THE UNIVERSITY OF CHICAGO

INFERENCE FOR SPATIAL GAUSSIAN RANDOM FIELDS WHEN THE OBJECTIVE IS PREDICTION

A DISSERTATION SUBMITTED TO
THE FACULTY OF DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
MARK STEPHEN HANDCOCK

CHICAGO, ILLINOIS
AUGUST 1989

ACKNOWLEDGMENTS

I would like to thank my advisor, Michael L. Stein, whose vigor and insight have been an inspiration to me and whose knowledge and patience were essential to this work. I am also indebted to the members of the Spatial Statistics & Time Series Interest group for their valuable comments and Ronald A. Thisted for helpful advice on numerical linear algebra.

I am grateful to Betty J. Flehinger and the members of the statistics group at T. J. Watson Research Center for their encouragement and the support of IBM through an 1986-88 Predoctoral Fellowship. Computations for this thesis were performed using computer facilities supported in part by the National Science Foundation under grant DMS 87-03942 awarded to the Department of Statistics at the University of Chicago, and by The University of Chicago Block Fund.

Finally, I am indebted to the faculty, fellow students and staff at the Department of Statistics for the many facets of a graduate education.

TABLE OF CONTENTS

Chapter

Chapter

Chapter

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS

ABSTRACT

This thesis is concerned with aspects of statistical inference for Gaussian random fields when the ultimate objective is prediction. If we wish to predict a value for a random field at unobserved locations in a bounded region it is essential to have knowledge about the covariance function for that field. This knowledge is important both for the quality of prediction and for the assessment of the quality of prediction. The perspective taken in this thesis is that a predictor is incomplete without an associated measure of uncertainty. That is, obtaining a good estimate of the prediction variance is as much of a concern as obtaining a good prediction.

Stein (1988a) has shown that the impact on the best linear unbiased predictor from not using the correct covariance function is asymptotically negligible as the number of observations increases if the covariance function used is "compatible", in a well defined sense, with the actual covariance function on the region of interest. The concept of compatibility thus plays a central role in determining how a covariance function should be estimated.

This thesis concentrates on likelihood based inference for the covariance function when the random field is known to be Gaussian and the mean function has known form. We find that the likelihood statistic tells us as much about the model chosen as the data we are analyzing. If we use a peculiar model the likelihood statistic will indicate this by exhibiting peculiar behavior. If we have difficulty accepting a peculiar likelihood, we should choose a different model. Choosing to ignore the likelihood and using an alternative estimation procedure for the same peculiar model will not make the peculiar behavior exhibited by the likelihoods go away. We address numerical issues that arise when an irregular likelihood is used for inference.

For spatial random fields observed on a fixed region, an increasing density asymptotic framework may lead to bounded information for some of the parameters. We investigate the approximation of discrete observation in a fixed interval by continuous observation on the same interval. We find that the distribution of the maximum likelihood estimates are well approximated by their continuous versions when the range of correlation is comparable to the length of the segment, in a sense made precise. In this setting the empirical covariance function fails as an effective estimator of procurable characteristics of the covariance function. We investigate the empirical spectral density and find that it also fails in a surprising manner.

We analyse the best linear unbiased prediction procedure from within a Bayesian framework. The objective is to monitor the performance of the procedure when the underlying model is misspecified. Particular attention is paid to the treatment of parameters in the covariance structure and their effect on the quality, both actual and perceived, of the prediction.

These ideas are implemented using topographical data from Davis (1973) as a forum.

Key words: Interpolation, Likelihood, Geostatistics, Compatibility, Robustness

CHAPTER 1

INTRODUCTION AND COMPATIBILITY

1.1 Introduction and Motivation

This thesis is concerned with aspects of statistical inference for Gaussian random fields when the ultimate objective is prediction. If we wish to predict a value for a random field at unobserved locations in a bounded region it is essential to have knowledge about the covariance function for that field. This knowledge is important both for the quality of prediction and for the assessment of the quality of prediction. This thesis concentrates on inference for the covariance function when the random field is known to be Gaussian and the mean function is of a known form.

The motivating problem was the prediction of ore grades within the Mount Charlotte gold mine on the basis of samples taken from drill cores. The samples were spatially distributed along drill holes that spanned the geologically defined ore body. The ore was mined in blocks of about 10 meters in dimension. The problem faced by the mining engineers was the estimation of the average grade, in grams of gold per tonne, within each of these blocks. This information is essential in deciding where and how much to mine to maintain a fixed average grade. Their standard procedure to determine overall tonnage was the volume weighted sample tonnage within the ore body. This estimate was typically within 5% of the of the overall tonnage determined when the ore was milled. The ore body extended far deeper than the current drilling and mining level. A drilling scheme was planned to decide if mining should continued. A principal question was how to design the scheme so that the ore distribution and overall grade could be predicted with a moderate amount of drilling. Our approach was to model the spatial distribution of the ore grade by a random field and predict the ore grades at unknown locations. The random field can then be used as a foundation for answering the questions posed by the mining engineers. The statistical issues raised by this formulation stimulated this thesis.

There are a multitude of scientific fields that now use the random field model as a basis for prediction. We mention Forestry (Matérn (1960)), Meteorology (Gandin (1963)), and Hydrology (Kitanidis (1983)). Of special note is the mathematical work for the French Geostatistical school led by Matheron (1963, 1973). They have independently developed methodology under the name 'kriging'. Much of this methodology has grown with little cross fertilization between fields and detached from the mainstream statistical community.

The statistical framework addressed in this thesis has a wide basis outside of spatial settings. A general problem is the modeling of a response as a function of input variables. Often the determination of the response for each set of input values is expensive or restricted. A usual objective is to predict the value of the response at additional values of the inputs based on observing the response at a moderate number of inputs. Sacks, Shiller & Welch (1989) consider the prediction problem when the response is the outcome of a complex computer model for chemical kinetics problems. In these cases the stochastic basis for the response is the statistician's lack of knowledge of the underlying physical model. This approach has been applied to the design of VLSI circuits and for numerical integration. The approach is a useful extension to the classical linear model in the design of experiments. As identical issues of statistical inference for the covariance structure arise, this thesis should be of interest to researchers in these areas.

The important design issue of the optimal choice of input locations is not discussed in this thesis. See, for example, Sacks & Ylvisaker (1966, 1968, 1970, 1971), Cambanis (1985) and Sacks & Schiller (1988).

The balance of this chapter concerns the statistical framework. The next sections specify the problem and discuss issues that have arisen in the literature such as the importance of the Gaussian assumption to a theory based only on the second order properties and asymptotic perspectives. A guiding concept is that of the 'compatibility' of covariance functions defined by Stein (1988a). The bulk of this chapter discusses easily verifiable conditions for compatibility.

1.2 The statistical formulation of the prediction problem

We conceptualize a quantity of interest $Z(x)$ at each location x in a region $R \subset \mathbb{R}^d$. As discussed above, we choose to take a stochastic view and let each $Z(x)$ be a random variable so that $Z(\cdot)$ is a random field. The stochastic nature could be due to physical sources, or provide a surrogate for the statistician's lack of knowledge. There are clearly philosophical issues that need to be addressed in each area of application of the statistical theory. Some authors appear to have fixated on this issue. See Philip & Watson (1986a,b,c).

Suppose $Z(x)$ is a real–valued Gaussian random field on R with mean

$$\mathbb{E}Z(x) = \beta' f(x), \tag{1.2.1}$$

where $f(x) = (f_1(x), \ldots, f_q(x))'$ is a known vector function, β is a vector of unknown regression coefficients, and covariance function

$$\mathrm{Cov}(Z(x), Z(x')) = \alpha K_\theta(x, x') \qquad \text{for } x, x' \in R$$

where $\alpha > 0$ is a scale parameter, $\theta \in \Theta$ is a $p \times 1$ vector of structural parameters and Θ is an open set in $\mathbb{R}^p$. The division is purely formal as θ may also determine aspects of scale. We observe, from a single realization, $\{Z(x_1),\ldots,Z(x_n)\} = Z'$ and, will focus on the prediction of $Z(x_0)$. If L_i is any linear functional of the field $Z(x)$ then the extension to observing $L_1,\ldots,L_n$ and predicting L_0 is, from a theoretical standpoint, straightforward. We will focus on the class of predictors that are linear combinations of the data of the form

$$\sum_{i=1}^{N} \lambda_i(\theta) Z(x_i).$$

The best linear unbiased (BLU) predictor, $\widehat{Z}_\theta(x_0)$, is the unbiased linear predictor that minimizes the variance of the prediction error. It is straightforward to show that the corresponding weight vector $\lambda(\theta)$ defining $\widehat{Z}_\theta(x_0)$ is given by

$$\lambda(\theta) = {K_\theta}^{-1} k_0 + {K_\theta}^{-1} F (F' {K_\theta}^{-1} F)^{-1} b_\theta, \tag{1.2.2}$$

where

$$F = \{f_j(x_i)\}_{n\times q},$$

$$k_0 = \{K_\theta(x_0, x_i)\}_{n\times 1},$$

$$K_\theta = \{K_\theta(x_i, x_j)\}_{n\times n},$$

$$b_\theta = f(x_0) - F' {K_\theta}^{-1} k_0.$$

It will be assumed that F and C have full rank. This theory is developed in, for example, Goldberger (1962). The underlying field need not be Gaussian for the predictor to be a BLU.

The quality of the prediction is determined by the distribution of the prediction error, $e_\theta(x_0) = Z(x_0) - \widehat{Z}_\theta(x_0)$. Note that neither the predictor nor the prediction error depend on α or β. This has been used to argue for methods of estimation for the covariance structure that do not depend on the unknown β. Of course, had we additional knowledge about β beyond the data this information should be used.

If we wish to base inference on a single realization of a random field then additional structure is necessary. The natural assumptions are based on symmetry.

We concentrate on homogeneity and isotropy of the covariance function, that is, $K(x, y) \equiv K(|x - y|),\ \forall x, y \in R$, so that the class is usually written as a function of a single scalar variable $K(x)\,, x \in \mathbb{R}$. These constraints can be weakened in natural ways. For example, Matheron (1973) proposes a class of Intrinsic random functions based on Generalized covariance functions. The idea is that only certain increments of the random field are required to be stationary. Another possibility is to consider geometric anisotropies (Journel & Huijbregts, 1978, p. 177) where $K(x) \equiv K(|Vx|)$ for a possibly unknown matrix V. Such approaches extend the coverage of homogeneous and isotropic random fields. However, sound approaches to unspecified complex nonstationarities in the random field have yet to be developed. An allied concern is the necessity for methods robust against data contamination (Cressie (1984)). These are intrinsically very difficult problems. It is hoped that progress in statistical inference for the isotropic and homogeneous random fields will provide a basis for inroads into these more problematic situations.

1.3 Why does the covariance theory focus on Gaussian fields?

In this section it is argued that if one studies a random field only through its mean and covariance characteristics, then this approach is highly suspect if the field is not Gaussian. This is not to argue that most random fields are Gaussian, or that analyzing non–Gaussian fields is improper, only that a different approach should be taken. A Gaussian random field is a field in which each finite subset is jointly Gaussian. Let N be a Poisson Process on $\mathbb{R}$ with mean 1 and observed discontinuity points $t_1, t_2, \ldots$

Consider the following three random fields.

A. Random Telegraph

If $2t_{i-1} \leq x < 2t_i$ then

$$Z_A(x) = \begin{cases} 1 & \text{if } i \text{ even} \\ \text{-1} & \text{if } i \text{ odd} \end{cases}$$

B. Point Process with Adjoined Random Variables

Let $X_1, X_2, \ldots$ be an independent and identically distributed sequence with mean zero and unit variance and set $Z_B(x) = X_i \quad t_{i-1} \leq x < t_i$.

C. Ornstein–Uhlenbeck Process

Let $W(t)$ be Brownian Motion on $\mathbb{R}$ under the Ornstein-Uhlenbeck theory, that is, a mean zero Gaussian process with covariance function

$$\text{Cov}\{W(t_1),\ W(t_2)\} = \min(t_1,\ t_2) + (e^{-\min(t_1,\ t_2)} - 1),$$

where $t_1,\ t_2 \geq 0$. Under this theory, $W(t)$ is differentiable and we can denote its velocity by Z_C.

Each of Z_A, Z_B, and Z_C has mean zero and covariance structure $K(x,y) = e^{-|x-y|}$. However, the three processes have vastly different behaviors that should be taken to account in any analysis. If the field is very non–Gaussian, linear prediction is dubious. The field of interest is the conditional field given Z. Quantities derived from this field, such as its mean and variance, are not very tractable unless the field is Gaussian. The unconditional quantities are usually easy to work with, but less useful. We will focus on Gaussian random fields in this thesis. The interesting and harder problem of non–Gaussian fields will not be considered.

1.4 Compatibility of covariance functions

The concept of compatibility for covariance structures of random fields was defined by Stein (1988a). In this section we review the definition and implications for spatial inference and prediction. The compatibility of covariance functions is a central motivating idea in this thesis and will be the subject of the remainder of this chapter.

Stein (1988a) has shown that the impact on the best linear unbiased predictor from not using the correct covariance function is asymptotically negligible as the number of observations increases, if the covariance function used is "compatible" with the actual covariance function on the region of interest. The concept of compatibility plays a central role in determining how a covariance function should be estimated. Compatibility reflects the intuitively sensible concept that, for purposes of best linear unbiased interpolation, usually only the behavior near the origin of the covariance function is critical. Compatibility has been used extensively by Stein (1987b,c, 1988a,b) to investigate the effect of misspecifying the covariance structure of a random process on the BLU predictor.

Let $Z(x)$ be a continuous, not necessarily Gaussian, random field on a bounded region in $\mathbb{R}^d$ with mean function $m(x)$ and covariance function $K(x,y)$. We explicitly state that $Z(x)$ is a finite real-valued function taking values on a probability space $(\Omega, \mathcal{F}, P)$. It is well known that for any such mean function and covariance function we can produce a unique Gaussian random field with those characteristics. Let $[m, K]$ denote the unique Gaussian probability measure defined by $m(x)$ and $K(x,y)$. The nature of this measure requires some clarification. Let V be the space of all real-valued functions on R. Let $\mathcal{G}$ be the $\sigma-$ field of sets $A' \subset V$ that have inverse images A in $\mathcal{F}$. Now $Z(x)$ defines a mapping from $(\Omega, \mathcal{F}, P)$ into V. The measure $[m, K]$ is the one induced on V by

$$[m, K](A') = P(A) \qquad \forall A' \in \mathcal{G}.$$

That is, $(V, \mathcal{G}, [m, K])$ is a probability space and $Z(x)$ takes values in V in accordance with the probability measure $[m, K]$.

Recall that two measures P_0 and P_1 on a space $(\Omega, \mathcal{F})$ are mutually absolutely continuous if $A \in \mathcal{F}, P_0(A) = 0 \iff P_1(A) = 0$. This is denoted by $P_0 \sim P_1$. That is, if we observe $\omega \in \Omega$ we can not distinguish between P_0 and P_1 with probability 1. P_0 and P_1 are orthogonal if $\exists A \in \mathcal{F}$ with $P_0(A) = 0$ and $P_1(A) = 1$. That is, if we observe $\omega \in \Omega$ we can tell P_0 and P_1 apart with probability 1. Based on a finite sample from ω we should be able to distinguish between P_0 and P_1 with high probability.

While two general probability measures on $(V, \mathcal{G})$ may be neither mutually singular nor mutually absolutely continuous, any two measures on $(V, \mathcal{G})$ corresponding to Gaussian random fields are necessarily either mutually singular or mutually absolutely continuous (Hajek (1958), p. 615). This property can be used to show that

$$[m_0, K_0] \sim [m_1, K_1] \iff \begin{cases} [m_0 - m_1, K_1] \sim [0, K_1] \text{ and} \\ [0, K_0] \sim [0, K_1]. \end{cases}$$

Hence to establish mutual absolute continuity in the Gaussian case we may consider two simpler cases:

1) K_0 is identical to K_1, but m_0 differs from m_1, and

2) K_0 differs from K_1, but m_0 is identical to m_1.

We are now in a position to state the definition of compatibility.

Definition: Let K_0 and K_1 be covariance functions defined on R, a bounded region in $\mathbb{R}^d$. Then K_1 is **compatible** with K_0 on R if $[0, K_1]$ and $[0, K_0]$ are mutually absolutely continuous. We will continue the notation $[0, K_0] \sim [0, K_1]$. In addition, say that K_1 is scale compatible with K_0 on R if $\exists \alpha > 0$ such that K_1 is compatible with αK_0 on R.

It should be emphasized that compatibility is a property of covariance functions with respect to a given region. The random function $Z(x)$ is not required to be Gaussian.

The importance of this concept for spatial prediction follows from a result from Stein (1988b). Let $e_i^n(x_0)$ be the prediction error of the BLU predictor based on Z and the covariance structure $\alpha_i K_{\theta_i}$, $i = 0, 1$. Let $\mathbb{V}_i(\cdot)$ denote the variance operator under the covariance structure given by $\alpha_i K_{\theta_i}$.

Theorem 1.4.1 Stein (1988b):

Let $x_1, x_2, \ldots, x_n$ have x_0 as a limit point. Suppose K_{θ_1} is scale compatible with K_{θ_0} on R. If

$$\mathbb{V}_0[e_0^n(x_0)] \to 0 \tag{1.4.1}$$

as $n \to \infty$, then

$$\frac{\mathbb{V}_0[e_0^n(x_0)]}{\mathbb{V}_0[e_1^n(x_0)]} \to \quad 1 \tag{1.4.2}$$

$$\frac{\mathbb{V}_1[e_1^n(x_0)]}{\mathbb{V}_0[e_1^n(x_0)]} \to \quad \gamma \tag{1.4.3}$$

as $n \to \infty$. Here γ is the constant such that K_{θ_0} is compatible with γK_{θ_1} on R. ∎

If $\alpha_0 K_{\theta_0}$ is the actual covariance structure of the random field and we misspecify it by a covariance structure that is scale compatible with K_{θ_0} then (1.4.2) indicates we will still obtain an asymptotically efficient predictor of $Z(x_0)$. The condition (1.4.1) requires that the predictor based on K_{θ_0} be consistent for $Z(x_0)$. As x_0 is a limit point of the observations this places only mild conditions on K_{θ_0} (Stein (1987b)). From (1.4.3) we see that the ratio of the perceived variance under the misspecified covariance structure to the actual variance under the actual covariance structure approaches a constant. If we can determine this constant and so find a covariance structure compatible with the actual covariance structure, we will also obtain an asymptotically accurate value for the prediction variance $\mathbb{V}_0[e_0^n(x_0)]$.

The practical importance of this result is that it is not important, asymptotically, to discriminate between the actual covariance function of the random field and a compatible covariance function when the objective is prediction.

The asymptotic perspective taken can be thought of as 'fixed region' in the sense that the density of observations in a neighbourhood of x_0 increases. This is a natural asymptotic perspective for spatial random fields. In spatial applications the observations are usually taken from a well defined spatial region that is defined by external factors. In mining the region is defined by the geologic extent of homogeneous mineralization. In the topological example of Chapter 5 the region is the north face of the hill. If additional sampling is foreseen it will be taken from the same region, thus increasing the density of observation within the region rather than extending the region itself. This is not to say that the increasing region perspective is always inappropriate for spatial random fields, only that it is not the foremost perspective.

In time-series the increasing region asymptotic approach is natural. Usually additional sampling will occur at future time points, increasing the size of the region and maintaining the constant spacing between the locations of the observations. Based on this asymptotic approach the information about the covariance structure grows unboundedly, so that the entire function can be estimated consistently. For some time the increasing region perspective was the default choice for spatial random fields (Mardia & Marshall (1984)), however the fixed region perspective is now receiving more attention (Yakowitz & Szidarovszky (1985), Stein (1987a,b, 1988a)). Yakowitz and Szidarovszky (1985) view the problem from a non-parametric regression standpoint and have noted that it is not in general possible to get consistent estimates of the covariance function based on observations in a fixed region.

The real test of an asymptotic perspective is how well it approximates the situations and quantities of data typically seen in practice. Often more than one perspective is useful, although we emphasize that we regard the asymptotic solutions as approximations to reality and not vice versa. This issue will be investigated in Chapter 3.

The underlying theme is that only those parameters that matter for predictive purposes are able to be estimated very well. That is, if a parameter is difficult to determine then the predictive inference is insensitive to its value. This is not a dictum but we will see that it is the pattern for the prediction situations considered in this thesis.

Of course incompatible covariance functions may still give very similar predictions for a given data set, but the "more" incompatible covariance functions are, the greater the differences in the predictions tend to be, even on small data sets and odd geometries. The examples in the next chapter indicate that it really does matter, both from an estimation and a prediction perspective, that the class of covariance functions correspond to a realistic model. There is some empirical and theoretical evidence that likelihood based methods yield estimated covariance functions that are as "close" to compatible to the true covariance function as possible within the class considered.

1.5 Assessing compatibility of covariance functions

We have seen in the last section the theoretical importance of compatibility for the statistical inference of covariance structures. To determine if two covariance functions are compatible it is required to verify the mutual absolute continuity of two measures. This is a difficult problem as general conditions require the solution of integral equations. While the application of compatibility is a recent idea, the general question of absolute continuity of Gaussian measures has been considered since Hajek (1958) and Feldman (1958). Their interest was in the properties of Radon–Nikodym derivatives based on observing the associated process on continuous segments. Rozanov (1968) presents conditions for stationary processes in terms of their covariance functions and spectral densities. Skorokhod & Yadrenko (1973) extended many of these results to homogeneous random fields. Ibragimov & Rozanov (1978, III) provide an update to Rozanov (1968). The most recent survey is Yadrenko (1983, §3.3).

In the next section we highlight three of their important results which will be used in later sections. In the following sections we summarize research into easily verifiable conditions for the compatibility of covariance structures.

1.5.1 *Highlights of the compatibility results in the literature*

Let $Z_i(x), i = 1, 2$ be mean zero Gaussian random fields on a compact region R in $\mathbb{R}^d$. Suppose $Z_i(x)$ has homogeneous covariance function $K_i(x)$ and spectral density $f_i(\lambda), \lambda \in \mathbb{R}^d$. The basic result is:

Theorem 1.5.1 Skorokhod & Yadrenko (1973, Theorem 2):

Suppose $f_1(\lambda)$ is bounded on $\mathbb{R}^d$. Then K_1 is compatible with K_2 on R if, and only if, $K_1(x-y) - K_2(x-y)$, $x, y \in R$ can be extended to a function on $\mathbb{R}^d \times \mathbb{R}^d$ that

1) is square integrable on $\mathbb{R}^d \times \mathbb{R}^d$.

2) has a Fourier transform $\phi(\lambda_1, \lambda_2)$ satisfying

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{|\phi(\lambda_1, \lambda_2)|^2}{f_1(\lambda_1) f_2(\lambda_2)} d\lambda_1 d\lambda_2 < \infty$$

∎

These conditions are difficult to verify in particular cases. An easily verifiable sufficient condition can be derived from the approximation theory of functions (Nikolskii (1975)). Let $C^\infty(R)$ be the set of infinitely differentiable functions with support in R. Consider, $W_2^s(R)$, the Sobolev class of order s over R. This is the closure of $C^\infty(R)$ with respect to the metric

$$\|f\|_{W_2^s(R)} = \int_R |f(x)|^2 dx + \sum_{|\alpha|=s} \int_R |D^\alpha f(x)|^2 dx$$

when s is an integer, and in the metric

$$\|f\|_{W_2^s(R)} = \int_R |f(x)|^2 dx + \sum_{|\alpha|=s} \int_R \int_R \frac{|D^\alpha f(x) - D^\alpha f(y)|^2}{|x-y|^{d+2\gamma}} dx dy$$

when $s = [s] + \gamma, 0 < \gamma < 1$. Here α is an $d-$ tuple of non-negative integers, $|\alpha| = \sum_1^d \alpha_i$ and

$$D^\alpha f(x) = \frac{\partial^{|\alpha|} f(x)}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d}$$

As s increases the metrics bestow increasing smoothness on the functions through their partial derivatives. When s is non-integral natural Lipschitz conditions are placed on the partial derivatives. Crudely put, $W_2^s(R)$ consists of those functions that have square integrable partial derivatives at up to order s. We assume that the boundary of R is smooth enough so that the Sobolev condition can be extended beyond R. This is purely a technical condition (Nikolskii (1975, p. 381)). We take $f(\lambda) \asymp g(\lambda)$ to mean that there exists c_1, c_2 such that $0 \leq c_1 < f(\lambda)/g(\lambda) \leq c_2 < \infty$.

Theorem 1.5.2 Yadrenko (1983, §3.3, Theorem 3):

Suppose

$$f_1(\lambda) \asymp \frac{1}{(1+|\lambda|^2)^{p+d/2}} \qquad p > 0 \tag{1.5.1}$$

then K_1 and K_2 are compatible on R if $K_1(x) - K_2(x)$ is a member of $W_2^{2p+d}(R)$.

∎

The condition (1.5.1) covers a wide range, but not all, behaviors for the covariance structure. Most reasonable structures satisfy this condition. This result indicates that the covariance functions will be compatible if their difference is smooth enough. In particular this result implies that compatible covariance functions must have similar behaviors at the origin. In Skorokhod & Yadrenko (1973) this condition was erroneously given as both necessary and sufficient.

The final condition from the literature is an analogue of the previous result in terms of the spectral densities:

Theorem 1.5.3 Yadrenko (1983, §3.3, Theorem 4):

Suppose

$$f_1(\lambda) \asymp |\phi(\lambda)|^2 \qquad (1.5.2)$$

where $\phi(x)$ is the Fourier transform of a function square integrable in some neighbourhood of the origin. Then K_1 and K_2 are compatible on R if

$$\int_{\mathbb{R}^d} [1 - \frac{f_2(\lambda)}{f_1(\lambda)}]^2 d\lambda < \infty \qquad (1.5.3)$$

∎

Note that this condition is independent of the compact region R. The condition (1.5.2) is a technical smoothness condition on the spectral density. Spectral densities that satisfy (1.5.1) also satisfy (1.5.2). The result indicates that if the ratio of the spectral densities does not vary all that much in the tails then the two covariance structures will be compatible.

1.5.2 *Compatibility within the Vecchia class*

Vecchia (1985) presents a parametric covariance class for two dimensional random fields. The class is most directly defined through its two dimensional spectral densities:

$$f_{M,N}(\lambda) = \sigma^2 \frac{\prod_{j=1}^{r} ||\lambda|^2 + \theta_j|^{2n_j} \prod_{j=r+1}^{q} (|\lambda|^2 + \theta_j)^{n_j}}{\prod_{j=1}^{s} ||\lambda|^2 + \phi_j|^{2m_j} \prod_{j=s+1}^{p} (|\lambda|^2 + \phi_j)^{m_j}}$$

where $\{\theta_j\}_{j=1}^{r}$ and $\{\phi\}_{j=1}^{s}$ are taken to be strictly complex while $\{\theta_j\}_{j=r+1}^{q}$ and $\{\phi\}_{j=s+1}^{p}$ are real. Let

$$M = \sum_{j=1}^{s} 2m_j - \sum_{j=s+1}^{p} m_j \quad \text{and} \quad N = \sum_{j=1}^{r} 2n_j - \sum_{j=r+1}^{q} n_j.$$

The parameters are constrained so the $f_{M,N}(\lambda)$ is a valid spectral density. In particular $M - N \geq 2$. Figure 1 reproduces the shapes of some simple members.

In this section we show that the compatibility of members of this class are not influenced by the values of $\{\theta_j\}_{j=1}^{q}$ and $\{\phi\}_{j=1}^{p}$. As it is not asymptotically important to distinguish between compatible covariance functions a much smaller subset should suffice when the objective is prediction. It would also be difficult to identify these parameters that do not effect compatibility of the members (Stein (1987a)).

This class is a subset of the rational spectral densities in two dimensions. The corresponding covariance functions are linear combinations of modified Bessel functions of the second kind and integral order. The class is intended for use as a general model for two dimension fields. We need the following lemma, the necessity of which is clear from (1.5.3).

Lemma 1.1:

Consider the spectral densities on $\mathbb{R}^d$:

$$f_j(\lambda) \asymp \frac{\sigma_j^2}{(1+|\lambda|^2)^{m_j+d/2}} \qquad m_j > 0$$

for $j = 1, 2$. $f_1(\lambda)$ is compatible with $f_2(\lambda)$ if, and only if, $\sigma_1^2 = \sigma_2^2$ and $m_1 = m_2$.

Proof: We give a proof because such results are less obvious for $n > 1$. By Krasnitskii (1973), Corollary 3 it suffices to show that

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \prod_{j=1}^{d} \frac{\sin^2(\lambda_j - \mu_j)}{(\lambda_j - \mu_j)^2} \Big(1 - \frac{f_2(\lambda)}{f_1(\lambda)}\Big)\Big(1 - \frac{f_1(\mu)}{f_2(\mu)}\Big) d\lambda d\mu$$

diverges. But

$$\Big(1 - \frac{f_2(\lambda)}{f_1(\lambda)}\Big)\Big(1 - \frac{f_1(\mu)}{f_2(\mu)}\Big) = 2 - \frac{\sigma_1^2}{\sigma_2^2}\Big(1+|\mu|^2\Big)^{m_2-m_1} - \frac{\sigma_2^2}{\sigma_1^2}\Big(1+|\lambda|^2\Big)^{m_1-m_2}$$

so the integral diverges unless $\sigma_1^2 = \sigma_2^2$ and $m_1 = m_2$. ∎

Lemma 1.2:

The spectral density $f_{M,N}(\lambda)$ is compatible with

$$f(\lambda) = \frac{\sigma^2}{(1+|\lambda|^2)^{M-N}} \tag{1.5.4}$$

This follows directly from an application of (1.5.3).

Result 1.1:

Two members of the Vecchia class, $f_{M_1,N_1}(\lambda)$ and $f_{M_2,N_2}(\lambda)$, are scale compatible if, and only if, $M_1 - N_1 = M_2 - N_2$. ∎

This follows directly from the two lemmas. This result indicates that the only parameters that are asymptotically important for prediction are σ^2 and the difference $M-N$. Thus the array of parameters $\{\theta_j,\ n_j\}_{j=1}^{q}$ and $\{\phi_j,\ m_j\}_{j=1}^{p}$ may not provide a broad range of covariance structures. The reduction to the class defined by (1.5.4) would also reduce the number of free parameters to those that are important for both the actual and perceived quality of prediction.

1.5.3 *The Matérn class of covariance functions*

In this section we describe a general class of covariance functions that we feel provides a sound foundation for the parametric modeling of Gaussian random fields. This class will be used extensively in the later chapters. The spectral density on $\mathbb{R}^d$ has the general form:

$$f(\lambda) = \alpha \frac{\Gamma(\theta_2 + d/2)}{\Gamma(\theta_2)\pi^{d/2}} \cdot \frac{{\theta_1}^{d}}{(1+(\theta_1\lambda)^2)^{\theta_2+d/2}}$$

The corresponding isotropic covariance functions have the form:

$$K_\theta(x) = \frac{\alpha}{2^{\theta_2-1}\Gamma(\theta_2)} \left(\frac{x}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{x}{\theta_1}\right)$$

where $\alpha > 0$ is a variance parameter, $\theta_1 > 0$ is a scale parameter controlling the range of correlation and $\theta_2 > 0$ is the parameter controlling the smoothness of the field. $\mathcal{K}_{\theta_2}$ is the modified Bessel function of the second kind and order θ_2 discussed in Abramowitz & Stegun (1964), §9.

The motivation for this class is the wide range of behaviors that arise from the spectral density. A field with this covariance function is $[\theta_2)$ times (mean-square) differentiable, corresponding to continuous $[\theta_2) - 1$ derivatives. The Exponential class corresponds to the sub-class with smoothness parameter $\theta_2 = \frac{1}{2}$. The sub-class defined by $\theta_2 = 1$ was introduced by Whittle (1954) as a model for two dimensional fields. A general treatment is given in the seminal work by Matérn (1960).

The calculation of $\mathcal{K}_{\theta_2}$ for non-integral θ_2 is quite difficult. Fortunately there exist stable, optimized and accurate algorithms to calculate them. One publicly available version is Amos (1986). All calculations of $\mathcal{K}_{\theta_2}$ in this thesis use the Amos (1986) algorithm. While the calculation is expensive relative to the other forms of covariance functions, this cost is negligible compared to the other computing costs involved in the analysis.

Note that the covariance functions are always positive, so that the class is inappropriate for fields with negative correlations. This is quite rare in spatial settings.

The class captures a wide range of behaviors at the origin. Let

$$c(\theta) = \frac{1}{(2\theta_1)^{2\theta_2}\Gamma(\theta_2)\Gamma(\theta_2+1)}.$$

If θ_2 is an integer,

$$K_\theta(x) = 2\alpha(-1)^{\theta_2+1}c(\theta)\cdot x^{2\theta_2+2}\log x + \alpha\{\sum_{j=0}^{\infty} a_j(\theta)x^{2j}\} + \alpha x^{2\theta_2+4}\log x\{\sum_{j=0}^{\infty} b_j(\theta)x^{2j}\}$$

where $\{a_j(\theta)\}_{j=0}^{\infty}$ and $\{b_j(\theta)\}_{j=0}^{\infty}$ are functions of θ alone. If θ_2 is not an integer then,

$$K_\theta(x) = \frac{\pi\alpha c(\theta)}{\sin(\pi\theta_2)}\cdot x^{2\theta_2} + \alpha\{\sum_{j=0}^{\infty} d_j(\theta)x^{2j}\} + \alpha x^{2\theta_2+2}\{\sum_{j=0}^{\infty} f_j(\theta)x^{2j}\}$$

where $\{d_j(\theta)\}_{j=0}^{\infty}$ and $\{f_j(\theta)\}_{j=0}^{\infty}$ are functions of θ alone.

Based on Lemma 1.1, two members of the Matérn class are scale compatible if, and only if, they have the same smoothness parameter. While the range parameter does not affect the compatibility of a member it provides an adjustment for the spatial scale of the field.

If $Z(x)$ is a Gaussian random field on $\mathbb{R}^d$ with covariance function $K_\theta(x)$ then it satisfies the stochastic partial differential differential equation (Whittle (1963)):

$$[\ \theta_1^2 \sum_{j=1}^{d} (\partial^2/\partial^2 x_i) - 1\ ]^{\theta_2 + d/4} \cdot Z(x) = \sigma dW(x)$$

where $\sigma^2 = \alpha\Gamma(\theta_2 + d/2)(2\sqrt{\pi}\theta_1)^d/\Gamma(\theta_2)$ and $W(\cdot)$ is the d dimensional Wiener random field. If $\theta_2 + d/4$ is an integer then this gives a physical basis for the covariance. If $\theta_2 + d/4$ is not an integer then the interpretation is more problematical. This equation has motivated Jones (1989) to use the member with $\theta_2 = 1$ to model Aquifer Head data and is commonly used in Hydrology (Mejía & Rodríguez-Iturbe (1974), Creutin & Obled (1982)).

1.6 Summary and conclusions

In this introductory chapter we present the motivation and statistical formulation of the problem addressed by this thesis.

In §1.4 we review the concept of compatibility of covariance structures defined by Stein (1988a). Our approach to statistical inference for the covariance structure is influenced by this idea. In §1.5 we consider some conditions in the literature for assessing the compatibility of covariance structures. There are currently few easily verifiable conditions for compatibility for random fields in more than one dimension. The final section defines the Matérn class of covariance functions for general use in the modeling of Gaussian random fields. This class is used in the subsequent chapters.

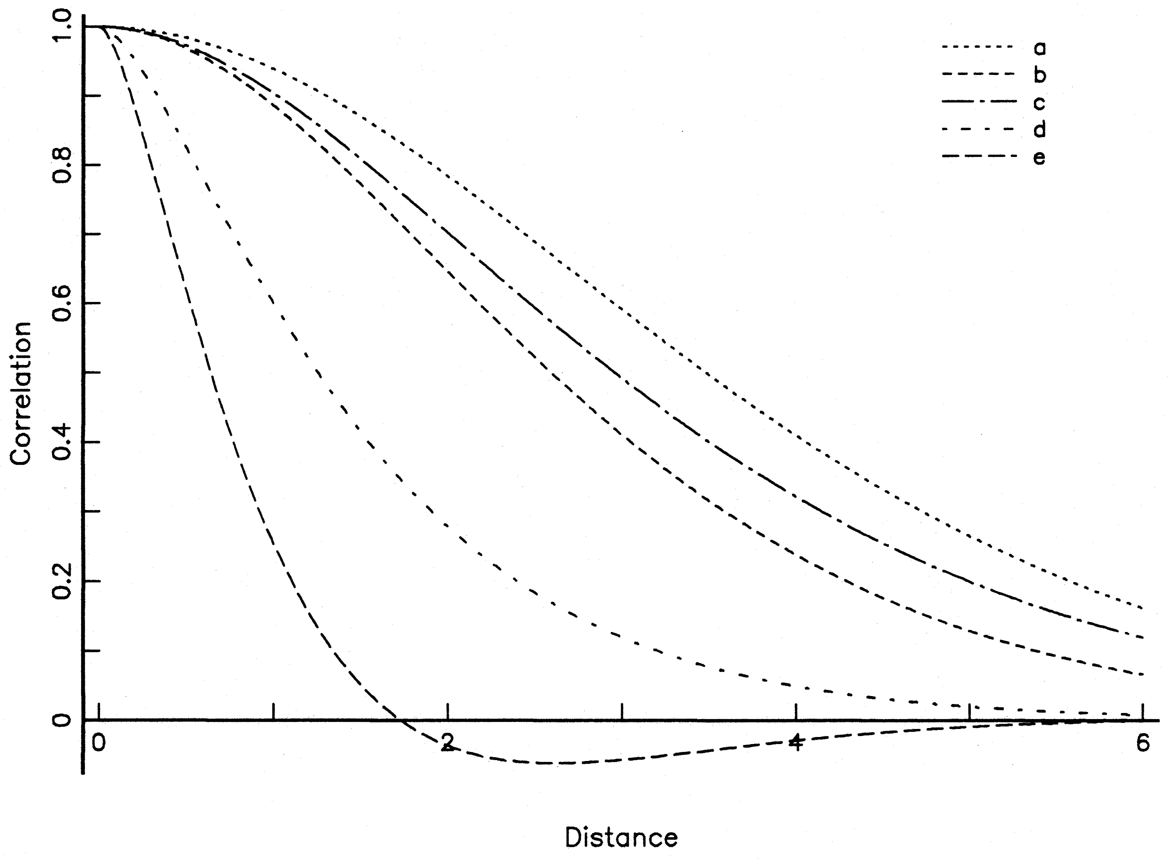


Fig. 1. Examples from Vecchia's correlation function class. The spectral densities corresponding to the correlation functions are: a) $1 / (1 + \lambda^2)^6$, b) $4(10 + \lambda^2)^2 / (1 + \lambda^2)^2(1 + 2\lambda^2)^2$, c) $1 / (1 + \lambda^2)^4$, d) $1 / (1 + \lambda^2)^2$, e) $\lambda^4 / (1 + \lambda^2)^4$.

CHAPTER 2

STATISTICAL INFERENCE FOR SPATIAL COVARIANCE STRUCTURES

2.1 Introduction

The central concern in spatial prediction by Gaussian random fields is the identification of the covariance structure. If the covariance structure is known, then in principle, the theory and practice of BLU prediction are straightforward.

In this chapter we will consider the various methods for inference for the covariance structure and develop methods based on the likelihood statistic.

In §2.2 the likelihood approach is described and arguments are made for the use of modified likelihoods. In §2.3, §2.4 and §2.5 we analyse two standard covariance classes for one dimensional random fields. Attention is given to both the increasing region and fixed region asymptotic approaches. The Exponential class is considered in §2.4 and the Triangular class in §2.5.

The Spherical class on the plane is considered in §2.6, where the existence of multiple modes is demonstrated and analyzed. Finally, §2.7 discusses the computational issues involved in maximum likelihood estimation for spatial random fields.

The bulk of this introduction is a critique of the traditional methods of inference for the covariance structure of Gaussian random fields. The notation used is the same as that defined in §1.2. We suppose $Z(x)$ is a real–valued stationary Gaussian random field on R with mean

$$\mathbb{E}Z(x) = \beta' f(x), \tag{2.1.1}$$

where $f(x) = (f_1(x), \dots, f_q(x))'$ is a known vector function, β is a vector of unknown regression coefficients, and covariance function

$$\text{Cov}(Z(x), Z(x')) = \alpha K_\theta(x, x') \qquad \text{for } x, x' \in R$$

where $\alpha > 0$ is a scale parameter, $\theta \in \Theta$ is a $q \times 1$ vector of structural parameters and Θ is an open set in $\mathbb{R}^p$. The division is purely formal as θ may also determine aspects of scale.

We observe, from a single realization, $\{Z(x_1), \dots, Z(x_n)\} = Z'$ and, as usual, will focus on the prediction of $Z(x_0)$. Some of the traditional methods implicitly take $f(x) \equiv 1$. It is possible to make adjustments to these methods if the mean of the field is a general regression function, although this will not be described here.

In kriging the covariance structure is parameterized by the variogram,[1] $\gamma(h)$, defined by

$$\gamma(h) = \alpha\{K_\theta(0) - K_\theta(h)\}.$$

Some of the methods discussed were originally developed for $\gamma(h)$.

The motivations usually given for the nonparametric approach are those for the empirical autocorrelation function in time-series. The major difference is that the spatial geometry of the observations plays a much more important role in random fields than it does in time-series where the geometry is simple and fixed. As a nonparametric estimator for $\gamma(h)$ Matheron (1963) considered the empirical variogram function

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i,j \in D(h)} \{Z(x_i) - Z(x_j)\}^2$$

where $D(h) = \{(i,j) : |x_i - x_j| = h\}$ and $N(h)$ is the number of pairs in $D(h)$. Although the unbiasedness of this appealing estimator is often touted, it often has

[1] In fact $\gamma(h)$ is the 'semi'-variogram and $2\gamma(h)$ is the variogram. This unfortunate notion will not be used in this thesis.

poor statistical properties. It is sensitive to deviation from the Gaussian assumption and the variability at different lags depends on the size of $D(h)$ and the locations of the observations. Alternative empirical estimators for $\gamma(h)$ have attempted to address the lack of robustness. Cressie & Hawkins (1980) and Cressie & Horton (1987) considered M-estimates, Armstrong & Delfiner (1980) suggest an estimator based on quantile estimation while Omre (1984) suggested a spatially weighted version of $\gamma(h)$. Similar comments apply to the empirical covariance and correlation functions. The major drawback of all these nonparametric estimators are the strong correlations between estimates at different lags making interpretation difficult.

The empirical estimate need not possess the essential properties expected in a covariance structure such as positive definiteness. In addition, the direct substitution of the empirical estimate for the true structure ignores the substantial sampling variability present in all but the largest sample sizes. The usual approach is to choose a parametric class and estimate the parameters for that class. This choice is always subjective. The traditional classes chosen are the Exponential, Spherical, Squared Exponential and sometimes the Triangular. The last three have implicit properties that would normally be considered physically unrealistic. The Exponential is a fine class for some one dimensional fields, although a larger class is desirable especially for multidimensional fields. The Exponential and Spherical classes are studied in this chapter.

Clarke (1979) suggests a non-statistical approach to parametric estimation by fitting the theoretical model to $\widehat{\gamma}(h)$ by eye. Given the correlated nature and differing variances in the points this approach can be very misleading. A common alternative is to use a least squares fit to $\widehat{\gamma}(h)$ (Journel & Huijbregts (1978)) or a weighted least squares method adjusting for the differing variances (Feinerman *et al.* (1986)). These methods have little statistical rationale.

Bastin & Gevers (1985) and Cressie (1985) suggest using generalized least squares to take into account the correlations between the lags. Let $y_{ij} = \frac{1}{2}(Z(x_i) - Z(x_j))^2$ and $d_{ij} = |x_i - x_j|$ for $i < j$. Then $\mathrm{Cov}(y_{ij}, y_{jk}) = 2\alpha^2\{K_\theta(d_{ij}) + K_\theta(d_{jk}) - K_\theta(d_{ik}) - K_\theta(d_{jl})\}^2$ Hence, given θ, we can calculate a minimum variance unbiased estimate of α by generalized least squares with respect to $\mathrm{Cov}(y_{ij}, y_{jk})$. We can go on to estimate θ by the value that produces the smallest value for the sum of squared deviations. Note that the value of θ is not itself a minimum variance unbiased estimator. Nor is it clear that minimum variance unbiased estimation of α and θ is preferable as the y_{ij} are non-Gaussian and generalized least squares requires the repeated inversion of an order $n(n-1)/2$ matrix.

Recent work has focused on optimization of global criteria in the cross-validated prediction errors. See for example Dowd (1984), Bastin & Gevers (1985) and Samper & Neuman (1989).

2.2 Likelihood based methods of inference

In this section likelihood methods for the inference of spatial covariance functions are developed. The approach was first applied in the Hydrological and Geological fields following Kitanidis (1983), Kitanidis & Lane (1985) and Hoeksema & Kitanidis (1985). Mardia & Marshall (1984) is a standard reference in the statistical literature. Assume the "true" covariance structure is $\alpha_0 K_{\theta_0}(\cdot,\cdot)$. Initially we do not assume anything about $K_\theta(\cdot,\cdot)$ beyond positive definiteness.

The log–likelihood of α, θ and β having observed Z is, up to an additive constant,

$$L(\alpha, \theta, \beta;\ Z) = -\frac{n}{2}\ln(\alpha) - \frac{1}{2}\ln(|K_\theta|) - \frac{1}{2\alpha}(Z - F\beta)' K_\theta^{-1}(Z - F\beta) \qquad (2.2.1)$$

where $K_\theta = \{K_\theta(x_i, x_j)\}_{n\times n}$ and the dependencies upon n have been suppressed.

For fixed θ, this is maximized over β by the generalized least squares estimator

$$\widehat{\beta}(\theta) = (F' K_\theta^{-1} F)^{-1} F' K_\theta^{-1} Z$$

where $F = \{f_j(x_i)\}_{n \times q}$. Given θ, the maximum likelihood estimate of α is

$$\widehat{\alpha}(\theta) = \frac{1}{n}(Z - F\widehat{\beta}(\theta))' K_\theta^{-1} (Z - F\widehat{\beta}(\theta))$$

and the profile log–likelihood

$$L_p(\theta;\ Z) \equiv L(\theta, \widehat{\alpha}(\theta), \widehat{\beta}(\theta);\ Z) \tag{2.2.2}$$

is maximized by $\widehat{\theta}$ if and only if $\widehat{\theta}$ maximizes (2.2.1), and in this case the maximum likelihood estimate of (α, β) is $(\widehat{\alpha}(\widehat{\theta}), \widehat{\beta}(\widehat{\theta}))$.

Define a **contrast** relative to F to be any linear combination of the data $Z'\mu$ such that $F'\mu = 0$. The weights μ will be called an **increment**. If we assume F is of full rank then clearly such contrasts exist. Let H_θ be the hat–matrix from the regression of Z on F assuming the covariance structure is given by $\alpha K_\theta(\cdot, \cdot)$. That is, $H_\theta = I - F(F' K_\theta^{-1} F)^{-1} F' K_\theta^{-1}$. Now $H_\theta Z = Z - F\widehat{\beta}(\theta)$ is a particular set of n contrasts relative to F, and the maximum likelihood estimate of θ_0 depends on the data only through these contrasts. Let $\bar{H}_\theta$ be derived from H_θ by dropping any q rows. As H_θ has rank $n - q$, $\bar{H}_\theta$ will have linearly independent rows and, $F'\bar{H}_\theta = 0$ as $F'H_\theta = 0$. Hence $Z^c = \bar{H}_\theta Z$ is a particular set of $n - q$ contrasts relative to F. By definition a contrast is unaffected by the addition of a mean to the underlying field of the form γF for any γ. Hence one might base inference about the covariance structure on Z^c instead of Z. This is the Modified Maximum Likelihood of Patterson & Thompson (1974). They argue that, if β is unknown, then there is no information loss in going from Z to Z^c as Z^c is marginally sufficient for α and θ.

The distribution of Z^c is $N(0, \alpha\bar{H}_\theta K_{\theta_0}\bar{H}_\theta')$. Hence, for the purposes of estimation, it is no longer necessary to completely specify $\{\alpha K_\theta(\cdot,\cdot)\}$, but it suffices to specify a class $\{G(\cdot,\cdot;\theta')\}$ such that for each θ there exists a θ' such that $C(G_{\theta'} - K_\theta)C' = 0$. This idea has taken form in the concept of generalized covariances (Matheron (1973, 1974)). The theory is based on a particular case of the generalized random functions with stationary increments of Gel'fand & Vilenkin (1964, §3.5). As an example suppose that $Z(x)$ is a random field on $\mathbb{R}$ with $f(x) = 1$ and covariance $\alpha K_\theta(x,y;\theta) = \alpha(\theta - |x-y|)^+$ where $\theta > 1$. We observe the field at $0, \frac{1}{n}, \ldots, 1$. It is easy to see that $Y_i = Z(\frac{i}{n}) - Z(\frac{i-1}{n}), \quad i = 1,\ldots,n$ is a set of contrasts and $Y \sim N(0, \frac{2\alpha}{n}I)$, so that the covariance of the contrasts does not depend on θ. Hence we can consider the class of functions $\alpha G(x,y) = -\alpha|x-y|$ instead of $K_\theta(x,y)$. In fact $G(x,y)$ is minus the variogram corresponding to $K_\theta(x,y)$.

Another method closely related to maximum likelihood is Minimum Norm Quadratic Estimation. This is used when the covariance class is linear in its parameters and is discussed by Kitanidis (1986), Mardia & Marshall (1986) and Stein (1987a). It will not be considered here.

There has been much written about maximum likelihood estimation in non–regular settings. See, for example, Barnard (1967), Smith (1985), Cheng & Iles (1987) and Smith & Naylor (1987). The focus of attention has been distributional classes such as the three–parameter Weibull, gamma and lognormal. Often the numerical and statistical issues are intermixed. The modified profile likelihood approach of Barndorff–Nielsen (1983) and Cox & Reid (1987) appears to be one avenue of attack and the Bayesian approach another. There is little work evaluating the merits of profile and modified likelihood approaches for spatial data. The Bayesian approach is nagged by the choice of distributions prior to the data. The approaches differ in the philosophical treatment of nuisance parameters. The marginal posterior density of the parameters of interest is obtained by integrating over the nuisance parameters, while the profile likelihood function is obtained by maximizing with respect to the nuisance parameters. In non–regular settings these two methods can lead to different results. Smith & Naylor (1987) give examples and argue for the Bayesian approach. A Bayesian approach to the estimation of covariance structure for spatial processes is developed in Chapter 4.

In their recent article, Warnes & Ripley (1987) find that likelihood surfaces for certain covariance functions have irregular behavior, including multiple modes. They find local maxima far, in the euclidean sense, from values they regard as reasonable and, on this basis, claim that maximum likelihood is a perilous method for inferring the covariance structure of spatial random fields. The impetus for their claims are two examples, one real and one simulated. Some irregularities are of a numerical nature, others are due to the unrealistic nature of the models. We find that the occurrence of these irregularities is consistent with the models chosen. The behavior of the likelihood surfaces can be better understood if the compatibility classes of the models are considered. This perspective is developed in Stein (1987a, 1987b, 1988). While likelihood methods are not the complete solution to the problem, they are, in the author's view, one of the best methods available. In §2.6 the simulated example from Warnes & Ripley (1987) is revisited and the irregular behavior is shown to be a result of the model and not an artifact of the likelihood approach. The real example using topological data from Davis (1973) is considered in Chapter 5.

