Biometrika (1989), **76**, 2, *pp.* 289-95
Printed in Great Britain

On multimodality of the likelihood in the spatial linear model

BY K. V. MARDIA
Department of Statistics, University of Leeds, Leeds LS2 9JT, U.K.

AND A. J. WATKINS
Department of Management Science and Statistics, University College, Swansea SA2 8PP, U.K.

SUMMARY

A popular covariance scheme used for the spatial linear model in geostatistics has spherical form. However, the likelihood is not twice differentiable with respect to the range parameter, and this raises some questions regarding the unimodality of the likelihood. We compare the likelihoods of the spatial linear model for small samples under this scheme and the doubly geometric scheme. Also, a power covariance with range parameter is proposed. In view of potential multimodal likelihoods for small samples for this model, a convenient profile likelihood is introduced and studied.

Some key words: Covariance scheme; Kriging; Maximum likelihood of estimation; Multimodality; Profile likelihood; Spatial linear model.

1. INTRODUCTION

Following Mardia & Marshall (1984) we consider the spatial linear model defined as follows. Suppose that we have a real-valued covariance stationary Gaussian process $\{Y(t); t \in T\}$ defined on an index set T. For all $t \in T$, $E\{Y(t)\} = \{x(t)\}'\beta$, where $x(t) = \{x_1(t), \ldots, x_q(t)\}'$ is a $q \times 1$ vector of nonrandom regressors based on the q functions $x_1, \ldots, x_q$ defined throughout T, and β is a $q \times 1$ vector of parameters. Further, for all $s, t \in T$,

$$\operatorname{cov}\{Y(s), Y(t)\} = \sigma^2 \rho(s-t; \theta) \quad (\sigma > 0),$$

where $\rho(.)$ is a positive-definite covariance function of known functional form, parameterized by a $p \times 1$ vector $\theta \in \Theta$, an open subset of R^p and $\rho(0; \theta) = 1$. Thus the covariance function has $p+1$ parameters. We suppose that the process is observed at each of the n distinct locations $t_1, \ldots, t_n$ in T, yielding the sample vector $Y = \{Y(t_1), \ldots, Y(t_n)\}'$. We let X be the $n \times q$ matrix with ith row $\{x(t_i)\}'$, and V with elements $\rho(t_i - t_j; \theta)$ be the $n \times n$ theoretical correlation matrix of Y. The log likelihood of Y, apart from an additive constant, is

$$l(Y, \beta, \sigma^2, \theta) = -n \log \sigma - \tfrac{1}{2} \log |V(\theta)| - \tfrac{1}{2}(Y - X\beta)'\{V(\theta)\}^{-1}(Y - X\beta)/\sigma^2. \quad (1\cdot1)$$

We deal here mainly with the estimation of σ^2, β and θ in the case $p = 1$, $q = 1$ and $X = 1$, so that $\{Y(t)\}$ is stationary, and β and θ are scalar parameters. Note that β is the constant mean and θ represents a correlation parameter. This three-parameter model will be called our basic model.

In § 2, we study through simulations the modality of the log likelihood for some specific covariance schemes for small samples. In § 3, a convenient profile log likelihood is introduced and a practical example given. In § 4, a discussion of our findings is given. All computations were performed by FORTRAN programs using NAG subroutines and double precision arithmetic. This paper was motivated by Warnes & Ripley (1987) who gave some examples of multimodal likelihoods for our basic model.

2. SOME COVARIANCE SCHEMES AND THE LOG LIKELIHOOD

2·1. *The index set T*

In this section we consider the possibility that (1·1) may be multimodal for samples of finite size. To emphasize the spatial nature of the problem we take the sample locations to be the nodes of a unit square rectangular $N \times N$ lattice in two dimensions. We choose $T = Z^2$ although we note the possibility $T = R^2$ and remark that, as in time series analysis, the relation between T and Θ means that our choice depends in part on the variety of covariance structures it is desired to deem admissible. We now give some important points related to three covariance schemes studied here.

