J. R. Statist. Soc. B (2004)
66, *Part* 2, *pp.* 275–296

Approximating likelihoods for large spatial data sets

Michael L. Stein, Zhiyi Chi and Leah J. Welty

University of Chicago, USA

[Received January 2003. Revised September 2003]

Summary. Likelihood methods are often difficult to use with large, irregularly sited spatial data sets, owing to the computational burden. Even for Gaussian models, exact calculations of the likelihood for n observations require $O(n^3)$ operations. Since any joint density can be written as a product of conditional densities based on some ordering of the observations, one way to lessen the computations is to condition on only some of the 'past' observations when computing the conditional densities. We show how this approach can be adapted to approximate the restricted likelihood and we demonstrate how an estimating equations approach allows us to judge the efficacy of the resulting approximation. Previous work has suggested conditioning on those past observations that are closest to the observation whose conditional density we are approximating. Through theoretical, numerical and practical examples, we show that there can often be considerable benefit in conditioning on some distant observations as well.

Keywords: Chlorophyll fluorescence; Estimating equations; Restricted maximum likelihood; Variogram estimation

1. Introduction

Estimating spatial covariance structures is fundamental to geostatistics. Spatial processes are often observed at irregular locations, which can make estimating the covariance structures of these processes rather difficult. In the geostatistical literature, inference about the second-order structure of an isotropic random field is often based on the empirical variogram: the average squared increment between all pairs of observations within some range of interpoint distances plotted as a function of distance. Since variograms must be conditionally negative definite (Chilès and Delfiner, 1999), practical variogram estimation generally requires the selection of some parametric class of variograms and the estimation of these parameters, although nonparametric methods using spectral approximations are available (Shapiro and Botha, 1991; Genton and Gorsich, 2002). It is natural to consider likelihood-based and Bayesian methods when estimating the parameters of a spatial covariance function. For Gaussian random fields, it is easy to write down an exact expression for the likelihood function. Although likelihood methods that are based on a Gaussian assumption have not been fully embraced by the geostatistical community (Chilès and Delfiner (1999), page 110), they have been used more frequently in recent years (Lark, 2000; Pardo-Igúzquiza, 1998; Pardo-Igúzquiza and Dowd, 1997, 1998; Park and Baek, 2001). Furthermore, the calculation of likelihoods for Gaussian random fields is an essential step in algorithms for making inferences on some kinds of non-Gaussian random fields (Diggle *et al.*, 1998). Unfortunately, even for Gaussian random fields, when the n observations are irregularly sited, each calculation of the likelihood function requires $O(n^3)$ operations and hence is impractical for large data sets.

Address for correspondence: Michael L. Stein, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, IL 60637, USA.
E-mail: stein@galton.uchicago.edu

 1369–7412/04/66275

Vecchia (1988) suggested a simple approximation to the likelihood for spatial data based on the fact that any joint density can be written as a product of conditional densities. Using p to indicate a generic density (possibly conditional), suppose that $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ has joint density $p(\mathbf{z}; \phi)$, where ϕ is an unknown vector-valued parameter. Partition $\mathbf{Z}$ into subvectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_b$ of possibly different lengths and define $\mathbf{Z}'_{(j)} = (\mathbf{Z}'_1 \ldots \mathbf{Z}'_j)$. We always have

$$p(\mathbf{z}; \phi) = p(\mathbf{z}_1; \phi) \prod_{j=2}^{b} p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \phi). \tag{1}$$

Vecchia (1988) noted that it may not be critical to condition on all components of $\mathbf{z}_{(j-1)}$ when calculating $p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \phi)$ and thereby reduce the computational effort. In particular, if, for $j = 1, \ldots, b-1$, $\mathbf{S}_{(j)}$ is some subvector of $\mathbf{Z}_{(j)}$, then

$$p(\mathbf{z}; \phi) \approx p(\mathbf{z}_1; \phi) \prod_{j=2}^{b} p(\mathbf{z}_j | \mathbf{s}_{(j-1)}; \phi)$$

is the general form for Vecchia's approximation to the likelihood. We shall call $\mathbf{Z}_j$ the jth prediction vector and $\mathbf{S}_{(j-1)}$ the corresponding conditioning vector. Vecchia (1988) only considered prediction vectors of length 1, so $\mathbf{Z}_j = Z_j$, but we shall find it useful to allow longer prediction vectors. Pardo-Igúzquiza and Dowd (1997) described software that uses Vecchia's approach to approximating likelihoods for Gaussian random fields with certain covariance structures. Eide *et al.* (2002), section 7, used an approximation of this form as one step of an implementation of a Markov chain Monte Carlo algorithm in a Bayesian analysis of seismic data.