2.3 Two representative covariance classes in one dimension

In the next two sections we will analyze two common covariance models for random fields in $\mathbb{R}$ for the case where the observations are regularly spaced. The Exponential and Triangular covariance classes on $\mathbb{R}$ are,

$$K_E(|x-y|;\theta_1,\theta_2) = \theta_1\theta_2 e^{-|x-y|/\theta_2} \tag{2.3.1}$$

$$K_T(|x-y|;\theta_1,\theta_2) = \theta_1(\theta_2 - |x-y|)^+ \tag{2.3.2}$$

where $\theta_1 > 0$, $\theta_2 > 0$, $x,\ y \in \mathbb{R}$. This parameterization is chosen to emphasize that we wish to estimate the behavior at the origin well. The "slope", θ_1, is the slope at the origin of the function, which controls the smoothness of the implied random field. The "range", θ_2, changes the rate of decrease of the correlation with distance. For the Triangular class, points separated by distances greater than the range are uncorrelated. See Figure 2. The Exponential is extensively used in practice.

Initially one might expect that given an Exponential covariance we should be able to find a Triangular covariance leading to similar predictive properties. The obvious candidates are those members with the value of θ_1. However, the "kink" in the Triangular class makes every member of the Triangular class incompatible with all members of the Exponential class on regions that include neighborhoods of points θ_2 apart. This leads to very different statistical properties. In fact, it can be shown that each Triangular covariance is incompatible with every other Triangular covariance on intervals longer than the minimum θ_2. Exponentials are compatible if and only if they have the same slope parameter. The log spectral densities are given in Figure 3, where the asymptotes of the Triangular are marked with vertical lines. The discrete process created by observing a continuous process at the integers has bounded symmetric log spectral density given in Figure 4. The effect of the kink on the Triangular spectral density is to create these asymptotes that lead to the irregular behavior. Looking at the covariance function itself one might initially believe it to be a well behaved approximation to reality, much as linear relationships are in regression context. However, one would be very wary of using this function based on the above spectral density.

Let $W(t)$ represent the location of Brownian motion at time t on $\mathbb{R}$ with the convention that $W(0) = 0$. Both classes have a physical interpretation in terms of $W(t)$. The interpretation for the Exponential class is given in §1.3, case C. Under the Einstein-Smoluchowski theory of Brownian Motion, $W(t)$ can be taken to be a mean zero Gaussian process with covariance function $\mathrm{Cov}\{W(t_1),\ W(t_2)\} = \theta_1\theta_2 \min(t_1,\ t_2)$. The Triangular class is then the moving average of white noise, the formal derivative of $W(t)$,

$$Z(x) = \int_x^{x+1} dW(t) = W(x+1) - W(x)$$

If the field has a mean of the form (2.1.1) then, in practice, one can consider the likelihood profiled over β . That is

$$L_{pm}(\alpha, \theta;\ Z) \equiv L(\alpha, \theta, \widehat{\beta}(\theta);\ Z)$$

amounting to replacing Z in (2.2.1) by $\tilde{Z} = Z - F\widehat{\beta}(\theta)$ for each value of θ . While $\mathbb{E}(\tilde{Z}) = 0$, $\tilde{Z}$ does not have the covariance structure K_θ . Unless the conditional and marginal likelihoods over β are similar substantial information is lost in using the profile likelihood as a surrogate for the full likelihood. This issue can be finessed by basing inference on the modified likelihood. For simplicity the examples in the next sections will have mean zero.

2.4 Analysis of the Exponential covariance class in one dimension

In this section we derive explicit expressions for the unique maximum likelihood estimates for the Exponential model (2.3.1) in one dimension. Both the increasing and fixed region asymptotic situations are considered.

2.4.1 *Observations regularly spaced in an increasing region*

Suppose $Z(x)$ is a random field with covariance function of the form (2.3.1) and zero mean. Suppose we observe the random field at $Z = (Z(1),\ Z(2), \ldots,\ Z(n))'$ and wish to infer the values of θ_1 and θ_2. In this setting $Z(x)$ may be viewed as a zero mean $AR(1)$ process with covariance function

$$\mathrm{Cov}(\ Z(i),\ Z(j)\ ) = \theta_1\theta_2 e^{-|i-j|/\theta_2} = \theta_1\theta_2\rho^{|i-j|}$$

where $\rho = e^{-1/\theta_2}$, $0 \le \rho < 1$. Note that not every $AR(1)$ process is of this form as ρ is restricted to be non-negative. The log-likelihood is of the form (2.2.1) where

$$K_\theta = \{\theta_1\theta_2\rho^{|i-j|}\}_{n\times n}$$

and it is easy to show that the inverse of K_θ is the tridiagonal symmetric matrix

$$K_\theta^{-1} = \frac{-\ln\rho}{\theta_1(1-\rho^2)}\begin{pmatrix} 1 & -\rho & 0 & \ldots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \ldots \\ 0 & -\rho & 1+\rho^2 & -\rho & \ldots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & -\rho & 1+\rho^2 & -\rho \\ 0 & \ldots & 0 & -\rho & 1 \end{pmatrix}$$

and the determinant of K_θ is $|K_\theta| = |K_\theta^{-1}|^{-1} = (-1)^n{\theta_1}^n(1-\rho^2)^{n-1}/\ln^n\rho$. Thus the log–likelihood of θ_1 and ρ given Z may be written

$$L(\theta_1, \rho;\ Z) = \frac{n}{2}\ln(-\ln\rho) - \frac{n}{2}\ln\theta_1 - \frac{n-1}{2}\ln(1-\rho^2) + \frac{g_Z(\rho)\ln\rho}{\theta_1(1-\rho^2)} \qquad (2.4.1)$$

where $0 \le \rho < 1$, $\theta_1 > 0$ and,

$$g_Z(\rho) = a_2\rho^2 + a_1\rho + a_0,$$

$$a_2 = \sum_{i=2}^{n-1} Z^2(i), \qquad a_1 = -2\sum_{i=1}^{n-1} Z(i)Z(i+1), \qquad a_0 = \sum_{i=1}^{n} Z^2(i)$$

Note that the only dependence on Z is through $g_Z(\rho)$, $\{a_2,\ a_1,\ a_0\}$ is a set of minimal sufficient statistics and $L(\theta_1, \rho;\ Z)$ is twice continuously differentiable. It is straightforward to show that

$$\widehat{\theta}_1(\rho) = -\frac{g_Z(\rho)\ln\rho}{n(1-\rho^2)} \tag{2.4.2}$$

is the unique solution curve of $\partial L(\theta_1, \rho;\ Z)/\partial\rho = 0$ in $0 \leq \rho < 1$, and that $\partial L(\theta_1, \rho;\ Z)/\partial\theta_1 = 0$ and $\partial L(\theta_1, \rho;\ Z)/\partial\rho = 0 \iff$

$$n(1-\rho^2){g_Z}'(\rho) + 2\rho g_Z(\rho) = 0 \tag{2.4.3}$$

These are just the roots of the cubic equation :

$$f(\rho) = -2a_2(n-1)\rho^3 - a_1(n-2)\rho^2 + 2(a_0 + na_2)\rho + na_1$$

Hence to investigate the existence and uniqueness of maximum likelihood estimates we can focus on (2.4.3). This equation was derived by Koopmans (1942) and discussed in White (1961) and Anderson (1971).

If $a_1 \leq 0$ then it is easy to show that $f(\rho)$ has three real roots. Hence the unique real root of $f(\rho)$ in $[0,\ 1]$ is

$$\widehat{\rho} = \sqrt{3q}\sin(\nu) - \sqrt{q}\cos(\nu) - s \tag{2.4.4}$$

where

$$q = \frac{na_2 + a_0}{3(n-1)a_2} + s^2,$$

$$r = \frac{(n-2)a_0a_1 - n(2n-1)a_1a_2}{12(n-1)^2{a_2}^2} + s^3,$$

$$s = \frac{(n-2)a_1}{6(n-1)a_2},$$

$$\cos(3\nu) = -\sqrt{\frac{r^2}{q^3}}, \qquad \pi/2 \leq 3\nu \leq \pi$$

This explicit expression first appears in Beach & MacKinnon (1978) in the Econometric literature. Hasza (1980) also considers this stationary AR(1) process and proves (2.4.4) is the unique maximum. The material from here on appears to be novel.

On the other hand, if $a_1 > 0$ and $0 \le \rho < 1$

$$\frac{\partial L(\theta_1, \rho;\ Z)}{\partial \theta_1} \begin{cases} > 0 & \text{if } \theta_1 < \widehat{\theta}_1(\rho) \\ < 0 & \text{if } \theta_1 > \widehat{\theta}_1(\rho) \end{cases},$$

$$\frac{\partial L(\theta_1, \rho;\ Z)}{\partial \rho} < 0 \quad \text{if } \ \theta_1 > 0$$

so the likelihood increases as $\theta_1 \to \infty$, $\theta_2 \to 0$. The maximum likelihood covariance is

$$\widehat{\mathrm{Cov}}(Z(x), Z(y)) = \begin{cases} a_0/n & \text{if } x = y \\ 0 & \text{if } x \neq y, \end{cases} \tag{2.4.5}$$

that is, a stochastic process with independent observations and an appropriate variance, and that occurs on the edge of the parameter space. Figure 5 is an example of a typical log–likelihood surface for this situation.

How often does this occur? Now $a_1 \le 0$ if and only if the sample first autocorrelation coefficient, $\widehat{\rho}(1) \equiv -a_1/2a_0$ is negative. This event is unaffected by the true value of the slope, θ_1. The intuition is that $\rho = e^{-1/\theta_2}$ only takes on positive values, so when the natural sample quantity is negative the likelihood suggests the best solution is to take $\rho \downarrow 0$. The distribution of $\widehat{\rho}(1)$ has been extensively studied. It is easy to show that

$$\widehat{\rho}(1) \sim \mathrm{AN}(\rho, \frac{1-\rho^2}{n}).$$

In fact $\widehat{\rho}(1)$ has close to a χ^2 shape for n small and ρ not close to 0. Clearly, when the mean is assumed zero, we can regard $\widehat{\rho}(1)$ as the empirical uncentered correlation of $\{0,\ Z(1), \ldots,\ Z(n)\}$ and $\{Z(1), \ldots,\ Z(n),\ 0\}$ so a better distributional approximation should be achieved by Fisher's inverse hyperbolic tangent transform.

That is,

$$\tanh^{-1}(\widehat{\rho}(1)) \sim \mathrm{AN}\Big(\ \tanh^{-1}(\rho) + \frac{\rho}{2n}\{1 + \frac{5+\rho^2}{4n}\},\quad \frac{(1+\rho^2)^2}{n}\{1 + \frac{4-\rho^2}{2n}\}\Big)$$

As $\Pr(\text{ M.L.E is on the boundary }) = \Pr(\widehat{\rho}(1) \le 0) = \Pr(\ \tanh^{-1}(\widehat{\rho}(1)) \le 0)$ we obtain the approximation

$$\Pr(\ \text{ M.L.E is on the boundary }) \simeq 1 - \Phi\Big(\ \frac{\tanh^{-1}(\rho)\sqrt{n}}{1+\rho^2} + \frac{2\rho - \tanh^{-1}(\rho)(4-\rho^2)}{4(1+\rho^2)\sqrt{n}}\ \Big)$$

We can show empirically that this approximation is accurate to within 2% uniformly over all values of θ_2 for n as small as 5. It is easy to calculate numerically the exact values for any given n. The values for $n=5$ are given in Figure 6. Each $\triangle$ represents an evaluation of the above approximation at a selected value of the range.

What can be said about the statistical properties of the maximum likelihood estimate? The information matrix for θ, $I = \{I_{ij}\}_{2\times 2}$, can be obtained directly as

$$I_{ij} \equiv -\mathbb{E}\Big(\frac{\partial^2 L(\theta;\ Z)}{\partial\theta_i \partial\theta_j}\Big) = \frac{1}{2}\mathrm{tr}\Big(K_\theta^{-1}\frac{\partial K_\theta}{\partial\theta_i}K_\theta^{-1}\frac{\partial K_\theta}{\partial\theta_j}\Big)$$

where the differentiation is element-wise. The asymptotic variances can then be directly computed from the inverse of the information matrix $I^{-1} = \{I^{ij}\}_{2\times 2}$:

$$\begin{aligned}
\mathbb{V}(\widehat{\theta}_1) &\sim \frac{\theta_1^2\{2\rho^2(1+\rho^2)\ln^2\rho + 4\rho^2(1-\rho^2)\ln\rho + (1-\rho^2)^2\}}{\rho^2(1-\rho^2)\ln^2\rho} \cdot \frac{1}{n} + O(\frac{1}{n^2}),\\
\mathbb{V}(\widehat{\theta}_2) &\sim \frac{(1-\rho^2)}{\rho^2\ln^4\rho} \cdot \frac{1}{n} + \frac{(1-3\rho^2)}{\rho^2\ln^4\rho} \cdot \frac{1}{n^2} + O(\frac{1}{n^3}),\\
\mathrm{Corr}^2(\widehat{\theta}_2, \widehat{\theta}_1) &\sim \frac{(2\rho^2\ln\rho + 1 - \rho^2)^2}{(1-\rho^2)^2(2\rho^2\ln\rho(2+\ln\rho) + 2\rho^4(1-\ln\rho)^2 + (1-\rho^2)^2)} + O(\frac{1}{n})
\end{aligned} \tag{2.4.6}$$

For our purposes one is better served by considering $\ln(\widehat{\theta}_1)$ and $\ln(\widehat{\theta}_2)$, as it can be empirically shown that the log–likelihood surface under this parameterization is much closer to quadratic in the region of high density. Figure 7 is an example of a typical log–likelihood surface parameterized by slope and range with $a_1 \leq 0$, so the maximum likelihood estimate is in the interior of the parameter space. The maximum likelihood estimate, $\widehat{\theta} = (1.32, 0.48)$ is marked with a cross. Figure 8 is the corresponding surface using the parameterization $\ln(\theta_1)$ and $\ln(\theta_2)$. Note that a quadratic approximation to the surface near the maximum under this parameterization will be much better. This behavior is typical for n as little as 5. The corresponding variances are just

$$\begin{aligned} \mathbb{V}(\ \ln(\widehat{\theta}_1)\ ) &\sim \frac{2\rho^2(1+\rho^2)\ln^2\rho + 4\rho^2(1-\rho^2)\ln\rho + (1-\rho^2)^2}{\rho^2(1-\rho^2)\ln^2\rho} \cdot \frac{1}{n} + O(\frac{1}{n^2}), \\ \mathbb{V}(\ \ln(\widehat{\theta}_2)\ ) &\sim \frac{(1-\rho^2)}{\rho^2\ln^2\rho} \cdot \frac{1}{n} + \frac{(1-3\rho^2)}{\rho^2\ln^2\rho} \cdot \frac{1}{n^2} + O(\frac{1}{n^3}), \end{aligned} \tag{2.4.7}$$

$$\mathrm{Corr}^2\{\ln(\widehat{\theta}_2), \ln(\widehat{\theta}_1)\} \sim \frac{(2\rho^2\ln\rho + 1 - \rho^2)^2}{(1-\rho^2)^2(2\rho^2\ln\rho(2+\ln\rho) + 2\rho^4(1-\ln\rho)^2 + (1-\rho^2)^2)} + O(\frac{1}{n})$$

Empirically, the bias of $\ln(\widehat{\theta}_1)$ and $\ln(\widehat{\theta}_2)$ is a significant part of the MSE for small n.

To summarize, the above equations (2.4.6) or (2.4.7) can be used to construct (joint) confidence intervals for $\widehat{\theta}_1$ and $\widehat{\theta}_2$. The estimates are weakly consistent and asymptotically (jointly) Gaussian with mean θ and covariance matrix given by (2.4.6).

The statistical description in this section holds few surprises. Indeed we would have been surprised by any odd occurrence from such a standard setting. We now consider a small alteration that will produce a more interesting situation.

2.4.2 *Observations evenly spaced in a fixed region*

Suppose we observe the random field on some fixed interval instead of at $\{1,\ 2, \ldots,\ n\}$. That is we observe $Z = (Z(0),\ Z(1/(n-1)), \ldots,\ Z(1))'$ and again wish to infer the values of θ_1 and θ_2. Using the same notation as in 2.2 and $a_0,\ a_1,\ a_2$

based on the above Z,

$$K_\theta = \{\theta_1\theta_2 e^{-|i-j|/(n-1)\theta_2}\}_{n\times n} = \{\theta_1'\theta_2' e^{-|i-j|/\theta_2'}\}_{n\times n}$$

where $\theta_1' = \theta_1/(n-1)$, $\theta_2' = (n-1)\theta_2$ and $\rho' = e^{-1/\theta_2'} = \rho^{1/(n-1)}$. Hence the problem of inferring θ_1 and θ_2 based on Z is equivalent to the problem of inferring θ_1 and θ_2 based on $(Z(1),\ Z(2),\ldots,\ Z(n))$ from a random function with parameters θ_1' and θ_2'. Hence we can immediately state that if $a_1 > 0$ the maximum likelihood solution is given by (2.4.5), and if $a_1 \le 0$ the maximum likelihood solution is given by (2.4.6) and (2.4.2) which provide expressions for θ_1' and θ_2'. It is tempting to try to estimate $\mathbb{V}(\ \ln(\widehat{\theta}_1)\ ), \mathbb{V}(\ \ln(\widehat{\theta}_2)\ )$ and $\mathrm{Corr}^2(\widehat{\theta}_2, \widehat{\theta}_1)$ by substituting in θ_1' and θ_2' for θ_1 and θ_2 in (2.4.7), that is:

$$\begin{aligned}\mathbb{V}(\ \ln(\widehat{\theta}_1)\ ) &\approx \frac{2}{n} + O(\frac{1}{n^2}),\\ \mathbb{V}(\ \ln(\widehat{\theta}_2)\ ) &\approx \frac{2\theta_2}{1+\theta_2} + + \frac{2}{n} + O(\frac{1}{n^2}),\\ \mathrm{Corr}^2\{\ln(\widehat{\theta}_2),\ \ln(\widehat{\theta}_1)\} &\approx \frac{1+\theta_2}{\theta_2}\cdot\frac{1}{n} + + O(\frac{1}{n^2})\end{aligned} \tag{2.4.8}$$

However, the basis of the asymptotic expansions is increasing information and in the fixed region case we do not have increasing information for θ_2. Empirically the above leading terms can be used as rough approximations to the actual variances for n greater than 5 and θ_2 not too much less than 1. The approximations tend to underestimate the variances for n small. The smaller θ_2 the larger n needs to be for the approximation to be good. For θ_2 less than 0.1, say, the approximation in (2.4.8) is quite good for moderate n. The distributions of $\ln(\widehat{\theta}_2)$ and the quality of this approximation will be studied in §3.3.

For the fixed region situation the maximum likelihood estimate of the slope parameter is $\sqrt{n}$–consistent, while the range can not be obtained consistently. This is in sharp contrast to the growing region situation, in which both the slope and range are $\sqrt{n}$–consistent. The intuition is clear. In the growing region situation the information content of the data for both the slope and range increases steadily as the number of observations increases, while in the fixed region case the shrinking gap between observations allows the slope to be obtained easily while information about the range remains inaccessible, even under extensive sampling of the region. The results of Mardia & Marshall (1984) on inference for spatial regression problems just do not apply.

It is easy to show that $K_{E}(x;\theta_1,\theta_2)$ is compatible with $K_{E}(x;\theta_1',\theta_2') \iff \theta_1 = \theta_1'$. Hence, for the purposes of prediction, we only need to estimate well θ_1 (which we can), and not θ_2 (which we can not). This result is an example of a meta-theorem: "For the purposes of spatial interpolation you can only estimate well those functions of the parameters that are important for the purposes of prediction". This idea has been explored by Stein (1987a) for covariances in which the parameters appear linearly and by Dawid (1984) in the context of prequential forecasting.

The clear and comforting picture for the Exponential class will be contrasted with that of the Triangular class in the next section.

2.5 Analysis of the Triangular covariance class in one dimension

In this section we derive expressions for the likelihood for covariance structure for the Triangular model (2.3.2) in one dimension. Both the increasing and fixed region asymptotic situations are considered. The likelihood is complex with the geometry of observations determining the structure. Multiple modes are the norm, occurring both on a network of curves defined by the geometry of the observations and between these curves.

2.5.1 *Observations regularly spaced in an increasing region*

Suppose $Z(x)$ is a mean zero random field with covariance function of the form (2.3.2). Suppose we observe the random field at $Z = (Z(1),\ Z(2), \ldots,\ Z(n))'$ and wish to infer the values of θ_1 and θ_2. For instance, if we had three points in $\mathbb{R}$ at 0, $9\theta_2/10$ and $11\theta_2/10$ then the first is uncorrelated with the last even though it is correlated with the second and the second is strongly correlated with the third. This "cut-off" effect results in irregular and unrealistic behavior. The Triangular log-likelihood is of the form (2.2.1) where

$$K_\theta = \theta_1\theta_2 11' - \theta_1 D_\theta$$

where $1 = \{1\}_{n\times 1}$, $D_\theta = \{|i-j| \cdot \mathrm{Ind}(|i-j| < \theta_2)\}_{n\times n}$ and $\mathrm{Ind}(A)$ indicates if the event A has occurred. Let $[\theta_2)$ be the greatest integer less than θ_2 and $\zeta = \theta_2 - [\theta_2)$ and then write the log–likelihood of θ_1 and θ_2 given Z as $L_{[\theta_2)}(\theta_1, \theta_2;\ Z)$ to emphasize that the likelihood changes character at each point where θ_2 is an integer. Consider $L_{[\theta_2)}(\theta_1, \theta_2;\ Z)$ as a function of θ_1 and θ_2 for fixed $[\theta_2)$. It is not difficult to show that

(1) $|K_\theta|$ is of the form $\theta_1^n P_{[\theta_2)}(\zeta)$ where $P_{[\theta_2)}$ is a positive polynomial on $[0,1]$ and $P_{[\theta_2)}(1) = P_{[\theta_2)+1}(0)$ for $[\theta_2) = 0,\ 1, \ldots,\ n-1$

(2) $Z'K_\theta^{-1}Z$ is a quadratic form of Z of the form

$$\frac{1}{\theta_1}\sum_{i,j=1}^{n} Z(i)Z(j)\frac{Q^{ij}_{[\theta_2)}(\zeta)}{P_{[\theta_2)}(\zeta)}$$

where $Q^{ij}_{[\theta_2)}$ is a polynomial of order $n-1$.

It is straightforward to show that

$$\widehat{\theta}_1(\theta_2) = \frac{1}{n}\sum_{i,j=1}^{n} Z(i)Z(j)\frac{Q^{ij}_{[\theta_2)}(\zeta)}{P_{[\theta_2)}(\zeta)}$$

is the unique solution curve of $\partial L(\theta_1, \theta_2;\ Z)/\partial\theta_1 = 0$ and that $\partial L(\theta_1, \theta_2;\ Z)/\partial\theta_1 = 0$ & $\partial L(\theta_1, \theta_2;\ Z)/\partial\theta_2 = 0 \iff$

$$(n-1)P_{[\theta_2)}{}'(\zeta)\sum_{i,j=1}^{n} Z(i)Z(j)Q^{ij}_{[\theta_2)}(\zeta) - nP_{[\theta_2)}(\zeta)\sum_{i,j=1}^{n} Z(i)Z(j)Q^{ij}_{[\theta_2)}{}'(\zeta) = 0$$

That is, to solve the likelihood equations one needs to find the roots of n polynomials each possibly of order $2n-1$, as well as checking the points where $\zeta = 0$. In summary, the log–likelihood is a piecewise infinitely differentiable function with seams along the lines $\zeta = 0$, where it is continuous. Figure 9 is an example of a log–likelihood profile at $\theta_1 = 1$ for $n = 5$. Figure 11 represents the probability of a unimodal likelihood for $n = 5$, $\theta_1 = 1$ and various values of θ_2. In Figure 11 and Figure 12 the Δs demark individual 95% confidence intervals based on $50,000$ simulations at each value. Note that the different trends depend on $[\theta_2)$ and that the probability is constant for ranges less than 1 as the data points are independent. Figure 12 represents the probabilities of modes occurring at the integers, which decreases from 70% under independence to about 45% for a range of 6. While multiple modes within a segment are possible they comprise less than 0.1% of those that occur. Figures 13 and 14 represent a profile log–likelihood at $\theta_1 = 1$ for data from $\theta = (1,\ 3)$. It has three modes, two of which fall in [1, 2), the first at $\theta_2 = 1.228$ and the second at $\theta_2 = 1.998$ As n increases the likelihoods become smoother overall, but not at the local level. Figure 10 is an example of a log–likelihood profile at $\theta_1 = 1$ for $n = 10$.

The problems occur at two levels. The primary concern is statistical. What do such likelihoods tell us about the model? Is this behavior an artifact of the likelihood approach to inference, or is it due to peculiarities in the particular model? The likelihood reflects the fact that as the range crosses each integer value less than n the model radically reassesses the importance of relationships among the data. It should be noted that if $[\theta_2)$ can be determined, then as the likelihood function between $[\theta_2)$ and $[\theta_2) + 1$ is well behaved the maximum likelihood estimate can be easily determined. Consider the closely related modified maximum likelihood problem referred to in the introduction where we consider the likelihood of the contrasts $Y_i = Z(i+1) - Z(i), \quad i = 1, \ldots, n-1$. The stationary covariance function of Y is

$$\mathrm{Cov}(Y_i, Y_j) = \theta_1 \begin{cases} 2 & \text{if } \left|i-j\right| = 0 \\ -\zeta & \text{if } \left|i-j\right| = [\theta_2) \\ \zeta - 1 & \text{if } \left|i-j\right| = [\theta_2) + 1 \\ 0 & \text{other wise} \end{cases}$$

That is very peculiar; an isosceles triangle with height $-\theta_2$ and width 2 sliding along the distance axis. Each Y_i is independent of all but the $2-4$ contrasts a distance $[\theta_2)+1$ or $[\theta_2)$ away from it. For instance, if $\theta_2 > n-1$ then $Y_1, \ldots, Y_{n-1}$ are independent and distributed as $N(0, 2\theta_1)$ so there is no additional information about θ_1. If $n-1 > \theta_2 > n-2$ then $Y_2, \ldots, Y_{n-2}$ are independent and distributed as $N(0, 2\theta_2)$ and the only information about θ_2 is in the bivariate Gaussian pair $\{Y_1, Y_{n-1}\}$ as $\mathrm{Cov}(Y_1, Y_{n-1}) = -\theta_1\zeta$. This pattern continues as θ_2 decreases through each integer value.

In summary, the Triangular covariance class is truly peculiar and should be avoided at all costs, especially because alternatives , such as the Exponential, exist.

2.5.2 *Observations evenly spaced in a fixed region*

Suppose we observe the random field at $Z = (Z(0),\ Z(\ 1/(n-1)\ ), \ldots,\ Z(1))'$ and again wish to infer the values of θ_1 and θ_2. Using the same notation as in §2.5.1,

$$K_\theta = \left\{\theta_1\left(\theta_2 - \frac{|i-j|}{n-1}\right)^+\right\}_{n\times n} = \left\{\theta_1'\left(\theta_2' - |i-j|\right)^+\right\}_{n\times n}$$

where $\theta_1' = \theta_1/(n-1)$, $\theta_2' = (n-1)\theta_2$ Hence, as in §2.4.2, the problem of inferring θ_1 and θ_2 based on Z is equivalent to the problem of inferring θ_1 and θ_2 based on $(Z(1),\ Z(2), \ldots,\ Z(n))$ from a random function with parameters θ_1' and θ_2'. That is, the behavior of the Triangular class observed in §2.5.1 carries over to the bounded region case.

Finally, consider the following example used in Stein & Handcock (1988) . Suppose we wished to predict the random field at $n/(n-1)$ based on Z using the BLUP, $\widehat{Z}_{\theta_2}(n/(n-1))$, based on the Triangular with a range of 1 . In particular we have

$$\widehat{Z}_1\Big(\frac{n}{n-1}\Big) = \frac{2n-3}{2(n-1)}Z\big(1\big) - \frac{1}{2}Z\Big(\frac{1}{n-1}\Big) + \frac{n-2}{2(n-1)}Z\big(0\big),$$

so that no matter how densely we observe the field on $[0,\ 1]$, the weights on the two observations furthermost from the point to be predicted remain substantial, while values closer receive no weight whatsoever. In addition note that $\mathbb{V}\Big\{Z\Big(\frac{n}{n-1}\Big) - \widehat{Z}_1\Big(\frac{n}{n-1}\Big)\Big\} = (7n-9)\theta_1/2(n-1)^2$ while $\mathbb{V}\Big\{Z\Big(\frac{n}{n-1}\Big) - Z\big(1\big)\Big\} = 2\theta_1/(n-1)$, so that the weights placed on on the two extreme points have a strong influence on the predictor. Figures 15 and 16 give the weights on $Z(\frac{i-1}{n-1})$ as a function of the range, for $n = 5,\ 20$.

2.6 The Spherical class on an square grid

In this section we will analyse a model commonly used for geological and hydrological applications in $\mathbb{R}^2$ and $\mathbb{R}^3$, and the one considered in §2.0 of Warnes & Ripley (1987) . The Spherical covariance class is the direct generalization of the Triangular class to three dimensions. Let $W(x)$ be three dimensional white noise, and $Z(x)$ be the random field obtained by integrating $W(x)$ over a ball of radius $\frac{1}{2}\theta_2$ centered at x . Then the isotropic covariance of $Z(x)$ has the general form:

$$K_S(x;\theta_1,\theta_2) = \begin{cases} \theta_1\theta_2\{\frac{2}{3} - \frac{|x|}{\theta_2} + \frac{1}{3}(\frac{|x|}{\theta_2})^3\} & \text{if } |x| < \theta_2 \\ 0 & \text{if } |x| \geq \theta_2 \end{cases}$$

Again, the parameterization is chosen so that θ_1 is the slope at the origin and θ_2 is the range. It can be shown that the Triangular is not an isotropic covariance function in $\mathbb{R}^2$, the Spherical is not an isotropic covariance function in $\mathbb{R}^4$, and the Exponential is an isotropic covariance function in any number of dimensions. A

Triangular, Spherical or Exponential random field is mean-square continuous, but not mean–square differentiable. One might expect that the Spherical in $\mathbb{R}^3$ might have similar properties as the Triangular in $\mathbb{R}$, and in Stein & Handcock (1988) an analogous example to the one given above suggests that the Spherical model exhibits behavior that is physically unrealistic for most fields in $\mathbb{R}^3$. We shall see the likelihoods based on a Spherical covariance exhibits irregularities similar to those of the Triangular in one dimension. As $K_S(x)$ has a continuous derivative at $x = \theta_2$ it is not hard to show that the corresponding Spherical likelihood exists and has a continuous derivative. However, the second derivatives of the likelihood are discontinuous and this leads to multiple modes. The behavior of the likelihood is best explained through a typical case comparing it to the Exponential.

A mean zero Spherical random field with $\theta = (0.5, 3)$ was observed on a 6×6 grid with unit spacing. Figure 17 is the profile log–likelihood at θ_1 set to its conditional maximum likelihood estimate. The vertical lines represent distances that separate the observations. For example, $2.236 = \sqrt{1^2 + 2^2}$. Figure 18 is the corresponding log–likelihood for the Exponential. Figures 19 and 20 are the derivatives of the log–likelihood with respect to θ_2 for the Spherical and Exponential, respectively. Figures 21 and 22 are the second derivatives of the log–likelihood with respect to θ_2 for the Spherical and Exponential, respectively. Note the discontinuous nature of the Spherical compared to that of the Exponential.

In summary, analysis of the Spherical likelihood on a regular grid in $\mathbb{R}^2$ indicates that the important parameter can be estimated well, while the unimportant parameter is much more difficult to tie down. The use of local smoothing of the log–likelihood is not unreasonable in this setting, purely as a means of deemphasizing the ripples. We believe that the likelihood surface is highly informative about the structure of the Spherical model. These ripples are purely a consequence of using the

Spherical model and are not an indication of a fault with the likelihood approach to spatial data. The Spherical model should not be applied to a data set unless there is a strong *a priori* belief that it is the correct model for the underlying process.

2.7 Computational issues in the calculation and maximization of likelihoods

In this section we discuss computational issues in the calculation of likelihoods using the Spherical model as an example of the pitfalls involved. After this section was written, a paper by Mardia & Watkins (1989) that discussed likelihood estimation for the Spherical model was brought to my attention. The material in this section extends the material in their paper.

The usual technique used to maximize the likelihood is the simple Newton–Raphson algorithm: If $L : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, and $\theta_0 \in \mathbb{R}^d$ is a starting value then

(1) Solve $\nabla^2 L(\theta_k) s_k = -\nabla L(\theta_k)$ for s_k.

(2) Set $\theta_{k+1} = \theta_k + s_k$.

(3) If convergence, stop; otherwise $k \to k+1$; go to (1).

This algorithm has a great many advantages numerically and statistically. If θ_0 is sufficiently close to a local maximizer θ_m of $L(\theta)$ with $\nabla^2 L(\theta_m)$ non-singular and $\nabla^2 L(\theta)$ Lipschitz continuous at θ_m then $\theta_1,\ \theta_2, \ldots$ will converge q–quadratically to θ_m. However, unless much stronger conditions, such as $\nabla^2 L(\theta_k)$ being negative definite for $k = 0,\ 1, \ldots,$ are satisfied the algorithm may proceed to a saddlepoint or even a minimum where $\nabla L(\theta)$ is also zero. That is, the Newton–Raphson algorithm at each step goes to the critical point of the current locally quadratic model, regardless of whether this point is a minimizer, maximizer or a saddlepoint of the local model.

As the second derivative is not continuous for the Spherical likelihood, it is not surprising that for some realizations problems occur. Indeed in about 50% of the cases tested the Newton–Raphson algorithm did not converge. Note that this particular problem is purely numerical in nature and not statistical.

A useful enhancement is to use a simple step-length search in the direction defined by the Newton–Raphson algorithm. A local cubic along this direction was maximized using the first derivative information at the current point and the point suggested by the Newton–Raphson algorithm. A simple check is included to ensure the chosen point is not absurd, and if not the Newton–Raphson point is used. This method is much less affected by the discontinuous second derivative and has proved very reliable for the Spherical log–likelihood. A mean zero Spherical random field with $\theta = (0.5, 3)$ was observed on a 6×6 grid with unit spacings. Figure 23 is a contour of a typical log–likelihood based on data from a mean zero Spherical random field with $\theta = (0.5, 3)$ observed on a 6×6 grid with unit spacings. For this data–set the Newton–Raphson algorithm moved in close to the maximum and then was caught in a repeating oscillation from one side of the maximum to the other. The same two oscillation points occur for almost all reasonable starting values. The path of the Newton–Raphson algorithm for a particular starting value is marked by $\times$. Although the Algorithm does not converge, it does settle down to a region quite close to the maximum value. The cubic search method proceeds directly to the maximum. The path, using the same starting value as the Newton–Raphson algorithm, is marked with $+$. Note that each step of the cubic search requires twice as many function evaluations as the Newton–Raphson algorithm. Figure 24 shows the corresponding profile log–likelihood. Figures 25 and 26 show the contours of the derivative with respect to θ_1 and θ_2. We see that Newton–Raphson algorithm oscillating along the zero contour for the derivative with respect to θ_1 and across the zero contour for the derivative with respect to θ_2. The heart of the reason Newton–Raphson algorithm has difficulties can be seen in Figure 27, the second derivative of the log–likelihood with respect to θ_2. The Δ's correspond to the oscillation points for Newton–Raphson algorithm, and the $+$ is the actual maximum. This behavior is repeatedly seen in situations where Newton–Raphson algorithm does not converge. The cubic method overcomes this simple numerical complication, but is not relevant to the statistical considerations.

2.7.1 *The occurrence of multiple modes*

Multiple modes do occur. For example for the mean zero Spherical random field with $\theta = (0.5, 3)$ observed on a 6×6 grid with unit spacings the breakdown is:

Table 1

Modes for the Likelihood for a Spherical on 6×6 grid

$\theta = (0.5, 3)$

Number of Modes $=$	1	2	3	4	5
Percentage $=$	26%	51%	21%	2%	0%

These values are determined by direct simulation. The standard errors are all less than 2%. As the range increases the proportion of unimodal likelihoods increases. As the slope increases, for a fixed range, the proportion of unimodal likelihoods decreases as the variance of the random field increases. As the number of observations increases, with the same unit spacing, the number of unimodal likelihoods decreases. Mardia & Watkins (1989) in their Table 1 provide an extension this table for other values of θ_2 and n. A typical example of a multiple modal likelihood is give in Figure 28. The true values are $\theta = (0.5, 3)$, and the local maxima are marked. Figure 29 is the log–likelihood profile when the slope is set to its conditional maximum likelihood estimate.

The existence of the multiple modes could naively be taken as a condemnation of likelihood based methods of inference when the purpose is interpolation. However, recall that in one or two dimensions, $K_S(x;\theta_1,\theta_2)$ is compatible with $K_S(x;\theta_1',\theta_2') \iff \theta_1 = \theta_1'$, that is, the slopes are the same. Hence, θ_1 is important, asymptotically, for predictive purposes, while θ_2 is much less important. The predictive distributions are insensitive to changes in θ_2, but not θ_1. Note that while the perceived prediction variance is proportional to θ_1, the linear predictor is unaltered by changes in θ_1. In Figure 28, the modes are all on a ridge with $\theta_1 \approx 0.53$. Moreover the likelihood along this ridge is quite flat over a wide span of ranges, with small ripples causing the local maxima. The drop–off across the ridge is more marked. In contrast, the likelihood for the Exponential model has the same general shape, but no ripples. The interpretation given to the ripples is the same as those of the Triangular in one dimension: artifacts of the peculiar model placed on the random field. Note that the behavior of the likelihood is not nearly as severe as the Triangular likelihood.

2.7.2 *Cross validation based on the likelihoods*

The use of a parametric model for the covariance assumes that the actual covariance falls in that class. If the field is Gaussian and covariance class is correctly specified then we have argued for inference based on the likelihood function. A major concern is if the covariance class is misspecified. In this section we introduce a diagnostic method to help detect if this assumption is incorrect.

Model validation for spatial random fields differs from model validation from time-series due to the influence of the geometry of locations. In time-series, and to some extent regularly spaced spatial fields, the geometry is fixed and a known quantity. This simplifies the analysis as attention can focus on the observed values of the field. When the locations are irregularly spaced the observed geometry typically has a large impact on the inference. The motivation for our approach is that if the model is correctly specified then the cross-validated likelihoods should provide inference similar to the full data likelihood. The cross-validated likelihoods should also be sensitive to outliers and influential values. Our approach appears to be novel for random fields.

We consider a simple cross-validation comprising of fitting the covariance model to the n data sets obtained by excluding successively just one location. In each case we produce a 'cross-validated' likelihood function and summarize it by its maximum likelihood value. We then calculate the full data likelihood of each of these cross-validated maximum likelihood estimates and use this set of n statistics to check the consistency of the model. This is but one of many approaches that can be taken. Considering the cross-validated likelihoods themselves is not a good idea as these are calculated using different data. An unexplored alternative is to look at the maximum likelihood estimates under the data set with the excluded value replaced by the value predicted under a model based on the other $n-1$ values.

To observe how the likelihoods surfaces change under covariance models corresponding to different smoothnesses we consider the Exponential class and the Matérn class with $\theta_2 = \frac{3}{2}$. We then generate realizations from the mean-zero Gaussian random fields on 7×7 grids under both these models. Our objective is to discover features that will enable us to distinguish between the correct model and a misspecified model. Figure 30 represents the spatial distribution of the full log-likelihoods under the Exponential model at the cross-validated maximum likelihood estimates for an Exponential realization with range parameter 3. The eye is drawn to the location $(5,\ 6)$ with likelihood substantially less, 0.12, than the other values. It corresponds to an estimate for the range much higher than for the other values. Figure 31 represents the plot when the Matérn model with $\theta_2 = \frac{3}{2}$ is applied to the same data. The location $(5,\ 6)$ again corresponds to an inflated value for the range, although a couple of other values are also low.

This pattern is typical of the realizations: a single location usually had substantially lower likelihood than the others, and the likelihoods plots are similar for the two models. Similar behavior was observed when the realizations were generated under the Matérn model with $\theta_2 = \frac{3}{2}$. A single location usually had substantially lower likelihood than the rest. It typically was on the edge of the region, and in a corner. We have also investigated plotting the likelihood against the cross-validated maximum likelihood estimates, again not finding features to discriminate between the covariance classes. As we might expect from the data geometry, the pattern of likelihoods is not substantially altered by the addition of a planar mean to the model.

In summary, this simple approach used for two dimensional random fields observed on grids indicate that changes in the smoothness of the field will be difficult to detect. This is partly due to the regular geometry of the observations. We will study irregularly spaced data in §5.5.1.

2.8 Summary and conclusions

In this chapter we analyse likelihood based methods of inference for three covariance classes when the random filed has been observed regularly in one and two dimensions.

Our findings indicate that the likelihood is telling us as much about the model chosen as the data we are analyzing. If we use a peculiar model the likelihood statistic will indicate this by exhibiting peculiar behavior. If we have difficulty accepting a peculiar likelihood, we should choose a different model. Choosing to ignore the likelihood and using an alternative estimation procedure for the same peculiar model will not make the peculiar behavior exhibited by the likelihoods go away. A second issue is practical estimation. If you accept the likelihood and wish to use it for inference then the unusual behavior will be an obstacle. In §2.7 we address the numerical issues involved in the determination of the maximum likelihood estimate. It is essential to consider the entire likelihood function as a basis for inference. Reference to the maximum likelihood estimate in isolation is analogous to summarizing a distribution by its modal value alone. Often this is expedient, but only under particular asymptotic scenarios is it adequate. Inference for the models in §2.5 and §2.6 are prime examples. To what extent can a single value, maximum likelihood or otherwise, serve as a surrogate for the information in the likelihood about the covariance structure? In spatial problems it is unwise to use a single point value without an analysis of the likelihood surface to see if it is appropriate. One approach is to look at the likelihoods produced from data simulated under the assumed model to check if the maximum likelihood estimate is an appropriate summary. The shape of the likelihood for the observed data can then be compared to that of simulated likelihoods.

Many of Ripley's (1987, 1988) arguments against likelihood analysis of spatial random fields are based on the fallacy that likelihood and maximum likelihood inference are synonymous. The optimal properties of the maximum likelihood estimate are based on asymptotic arguments, which are descriptions of how the likelihood surface degenerates as the information about each of the parameters grows unboundedly. In this case the maximum likelihood estimate is a good surrogate for the entire likelihood surface. Our interest is in the sample sizes that occur in practice and under the asymptotic scheme(s) that is most appropriate.

If there is unbounded growth of the information in all the parameters it is natural to write down theorems confirming the maximum likelihood estimate is weakly consistent and uniformly asymptotically Gaussian. Mardia & Marshall (1984) provide theorems based on conditions from Sweeting (1980). The theorems also require the covariance function to be smooth and that the observed information satisfy an appropriate convergence property. These results do not apply to many asymptotic situations of interest, in particular to fixed region asymptotics where information for some parameters is not increasing. In particular, their results do not apply to the asymptotic situations in §2.4.2 and §2.5.2 because the information is bounded for the range parameter, θ_2. They also do not apply to the Spherical class in §2.6 because the smoothness condition is not satisfied.

The Exponential model indicates the pivotal nature of the asymptotic framework. It easy to fall in a trap of taking '$n \to \infty$' instead of thinking about the relationship of future observations to those presently available. For time–series situations the increasing region, unbounded information approach is usually appropriate. For spatial random fields observed in a fixed region, the increasing density perspective is often more appropriate. Of course asymptotic results are themselves reductions, and are only as interesting in as much as they tell us about behavior in the sample sizes we have in practice.

Another issue is the relative usefulness of the conditional, marginal and profile likelihoods. Experience indicates that profiling over the mean parameters β is reasonable. Profiling with respect to the scale parameter α is appropriate under the well behaved models such as those in the Matérn class. For eccentric models it does hide some of the features, as in Figures 28 and 29. If we profile over both α and β then typically only one or two structural parameters remain.

If we misspecify α we will misspecify the prediction variance proportionately. However the prediction weights, (1.2.2) will not be affected. If we misspecify θ then, in general, this will effect both the prediction weights and the perceived prediction variance. The perspective taken in this thesis is that a predictor is incomplete without an associated measure of uncertainty. That is, obtaining a good estimate of the prediction variance is as much of a concern as obtaining a good prediction.

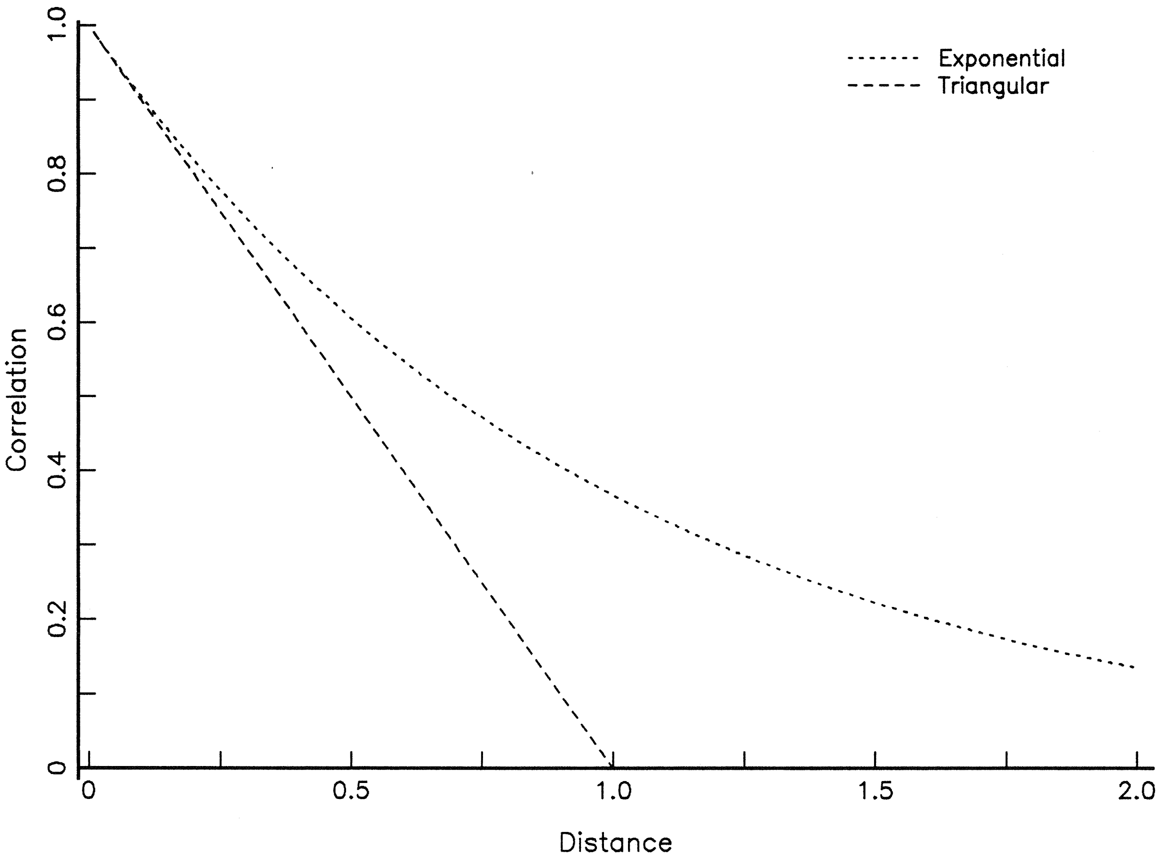


Fig. 2. The Exponential and Triangular correlation functions.

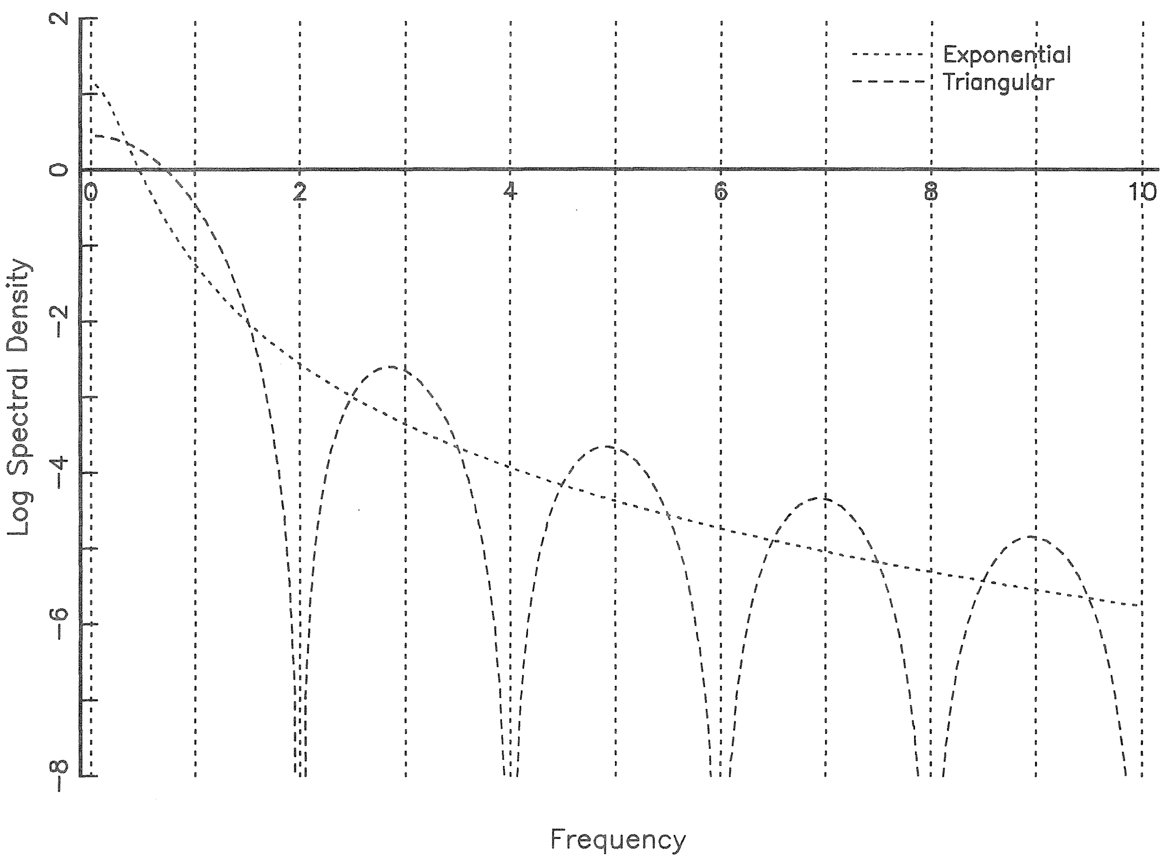


Fig. 3. The log spectral densities for the continuous processes. The frequency scale is in units of π.

