) *Principles of Regression Analysis*. Oxford: Clarendon Press.

Ripley, B. D. (1988) *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Stein, M. L. (1986) A modification of the minimum norm quadratic estimation of a generalised covariance function for use with large data sets. *Mathematical Geology*, **18**: 625-633.

Stein, M. L. (1987) Minimum norm quadratic estimation of spatial variograms. *Journal of the American Statistical Association*, **82**: 765-772.

Stein, M. L. (1988) Asymptotically efficient prediction of a random field with a misspecified covariance function. *Annals of Statistics*, **16**: 55-63.

Tunnicliffe–Wilson, G. (1989) On the use of marginal likelihood in time series estimation. *Journal of the Royal Statistical Society*, **51B**: 15-27.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society*, **50B**: 297-312.

Warnes, J. J. and B. D. Ripley. (1987) Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74**: 640-42.

Watkins, A. J. and K. V. Mardia. (1989) Some problems in spatial inference, to appear.

Whittle, D. (1954) On stationary processes in the plane. *Biometrika*, **41**: 434-449.

DISCUSSION

"Maximum likelihood estimators for spatial models"

by Kanti V. Mardia

As an intuitively suggested extension of one–dimensional time series analysis, spatial analysis in two dimensions calls itself to attention; however, as is often the case, the introduction of an additional dimension gives rise to a series of new problems.

The first is the operational *specification* of a computable model; as is the case with spectral analysis, the presence of a trend should be explicitly recognized and taken care of. Various possibilities are suggested by Mardia: direct representation, conditional autoregression, and simultaneous autoregression. As is usual in spatial statistics (and spatial econometrics should take this more often into account), boundary problems are taken up, which can be done in various manners (torus–solution, free boundaries).

The next problem is that of *estimating* the parameters of one of the above–mentioned specifications. Maximum likelihood is a good candidate (for spatial econometrics one is referred to Paelinck and Klaassen (1979, Ch. 3) but beset with a certain number of difficulties. Is there a single global maximum (hence the suggestions of plotting the profile likelihood)? How does one compute it in the many parameters case? Furthermore, do the resulting estimators have acceptable asymptotic properties, if small sample imperfections (*e. g.*, bias) are present, as is often the case when so–called Whittle–type approximations are used? The above remarks also are applicable to intrinsic processes. A useful addition to specification and estimation is the study of errors in the variables (measurement errors), which all too often are present in empirical exercises.

Finally *testing* is derived from asymptotic normality.

The examples given by Mardia show the operational character of the methods developed and advocated; they put into light one of the difficulties besetting spatial data, to wit anisotropy (together with the trend already mentioned above). The analysis of topographic and landsat data is revealing from those points of view.

In his conclusion the author rightly draws the reader's attention to a number of key points, which may be summarized as follows:

1. the possible presence of identifiability problems (neglecting trend terms can lead to erroneously estimating long–range correlations);
2. the necessity to explicitly include anisotropy (this matches a problem in spatial econometrics, that of asymmetry of the relations postulated); and,
3. the need to tackle second–order conditions, or at least to graphically inspect the likelihood function.

References

Paelinck, J., and L. Klaassen. (1979) *Spatial Econometrics*. Farnborough: Saxon House.

J. H. P. Paelinck, Erasmus University