Throughout this work we shall suppose that $\mathbf{Z} \sim N\{F\boldsymbol{\beta}, K(\boldsymbol{\theta})\}$, where F is a known $n \times p$ matrix of rank p, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown coefficients and $\boldsymbol{\theta} \in \Theta$ is a vector of length q of unknown parameters for the covariance matrix of $\mathbf{Z}$, so that $\phi = (\boldsymbol{\beta}, \boldsymbol{\theta})$. For estimating $\boldsymbol{\theta}$, maximum likelihood acts as if $\boldsymbol{\beta}$ were known and, hence, tends to underestimate the variation of the spatial process (Stein (1999), section 6.4). Restricted maximum likelihood (REML) avoids this problem by estimating $\boldsymbol{\theta}$ by using only contrasts, or linear combinations of the observations whose means do not depend on $\boldsymbol{\beta}$. Kitanidis (1983) was the first to suggest the use of REML to estimate parameters of a spatial covariance function. Further remarks comparing REML and maximum likelihood are made at the end of Section 2.

We show how the approach of Vecchia (1988) can be adapted to approximate the restricted likelihood for Gaussian observations. To motivate our approximation, first note that, for Gaussian $\mathbf{Z}$, $p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \phi)$ is the density of the error of the best linear predictor (BLP) of $\mathbf{Z}_j$ in terms of $\mathbf{Z}_{(j-1)}$ as a function of ϕ. Thus, Vecchia's approximation involves replacing the density of the error of the BLP of $\mathbf{Z}_j$ in terms of $\mathbf{Z}_{(j-1)}$ with the density of the error of the BLP of $\mathbf{Z}_j$ in terms of $\mathbf{S}_{(j-1)}$. Just as the full likelihood can be written in terms of the densities of errors of BLPs, the restricted likelihood can be written in terms of the densities of errors of best linear unbiased predictors (BLUPs). The BLUP of, say, Z_1 given some subvector $\mathbf{S}$ of $\mathbf{Z}$ not containing Z_1 is just the linear combination $\boldsymbol{\lambda}'\mathbf{S}$ that minimizes the variance of $Z_1 - \boldsymbol{\lambda}'\mathbf{S}$ subject to the constraint that the mean of $Z_1 - \boldsymbol{\lambda}'\mathbf{S}$ is 0 for all values of $\boldsymbol{\beta}$. Proposition 1 in Section 2 shows that the restricted likelihood can be computed by using an expression that is similar to equation (1). Specifically, if we replace $p(\mathbf{z}_1; \phi)$ by the joint density of a linearly independent set of contrasts of $\mathbf{Z}_1$ and replace $p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \phi)$ by the joint density of the error of the BLUP of $\mathbf{Z}_j$ in terms of $\mathbf{Z}_{(j-1)}$, we obtain the restricted likelihood, which only depends on ϕ through $\boldsymbol{\theta}$. We thus obtain a natural analogue to Vecchia's approximation to the likelihood by considering the density of the error of the BLUP of $\mathbf{Z}_j$ in terms of $\mathbf{S}_{(j-1)}$ rather than $\mathbf{Z}_{(j-1)}$.

To implement this approximation, we need to order the observations in some manner and to choose the prediction and conditioning vectors. We concur with Vecchia (1988) that the ordering of the observations is not crucial and we use simple orderings such as the rank of the projections of the observation locations along some axis. Vecchia (1988) recommended choosing $\mathbf{S}_{(j-1)}$ to be made up of the m observations that are nearest to Z_j. This choice has the considerable virtue of simplicity. Furthermore, as Vecchia (1988) pointed out, a more statistically relevant criterion for choosing $\mathbf{S}_{(j-1)}$ will generally depend on the unknown parameters, making it difficult to find conditioning vectors that work well throughout the parameter space. Nevertheless, we shall argue that the gain in efficiency in choosing conditioning vectors that contain some distant observations can be sufficiently great in many circumstances to make such a choice worthwhile.