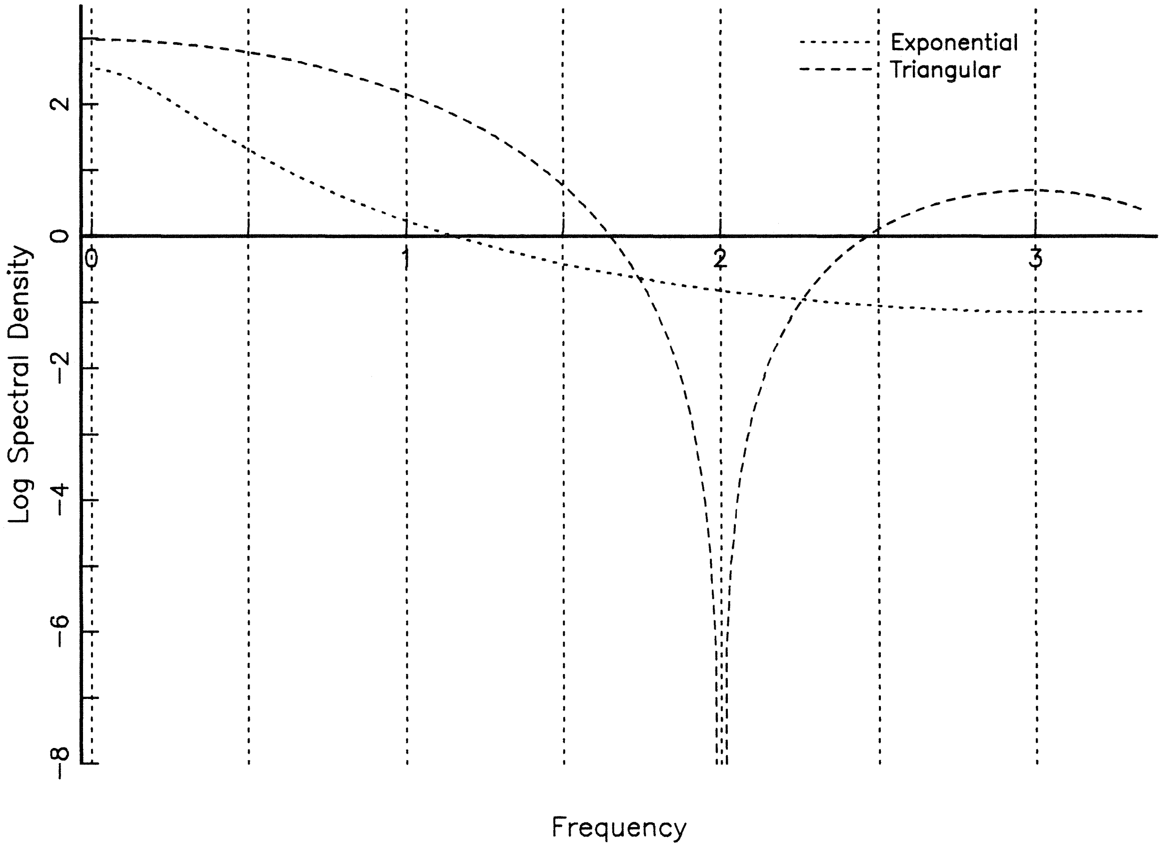


Fig. 4. The log spectral densities for the discrete processes. The frequency scale is in units of π.

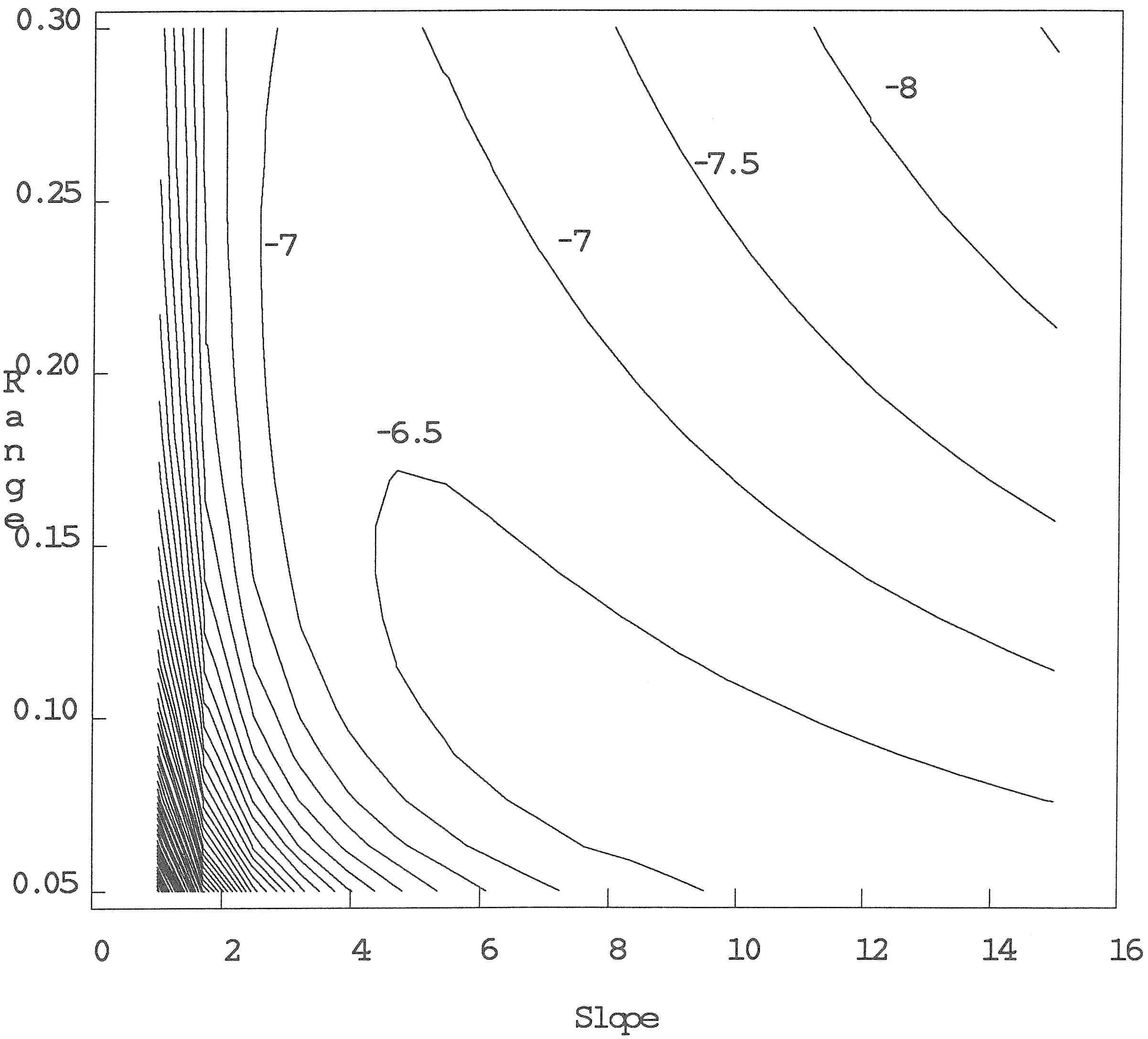


Fig. 5. Typical log-likelihood contour with the MLE identified on the boundary. The process was observed at 0, 0.25, 0.5, 0.75 and 1. The true slope is 5 and the true range is 0.2.

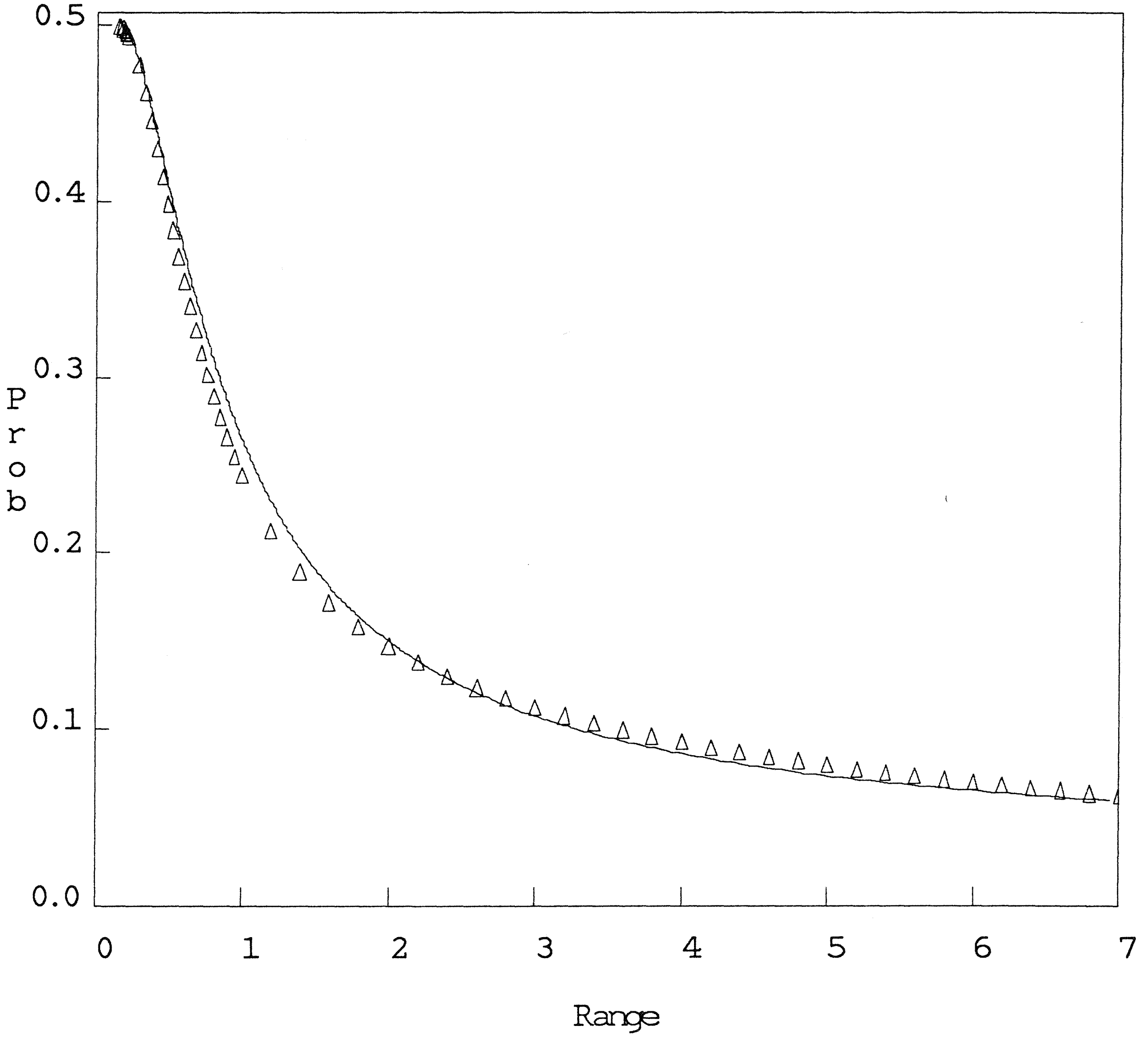


Fig. 6. Probability of the MLE occurring on the boundary as a function of the range for n = 5.

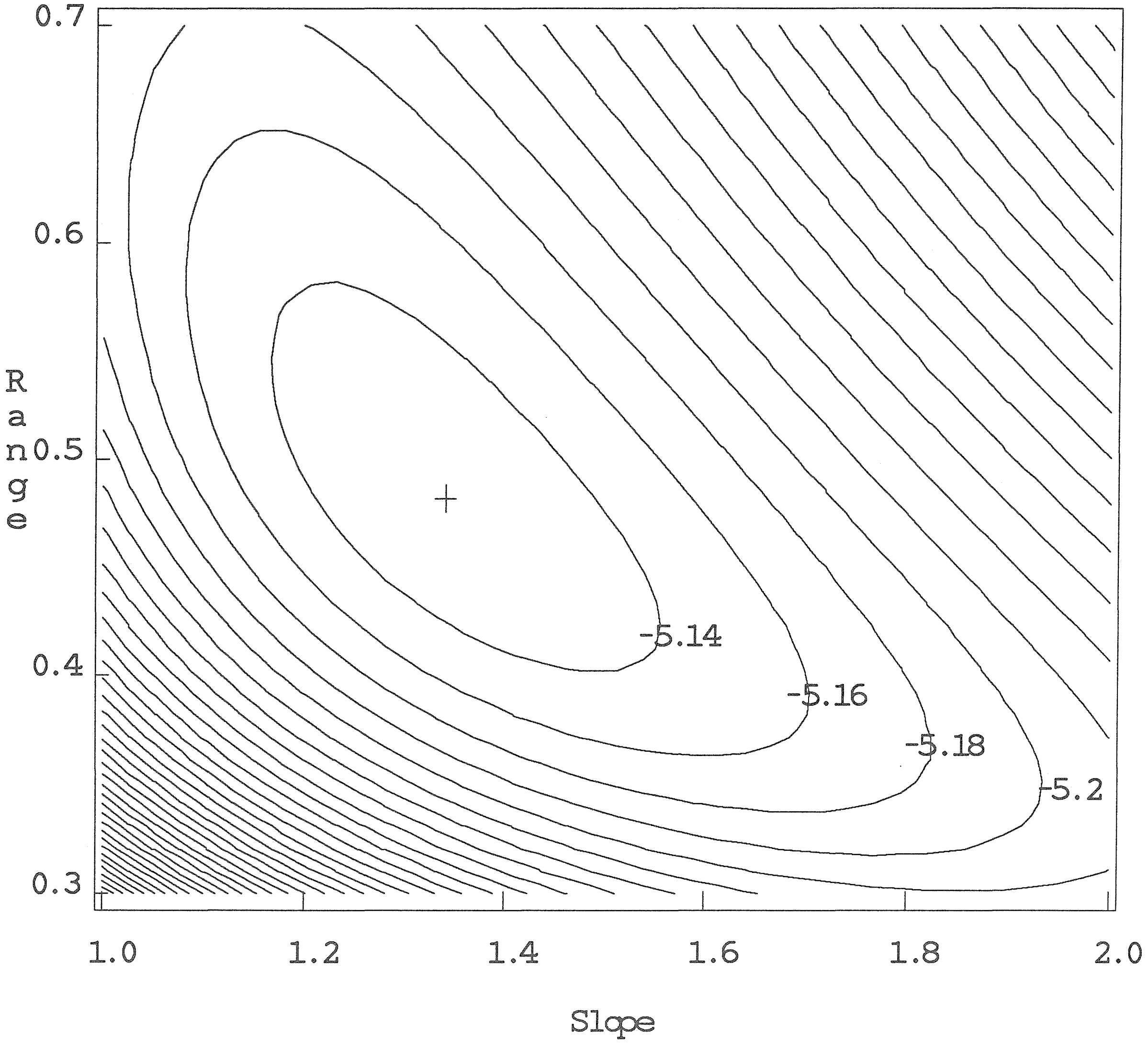


Fig. 7. Typical log-likelihood contour with the MLE identified in the interior. The process was observed at 0, 0.25, 0.5, 0.75, and 1. The true slope is 5 and the true range is 0.2.

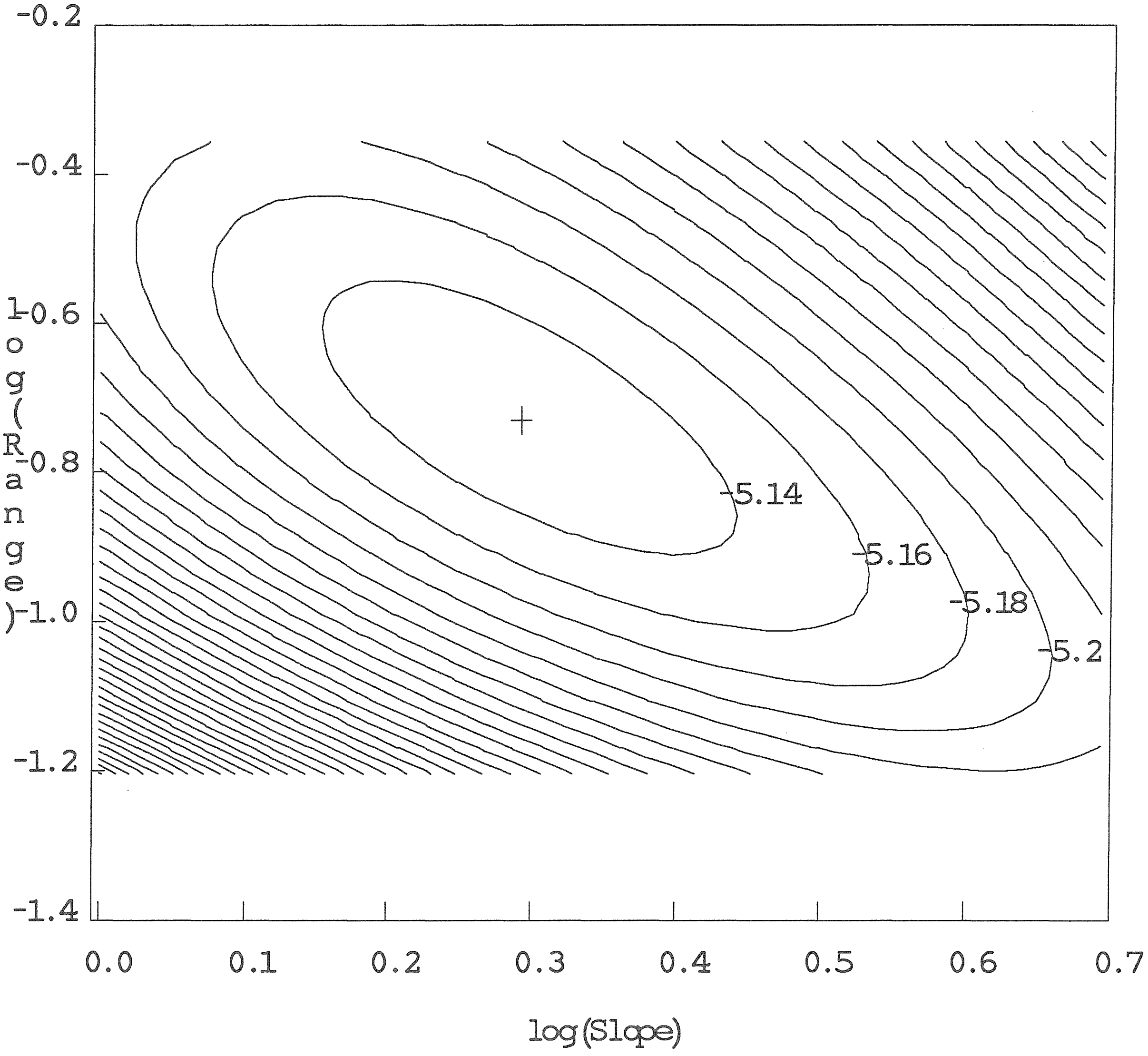


Fig. 8. Typical contour under the log parameterization with the MLE identified in the interior. The process was observed at 0, 0.25, 0.5, 0.75, and 1. The true slope is 5 and the true range is 0.2.

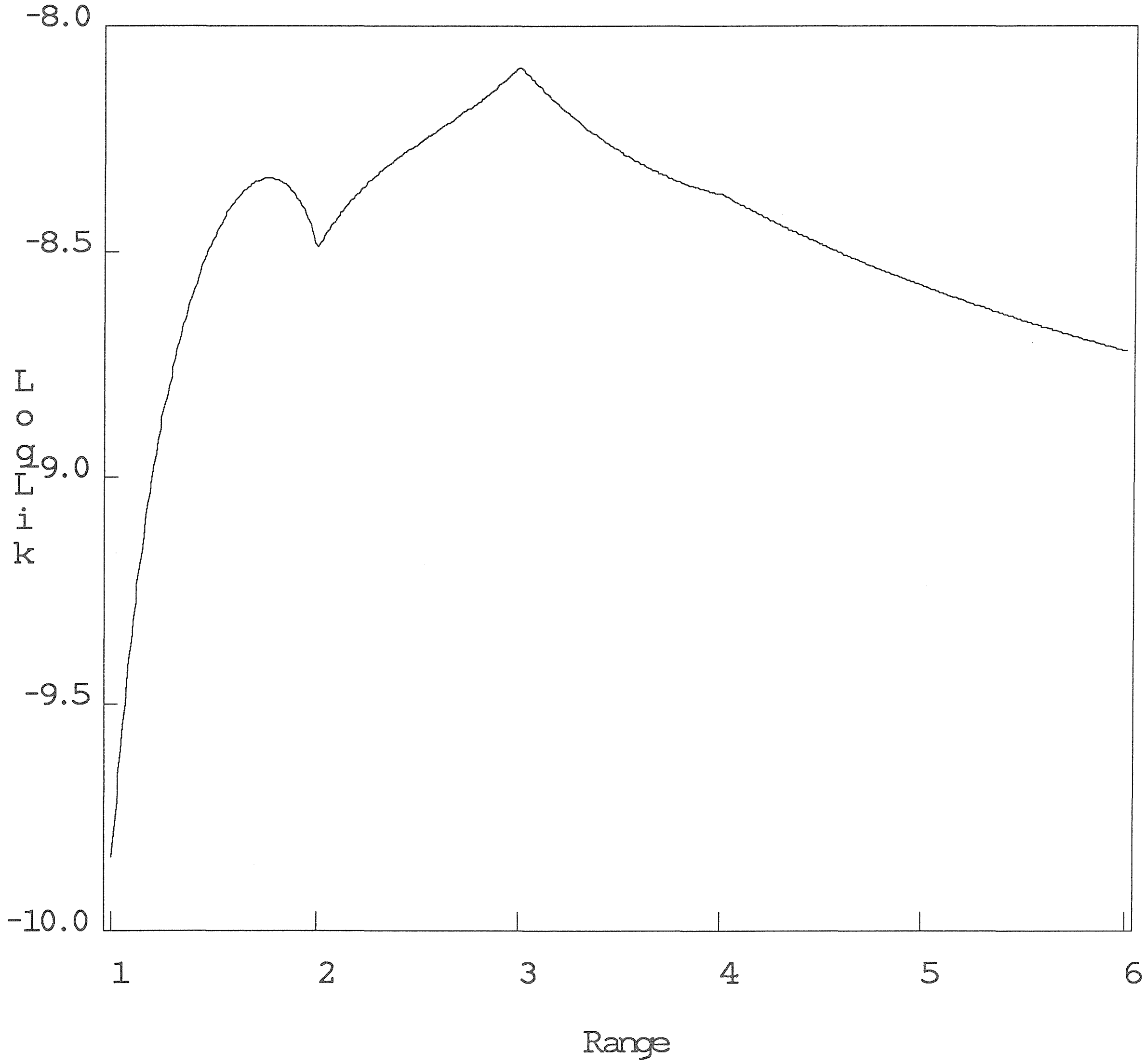


Fig. 9. Typical profile log-likelihood for the Triangular process observed at 1, 2, 3, 4, and 5. The true slope is 1 and the true range is 3.

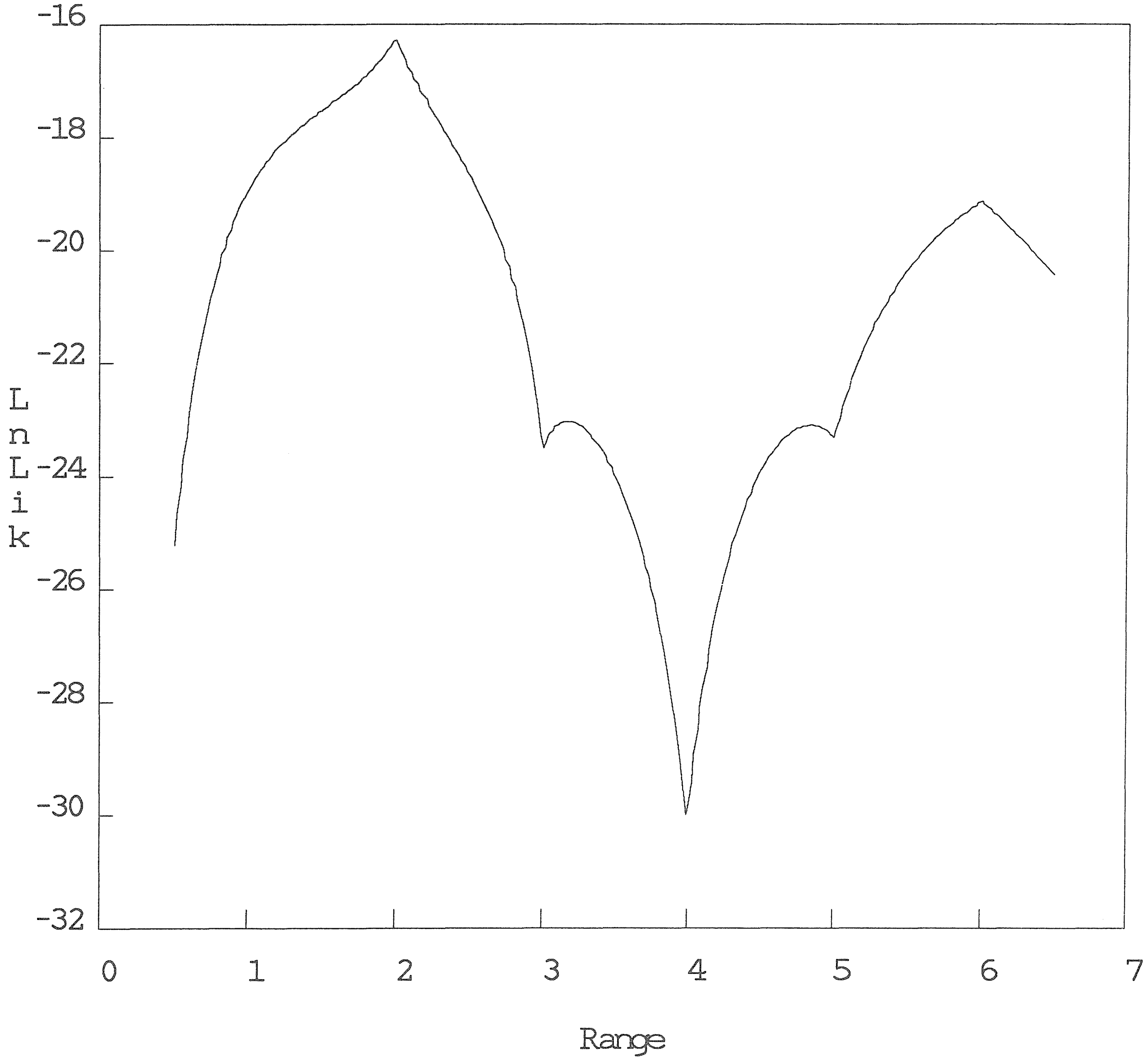


Fig. 10. Typical profile log-likelihood for the Triangular process observed at 1, 2, ..., 10. The true slope is 1 and the true range is 3.

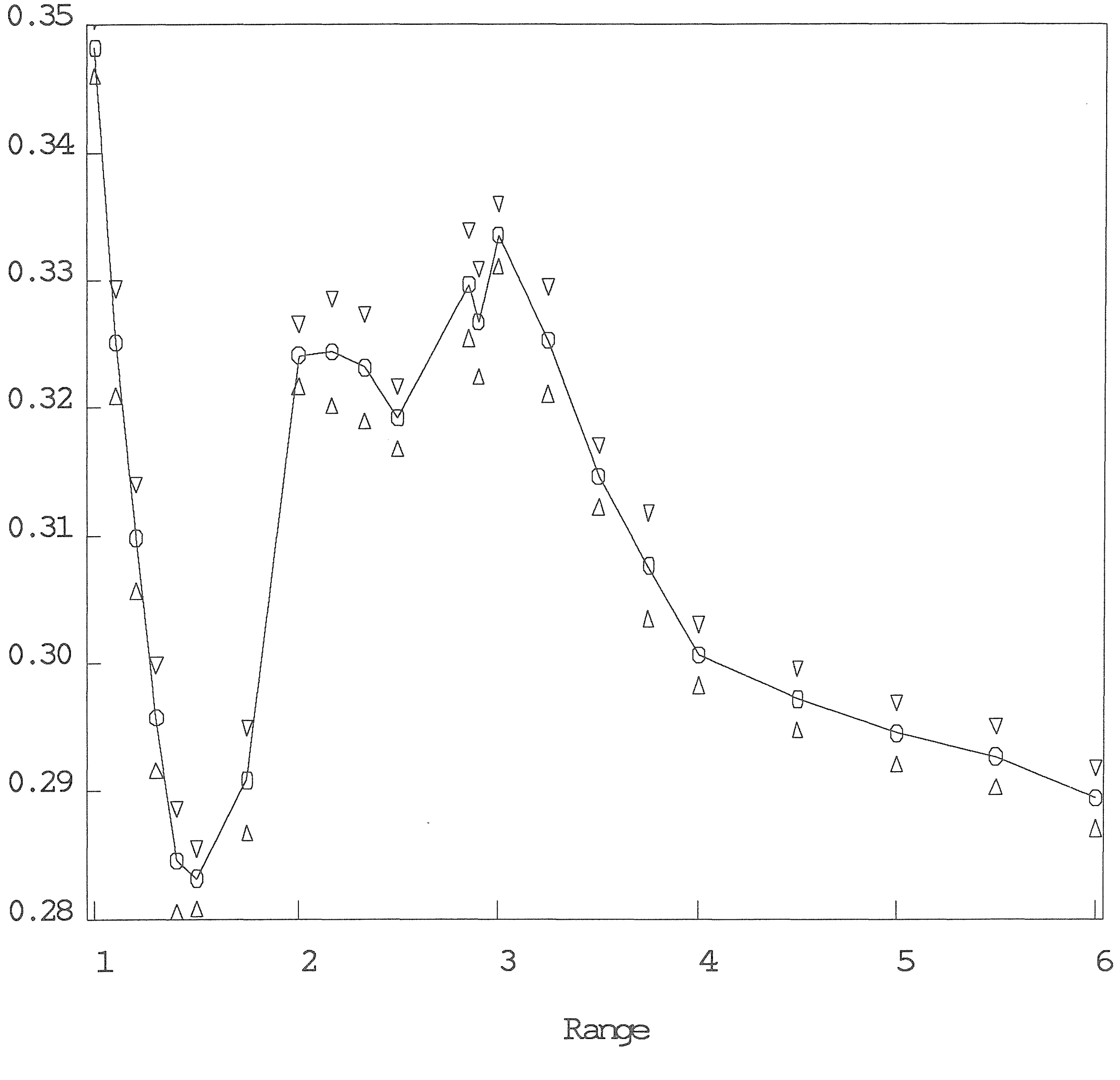


Fig. 11. Probability of a unimodal likelihood for a Triangular process observed at 1, 2, 3, 4, and 5. The true slope is 1.

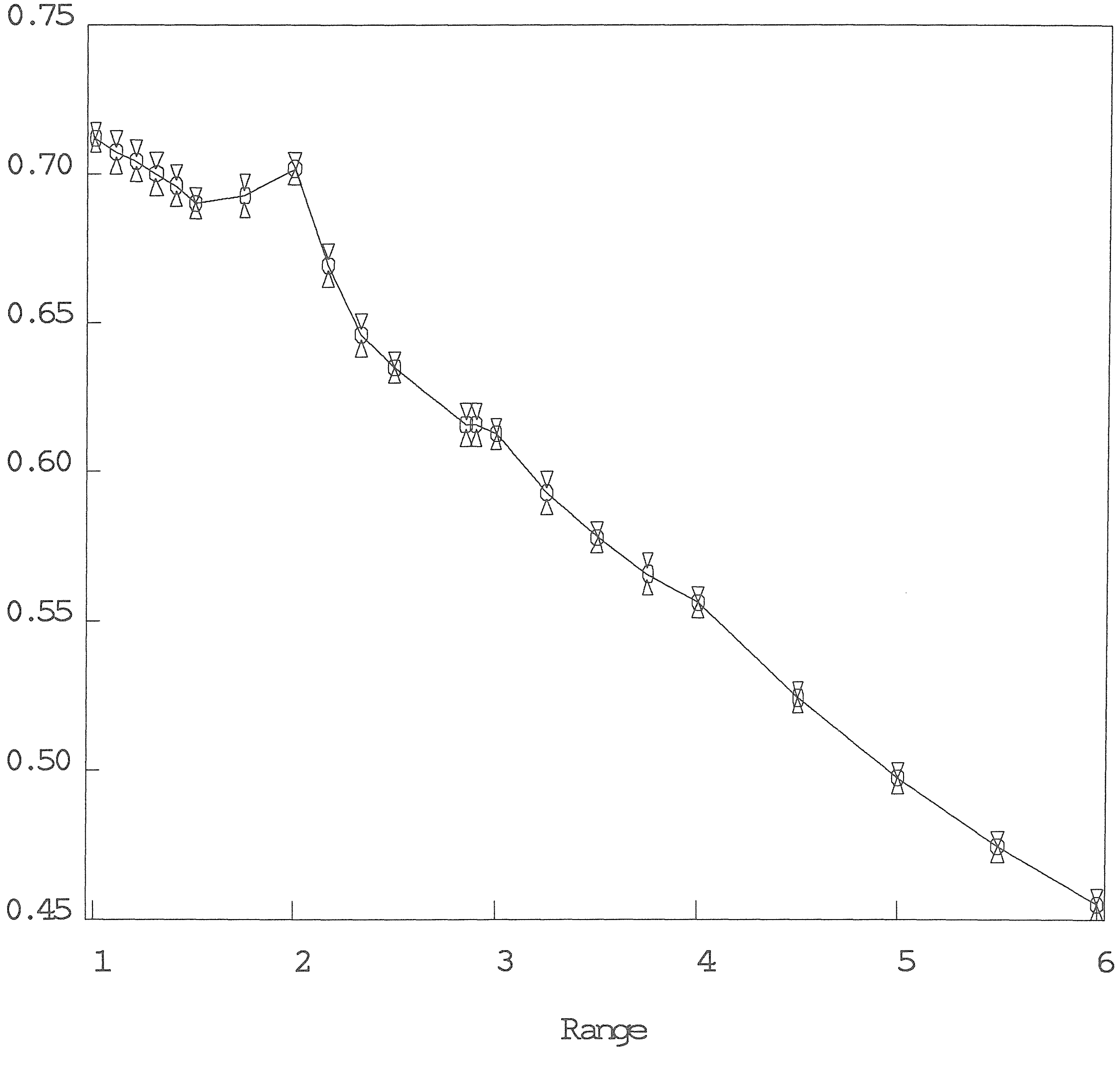


Fig. 12. Probability of a modes at integer values for the likelihood of a Triangular process observed at 1, 2, 3, 4, and 5. The true slope is 1.

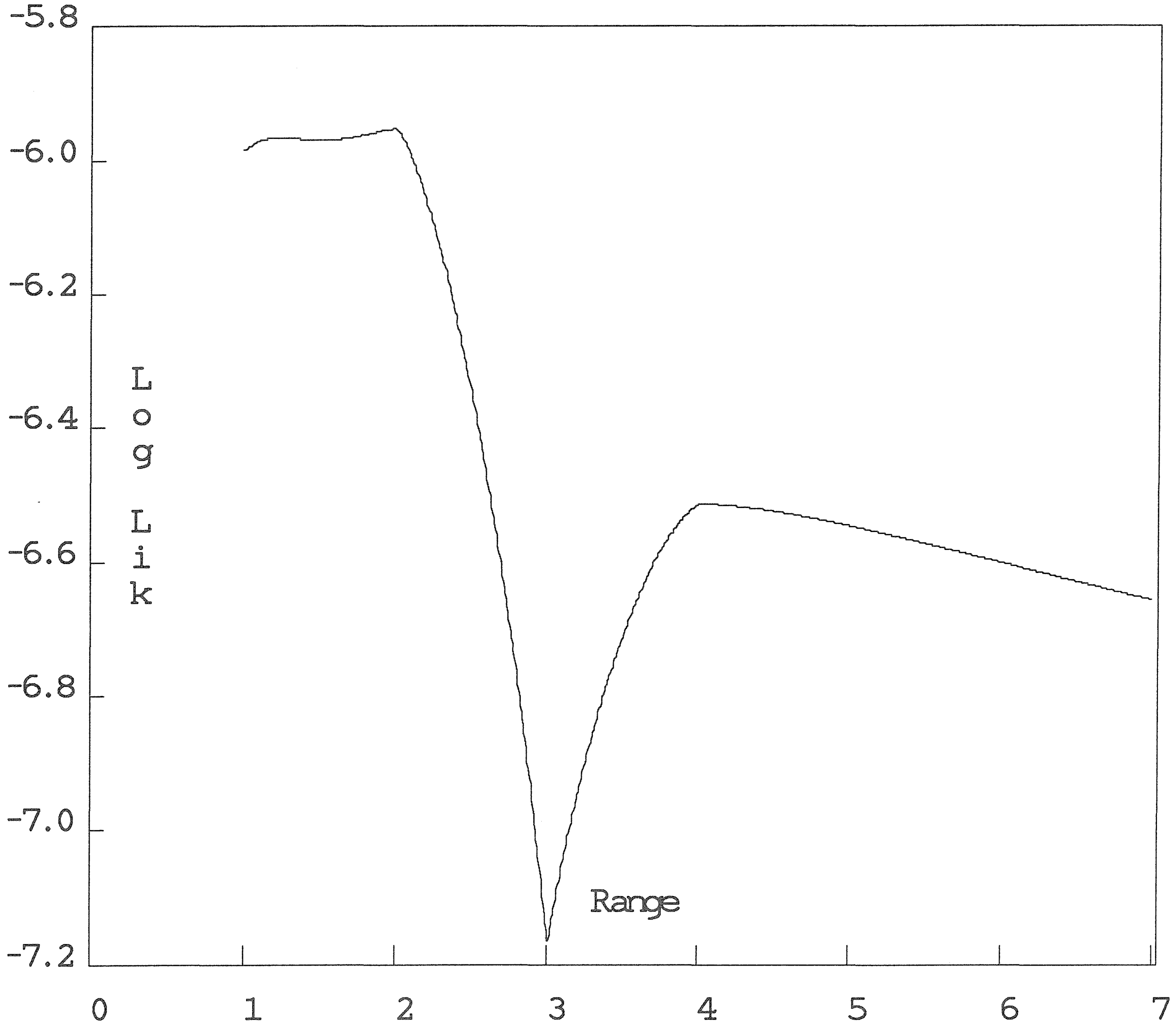


Fig. 13. Typical profile log-likelihood for the Triangular process with modes between the integers. The process was observed at 1, 2, 3, 4, and 5. The true slope is 1 and the true range is 3.

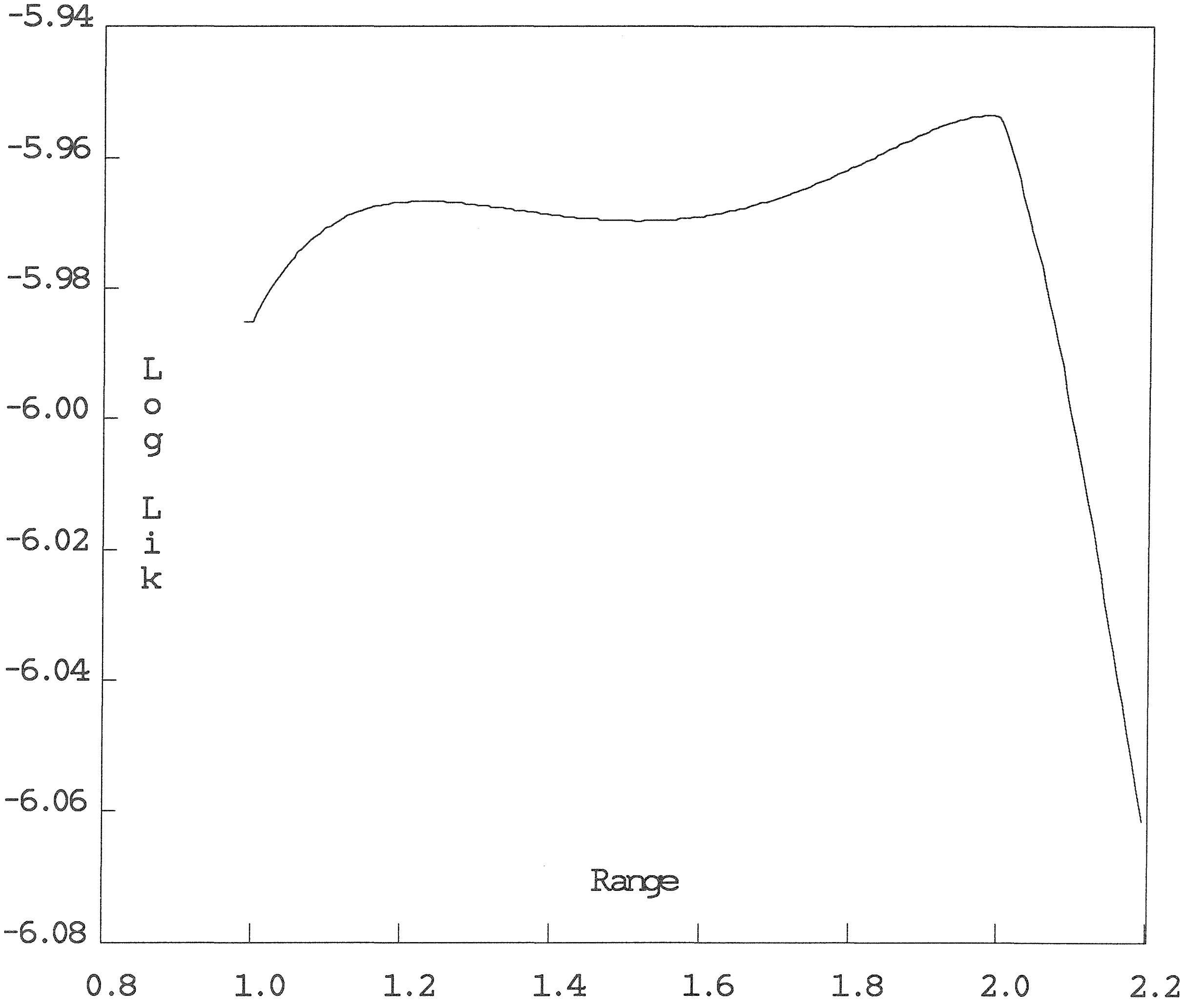


Fig. 14. Close up section of typical profile log-likelihood for the Triangular process with modes between the integers. The process was observed at 1, 2, 3, 4, and 5. The true slope is 1 and the true range is 3.

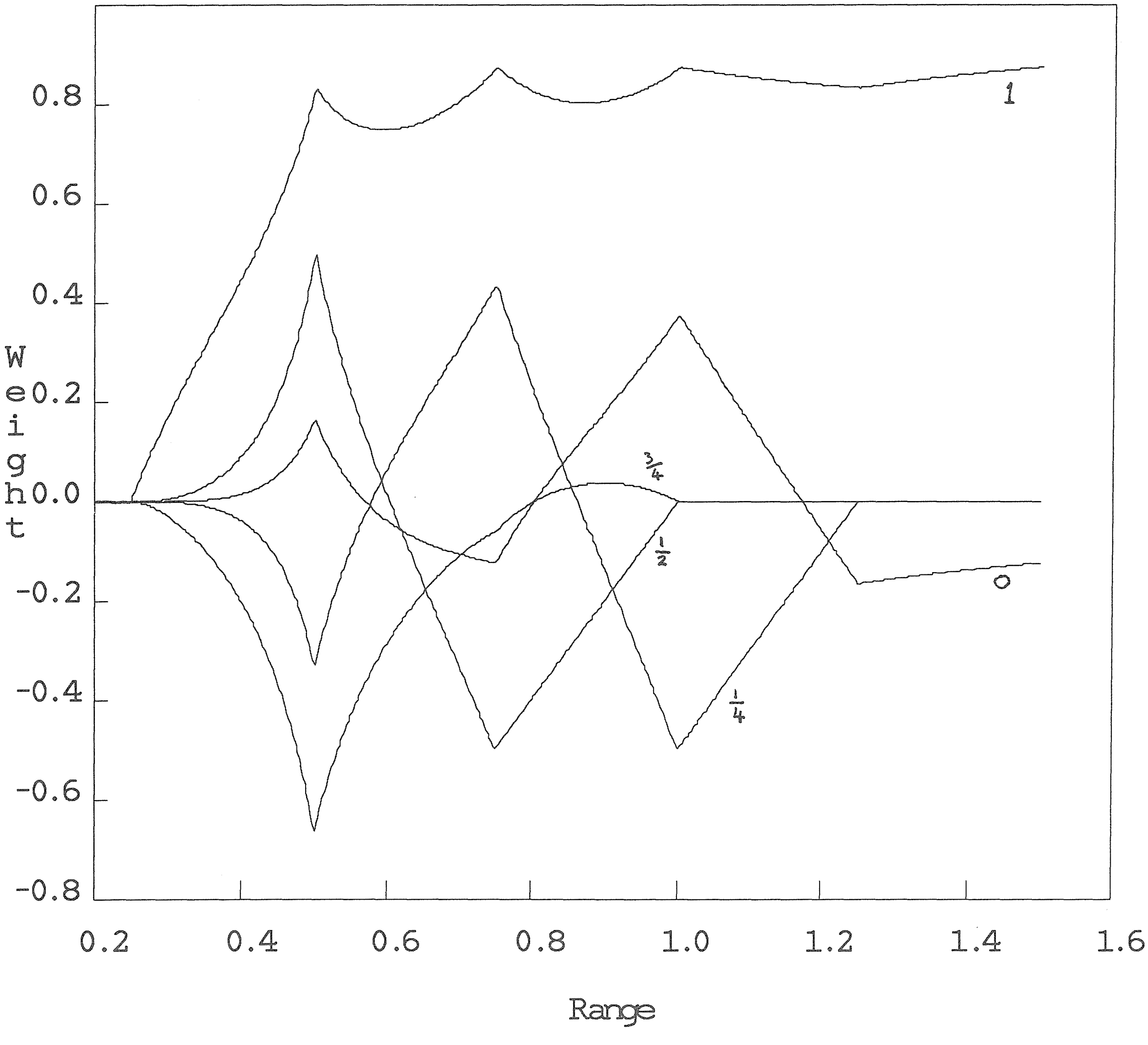


Fig. 15. Weights on each location for a Triangular process observed at 0, 0.25, 0.5, 0.75, and 1 as the range changes. The true range is 1 and the weights do not depend on the slope.

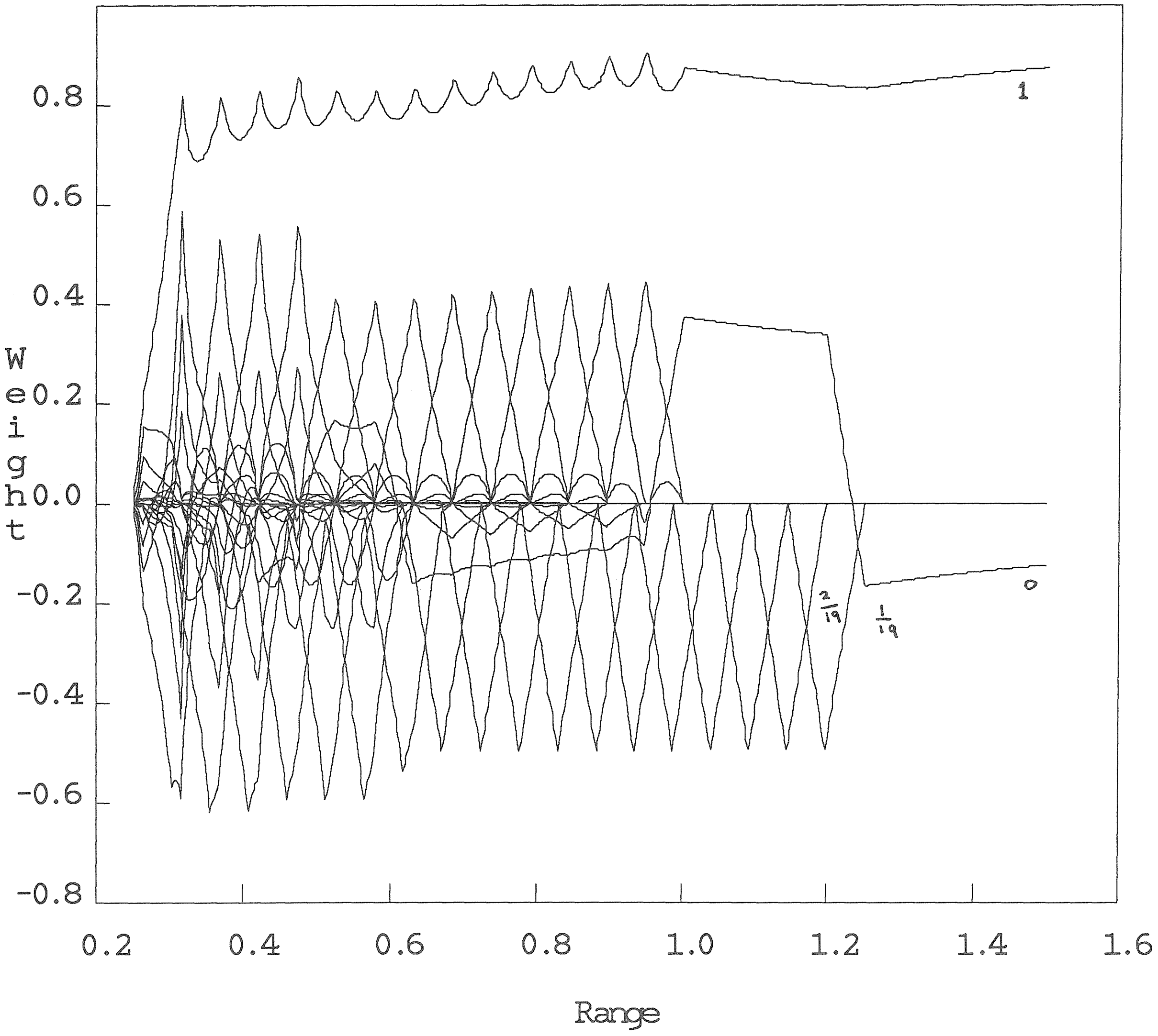


Fig. 16. Weights on each location for a Triangular process observed at 0, 1/19, ..., 18/19, 1 as the range changes. The true range is 1 and the weights do not depend on the slope.

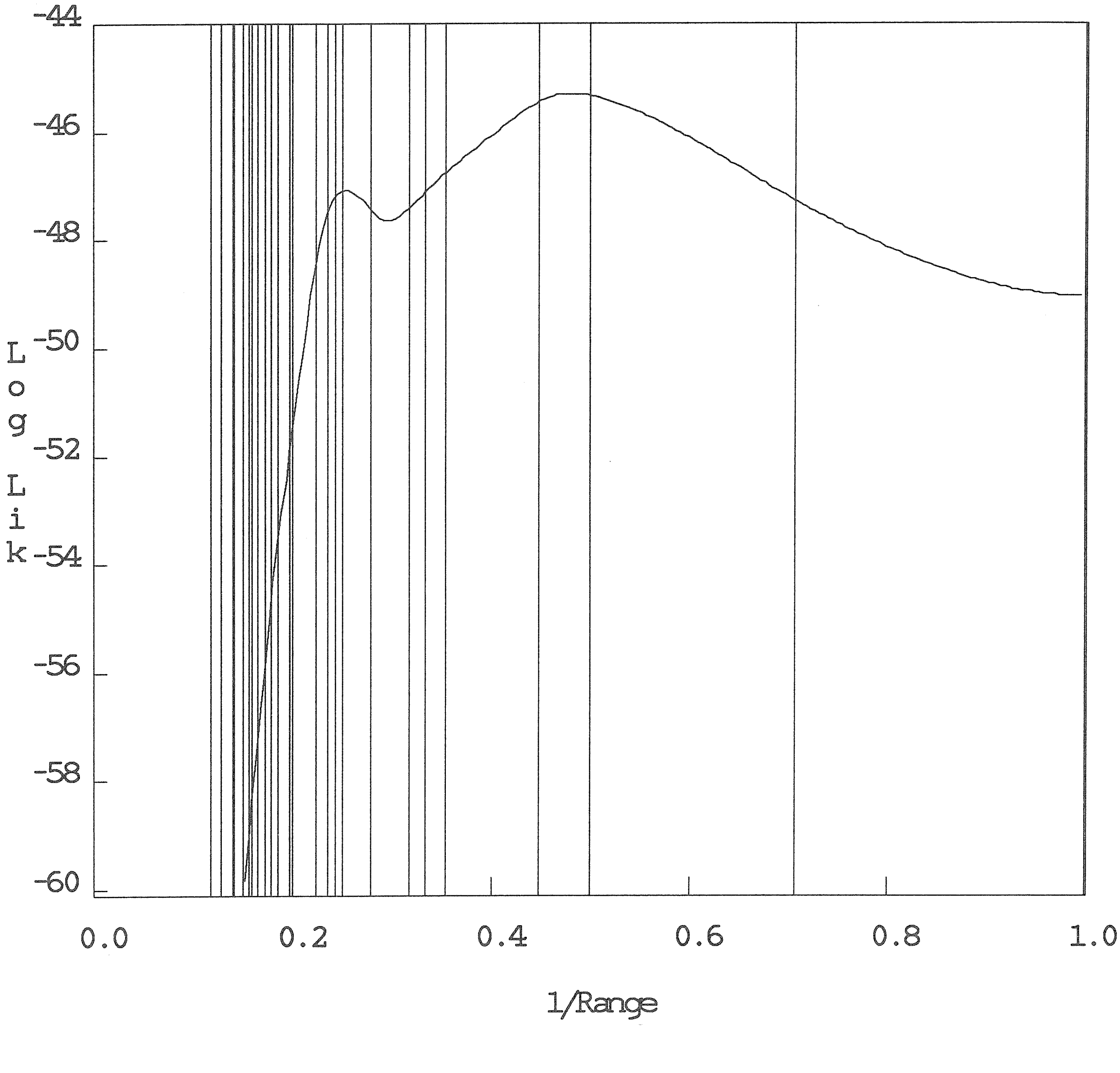


Fig. 17. Typical profile log-likelihood for the Spherical random field on a 6 x 6 grid. The vertical lines represent the distances separating the observations. The true slope is 0.5 and the true range is 3.