2·2. *Covariance schemes*

The spherical correlation scheme is defined by

$$\rho_s(h;\alpha) = \begin{cases} 1 - \tfrac{3}{2}(|h|/\alpha) + \tfrac{1}{2}(|h|^3/\alpha^3) & (|h| \leq \alpha), \\ 0 & (|h| > \alpha), \end{cases} \tag{2·1}$$

where $|h|$ is the Euclidean norm of h. It can be shown for $n = 2$ and $|t_1 - t_2| = 1$ that

$$[d^2 l/d\alpha^2]_{\alpha \to 1^-} = 0,$$

but

$$[d^2 l/d\alpha^2]_{\alpha \to 1^+} = 3\,Y(t_1)\,Y(t_2).$$

Hence l is not twice differentiable at $\alpha = 1$. Similarly, in general, the likelihood need not be twice differentiable at points of

$$S = \{\alpha\colon \alpha = |t_i - t_j|;\ i = 1, \ldots, n-1;\ j = i+1, \ldots, n\},$$

although the first derivative is continuous, and the second derivative exists at all points of the complement of S in R^+. For large α, all elements of S may be less than the true value of α, although we note that the number of elements increases with n. For $T = Z^2$, we have

$$S = \{(i^2 + j^2)^{\frac{1}{2}};\ i = 1, \ldots, N-1;\ j = 0, \ldots, i\}, \tag{2·2}$$

where the largest element is $2^{\frac{1}{2}}(N-1)$.

Figure 2·1 shows the first derivative of a log likelihood under (2·1) over $\alpha \in [2·6, 3·8]$ for a sample with $N = 5$ and $\alpha = 3$. From (2·2) the points of S in this interval are $8^{\frac{1}{2}}$, 3, $10^{\frac{1}{2}}$, $13^{\frac{1}{2}}$; that is, 2·8, 3, 3·2, 3·6. These points of S partition (2·6, 3·8) into subintervals; the three zeros of the first derivative shown in Fig. 2·1 occur in distinct subintervals, so that these discontinuities in the second derivative are here associated with the multimodality of the likelihood.

The power correlation scheme is defined by

$$\rho(h;\gamma) = \begin{cases} (1 - |h|/\gamma)^m & (|h| \leq \gamma), \\ 0 & (|h| \geq \gamma), \end{cases}$$

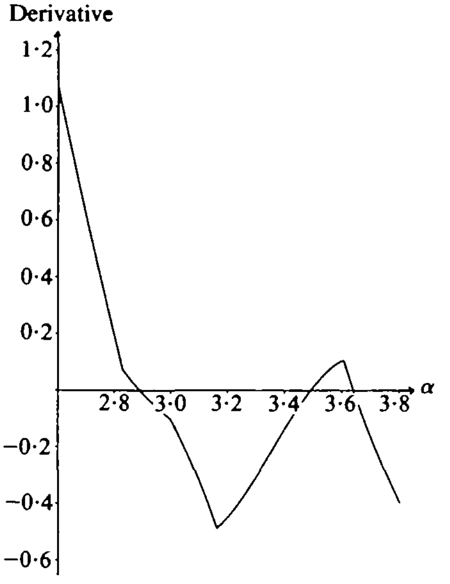

Fig. 2·1. Derivative of log likelihood (1·1) under $\rho_s(h;\alpha)$ with $\alpha \in (2\cdot6, 3\cdot8)$ for simulated data, $\alpha = 3$.

where $m \geqslant 2$ is an integer. We show that this function is positive-definite in R^3 and hence also in R and R^2. We first consider $m = 2$. It can be shown that the Fourier transform of $\{f(t)\}^2$ with

$$f(t) = \begin{cases} 1-|t| & (|t| \leqslant 1), \\ 0 & (|t| \geqslant 1), \end{cases}$$

is for $t \in R^3$

$$\begin{cases} (2\pi^{-1})^{\frac{1}{2}}/30 & (r=0), \\ (2\pi^{-1})^{\frac{1}{2}} 2r^{-5}\{r(2+\cos r) - 3\sin r\} & (r>0), \end{cases}$$

which is positive for all r. Hence $\{f(t)\}^2$ is positive-definite. We can similarly show that $\{f(t)\}^3$ is positive-definite, and it now follows by multiplication that $\{f(t)\}^m$ is positive for all integer $m \geqslant 2$. We consider here

$$\rho_p(h;\gamma) = \begin{cases} (1-|h|/\gamma)^4 & (|h| \leqslant \gamma), \\ 0 & (|h| \geqslant \gamma), \end{cases} \tag{2·3}$$

so that the first three derivatives of $\rho_p(.)$ with respect to γ are ensured.