Jones and Zhang (1997) discussed applications of Vecchia's approach to space–time processes. Because of the incommensurability of space and time, they suggested defining nearest neighbours by the strength of correlation based on some preliminary estimate of the space–time correlation function. This choice would not remedy any of the problems that we find with Vecchia's approach in Sections 3 and 4 since the correlation functions that we consider are all monotone functions of distance. However, it does point out the difficulty in choosing conditioning sets when different co-ordinates are not commensurable. We face this problem in our application in Section 6 in which the variations in the horizontal dimension are fundamentally different from those in the vertical dimension.

Lark (2000) noted that there have not been any efforts to date to evaluate the effect of Vecchia's approximation to the likelihood on the resulting parameter estimates. We shall show in Section 2 that an estimating equations approach provides a natural way to evaluate the approximation. Specifically, by setting the derivatives of the approximate restricted likelihood with respect to the components of $\boldsymbol{\theta}$ to 0, we obtain a set of unbiased estimating equations for $\boldsymbol{\theta}$. We can then use the robust information criterion (Heyde (1997), page 12) to assess the statistical efficiency of various choices for the prediction and conditioning vectors. However, for sufficiently large data sets, even evaluating this information measure will become prohibitive and, in this case, we propose in Section 2 to approximate it by a sampling approach.

Section 3 considers two simple examples in which it is possible to contrast the following three criteria for choosing conditioning vectors of a given length when the prediction vector is a scalar: choosing the nearest neighbours to the predictand (the quantity being predicted), choosing the conditioning vector that minimizes the prediction error variance and choosing the conditioning vector that is best for estimating the unknown parameters. The first criterion always gives the same conditioning vectors independently of the true values of the parameters, but the second and third generally do not. However, for the first example in Section 3, the prediction error criterion chooses the same conditioning vector irrespectively of the true parameter values, and this choice is not made up of the nearest neighbours. The second example shows that the prediction error criterion can pick the same conditioning vectors irrespectively of the parameter values but that this choice of conditioning vectors is poor for parameter estimation.

Section 4 compares the efficiency of different ways for choosing conditioning and prediction vectors by using the information measure of the resulting estimating equations. All the examples are based on 1000 irregularly sited observations in a square region. The conditioning vectors are of length $m = 8$, 16, 32 and have either all, three-quarters or half of their components chosen to be the nearest neighbours to the prediction vectors and the remaining components more distant observations. In many circumstances, choosing some components of the conditioning vectors not to be nearest neighbours can lead to dramatic improvements in the efficiencies of the resulting estimators. For $m = 8$, the reverse is sometimes true but, for $m = 32$, choosing some distant observations is nearly uniformly superior to choosing only nearest neighbours. For stationary

covariance functions, the nearest neighbour designs tend to perform better when the spatial correlations are weakest, but further generalizations are difficult to make. In practice, we may want to obtain preliminary parameter estimates before making a final choice of conditioning vectors.

Section 5 discusses computational issues in implementing our procedure. We argue that conditioning sets that are twice the size of the prediction sets may be a good choice in some circumstances, suggesting that prediction sets with more than one observation will often be desirable.

Section 6 describes an application of our approximate restricted likelihood method to a data set of over 13 000 measurements of levels of chlorophyll in Lake Michigan. The data are taken in an irregular saw-tooth-like pattern and the variations in the chlorophyll levels in the horizontal and vertical dimensions are distinctly different, providing interesting challenges in the choice of conditioning sets. We find that including some distant observations in the conditioning sets leads to a dramatic improvement in the efficiency of some of the parameter estimates.

Section 7 discusses some computational and inferential issues arising from this work that deserve further attention.

2. Methodology

This section describes how the restricted likelihood can be written in a form that is analogous to equation (1) by using BLUPs, which then leads to an approximation of the restricted likelihood that is similar to Vecchia's (1988) for the full likelihood. We show that the derivatives with respect to the unknown covariance parameters of this approximation yield unbiased estimating equations for these parameters. Furthermore, the estimating equations framework provides a natural way of assessing the efficacy of the approximation.

We shall assume throughout this work that the covariance matrix $K(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \Theta$. Let $\mathbf{Z}_i$ have length n_i and take F_i to be the corresponding n_i rows of F so that $E(\mathbf{Z}_i) = F_i\boldsymbol{\beta}$. Assume that $\text{rank}(F_1) = p$ and define $n_{(j)} = n_1 + \ldots + n_j$. For $j > 1$, the BLUP of $\mathbf{Z}_j$ in terms of $\mathbf{Z}_{(j-1)}$ exists for all $\boldsymbol{\theta} \in \Theta$.