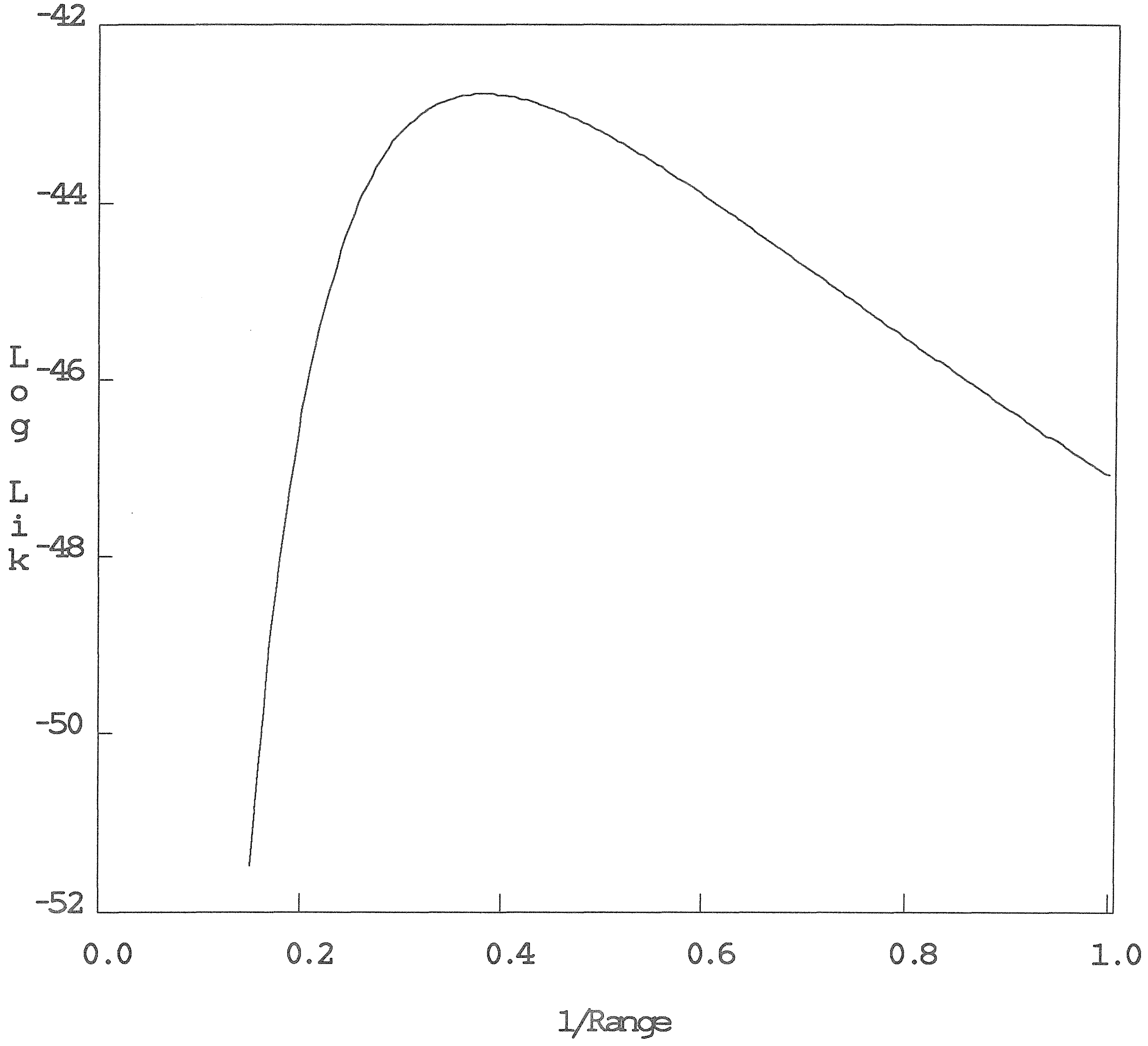


Fig. 18. The profile log-likelihood under the Exponential model for the same data.

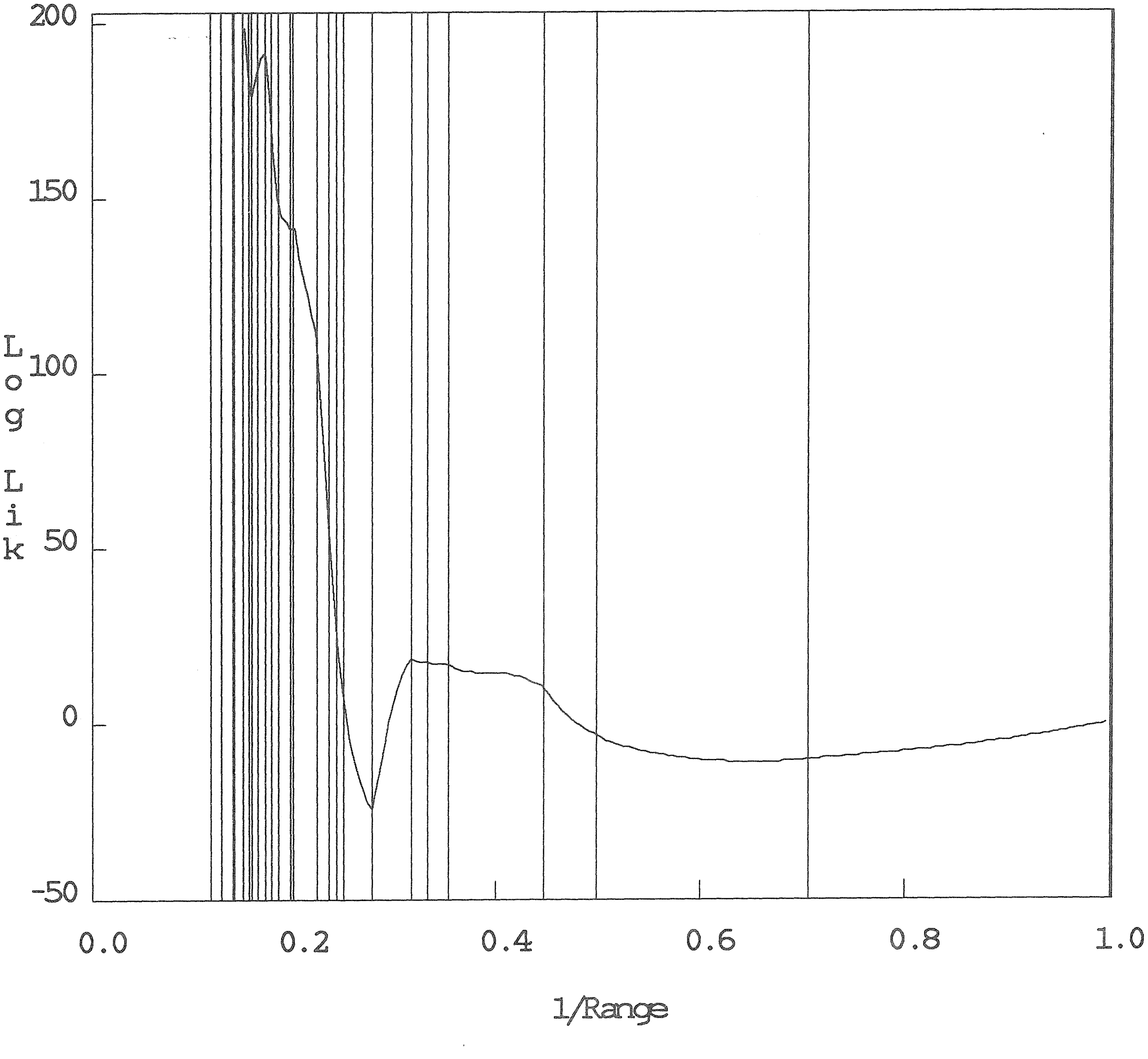


Fig. 19. Derivatives of the log-likelihood for the Spherical random field on a 6 x 6 grid. The vertical lines represent the distances separating the observations.

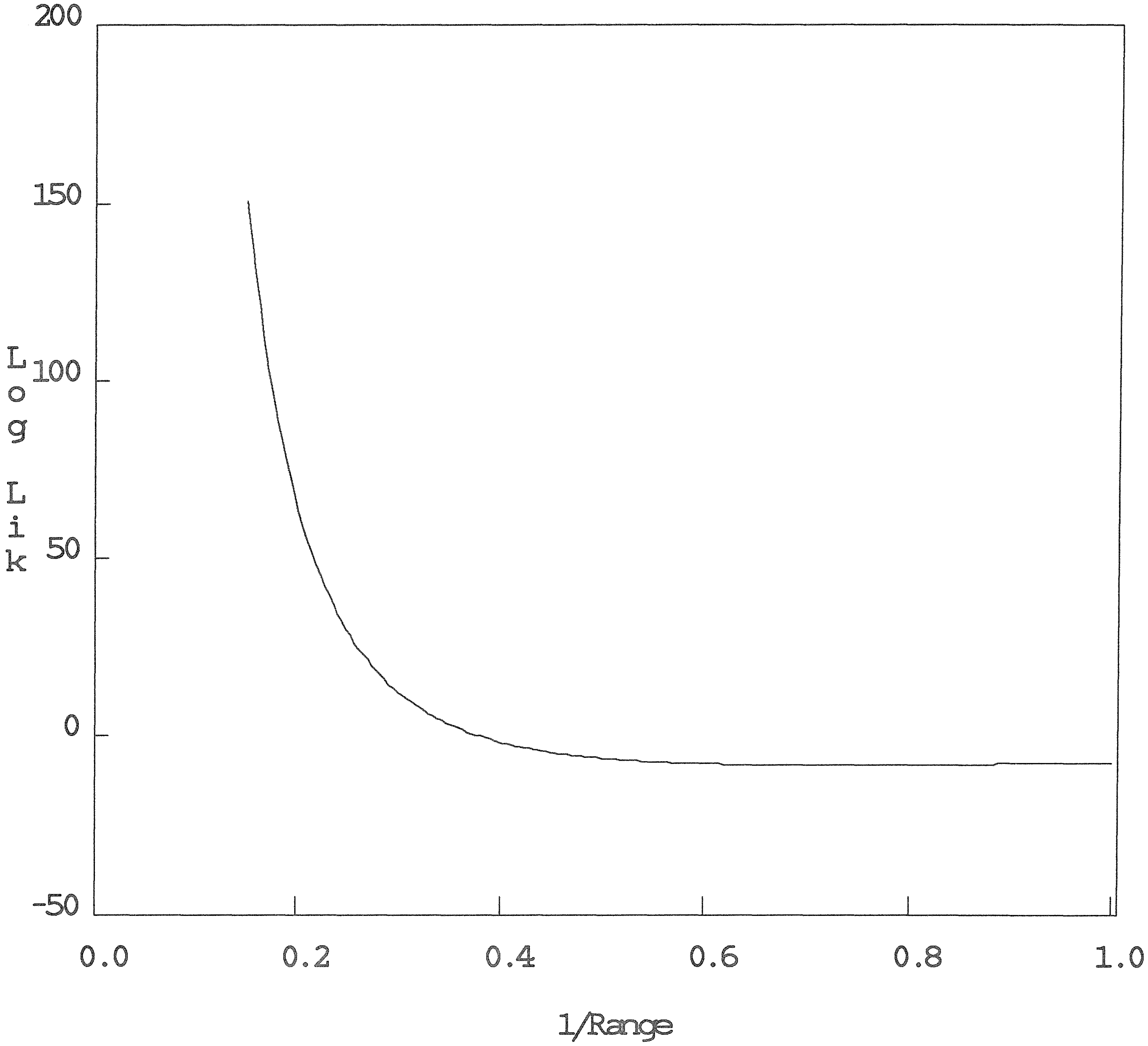


Fig. 20. The derivatives of the log-likelihood under the Exponential model for the same data.

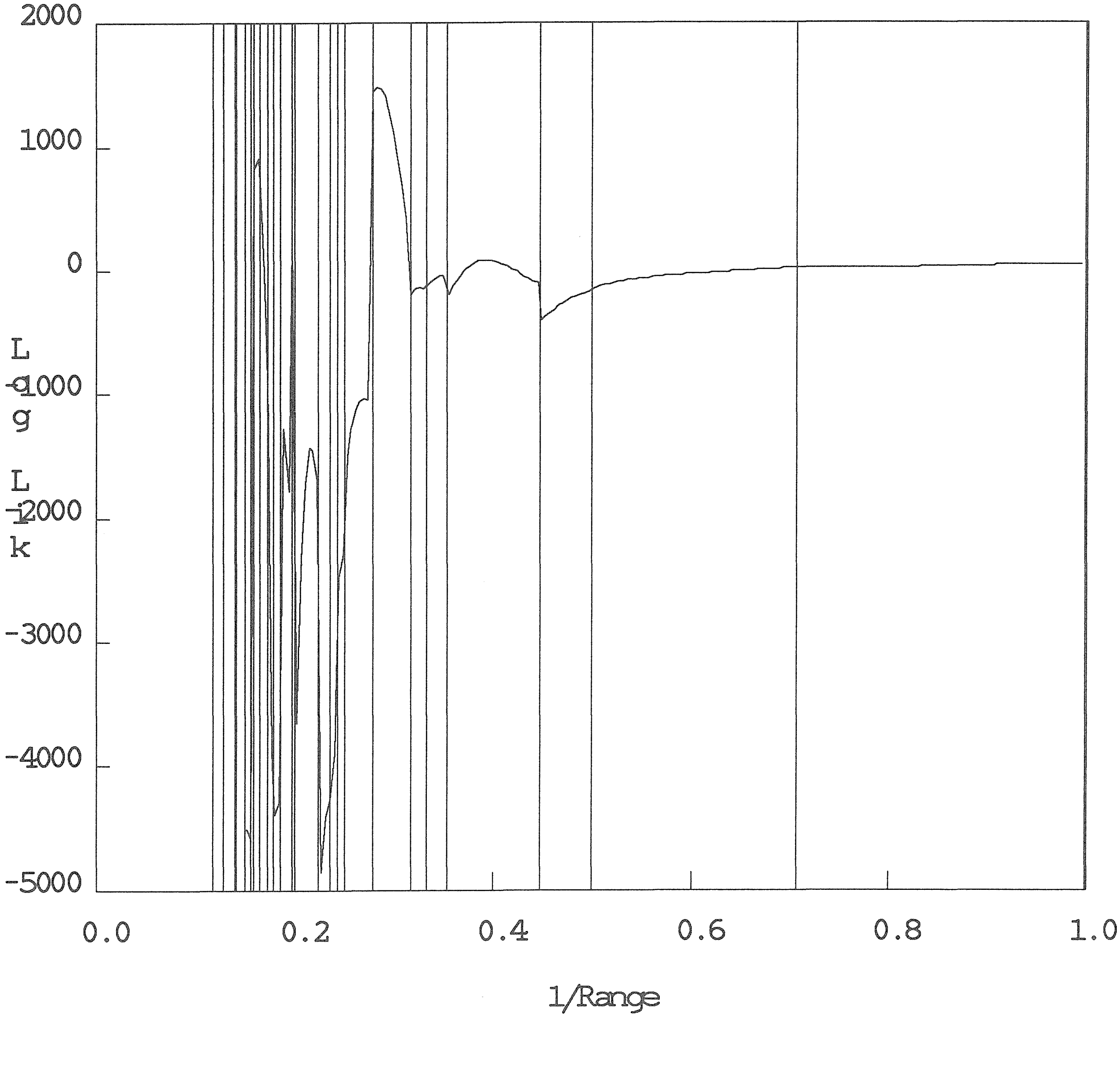


Fig. 21. Second Derivatives of the log-likelihood for the Spherical random field on a 6 x 6 grid. The vertical lines represent the distances separating the observations. The true slope is 0.5 and the true range is 3.

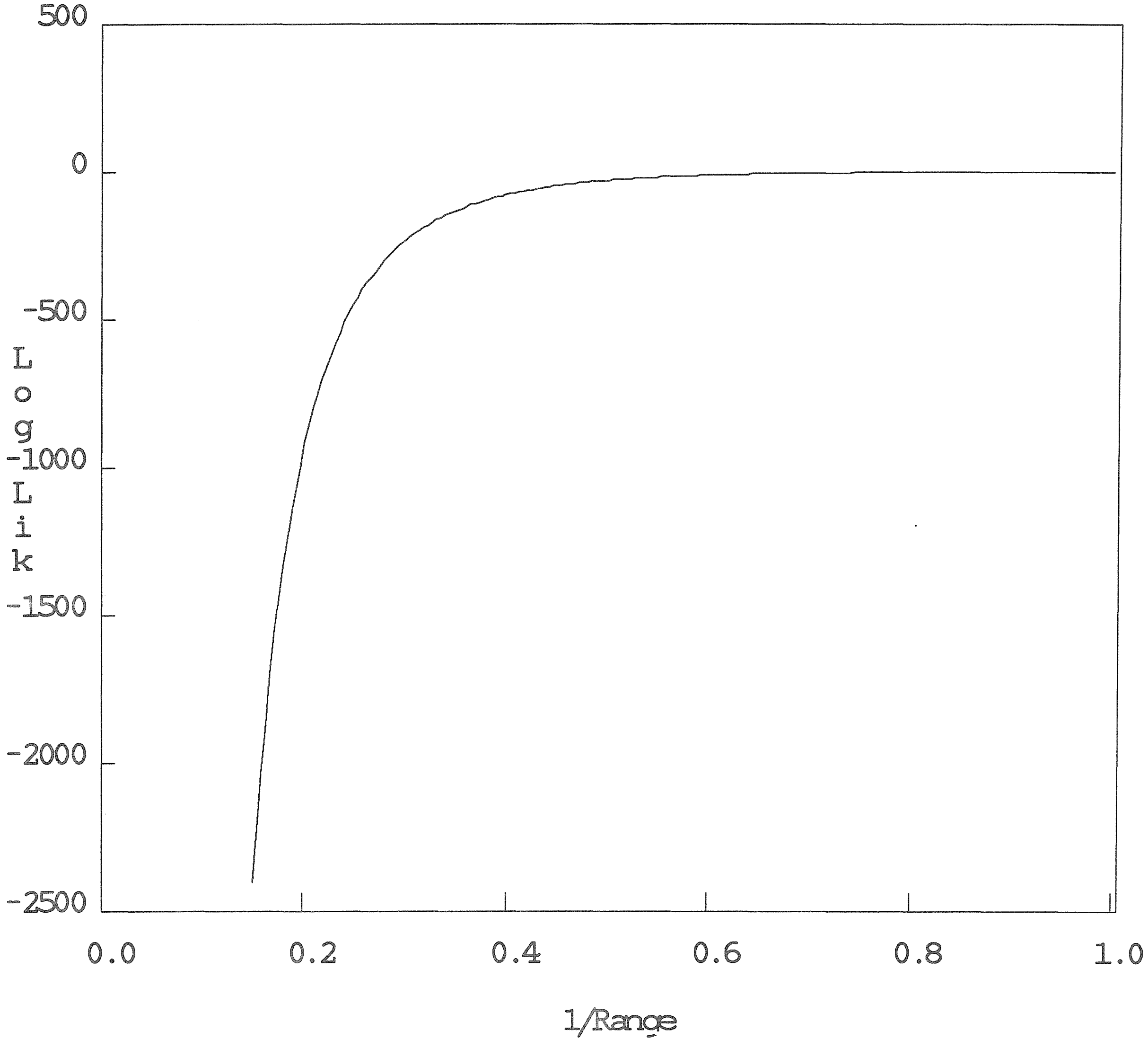


Fig. 22. The second derivatives of the log-likelihood under the Exponential model for the same data.

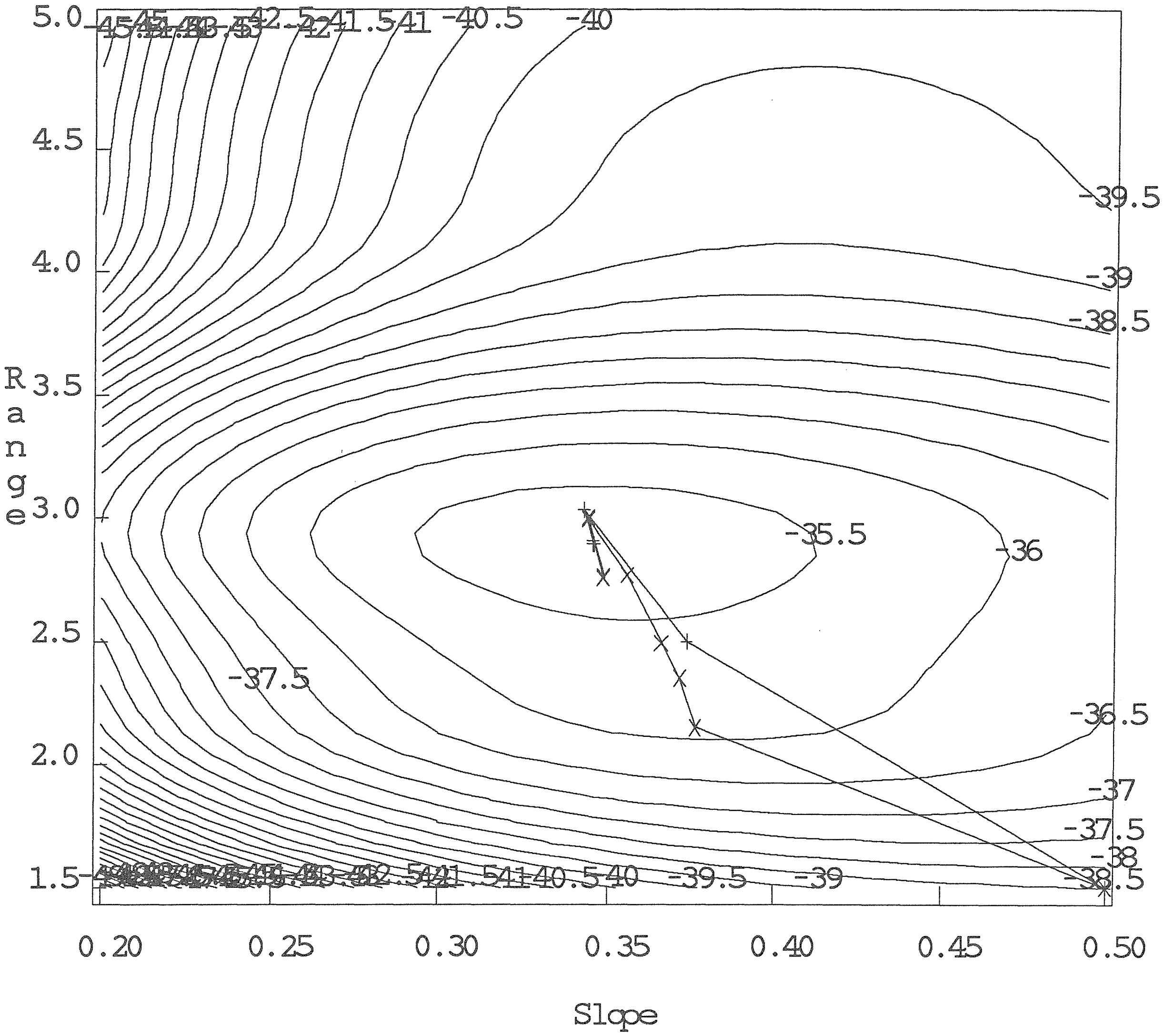


Fig. 23. Typical contour log-likelihood for a Spherical random field on a 6 x 6 grid. The x indicate the path of a Newton-Raphson method. The + indicate the path followed from a cubic search method from the same starting point. The true slope is 0.5 and the true range is 3.

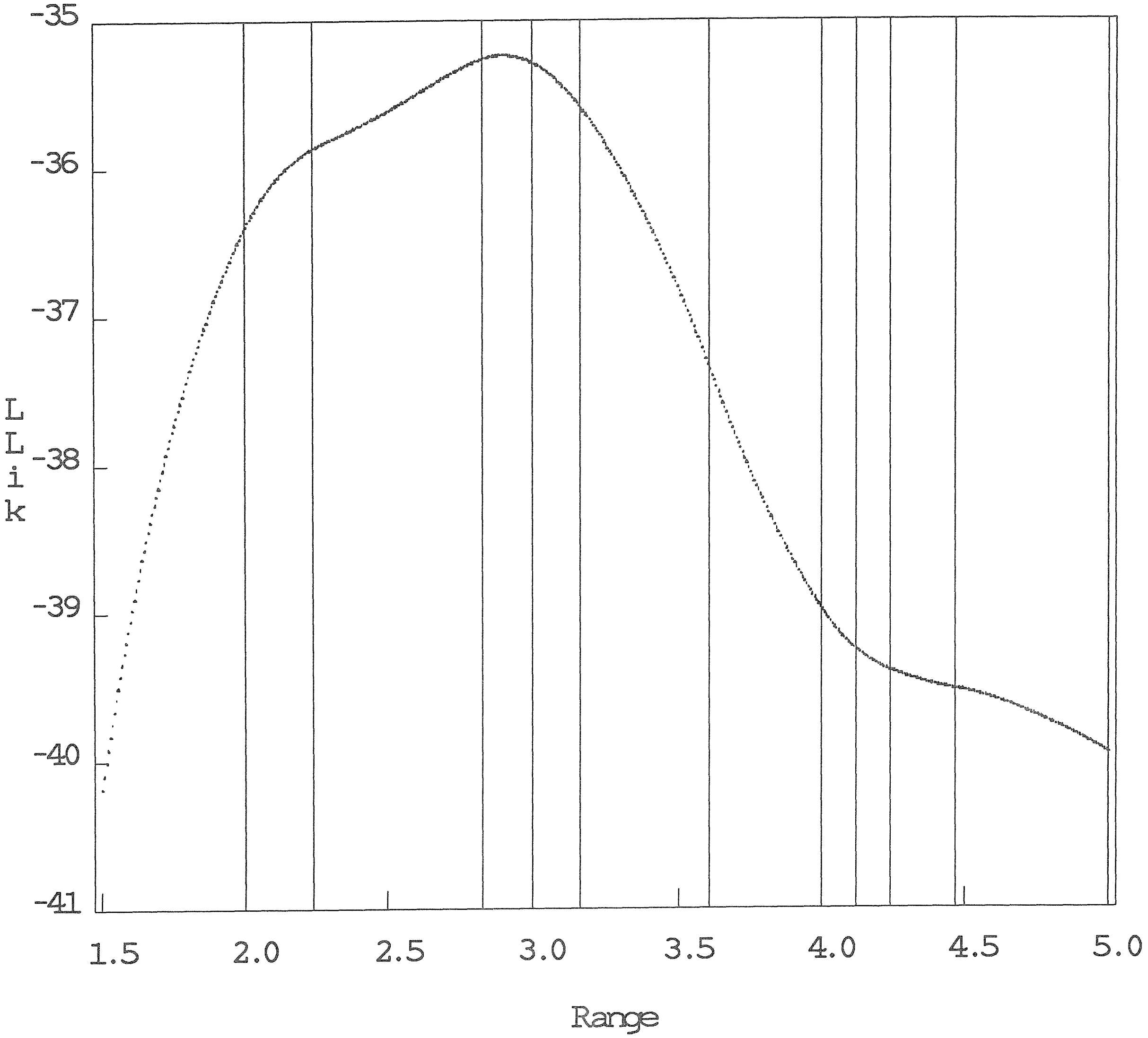


Fig. 24. The profile log-likelihood when the slope is set to its MLE for the same data.

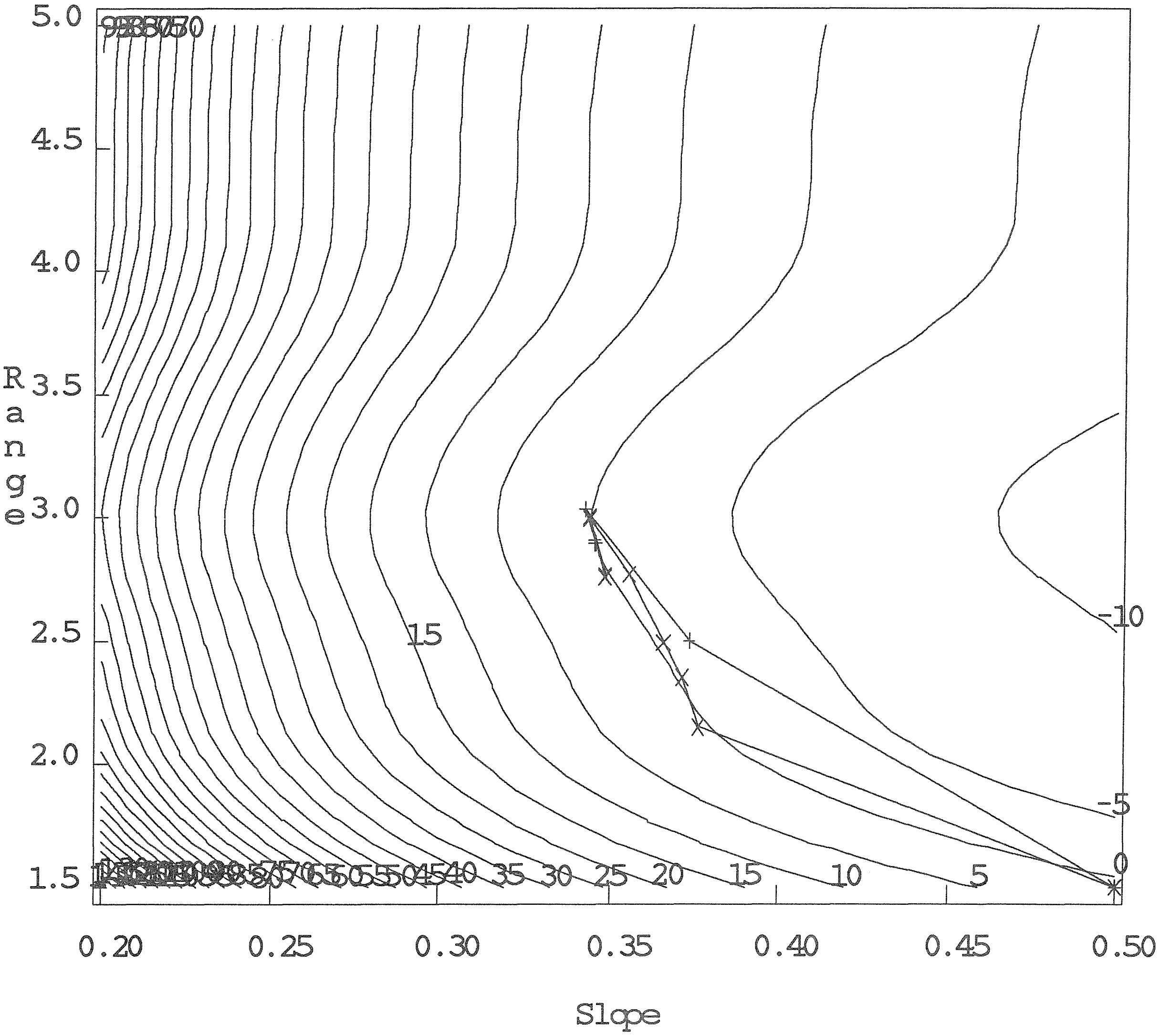


Fig. 25. Contour of the derivative of the log-likelihood with respect to slope for the same realization.

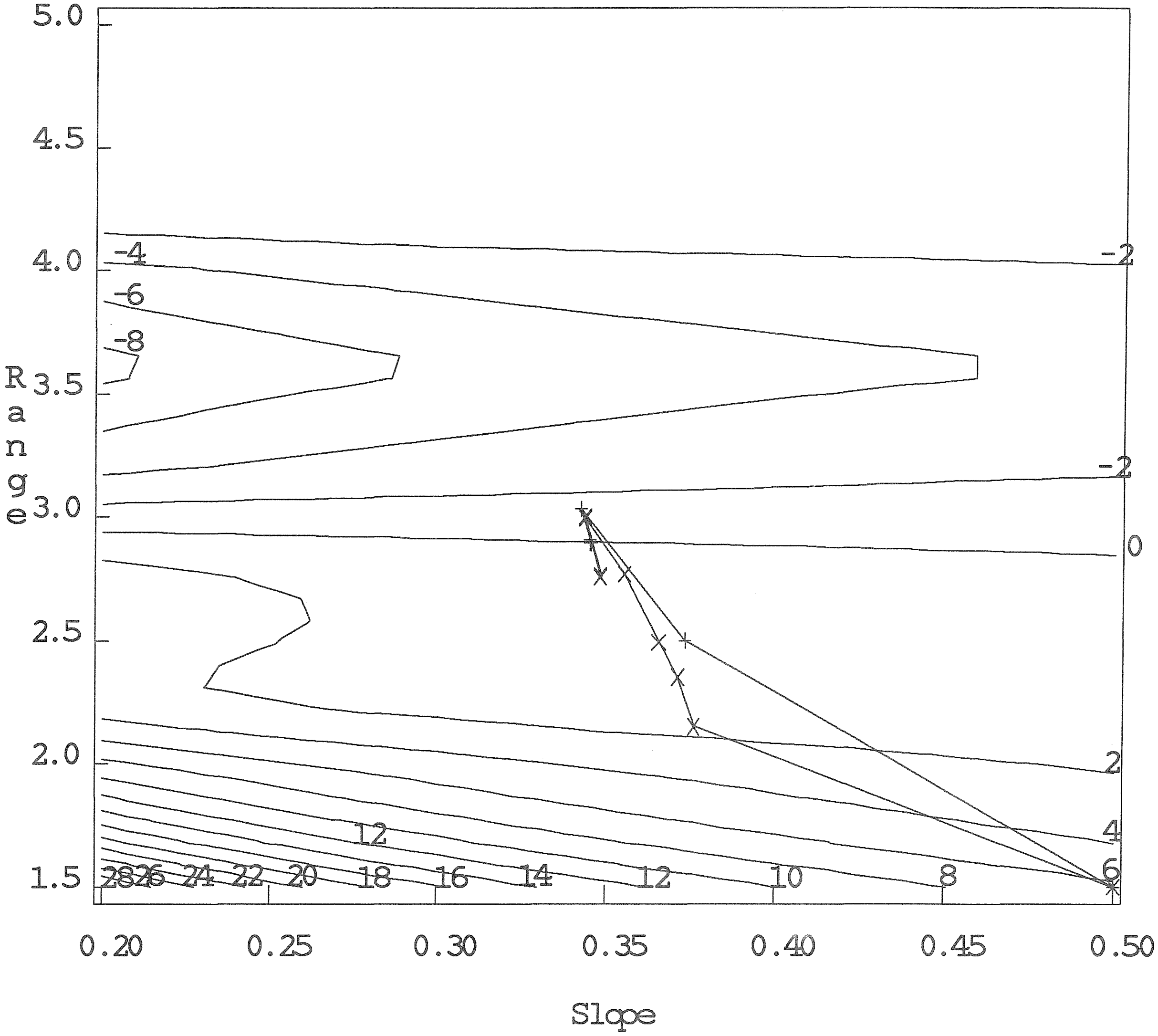


Fig. 26. Contour of the derivative of the log-likelihood with respect to range for the same realization.

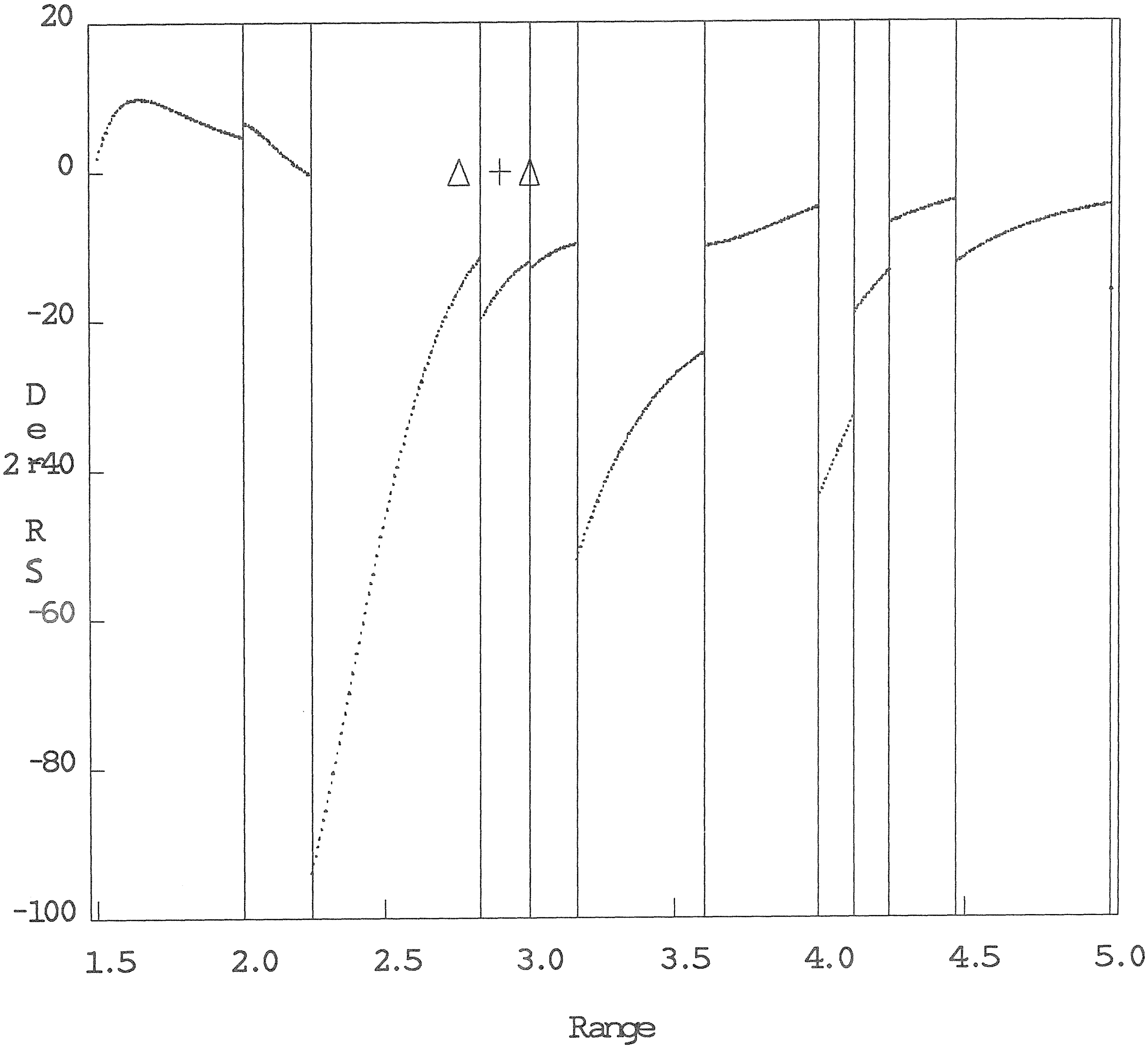


Fig. 27. The profile of the second derivative of the log-likelihood with respect to the range, when the slope is set to its MLE.

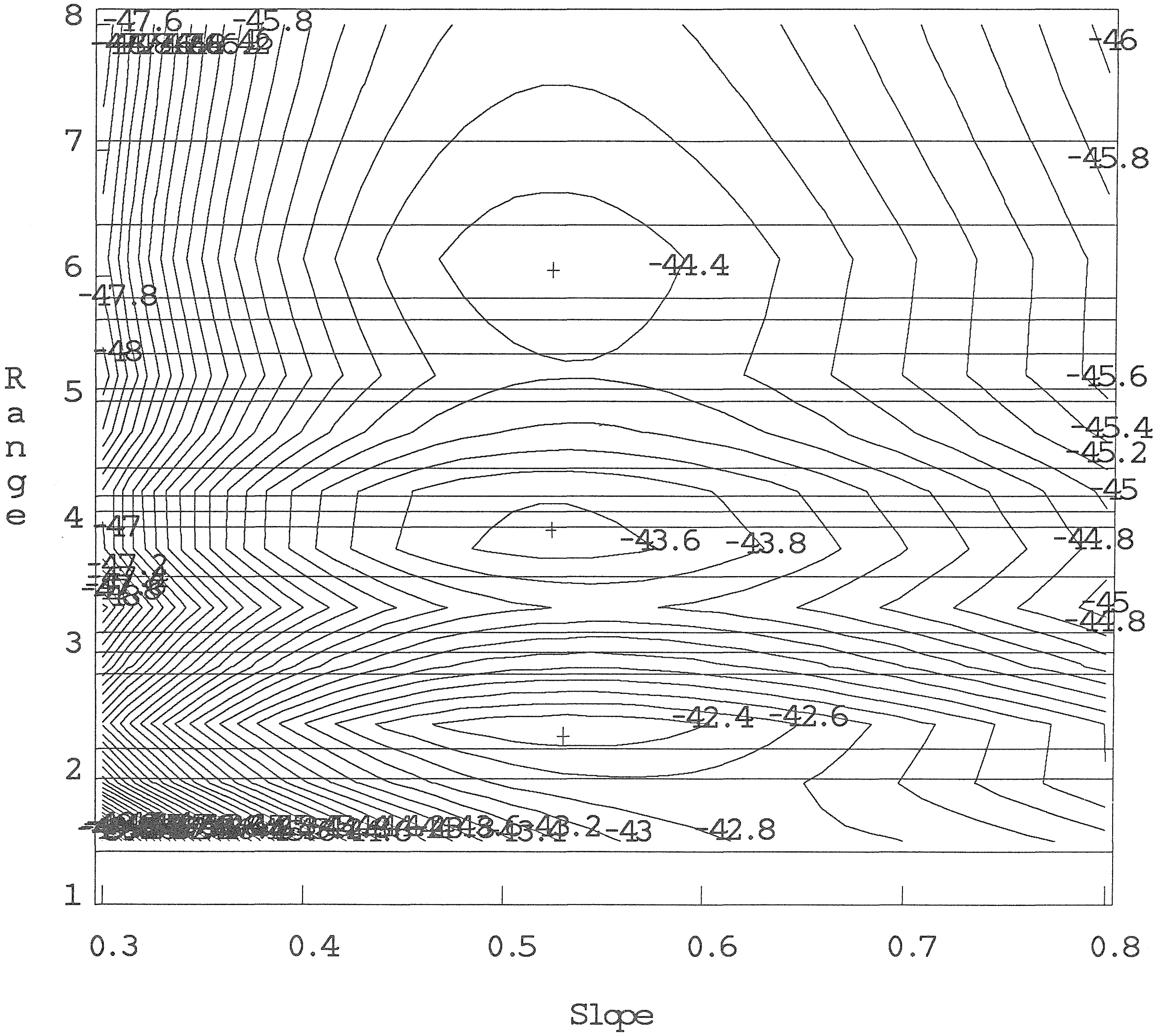


Fig. 28. Typical contour for a multimodal log-likelihood for a Spherical random field on a 6 x 6 grid. The horizontal lines represent the interpoint distances. The true slope is 0.5 and the true range is 3.

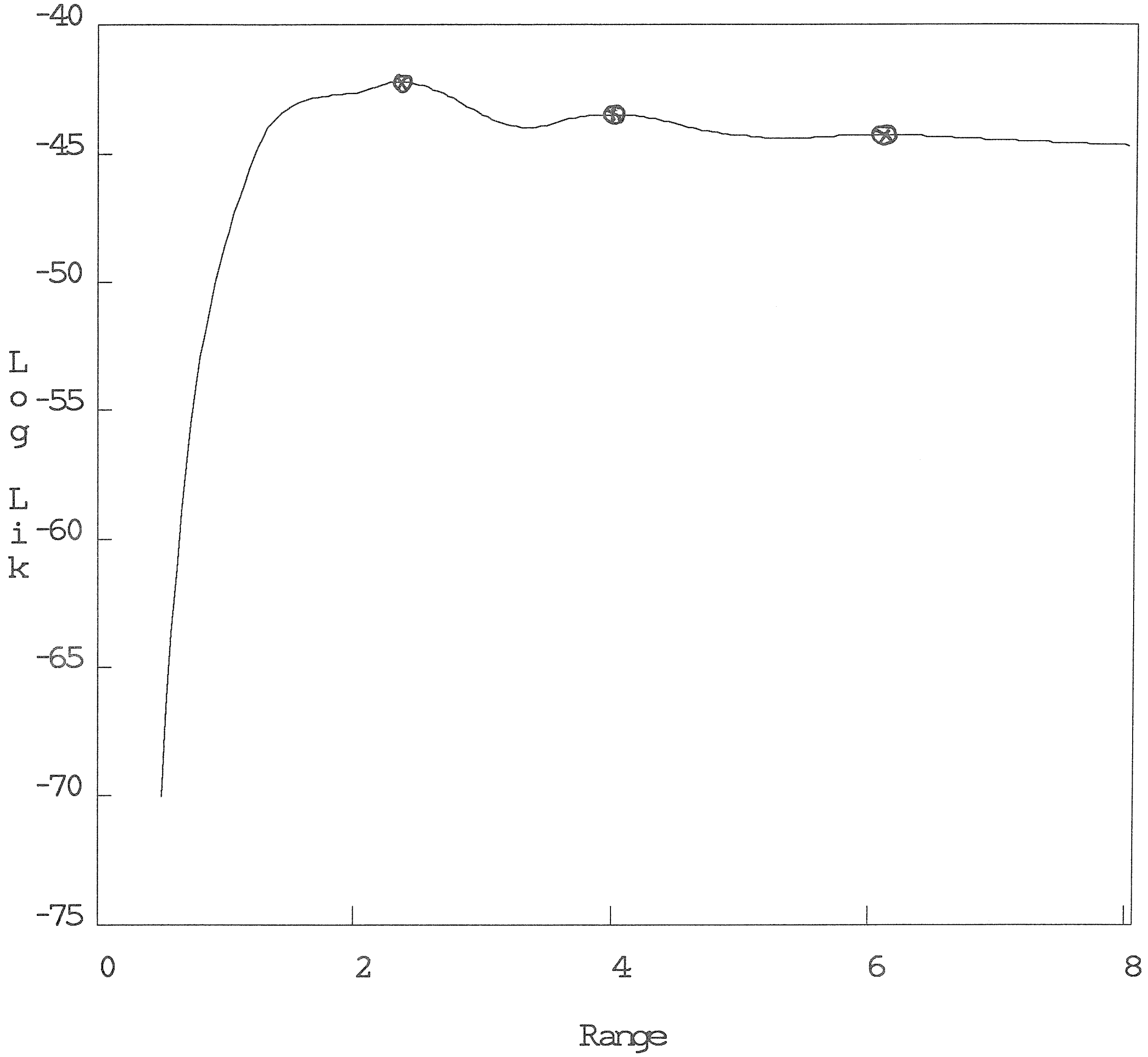


Fig. 29. The profile log-likelihood when the slope is set to its MLE for the same data. The three modes are indicated.

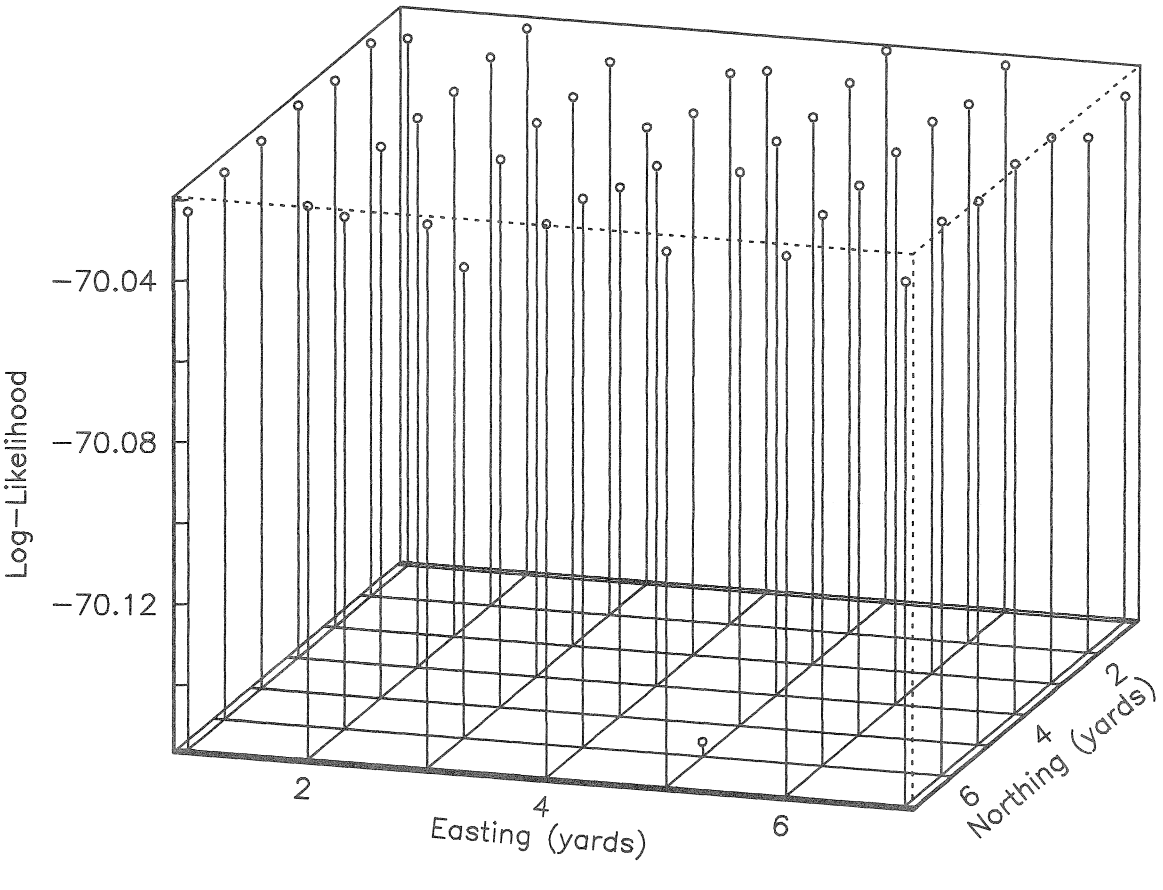


Fig. 30. Full log-likelihoods at the cross-validated MLEs for the Exponential data, based on the Exponential model.