Consider also the doubly geometric scheme defined by

$$\rho_g(h;\lambda) = \lambda^{|k|+|l|} \quad (|\lambda|<1),$$

where $h = (k, l)'$. This scheme is symmetric under permutations of the coordinates, and reflections through the coordinate axes, but is not completely isotropic like ρ_s and ρ_p.

For Z^2, see Martin (1979) and Mardia (1980). In this case, we can write the maximum likelihood equation for λ, with σ^2 estimated or not, as a quintic in λ. Thus, for the regular lattice case, it is easier to study modality for this scheme than for the previous two schemes. We could write $\rho_g(h;\lambda)$ as $\exp\{-a(|k|+|l|)\}$.

2·3. *A simulation study*

We first study the log likelihood for $N = 3$, 4 and 5, for $\beta = 0$ and known, $\sigma^2 = 1$, and for varying values of the correlation parameters in the three schemes described above. For each combination of sample size and correlation parameter we simulated 5000 sets of data from the corresponding population. We examined numerically the derivative of the log likelihood for each set of data for $\rho_s(.)$ and $\rho_p(.)$, and classified each log likelihood as unimodal or multimodal. For $\rho_g(.)$, we used the roots of the polynomial equation. Table 2·1 gives the percentage of multimodal log likelihoods. The values of the correlation parameters are selected so that the three correlation schemes are roughly equal at unit distance.

Table 2·1. *Percentage of multimodal likelihoods for spherical, geometric and power schemes based on an $N \times N$ grid with unit spacing when only correlation parameter is used*

Spherical α	Geom. λ	Power γ	Percentage of multimodal likelihoods $N=3$	$N=4$	$N=5$
1·5	0·15	2·5	11·1, 5·8, 2·5	15·4, 2·2, 0·6	16·3, 0·6, 0·2
2·0	0·31	4·0	11·1, 6·5, 3·5	15·7, 2·5, 0·9	20·0, 0·8, 0·4
2·5	0·43	5·0	10·6, 7·9, 3·4	17·9, 3·3, 0·6	21·7, 1·0, 0·3
3·0	0·52	7·0	8·8, 9·4, 3·3	13·7, 3·1, 1·2	23·6, 0·9, 0·3
3·5	0·59	8·0	7·9, 10·0, 4·2	13·9, 3·6, 1·5	24·1, 0·9, 0·4
4·0	0·63	9·0	7·5, 10·8, 3·7	12·0, 4·0, 1·6	25·3, 1·1, 0·7
4·5	0·67	11·0	6·3, 11·6, 4·5	9·9, 4·0, 1·7	23·2, 1·0, 0·7
5·0	0·70	12·0	5·0, 11·2, 4·2	8·5, 4·3, 1·7	18·6, 1·4, 0·9
5·5	0·73	13·0	4·8, 11·4, 4·0	7·9, 4·9, 1·5	13·0, 1·5, 0·9
6·0	0·75	14·0	4·1, 12·4, 3·2	4·8, 4·9, 1·6	9·3, 1·6, 0·9
6·5	0·77	15·0	3·9, 11·8, 2·9	4·5, 5·5, 1·5	7·6, 2·2, 1·0
7·0	0·79	17·0	3·4, 13·6, 2·7	4·2, 5·3, 1·4	6·4, 2·1, 0·9
10·0	0·85	25·0	2·1, 14·2, 2·0	2·5, 8·0, 1·1	2·7, 3·6, 1·0
15·0	0·90	40·0	2·0, 16·8, 2·1	2·1, 10·4, 1·5	2·4, 6·5, 1·2
30·0	0·95	80·0	2·2, 21·1, 2·4	2·1, 16·7, 1·7	2·2, 12·0, 1·4
150·0	0·99	400·0	2·1, 24·8, 2·0	2·0, 23·5, 2·1	2·0, 21·7, 1·8