As a function of $\boldsymbol{\theta}$, for $j > 1$, let $B_j(\boldsymbol{\theta})$ be the $n_j \times n$ matrix such that $\mathbf{W}_j(\boldsymbol{\theta}) = B_j(\boldsymbol{\theta})\mathbf{Z}$ is the vector of errors of the BLUP of $\mathbf{Z}_j$ based on $\mathbf{Z}_{(j-1)}$. Thus, the last $n - n_{(j-1)}$ columns of $B_j(\boldsymbol{\theta})$ equal $(I_{n_j}\ O)$, where I_{n_j} is the identity matrix of order n_j and O is a matrix of 0s. For $j = 1$, take $B_1(\boldsymbol{\theta})$ to be a fixed matrix (independent of $\boldsymbol{\theta}$) of size $(n_1 - p) \times n$ with rank $n_1 - p$ such that $\mathbf{W}_1(\boldsymbol{\theta}) = B_1(\boldsymbol{\theta})\mathbf{Z}$ is a set of contrasts of $\mathbf{Z}_1$. Set $B(\boldsymbol{\theta})' = (B_1(\boldsymbol{\theta})' \ldots B_b(\boldsymbol{\theta})')$ and $\mathbf{W}(\boldsymbol{\theta})' = (\mathbf{W}_1(\boldsymbol{\theta})' \ldots \mathbf{W}_b(\boldsymbol{\theta})')$. Then $\mathbf{W}_j(\boldsymbol{\theta}) \sim N\{\mathbf{0}, V_j(\boldsymbol{\theta})\}$ where

$$V_j(\boldsymbol{\theta}) = B_j(\boldsymbol{\theta})\ K(\boldsymbol{\theta})\ B_j(\boldsymbol{\theta})'.$$

From now on, we shall suppress the dependence of V_j and $\mathbf{W}_j$ on $\boldsymbol{\theta}$ when this will not lead to confusion. It turns out that $\mathbf{W}_1, \ldots, \mathbf{W}_b$ are independent, leading to the following result (see Appendix A.1 for a proof).

Proposition 1. The restricted log-likelihood of $\boldsymbol{\theta}$ in terms of $\mathbf{Z}$ is given by

$$\text{rl}(\boldsymbol{\theta}; \mathbf{Z}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{b}[\log\{\det(V_j)\} + \mathbf{W}_j' V_j^{-1}\mathbf{W}_j].$$

Now consider approximating the restricted likelihood by computing the BLUP of $\mathbf{Z}_j$ in terms of just some subvector of $\mathbf{Z}_{(j-1)}$, taking care to make sure that a linear unbiased predictor of $\mathbf{Z}_j$ based on this subvector exists. For $j > 1$, the conditioning vector $\mathbf{S}_{(j-1)}$ is now the subvector of $\mathbf{Z}_{(j-1)}$ on which the BLUP of $\mathbf{Z}_j$ is based. Let S be the collection of subvectors $\mathbf{S}_{(1)}, \ldots, \mathbf{S}_{(b-1)}$.

Define $\mathbf{W}_1(S) = \mathbf{W}_1$ and, for $j > 1$, $\mathbf{W}_j(S)$ is the error of the BLUP of $\mathbf{Z}_j$ based on $\mathbf{S}_{(j-1)}$. Let $V_j(S)$ be the covariance matrix of $\mathbf{W}_j(S)$. Consider approximations to $\mathrm{rl}(\boldsymbol{\theta}; \mathbf{Z})$ of the form

$$\mathrm{rl}(\boldsymbol{\theta}; S) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{b}(\log[\det\{V_j(S)\}] + \mathbf{W}_j(S)'\, V_j(S)^{-1}\, \mathbf{W}_j(S)). \tag{2}$$

This approximation includes two innovations beyond what Vecchia (1988) proposed. The first is that it applies to the restricted likelihood and not the full likelihood, which depends on both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The second is that Vecchia (1988) considered only the special case in which each prediction vector has a single observation.