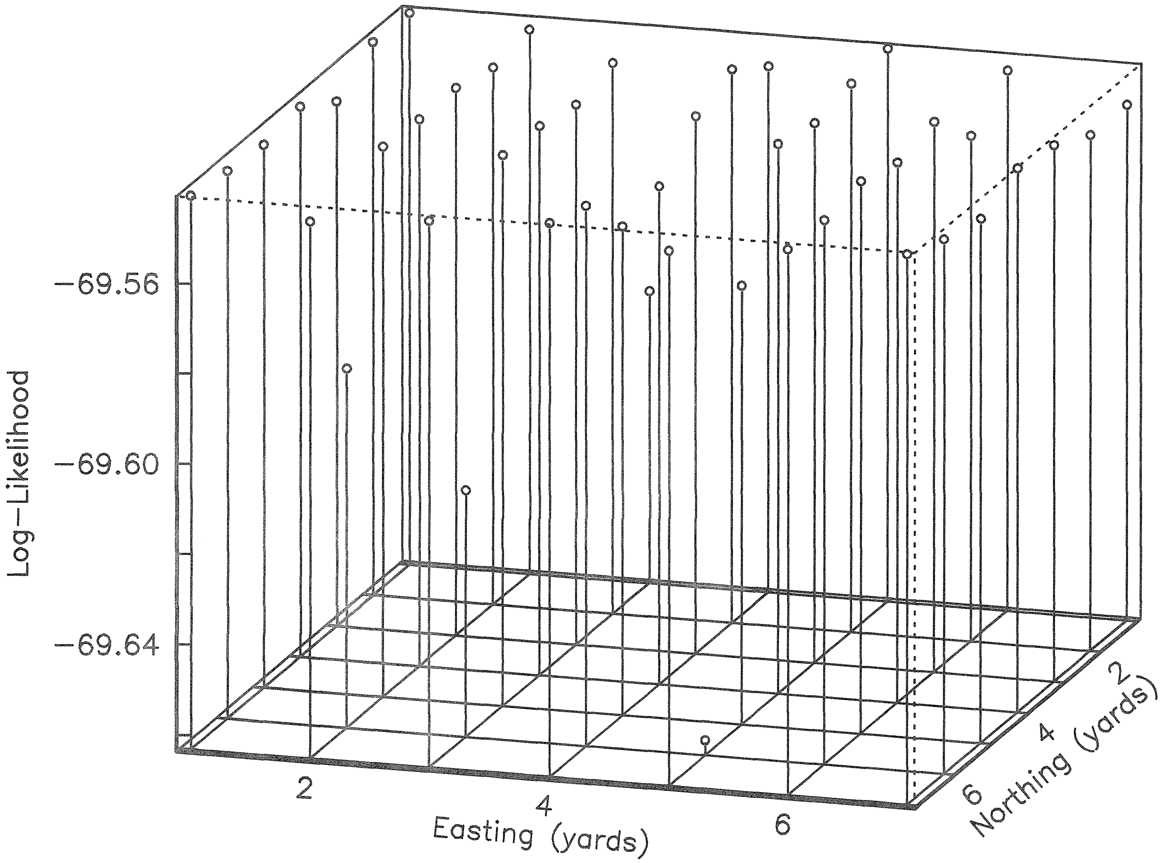


Fig. 31. Full log-likelihoods at the cross-validated MLEs for the Exponential data, based on the Matérn model with smoothness 3/2.

CHAPTER 3

ALTERNATIVE APPROACHES TO INFERENCE BASED ON OBSERVING A FINITE SEGMENT

3.1 Introduction

Suppose $Z(x)$ is a real–valued stationary Gaussian random field on R with mean

$$\mathbb{E}Z(x) = 0$$

and covariance function,

$$\mathrm{Cov}(Z(x), Z(x')) = R(|x - x'|) \qquad \text{for } x, x' \in R$$

In practice we observe $Z(x)$ at a finite number of points in $[0,\ T]$ $T > 0$, and wish to make inference about $R(x)$. It is usual to assume that $Z(x)$ is mean–square continuous, so that $R(x)$ is continuous and we have the representation,

$$R(x) = \int_{-\infty}^{\infty} e^{ix\lambda} dF(\lambda),$$

where $F(\lambda)$, the spectrum, is a nonnegative nondecreasing function with $F(\infty) < \infty$. For ease of exposition we will assume that $R(x)$ falls off rapidly enough at infinity to be Lebesgue integrable, $\int_{-\infty}^{\infty} |R(x)|\, dx < \infty$, a condition that is almost certainly satisfied in practice. In this case $F(x)$ is absolutely continuous with continuous and bounded derivative $f(x)$, the spectral density, satisfying

$$R(x) = \int_{-\infty}^{\infty} e^{ix\lambda} f(\lambda) d\lambda. \tag{3.1.1}$$

When the data are regularly spaced, standard texts on stochastic processes such as Bartlett (1978) and Yaglom (1987a) present the (discrete) empirical covariance function

$$B_n(k) = \frac{1}{n}\sum_{i=0}^{n-k} Z(x_i)Z(x_{k+i}) \qquad 0 \le k \le n$$

as the workhorse for inference about $R(x)$. Under the classical time-series asymptotics and the above conditions on $R(x)$ it is a consistent estimator of $R(k)$ as $n \to \infty$. The explicit mathematics hides the implicit assumption that $R(x)$ dies out over a distance small with respect to the typical length of observation.

As we have discussed in §2.1 and observe in §5.3 it is very difficult to extract information about $R(k)$ from $B_n(k)$ for some spatial processes. The reason for this is that the range of correlation of $R(x)$ is comparable to the length of observation so that the information in the data about $R(x)$ is obscured by $B_n(k)$. The 'second order' effects, negligible from the information rich standpoint now play an important role. We now quantify our use of the terms "information rich" and "strongly dependent". Consider the so called "correlation length",

$$T_c \equiv \frac{2}{R(0)}\int_0^\infty R(x)dx = \frac{\pi f(0)}{\int_0^\infty f(\lambda)d\lambda} < \infty,$$

which is a length scale characterizing the strength of correlation between $Z(x)$ and $Z(x+h)$ as h increases. As an example of the relevance of T_c, consider estimating the constant unknown mean of $Z(x)$. We can then interpret T/T_c as the effective number of independent observations in the sense that the variance of the sample mean is approximately $T_cR(0)/T$ (See Yaglom(1987a), §16). The conventional wisdom from time-series is that unless $T \gg T_c$, unknown parameters will be difficult to determine and the estimates would have little value. T_c provides a scale to calibrate the oft used term "for large T". When $T \asymp T_c$,[1]we find that the conventional wisdom

[1] Recall that $f(T) \asymp g(T)$ means $\exists c_1, c_2$ s.t. $0 < c_1 \le f(T)/g(T) \le c_2 < \infty$ for relevant values of T. As a rule-of-thumb we envision $0.2 < c_1 < c_2 < 5$, say.

is not final because the usual objective in spatial prediction is interpolation and not extrapolation. This difference in objective allows a spatial statistician to operate when T is of the same order as T_c. This theme is developed in Stein(1988, 1987b) and will be explored further in this chapter. It is important to note that $T \asymp T_c$ does not mean that we have little information about all characteristics of the covariance structure. In fact, as we shall see in the next section some characteristics can be determined with probability one for *any* T. Understandably this situation has received little research focus because of its reduced importance to time-series.

Much of the work in this chapter was motivated by its close relationship with nearly non-stationary time-series. They have received considerable attention in the Economics and Statistics literature (Phillips (1987a,b), Solo (1984), Dickey & Fuller (1979)). The maximum likelihood and least squares estimates derived in section §3.3 are in fact the asymptotic limits of the corresponding parameters from the nearly non-stationary AR(1) process. In fact the triangular array discussed in §3.2 represents such a series. The results of this chapter should be of interest to researchers in this area.

In the first half of this chapter we investigate maximum likelihood estimation of the covariance structure of the Ornstein–Uhlenbeck process on the basis of observing a single realization continuously on $[0,\ T]$. This has received considerable attention in the literature, especially Arató (1964a,b). We extend and apply this work to obtain the distribution of the maximum likelihood estimates.

When $T \asymp T_c$, the information about certain parameters obtained by discrete observation on $[0, T]$ is bounded. This bound is the information available from a continuously observed record over $[0,\ T]$. In §3.4 we compare the continuous maximum likelihood estimate to the maximum likelihood estimate based on the usual discrete measurement when $T \asymp T_c$. The analysis of the entire segment provides independence from particular sampling schemes as well as providing mathematical tractability at the cost of direct applicability. We shall see that properties of estimators based on continuous observation serve as a useful guide to the properties of their discrete counterparts.

In the second half of this chapter we investigate the use of the spectral density as an alternative estimator for the covariance structure when $T \asymp T_c$. Exact formulas for the covariance of the empirical spectral density process are derived. The behavior of the Ornstein–Uhlenbeck process is studied under both fixed and increasing region asymptotic schemes.

3.2 Some aspects of likelihood estimation for the Ornstein–Uhlenbeck process

Let $Z(x)$ be the Ornstein–Uhlenbeck process with mean β and covariance structure defined by the Exponential covariance function,

$$R(x) = \theta_1 \theta_2 e^{-\frac{x}{\theta_2}},$$

or equivalently the spectral density,

$$f(\lambda) = \frac{\theta_1 \theta_2^2}{\pi(1 + \theta_2^2 \lambda^2)} \tag{3.2.1}$$

This process is the Exponential Gaussian process studied in previous chapters.

In this section we review the maximum likelihood estimation of the parameters $\beta,\ \theta_1,\ \theta_2$ on the basis of observing a single realization continuously on $[0,\ T]$. Let $\{\pi_n\}_{n=1}^{\infty}$ be a sequence of interval partitions on $[0, T]$, $\pi_n = \{0 = t_0^{(n)} < \cdots < t_{n+1}^{(n)} = T\}$. Consider the statistic

$$S_n = \frac{1}{2} \sum_{k=0}^{n} \{Z(t_{k+1}^{(n)}) - Z(t_k^{(n)})\}^2 \tag{3.2.2}$$

Let m_n be the mesh size of π_n, that is, $m_n = \max_{0\leq k\leq n}\{t_{k+1}^{(n)} - t_k^{(n)}\}$. If $m_n \to 0$ then $S_n \xrightarrow{P} \theta_1$ as $n \to \infty$. Further if $m_n \log(n) \to 0$ then $S_n \overset{\text{as}}{\to} \theta_1$ as $n \to \infty$ (Klein & Giné (1975)). In practice this means that on the basis of observing $Z(x)$ on any dense sequence in $[0,T]$ we can determine θ_1 exactly. Hence we need only consider the estimation of β and θ_2. Note also that Ornstein–Uhlenbeck processes with different slope parameters are incompatible, in the sense of Stein (1988a), on any dense sequence of observations in $[0,T]$.

Let P_W be the product of one-dimensional Lebesgue measure and the standard conditional Wiener measure on the space of functions on $[0,T]$. Let $P_{\theta_2,\beta}$ be the measure generated by $Z(x)$ on the product of the real line $Z(0)$ and the space of realizations of $Z(x)-Z(0)$, that is continuous functions on $[0,T]$. Then the Radon–Nikodym derivative of $P_{\theta_2,\beta}$ with respect to P_W is, Striebel (1959),

$$\frac{dP_{\theta_2,\beta}}{dP_W}(Z(x)) = \frac{1}{2\pi\theta_1\theta_2}\exp\{-\frac{1}{2\theta_1\theta_2}[s_1^2 - \theta_1 T + \frac{1}{2\theta_2}s_2^2]\} \tag{3.2.3}$$

where

$$s_1^2 = \frac{1}{2}\{(Z(0)-\beta)^2 + (Z(T)-\beta)^2\},$$
$$s_2^2 = \int_0^T (Z(x)-\beta)^2 dx$$

This result is easy to derive directly from measure–theoretic considerations. However, insight may be gained by considering the triangular arrays of random variables defined by

$$Z_k^n = Z(T(k-1)/n), \quad W_k^n = W(T(k-1)/n) \qquad k = 1,2,\dots n. \tag{3.2.4}$$

For each fixed n, the sequence Z_k^n satisfy the difference equation

$$Z_{k+1}^n = \beta_n Z_k^n + \varepsilon_{k+1}^n \qquad k = 1,2,\dots n$$

where ε_k^n are independent and Gaussian with mean zero and variance $\theta_1\theta_2(1-\beta_n^2)$, $\beta_n = e^{-T/n\theta_2}$.

Let $f_n(Z)$ and $f_n(W)$ be the probability densities of the sequences $Z_1^n, \ldots Z_n^n$ and $W_1^n, \ldots W_n^n$, respectively. One can then derive the result (3.2.3) using a functional Central Limit Theorem (Billingsley (1968), §10) on the likelihood ratio $f_n(Z)/f_n(W)$. The functional Central Limit Theorem and the continuous mapping theorem (Billingsley (1968), §5) are the basis of the bonds between discrete time-series and continuous time-series, and will appear again and again.

To simplify the exposition we will initially focus on the situation where the mean is known. We will take $\beta = 0$. From (3.2.3) the maximum likelihood estimator of $\widehat{\theta}_2$ satisfies

$$\theta_1 \widehat{\theta}_2^2 - \widehat{\theta}_2(s_1^2 - \theta_1 T) - T s_2^2 = 0,$$

so that,

$$\widehat{\theta}_2 = \frac{(s_1^2 - \theta_1 T) + \sqrt{(s_1^2 - \theta_1 T)^2 + 4T\theta_1 s_2^2}}{2\theta_1} \tag{3.2.5}$$

This result was first derived by Striebel (1959). The development given here follows Arató (1964a). Under the assumption that $Z(0) = z_0$, we can show using a triangular array of random variables similar to (3.2.4) that the Radon–Nikodym derivative for this case is

$$\frac{dP^c_{\theta_2}}{dP_W}(Z(x)) = \exp\{-\frac{1}{2{\theta_2}^2}[s_1^2 + 2\theta_2 s_3]\}$$

where $s_3 = \int_0^T Z(x)dZ(x) = \frac{1}{2}\{Z(T)^2 - z_0^2 - T\}$. The conditional maximum likelihood estimator of $\widehat{\theta}_2$ is then

$$\widehat{\theta}^c_2 = -\frac{s_1^2}{s_3} = -\frac{2\int_0^T Z(x)^2 dx}{Z(T)^2 - z_0^2 - T} \tag{3.2.6}$$

which is also the well known least–squares estimator, Bartlett (1978). We will focus on $z_0 = 0$.

3.3 The distributions of the MLE, $\widehat{\theta}_2$, and the conditional MLE, $\widehat{\theta}_2^c$

In the previous section we gave expressions for the maximum likelihood estimates $\widehat{\theta}_2$ and $\widehat{\theta}_2^c$ in (3.2.5) and (3.2.6), respectively. In this section we express their distributions in terms of characteristic functions and use the method of Davies (1973) to determine them numerically. Our approach follows that of Arató (1964b) who proved Theorem 3.3.1 and used the inverse Laplace transform to create a table of quantiles for the distribution of $\widehat{\theta}_2$. The determination of the distribution of $\widehat{\theta}_2^c$ and the graphical comparison between $\widehat{\theta}_2$ and $\widehat{\theta}_2^c$ appear to be new.

Let $\phi_2(z_1, z_2)$ denote the joint characteristic function of s_1^2 and s_2^2. For fixed x, let $\eta_1(x) = xs_1^2 + s_2^2$ then as

$$2\theta_1 x - (s_1^2 - \theta_1 T) \geq 2\theta_1 x - \sqrt{(s_1^2 - \theta_1 T)^2 + 4\theta_1 s_2^2} \geq -\sqrt{(s_1^2 - \theta_1 T)^2 + 4\theta_1 s_2^2}$$

we have,

$$\begin{aligned} P(\widehat{\theta}_2 \leq x) &= P((s_1^2 - \theta_1 T)^2 + 4\theta_1 s_2^2 \leq \{2\theta_1 x - (s_1^2 - \theta_1 T)\}^2) \\ &= P(\eta_1(x) \leq \theta_1 x(x+T)) \end{aligned}$$

Now the characteristic function of $\eta_1(x)$ is $\phi_3(z) = \phi_2(xz, z)$, so that we can determine the distribution of $\widehat{\theta}_2$ at x by inverting $\phi_3(z)$ at $\theta_1 x(x+T)$. Similarly, if we let $\phi_1(z_1, z_2)$ denote the joint conditional characteristic function of s_3 and s_2^2 and $\eta_2(x) \equiv s_2^2 - xs_3$ then

$$\begin{aligned} P(\widehat{\theta}_2^c \leq x) &= P(s_2^2 - xs_3 \leq 0 | Z(0) = 0) \\ &= P(\eta_2(x) \leq 0 | Z(0) = 0) \end{aligned}$$

The conditional characteristic function of $\eta_2(x)$ is $\phi_4(z) = \phi_1(-xz, z)$.

Based on the Gil–Pelaes (1961) formula

$$P(X \leq x) = \frac{1}{2} - \frac{1}{2\pi}\int_{-\infty}^{\infty} Im\Big\{\frac{\phi(t)e^{itx}}{t}\Big\}dt$$

we can express the distributions in terms of the characteristic functions:

$$P(\widehat{\theta}_2 \leq x) = \frac{1}{2} - \frac{1}{2\pi}\int_{-\infty}^{\infty} Im\Big\{\frac{\phi_1(-xt,t)}{t}\Big\}dt \tag{3.3.1}$$

and

$$P(\widehat{\theta}_2^c \leq x) = \frac{1}{2} - \frac{1}{2\pi}\int_{-\infty}^{\infty} Im\Big\{\frac{\phi_2(xt,t)e^{it\theta_1 x(x+T)}}{t}\Big\}dt \tag{3.3.2}$$

The unidentified components are $\phi_1(z_1,z_2)$ and $\phi_2(z_1,z_2)$. Arató (1964a) derives an explicit formula for $\phi_2(z_1,z_2)$ by considering a differential equation in the conditional characteristic function of s_1^2 and s_2^2. His result is:

Theorem 3.3.1 (Arató(1964a), §2):

The joint characteristic function of s_1^2 and s_2^2 is

$$\begin{aligned}\phi_2(z_1,z_2) &\equiv \mathbb{E}(\exp(iz_1 s_1^2 + iz_2 s_2^2))\\ &= \frac{2\sqrt{\theta_2}\Delta^{1/2}e^{\frac{T}{2\theta_2}}}{\{e^{\Delta/\theta_2}(\Delta - 2z_2\theta_2 + T)^2 - e^{-\Delta/\theta_2}(\Delta + 2z_2\theta_2 - T)^2\}^{\frac{1}{2}}}\end{aligned} \tag{3.3.3}$$

where $\Delta = (T^2 - 2z_2\theta_2^2)^{\frac{1}{2}}$.

A similar result holds for the characteristic function $\phi_1(z_1,z_2)$:

Theorem 3.3.2:

The joint conditional characteristic function of s_3 and s_2^2 is

$$\begin{aligned}\phi_1(z_1,z_2) &\equiv \mathbb{E}(\exp(iz_1 s_3 + iz_2 s_2^2 \mid Z(0)=0))\\ &= \frac{\sqrt{2}\Delta^{\frac{1}{2}}e^{\frac{2z_1\theta_2+T}{2\theta_2}}}{\{e^{-\Delta/\theta_2}(\Delta + 2z_1\theta_2 - T) + e^{\Delta/\theta_2}(\Delta - 2z_1\theta_2 + T)\}^{\frac{1}{2}}}\end{aligned} \tag{3.3.4}$$

where $\Delta = (T^2 - 2z_2\theta_2^2)^{\frac{1}{2}}$.

Proof :

The most instructive proof uses the triangular array of random variables (3.2.4). The least–squares estimate of the serial correlation for a discrete time-series under increasing region asymptotics is considered by White (1958). For fixed n, he shows that the characteristic function of

$$\frac{1}{n}\sum_{k=1}^{n} Z_k^n Z_{k-1}^n - \frac{e^{-T/n\theta_2}}{n}\sum_{k=1}^{n}(Z_{k-1}^n)^2 \quad \text{and} \quad \frac{1}{n^2}\sum_{k=1}^{n}(Z_{k-1}^n)^2$$

is $D_n^{-1/2}$ (White (1958), §2) where $D_n(z_1, z_2)$ satisfies

$$\begin{aligned} D_n &= \frac{(1-s_n)r_n^n - (1-r_n)s_n^n}{r_n - s_n} \\ D_1 &= 1 \\ D_2 &= p_n - q_n^2 \end{aligned}$$

and s_n, r_n are the solutions to $x^2 - p_n x + q_n^2 = 0$ for

$$\begin{aligned} p_n &= 1 + \beta_n^2 - \frac{2z_2}{n^2} + \frac{2\beta_n z_1}{n} \\ q_n &= -b_n - \frac{z_1}{n} \\ \beta_n &= e^{-T/n\theta_2} \end{aligned}$$

as before. Of course, White considered β_n constant in his construction. As $n \to \infty$,

$$\begin{aligned} r_n &= 1 + (T\theta_2 z_1 - T^2 + \theta_2\Delta) \cdot \frac{1}{n\theta_2 T} + o(\frac{1}{n}) \\ s_n &= 1 + (T\theta_2 z_1 - T^2 - \theta_2\Delta) \cdot \frac{1}{n\theta_2 T} + o(\frac{1}{n}) \end{aligned}$$

where $\Delta = (T^2 - 2T\theta_2 z_1 - 2\theta_2^2 z_2)^{1/2}$, so that

$$\begin{aligned} D(z_1, z_2) &\equiv \lim_{n\to\infty} D_n(z_1, z_2) \\ &= \frac{e^{(-\Delta+\theta_2 z_1 - T)/\theta_2}(\Delta + \theta_2 z_1 - T) + e^{(\Delta+\theta_2 z_1 - T)/\theta_2}(\Delta - \theta_2 z_1 + T)}{2\Delta} \end{aligned}$$

We now apply the functional central limit theorem to D_n to find that the joint conditional characteristic function of $\int_0^T Z(x)dZ(x) + \frac{T}{\theta_2}\int_0^T Z(x)^2 dx$ and $\int_0^T Z(x)^2 dx$ is D. Finally,

$$\begin{aligned} \phi_1(z_1, z_2) &= \mathbb{E}(\exp(iz_1 s_3 + iz_2 s_2^2 | Z(0) = 0) \\ &= \mathbb{E}(\exp(iz_1(s_3 + Ts_2^2/\theta_2) + is_2^2(z_2 - Tz_1/\theta_2) \ | Z(0) = 0) \\ &= D(z_1, z_2 - Tz_1/\theta_2) \end{aligned}$$

which evaluates to (3.3.4). ∎

3.3.1 *Computation of the distributions of $\widehat{\theta}_2$ and $\widehat{\theta}_2^c$*

Using the expressions (3.3.1) and (3.3.2) we are now in a position to determine the distributions of $\widehat{\theta}_2$ and $\widehat{\theta}_2^c$ numerically. Davies(1973) presents a method for determining the distribution function of a random variable based on the use of the trapezoidal rule to numerically evaluate the integral in the Gil–Pelaes (1961) formula. The advantage of this method is the direct calculation of the distribution function and the possibility of bounding the calculation error.

The number of terms necessary for convergence is huge, of the order of $20,000$ for 5 significant figures accuracy. Clearly, if one wished to invert a large number of these characteristic functions the method used should exploit the structure of the characteristic function to a greater extent, especially in the slowly wavering tails. Both $\phi_1(z_1, z_2)$ and $\phi_1(z_1, z_2)$ involve complex roots, only one path of which gives the correct characteristic function. Theoretically, this path is defined by continuity, but in practice the path is difficult to follow as the arguments in the characteristic functions change.

One drawback of $\widehat{\theta}_2^c$ is that it can produce negative estimates for θ_2. As T_c/T increases the probability of a negative values increases. These values are unacceptable, and meaningless within the definition of the Ornstein–Uhlenbeck process.

Consider first the situation where the length of observation is the same as the correlation length, that is, $T_c = T$.

Figure 32 compares the distribution of the conditional maximum likelihood estimate for θ_2, $\widehat{\theta}_2^c$ to the distribution of the full maximum likelihood estimate, $\widehat{\theta}_2$. Although it is obscured in the graph, 3% of the values for $\widehat{\theta}_2^c$ are negative. The conditional maximum likelihood estimate is also extremely heavy right tailed, with values for θ_2 of 50 or more arising occasionally.

If we truncate the conditional maximum likelihood estimate at zero, we can compare the two estimates for $\log(\theta_2)$, and so reduce the skewness. Figure 33 compares the distribution of the truncated conditional maximum likelihood estimate for $\log(\theta_2)$, $\log(\widehat{\theta}_2^c)$ to the distribution of the full maximum likelihood estimate, $\log(\widehat{\theta}_2)$. The truncated values could be represented by a point mass of 3% at negative infinity. It is clear that $\log(\widehat{\theta}_2)$ is very well approximated by the Gaussian distribution and that it is slightly biased downward. The conditional maximum likelihood estimate is still slightly right skewed, although it is reasonably close to the full maximum likelihood estimate.

Consider next the situation where the dependence between the observations is increased to $T_c = 2T$. In this case 10% of the values of the conditional maximum likelihood estimate are negative. Figure 34 compares the distribution of the truncated conditional maximum likelihood estimate for $\log(\theta_2), \log(\widehat{\theta}_2^c)$ to the distribution of the full maximum likelihood estimate, $\log(\widehat{\theta}_2)$. The truncated values could be represented by a point mass of 10% at negative infinity. Again the $\log(\widehat{\theta}_2)$ is very well approximated by the Gaussian distribution. The right skewness of the conditional maximum likelihood estimate has increased and there is greater difference between distributions.

In summary, the maximum likelihood estimate of $\log(\theta_2)$ is well approximated by a Gaussian distribution, while the conditional maximum likelihood estimate is slightly skewed right. The non-negligible probability of obtaining a negative estimate from the conditional maximum likelihood estimate is reason enough for preferring the full maximum likelihood estimate. As the dependence among the observations increases the conditional maximum likelihood estimate deteriorates, tending to produce occasional large estimates.

3.4 Approximation of the discrete observation MLE by the MLE based on a finite segment of a single realization

In this section we compare the maximum likelihood estimate based on observing the segment $[0, T]$ of a realization to the maximum likelihood estimate based on the usual discrete measurement in $[0, T]$ when $T \asymp T_c$. When $T \asymp T_c$, the information about certain parameters obtained by discrete observation on $[0, T]$ is bounded. This bound is the information available from a continuously observed record over $[0, T]$. The continuous version is in fact an asymptotic limit of the discrete version in a sense that can be made precise. We shall see that properties of estimators based on continuous observation serve as a useful guide to the properties of their discrete counterparts.

As indicated the slope parameter, θ_1, for the Ornstein–Uhlenbeck process can be determined with arbitrary precision based on a continuous record. For regular discrete observation the natural estimator S_n from (3.2.2) has variance $2\theta_1^2/n + O(1/n^2)$ as $n \to \infty$. Hence interest focuses on the maximum likelihood estimate of the range parameter, θ_2. As in §3.2 we can consider the situation where $\theta_1 = T = 1$ and re-parameterize to recover the general situation. The distribution of the continuous version has been studied in §3.3. The properties of the discrete version has been studied in §2.4. These distributions do not have closed form representations. The correlation length for the Ornstein–Uhlenbeck process is $T_c = 2\pi f(0)/R(0) = 2\theta_2$.

Consider first the situation where the length of observation is the same as the correlation length, that is, $T_c = T$. Figure 35 compares the distribution of the discrete maximum likelihood estimate, $\widehat{\theta}_2^n$ for $n = 10,\ 25,\ 50$ to the distribution of the continuous version, $\widehat{\theta}_2$. We see that for n as small as 10 the distributions are similar. For $n = 50$ the distributions are almost identical. As n increases the distributions become less right-skewed and, as expected, the variance decreases. Figure 36 represents the theoretical (rotated) quantile plot for $\log(\widehat{\theta}_2)$. This is the usual quantile plot where the $45°$ reference line has been rotated to the horizontal. The log-transform removes the right-skewedness so that $\log(\widehat{\theta}_2)$ is very well approximated by a Gaussian distribution.

Consider next the situation where the length of observation is the half the correlation length, $T_c = 2T$, so that the dependence between the observations is increased. Figure 38 compares the distribution of the discrete maximum likelihood estimate, $\widehat{\theta}_2^n$ for $n = 5,\ 10,\ 25$ to the distribution of the continuous version, $\widehat{\theta}_2$. We see faster convergence of the discrete maximum likelihood estimate to the continuous version. For $n = 25$ the distributions are almost identical. The path of convergence is similar to the previous situation. Figure 37 represents the theoretical (rotated) quantile plot for $\log(\widehat{\theta}_2)$. The log-transform removes the right-skewedness, but shortens the tails a little too much. Still $\log(\widehat{\theta}_2)$ is well approximated by a Gaussian distribution.

As T_c/T increases, the continuous approximation becomes suitable for smaller sample sizes. In addition the skewness of the distribution itself increases. As T_c/T decreases the sample size required for a good approximation increases. Of course, under the fixed region sampling scheme, the distribution of the limiting form is always the continuous version. If $T_c/T < 5\%$ the approximation will be poor for all but the largest sample sizes.

We are now in a position to check the adequacy of the approximation (2.4.8) to the variance of the maximum likelihood estimate of $\log(\widehat{\theta}_2)$ calculated in §2.4.2.

Figure 39 plots the variance of $\log(\widehat{\theta}_2^n)$ as a function of n, the density of the observations. The true range is $\theta_2 = \frac{1}{2}$, that is $T_c = T$. Under this transformation the distribution is approximately Gaussian. Also plotted is the mean-squared error of $\log(\widehat{\theta}_2^n)$. The values are calculated based on 20000 simulations. The dither in the plot is due to the residual sample variation from the simulation. For $n < 40$ the function decreases like $1/n^2$ and levels off completely for n greater than 60. The dashed line is the approximation of (2.4.8), $\frac{2T_c}{2T+T_c} = 0.667$. We see that the approximation overestimates the true values by about 10%, but provides a better indication to the mean-squared error. The approximation is surprisingly good.

Note that the bias of the maximum likelihood estimate does not vanish as n increases, and in fact increases as n increases, leveling off for $n > 30$ at the bias of the continuous maximum likelihood estimate.

Figure 40 plots the variance of $\log(\widehat{\theta}_2^n)$ when the dependence is increased to $T_c = 2T$. For $n < 60$ the function decreases like $\frac{1}{n^2}$ and levels off completely for n greater than 60. The approximation of (2.4.8) is $\frac{2T_c}{2T+T_c} = 1.0$. We see that the approximation again overestimates the true values by about 10% and underestimates the mean-squared error by about 5%.

In summary, if $T_c \asymp T$, the continuous approximation to discrete sampling is quite good for small to moderate values of n. The distribution of the maximum likelihood estimate of the range parameter is approximately log-Gaussian. Both $\widehat{\theta}_2$ and $\log(\widehat{\theta}_2)$ suffer from moderate bias.

3.5 On the determination of the spectrum based on a finite segment of a single realization

In this section we study the empirical spectral density as a guide to the covariance structure when $T_c \asymp T$. We consider the empirical spectral density with the objective of inferring the covariance structure of $Z(x)$. It is hoped that the spectral domain will yield information that the time domain obscures. In particular the objective is to derive a graphical tool by which the covariance structure can be surmised. It is motivated by the success of the empirical spectral density in time-series when

$T_c \ll T$.

3.5.1 *The empirical spectral density,* $I_T(\lambda)$

Suppose we observe one realization of $Z(x)$ completely on $[0,T]$, $T>0$. Define the empirical spectral density, $I_T(\lambda)$ on $[0,T]$ to be the random process,

$$I_T(\lambda) = \frac{1}{2\pi T}\left|\int_0^T e^{ix\lambda} Z(x)dx\right|^2$$

This is the periodogram from the time-series literature. We call it the empirical spectral density to emphasize its relation to the empirical covariance function and the theoretical spectral density. Consider the Fourier transform of $I_T(\lambda)$

$$\begin{aligned} B_T(x) &= \int_{-\infty}^{\infty} e^{-ix\omega} I_T(\omega)d\omega \\ &= \frac{1}{T}\int_0^{T-x} Z(t)Z(t+x)dt \qquad 0 \le x \le T \end{aligned}$$

We see that $B_T(x)$ is the empirical covariance function and $I_T(\lambda)$ is the empirical spectral density. More importantly, they are a Fourier transform pair, so that the information in the empirical spectral density is the same as in the empirical covariance function. This relationship is not as well known as it should be. The above mentioned convergence properties of $B_T(x)$ to $R(x)$ are consequences of a result of Bartlett (1946):

Result 3.5.1:

The non-Gaussian process $\{B_T(x) : 0 \le x \le T\}$ has mean function

$$\mathbb{E}B_T(x) = \frac{T-x}{T}R(x)$$

and (non-stationary) covariance function

$$\text{Cov}(B_T(x_1), B_T(x_2)) = \frac{1}{T^2}\int_{-(T-x_1)}^{T-x_2} \phi(x)\{R(x)R(x+x_2-x_1)+R(x+x_2)R(x-x_2)\}\, dx$$

where $0 \le x_1 \le x_2$ and,

$$\phi(x) = \left\{ \begin{array}{lll} T - x_1 + x - (T - x_1) \le & x \le -(x_2 - x_1) \\ T - x_2 - (x_2 - x_1) \le & x \le 0 \\ T - x_2 - x \quad\; 0 \le & x \le T - x_2 \end{array} \right.$$

Proof : Both equations can be obtained by direct calculation. See, for example, Bartlett (1946) or Jenkins & Watts (1968),§5.3. One can show that this covariance function is strictly positive. ∎

Our primary objective is to use the observed values of $I_T(\lambda)$ for $\lambda > 0$ to determine $R(x)$ or equivalently estimate the characteristics of the spectrum and the spectral density. It is clear that for fixed T it is impossible, without additional restrictions on the form of $R(x)$, to completely determine $R(x)$ at distances greater than T. Hence we expect that the behavior of $f(\lambda)$ for small frequencies will be difficult to determine, but hope that the behavior at high frequencies will provide useful information.

3.5.2 *The frequency domain process,* $J_T(\omega)$

In this section we consider the properties of the a stochastic process closely related to $I_T(\lambda)$. The results will be used in later sections of this chapter.

Consider the stochastic process $\{J_T(\omega) : \omega > 0\}$ defined by

$$\begin{aligned} J_T(\omega) &= \frac{1}{\sqrt{2\pi T}} \int_0^T e^{ix\omega} Z(x)dx \\ &= J_R(\omega) + iJ_I(\omega) \end{aligned}$$

where

$$J_R(\omega) = \frac{1}{\sqrt{2\pi T}} \int_0^T \cos(x\omega)\ Z(x)dx, \qquad J_I(\omega) = \frac{1}{\sqrt{2\pi T}} \int_0^T \sin(x\omega)\ Z(x)dx$$

For fixed T, this process is the Fourier transform of $Z(x)$ on $[0,\ T]$ and represents a spectral version of the information in $Z(x)$ on $[0,\ T]$. It is natural to use $J_T(\omega)$ when exploring the spectral properties of $Z(x)$. This process, and its discrete analogue, are used implicitly in text books such as Jenkins & Watts (1968), Yaglom (1987a), Brillinger (1975), Priestley (1981), etc. We can summarize the statistical characteristics in:

Result 3.5.2:

The complex Gaussian process $\{J_T(\omega) : \omega > 0\}$ has zero mean and real (non-stationary) covariance function

$$C_f(\omega_1, \omega_2) \equiv \mathrm{Cov}(J_T(\omega_1), J_T(\omega_2)) = \frac{T}{2\pi} \int_{-\infty}^{\infty} h_T(\lambda - \omega_1) \cdot h_T(\lambda - \omega_2)\ f(\lambda)\ d\lambda \quad (3.5.1)$$

where $h_T(\omega) = \frac{\sin(T\omega/2)}{T\omega/2}$.

In particular

$$\begin{aligned}\mathrm{Cov}(J_R(\omega_1), J_R(\omega_2)) &= \frac{1}{2}\{C_f(\omega_1, \omega_2) + C_f(\omega_1, -\omega_2)\}\\ \mathrm{Cov}(J_I(\omega_1), J_I(\omega_2)) &= \frac{1}{2}\{C_f(\omega_1, \omega_2) - C_f(\omega_1, -\omega_2)\}\\ \mathrm{Cov}(J_R(\omega_1), J_I(\omega_2)) &= 0\end{aligned}$$

Proof : We note that $J_T(-\omega) = \overline{J_T(\omega)}$ This proof is derived from Yaglom (1987a), Ch.3, footnote 37. Using the spectral representation (3.1.1) for $R(x)$ and a little algebra we can show that

$$C_f(\omega_1, \omega_2) = \frac{1}{2\pi T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} \cos(\omega_1 t - \omega_2 s) R(t - s)\ dtds$$

and hence, for example, $\mathrm{Cov}(J_R(\omega_1), J_R(\omega_2)) = \frac{1}{2}\{C_f(\omega_1, \omega_2) + C_f(\omega_1, -\omega_2)\}$. Finally,

$$\mathrm{Cov}(J_T(\omega_1), J_T(\omega_2)) = \mathrm{Cov}(J_R(\omega_1), J_R(\omega_2)) + \mathrm{Cov}(J_I(\omega_1), J_I(\omega_2)) = C_f(\omega_1, \omega_2).$$

∎

The empirical spectral density process,

$$I_T(\lambda) = |J_T(\omega)|^2 = J_R^2(\omega) + J_I^2(\omega),$$

is just the magnitude squared of $J_T(\omega)$, so by considering the empirical spectral density process alone we are ignoring any information in the phase of $J_T(\omega)$.

3.6 The covariance structure of the empirical spectral density process

In this section we derive expressions for the mean and covariance for the stochastic process $I_T(\lambda)$. These expressions are the basis for the analysis of $I_T(\lambda)$ in the subsequent sections. From Result 3.5.2 we have:

$$\begin{aligned} \mathbb{E} I_T(\lambda) &= \frac{T}{2\pi} \int_{-\infty}^{\infty} h_T^2(\omega - \lambda) f(\omega)\, d\omega = C_f(\lambda, \lambda), \\ \mathrm{Cov}(\ I_T(\omega_1), I_T(\omega_2)\ ) &= C_f^2(\omega_1, \omega_2) + C_f^2(\omega_1, -\omega_2), \end{aligned} \tag{3.6.1}$$

where $\omega_1,\ \omega_2 > 0$.

The following identity is essential:

Lemma 3.6.1:

$$\frac{T}{2\pi} \int_{-\infty}^{\infty} h_T(\lambda - \omega_1) \cdot h_T(\lambda - \omega_2)\, d\lambda \quad = \quad h_T(\omega_2 - \omega_1) \qquad \forall \omega_1,\ \omega_2 > 0$$

Proof : First note that

$$\begin{aligned} \frac{T^2}{4\pi^2} \Big| \int_{-\infty}^{\infty} h_T(\lambda - \omega_1) \cdot h_T(\lambda - \omega_2)\, d\lambda\ \Big|^2 &\le \frac{T^2}{4\pi^2} \int_{-\infty}^{\infty} h_T^2(\lambda - \omega_1) d\lambda \cdot \int_{-\infty}^{\infty} h_T^2(\lambda - \omega_2) d\lambda \\ &= 1\ < \infty, \end{aligned}$$

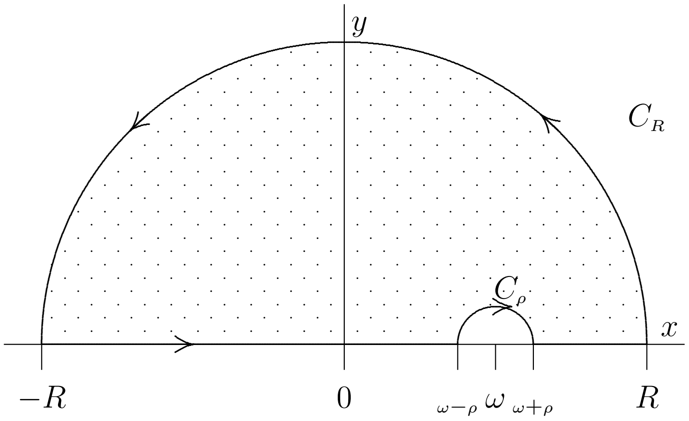
as $\frac{T}{2\pi} \int_{-\infty}^{\infty} h_T^2(\lambda)\ d\lambda = 1$ and using the Cauchy-Schwarz inequality, so the integral exists. Let $\omega = \frac{T(\omega_2 - \omega_1)}{4}$ then

$$\frac{T}{2\pi} \int_{-\infty}^{\infty} h_T(\lambda - \omega_1) \cdot h_T(\lambda - \omega_2)\, d\lambda = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin(x + \omega) \sin(x - \omega)\ dx}{(x + \omega)(x - \omega)}.$$

In order to evaluate this last integral we can consider the result of integrating the function

$$f(z) = \frac{e^{2i\omega} - e^{2iz}}{2\pi (z + \omega)(z - \omega)}$$

around the following closed contour:



where we can take $0 < \omega < R$. We note that

$$\lim_{z\to\omega}(z-\omega)f(z) = \frac{e^{-2i\omega}-e^{2i\omega}}{4\pi\omega} \quad \text{and} \quad \lim_{z\to-\omega} f(z) = \frac{ie^{-2i\omega}}{2\pi\omega},$$

so that $f(z)$ has the single simple pole on the real axis at $z = \omega$. As $f(z)$ is analytic inside the contour, Cauchy's residue theorem gives

$$\int_{-R}^{\omega-\rho} f(x)\ dx + \int_{C_\rho} f(z)\ dz + \int_{\omega+\rho}^{R} f(x)\ dx + \int_{C_R} f(z)\ dz \quad = \quad 0 \tag{3.6.2}$$

By Jordan's Lemma,

$$\lim_{R\to\infty}\int_{C_R} f(z)\ dz = 0.$$

As C_ρ is traveled in the negative sense, by Cauchy's residue theorem, we have

$$\begin{aligned}\lim_{\rho\to 0}\int_{C_\rho} f(z)\ dz &= -\pi i Res(\omega)\\ &= \frac{-i(e^{-2i\omega}-e^{2i\omega})}{4\omega} = -\frac{\sin(2\omega)}{2\omega}\end{aligned}$$

Thus, proceeding to the limit as $\rho \to 0$ and $R \to \infty$, we have

$$-\!\!\!\!\!\!\int_{-\infty}^{\infty} \frac{e^{2i\omega}-e^{2ix}}{2\pi(x^2-\omega^2)}\ dx = \frac{\sin(2\omega)}{2\omega}$$

Equating real and imaginary parts of both sides we obtain

$$\begin{aligned}\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sin(x+\omega)\sin(x-\omega)\ dx}{(x^2-\omega^2)} &= \frac{\sin(2\omega)}{2\omega}\\ \frac{1}{\pi}-\!\!\!\!\!\!\int_{-\infty}^{\infty} \frac{\sin(x+\omega)\cos(x-\omega)\ dx}{(x^2-\omega^2)} &= 0\end{aligned}$$

where the principal value sign is omitted on the first integral as we have shown it converges, the integrand being finite at all x. ∎

We now rewrite the expression for the covariance of $I_T(\omega_1)$ and $I_T(\omega_2)$ in (3.6.1) in a form better suited for calculations when T is fixed and $\omega_2 \geq \omega_1 \to \infty$. Let

$$\gamma = \frac{T(\omega_2 + \omega_1)}{2}, \quad \beta = \frac{T(\omega_2 - \omega_1)}{2} \tag{3.6.3}$$

then from equation (3.5.1),

$$\begin{aligned} h_T(\lambda/T - \omega_1) \cdot h_T(\lambda/T - \omega_2) &= 2 \cdot \frac{\cos(\frac{T(\omega_2-\omega_1)}{2}) - \cos(\lambda - \frac{T(\omega_2+\omega_1)}{2})}{(\lambda - T\omega_1)(\lambda - T\omega_2)} \\ &= 2 \cdot \frac{\cos(\beta) - \cos(\lambda - \gamma)}{(\lambda - \gamma + \beta)(\lambda - \gamma - \beta)} \end{aligned}$$

Then

$$\begin{aligned} C_f(\omega_1, \omega_2) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} h_T(\lambda/T - \omega_1) \cdot h_T(\lambda/T - \omega_2)\ f(\lambda/T)\ d\lambda \\ &= \frac{T}{4\pi} \int_{-\infty}^{\infty} \frac{(\cos(\beta) - \cos(\lambda - \gamma)) \cdot f(\lambda/T)}{(\lambda - \gamma + \beta)(\lambda - \gamma - \beta)}\ d\lambda \\ &= \frac{T}{4\pi} \int_{-\infty}^{\infty} \frac{(\cos(\beta) - \cos(\lambda)) \cdot f(\frac{\lambda+\gamma}{T})}{(\lambda^2 - \beta^2)}\ d\lambda \\ &\equiv Q_f(\beta, \gamma) \end{aligned}$$

Finally, we have:

Result 3.6.1:

$$\begin{aligned} \mathbb{E} I_T(\lambda) &= Q_f(0, T\lambda), \\ \mathbb{V}(I_T(\lambda)) &= Q_f^2(0, T\lambda) + Q_f^2(T\lambda, 0) \\ \mathrm{Cov}_f(\ I_T(\omega_1), I_T(\omega_2)\ ) &= Q_f^2(\beta, \gamma) + Q_f^2(\gamma, \beta) \end{aligned} \tag{3.6.4}$$

where β and γ are given in (3.6.3). The last equality follows because $f(\lambda)$ is symmetric about zero. β represents the difference between frequencies and γ the average frequency.

3.7 The behavior of $I_T(\lambda)$ for the Ornstein–Uhlenbeck process

If $f(\lambda)$ is arbitrary and T fixed, (3.6.4) results in intractable expressions for the covariance structure of $I_T(\lambda)$. In this section we consider the behavior of $I_T(\lambda)$ for the Ornstein–Uhlenbeck process defined in §3.2. For the spectral density (3.2.1), $Q(\beta,\gamma)$ becomes

$$\frac{\theta_1 T^2}{\pi^2}\int_{-\infty}^{\infty}\frac{\cos(\beta)-\cos(\lambda)}{(\lambda^2-\beta^2)(\alpha^2+(\lambda+\gamma)^2)}\,d\lambda$$

where $\alpha = T/\theta_2$. The following algebraic identity can be easily checked by multiplication:

$$\frac{1}{(\lambda^2-\beta^2)(\alpha^2+(\lambda+\gamma)^2)} = \frac{1}{\eta(\alpha,\beta,\gamma)}\cdot\left\{\frac{2\gamma\lambda+2\gamma^2-\delta}{\alpha^2+(\lambda+\gamma)^2} - \frac{2\gamma\lambda-2\gamma^2-\delta}{\lambda^2-\beta^2}\right\}$$

where $\eta(\alpha,\beta,\gamma) = (\alpha^2+(\gamma+\beta)^2)(\alpha^2+(\gamma-\beta)^2)$ and $\delta = \beta^2-\gamma^2+\alpha^2$. The following integrals are useful below:

$$\int_0^\infty \frac{\cos(\lambda)}{\alpha^2+\lambda^2}\,d\lambda = \frac{\pi}{2\alpha}e^{-\alpha} \qquad \int_0^\infty \frac{d\lambda}{\alpha^2+\lambda^2} = \frac{\pi}{2\alpha} \qquad \int_0^\infty \frac{\lambda\sin(\lambda)}{\alpha^2+\lambda^2}\,d\lambda = \frac{\pi}{2}e^{-\alpha}$$

We can then write

$$\frac{\pi^2}{\theta_1 T^2}\cdot Q(\beta,\gamma)\cdot\eta(\alpha,\beta,\gamma) =$$

$$\int_{-\infty}^{\infty}\frac{(\cos(\beta)-\cos(\lambda-\gamma))\cdot(2\gamma\lambda-\delta)}{\alpha^2+\lambda^2}\,d\lambda \quad + \quad (2\gamma^2+\delta)\int_0^\infty\frac{\cos(\beta)-\cos(\lambda)}{\lambda^2-\beta^2}\,d\lambda$$

$$= \quad 2\delta\cos(\gamma)\int_0^\infty\frac{\cos(\lambda)}{\alpha^2+\lambda^2}\,d\lambda \quad - \quad 2\delta\cos(\beta)\int_0^\infty\frac{d\lambda}{\alpha^2+\lambda^2}$$

$$+\,(2\gamma^2+\delta)\int_0^\infty\frac{\cos(\beta)-\cos(\lambda)}{\lambda^2-\beta^2}\,d\lambda \quad - \quad 4\gamma\sin(\gamma)\int_0^\infty\frac{\lambda\sin(\lambda)}{\alpha^2+\lambda^2}\,d\lambda$$

$$= \frac{\pi}{\alpha}\{\delta\cos(\gamma)e^{-\alpha} - \delta\cos(\beta) + \alpha(2\gamma^2+\delta)\sin(\beta)/\beta - 2\alpha\gamma\sin(\gamma)e^{-\alpha}\}$$

$$= \frac{\pi}{\alpha}\{(\gamma^2-\beta^2-\alpha^2)[\cos(\beta)-e^{-\alpha}\cos(\gamma)] - 2\alpha e^{-\alpha}\gamma\sin(\gamma) + \alpha(\alpha^2+\beta^2+\gamma^2)\sin(\beta)/\beta\}$$

where we have used Lemma 3.6.1 in the second to the last line. Finally, we have

$$Q(\beta,\gamma) = \frac{\theta_1 T^2\{(\gamma^2-\beta^2-\alpha^2)[\cos(\beta)-e^{-\alpha}\cos(\gamma)] - 2\alpha e^{-\alpha}\gamma\sin(\gamma) + \alpha(\alpha^2+\beta^2+\gamma^2)\sin(\beta)/\beta\}}{\pi\alpha[\alpha^2+(\gamma+\beta)^2][\alpha^2+(\gamma-\beta)^2]}$$

$$Q(\gamma,\beta) = \frac{\theta_1 T^2\{(\beta^2-\gamma^2-\alpha^2)[\cos(\gamma)-e^{-\alpha}\cos(\beta)] - 2\alpha e^{-\alpha}\beta\sin(\beta) + \alpha(\alpha^2+\gamma^2+\beta^2)\sin(\gamma)/\gamma\}}{\pi\alpha[\alpha^2+(\gamma+\beta)^2][\alpha^2+(\gamma-\beta)^2]}$$

and

$$Q(0,\gamma) = \frac{\theta_1 T^2\Big\{(\gamma^2-\alpha^2)[1-e^{-\alpha}\cos(\gamma)] - 2\alpha e^{-\alpha}\gamma\sin(\gamma) + \alpha(\alpha^2+\gamma^2)\Big\}}{\pi\alpha(\alpha^2+\gamma^2)^2}$$

$$Q(\gamma,0) = \frac{\theta_1 T^2\Big\{\alpha\sin(\gamma)/\gamma - \cos(\gamma) + e^{-\alpha}\Big\}}{\pi\alpha(\alpha^2+\gamma^2)}$$

From (3.6.4) we have,

$$\begin{aligned}\mathbb{E}I_T(\lambda) &= Q(0,T\lambda)\\ &= \frac{\theta_1 T^2\Big\{(T^2\lambda^2-\alpha^2)[1-e^{-\alpha}\cos(T\lambda)] - 2\alpha e^{-\alpha}T\lambda\sin(T\lambda) + \alpha(\alpha^2+T^2\lambda^2)\Big\}}{\pi\alpha(\alpha^2+T^2\lambda^2)^2}\end{aligned} \tag{3.7.1}$$

Similarly,

$$Q(T\lambda,0) = \frac{\theta_1\theta_2^2}{\pi T(1+\theta_2^2\lambda^2)}\left\{\theta_2 e^{-T/\theta_2} - \theta_2\cos(T\lambda) + \frac{\sin(T\lambda)}{\lambda}\right\}$$

so we can fully express $\mathbb{V}(I_T(\lambda)) = Q^2(T\lambda,0)+Q^2(0,T\lambda)$ and $\mathrm{Cov}(\ I_T(\omega_1), I_T(\omega_2)\ )$.