Let l_s and l_g be likelihoods with schemes ρ_s and ρ_g respectively. Table 2·1 shows that, for small N, multimodal likelihoods are possible for all three schemes, but for ρ_p and ρ_g the percentage of multimodal likelihoods decreases as N increases, whereas for ρ_s this percentage increases. Also, there is a larger percentage of multimodal likelihoods for ρ_s when $\alpha \leq 2^{\frac{1}{2}}(N-1)$. In § 2·1, we observed that the number of discontinuities in the second derivative of l_s increases with N, and that all discontinuities are in the range $1 \leq \alpha \leq 2^{\frac{1}{2}}(N-1)$. Hence, the discontinuities in the second derivative of l_s seem to be associated with multimodality in the likelihood.

It is known for random samples from the Cauchy distribution that the modality of the likelihood depends on whether we are estimating the location parameter alone, or jointly with the scale parameter (Barnett, 1966; Copas, 1975; Reeds, 1985). Table 2·2 gives a comparative simulation study for l_g when both σ^2 and λ are estimated. Again, 5000 simulations were used for each value of λ, and the roots of the appropriate quintic examined. The approach to unimodality is much faster than when the correlation parameter is estimated alone; the same comment applies to the other two schemes.

In view of the possibility of multimodality of the likelihood with small N, the scoring method to find the maximum likelihood estimators may not work, since this can locate

Table 2·2. *Number of multimodal likelihoods for doubly geometric scheme based on $N \times N$ grid when λ and σ^2 both estimated*

λ	$N=3$	$N=4$	$N=5$	λ	$N=3$	$N=4$	$N=5$
0·0	10	3	0	0·5	7	4	0
0·1	9	2	0	0·6	17	1	0
0·2	14	0	0	0·7	16	1	0
0·3	10	2	0	0·8	7	1	0
0·4	19	0	0	0·9	14	1	0

minima as well as maxima. This is particularly so for ρ_s, where even large N may not guarantee unimodality. We note that the various simulations reported by Mardia & Marshall (1984) for small N are based on potentially spurious estimates of α wherever the scoring method is not supplemented. Further, our study also explains the oscillation in the scoring method for ρ_s reported by Warnes & Ripley (1987).

3. Profile log likelihoods

In view of § 2, we may wish to examine plots of the log likelihood, which of course can be cumbersome. We introduce a convenient profile log likelihood for vector-valued β and θ. We have (Mardia & Marshall, 1984) the maximum likelihood estimators of β and σ^2 for fixed θ as

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y, \quad \hat{\sigma}^2 = (Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta})/n. \tag{3·1}$$

Therefore, from (1·1), the profile log likelihood for θ is, apart from an additive constant,

$$l_p^*(Y; \theta) = -\tfrac{1}{2}\log|V| - \tfrac{1}{2}n \log\{(Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta})\}. \tag{3·2}$$

Note that for $p = 1$ this is just a univariate function of θ and thus contour plots are not needed; compare with Warnes & Ripley (1987). Clearly, if $(\hat{\beta}, \hat{\sigma}^2, \hat{\theta})$ maximizes $l(Y; \beta, \sigma^2, \theta)$ then $\hat{\theta}$ maximizes $l_p^*(Y; \theta)$. Further, once $\hat{\theta}$ is found from (3·2), we can obtain $\hat{\beta}$ and $\hat{\sigma}^2$ from (3·1).

We consider Warnes & Ripley's (1987) analysis of the 52 data points in Table 6.4 of Davis (1973). Their model is, with the exponential correlation scheme,

$$\rho_e(h; \alpha) = \exp(-|h|/\alpha) \quad (\alpha > 0). \tag{3·3}$$

For this, Fig. 3·1 shows l_p^* for $\alpha \in [3, 9]$ and indicates that l_p^* is unimodal, having its maximum at $\alpha = 6·12$. Our other calculations show that the maximum is not outside this range. Thus, we conclude that the likelihood is unimodal, and we have $\hat{\alpha} = 6·12$, $\hat{\beta} = 863·7$ and $\hat{\sigma}^2 = 4086·7$.