We can use equation (2) to define a set of unbiased estimating equations for $\boldsymbol{\theta}$. Writing ∂_k for $\partial/\partial\theta_k$ and defining $g_k(S) = \partial_k \mathrm{rl}(\boldsymbol{\theta}; S)$, we have

$$\begin{aligned} g_k(S) = -\frac{1}{2}\sum_{j=1}^{b}[&\mathrm{tr}\{V_j(S)^{-1}\,\partial_k V_j(S)\} + 2\,\mathbf{W}_j(S)'\, V_j(S)^{-1}\,\partial_k \mathbf{W}_j(S) \\ &- \mathbf{W}_j(S)'\, V_j(S)^{-1}\{\partial_k V_j(S)\}\, V_j(S)^{-1}\,\mathbf{W}_j(S)] \end{aligned} \tag{3}$$

and $\mathbf{G}(S) = (g_1(S) \ldots g_q(S))'$. Appendix B gives explicit expressions for $\partial_k V_j(S)$ and $\partial_k \mathbf{W}_j(S)$. Then $\mathbf{G}(S)$ is an estimating function for $\boldsymbol{\theta}$ and $\mathbf{G}(S) = \mathbf{0}$ an estimating equation. Now, for any $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ in Θ, for $j > 1$, $\mathbf{W}_j(\boldsymbol{\theta}_0; S) - \mathbf{W}_j(\boldsymbol{\theta}_1; S)$ is a contrast depending only on $\mathbf{S}_{(j-1)}$, so $\partial_k \mathbf{W}_j(S)$ is also a contrast of $\mathbf{S}_{(j-1)}$ and, hence, $\partial_k \mathbf{W}_j(S)$ and $\mathbf{W}_j(S)$ are independent under $\boldsymbol{\theta}$ since the error of a BLUP is independent of all contrasts of the observations. Furthermore, $\mathbf{W}_1(S)$ does not depend on $\boldsymbol{\theta}$, so $\partial_k \mathbf{W}_1(S) = \mathbf{0}$ and is trivially independent of $\mathbf{W}_1(S)$. It follows from equation (3) that $E\{\mathbf{G}(S)\} = \mathbf{0}$ for all $\boldsymbol{\theta} \in \Theta$ and $\mathbf{G}(S) = \mathbf{0}$ is an unbiased estimating equation for $\boldsymbol{\theta}$.

We can now use the well-developed theory of estimating equations to study the properties of solutions to $\mathbf{G}(S) = \mathbf{0}$, providing us with a natural way of investigating the effectiveness of various choices for S. As in Heyde (1997), define $\dot{\mathbf{G}}(S)$ to be the $q \times q$ matrix with (k, l)th element $\partial g_k(S)/\partial\theta_l$. The robust information measure about $\boldsymbol{\theta}$ in the estimating function $\mathbf{G}(S)$ is

$$\mathcal{E}\{\mathbf{G}(S)\} = (E\{\dot{\mathbf{G}}(S)\})'[E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}]^{-1} E\{\dot{\mathbf{G}}(S)\}$$

(Heyde (1997), page 12). A basic goal in estimating equations is to make $\mathcal{E}\{\mathbf{G}(S)\}$ large in the partial ordering of non-negative definite matrices. If there is no restriction on S, then we can set $\mathbf{S}_{(j)} = \mathbf{Z}_{(j)}$ for $j = 1, \ldots, b-1$, in which case $\mathcal{E}\{\mathbf{G}(S)\}$ is just the Fisher information matrix for the contrasts, which is the maximizer of $\mathcal{E}\{\mathbf{G}(S)\}$ under the partial ordering.

When n is too large to make calculating $\mathrm{rl}(\boldsymbol{\theta}; \mathbf{Z})$ feasible, we may then seek to make $\mathcal{E}\{\mathbf{G}(S)\}$ large subject to, for example, some bounds on the lengths of the prediction and conditioning vectors. As in Vecchia (1988), we seek simple rules that yield good results. Vecchia only considered $\mathbf{Z}_j$s of length 1 and m fairly small (at most 10 in his examples). For $j > m$, he let $\mathbf{S}_{(j-1)}$ be the m nearest neighbours to $\mathbf{Z}_j$ among $\mathbf{Z}_{(j-1)}$. Vecchia (1988) showed that this choice yields good approximations to the likelihood in some circumstances, but in his examples the spatial correlations are only non-negligible at distances that are much smaller than the diameter of the observation region. Examples in Section 4 show that there can sometimes be a considerable advantage in selecting some of the observations in the conditioning vectors to be rather distant from the observations in the prediction vectors.