These expressions describe the mean and covariance structure of $I_T(\lambda)$ for any T. Even with $f(\lambda)$ of such a simple form these expressions are very complex. For alternative interesting $f(\lambda)$ the analogous expressions become rapidly intractable.

3.8 Using $I_T(\lambda)$ to determine the local behaviour of a stochastic process

In this section we consider using $I_T(\lambda)$, or equivalently $B_T(x)$, to estimate the local behavior of the random field when $T \asymp T_c$. In addition to the assumption that $R(x)$ be Lebesgue integrable we will insist that $R(x)$ has uniformly bounded second derivative on $(0,T]$, so that the left and right derivatives of $R(x)$ exist and are finite. We define the variogram of $Z(x)$ by

$$2\gamma(h) \equiv \mathbb{E}\{(Z(x+h)-Z(x))^2\}$$

so that we have

$$\lim_{h\to 0^+} \frac{\gamma(h)}{h} = -\frac{1}{2}R^{'}(0^+) \equiv \theta \geq 0. \tag{3.8.1}$$

If $Z(x)$ is mean-square differentiable then necessarily $\theta = 0$. We will consider only those random fields for which $\theta > 0$. If $Z(x)$ is $n \geq 1$ times differentiable then the n^{th} mean-square derivative of $Z(x)$ is non-differentiable with covariance function $R^{(2n)}(x)$. Thus we consider a class of stochastic processes that are mean–square continuous but not mean–square differentiable. Furthermore, the realizations of $Z(x)$ are a.s. continuous and, in fact, satisfy a Lipschitz condition of order up to $\frac{1}{2}$. For this class of processes, θ determines the behavior at the origin of $R(x)$ and hence the local behavior of the process. On the basis of observing $Z(x)$ on any dense sequence in $[0, T]$ we can use S_n from (3.2.2) to determine θ with arbitrary precision.

It is possible to express (3.8.1) in terms of the spectral density,

$$\gamma(h) = 2\int_{-\infty}^{\infty} \sin^2 (h\lambda/2) f(\lambda) d\lambda$$

or conversely

$$\pi\lambda\Big\{1 - F(\lambda)\Big\} = \int_{-\infty}^{\infty} \sin (h\lambda) \frac{\gamma(h)}{h} dh$$

Using approximation theory for this singular integral it is possible to confirm that

$$\lim_{\lambda\to\infty} \pi\lambda \int_{\lambda}^{\infty} f(\omega) d\omega = \theta \tag{3.8.2}$$

This equation reflects the relationship between the behavior of $R(x)$ at the origin and the behavior of the spectral density at infinity.

It is clear that based on observing one realization of $Z(x)$ completely on $[0, T]$ the covariance $R(x)$ can not be determined completely.

Our primary objective, then, is to use the observed values of the empirical spectral density process $\{I_T(\lambda) : \lambda > 0\}$ to determine θ. In the following subsections we will argue that this is not possible and that θ can not be recovered from $I_T(\lambda)$ or $B_T(x)$ alone.

3.8.1 *Determination of the local behavior by weighting* $I_T(\lambda)$

In this section we investigate the estimation of θ by weighted integrals of $I_T(\lambda)$. Based on (3.8.2) we might expect

$$\pi\lambda\int_{\lambda}^{\infty} I_T(\omega)d\omega \overset{\text{as}}{\to} \theta \quad \text{as} \quad \lambda\to\infty$$

and we would be surprised if a suitable weight function $g(\omega)$ cannot be found so that

$$\pi\lambda\int_{\lambda}^{\infty} g(\omega)I_T(\omega)d\omega \overset{\text{as}}{\to} \theta \quad \text{as} \quad \lambda\to\infty$$

Hence it is natural to consider the statistics $\tau_n = \sum_{i=1}^{n} a_i I_T(\omega_i)$ where ω_i are arbitrary frequencies and a_i are arbitrary weights. Our intent is to estimate θ by τ_n.

We should choose the weights and frequencies so that

$$\frac{\mathbb{V}(\tau_n)}{\mathbb{E}^2(\tau_n)} \to 0,$$

as $n\to\infty$. If we define

$$\Sigma_n = \left\{ \text{Corr}(\ I_T(\omega_i), I_T(\omega_j)\ ) \right\}_{n\times n} \qquad \nu_n = \left\{ \frac{\mathbb{E}(I_T(\omega_i))}{\sqrt{\mathbb{V}(I_T(\omega_i))}} \right\}_{n\times 1}$$

then we can show we must choose the sequence of frequencies, $\omega_1, \omega_2, \ldots$ so that $\nu_n'\Sigma_n^{-1}\nu_n$ is unbounded. Unfortunately it is does not appear possible to divine from (3.6.1) and (3.5.1) the structure of Σ_n for general $f(\lambda)$. Because of this we return again to the Ornstein–Uhlenbeck process for which we do have an exact expression for Σ_n.

Result 3.8.1:

For the Ornstein–Uhlenbeck process of (3.2.1) we have:

$$\begin{aligned}\mathbb{E}(I_{\scriptscriptstyle T}(\omega)) =& \frac{\theta_1(1+\alpha-e^{-\alpha}\cos(T\omega))}{\pi\alpha\omega^2} + O(\frac{1}{\omega^3}) \\ \mathrm{Corr}(\ I_{\scriptscriptstyle T}(\omega_1), I_{\scriptscriptstyle T}(\omega_2)\ ) =& \frac{1}{\sqrt{d(\omega_1)\cdot d(\omega_2)}}[\ a^2(\beta,\gamma)+a^2(\gamma,\beta)]\ + O(1/\omega_1)+O(1/\omega_2)\end{aligned}$$

as $\omega_1, \omega_2 \to \infty$, where

$$\gamma = \frac{T(\omega_2+\omega_1)}{2}, \qquad \beta = \frac{T(\omega_2-\omega_1)}{2}, \qquad \alpha = T/\theta_2 = 2T/T_c$$

and

$$\begin{aligned}d(\omega) &= (1+\alpha-e^{-\alpha}\cos T\omega)^2 + (e^{-\alpha}-\cos T\omega)^2 \\ a(\beta,\gamma) &= \cos\gamma - e^{-\alpha}\cos\beta + \frac{\alpha\sin\gamma}{\gamma}\end{aligned}$$

Proof : The parameter α is a measure of the amount of information for θ in the sense that $T_c = 2T/\alpha$. The expressions are derived from the expressions in §3.7 by careful and tedious calculations. They have been reformulated in terms of β and γ to emphasize the periodic structure of the covariance. Note that the correlation is always positive. ∎

We see that $\mathbb{E}(I_{\scriptscriptstyle T}(\omega))$ decreases to zero like $f(\omega)$. More importantly the correlation structure does *not* die out as β or γ increases. In fact $\mathrm{Corr}(\ I_{\scriptscriptstyle T}(\omega_1), I_{\scriptscriptstyle T}(\omega_2)\ )$ approaches a periodic function as ω_1 and ω_2 increase. This is in sharp contrast to the behavior for increasing region asymptotics, described in Result 3.9.1e, where the correlation drops to zero as the difference between frequencies increases. Viewed as a function of ω_1 and ω_2 , $\mathrm{Corr}(\ I_{\scriptscriptstyle T}(\omega_1), I_{\scriptscriptstyle T}(\omega_2)\ )$ looks like an upside down egg carton stretching out to infinity. The dips almost go down to zero and the hills raise up to values depending on T_c/T.

We must choose the frequencies in τ_n so that the successive components are as weakly dependent as possible, so that the total information in them about θ

grows. To minimize the correlation we need to minimize $a(\beta, \gamma)$ as $\gamma \to \infty$. For a given frequency, ω_1, we can show that the sequence of frequencies, ω_2, with $I_T(\omega_1)$ approximately uncorrelated with $I_T(\omega_2)$ satisfy,

$$T(\omega_2 - \omega_1) = (2n+1)\pi + \frac{2(1-e^{-2\alpha})}{\alpha\pi(n+\frac{1}{2})} + O(\frac{1}{n^2}) \qquad n = 1, 2, \ldots$$

However, if we choose a third frequency, ω_3, so that $I_T(\omega_3)$ is approximately uncorrelated with $I_T(\omega_2)$ then it will differ from ω_1 by approximately a multiple of $2\pi/T$. From Result 3.8.1,

$$\mathrm{Corr}(\ I_T(\omega_1), I_T(\omega_3)\ ) \geq \frac{(1-e^{-\alpha}\cos(T\omega_1))^2 + (\cos(T\omega_1) - e^{-\alpha})^2}{(1+\alpha-e^{-\alpha}\cos(T\omega_1))^2 + (\cos(T\omega_1)-e^{-\alpha})^2} > 0 \quad (3.8.3)$$

That is, we can choose two frequencies at which $I_T(\cdot)$ has arbitrarily small correlation, but $I_T(\cdot)$ at any third frequency will necessarily have a non-trivial correlation with one of the first two. Hence it is impossible to choose a sequence of frequencies so that the components are approximately uncorrelated. A consequence of this result is that any estimate based on solely non-negative weights will not be consistent because the correlation between $I_T(\lambda)$ at any three frequencies is bounded away from zero.

To see that this inherent dependence is enough to ruin τ_n, consider a sequence w_i consisting of large even multiples of π/T. We can show that

$$\nu_n \asymp \nu_0^{\frac{1}{2}}\mathbf{1}$$

$$\Sigma_n \asymp (1-\rho)I_n + \rho\mathbf{1_n}\mathbf{1_n}'$$

where from (3.8.3),

$$\rho = \frac{2(1-e^{-\alpha})^2}{(1+\alpha-e^{-\alpha})^2 + (1-e^{-\alpha})^2}$$

$$\nu_0 = \frac{(1+\alpha-e^{-\alpha})^2}{(1+\alpha-e^{-\alpha})^2 + (1-e^{-\alpha})^2}$$

$$\mathbf{1_n} = \{1\}_{n\times 1}$$

A little work shows that $\nu_n'\Sigma_n^{-1}\nu_n = n\nu_0/(1+(n-1)\rho)$ and so

$$\inf_a \frac{\mathbb{V}(\tau_n)}{\mathbb{E}^2(\tau_n)} = \frac{\rho}{\nu_0} + \frac{1-\rho}{n\nu_0}.$$

Hence the best weighted linear estimator based on frequencies that are multiples of π/T will not do better than ρ/ν_0, no matter the number of frequencies used.

3.8.2 *Determination of the local behavior by smoothing* $B_T(x)$

In the previous sub-section it was shown that θ could not be estimated using a natural linear weighted statistic of $I_T(\lambda)$ at large frequencies. This objective can be stated in terms of $B_T(x)$: Is it possible to determine θ exactly based on observing one realization of $\{B_T(x) : 0 \leq x \leq T\}$ by smoothing $B_T(x)$ at the origin?

Explicitly we will consider weighted versions of $B_T(x)$,

$$B_T^{(a)}(x) = a_T(x) B_T(x) \qquad 0 \leq x \leq T$$

where the weight function $a_T(x)$ smooths $B_T(x)$ at the origin so that estimates for θ of the form

$$S_T^{(a)}(x) = \frac{B_T^{(a)}(0) - B_T^{(a)}(x)}{x} \tag{3.8.4}$$

can be considered.

It is important to note that the class of smoothed empirical spectral density estimators is equivalent to the class of smoothed empirical covariance function estimators. Their relationship is expressed by

$$\begin{aligned}\Phi_T^A(\omega) =& \int_{-\infty}^{\infty} A_T(\omega - \lambda) I_T(\lambda) d\lambda \\ =& \mathcal{F}\{B_T^{(a)}(x)\}\end{aligned}$$

where $A_T(\lambda) = \mathcal{F}\{a_T(x)\}$ and $\mathcal{F}$ denotes the Fourier transformation. This relationship is seldom exploited in the literature.

We choose $a_T(x)$ to be continuous and non-negative definite so that $B_T^{(a)}(x)$ will have the same properties. In fact $a_T(x)$ is a lag-window from time-series literature. Suppose $a_T(x)$ has continuous first derivative and its second derivative is finite in some neighbourhood of the origin. Then we can write $a_T(x) = 1 + sx + \nu x^2 + o(x^2)$ as $x \to 0$.

The quality of $S_T^{(a)}(x)$ as an estimator for θ can be gauged by its mean-squared error. After some careful algebra based on the expression for the covariance in Result 3.5.1 we can show that

$$\mathrm{MSE}[S_T^{(a)}(x)] = \frac{\theta^2}{\alpha^4}\Big\{1 - 2\alpha s + \alpha^2(1 - 2s + 2\alpha s^2 + e^{-2\alpha}(1+s)^2)\Big\} + O(x)$$

as $x \to 0$. The value of s that minimizes the mean-squared error is

$$s = \frac{1 + \alpha + \alpha e^{-2\alpha}}{\alpha(2\alpha + e^{-2\alpha})}$$

corresponding to a mean-squared error we hesitate to write down. Figure 41 is a plot of this mean-squared error as a function of s when $\alpha = 1$. The natural unsmoothed estimator based on $a_T(x) \equiv 1$ has mean-squared error $\theta^2\{1 + \alpha^2(1 + e^{-2\alpha})\}/\alpha^4$.

In summary, the smooth estimators based on $B_T(x)$ have mean-squared error that is bounded above zero, independently of the smoothing function used.

It is interesting to note the existence of a closely related process motivated by (3.8.1),

$$\begin{aligned}\gamma_T(x) =& \frac{1}{2T}\int_0^{T-x}\{Z(t) - Z(t+x)\}^2 dt \qquad 0 \le x \le T\\ =& B_T(0) - B_T(x) - \frac{1}{2T}\int_0^x Z^2(t)dt - \frac{1}{2T}\int_{T-x}^T Z^2(t)dt\end{aligned}$$

The estimate $\gamma_T(x)/x$ appears superficially close to the form of $S_T^{(a)}(x)$ given in (3.8.4). We can show using some more careful algebra based on Result 3.5.1 that

$$\begin{aligned}\mathbb{E}(\gamma_T(x)/x) =& \frac{T-x}{T}\cdot\frac{R(0) - R(x)}{x}\\ =& \theta + O(x)\\ \mathbb{V}(\gamma_T(x)/x) =& \frac{4\theta^2 x}{3\alpha^2 T^3} + O(x^2)\end{aligned}$$

as $x \to 0$. Hence by taking x small enough we can approximate θ with arbitrary precision using $\gamma_T(x)/x$.

The rationale for the failure of the empirical covariance function can be seen from the plot of $\mathrm{Corr}(\ B_T(x_1), B_T(x_2)\ )$ given in Figure 42 where $T_c = 2T$. The correlation is bounded below by 0.85. It is straightforward to derive this lower bound from the exact expressions. It is always the limit of $\mathrm{Corr}(\ I_T(0), I_T(T-x)\ )$ as $x \to 0$. As T_c/T increases the bound increases.

In summary, on the basis of observing Z on $[0,T]$ it is well known that it is possible to estimate the local behavior of the random field arbitrarily well. However we find that we apparently can not determine the local behavior based on the empirical spectral density alone. Thus the empirical spectral density, or equivalently the empirical covariance, appears to throw out much of the information available in the data, in contrast to the increasing region asymptotic setting.

3.9 Asymptotic behaviour of $I_T(\omega)$ as $T \to \infty$

In this section we consider the asymptotic properties of the empirical spectral density as the region of observation grows, so that the information about $I_T(\lambda)$ is unbounded. Our objective is to provide a foundation to which our results for $T_c \asymp T$ can be compared. Many authors have described the behavior of $I_T(\omega)$ asymptotically as $T \to \infty$, usually for discrete–time processes. We re-prove the major results using approximation theory for singular integrals. Let C be the set of uniformly continuous functions that are bounded on $\mathrm{I\!R}$, and as usual let L_p denote the set of functions which are Lebesgue integrable to the pth power over $\mathrm{I\!R}$, where $1 \le p \le \infty$. C is endowed with the norm $\|f\|_C = \sup_{x\in\mathrm{I\!R}} |f(x)|$, and

$$\|f\|_p = \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |f(x)|^p dx \right\}^{1/p} \quad \text{for} \quad 1 \le p < \infty$$

while L_∞ has norm $\|f\|_\infty = \operatorname{ess\,sup}_{x\in\mathrm{I\!R}} |f(x)|$. The symbol X will be used to denote one of the spaces C or $L_p, 1 \le p < \infty$. Finally, if we let C_0 be $\{f \in C :$

$\lim_{|x|\to\infty} f(x) = 0\}$, then as $f(\omega)$ is the Fourier transform of $R(x) \in L_1$ we clearly have $R(x), f(\omega) \in L_1 \cap C_0$ and in particular $R(x), f(\omega) \in L_p$, $1 \le p \le \infty$.

Result 3.9.1:

a) $\|\mathbb{E}I_T(\omega) - f(\omega)\|_X = O(\frac{1}{T}) \iff \|f(\cdot + h) - f(\cdot)\|_X = O(h) \quad \text{as } h \to 0^+$

In particular, if $\int_{-\infty}^{\infty} |xR(x)|\, dx \;<\; \infty$ then,

$$\mathbb{E}I_T(\omega) = f(\omega) + O(\tfrac{1}{T})$$

where the $O(1/T)$ term is uniform with respect to ω, as $T \to \infty$.

b) $\|\mathbb{E}I_T(\omega) - f(\omega)\|_X = o(\frac{1}{T}) \Rightarrow \|\mathbb{E}I_T(\omega) - f(\omega)\|_X = 0 \quad \forall T$

That is, the rate of convergence of $I_T(\omega)$ to $f(\omega)$ is never faster than $O(\frac{1}{T})$.

c) If $\int_{-\infty}^{\infty} |xR(x)|\, dx \;<\; \infty$ then,

$$\mathbb{V}(I_T(\omega)) = f^2(\omega) + O(\tfrac{1}{T})$$

uniformly on any set of ω bounded away from zero, as $T \to \infty$.

d) If $\int_{-\infty}^{\infty} |xR(x)|\, dx \;<\; \infty$ then,

$$\mathbb{E}\{[\ I_T(\omega) - f(\omega)\ ]^2\} = f^2(\omega) + O(\tfrac{1}{T})$$

uniformly on any set of ω bounded away from zero, as $T \to \infty$.

e) If $\int_{-\infty}^{\infty} |xR(x)|\, dx \;<\; \infty$ then,

$\text{Cov}(\ I_T(\omega_1), I_T(\omega_2)\ ) = \Big\{h_T^2(\omega_1 - \omega_2) + h_T^2(\omega_1 + \omega_2)\Big\} \cdot f(\omega_1)f(\omega_2) \;+\; O(\frac{1}{T})$

for $\omega_1, \omega_2 > 0$. Note that $\text{Cov}(\ I_T(\omega_1), I_T(\omega_2)\ )$ is $O(\frac{1}{T^2})$ unless $\omega_1 = \omega_2$ and that the $O(\frac{1}{T})$ term is uniform on any set of ω_1, ω_2 bounded away from zero.

f) $I_T(\omega) \quad \overset{\mathcal{D}}{\to} \quad \frac{1}{2} f(\omega) \cdot \chi_2^2 \qquad \text{for } \omega > 0.$

Proof : In the terminology of approximation theory (Butzer & Nessel (1971), §3) the set of functions $\{\chi_T(x) : T > 0\}$ is called a kernel on $\mathbb{R}$ if $\chi_T(x) \in L_1$ for each T and $\int_{-\infty}^{\infty} \chi_T(x)\, dx \;=\; \sqrt{2\pi}$. A kernel is called an approximate identity on $\mathbb{R}$ if there is some constant $M > 0$ with $\|\chi_T(\cdot)\|_1 < M$ for $T > 0$ and $\int_{|x|\geq\delta} |\chi_T(x)|\, dx \to 0$ as $T \to \infty$ for each $\delta > 0$.

It is easy to show that the function

$$\mathcal{F}(x) = \frac{1}{\sqrt{2\pi}} h_T^2(x/T) \tag{3.9.1}$$

defines the approximate identity $\{T\mathcal{F}(Tx) : T > 0\}$ on $\mathbb{R}$. $\mathcal{F}(x)$ is usually referred to as the Fejér kernel. Further,

$$\begin{aligned} \mathbb{E}(I_T(\omega)) =& \frac{T}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathcal{F}(Tx) f(x - \lambda) dx \\ =& \sigma(f; \lambda; T), \end{aligned}$$

that is $\mathbb{E}(I_T(\omega))$, as a function of $f(\cdot)$, λ and T, is the singular integral of Fejér (Butzer & Nessel (1971), §3.1.2). We can then apply the results of Corollary 3.5.4, Proposition 12.4.1 and Proposition 13.2.5 to deduce a). Similarly b) follows from Proposition 12.4.1(a) and Proposition 13.2.1(a).

From (3.6.1), $\mathbb{V}(I_T(\omega)) = C_f^2(\omega, \omega) + C_f^2(\omega, -\omega)$. Using the definition,

$$C_f(\omega, -\omega) = \frac{T}{2\pi} \int_{-\infty}^{\infty} h_T(\lambda - \omega) \cdot h_T(\lambda + \omega)\ f(\lambda)\ d\lambda,$$

and Lemma 3.6.1 one can easily show that $C_f(\omega, -\omega) = O(\frac{1}{T})$ uniformly on any set of ω bounded away from zero. c) then follows from (3.6.1). As $\mathbb{E}\{[\ I_T(\omega) - f(\omega)\ ]^2\} = \{\mathbb{E}I_T(\omega) - f(\omega)\}^2 + \mathbb{V}(I_T(\omega))$, d) follows from a) and c). A direct derivation of e) is possible using the arguments in §6 of Jenkins & Watts (1968), or altering the discrete time arguments of Brillinger (1975), §5.2. Note also the comments of Yaglom (1987b), p86. As an alternative one can work with the kernels,

$$\chi_T(x) = \frac{T}{\sqrt{2\pi}} \frac{h_T(x - \omega/T) \cdot h_T(x + \omega/T)}{h_T(2\omega/T)} \qquad \text{for each fixed } \omega$$

From a), c) and Result 3.5.2, $\mathbb{V}(J_R(\omega)) = \frac{1}{2} f(\omega) + O(\frac{1}{T})$ and $\mathbb{V}(J_I(\omega)) = \frac{1}{2} f(\omega) + O(\frac{1}{T})$ so that $I_T(\omega) = J_R^2(\omega) + J_I^2(\omega)$ is asymptotically the sum of two independent $\frac{1}{2} f(\omega)\chi_1^2$ random variables, and f) follows.

Note that a) and b) are indicative that additional local smoothness conditions on $f(x)$, for example $f^{(r)}(x) \in Lip(\alpha)$, or equivalently global conditions on $B(x)$, for example $x^r B(x) \in L^1$, will not improve the order of approximation.

It is possible to produce most of these results directly from the proofs of the analogous statements in discrete time-series as given in, for example, Brillinger (1975), Theorems 5.2.2, 5.2.4 & 5.2.5 or Priestley (1981), §6.3.2. ∎

The implications of these results are usually stated as:

- $I_T(\omega)$ is asymptotically unbiased for $f(\omega)$.
- Asymptotically, $I_T(\omega)$ behaves like an independent process in the continuous parameter ω, because e) and f) say that, asymptotically, $I_T(\omega)$ forms an uncorrelated process with each member distributed as a multiple of a χ^2_2.
- From results d) and f), asymptotically, the random fluctuations of $I_T(\omega)$ about a mean near $f(\omega)$ have about the same magnitude as $f(\omega)$ itself, even for large values of T.

Hence $I_T(\omega)$ itself is a poor estimate for $f(\omega)$. The standard method of improving the performance of $I_T(\omega)$ is to exploit the approximate independence and use a smoothed version:

$$\widehat{f_T}(\omega) = \int_{-\infty}^{\infty} W_T(\omega - \lambda) \cdot I_T(\lambda)\, d\lambda$$

where $W_T(\omega)$ is a weight function emphasizing values of λ near zero. The asymptotic properties of such estimates have been extensively studied and it has been shown that they provide adequate estimates of $f(\omega)$ as $T \to \infty$. See for example the excellent review paper Dzhaparidze & Yaglom (1983), Bentkus & Rudzkis (1982) and the references therein.

As we have see in the previous sections, the situation for $T \asymp T_c$ is much less satisfactory.

3.9.1 *The behavior of $I_T(\lambda)$ for the Ornstein–Uhlenbeck process as $T \to \infty$*

In this section we compare the exact expressions (3.7.1) as $T \to \infty$ to the general results in the previous section. It is interesting to see that the Ornstein–Uhlenbeck process achieves the bounds in the results. We can write

$$\begin{aligned}\mathbb{E}I_T(\lambda) &= f(\lambda)\cdot\left\{1+\frac{\theta_2}{T}(1-\frac{2}{(1+\theta_2^2\lambda^2)})(1-e^{-T/\theta_2}\cos(T\lambda))-\frac{2\theta_2^2e^{-T/\theta_2}\lambda\sin(T\lambda)}{T(1+\theta_2^2\lambda^2)}\right\}\\ &= f(\lambda)\cdot\left\{1+\frac{\theta_2}{T}(1-\frac{2}{(1+\theta_2^2\lambda^2)})\right\}+O(e^{-T/\theta_2})\end{aligned}$$

as $T \to \infty$. This verifies Result 3.9.1 a). Now the variance,

$$\mathbb{V}(I_T(\lambda)) = f^2(\lambda)\cdot\left\{1+\frac{2\theta_2}{T}(1-\frac{2}{(1+\theta_2^2\lambda^2)})\right\}+O(\frac{1}{T^2})$$

as $T \to \infty$. This verifies c) and d). We can also easily verify b) and e) based on the Result 3.8.1.

3.10 Summary and conclusions

In this chapter we consider inference for the covariance structure of a stochastic process on the basis of observing the process continuously on a finite segment of a single realization. The objective is to approximate the inference based on discrete observation on the same segment. It is shown that under the condition that the range of correlation is comparable to the length of the segment, in a sense made precise, the approximation is appropriate. In the early sections maximum likelihood estimation for the Ornstein–Uhlenbeck process is used to illustrate the point. In §3.5 through §3.6 the empirical spectral density process is considered and exact expressions for the covariance structure are given. In §3.7 an explicit expression is given for the Ornstein–Uhlenbeck process.

In the §3.8 the focus is on the estimation of the local behavior of the random field based on the empirical spectral density alone. It is shown that the local behavior apparently can not be recovered on the basis of either the empirical spectral density or the empirical covariance function alone.

In §3.9 we derive some results describing the behavior of the empirical spectral density process as $T \to \infty$. This behavior is contrasted to the behavior when $T \asymp T_c$.

CHAPTER 4

A BAYESIAN ANALYSIS OF KRIGING

4.1 Introduction

In this chapter we will view the kriging procedure for the prediction of Gaussian random fields within the Bayesian framework. The objective is to monitor the performance of kriging when the underlying model is misspecified. Particular attention is paid to the treatment of parameters in the covariance structure and their effect on the quality, both actual and perceived, of the prediction.

The use of the notion of subjective probability can be viewed as a convenient formal platform for inference. Indeed, if the random field arises as a manifestation of the statistician's uncertainty, this approach is most natural.

Bayesian analyses of kriging procedures are relatively new. Except for the work of Omre (1987), Omre & Halvarsen (1989) and Woodbury (1989) there appears to be no work from within the Geostatistical community using the Bayesian perspective. Omre & Halvarsen (1989) describe a Bayesian approach to predicting the depth of geologic horizons based on seismic reflection times. They utilize prior information about the mean function to bridge the gap between simple kriging, that is assuming the mean function is known, and universal kriging, that is assuming only that the mean is of the regression type.[1] Of course, the situation is a direct extension of standard Bayesian work in linear models where, for example, Box & Tiao (1973) §2.7 and

[1] See §1.2

Zellner (1971) §7 are textbook references. Both sketch results using natural vague prior distributions for the parameters. Much of the work in Bayesian Time–series focuses on the estimation of the parameters of particular ARMA models. Zellner (1971) §5 derives the predictive distribution of a future observation from a not necessarily stationary AR(1) process. Broemeling (1985) gives a discussion of standard regression and mixed models. His §5 extends Zellner's (1971) work by using the proper conjugate prior distributions for autoregressive series. He does not extend the work on prediction. All these analyses assume that the ARMA orders are known. Lahiff (1984) considers prediction and estimation for the AR(1) model with known mean and variance. Her §1.3 provides a convenient summary of previous work. Also of interest is the work of Harrison & Stevens (1976) and West & Harrison (1986) on forecasting using dynamic fully Bayesian models. The last few years have seen an explosion of work in Bayesian time–series, especially on state space approaches using the Kalman filter. For a summary consider Geol & Zellner (1986). What is novel about this chapter is the general treatment of parameters in the covariance structure other than location and scale.

A related non–Bayesian approach to prediction is that of Hinkley (1979) and Butler (1986) based on "predictive" likelihoods. Their approach looks at a component of the full likelihood that can be used to infer a value to be predicted without prior knowledge of the parameters. The inference tends to be close to that produced using Jeffreys's non-informative prior for the parameters.

Note that the underlying kriging procedure is motivated by sampling considerations, producing point predictions and associated measures of uncertainty for those predictions both based on sampling distributions. However, kriging, when the mean is of known regression form, can be given a Bayesian interpretation. This will be commented on in the next section.

4.2 Traditional kriging: covariance parameters assumed known

The notation defined here will be used throughout the chapter and is consistent with the notation in §1.2. Suppose $Z(x)$ is a real–valued stationary Gaussian random field on R with mean

$$\mathbb{E}Z(x) = \beta' f(x), \tag{4.2.1}$$

where $f(x) = (f_1(x), \ldots, f_q(x))'$ is a known vector function, β is a vector of unknown regression coefficients, and covariance function

$$\mathrm{Cov}(Z(x), Z(x')) = \alpha K_\theta(x, x') \qquad \text{for } x, x' \in R$$

where $\alpha > 0$ is a scale parameter, $\theta \in \Theta$ is a $q\times1$ vector of structural parameters and Θ is an open set in $\mathbb{R}^p$. The division is purely formal as θ may also determine aspects of scale. We observe, from a single realization, $\{Z(x_1), \ldots, Z(x_n)\} = Z'$ and, as usual, will focus on the prediction of $Z(x_0)$. As defined in §1.2, the kriging predictor is the best linear unbiased predictor (BLUP) of the form $\widehat{Z}_\theta(x_0) = \sum_{i=1}^N \lambda_i(\theta) Z(x_i)$, that is, the unbiased linear combination of the observations that minimizes the variance of the prediction error. It is straightforward to show that the corresponding weight vector $\lambda(\theta)$ defining $\widehat{Z}_\theta(x_0)$ is given by

$$\lambda(\theta) = {K_\theta}^{-1} k_0 + {K_\theta}^{-1} F (F' {K_\theta}^{-1} F)^{-1} b_\theta, \tag{4.2.2}$$

where

$$\begin{aligned} F &= \{f_j(x_i)\}_{n\times q}, \\ k_0 &= \{K_\theta(x_0, x_i)\}_{n\times 1}, \\ K_\theta &= \{K_\theta(x_i, x_j)\}_{n\times n}, \\ b_\theta &= f(x_0) - F' {K_\theta}^{-1} k_0. \end{aligned}$$

The quality of the prediction is determined by the distribution of the prediction error, $e(x_0) = Z(x_0) - \widehat{Z}_\theta(x_0)$. Note that the prediction error does not depend on α

or β. Under our Gaussian model, the conditional distribution of $e(x_0)$ given α, β, θ and Z is

$$P(e(x_0) \mid \alpha, \beta, \theta, Z) \sim N(\ b'_\theta(\beta - \hat{\beta}(\theta)),\ \alpha\{K_\theta(x_0, x_0) - k'_\theta K_\theta^{-1} k_\theta\ \}\ )$$

where $\hat{\beta}(\theta) = (F' K_\theta{}^{-1} F)^{-1} F' K_\theta{}^{-1} Z$ and $N(\cdot, \cdot)$ denotes the Gaussian distribution. The usual sampling distribution for $e(x_0)$ follows if we condition on α, θ and Z alone:

$$P(e(x_0) \mid \alpha, \theta, Z) \sim N(\ 0,\ \alpha V_\theta\ ) \tag{4.2.3}$$

where $V_\theta = K(x_0, x_0) - k'_\theta K_\theta^{-1} k_\theta + b'_\theta (F' K_\theta{}^{-1} F)^{-1} b_\theta$ and αV_θ is the usual prediction variance.

If β were known we would do better to use the conditional expectation of $Z(x_0)$ given Z and under our model

$$\mathbb{E}(Z(x_0) \mid \alpha, \beta, \theta, Z) = k'_\theta K_\theta^{-1} Z + b'_\theta \beta$$

with prediction error distribution

$$P(Z(x_0) - \mathbb{E}(Z(x_0) \mid \alpha, \beta, \theta, Z) \mid \alpha, \beta, \theta, Z) \sim N(\ 0,\ \alpha\{K_\theta(x_0, x_0) - k'_\theta K_\theta^{-1} k_\theta\ \}\ ) \tag{4.2.4}$$

This is also the approach taken by simple kriging. From this perspective, the loss due to uncertainty about β is the conditional bias, determined by the deviation of $\hat{\beta}(\theta)$ from β. The optimal predictor corrects for this bias by subtracting $b'_\theta(\beta - \hat{\beta}(\theta))$ from the kriging predictor, a term clearly dependent on β.

The log–likelihood of θ, α and β having observed Z is, up to an additive constant,

$$\begin{aligned} L(\alpha, \beta, \theta;\ Z) &= -\frac{n}{2}\ln(\alpha) - \frac{1}{2}\ln(|K_\theta|) - \frac{1}{2\alpha}(Z - F\beta)' K_\theta^{-1} (Z - F\beta) \\ &= -\frac{n}{2}\ln(\alpha) - \frac{1}{2}\ln(|K_\theta|) - \frac{1}{2\alpha}\{\nu S_\theta^2 + (\beta - \hat{\beta}(\theta))'(F' K_\theta{}^{-1} F)(\beta - \hat{\beta}(\theta)\} \end{aligned} \tag{4.2.5}$$

where $S_\theta^2 = \frac{1}{\nu}(Z - F\hat{\beta}(\theta))' K_\theta^{-1}(Z - F\hat{\beta}(\theta))$, $\nu = n - q$ and the dependencies upon n have been suppressed. For fixed θ, this is maximized over β by $\widehat{\beta}(\theta)$, the generalized least squares estimator and the profile log–likelihood

$$L_{pm}(\alpha, \theta;\ Z) \equiv L(\alpha, \theta, \widehat{\beta}(\theta);\ Z)$$

is maximized by $\widehat{\theta}$ if and only if $\widehat{\theta}$ maximizes (4.2.5), and in this case the maximum likelihood estimate of β is $\widehat{\beta}(\widehat{\theta})$. For fixed θ, S_θ^2 is the natural estimate of α and ν represents the degrees of freedom.

Traditionally it is assumed that the covariance function is known exactly and the investigator has little knowledge about β prior to analyzing the data. The underlying kriging approach usually presumes ignorance about β and the unrelatedness of β to the behavior of the covariance function. This philosophy will be followed throughout the chapter. Under these assumptions, the likelihood is data–translated and an appropriate prior distribution has $P(\beta \mid \alpha, \theta)$ locally uniform. Using (4.2.5) the posterior distribution of β is

$$P(\beta \mid \alpha, \theta, Z) \sim N_q(\ \widehat{\beta}(\theta),\ \alpha(F' K_\theta{}^{-1} F)^{-1}) \tag{4.2.6}$$

The posterior distribution of the prediction error is then

$$P(e(x_0) \mid \alpha, \theta, Z) \propto \int_\beta P(e(x_0) \mid \alpha, \beta, \theta, Z) P(\beta \mid \alpha, \theta, Z) d\beta$$

that is, by direct calculation,

$$P(e(x_0) \mid \alpha, \theta, Z) \sim N(\ 0,\ \alpha V_\theta\ ) \tag{4.2.7}$$

the same as the sampling distribution (4.2.3). This distribution forms the basis for all inferential statements about the prediction error. Hence, except for the usual differences in interpretation, we end up with the same analysis as the traditional approach. This comparison may be loosely stated as: ordinary kriging is 'Bayesian' with the non-informative prior for the mean parameter. We note that even if (α, θ) are known there is still vestigial uncertainty about $Z(x_0)$.

In practice, one does not know α and θ and estimates them by either likelihood methods or various *ad hoc* methods. Usually the predictor and the prediction error are themselves estimated by "plugging–in" the estimates into (4.2.2) and (4.2.3). The statistical properties of such estimates are considered in §4.6 and in §3.7.

4.3 Kriging when the scale parameter is unknown

In the section the assumption that the covariance function is known exactly is relaxed slightly by assuming that the scale factor α is unknown, as well as the regression parameter β. As β is a location parameter we expect that our prior opinions about β bear no relationship to those about α and *a priori* might expect α and β to be independent, leading to the use of Jeffreys's prior,

$$P(\alpha,\ \beta \mid \theta) \propto 1/\alpha$$

Their joint posterior distribution is then

$$P(\alpha,\ \beta \mid \theta, Z) \propto$$

$$\exp\left\{ -\frac{n+1}{2}\ \ln\alpha - \frac{1}{2}\ln(|K_\theta|) - \frac{1}{2\alpha}\{\nu S_\theta^2 + (\beta - \hat{\beta}(\theta))'(F'K_\theta{}^{-1}F)(\beta - \hat{\beta}(\theta))\}\right\} \tag{4.3.1}$$

This can be factored as

$$P(\alpha,\ \beta \mid \theta, Z) \propto P(\beta \mid \alpha,\ \beta, Z) \cdot P(\alpha \mid \theta, Z)$$

where from (4.3.1), $P(\alpha \mid \theta, Z)$ is $\nu S_\theta^2 \chi_\nu^{-2}$, a scaled inverted chi-squared distribution and,

$$P(\beta \mid \alpha, \theta, Z) \sim N_q(\ \widehat{\beta}(\theta),\ \alpha(F'K_\theta{}^{-1}F)^{-1}) \tag{4.3.2}$$

from (4.2.6). The posterior distribution of the prediction error is then

$$P(e(x_0) \mid \theta, Z) \propto \int_0^\infty P(e(x_0) \mid \alpha, \theta, Z) P(\alpha \mid \theta, Z) d\alpha,$$

which is well known to be a univariate t distribution on ν degrees of freedom. That is,

$$P(e(x_0) \mid \theta, Z) \sim t_1(\ 0,\ S_\theta^2 V_\theta; \nu) \tag{4.3.3}$$

Comparing (4.3.3) with (4.2.7) we see that the ignorance about the scale parameter α expresses it self as the difference between a t distribution on ν degrees of freedom using the natural estimate of αV_θ , $S_\theta^2 V_\theta$ and a Gaussian distribution with variance αV_θ . The ratio of variances is approximately $\frac{\nu}{\nu-2}$, a small difference if ν is moderately large. This indicates that ignorance about the scale parameter alone will not change the posterior distribution of the prediction error very much.

4.4 Real kriging: covariance parameters are unknown

In this section we allow the covariance function to be unknown, but still a member of the parametric class Θ . This situation is much more realistic than both the traditional and scale cases dealt with above. If θ is known so that only the location parameter β and the scale parameter α are uncertain then we are in a standard regression setting. The distinction between the regression setting and the spatial random field setting is the uncertainty in the structural parameter θ. While the restriction to a parametric class is a significant assumption, it still allows great latitude. In general it will no longer be possible to write down meaningful closed form expressions for the posterior distributions, and the use of numerical methods is required.

Consider the covariance function on $\mathbb{R}$,

$$K_E(|x-y|; \alpha, \theta) = \alpha\theta e^{-|x-y|/\theta}$$

where $\alpha > 0$, $\theta > 0$, $x,\ y \in \mathbb{R}$. This parameterization is chosen to emphasize that we wish to estimate the behavior at the origin well. The "slope", α, is the slope at the origin of the function, and controls the smoothness of the implied random field. The "range", θ, changes the rate of decrease of the correlation with distance. The variance of the random field is $\alpha\theta$. For this example one might *a priori* expect θ to be independent of α and β. In general, it is helpful to choose the parameterization so that it is appropriate to assume independence between α and θ. Additional comments on the choice of distributions prior to the data are made in §4.8.

If we had specific information on the form of K_θ then we could determine Jeffreys's prior based on I. The emphasis here is on situations where the information for each component of θ is not necessarily increasing to infinity, so that the use of Jeffreys's prior is less appropriate. Partly for convenience, the form of the prior used here will be

$$P(\alpha,\ \beta, \theta) \propto P(\theta)/\alpha$$

Using (4.2.5) the marginal posterior distribution of θ can be shown to be

$$P(\theta \mid Z) \propto P(\theta) \cdot |K_\theta|^{-\frac{1}{2}} |F' K_\theta{}^{-1} F|^{-\frac{1}{2}} (\nu S_\theta^2)^{-\nu/2} \tag{4.4.1}$$

Note that $P(\theta \mid \beta = \hat{\beta},\ Z) \propto P(\theta) \cdot |K_\theta|^{-\frac{1}{2}} (\nu S_\theta^2)^{-n/2}$. Note also from (4.3.2) that $P(\beta \mid \theta, Z)$ is the appropriate multivariate t distribution while $P(\beta \mid Z)$ will in general not have a simple form.

The posterior distribution of the prediction error is

$$P(e(x_0) \mid Z) \propto \int_\Theta P(e(x_0) \mid \theta, Z) \cdot P(\theta \mid Z) d\theta \tag{4.4.2}$$

where the integrand is given by (4.3.3) and (4.4.1). As the dependence of K_θ on θ is not specified this expression can not be further simplified and further exploration will in general require numerical computation. If prior information is available it can be directly incorporated into (4.4.1), although additional numerical integration maybe necessary if prior dependencies among (α, β, θ) are envisaged.

4.5 Making multiple predictions

Within this paradigm the extension to multiple predictions is straightforward. Suppose we wish to predict at $y_1, \ldots, y_m \in R$. Let $Z_0 = \{Z(y_1), \ldots, Z(y_m)\}'$ and $e_0 = \{e(y_1), \ldots, e(y_m)\}'$ then

$$\begin{pmatrix} Z \\ - \\ Z_0 \end{pmatrix} \sim N_{n+m}\Big[\begin{pmatrix} F\beta \\ - \\ \tilde{F}\beta \end{pmatrix}, \alpha \begin{pmatrix} K_\theta & | & H_\theta \\ - & \cdot & - \\ H'_\theta & | & J_\theta \end{pmatrix}\Big]$$

Based on the distribution it is possible to derive the following distributions:

$$\begin{aligned} P(e_0 \mid \alpha, \theta, Z) &\sim N_m\Big(\ 0,\ \alpha\{J_\theta - H'_\theta {K_\theta}^{-1} H_\theta + B'_\theta (F' {K_\theta}^{-1} F)^{-1} B_\theta\}\Big) \\ P(e_0 \mid \theta, Z) &\sim t_m\Big(\ 0,\ S^2_\theta\{J_\theta - H'_\theta {K_\theta}^{-1} H_\theta + B'_\theta (F' {K_\theta}^{-1} F)^{-1} B_\theta\}\Big) \\ P(e_0 \mid Z) &= \int_\Theta P(e_0) \mid \theta, Z) P(\theta \mid Z) d\theta \end{aligned}$$

where $B_\theta = \tilde{F}' - H'_\theta {K_\theta}^{-1} H_\theta$

For the first two distributions, both the marginal and conditional posterior distributions of subsets are multivariate Gaussian or t with the appropriate covariance matrices. This behavior is extremely useful in practice where the prediction locations are usually many and of a systematic nature. If, for example, we construct a predictor of the form WZ_0 for, say, a set of m area means, where W is an $m \times p$ matrix $(m \leq p)$, then the posterior distribution of WZ_0 is again multivariate Gaussian (or t) with the natural location and scale parameters. An additional practical advantage is the direct availability of posterior probability regions. No such convenient analytic results exist for the case where the structural covariance parameters are unknown.

4.6 Actual performance of plug–in estimates

The results given in §4.2 are with respect to the perceived distributions; that is the distributions assuming the plug-in model is the true model. Of greater relevance is the posterior distributions of the predictor from the plug-in model under the Bayesian model.

Suppose we assume that the parameters of the covariance structure are $(\tilde{\alpha}, \tilde{\theta})$. These may be arrived at by any procedure, although the usual methods are maximum likelihood, weighted least squares or derived from empirical correlation functions. The perceived posterior distribution of the prediction error, $\tilde{e}(x_0) = Z(x_0) - \widehat{Z}_{\tilde{\theta}}(x_0)$, is then

$$P(\tilde{e}(x_0) \mid \tilde{\alpha}, \tilde{\theta}, Z) \sim N(\ 0,\ \tilde{\alpha} V_{\tilde{\theta}}\ ) \tag{4.6.1}$$

By perceived posterior distribution we mean the distribution that an investigator would use as a basis for inference if she had a degenerate prior for (α, θ) at $(\tilde{\alpha}, \tilde{\theta})$. Now, for a given θ, $\tilde{e}(x_0)$ is independent of $\tilde{\alpha}$ so that,

$$P(\tilde{e}(x_0) \mid \theta, Z) \sim t_1(\ \widehat{Z}_\theta(x_0) - \widehat{Z}_{\tilde{\theta}}(x_0),\ S^2_\theta V_{\tilde{\theta}};\ \nu\ ) \tag{4.6.2}$$

Observe that the specification of $\tilde{\alpha}$ acts only as a multiplier on the variance of the *perceived* distribution, and that the actual distribution of the predictor is independent of this choice. By actual posterior distribution we mean the posterior distribution of the quantity $Z(x_0) - \widehat{Z}_{\tilde{\theta}}(x_0)$ given Z and with respect to the full posterior for (α, β, θ). We note that $\widehat{Z}_{\tilde{\theta}}(x_0)$ is a constant. Under the assumption that the covariance class is correctly specified, this posterior provides a basis for valid inference for the plug-in prediction error $\tilde{e}(x_0)$.

Thus the actual posterior distribution is

$$P(\tilde{e}(x_0) \mid Z) = \int_\Theta P(\tilde{e}(x_0) \mid \theta, Z) P(\theta \mid Z) d\theta \tag{4.6.3}$$

where $P(\theta \mid Z)$ is the full posterior for θ and given is in (4.4.1). This expression is a clear expression of the effect of the plug–in estimates. The uncertainty in α manifests itself in the conversion from a Gaussian to a t distribution. The uncertainty in θ manifests itself through the weighting of each of these t distributions by the posterior for θ. Depending on the influence of θ on the spread and location of the t distribution, the actual posterior might be wider or narrower than the perceived posterior. Misspecification leads also to bias. Typically the actual distribution will have no simple analytic form and must be determined numerically.

It should be noted that in practice the actual posterior would not be used as a basis for inference. The Bayesian approach would be to use the complete posterior of the prediction error. The traditional kriging approach would be to use the perceived posterior of the prediction error. The actual posterior of the prediction error is used to evaluate the adequacy of the perceived posterior for inference when the plug-in predictor is used. We would like the perceived posterior to be closer to the actual posterior than to the complete posterior. The difference between the perceived and complete posteriors represents the difference in inference between the traditional kriging approach and the fully Bayesian approach. For this reason it is inappropriate to compare the actual posterior to the complete posterior.

4.7 Application to two dimensional random fields on a grid

In this section the theory of the previous sections is applied to kriging where the continuous field is observed on a planar grid set over the unit square. Three covariance classes are investigated, the Squared Exponential, Spherical and Matérn. The first two are chosen because they have been shown to exhibit unusual behavior both in terms of prediction and the estimation of the covariance structure. The Bayesian paradigm provides an additional perspective on this behavior. The rich Matérn class is introduced for comparison and as a surrogate for the Spherical class when the underlying field is a member of the Spherical class.

For focus the fields in this section are mean-zero, isotropic and Gaussian. In our examples 36 observations will be taken on the 6×6 grid on the unit square. Larger 8×8 grids were also used, providing results substantially similar to those reported.

4.7.1 *The Spherical covariance structure*

The Spherical covariance class has been consider in §2.6. It is commonly used for geological and hydrological applications in $\mathbb{R}^2$. The isotropic covariance has the general form:

$$K_\theta(x) = \begin{cases} \frac{2}{3} - \frac{|x|}{\theta} + \frac{1}{3}\{\frac{|x|}{\theta}\}^3 & \text{if } |x| < \theta \\ 0 & \text{if } |x| \geq \theta \end{cases}$$

where θ is a range parameter defining the limit of direct correlation. It corresponds to a field that is mean-square continuous, but not mean–square differentiable. As we have seen in §2.6 the corresponding likelihood surface exists and has a continuous derivative. However, the second derivatives of the likelihood are discontinuous leading to multiple modes.

Parameter values were chosen to be both realistic and interesting. The range was set to $\theta = 0.6$ and α, proportional to the variance, was set to 1.5. If the range is set below $\frac{1}{5}$ or above 1 multiple modes will not occur. Realizations were generated from this model and the profile likelihood surfaces over (α, θ) were constructed. If the likelihood surfaces exhibit multiple modes then the question arises about how both estimation and prediction should be done.

Consider first unimodal likelihoods, so that $\widehat{\theta}$, the maximum likelihood estimate for θ can be defined by standard numerical techniques.

Traditional kriging uses $\widehat{\theta}$ as a surrogate for θ and expresses the uncertainty in the prediction by the perceived posterior error distribution $(4.6.1)$ with $(\tilde{\alpha}, \tilde{\theta}) =$

$(\widehat{\alpha}, \widehat{\theta})$. Of course any alternative plug-in estimation procedure could be used. The actual posterior predictive distribution of this predictor is given by (4.6.3) and the complete posterior predictive distribution is given by (4.4.2).

An example is given in Figure 43. The location the field is predicted at was arbitrarily chosen within the central grid square to be $(0.53, 0.58)$. The maximum likelihood estimate is $(\widehat{\alpha}, \widehat{\theta}) = (1.32, 0.5)$. The above three posterior densities are represented. As can be seen they are very similar in shape with a standard deviation of about 0.6. The perceived posterior is always a centered Gaussian while the actual and complete posteriors are mixtures of non-central and central t-distributions respectively. The complete posterior reflects the full uncertainty in the covariance structure and will be regarded here as the correct reference for inference. It is always symmetric about zero. The perceived posterior is based on an incorrect model, and can be wider or narrower than the complete posterior depending on the plug-in estimates used. In general it tends to underestimate the uncertainty by a small amount. The actual posterior of the plug-in predictor is in general symmetric about a non-centered value, indicating that the plug-in predictor is slightly biased and slightly underestimates the uncertainty. Figure 43 is a good representation of the absolute shapes of the posteriors, but does allow accurate relative comparisons of the densities. Figure 44 represents the ratios of the perceived and actual posteriors to the complete posterior. The vertical axis has a logarithmic scale. Notice that the 99% probability interval has a width of about 3 units, and that values outside of 2 have negligible weight. Over this interval the ratio is close to one for both posteriors. In the tails the perceived posterior thins rapidly. Probability regions based on these posteriors will be very similar.