In passing, note that our study casts doubt on the accuracy of the contour plot of a profile log likelihood of Warnes & Ripley (1987, Fig. 1). Our doubts are strengthened when we note that (3·1) implies that, for α and hence β fixed, their profile should have a single turning point over σ^2, whereas in their figure, there are values of α for which the contoured function has three turning points, for example $\alpha = 6·1$. The basic model with the power correlation scheme also has a unimodal likelihood. Further, contour plots from the maximum likelihood predictor (Mardia & Marshall, 1984) with (3·3) as well as (2·3) are very similar to both Davis (1973, Fig. 6.9) and Ripley (1981, Fig. 4.14). This is not surprising in view of the sensitivity analysis by Warnes (1986). All contour plots indicate that the data is nonstationary, having a global basin shape.

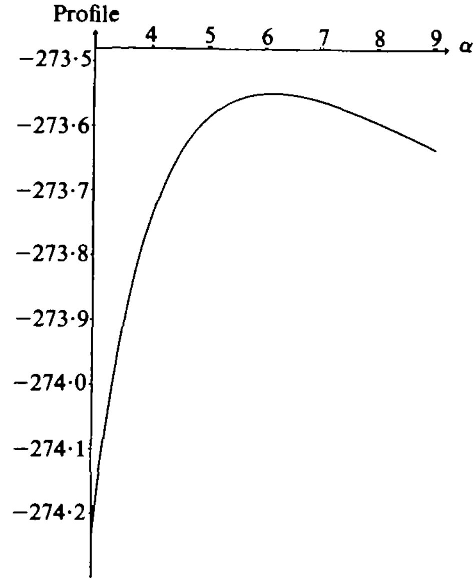

Fig. 3·1. Profile log likelihood (3·2) for Davis's (1973) data over $\alpha \in (3, 9)$, under exponential scheme (3·3).

We comment finally on the discrepancy between the maximum likelihood estimate of α and the value apparently indicated by empirical covariance plots of Ripley (1981, Fig. 4.16). This discrepancy may well be chiefly caused by the unrealistic assumption of stationarity. Somewhat more formally, calculation of the empirical semivariograms in the four main directions serves to confirm this intuitive feeling that a nonconstant mean is more appropriate here.

4. Discussion

We have sought to explain some difficulties with the method of maximum likelihood estimation for the spatial linear model. It seems that the lack of a second derivative with respect to the range parameter for the spherical correlation scheme is associated with multimodality of the likelihood there, but, as a referee points out, this is not the complete story, as we can use a mixture to make (2·1) smoother. However, for small samples, even with a regular covariance scheme, the likelihood can be multimodal and it is worth examining a suitable profile log likelihood. Also (Warnes & Ripley, 1987) the scoring method may not work. There is considerable scope for further work.

Acknowledgements

The authors are grateful to Jeremy Warnes for preprints of Warnes & Ripley (1987), to Tim Hainsworth for his help, and to John Kent and the referees for their helpful comments.

References

BARNETT, V. D. (1966). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika* **53**, 151–65.

COPAS, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika* **62**, 701–4.

DAVIS, J. C. (1973). *Statistics and Data Analysis in Geology.* New York: Wiley.
MARDIA, K. V. (1980). Some statistical inference problems in kriging: II Theory. In *Proceedings of the 26th International Geology Congress Sciences de la Terre* Serie "Informatique Geologique" **15**, 113-31.
MARDIA, K. V. & MARSHALL, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-46.
MARTIN, R. J. (1979). A subclass of lattice processes applied to a problem in planar sampling. *Biometrika* **66**, 209-17.
REEDS, J. A. (1985). Asymptotic number of roots of Cauchy location likelihood equations. *Ann. Statist.* **13**, 775-84.
RIPLEY, B. D. (1981). *Spatial Statistics.* New York: Wiley.
WARNES, J. J. (1986). A sensitivity analysis for universal kriging. *Math. Geol.* **18**, 653-76.
WARNES, J. J. & RIPLEY, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika* **74**, 640-2.

[*Received July* 1987. *Revised October* 1988]