Let us now consider evaluating $\mathcal{E}\{\mathbf{G}(S)\}$. We suppress the dependence of V_j and $\mathbf{W}_j$ on S as well as on $\boldsymbol{\theta}$. Again using the independence of $\mathbf{W}_j$ with any contrast of $\mathbf{S}_{(j-1)}$ yields

$$E\{\partial_l g_k(S)\} = -\sum_{j=1}^{b}[\tfrac{1}{2}\mathrm{tr}\{V_j^{-1}(\partial_l V_j)\, V_j^{-1}(\partial_k V_j)\} + \mathrm{tr}\{V_j^{-1}\,\mathrm{cov}(\partial_k \mathbf{W}_j, \partial_l \mathbf{W}_j')\}]. \quad (4)$$

Furthermore,

$$\begin{aligned} E\{g_k(S)\, g_l(S)\} = \sum_{i,j=1}^{b} & [\mathrm{tr}\{V_i^{-1}\,\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j')V_j^{-1}\,\mathrm{cov}(\partial_l \mathbf{W}_j, \partial_k \mathbf{W}_i')\} \\ & + \mathrm{tr}\{V_i^{-1}\,\mathrm{cov}(\mathbf{W}_i, \partial_l \mathbf{W}_j')V_j^{-1}\,\mathrm{cov}(\mathbf{W}_j, \partial_k \mathbf{W}_i')\} \\ & - \mathrm{tr}\{V_i^{-1}\,\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j')V_j^{-1}(\partial_l V_j)V_j^{-1}\,\mathrm{cov}(\mathbf{W}_j, \partial_k \mathbf{W}_i')\} \\ & - \mathrm{tr}\{V_j^{-1}\,\mathrm{cov}(\mathbf{W}_j, \mathbf{W}_i')V_i^{-1}(\partial_k V_i)V_i^{-1}\,\mathrm{cov}(\mathbf{W}_i, \partial_l \mathbf{W}_j')\} \\ & + \tfrac{1}{2}\mathrm{tr}\{V_i^{-1}(\partial_k V_i)V_i^{-1}\,\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j')V_j^{-1}(\partial_l V_j)V_j^{-1}\,\mathrm{cov}(\mathbf{W}_j, \mathbf{W}_i')\}], \end{aligned} \quad (5)$$

by repeated application of

$$\begin{aligned} \mathrm{cov}(\mathbf{X}_1' A_1 \mathbf{Y}_1, \mathbf{X}_2' A_2 \mathbf{Y}_2) = & \mathrm{tr}\{A_1\,\mathrm{cov}(\mathbf{Y}_1, \mathbf{Y}_2')A_2'\,\mathrm{cov}(\mathbf{X}_2, \mathbf{X}_1')\} \\ & + \mathrm{tr}\{A_1\,\mathrm{cov}(\mathbf{Y}_1, \mathbf{X}_2')A_2\,\mathrm{cov}(\mathbf{Y}_2, \mathbf{X}_1')\} \end{aligned}$$

for $(\mathbf{X}_1', \mathbf{Y}_1', \mathbf{X}_2', \mathbf{Y}_2')$ multivariate normal with mean $\mathbf{0}$ and A_1 and A_2 fixed matrices of appropriate order. Equation (5) greatly simplifies when $\mathbf{S}_{(j)} = \mathbf{Z}_{(j)}$ for $j = 1, \ldots, b-1$ so that $\mathbf{G}(S)$ is the score function. In this case, when $i < j$, since $\mathbf{W}_i$ and $\partial_k \mathbf{W}_i$ are contrasts that are functions of $\mathbf{S}_{(j)}$, then $\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j') = \mathrm{cov}(\partial_k \mathbf{W}_i, \mathbf{W}_j') = O$. Thus, $E\{g_k(S)\, g_l(S)\}$ reduces to $-E\{\partial_l g_k(S)\}$ and $\mathcal{E}\{\mathbf{G}(S)\} = E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$, the Fisher information matrix of the contrasts.

The matrix $\mathcal{E}\{\mathbf{G}(S)\}$ is valuable not just as a measure of information about the estimating function $\mathbf{G}(S)$ but also for inferential purposes. Let $\hat{\boldsymbol{\theta}}(S)$ be a solution to $\mathbf{G}(S) = \mathbf{0}$. In many circumstances, for sufficiently large sample sizes, $\hat{\boldsymbol{\theta}}(S)$ will be approximately normally distributed with mean $\boldsymbol{\theta}$ and covariance matrix $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ (Heyde (1997), chapter 4). The available theorems do not apply to the present circumstances of irregularly sited spatial data, but it is at least plausible that, if the diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ are all sufficiently small, $\hat{\boldsymbol{\theta}}(S)$ should be approximately $N[\boldsymbol{\theta}, \mathcal{E}\{\mathbf{G}(S)\}^{-1}]$ for observations from a Gaussian random field.