Another realization of this random field is given in Figure 45. This was chosen as the worst case of 10. The actual posterior indicates that the plug-in predictor has a bias of about 0.2 units and is slightly thicker in the tails.

Consider now realizations that have likelihoods with multiple modes, so that the determination of the maxima is problematical. An example of such a likelihood is given in Figure 46. The natural slope parameterization used is that of §2.6 so that Slope is α/θ. Note that the global maximum likelihood estimate of $(2.75, 0.5)$ corresponding to $(\widehat{\alpha}, \widehat{\theta}) = (1.44, 0.5)$ is close to the generating value $(1.5, 0.6)$, while a substantial local maximum exists with an inflated range. This behavior in the likelihood is typical. There can be 3 or more local maxima each with a larger range and similar slopes. A search routine starting with a long range might converge to the local maxima instead of the global maxima. Hence the estimate of slope is about right, while the range could easily be far from the global maxima.

The behavior of the kriging predictor based on the global maxima will be similar to that of the unimodal likelihood discussed above. We will focus on the local maxima at $(2.75, 1)$ corresponding to $(\widehat{\alpha}, \widehat{\theta}) = (2.75, 1)$. Figure 47 presents the perceived, actual and complete posteriors. We see that the plug-in predictor does not have appreciable bias, but the perceived posterior under-represents its uncertainty. This can be seen better in Figure 48 representing the relative comparisons. We see that this choice of the range parameter does not hurt performance very much. As an extreme example consider Figure 49 reporting the posteriors when the range is set to 3 corresponding to $(\widehat{\alpha}, \widehat{\theta}) = (8.28, 3)$. The perceived and actual posteriors are almost identical to those based on a range of 1. Hence even if an extreme mode exists it will not alter the inference very much.

In summary, the actual and perceived performance of the plug-in kriging predictor is insensitive to the specification of the range as long as the slope parameter is reasonable. Poor specification of the slope will have little effect on the actual posterior, but will have a multiplicative effect on the variance of the perceived posterior.

This is nice because it is the slope parameter that can be determined well while the range parameter is difficult to determine.

4.7.2 *The Squared Exponential covariance structure*

In this section the analysis of the previous section is repeated using random fields based on the Squared Exponential covariance class. This class is often used in applications although the rationale for it is weak. The isotropic covariance has the general form:

$$K_\theta(x) = e^{-(x/\theta)^2}$$

where θ is a range parameter. Unfortunately this class has also acquired the designation "Gaussian covariance class", which is neither historically correct nor appropriate. It corresponds to a field with analytic realizations, a very severe restriction. The corresponding likelihood surface exists and has a continuous second derivative.

The random field used in the examples has variance $\alpha = 0.75$ and range $\theta = 0.3$ in line with the values chosen for the Spherical example. Realizations were generated from this model and the profile likelihood surfaces over (α, θ) were constructed. All the likelihood surfaces observed were unimodal and the maximum likelihood estimate was found by a numerical search routine. The location to be predicted at was arbitrarily perturbed to $(0.44, 0.43)$. Posteriors for a typical realization are given in Figure 50. The maximum likelihood estimate was $(\widehat{\alpha}, \widehat{\theta}) = (0.62, 0.31)$. Their shapes are similar with the perceived having much narrower tails than the complete posterior. In general the perceived posterior varies and often has tails longer than the complete posterior. The actual posterior of the predictor indicates that the plug-in predictor based on the maximum likelihood estimates has a small bias and the perceived posterior is too thin in the tails. Figure 51 represents a relative comparison for the same realization.

Another realization of this random field is given in Figure 52. This was chosen as the worst case of 10 . The actual posterior indicates that the plug-in predictor has a bias of 0.02 units, while the perceived is again much thinner in the tails.

In general the perceived posterior from the Squared Exponential covariance class is close to the complete posterior, while the actual posterior indicates that the predictor is slightly biased. The predictions themselves tend to be much sharper than for the Spherical model because this model assumes that the realizations are very smooth.

4.7.3 *The effect of misspecifying the Spherical class by the Matérn class*

Suppose we are given data on a grid from a random field with a covariance from the Spherical class. The Spherical model is atypical and unless we had good reason to believe that the data came from this class we might try a model from an omnibus class such as the Matérn class of §1.5.4. We would then proceed to krige based on this incorrect model. What is the effect of misspecifying the Spherical class by the Matérn or the Squared Exponential class?

Under the alternative class we will consider the plug-in predictor using parameters estimated using maximum likelihood. We can then construct the perceived posterior distribution of the prediction error under the alternative model and the posterior distribution under the correct Spherical model. We will compare these distributions to the posteriors using the correct Spherical model for both estimation and prediction.

Figure 53 is a representative example of a realization from the same field as used above. The perceived, actual and complete posteriors under the Spherical model are reported. Each represents different stages of enlightenment about the underlying field. The perceived and complete posteriors under the Matérn model have been added. We see that the complete posterior under the Matérn model is close to the complete model under the Spherical model. That is, the effect of misspecification is small if a fully Bayesian procedure is followed.

The perceived posterior based on the Matérn maximum likelihood estimates plugged in to the kriging equation is substantially narrower than the complete posterior, and in particular provides much worse inference than the perceived posterior under the Spherical model. In this example the perceived posterior leads to a non-conservative inference, while in others the perceived posterior leads to a conservative inference. The marginal posterior distribution for the smoothness parameter of the Matérn covariance is extremely right skewed, indicating the smoothness parameter itself is not well determined from the 36 values. Given that the generating Spherical field is not differentiable we might expect that the bulk of the posterior weight for the smoothness parameter on values less than two while in practice values greater than two receive non-negligible weight. As the number of values in the field increases this deviation decreases slowly. The actual performance of the Matérn plug-in predictor is similar to the actual performance of the Spherical plug-in predictor, although it indicates that the Matérn predictor has a bias of about 0.2 units.

If the maximum likelihood estimate of the smoothness parameter of the Matérn model is large, then the quality of prediction can be quite bad. An example is given in Figure 53, where the maximum likelihood estimate of the smoothness parameter is 5.05. The predictor is optimal under a model with four times differentiable realizations while the data is generated from a model with non-differentiable realizations. Hence the perceived posterior under this smooth model tends to be narrower than the posterior under the correct Spherical model. The complete posterior makes an adjustment towards the correct posterior, but is still inadequate.

The Matérn class covers a wide range of behaviors, and so we might expect it will do well in approximating the Spherical. An alternative model that can be tested is the squared Exponential. The above procedure can be repeated with the Squared Exponential class replacing the Matérn class. The posteriors are given in Figure 54. The same realization from the Spherical random field is analyzed. The perceived posterior is even more non-conservative than for the Matérn class. In addition the complete posterior under the Squared Exponential model is very non-conservative, so that under the Squared Exponential model we believe the predictions to be much better than they really are. This is a general pattern where the perceived posterior is extremely non-conservative and the complete posterior only goes part way toward the complete posterior under the Spherical model. The reason would seem to be that the Squared Exponential model can only view predictors as being very precise. The Matérn class can view predictors as having a wide range of precision. The particular estimated Matérn predictor is viewed as very narrow because it is optimal under a very smooth model. However the complete posterior under the Matérn model takes into account a wide range of smoothness and is usually able to compensate.

4.8 Issues in the choice of distributions prior to the data

The expression of prior knowledge about the covariance structures is a fundamental issue that requires a balance between generality and practicality. In this section we make some guidelines on how prior distributions should be chosen for spatial random fields. The covariance structures are almost exclusively viewed from within a parametric framework.

If the investigator has real prior knowledge about the parameters individually and their relationships then the issue is the expression of that knowledge in terms of distributions. Given a joint prior distribution for (α, β, θ) one would use (4.2.5) and (4.2.4) to derive the posterior distributions in the same fashion as in §4.4.

It is difficult to incorporate geological knowledge directly into kriging studies as this knowledge usually takes the form of location of fault lines, changes in the composition of rock or consistency of mineralization. The parameters in the covariance model should correspond to easily interpretable quantities. For example one way to incorporate the location of fault lines and rock composition measurements is to use regression covariates and have prior distributions for their coefficients. Similarly the extent of dependence (range of the covariance model), existence of measurement error and micro-structures (size of the nugget effect term) and smoothness of the random field (degree of differentiability implied by the covariance function) can all be included. This incorporation is very situation dependent.

For spatial random fields it is usual to regard the mean parameter β independently of the covariance parameters α and θ. In the spirit of kriging the usual marginal prior for β can be taken as non-informative, which almost always means uniform and improper. If the covariance structure has θ degenerate then one could reasonably express *a prior* knowledge about (β, α) via the usual Gaussian-Gamma conjugate prior from generalized least squares.

Fisher's Information matrix for these parameters is $I = \operatorname{diag}\{I_\beta,\ I_{\alpha\theta}\}$ where

$$I_\beta = \frac{1}{\alpha} F' {K_\theta}^{-1} F$$

$$I_{\alpha\theta} = \frac{1}{2\alpha^2} \begin{pmatrix} n & \operatorname{tr}(K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_j}) \\ \operatorname{tr}(K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_i}) & \operatorname{tr}(K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_i} K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_j}) \end{pmatrix}$$

In the general case one should choose a parameterization in which α and θ can *a priori* be regarded as approximately independent, so that one could use the conjugate prior for (β, α) and a marginal prior for θ dependent on the particular covariance structure. Consider, for example, the Matérn class where θ is the parameter controlling the smoothness of the random field. One might believe the random field is smoother than Brownian motion, but not smoother than twice integrated Brownian motion and so consider a prior for θ with support from 0.5 to 2.5.

In the vast majority of situations the choice of prior distributions is extremely difficult because the investigator is unskilled in translating his knowledge into statistically meaningful distributions. It is therefore most important to check the inference for sensitivity to the prior distributions chosen. If the inference is sensitive to the particular prior distributions chosen then unless the investigator is confident that the prior is correctly calibrated the value of the final inference itself should be discounted. If one has previous data then by exploring the shape of the likelihood surfaces under reparameterizations one can develop prior distributions for the current data.

One difficulty in specifying a prior distribution for (β, α, θ) is that it is unobservable. However we can observe the random field, in principle, even at the location to be predicted, x_0. We are more comfortable expressing prior knowledge directly in terms of the potentially observable $Z(x_0)$ than the parameters themselves. An intriguing approach to the selection of prior distributions is the so called "device of imaginary results" considered in Good(1965), Winkler(1980) and Stigler (1982). The distribution of $Z(x_0)$ given (β, α, θ) is

$$P(Z(x_0)) = \int_{\Theta} \int_0^{\infty} \int_{R^q} P(Z(x_0) \mid \beta, \alpha, \theta) P(\beta, \alpha, \theta) \; d\beta d\alpha d\theta$$

Hence given that our prior knowledge about $Z(x_0)$ can be expressed as $P(Z(x_0))$ we can then indirectly evaluate priors for $(\beta, \alpha, \theta), P(\beta, \alpha, \theta),$ by solving this integral equation. This will be difficult to achieve in practice, but analysis might suggest families of priors to be explored by other methods. Interestingly, Stigler (1982,§5) suggests that this is the approach that Thomas Bayes would have applied rather than a direct appeal to the principle of "insufficient reason" as previous commentators had inferred.

4.9 Summary and conclusion

The kriging procedure is motivated because it produces optimal predictions when the covariance structure of the random field is known. If the covariance structure is not known and needs to be estimated then the primary motivation for kriging is in doubt. In this chapter we have seen that the Bayesian paradigm provides a framework in which to analyse the performance of the estimated kriging predictor.

As important as the quality of the predictor itself is the quality of the measure of uncertainty attached to that predictor. Many prediction procedures provide reasonable predictions, but supply dubious estimates of uncertainty if they provide estimates at all. In this chapter comparisons between perceived and actual measures of uncertainty were made.

The results of §4.7.1 indicate that the use of kriging predictors based on maximum likelihood covariance structures for the Spherical class usually produces accurate inferences, both actual and perceived. The perceived posterior is insensitive to the specification of the range parameter, but not to the slope parameter. The effect of inadequate specification manifests itself by bias in the plug-in predictor that is not reflected in the perceived posterior.

The results of §4.7.3 indicate that the Matérn class provides adequate inference when used as a surrogate for the Spherical class, while the Squared Exponential class produces inference with unwarranted precision. For both surrogates the perceived inference was much too precise. The maximum likelihood estimate, under the Matérn model, tended to choose a model that was much smoother than the underlying Spherical field. The marginal posterior for the smoothness parameter was very flat and the complete predictive posterior produces close to the correct inference.

The Squared Exponential model was excessively smooth and tended to overstate the precision of the predictors.

CHAPTER 5

A LIKELIHOOD APPROACH TO THE ANALYSIS OF DAVIS' TOPOGRAPHIC DATA

5.1 Introduction

In this chapter we analyse data originally from Davis (1973) that has recently attracted a lot of interest. The data are topological elevations over a small area on the northern side of a hill. The data were measure by a surveying class, using a plane table and alidade. Davis was interested in the analysis of maps and used the survey to produce contours of the region. Our purpose is to demonstrate the value of likelihood based methods in the analysis of spatial data when the objective is prediction.

The data are reproduced from Davis' book in Figure 55. This is a view looking slightly West of South. The region is about 300 yards by 300 yards. The 52 surveyed locations are marked by the drop lines and the symbols represent their elevations in feet above sea level. An important feature is the small streams running northward down the hill and joining together at the base of the region. They are indicated on the map by solid lines. The procedure used by Davis for contouring did not incorporate the information about the topography in the streams, although he recognized their importance. A map manually contoured by hand, given in his Figure 6.10, clearly attempted to use the information in the streams.

The data have been studied by Ripley (1981, pp. 58–72), and subsequently by Warnes (1986), Warnes & Ripley (1987) and Ripley(1988, pp. 15–21). The original data were scaled so that 50 yards in location corresponds to one map unit. This scaling has been carried through the later studies and for this reason will also be used here, even though a scaling in yards is more desirable. All locations are referenced as (Easting, Northing) measured from the South-West corner of the region. The survey locations are recorded to two significant figures and the elevations to three significant figures.

One should note that in Ripley(1981) there are two errors in transcribing the data. The elevation at (315, 110) should be 875, not 855; the former elevation appears in both the 1973 and 1978 editions of Davis' book. On p. 58, the author states that there are 51 observations, although the original data have 52 and they all appear in his diagrams. All the analysis in this chapter has been repeated using the data as reported in Ripley (1981) with no change in the substantive results. Also, Davis (1973) originally refers to the location scale in feet, although it is actually yards.

Ripley (1981) suggests both the Exponential and the Squared Exponential classes as models for the covariance structure. The former class was introduced in §2.3 as a basic model for one-dimensional processes. The later class was proposed by Thompson (1956) as a general model for two dimensional fields. The isotropic covariance has the general form:

$$K_G(x;\theta_1,\theta_2) = \theta_1\theta_2 e^{-x^2/\theta_2^2}$$

where θ_1 is the slope at the origin of the correlation function and θ_2 is is a range parameter. At that time the class became known as the "Gaussian covariance" based on its mathematical form. However this association is not historically accurate and can lead to confusion with the distribution of the field. Perhaps partially based on its familiar name and classical shape it has received extensive use in Meteorology, Hydrology and Geology. Unfortunately it corresponds to a field with analytic realizations, a very severe restriction. The corresponding likelihood surface exists and has a continuous second derivative.

In these previous studies the mean function is based on powers of the Northing and Easting for each location. The information available in the streams was not exploited.

In Ripley(1981) covariance functions are investigated based on fitting by eye the empirical correlation function. The model suggested in Warnes & Ripley (1987) and Ripley(1988), again based on empirical correlation plots, is Exponential with $(\theta_1, \theta_2) = (2112, 2)$ and flat mean.

In Warnes(1986) the focus is the effect of perturbations in the covariance structure on the prediction surface. He uses a flat mean and $\theta_2 = 2$. He finds that the predictions under the Exponential class are insensitive to perturbations in the range parameter and that the reverse is true for the Squared Exponential class. It is argued in Stein & Handcock (1989) that this behavior can be understood in terms of the compatibility of the models in the respective classes. This issue will be returned to in §5.3.

Warnes & Ripley (1987) and Ripley (1988) return to the question of parameter estimation. Their focus is likelihood estimation and they argue that use of the likelihood statistic can lead to misleading inference.

The bulk of this chapter uses this topological data as a forum for the issues raised in Warnes & Ripley (1987) and Ripley(1988). The focus is on the applicability of likelihood methods to the analysis of spatial data.

5.2 Choosing a modeling approach

How should we analyse the data if our objective is to predict the elevations within the region surveyed?

The major assumptions implicit in the model are stationarity of the Gaussian random field, isotropy of the correlations and the correct specification of the mean. These are interdependent so that checking them individually is usually not the best approach.

In general it is difficult to determine if the field is Gaussian because the observations are spatially dependent and the correlation structure is unknown. In particular the marginal distribution of the observations is little guide to the joint distribution. The realizations of the random field can be assumed to be smooth, at least continuous and maybe even differentiable. Two potential models for the covariance structure are mentioned above. In §5.4 a third general model will be investigated. Given the nature of the data and the measurement procedure it will be assumed that the measurement error is small so that the (observed) field is continuous.

The mean function should clearly include the Northing and Easting of the survey locations. In addition, there is information in the locations of the streams that should be taken into account. One crude way is to use the horizontal distance of the survey point to the closest stream as a covariate. This can be measured from Figure 55. We can investigate different approximations to the mean function by using polynomials of Northing, Easting and “Distance to Stream”. An informal way of discriminating between nested models for the mean function is to look at differences in log–likelihood.

The approach used here is to check the assumptions in the context of particular models. Conditional on the correctness of a particular model there are verifiable properties that can be checked. For example, cross validation is used in the next section. As a preliminary check the data were screened for obvious deviations from the assumptions. Some corrected errors are mentioned in §5.1. Initial exploration of univariate data transformations such as square-root and log indicate similar results. The analysis presented in the following sections is based on the untransformed data.

5.3 Model selection and evaluation within the Exponential class

One of the most common methods for fitting a covariance model to data is to match "by eye" a theoretical curve to the empirical correlation plot of the de-trended observations. This guide to intuition is very dubious for three reasons. Firstly, the values in the plot are very highly correlated so that the additional information in the latter points is very small. Secondly, each point is based on the average of greatly differing numbers of pairs of points. Thirdly, misspecification of the mean function will have a big effect upon the points at medium to large lags. As much of the information in the latter points is problematical, one good approach is to draw a line through the first few points so as to gauge the slope at the origin. For the same reasons direct curve fitting by optimization should be avoided.

Ripley (1981), Warnes & Ripley (1987) and Ripley (1988) suggest a value for the range, θ_1 , of about 2 , based on the empirical correlation plot in Figure 56. Two theoretical models have been superimposed for comparison. An Exponential with $\theta_2 = 2$ is below the points for the first few lags, but is a better 'overall' fit to the positive empirical correlations. Some researchers regard fitting 'by eye' to be a better guide to the correlation structure than the maximum likelihood estimate. However, it is based on a regression mind-set that is inappropriate. We will see in §5.5 that fitting 'by eye' is a estimation procedure substantially inferior to maximum likelihood estimate. The maximum likelihood correlation function, $\theta_2 = 6.12$, matches only at the first few lags. The vast differences at larger lags can be indicative of a misspecified mean function. This hypothesis is borne out later when models are compared.

Table 2 summarizes the results of fitting a number of different regression functions using the Exponential model. In this and later summaries of the likelihood we resist the temptation to give standard errors based on the curvature of the likelihood.

Figure 57 is the empirical correlation plot for Model 6, on which the maximum likelihood estimate of the correlation function has been superimposed. Again note the closeness of the maximum likelihood estimate to the empirical values at small lags.

5.3.1 *Cross validation based on the prediction errors*

How can we discriminate between different models given that the "truth" is unlikely to be represented by one of the models considered? A simple cross validation in this situation comprises of fitting the model to the n data sets obtained by excluding successively just one location. In each case predict the elevation at the excluded location from the fitted model, and use these values to check the consistency of the model.

Cross validation is often used to check the consistency of models using the prediction errors,

$$Z(x_i) - \hat{Z}(x_i), \tag{5.3.1}$$

or the standardized prediction errors,

$$\frac{Z(x_i) - \hat{Z}(x_i)}{\sqrt{\widehat{\mathbb{V}}_{\text{Model}}\{Z(x_i) - \hat{Z}(x_i)\}}}, \tag{5.3.2}$$

or the prediction errors standardized by the prediction standard deviations of the largest reasonable model,

$$\frac{Z(x_i) - \hat{Z}(x_i)}{\sqrt{\widehat{\mathbb{V}}_{\text{Model 6}}\{Z(x_i) - \hat{Z}(x_i)\}}}, \tag{5.3.3}$$

as a guide. Figures 58 are plots of the relative prediction error (5.3.3) against the elevation. Note the tendency of the errors to increase with elevation. The models struggle with the value at $(205,\ 40)$ with elevation 960, which is isolated on a high ridge between two branches of the stream. Figures 59 are plots of the standardized prediction error (5.3.2) against the elevation. Conditional on the fitted model being the truth, these values should be mean zero Gaussian with unit variance. However they are still correlated.

It is tempting to define overall measures of fit and the quality of the error precision. One such measure is the average sum of squares of the errors in (5.3.2) where values near 1 are healthy and values much larger than 1 indicate that the model tends to underpredict the true errors. However this measure favors the overprediction of standard errors. Another measure is the average sum of squares of errors in (5.3.1), which measures straight fit, but may be dominated by a few hard to predict survey locations. A better measure standardizes by the prediction standard deviations of the largest reasonable mode. Dividing by the degrees of freedom of the model, instead of n, is a rough adjustment for the complexity of the model. All these measures are reported in Table 2. Model 3, using only Northing and Stream as regressors is very competitive by all these measures, which are influenced by the value at $(205,\ 40)$. Model 3 represents this value best, both absolutely and relatively to its predicted variance. These measures are tied to the observed locations, and represent the entire area only as much as the observations themselves do. In general, the measure should be tuned for the purpose of prediction and in practice we will not be predicting at the observed locations. The Generalized Cross–validation developed by Craven & Wahba (1979) is an interesting approach that adjusts for the "equivalent degrees of freedom" of the model, but is not considered here.

Another model check is possible using the fact that, conditional on the estimated model, the whitened residuals $K_{\hat\theta}^{-\frac{1}{2}}\{Z-F\widehat{\beta}\}$ are independent and standard Gaussian. Figures 60 are plots of the whitened residuals against the elevation, where again the more complicated models are overdispersed. The χ^2 values are given in Table 2. It should be noted that this vector is not unique and we could just as well consider any orthogonal transformation of $K_{\hat\theta}^{-\frac{1}{2}}\{Z-F\widehat{\beta}\}$. The weakness of this check is that the whitened residuals no longer reflect the spatial information.

Overall, there is a weight of evidence for Model 3, although predictions from Model 6 should also be consulted. Clearly this analysis is not the last word on this data; much more can be achieved via models outside the structural confines that have been set. A more honest to earth model would use topographic subject knowledge to involve the stream patterns structurally.

5.4 Model selection and evaluation within the Matérn class

The analysis of the previous chapter indicated that there is room for improvement in the choice of covariance structure. The Exponential, while providing a reasonable initial covariance class, does not allow the field to have differentiable realizations. It is a natural class for one dimensional fields, but does not have special significance for random fields on the plane (Whittle (1954)). Given that *a priori* the form of the covariance is unknown it is unreasonable to exclude the possibility of smoother random fields. The Matérn class, introduced in §1.5.3, is a much richer candidate class. It covers a wide range of random fields, provides smooth transitions between dimensions and has stable numerical properties.

In this section we consider the Matérn class using the same regression model for the mean as for the Exponential. The results of model selection are compared to the Exponential subclass. The corresponding likelihood surfaces exist and are smooth. No evidence of multiple modal behavior has been observed, although there is no proof of unimodality.

Figure 61 reports the log-likelihood for (θ_1, θ_2) profiled over the variance parameter, α and a flat mean. The contours are approximately elliptical and indicate a strong correlation between the estimates θ_1 and θ_2. The smoothness parameter at the maximum is about 1 indicating an almost differentiable field. Note that values for the smoothness parameter between 0.75 and 1.5 and ranges between 1 and 4 have strong likelihoods. Given this strong dependence a re-parameterization is desirable. Figure 62 reports the log-likelihood for (α, θ_2) profiled over the range parameter, θ_1 and a flat mean. The axes of the elliptical contours are now closely aligned with the coordinate axes, indicating that the dependence between the parameters has been reduced. This parameterization will be used from now on. The best Exponential model has a range of 6.12 and a a log-likelihood of -244.60 while the best Exponential model with a range of 2 has a log-likelihood of -239.15.

Figure 63 reports the log-likelihood for (α, θ_2) profiled over the range parameter, θ_1 and a mean based on Northing and distance to closest stream as regressors. The striking feature is the flat ridge over a wide range of the smoothness parameter. The Orders between 2 and 30 have strong likelihood. The additional regressors have accounted for most of the variation, so that the smoothness of the covariance model is difficult to identify. As the smoothness increases the range decreases. We will comment on this lack of identifiability shortly.

Table 3 summarizes the results of fitting a number of different regression functions using the Matérn class. Model 0 is the model suggested by Ripley (1988) and has a log-likelihood of -254.92. As the model becomes more complex the maximum likelihood estimate of the smoothness parameter tends to increase and the range of the covariance function tends to decrease. Under the more complex regression model the covariance structure is essentially non-existent with a range of 0.20 map units and an Order of 11. The closest survey points are 10 yards or 0.2 map units apart. At the same time the likelihood is very flat over the smoothness parameter.

Table 4 summarizes the results of fitting using modified log-likelihoods over a number of different regression models. The likelihood surfaces are similar in shape to the corresponding profiled likelihood surfaces. The modified maximum likelihood estimates tend to be less smooth and have longer ranges than the profile maximum likelihood estimates. The modified likelihood surfaces tend to be very flat over a range of smoothnesses. For example, the maximum likelihood estimates for the range at $\theta_2 = \frac{1}{2}$ are 6.12 under the profile likelihood and 25.6 under the modified likelihood, but the modified likelihood for 6.12 is only 0.3 units less than that at the maximum.

By comparison with Table 2 we see that the differences in log-likelihoods for the Matérn class over the Exponential subclass is at least 2. In addition we see that the range parameter has been greatly reduced.

The flatness of the likelihood surfaces for the more complex mean functions requires more comment. These surfaces are very sensitive to the precise geometric location of observations that are close. For example, suppose we perturb the location of the observation at $(125,\ 225)$ by 5 yards towards the point $(115,\ 240)$. This shift represents the precision at which the locations were recorded. We then recalculated the maximum likelihood estimates for each of the models. The maximum likelihood estimate for the covariance structure for the model with a mean based on Northing and distance to closest stream as regressors is $(\widehat{\alpha}, \widehat{\theta}) = (2181, 0.48, 2.82)$. The original maximum likelihood estimate, reported as Model 4 in Table 3, has $(\widehat{\alpha}, \widehat{\theta}) = (4817, 0.20, 11.12)$. This illustrates the extreme sensitivity of the smoothness parameter to the geometry of the locations. The likelihood of this point under the original data likelihood is 0.89 less than the maximum value. The maximum likelihood estimates of the Matérn model with flat mean and the Exponential Model with each mean change only slightly with this perturbed value. These surfaces are also sensitive to the precise elevations of observations that are close. For example, suppose we perturb the elevation of the observation at $(125,\ 225)$ down by 1 foot towards the elevation at $(115,\ 240)$. This shift represents the precision at which the elevations were recorded. The maximum likelihood estimate for the covariance structure for the model with a mean based on Northing and distance to closest stream as regressors is $(\widehat{\alpha}, \widehat{\theta}) = (4091, 0.23, 8.28)$. Thus the smoothness parameter has dropped by 3 with a negligible drop in likelihood. Again the maximum likelihood estimates of the Matérn model with flat mean and the Exponential Model with each mean change only slightly. To understand this a little more consider Figure 64, the residual elevations after subtracting off the mean function at the maximum likelihood estimate. The pair of observations perturbed are marked with a $\times$. As one point is on the stream, the perturbation strongly alters the contribution to the mean function of the distance to stream variable.

Should we choose between these models? While the maximum likelihood estimate is a good representative value the overall flatness of the likelihoods would suggest against choosing a particular member as the “truth”. Clearly we need additional information before we can choose between member of the same class. The same comments apply to the choice of regression model. It is tempting to base the decisions on the changes in log-likelihood. It is still an open question as to the validity of this decision rule in the face of the interdependence of the mean and covariance structures.

In summary, the Matérn class appears to be an appropriate model for this topographical data because of the wide range of random fields it covers. As the generality of the model for the mean increases the level of identifiability of the smoothness parameter decreases. This suggests against using the maximum likelihood covariance structure alone as a surrogate for the information in the data about the covariance structure. As the log-likelihood is not close to elliptical, the usual measures of uncertainty based on curvature will be effected. Unless the inference is insensitive to the particular choice of model, inference based on the maximum likelihood estimate alone may be inadequate.

5.4.1 *Cross validation based on the likelihoods*

In this section we consider checking the consistency of the Matérn model for the topographical data using cross-validation of likelihoods. If the field is Gaussian and covariance class correctly specified then inference based on the likelihood function, we have argued, is sensible. Our concern is the possible misspecification of the covariance class and the identification of influential observations. In §4.7.3 we investigated the effect on the prediction error of misspecifying the Spherical class by the Matérn class. In §5.3.1 we considered cross–validation based on the prediction errors. We have introduced the topic in §2.7.2, where regularly spaced random fields in two dimensions are considered.

We have seen in §5.4 that the maximum likelihood estimates under the Matérn model with complex mean are sensitive to perturbations in the geometry and elevations of the data. Figure 65 represents the spatial distribution of the full log-likelihoods at the cross-validated maximum likelihood estimates for this model. In general the values are within 0.2 units of the maximum. The feature that stands out are the six very low values. Each of these corresponded to a smoothness parameter estimate above 70. This indicates that dropping out each of these points leads to a great change in the estimate of the smoothness parameter. Each of these values is located on a branch of the stream. This plot provides additional evidence for the sensitivity of the model to the data geometry. We believe that this sensitivity is a strike against using the model as a basis for prediction, as the less complex models are less sensitive.

For comparison we can consider the Matérn model with flat mean. The spatial distribution of the full log-likelihoods is given in Figure 66. We no longer see an extreme set of points, although our eyes are drawn to the location (205, 40) with elevation 960. This value was also sighted as hard to predict when we considered cross-validation based on the prediction error. Another influential point is at (15, 305) probably because it is on the edge of the region and is thus influential on the mean function.

The results for the Matérn model with flat mean and the particular members with $\theta_2 = 1$ and $\theta_2 = \frac{1}{2}$ show less sensitivity, although the observation located at (205, 40) is influential. We have also considered the log-likelihood plotted against the cross-validated maximum likelihood estimates, although it is unclear what they say about the misspecification of the modeling class.

5.5 Bayesian analysis of the topographic data

In this section the Bayesian viewpoint of Chapter 4 will be applied to the topological data. The main question is the quality of prediction, both actual and perceived, achieved by using maximum likelihood estimates of the covariance structure in place of the correct structure. The Matérn modeling class will be the reference class. It is implicitly assumed that the true model is an unknown member of this class. The constant model for the mean is used. This section is conceptually an extension of the discussion in §5.4.

The prior distribution used for the smoothness parameter, θ_2, is uniform from 0.25 to 2.5, the rationale being that we do not expect the realizations to be discontinuous or much smoother than once differentiable. An informative prior for θ_2 is given in Figure 67. Values between $\frac{1}{2}$ and 1 are regarded highly. Values rougher than 0.25 are excluded. The tails on either side of $\frac{1}{2}$ and 1 drop off like an inverse square. This prior will be used later in this section. One could consider placing point masses at the 'knots' $\frac{1}{2}, \frac{3}{2}$ or favoring the smoother structures. The joint prior for (β, α) is inversely proportional to α. This is consistent with the lack of information about β. The issue of choice of distributions prior to the data is also considered in §4.8.

Marginal posteriors for the smoothness parameter are given in Figure 68. Consider the posterior based on the convenience prior. The mode is near $\theta_2 = 1$ sometimes called Whittle's covariance function after Whittle (1954, 1962). Interestingly, Whittle regarded this model as the natural extension of the Exponential model $(\theta_2 = \frac{1}{2})$ from one to two dimensions. It corresponds to a random field with continuous realizations that are on the margin of mean-square differentiability. For $\theta_2 > 1$ the field is mean-square differentiable. The distribution fades out near 0.25 and 2.5, compatible with the prior specification. It is interesting to note that the ratio of the density at the mode to the density at the Exponential model is about $85 : 1$, so that the Exponential appears too rough for this field. The distribution is right skewed, so that the density for $\theta_2 = \frac{1}{2}$ is more than that at $\theta_2 = \frac{3}{2}$. The posterior based on the informative data places less emphasis on the larger values of the smoothness parameter. It is also centered about $\theta_2 = 1$.

Clearly these posterior densities are a useful tool for understanding the data.

The location chosen to be predicted is marked on Figure 55 in the center of the region. It was chosen to be reasonably distant from the survey locations. A second location was chosen on the branch of the stream because it would be potentially useful to a surveyor and because of the close proximity of survey locations.

The traditional kriging approach estimates the covariance parameters and proceeds as if the estimated covariance structure is known to be the correct covariance for the field. In this section we will use the maximum likelihood estimate of the covariance parameters. Measures of uncertainty for the distribution are then based on the perceived error distribution (7.6.1). The actual posterior predictive distribution of this predictor is, under the full model, given by (7.6.3) and the complete posterior predictive distribution is given by (7.4.2). The complete posterior weighs each covariance structure by the posterior density for the covariance structure under the Matérn model.

Figure 69 presents the posterior predictive densities for the model with constant mean function. The perceived posterior is a centered Gaussian with a standard deviation of about 25 feet. The maximum likelihood estimate is $(\widehat{\alpha}, \widehat{\theta}) = (3881, 1.95, 0.97)$. The actual and complete posteriors are mixtures of non-central and central t-distributions respectively. The complete posterior is a better reflection of the uncertainty in the covariance structure and should be regarded as a superior reference for inference. It is always symmetric about zero. The perceived posterior is based on an incorrect model, and can be wider or narrower than the complete posterior depending on the plug-in estimates used. In general it tends to underestimate the uncertainty. The actual posterior of the plug-in predictor indicates that it has a downward bias of about -3 feet. It also indicates that the perceived posterior slightly underestimates the uncertainty of the plug-in predictor. Figure 70 provides relative comparisons of the densities. The vertical axis has a logarithmic scale. The perceived posterior has lighter tails than the complete posterior and the actual posterior of the estimated predictor. Notice that values outside of 75 feet have negligible weight. Hence probability regions based on the perceived and complete posteriors will be similar.

The estimate for the covariance structure proposed in Warnes & Ripley (1987) and Ripley (1988) was $(\tilde{\alpha}, \tilde{\theta}) = (2112, 2, \frac{1}{2})$. There they claimed that the maximum likelihood estimate is “nonsensical” and that the “posterior density will have its mass concentrated on unrealistic values”. We can evaluate this claim within the Bayesian framework.

Figure 71 compares the perceived and actual performance of kriging predictor based on the parameters suggested by Warnes & Ripley (1987) to the complete posterior under the Matérn model. The perceived posterior is much broader than the actual posterior providing conservative inference. The actual posterior of this predictor indicates that it has a bias of about 5 feet. Figure 72 provides a relative comparison. Probability intervals based on the perceived posterior will be markedly wrong under the Bayesian model and will differ substantially from those based on the complete posterior.

If we base the parameter estimates on values suggested by the empirical correlation curves, as Warnes & Ripley (1987) have, then we will tend to obtain a misleading perceived performance and a diminished actual performance compared to the point predictor based on the maximum likelihood estimate. Given agreement on the model and prior distributions for the parameters it is absurd to suggest that the posterior will have its mass concentrated on unrealistic values.

How sensitive is our inference to the choice of prior distributions? In these examples a flat prior distribution for the smoothness parameter is used. Figure 73 is the analogue of Figure 69 using the informative prior in Figure 68. This prior places greater emphasis on values between $\frac{1}{2}$ and 1. The posteriors in Figure 73 and Figure 69 are quite similar. The actual distribution of the prediction error indicates that some of the bias has been removed from the predictor. Figure 74 represents a relative comparison of the complete posteriors using informative priors to the complete posterior using the convenience prior. The inference appears to be insensitive to moderate changes in the prior for θ_2. In general the posterior distribution of the prediction error will be less sensitive to alternative prior distributions than the posterior distribution of the smoothness parameter. The dotted line in Figure 74 represents a relative comparison of the complete posterior using a flat prior for α instead of the usual $1/\alpha$. The resulting posterior has slightly thinner tails. The posterior is insensitive to changes in the prior for β.

The likelihood values in Table 4 suggest that the model with a flat mean may be inadequate as compared to the models including the survey locations and distance to streams as regressors. Figure 75 summarizes the performance of Model 4 in Table 3. The posterior standard deviation is about 12 feet. The perceived posterior using the maximum likelihood estimates is again narrower than the complete posterior. The actual posterior of the plug-in predictor indicates that it has a bias of about 5 feet and that the perceived posterior substantially underestimates the true uncertainty of the predictor. Overall the plug-in predictor has greater deviation from the complete posterior than in the situation of a flat mean. However the performance of all predictors is better, that is, the posteriors are tighter. Figure 76 provides a relative comparison.

Figure 77 is the analogue of Figure 72 using a more complex model for the mean and analyzing the performance of the plug-in predictor based on the values suggested in Ripley (1988). The perceived posterior has a standard deviation of about 25 feet. This is an unacceptable difference. The actual posterior indicates that the plug-in predictor also has a bias of about 10 feet. The perceived performance of this predictor is very misleading and the overall quality of the plug-in predictor is poor compared to the complete Bayesian approach.

This analysis provides insight into the performance of the kriging procedure and the effect of misspecification of the covariance structure.

5.6 Numerical accuracy considerations in the calculation of likelihoods

In this section we consider some accuracy issues in the computation of the likelihood for spatial random fields. The log-likelihood based on observing the random field at n locations was given in (2.2.1). It involves the inverse and determinant of the $n \times n$ matrix K_θ. As K_θ is a covariance matrix it is positive definite and so, in principle, these operations present no difficulties. As the number of observations increases K_θ approaches numerical singularity so that numerical stability becomes important. In addition the computational effort required for these operations increases like n^3. In calculating the log-likelihood it is unnecessary to invert K_θ directly. All that is needed is the log determinant of K_θ and a quadratic form. These may be determined from the Cholesky factorization and solving linear systems in the Cholesky triangle using back-substitution. This is more efficient and numerically stable than calculating the inverse directly. However the Cholesky factorization still requires $n^3/6 + O(n^2)$ operations.

We can monitor how close K_θ is to singular by, κ, the condition number for the inversion problem. It can be defined using the matrix 2-norm, $\| K_\theta \| = \sup_{|x|=1} |K_\theta x|$ where $|x| = \sqrt{x'x}$, so that $\kappa = \| K_\theta \| \cdot \| K_\theta^{-1} \|$. The condition number measures the closeness of K_θ to singularity in the sense that $\kappa^{-1} = \| E \| \, / \, \| K_\theta \|$, where E is the smallest matrix (in the $\| \cdot \|$ sense) for which $K_\theta + E$ is singular. The error in the finite-precision arithmetic of linear systems in K_θ is bounded by a constant times ($\kappa \cdot$ machine precision). The usual rule of thumb is to keep $\kappa <$ (machine precision)$^{-\frac{1}{2}}$. All our calculations were done on a Sun 3/60 using double precision arithmetic corresponding to a machine precision of 10^{-19}. The issue of accuracy of calculation is usually ignored by practitioners.

We now consider the accuracy of the calculation of the log-likelihood for the topological data using a Exponential model with flat mean. Our interest is sparked by a finding of Warnes & Ripley (1987) . They model the data by a Gaussian random field with covariance from the Exponential class using the scale parameterization (σ, θ_2) where $\sigma = \sqrt{\theta_1 \theta_2}$. We can calculate the log-likelihood for (θ_1, θ_2) profiled over a flat mean. For the covariance matrices considered the condition number is about 2×10^3. As a comparison the n x n matrix with $n+1$ on the diagonal and n elsewhere has condition number 2.7×10^3 for $n = 52$ and 10^4 for $n = 100$. We note that that the condition $\kappa <$ (machine precision)$^{-\frac{1}{2}}$ is easily satisfied. Hence we should not run into numerical accuracy problems in calculation the log-likelihood.

Warnes & Ripley (1987) claim the log–likelihood surface is given by Figure 78. We have made exhaustive efforts to reproduce this figure with no success. The actual log–likelihood appears to be Figure 79, with the unique maximum marked with a $\times$. No ripples were found even going to an additional two decimal places than the ripples in Warnes & Ripley (1987) . Figure 80 is a close up of the central region that should show in detail two local maxima. All contour plots are based on independent evaluations on a 40 x 40 grid of points. We note that as the log-likelihood is unimodal in θ_1 it suffices to consider the log-likelihood (2.2.2) also profiled over θ_1. This is given in Figure 81. There is no sign of multiple modes. After this section was written a paper by Mardia & Watkins (1989) addressing the issue of ripples in the likelihood was brought to our attention. Their investigation arrives at the same conclusions as we do.

The condition number of K_θ is sensitive to the geometry of the sites. Typically the condition number increases as the minimum distance between locations decreases. We now investigate the effect of perturbations in the locations of the 52 sites on the condition number of K_θ. As the survey locations are recorded to 2 significant digits, one approach is to randomly move the last digit up or down one and look at the log–likelihood surface produced. The effect is to move the surface around, while retaining the basic shape. The maximum is perturbed about on the line approximately joining (640,7) to (800,4). The condition number of K_θ is approximately inversely proportional to the smallest distance between sites. Hence, unless sites coincide, the condition number is still of the order 10^3 and is of little numerical concern. When the sites were perturbed in a random direction a distance of 5 yards similar results occurred. This indicates that the numerical effects are not sensitive to the exact definition of the observations.

Using the natural slope parameterization of §2.3, the log–likelihood contour is given in Figure 82. Under this parameterization the eccentricity of the contours is greatly reduced. This has obvious advantages for estimation and inference.

In conclusion, we have has found no evidence of ripples in the likelihood surface for the topological data. All attempts to numerically produce them have failed and the computation, while difficult, appears to be numerically stable. Quite apart from the numerical evidence their is no substantive rationale for ripples of this kind. Unlike the situations studied in §2.5 and §2.6 the form of the covariance structure and the geometry of the sites provide no hint that irregular behavior could arise.

5.7 Summary and conclusions

In this chapter topological data from Davis (1973) is used as a forum for the analysis of spatial data when the objective is prediction. Our focus is likelihood methods including Bayesian analysis.

The results of §5.3 indicate that the Exponential class, while adequate, leaves room for improvement. The model selection is by maximum likelihood and evaluation is by cross validation.

The results of §5.4 indicate that the Matérn class leads to improved modelling of the data. The Exponential class is a sub-class of the Matérn class. As the model for the mean became more sophisticated the estimated covariance structures became shorter ranged and smoother. In addition the parameters became less identifiable.

Section 5.6 describes a Bayesian analysis of the kriging predictor. This approach is more sensitive to the complete likelihood surface than plugging in the maximum likelihood estimate of the covariance structure. It allows the performance of the plug-in predictor to be critiqued within a larger framework.

In conclusion, one should ideally base inference on the complete posterior distribution of the prediction error. Usually, inference is based on the perceived posterior of the prediction error based on an estimated covariance structure. In §5.5 we see that kriging based on on the maximum likelihood covariance structure provides an adequate perceived posterior. However there is a definite loss incurred in the use of a single covariance to represent the posterior knowledge of the covariance structure. The maximum likelihood estimate may be the best single representative available, but this reduction itself can be detrimental to the inference.

Ripley (1981), Warnes & Ripley (1987) and Ripley(1988) promote the use of covariance structures suggested by the empirical covariance functions. This chapter suggests that the perceived posteriors based on such estimated covariance structures differ markedly from the complete posteriors. In addition the perceived posterior are quite different for the actual performance of the predictors. Overall they are markedly worse than the plug-in kriging predictions based on the maximum likelihood estimates.

These conclusions provide support for the arguments in §5.5 against comparing empirical correlation plots to the theoretical curves as a means of estimating covariance parameters.

In §5.6 we consider some issues of numerical accuracy in the calculations of likelihoods. As a side benefit some claims of Warnes & Ripley (1987) and Ripley (1988) against likelihood methods have been evaluated. They claim that the likelihood surface of this data for an Exponential covariance model with flat mean is multimodal.

The results of §5.6 indicate that there is no evidence, numerical or substantive, to support this claim.

CHAPTER 5

A LIKELIHOOD APPROACH TO THE ANALYSIS OF DAVIS' TOPOGRAPHIC DATA

5.1 Introduction

In this chapter we analyse data originally from Davis (1973) that has recently attracted a lot of interest. The data are topological elevations over a small area on the northern side of a hill. The data were measure by a surveying class, using a plane table and alidade. Davis was interested in the analysis of maps and used the survey to produce contours of the region. Our purpose is to demonstrate the value of likelihood based methods in the analysis of spatial data when the objective is prediction.

The data are reproduced from Davis' book in Figure 55. This is a view looking slightly West of South. The region is about 300 yards by 300 yards. The 52 surveyed locations are marked by the drop lines and the symbols represent their elevations in feet above sea level. An important feature is the small streams running northward down the hill and joining together at the base of the region. They are indicated on the map by solid lines. The procedure used by Davis for contouring did not incorporate the information about the topography in the streams, although he recognized their importance. A map manually contoured by hand, given in his Figure 6.10, clearly attempted to use the information in the streams.

The data have been studied by Ripley (1981, pp. 58–72), and subsequently by Warnes (1986), Warnes & Ripley (1987) and Ripley(1988, pp. 15–21). The original data were scaled so that 50 yards in location corresponds to one map unit. This scaling has been carried through the later studies and for this reason will also be used here, even though a scaling in yards is more desirable. All locations are referenced as (Easting, Northing) measured from the South-West corner of the region. The survey locations are recorded to two significant figures and the elevations to three significant figures.

One should note that in Ripley(1981) there are two errors in transcribing the data. The elevation at (315, 110) should be 875, not 855; the former elevation appears in both the 1973 and 1978 editions of Davis' book. On p. 58, the author states that there are 51 observations, although the original data have 52 and they all appear in his diagrams. All the analysis in this chapter has been repeated using the data as reported in Ripley (1981) with no change in the substantive results. Also, Davis (1973) originally refers to the location scale in feet, although it is actually yards.

Ripley (1981) suggests both the Exponential and the Squared Exponential classes as models for the covariance structure. The former class was introduced in §2.3 as a basic model for one-dimensional processes. The later class was proposed by Thompson (1956) as a general model for two dimensional fields. The isotropic covariance has the general form:

$$K_G(x; \theta_1, \theta_2) = \theta_1 \theta_2 e^{-x^2/\theta_2^2}$$

where θ_1 is the slope at the origin of the correlation function and θ_2 is is a range parameter. At that time the class became known as the "Gaussian covariance" based on its mathematical form. However this association is not historically accurate and can lead to confusion with the distribution of the field. Perhaps partially based on its familiar name and classical shape it has received extensive use in Meteorology, Hydrology and Geology. Unfortunately it corresponds to a field with analytic realizations, a very severe restriction. The corresponding likelihood surface exists and has a continuous second derivative.

In these previous studies the mean function is based on powers of the Northing and Easting for each location. The information available in the streams was not exploited.

In Ripley(1981) covariance functions are investigated based on fitting by eye the empirical correlation function. The model suggested in Warnes & Ripley (1987) and Ripley(1988), again based on empirical correlation plots, is Exponential with $(\theta_1, \theta_2) = (2112, 2)$ and flat mean.

In Warnes(1986) the focus is the effect of perturbations in the covariance structure on the prediction surface. He uses a flat mean and $\theta_2 = 2$. He finds that the predictions under the Exponential class are insensitive to perturbations in the range parameter and that the reverse is true for the Squared Exponential class. It is argued in Stein & Handcock (1989) that this behavior can be understood in terms of the compatibility of the models in the respective classes. This issue will be returned to in §5.3.

Warnes & Ripley (1987) and Ripley (1988) return to the question of parameter estimation. Their focus is likelihood estimation and they argue that use of the likelihood statistic can lead to misleading inference.

The bulk of this chapter uses this topological data as a forum for the issues raised in Warnes & Ripley (1987) and Ripley(1988). The focus is on the applicability of likelihood methods to the analysis of spatial data.

5.2 Choosing a modeling approach

How should we analyse the data if our objective is to predict the elevations within the region surveyed?

The major assumptions implicit in the model are stationarity of the Gaussian random field, isotropy of the correlations and the correct specification of the mean. These are interdependent so that checking them individually is usually not the best approach.

In general it is difficult to determine if the field is Gaussian because the observations are spatially dependent and the correlation structure is unknown. In particular the marginal distribution of the observations is little guide to the joint distribution. The realizations of the random field can be assumed to be smooth, at least continuous and maybe even differentiable. Two potential models for the covariance structure are mentioned above. In §5.4 a third general model will be investigated. Given the nature of the data and the measurement procedure it will be assumed that the measurement error is small so that the (observed) field is continuous.

The mean function should clearly include the Northing and Easting of the survey locations. In addition, there is information in the locations of the streams that should be taken into account. One crude way is to use the horizontal distance of the survey point to the closest stream as a covariate. This can be measured from Figure 55. We can investigate different approximations to the mean function by using polynomials of Northing, Easting and "Distance to Stream". An informal way of discriminating between nested models for the mean function is to look at differences in log–likelihood.

The approach used here is to check the assumptions in the context of particular models. Conditional on the correctness of a particular model there are verifiable properties that can be checked. For example, cross validation is used in the next section. As a preliminary check the data were screened for obvious deviations from the assumptions. Some corrected errors are mentioned in §5.1. Initial exploration of univariate data transformations such as square-root and log indicate similar results. The analysis presented in the following sections is based on the untransformed data.

5.3 Model selection and evaluation within the Exponential class

One of the most common methods for fitting a covariance model to data is to match “by eye” a theoretical curve to the empirical correlation plot of the de-trended observations. This guide to intuition is very dubious for three reasons. Firstly, the values in the plot are very highly correlated so that the additional information in the latter points is very small. Secondly, each point is based on the average of greatly differing numbers of pairs of points. Thirdly, misspecification of the mean function will have a big effect upon the points at medium to large lags. As much of the information in the latter points is problematical, one good approach is to draw a line through the first few points so as to gauge the slope at the origin. For the same reasons direct curve fitting by optimization should be avoided.

Ripley (1981), Warnes & Ripley (1987) and Ripley (1988) suggest a value for the range, θ_1 , of about 2 , based on the empirical correlation plot in Figure 56. Two theoretical models have been superimposed for comparison. An Exponential with $\theta_2 = 2$ is below the points for the first few lags, but is a better ‘overall’ fit to the positive empirical correlations. Some researchers regard fitting ‘by eye’ to be a better guide to the correlation structure than the maximum likelihood estimate. However, it is based on a regression mind-set that is inappropriate. We will see in §5.5 that fitting ‘by eye’ is a estimation procedure substantially inferior to maximum likelihood estimate. The maximum likelihood correlation function, $\theta_2 = 6.12$, matches only at the first few lags. The vast differences at larger lags can be indicative of a misspecified mean function. This hypothesis is borne out later when models are compared.

Table 2 summarizes the results of fitting a number of different regression functions using the Exponential model. In this and later summaries of the likelihood we resist the temptation to give standard errors based on the curvature of the likelihood.

Figure 57 is the empirical correlation plot for Model 6, on which the maximum likelihood estimate of the correlation function has been superimposed. Again note the closeness of the maximum likelihood estimate to the empirical values at small lags.

5.3.1 *Cross validation based on the prediction errors*

How can we discriminate between different models given that the "truth" is unlikely to be represented by one of the models considered? A simple cross validation in this situation comprises of fitting the model to the n data sets obtained by excluding successively just one location. In each case predict the elevation at the excluded location from the fitted model, and use these values to check the consistency of the model.

Cross validation is often used to check the consistency of models using the prediction errors,

$$Z(x_i) - \hat{Z}(x_i), \tag{5.3.1}$$

or the standardized prediction errors,

$$\frac{Z(x_i) - \hat{Z}(x_i)}{\sqrt{\widehat{\mathbb{V}}_{\text{Model}}\{Z(x_i) - \hat{Z}(x_i)\}}}, \tag{5.3.2}$$

or the prediction errors standardized by the prediction standard deviations of the largest reasonable model,

$$\frac{Z(x_i) - \hat{Z}(x_i)}{\sqrt{\widehat{\mathbb{V}}_{\text{Model 6}}\{Z(x_i) - \hat{Z}(x_i)\}}}, \tag{5.3.3}$$

as a guide. Figures 58 are plots of the relative prediction error (5.3.3) against the elevation. Note the tendency of the errors to increase with elevation. The models struggle with the value at $(205,\ 40)$ with elevation 960, which is isolated on a high ridge between two branches of the stream. Figures 59 are plots of the standardized prediction error (5.3.2) against the elevation. Conditional on the fitted model being the truth, these values should be mean zero Gaussian with unit variance. However they are still correlated.

It is tempting to define overall measures of fit and the quality of the error precision. One such measure is the average sum of squares of the errors in (5.3.2) where values near 1 are healthy and values much larger than 1 indicate that the model tends to underpredict the true errors. However this measure favors the overprediction of standard errors. Another measure is the average sum of squares of errors in (5.3.1), which measures straight fit, but may be dominated by a few hard to predict survey locations. A better measure standardizes by the prediction standard deviations of the largest reasonable mode. Dividing by the degrees of freedom of the model, instead of n, is a rough adjustment for the complexity of the model. All these measures are reported in Table 2. Model 3, using only Northing and Stream as regressors is very competitive by all these measures, which are influenced by the value at $(205,\ 40)$. Model 3 represents this value best, both absolutely and relatively to its predicted variance. These measures are tied to the observed locations, and represent the entire area only as much as the observations themselves do. In general, the measure should be tuned for the purpose of prediction and in practice we will not be predicting at the observed locations. The Generalized Cross–validation developed by Craven & Wahba (1979) is an interesting approach that adjusts for the "equivalent degrees of freedom" of the model, but is not considered here.

Another model check is possible using the fact that, conditional on the estimated model, the whitened residuals $K_{\hat\theta}^{-\frac{1}{2}}\{Z-F\widehat{\beta}\}$ are independent and standard Gaussian. Figures 60 are plots of the whitened residuals against the elevation, where again the more complicated models are overdispersed. The χ^2 values are given in Table 2. It should be noted that this vector is not unique and we could just as well consider any orthogonal transformation of $K_{\hat\theta}^{-\frac{1}{2}}\{Z-F\widehat{\beta}\}$. The weakness of this check is that the whitened residuals no longer reflect the spatial information.

Overall, there is a weight of evidence for Model 3, although predictions from Model 6 should also be consulted. Clearly this analysis is not the last word on this data; much more can be achieved via models outside the structural confines that have been set. A more honest to earth model would use topographic subject knowledge to involve the stream patterns structurally.

5.4 Model selection and evaluation within the Matérn class

The analysis of the previous chapter indicated that there is room for improvement in the choice of covariance structure. The Exponential, while providing a reasonable initial covariance class, does not allow the field to have differentiable realizations. It is a natural class for one dimensional fields, but does not have special significance for random fields on the plane (Whittle (1954)). Given that *a priori* the form of the covariance is unknown it is unreasonable to exclude the possibility of smoother random fields. The Matérn class, introduced in §1.5.3, is a much richer candidate class. It covers a wide range of random fields, provides smooth transitions between dimensions and has stable numerical properties.

In this section we consider the Matérn class using the same regression model for the mean as for the Exponential. The results of model selection are compared to the Exponential subclass. The corresponding likelihood surfaces exist and are smooth. No evidence of multiple modal behavior has been observed, although there is no proof of unimodality.

Figure 61 reports the log-likelihood for (θ_1, θ_2) profiled over the variance parameter, α and a flat mean. The contours are approximately elliptical and indicate a strong correlation between the estimates θ_1 and θ_2. The smoothness parameter at the maximum is about 1 indicating an almost differentiable field. Note that values for the smoothness parameter between 0.75 and 1.5 and ranges between 1 and 4 have strong likelihoods. Given this strong dependence a re-parameterization is desirable. Figure 62 reports the log-likelihood for (α, θ_2) profiled over the range parameter, θ_1 and a flat mean. The axes of the elliptical contours are now closely aligned with the coordinate axes, indicating that the dependence between the parameters has been reduced. This parameterization will be used from now on. The best Exponential model has a range of 6.12 and a a log-likelihood of -244.60 while the best Exponential model with a range of 2 has a log-likelihood of -239.15.

Figure 63 reports the log-likelihood for (α, θ_2) profiled over the range parameter, θ_1 and a mean based on Northing and distance to closest stream as regressors. The striking feature is the flat ridge over a wide range of the smoothness parameter. The Orders between 2 and 30 have strong likelihood. The additional regressors have accounted for most of the variation, so that the smoothness of the covariance model is difficult to identify. As the smoothness increases the range decreases. We will comment on this lack of identifiability shortly.

Table 3 summarizes the results of fitting a number of different regression functions using the Matérn class. Model 0 is the model suggested by Ripley (1988) and has a log-likelihood of -254.92. As the model becomes more complex the maximum likelihood estimate of the smoothness parameter tends to increase and the range of the covariance function tends to decrease. Under the more complex regression model the covariance structure is essentially non-existent with a range of 0.20 map units and an Order of 11. The closest survey points are 10 yards or 0.2 map units apart. At the same time the likelihood is very flat over the smoothness parameter.

Table 4 summarizes the results of fitting using modified log-likelihoods over a number of different regression models. The likelihood surfaces are similar in shape to the corresponding profiled likelihood surfaces. The modified maximum likelihood estimates tend to be less smooth and have longer ranges than the profile maximum likelihood estimates. The modified likelihood surfaces tend to be very flat over a range of smoothnesses. For example, the maximum likelihood estimates for the range at $\theta_2 = \frac{1}{2}$ are 6.12 under the profile likelihood and 25.6 under the modified likelihood, but the modified likelihood for 6.12 is only 0.3 units less than that at the maximum.

By comparison with Table 2 we see that the differences in log-likelihoods for the Matérn class over the Exponential subclass is at least 2. In addition we see that the range parameter has been greatly reduced.

The flatness of the likelihood surfaces for the more complex mean functions requires more comment. These surfaces are very sensitive to the precise geometric location of observations that are close. For example, suppose we perturb the location of the observation at $(125,\ 225)$ by 5 yards towards the point $(115,\ 240)$. This shift represents the precision at which the locations were recorded. We then recalculated the maximum likelihood estimates for each of the models. The maximum likelihood estimate for the covariance structure for the model with a mean based on Northing and distance to closest stream as regressors is $(\widehat{\alpha}, \widehat{\theta}) = (2181, 0.48, 2.82)$. The original maximum likelihood estimate, reported as Model 4 in Table 3, has $(\widehat{\alpha}, \widehat{\theta}) = (4817, 0.20, 11.12)$. This illustrates the extreme sensitivity of the smoothness parameter to the geometry of the locations. The likelihood of this point under the original data likelihood is 0.89 less than the maximum value. The maximum likelihood estimates of the Matérn model with flat mean and the Exponential Model with each mean change only slightly with this perturbed value. These surfaces are also sensitive to the precise elevations of observations that are close. For example, suppose we perturb the elevation of the observation at $(125,\ 225)$ down by 1 foot towards the elevation at $(115,\ 240)$. This shift represents the precision at which the elevations were recorded. The maximum likelihood estimate for the covariance structure for the model with a mean based on Northing and distance to closest stream as regressors is $(\widehat{\alpha}, \widehat{\theta}) = (4091, 0.23, 8.28)$. Thus the smoothness parameter has dropped by 3 with a negligible drop in likelihood. Again the maximum likelihood estimates of the Matérn model with flat mean and the Exponential Model with each mean change only slightly. To understand this a little more consider Figure 64, the residual elevations after subtracting off the mean function at the maximum likelihood estimate. The pair of observations perturbed are marked with a ×. As one point is on the stream, the perturbation strongly alters the contribution to the mean function of the distance to stream variable.

Should we choose between these models? While the maximum likelihood estimate is a good representative value the overall flatness of the likelihoods would suggest against choosing a particular member as the “truth”. Clearly we need additional information before we can choose between member of the same class. The same comments apply to the choice of regression model. It is tempting to base the decisions on the changes in log-likelihood. It is still an open question as to the validity of this decision rule in the face of the interdependence of the mean and covariance structures.

In summary, the Matérn class appears to be an appropriate model for this topographical data because of the wide range of random fields it covers. As the generality of the model for the mean increases the level of identifiability of the smoothness parameter decreases. This suggests against using the maximum likelihood covariance structure alone as a surrogate for the information in the data about the covariance structure. As the log-likelihood is not close to elliptical, the usual measures of uncertainty based on curvature will be effected. Unless the inference is insensitive to the particular choice of model, inference based on the maximum likelihood estimate alone may be inadequate.

5.4.1 *Cross validation based on the likelihoods*

In this section we consider checking the consistency of the Matérn model for the topographical data using cross-validation of likelihoods. If the field is Gaussian and covariance class correctly specified then inference based on the likelihood function, we have argued, is sensible. Our concern is the possible misspecification of the covariance class and the identification of influential observations. In §4.7.3 we investigated the effect on the prediction error of misspecifying the Spherical class by the Matérn class. In §5.3.1 we considered cross–validation based on the prediction errors. We have introduced the topic in §2.7.2, where regularly spaced random fields in two dimensions are considered.

We have seen in §5.4 that the maximum likelihood estimates under the Matérn model with complex mean are sensitive to perturbations in the geometry and elevations of the data. Figure 65 represents the spatial distribution of the full log-likelihoods at the cross-validated maximum likelihood estimates for this model. In general the values are within 0.2 units of the maximum. The feature that stands out are the six very low values. Each of these corresponded to a smoothness parameter estimate above 70. This indicates that dropping out each of these points leads to a great change in the estimate of the smoothness parameter. Each of these values is located on a branch of the stream. This plot provides additional evidence for the sensitivity of the model to the data geometry. We believe that this sensitivity is a strike against using the model as a basis for prediction, as the less complex models are less sensitive.

For comparison we can consider the Matérn model with flat mean. The spatial distribution of the full log-likelihoods is given in Figure 66. We no longer see an extreme set of points, although our eyes are drawn to the location (205, 40) with elevation 960. This value was also sighted as hard to predict when we considered cross-validation based on the prediction error. Another influential point is at (15, 305) probably because it is on the edge of the region and is thus influential on the mean function.

The results for the Matérn model with flat mean and the particular members with $\theta_2 = 1$ and $\theta_2 = \frac{1}{2}$ show less sensitivity, although the observation located at (205, 40) is influential. We have also considered the log-likelihood plotted against the cross-validated maximum likelihood estimates, although it is unclear what they say about the misspecification of the modeling class.

5.5 Bayesian analysis of the topographic data

In this section the Bayesian viewpoint of Chapter 4 will be applied to the topological data. The main question is the quality of prediction, both actual and perceived, achieved by using maximum likelihood estimates of the covariance structure in place of the correct structure. The Matérn modeling class will be the reference class. It is implicitly assumed that the true model is an unknown member of this class. The constant model for the mean is used. This section is conceptually an extension of the discussion in §5.4.

The prior distribution used for the smoothness parameter, θ_2, is uniform from 0.25 to 2.5, the rationale being that we do not expect the realizations to be discontinuous or much smoother than once differentiable. An informative prior for θ_2 is given in Figure 67. Values between $\frac{1}{2}$ and 1 are regarded highly. Values rougher than 0.25 are excluded. The tails on either side of $\frac{1}{2}$ and 1 drop off like an inverse square. This prior will be used later in this section. One could consider placing point masses at the 'knots' $\frac{1}{2}, \frac{3}{2}$ or favoring the smoother structures. The joint prior for (β, α) is inversely proportional to α. This is consistent with the lack of information about β. The issue of choice of distributions prior to the data is also considered in §4.8.

Marginal posteriors for the smoothness parameter are given in Figure 68. Consider the posterior based on the convenience prior. The mode is near $\theta_2 = 1$ sometimes called Whittle's covariance function after Whittle (1954, 1962). Interestingly, Whittle regarded this model as the natural extension of the Exponential model $(\theta_2 = \frac{1}{2})$ from one to two dimensions. It corresponds to a random field with continuous realizations that are on the margin of mean-square differentiability. For $\theta_2 > 1$ the field is mean-square differentiable. The distribution fades out near 0.25 and 2.5, compatible with the prior specification. It is interesting to note that the ratio of the density at the mode to the density at the Exponential model is about $85 : 1$, so that the Exponential appears too rough for this field. The distribution is right skewed, so that the density for $\theta_2 = \frac{1}{2}$ is more than that at $\theta_2 = \frac{3}{2}$. The posterior based on the informative data places less emphasis on the larger values of the smoothness parameter. It is also centered about $\theta_2 = 1$.

Clearly these posterior densities are a useful tool for understanding the data.

The location chosen to be predicted is marked on Figure 55 in the center of the region. It was chosen to be reasonably distant from the survey locations. A second location was chosen on the branch of the stream because it would be potentially useful to a surveyor and because of the close proximity of survey locations.

The traditional kriging approach estimates the covariance parameters and proceeds as if the estimated covariance structure is known to be the correct covariance for the field. In this section we will use the maximum likelihood estimate of the covariance parameters. Measures of uncertainty for the distribution are then based on the perceived error distribution (7.6.1). The actual posterior predictive distribution of this predictor is, under the full model, given by (7.6.3) and the complete posterior predictive distribution is given by (7.4.2). The complete posterior weighs each covariance structure by the posterior density for the covariance structure under the Matérn model.

Figure 69 presents the posterior predictive densities for the model with constant mean function. The perceived posterior is a centered Gaussian with a standard deviation of about 25 feet. The maximum likelihood estimate is $(\widehat{\alpha}, \widehat{\theta}) = (3881, 1.95, 0.97)$. The actual and complete posteriors are mixtures of non-central and central t-distributions respectively. The complete posterior is a better reflection of the uncertainty in the covariance structure and should be regarded as a superior reference for inference. It is always symmetric about zero. The perceived posterior is based on an incorrect model, and can be wider or narrower than the complete posterior depending on the plug-in estimates used. In general it tends to underestimate the uncertainty. The actual posterior of the plug-in predictor indicates that it has a downward bias of about -3 feet. It also indicates that the perceived posterior slightly underestimates the uncertainty of the plug-in predictor. Figure 70 provides relative comparisons of the densities. The vertical axis has a logarithmic scale. The perceived posterior has lighter tails than the complete posterior and the actual posterior of the estimated predictor. Notice that values outside of 75 feet have negligible weight. Hence probability regions based on the perceived and complete posteriors will be similar.

The estimate for the covariance structure proposed in Warnes & Ripley (1987) and Ripley (1988) was $(\tilde{\alpha}, \tilde{\theta}) = (2112, 2, \frac{1}{2})$. There they claimed that the maximum likelihood estimate is “nonsensical” and that the “posterior density will have its mass concentrated on unrealistic values”. We can evaluate this claim within the Bayesian framework.

Figure 71 compares the perceived and actual performance of kriging predictor based on the parameters suggested by Warnes & Ripley (1987) to the complete posterior under the Matérn model. The perceived posterior is much broader than the actual posterior providing conservative inference. The actual posterior of this predictor indicates that it has a bias of about 5 feet. Figure 72 provides a relative comparison. Probability intervals based on the perceived posterior will be markedly wrong under the Bayesian model and will differ substantially from those based on the complete posterior.

If we base the parameter estimates on values suggested by the empirical correlation curves, as Warnes & Ripley (1987) have, then we will tend to obtain a misleading perceived performance and a diminished actual performance compared to the point predictor based on the maximum likelihood estimate. Given agreement on the model and prior distributions for the parameters it is absurd to suggest that the posterior will have its mass concentrated on unrealistic values.

How sensitive is our inference to the choice of prior distributions? In these examples a flat prior distribution for the smoothness parameter is used. Figure 73 is the analogue of Figure 69 using the informative prior in Figure 68. This prior places greater emphasis on values between $\frac{1}{2}$ and 1. The posteriors in Figure 73 and Figure 69 are quite similar. The actual distribution of the prediction error indicates that some of the bias has been removed from the predictor. Figure 74 represents a relative comparison of the complete posteriors using informative priors to the complete posterior using the convenience prior. The inference appears to be insensitive to moderate changes in the prior for θ_2. In general the posterior distribution of the prediction error will be less sensitive to alternative prior distributions than the posterior distribution of the smoothness parameter. The dotted line in Figure 74 represents a relative comparison of the complete posterior using a flat prior for α instead of the usual $1/\alpha$. The resulting posterior has slightly thinner tails. The posterior is insensitive to changes in the prior for β.

The likelihood values in Table 4 suggest that the model with a flat mean may be inadequate as compared to the models including the survey locations and distance to streams as regressors. Figure 75 summarizes the performance of Model 4 in Table 3. The posterior standard deviation is about 12 feet. The perceived posterior using the maximum likelihood estimates is again narrower than the complete posterior. The actual posterior of the plug-in predictor indicates that it has a bias of about 5 feet and that the perceived posterior substantially underestimates the true uncertainty of the predictor. Overall the plug-in predictor has greater deviation from the complete posterior than in the situation of a flat mean. However the performance of all predictors is better, that is, the posteriors are tighter. Figure 76 provides a relative comparison.

Figure 77 is the analogue of Figure 72 using a more complex model for the mean and analyzing the performance of the plug-in predictor based on the values suggested in Ripley (1988). The perceived posterior has a standard deviation of about 25 feet. This is an unacceptable difference. The actual posterior indicates that the plug-in predictor also has a bias of about 10 feet. The perceived performance of this predictor is very misleading and the overall quality of the plug-in predictor is poor compared to the complete Bayesian approach.

This analysis provides insight into the performance of the kriging procedure and the effect of misspecification of the covariance structure.

5.6 Numerical accuracy considerations in the calculation of likelihoods

In this section we consider some accuracy issues in the computation of the likelihood for spatial random fields. The log-likelihood based on observing the random field at n locations was given in (2.2.1). It involves the inverse and determinant of the $n \times n$ matrix K_θ. As K_θ is a covariance matrix it is positive definite and so, in principle, these operations present no difficulties. As the number of observations increases K_θ approaches numerical singularity so that numerical stability becomes important. In addition the computational effort required for these operations increases like n^3. In calculating the log-likelihood it is unnecessary to invert K_θ directly. All that is needed is the log determinant of K_θ and a quadratic form. These may be determined from the Cholesky factorization and solving linear systems in the Cholesky triangle using back-substitution. This is more efficient and numerically stable than calculating the inverse directly. However the Cholesky factorization still requires $n^3/6 + O(n^2)$ operations.

We can monitor how close K_θ is to singular by, κ, the condition number for the inversion problem. It can be defined using the matrix 2-norm, $\| K_\theta \| = \sup_{|x|=1} |K_\theta x|$ where $|x| = \sqrt{x'x}$, so that $\kappa = \| K_\theta \| \cdot \| K_\theta^{-1} \|$. The condition number measures the closeness of K_θ to singularity in the sense that $\kappa^{-1} = \| E \| \,/\, \| K_\theta \|$, where E is the smallest matrix (in the $\| \cdot \|$ sense) for which $K_\theta + E$ is singular. The error in the finite-precision arithmetic of linear systems in K_θ is bounded by a constant times ($\kappa \cdot$ machine precision). The usual rule of thumb is to keep $\kappa <$ (machine precision)$^{-\frac{1}{2}}$. All our calculations were done on a Sun 3/60 using double precision arithmetic corresponding to a machine precision of 10^{-19}. The issue of accuracy of calculation is usually ignored by practitioners.

We now consider the accuracy of the calculation of the log-likelihood for the topological data using a Exponential model with flat mean. Our interest is sparked by a finding of Warnes & Ripley (1987) . They model the data by a Gaussian random field with covariance from the Exponential class using the scale parameterization (σ, θ_2) where $\sigma = \sqrt{\theta_1 \theta_2}$. We can calculate the log-likelihood for (θ_1, θ_2) profiled over a flat mean. For the covariance matrices considered the condition number is about 2×10^3. As a comparison the n x n matrix with $n+1$ on the diagonal and n elsewhere has condition number 2.7×10^3 for $n = 52$ and 10^4 for $n = 100$. We note that that the condition $\kappa <$ (machine precision)$^{-\frac{1}{2}}$ is easily satisfied. Hence we should not run into numerical accuracy problems in calculation the log-likelihood.

Warnes & Ripley (1987) claim the log–likelihood surface is given by Figure 78. We have made exhaustive efforts to reproduce this figure with no success. The actual log–likelihood appears to be Figure 79, with the unique maximum marked with a ×. No ripples were found even going to an additional two decimal places than the ripples in Warnes & Ripley (1987) . Figure 80 is a close up of the central region that should show in detail two local maxima. All contour plots are based on independent evaluations on a 40 x 40 grid of points. We note that as the log-likelihood is unimodal in θ_1 it suffices to consider the log-likelihood (2.2.2) also profiled over θ_1. This is given in Figure 81. There is no sign of multiple modes. After this section was written a paper by Mardia & Watkins (1989) addressing the issue of ripples in the likelihood was brought to our attention. Their investigation arrives at the same conclusions as we do.

The condition number of K_θ is sensitive to the geometry of the sites. Typically the condition number increases as the minimum distance between locations decreases. We now investigate the effect of perturbations in the locations of the 52 sites on the condition number of K_θ. As the survey locations are recorded to 2 significant digits, one approach is to randomly move the last digit up or down one and look at the log–likelihood surface produced. The effect is to move the surface around, while retaining the basic shape. The maximum is perturbed about on the line approximately joining (640,7) to (800,4). The condition number of K_θ is approximately inversely proportional to the smallest distance between sites. Hence, unless sites coincide, the condition number is still of the order 10^3 and is of little numerical concern. When the sites were perturbed in a random direction a distance of 5 yards similar results occurred. This indicates that the numerical effects are not sensitive to the exact definition of the observations.

Using the natural slope parameterization of §2.3, the log–likelihood contour is given in Figure 82. Under this parameterization the eccentricity of the contours is greatly reduced. This has obvious advantages for estimation and inference.

In conclusion, we have has found no evidence of ripples in the likelihood surface for the topological data. All attempts to numerically produce them have failed and the computation, while difficult, appears to be numerically stable. Quite apart from the numerical evidence their is no substantive rationale for ripples of this kind. Unlike the situations studied in §2.5 and §2.6 the form of the covariance structure and the geometry of the sites provide no hint that irregular behavior could arise.

5.7 Summary and conclusions

In this chapter topological data from Davis (1973) is used as a forum for the analysis of spatial data when the objective is prediction. Our focus is likelihood methods including Bayesian analysis.

The results of §5.3 indicate that the Exponential class, while adequate, leaves room for improvement. The model selection is by maximum likelihood and evaluation is by cross validation.

The results of §5.4 indicate that the Matérn class leads to improved modelling of the data. The Exponential class is a sub-class of the Matérn class. As the model for the mean became more sophisticated the estimated covariance structures became shorter ranged and smoother. In addition the parameters became less identifiable.

Section 5.6 describes a Bayesian analysis of the kriging predictor. This approach is more sensitive to the complete likelihood surface than plugging in the maximum likelihood estimate of the covariance structure. It allows the performance of the plug-in predictor to be critiqued within a larger framework.

In conclusion, one should ideally base inference on the complete posterior distribution of the prediction error. Usually, inference is based on the perceived posterior of the prediction error based on an estimated covariance structure. In §5.5 we see that kriging based on on the maximum likelihood covariance structure provides an adequate perceived posterior. However there is a definite loss incurred in the use of a single covariance to represent the posterior knowledge of the covariance structure. The maximum likelihood estimate may be the best single representative available, but this reduction itself can be detrimental to the inference.

Ripley (1981), Warnes & Ripley (1987) and Ripley(1988) promote the use of covariance structures suggested by the empirical covariance functions. This chapter suggests that the perceived posteriors based on such estimated covariance structures differ markedly from the complete posteriors. In addition the perceived posterior are quite different for the actual performance of the predictors. Overall they are markedly worse than the plug-in kriging predictions based on the maximum likelihood estimates.

These conclusions provide support for the arguments in §5.5 against comparing empirical correlation plots to the theoretical curves as a means of estimating covariance parameters.

In §5.6 we consider some issues of numerical accuracy in the calculations of likelihoods. As a side benefit some claims of Warnes & Ripley (1987) and Ripley (1988) against likelihood methods have been evaluated. They claim that the likelihood surface of this data for an Exponential covariance model with flat mean is multimodal.

The results of §5.6 indicate that there is no evidence, numerical or substantive, to support this claim.

CHAPTER 6

SUMMARY AND FUTURE RESEARCH

6.1 Summary

In this section we will give a short overview of the thesis. It provides a complement to the primary five chapter summaries, which focus on the individual findings.

As indicated by the nature of the contributions in this thesis, foundational issues of statistical inference for spatial random fields have yet to be resolved. Our level of comprehension is usually 'What are the right things to think about?', sometimes progressing to 'What are the right things to do?' and rarely to 'Is it possible to prove this is the right thing to do?'.

There are two factors that make the statistical issues challenging compared to traditional statistical problems. The first is the dependence structure of the observations. If this structure is known up to location and scale parameters, then we can deal with the issues in familiar ways. In the majority of situations the dependence structure is unknown so that we must adapt our inference to account for this uncertainty. Until recently the impact of this uncertainty was largely unexplored and usually ignored. An exception is time-series where there is a vast literature (Priestley (1981)). The second factor is the pertinence of the spatial location of the observations to the inference. When the observations are regularly spaced, as in time-series, symmetry can be utilized to motivate and substantiate approaches to inference. When the observations are irregularly spaced it is necessary for inference to reflect the individual geometries. In particular the mathematics for the interesting situations is often intractable. This is reflected in this thesis where the theoretical contributions are based on either discrete equally spaced observation, continuous observation on a segment, or asymptotics. In all three approaches the data geometry has been finessed.

The findings of Chapter 2, summarized in §2.8, are central to the thesis and motivate the other chapters. In Chapter 3, we investigate the approximation of discrete observation in a fixed interval by continuous observation on the same interval. We find that the distribution of the maximum likelihood estimates are well approximated by their continuous versions when the range of correlation is comparable to the length of the segment, in a sense made precise. These distributions are surprising close to log-Gaussian. We see in Chapter 4 that much can be gained by viewing the kriging procedure for the prediction of Gaussian random fields within the Bayesian framework. In particular the effects of model misspecification are readily observable. In Chapter 5 topographical data from Davis (1973) is used as a forum for for the analysis of spatial data when the objective is prediction.

6.2 Future research

We have focussed on parametric approaches and likelihood based inference. Based on the findings of this thesis, our future research will continue on this path. Of course, there are other paths to follow and one may lead to better solutions. The primary focus of future research will be model validation. That is, how can we tell if we have misspecified a hypothesized parametric model. We have explored some simple approaches in Chapter 5. In §2.7.2 and §5.4.1 we considered cross-validation based on the likelihood statistic. In §5.3.1 we considered cross-validation based on the prediction errors. The study of these approaches is in its infancy. Another approach is based on resampling the observations within successively finer sub-regions. The analysis is redone on each subsample for a given resampling and the variation in the inference within a resampling is compared to that for successively finer resamplings. For example, consider observing data on a 8×8 grid in two dimensions. We can look at the 4 overlapping 6×6 blocks and for each estimate the parameters under a hypothesized model. These 4 estimates can be compared to those obtained under 4×4, 3×3 and 8×8 blockings. Deviations in the pattern of estimates as the resampling changes can be evidence for model misspecification. Clearly the type of resampling should be tailored to the parameter of interest. For example, non-contiguous blocks might be better for parameters measuring the range of dependence. The individual estimates will be dependent even if the subsamples are nonintersecting, so care must be taken in calibrating the procedure. Initial results indicate that such resampling techniques can be very informative.

Choosing good modeling classes is vital. This thesis promotes the Matérn class of covariance functions for use as an omnibus model. We still need to extensively apply it to real phenomena. Its statistical properties require more exploration, especially the geometry of the likelihood. For example, we do not have a proof that the likelihood is unimodal although the weight of empirical evidence indicates that it is. The properties of peculiar models such as the Spherical and Square Exponential should be publicized, so that researchers better understand their properties.

Advances in statistical computations and computer hardware are having an impact on the theory of spatial prediction. Techniques regarded as extravagant today will not be in the future. This is particularly true for graphical approaches. As the figures in this thesis indicate, the analysis of spatial data is especially amenable to graphical methods. We have considered in §2.7 and §5.6 numerical issues in the calculation of likelihoods. These issues have received careful attention from numerical analysts and our research should reflect their contributions.

In Chapter 3, we consider using the empirical spectral density as a tool for inference about the covariance structure when the correlation length of the process is of the same magnitude as the length of observation. There we do not discuss the estimation of the empirical spectral density based on observing the random field as irregular locations. One approach is called Direct Quadratic Spectrum Estimation and advocated by Marquardt & Acuff (1982, 1984). This amounts to the direct approximation of the integrals defining the empirical spectral density. The issue has been addressed in the literature on irregularly spaced time-series (Masry (1984)). Our objective is to use the estimate of the empirical spectral density as a diagnostic tool. It can also be used to estimate the covariance structure (Wahba (1980)).

The motivating problem for this thesis was the prediction of ore grades within the Mount Charlotte gold mine. A complete and satisfactory analysis remains for future research.

REFERENCES

Abramowitz, M. & Stegun, I. A. (1964). *Handbook of Mathematical Functions,* Vol. 55, Washington Bureau of Standards; reprinted in 1968 by Dover Publications, New York, 1087p.

Alekseev, V. G. (1976). Some Problems in the Spectral Analysis of Gaussian Random Processes. *Theo. Prob. and Math. Stat. 10,* 3-11.

Amos, D. E. (1986). A Portable Package for Bessel Functions of Complex Argument and Nonnegative Order. *ACM Trans. Math. Soft. 12,* 265-273.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* John Wiley, New York, 704p.

Aratò, M. (1964a; 1978). On the Statistical Examination of Continuous State Markov Processes: I. In *Selected Translations of Mathematical Statistics and Probability, 14.* American Mathematics Society, Providence, Rhode Island, 203-225.

Aratò, M. (1964b; 1978). On the Statistical Examination of Continuous State Markov Processes: III. In *Selected Translations of Mathematical Statistics and Probability, 14.* American Mathematics Society, Providence, Rhode Island, 253-267.

Armstrong, M. & Delfiner, P. (1980). Towards a More Robust Variogram: A Case Study in Coal. Centre de Morphologie Mathématique, Fontainebleau, France, N–671, 180p.

Barnard, G. A. (1967). The Use of the Likelihood Function in Statistical Practice. *5th Berkeley Symp. Math. Statist. & Prob. 1,* 27-40.

Barndorff–Nielsen, O. E. (1983). On a Formula for the Distribution of a Maximum Likelihood Estimator. *Biometrika 70,* 343-65.

Bartlett, M. S. (1946). On the Theoretical Specification and Sampling Properties of Autocorrelated Time Series. *J. R. Statist. Soc. (Suppl.). 8,* 27.

Bartlett, M. S. (1978). *An Introduction to Stochastic Processes.* 3rd Ed., Cambridge University Press, Cambridge, 388p.

Bastin, G. & Gevers, M. (1985). Identification and Optimal Estimation of Random Fields from Scattered Point-wise Data. *Automatica 21, 2,* 139-155.

Baxter, G. (1956). A Strong Limit Theorem for Gaussian Processes. *Proc. Amer. Math. Soc. 7,* 522-525.

Beach, C. M. & MacKinnon, J. G. (1983). A Maximum Likelihood Procedure for Regression with Autocorrelated Errors. *Econometrica 46, 1,* 51-58.

Belayev, Y. K. (1961). Continuity and Hölder's Conditions for Sample Functions of Stationary Gaussian Processes. *4th Berkeley Symp. Math. Statist. & Prob. 2,* 23-33.

Bentkus, R. Yu. & Rudzkis, R. A. (1982). On the Distribution of Some Statistical Estimates of Spectral Density. *Theo. Prob. and its Appl. 27,* 795-814.

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley, New York, 253p.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison–Wesley, Reading, Mass., 588p.

Brillinger, D. R. (1970). The Frequency Analysis of Relations between Stationary Spatial Series. *Proc. 12th Bien. Sem. Canad. Math. Cong.,* 39-82.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory.* Holt, Rinehart and Winston, New York, 552p.

Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models.* Marcel Dekker, 472p.

Butler, R. W. (1986). Predictive Likelihood Inference with Applications (with discussion). *J. Royal Statist. Soc. B, 48,* 1-38.

Butzer, P. L. & Nessel, R. J. (1971). *Fourier Analysis and Approximation.* Academic Press, New York, 553p.

Cambanis, S. (1985). Sampling Designs in Time Series. In *Time Series in the Time Domain (Handbook of Statistics, Vol. 5),* edited by Hannan, E. J., Krishnaiah, P. R., and Rao, M. M., Elsevier, Amsterdam, Netherlands, 337-362.

Cheng, R. C. & Iles, T. C. (1987). Corrected Maximum Likelihood on Non–regular Problems. *J. R. Statist. Soc. B, 49,* 95-101.

Clark, I. (1979). *Practical Geostatistics.* Applied Science, Essex, England, 169p.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc. B, 49,* 1-39.

Cramèr, H. & Leadbetter, M. R. (1967). *Stationary and Related Stochastic Processes.* John Wiley, New York, 348p.

Craven, P. & Wahba, G. (1979). Smoothing Noisy Data with Spline Functions. *Numerisch Mathematic. 31,* 377-403.

Cressie, N. (1984). Towards Resistant Geostatistics. In *NATO ASI Series, Geostatistics for Natural Resource Characterization,* edited by Verly, G., David, M., Journel, A. G. and Marechal, A., Reidel Publishing Co., Boston, Mass., 21-44.

Cressie, N. (1985). Fitting Variogram Models using Weighted Least Squares. *Math. Geol. 17,* 693-702.

Cressie, N. & Hawkins, J. (1980). Robust Estimation of the Variogram. *Math. Geol. 12, 2,* 115-126.

Cressie, N. & Horton, A. (1987). A Robust-Resistant Spatial Analysis of Soil Water Infiltration. *Water. Resour. Res. 23, 5,* 911-917.

Creutin, J. D. & Obled, C. (1982). Objective Analyses and Mapping Techniques for Rainfall Fields: An Objective Comparison. *Water. Resour. Res. 18, 2,* 413-431.

Davis, J. C. (1973). *Statistics and Data Analysis in Geology.* John Wiley, New York, 550p.

Davis, R. B. (1973). Numerical Inversion of a Characteristic Function. *Biometrika 60,* 415-417.

Dawid, A. P. (1984). Present Position and Potential Developments: Some Personal Views. *J. R. Statist. Soc. A, 147,* 278-292.

Delfiner, P. (1976). Linear Estimation of Non-Stationary Spatial Phenomena. In *Advanced Geostatistics in the Mining Industry,* edited by Gurascio, M., David, M., Huijbregts, C. Reidel Publishing Co., Dordrecht, 49-68.

Dickey, D. A. & Fuller, W. A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica 49,* 1057-1072.

Diaconis, P. (1988). Bayesian Numerical Analysis. In *Statistical Decision Theory and Related Topics IV, Vol. 1,* edited by Gupta, S. S. & Berger, J. O., Springer-Verlag, New York, 163-175.

Dowd, P. A. (1984). The Variogram and Kriging: Robust and Resistant Estimators. In *NATO ASI Series, Geostatistics for Natural Resource Characterization,* edited by Verly, G., David, M., Journel, A. G. and Marechal, A., Reidel Publishing Co., Hingham, Mass., 91-106.

Dzhaparidze, K. O. & Yaglom, A. M. (1983). Spectrum Parameter Analysis in Time–Series. In *Developments in Statistics, Vol. 4,* edited by Krishnaiah, P. R., Academic Press, New York, 1-96.

Feinerman, E., Dagan, G. & Bresler, E. (1986). Statistical Inference of Spatial Random Functions. *Water. Resour. Res. 22, 6,* 935-942.

Feldman, J. (1958). Equivalence and Perpendicularity of Gaussian Processes. *Pacif. J. Math. 8,* 699-708.

Gandin, L. S. (1963; 1965). *Objective Analysis of Meterological Fields.* GIMIZ, Leningrad, 238p. Translated from Russian by Hardin, R., Israel program for Scientific Translations, Jerusalem, 242p.

Gel'fand, I. M. & Vilenkin, N. Ya. (1964). *Generalized Functions Vol. 4: Applications of Harmonic Analysis.* Translated from Russian by Feinstein, A., Academic Press, New York, 384p.

Geol, P. K. & Zellner, A. (1986). In *Bayesian Inference and Decision Techniques: Essays in honor of Bruno de Finetti,* edited by Zellner, A., North–Holland, Amsterdam, Neth., 401-422.

Gil–Pelaes, J. (1951). Note on the Inversion Theorem. *Biometrika 38,* 481-482.

Goldberger, A. (1962). Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *J. Amer. Statist. Assoc. 57,* 369-375.

Good, I. J. (1965). *The Estimation of Probabilities; An Essay on Modern Bayesian Methods.* M. I. T. Press, Cambridge, Mass, 109p.

Hajek, J. (1958). On a Property of Normal Distributions of an Arbitrary Stochastic Process. *Czec. Math. J. 8,* 610-618.

Harrison, P. J. & Stephens, C. F. (1976). Bayesian Forecasting (with discussion). *J. Royal Statist. Soc. B, 38,* 205-247.

Hasza, D. P. (1980). A Note on Maximum Likelihood Estimation for the First–Order Autoregressive Process. *Commun. Statist.-Theor. Meth. A, 9, 13,* 1411-1415.

Hinkley, D. V. (1979). Predictive Likelihood. *Ann. Statist. 7,* 718-728 (corrig., 8, 694).

Hoeksema, R. J. & Kitanidis, P. K. (1985). Analysis of the Spatial Structure of Properties of Selected Aquifers. *Water Resourc. Res. 21, 4,* 563-572.

Ibragimov, I. A. (1963). On Estimation of the Spectral Function of a Stationary Gaussian Process. *Theo. Prob. and its Appl. 8,* 366-401.

Ibragimov, I. A. & Rozanov, Y. A. (1978). *Gaussian Random Processes.* Translated from Russian by Aries, A. B., Springer-Verlag, New York, 275p.

Jenkins, G. M. & Watts, D. G. (1968). *Spectral Analysis and its Applications.* Holden–Day, San Francisco, 523p.

Jones, R. H. (1989). Fitting a Stochastic Partial Differential Equation to Aquifer Head Data. To appear in *Stochastic Hydrol. Hydraul. 3.*

Journel, A. G. & Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press, London, 600p.

Kitanidis, P. K. (1983). Statistical Estimation of Polynomial Generalized Covariance Functions and Hydrologic Applications. *Water Resourc. Res. 19,* 909-921.

Kitanidis, P. K. & Lane, R. W. (1985). Maximum Likelihood Parameter Estimation of Hydrologic Spatial Processes by the Gauss-Newton Method. *J. Hydrol. 17,* 31-56.

Klein, R. & Giné, E. (1975). On Quadratic Variation of Processes with Gaussian Increments. *Ann. Prob. 3,* 716-721.

Koopmans, T. (1942). Serial Correlation and Quadratic Forms in Normal Variates. *Ann. Math. Statist. 13,* 14-33.

Krasnitskii, S. M. (1973). On Conditions of Equivalence and Perpendicularity of Measures corresponding to Homogeneous Gaussian fields. *Theo. Prob. and its Appl. 18,* 588-592.

Krasnitskii, S. M. (1979). Some Examples of Equivalent and Orthogonal Gaussian Distributions. *Theo. Prob. and its Appl. 24,* 161-165.

Lahiff, M. (1984). Bayes and Likelihood Methods for Prediction and Estimation in the AR(1) model. Ph.D. thesis, Department of Statistics, University of Chicago, Chicago, Illinois.

McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman & Hall, London, 285p.

Mardia, K. V. & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika 73,* 135-146.

Mardia, K. V. & Watkins, A. J. (1989). On Multimodality of the Likelihood in the Spatial Linear Model. To appear in *Biometrika 78.*

Marquardt, D. W. & Acuff, S. K. (1982). Direct Quadratic Spectrum Estimation from Unequally Spaced Data. In *Applied Time Series Analysis,* edited by Anderson, O. D. & Perryman, M. R., North–Holland, Amsterdam, Neth., 199-227.

Marquardt, D. W. & Acuff, S. K. (1984). Direct Quadratic Spectrum Estimation from Irregularly Spaced Data. In *Time Series Analysis of Irregularly Observed Data,* edited by Parzen, E., Lecture Notes in Statistics, Vol. 25, Springer-Verlag, New York, 211-223.

Masry, E. (1984). Spectral and Probability Density Estimation from Irregularly Observed Data. In *Time Series Analysis of Irregularly Observed Data,* edited by Parzen, E., Lecture Notes in Statistics, Vol. 25, Springer-Verlag, New York, 363p.

Mejía, J. M. & Rodríguez-Iturbe, I. (1974). On the Synthesis of Random Field Sampling From the Spectrum: An Application to the Generation of Hydrologic Spatial Processes. *Water. Resour. Res. 10, 4,* 705-711.

Matérn, B. (1947). *Metoder att Uppskatta Noggranheten vid Linje- och Provytetaxering.* (Summary: Methods of estimating the accuracy of line and sample plot surveys.) *Medd. från statens skogsforskningsinst., 36, nr. 1,* { in Swedish }.

Matérn, B. (1960; 1986). *Spatial Variation Medd. från statens skogsforskningsinst., 49, nr. 5,* Second Ed. Published as *Lecture Notes in Statistics,* 36, Springer-Verlag, Berlin.

Matheron, G. (1963). Principles of Geostatistics. *Econ. Geol. 58,* 1246-1266.

Matheron, G. (1973). The Intrinsic Random Functions and their Applications. *Adv. Appl. Probab. 5,* 437-468.

Matheron, G. (1974). Representations Stationnaires et Representations Minimales pour les F.A.I.–k. Centre de Morphologie Mathématique, Fontainebleau, France, N–377, 47p.

Nikolskii, S. M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems.* Translated from Russian by Danskin, J., Springer-Verlag, New York, 450p.

Omre, H. (1984). The Variogram and its Estimation. In *NATO ASI Series, Geostatistics for Natural Resource Characterization,* edited by Verly, G., David, M., Journel, A. G. and Marechal, A., Reidel Publishing Co., Boston, Mass., 107-125.

Omre, H. (1987). Bayesian Kriging - Merging Observations & Qualified Guesses in Kriging. *Math. Geol. 19, 1,* 25-39.

Omre, G. M. & Halvarsen, D. F. (1989). A Bayesian Approach to Kriging. In *Proc. Third Geostat. Congress I,* edited by Armstrong, M., Academic Publishers, Dordrecht, 49-68.

Parzen, E. (1984). *Time Series Analysis of Irregularly Observed Data.* Lecture Notes in Statistics, Vol. 25, Springer-Verlag, New York, 363p.

Patterson, H. D. & Thompson, R. (1974). Maximum Likelihood Estimation of Components of Variance. *Proc. 8th Int. Biometric Conf.,* 197-207.

Philip, G. M. & Watson, D. F. (1986a). Matheronian Geostatistics – Quo Vadis? *Math. Geol. 18,* 93-117.

Philip, G. M. & Watson, D. F. (1986b). Geostatistics & Spatial Data Analysis. *Math. Geol. 18,* 505-509.

Philip, G. M. & Watson, D. F. (1986c). A Method for Assessing Local Variation Among Scattered Measurements. *Math. Geol. 18,* 759-764.

Phillips, P. C. B. (1987a). Time Series Regression with a Unit Root. *Econometrica 55,* 277-301.

Phillips, P. C. B. (1987b). Towards a Unified Asymptotic Theory for Autoregression. *Biometrika 74,* 535-547.

Pochekuev, V. P. (1988). On a Method of Randomized Estimation of the Spectral Density of a Stationary Random Process. *Theo. Prob. and Math. Stat. 36,* 119-125.

Priestley, M. B. (1981). *Spectral Analysis and Time Series, Vol 1.* Academic Press, New York, 890p.

Ripley, B. D. (1981). *Spatial Statistics.* John Wiley, New York, 252p.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge, 148p.

Rozanov, Y. A. (1967). *Stationary Random Processes.* Holden–Day, San Francisco, 211p.

Rozanov, Y. A. (1968). *Infinite Dimensional Gaussian Distributions.* Proceedings of the Steklov Institute of Mathematics, *Amer. Math. Soc. 108,* 1-136.

Sacks, J. & Schiller, S. B. (1988). Spatial Designs. In *Statistical Decision Theory and Related Topics IV, Vol. 2,* edited by Gupta, S. S. & Berger, J. O., Springer-Verlag, New York, 385-389.

Sacks, J., Schiller, S. B. & Welch, W. J. (1989). Designs for Computer Experiments. *Technometrics 31,* 41-47.

Sacks, J. & Ylvisaker, D. (1966). Designs for Regression Problems with Correlated Errors. *Ann. Math. Statist. 37,* 66-89.

Sacks, J. & Ylvisaker, D. (1968). Designs for Regression Problems with Correlated Errors: Many Parameters. *Ann. Math. Statist. 39,* 49-69.

Sacks, J. & Ylvisaker, D. (1970). Designs for Regression Problems with Correlated Errors III. *Ann. Math. Statist. 41,* 2057-2074.

Sacks, J. & Ylvisaker, D. (1971). Statistical Designs and Integral Approximation. *Proc. 12th Bien. Sem. Canad. Math. Soc.,* 115-136.

Samper, F. J. & Neuman, S. P. (1989). Estimation of Spatial Covariance Structures by Adjoint State Maximum Likelihood Cross Validation 1. Theory. *Water. Resour. Res. 25, 3,* 351-362.

Skorokhod, A. V. & Yadrenko, M. I. (1973). On Absolute Continuity of Measures Corresponding to Homogeneous Gaussian Fields. *Theo. Prob. and its Appl. 18,* 27-40.

Smith, R. L. (1985). Maximum Likelihood Estimation in a Class of Non–Regular Cases. *Biometrika 72,* 67-90.

Smith, R. L. & Naylor, J. C. (1987). A Comparison of Maximum Likelihood and Bayesian Estimators for the Three–Parameter Weibull Distribution. *Applied Statistics 36,* 358-369.

Solo, V. (1984). The Order of Differencing in ARIMA Models. *J. Amer. Statist. Assoc. 79,* 916-921.

Stein, M. L. (1987a). Minimum Norm Quadratic Estimation of Spatial Variograms. *J. Amer. Statist. Assoc. 82,* 765-772.

Stein, M. L. (1987b). Uniform Asymptotic Optimality of Linear Predictions of a Random Field using an Incorrect Second-Order Structure. University of Chicago Tech. Report 214. Submitted to *Ann. Statist.*

Stein, M. L. (1987c). Large Sample Properties of Simulations Using Latin Hypercube Sampling. *Technometrics 29,* 143-151.

Stein, M. L. (1988a). Asymptotically Efficient Prediction of a Random Field with a Misspecified Covariance Function. *Ann. Statist. 16,* 55-63.

Stein, M. L. (1988b). Bounds on the Efficiency of Linear Predictions Using an Incorrect Covariance Function. University of Chicago Tech. Report 237.

Stein, M. L. & Handcock, M. S. (1989). Some Asymptotic Properties of Kriging When the Covariance Function is Misspecified. *Math. Geol. 21,* 171-190.

Stigler, S. M. (1982). Thomas Bayes's Bayesian Inference. *J. Royal. Statist. Soc. A, 145, 2,* 250-258.

Striebel, C. (1959). Densities for Stochastic Processes. *Ann. Math. Statist. 30,* 559-567.

Thompson, P. D. (1956). Optimal Smoothing of two-dimensional fields. *Tellus 8,* 384-393.

van de Laan, C. G. & Temme, N. M. (1980). *Calculation of Special Functions: the Gamma function, the Exponential Integrals and Error-like Functions.* CWI tract, 10, Centre for Mathematics and Computer Science, Netherlands, 231p.

Vecchia, A. V. (1985). A General Class of Models for Stationary Two-Dimensional Random Processes. *Biometrika 72,* 281-91.

Vecchia, A. V. (1988). Estimation and Model Identification for Continuous Spatial Processes. *J. Royal Statist. Soc. B, 50,* 297-312.

Wahba, G. (1980). Automatic Smoothing of the Log Periodogram. *J. Amer. Statist. Assoc. 75,* 122-132.

Warnes, J. J. (1986). A Sensitivity Analysis for Universal Kriging. Mathematical Geology, *Math. Geol. 18,* 653-676.

Warnes, J. J. & Ripley, B. D. (1987). Problems with Likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika 74,* 3, 640-2.

West, M. & Harrison, P. J. (1986). Monitoring and Adaptation in Bayesian Forecasting Models. *J. of the Amer. Stat. Assoc. 81,* 741-750.

White, J. S. (1958). The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case. *Ann. Math. Statist. 29,* 1188-1197.

White, J. S. (1961). Asymptotic Expansions for the Mean and Variance of the Serial Correlation Coefficient. *Biometrika 48, 1,* 85-94.

Whittle, P. (1954). On Stationary Processes in the Plane. *Biometrika 41,* 434-449.

Whittle, P. (1956). On the Variation of Yield variance with Plot Size. *Biometrika 43,* 337-343.

Whittle, P. (1962). Topographic Correlation, Power-Law Covariance Functions, and Diffusion. *Biometrika 49,* 305-314.

Whittle, P. (1963). Stochastic Processes in Several Dimensions. *Bull. Intern. Statist. Inst. 40, 1,* 974-994.

Winkler, R. L. (1980). Prior Information, Predictive Distributions and Bayesian Model-Building. In *Bayesian Analysis in Econometrics and Statistics,* edited by Zellner, A., North–Holland, Amsterdam, Neth., 95-109.

Yadrenko, M. I. (1983). *Spectral Theory of Random Fields.* Optimization Software, New York, 259p.

Yaglom, A. M. (1987a). *Correlation Theory of Stationary and Related Random Fields, Vol. I.* Springer Series on Statistics, Springer-Verlag, New York, 526p.

Yaglom, A. M. (1987b). *Correlation Theory of Stationary and Related Random Fields, Vol. II.* Springer Series on Statistics, Springer-Verlag, New York, 258p.

Yakowitz, S. J. & Szidarovszky, F. (1985). A Comparison of Kriging with Nonparametric Regression Methods. *J. of Mult. Anal. 16,* 21-53.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* John Wiley, New York, 178p.