For n sufficiently large, the exact calculation of $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$ via equation (3) will not be feasible. One possible approximation would be to ignore the terms $i \neq j$ in equation (5), since the whole idea in choosing S is to approximate the setting in which $\mathbf{S}_{(j)} = \mathbf{Z}_{(j)}$ for all j, in which case the $i \neq j$ terms are all 0. This corresponds to acting as if $\mathbf{G}(S)$ is the true score function. However, as we describe in Section 4.3, this can lead to highly overoptimistic approximations to $\mathcal{E}\{\mathbf{G}(S)\}$.

A better approach is to sample randomly from among the $i \neq j$ terms and to use sampling theory to obtain an unbiased estimator for $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$. We found the following simple stratified sampling scheme to be often effective (see Section 4.3). For each $i = 1, \ldots, b$, randomly select without replacement some small number $j_1(i), \ldots, j_r(i)$ out of $1, \ldots, i-1, i+1, \ldots, b$. Defining $u_{kl}(i, j)$ to be the summand in equation (5), for $k, l = 1, \ldots, q$, estimate $E\{g_k(S)\, g_l(S)\}$ by

$$\sum_{i=1}^{b} u_{kl}(i, i) + \frac{b-1}{r}\sum_{i=1}^{b}\sum_{t=1}^{r} u_{kl}\{i, j_t(i)\}. \quad (6)$$

The ideas on how to choose prediction and conditioning sets and the use of estimating equations to evaluate the efficiency of designs can be applied to maximum likelihood as well as to REML. However, REML has several advantages over maximum likelihood. First, in the time series setting, both theoretical (Kang *et al.*, 2003) and simulation studies (Wilson, 1988; McGilchrist, 1989; Tunnicliffe Wilson, 1989) have shown the superior statistical properties of

REML, although it would require further work to verify that these results carry over to the approximate likelihoods that are considered here. Nevertheless, if p is small and n is very large, there is unlikely to be much difference between the maximum likelihood and REML estimates. As with exact REML, the results in this section only require that we specify the variances of contrasts. Thus, in some of our examples in Section 4 and the application in Section 6, we specify only the variogram $\frac{1}{2}\operatorname{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\}$. As long as the mean function includes an unknown constant term, the variogram is sufficient for computing the approximate restricted likelihood. Specifically, wherever $\operatorname{cov}\{Z(\mathbf{x}), Z(\mathbf{y})\}$ is required in a formula, replacing it by $-\frac{1}{2}\operatorname{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\}$ yields the result desired. In contrast, neither the full likelihood nor Vecchia's approximation to it is even defined when only the variogram is given. Finally, computing the (exact or approximate) REML estimate may be slightly easier in practice because numerical maximization of the full likelihood generally requires iterating between estimating $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ (Mardia and Marshall, 1984); in contrast, REML just maximizes over $\boldsymbol{\theta}$ and then, if required, estimates $\boldsymbol{\beta}$ explicitly in terms of the REML estimate of $\boldsymbol{\theta}$.

3. Nearest neighbours, prediction and estimation

When all the $\mathbf{Z}_j$s have length 1, the critical aspect of the methodology of the previous section is the choice of the conditioning vectors $\mathbf{S}_{(1)}, \ldots, \mathbf{S}_{(b-1)}$. We might plausibly imagine that, subject to a size constraint on $\mathbf{S}_{(j-1)}$, selecting the vector to make $V_j(\boldsymbol{\theta}, S)$ as small as possible would be a good idea. However, as Vecchia (1988), page 301, noted, the minimizing vector will generally depend on $\boldsymbol{\theta}$. It is for this reason that Vecchia recommended selecting the vector by using those observations that are closest to the predictand in Euclidean or some other distance. Furthermore, even if we could pick the conditioning vectors to minimize the prediction variance, such designs are not necessarily good choices for purposes of estimation. In this section we describe two simple examples to demonstrate potential conflicts between nearest neighbour designs, designs that are good for prediction and designs that are good for estimation.

Our first example shows that nearest neighbour designs can be uniformly suboptimal for purposes of prediction. Suppose that Z is a fractional Brownian motion: a Gaussian process on the real line with unknown constant mean and, for all x and y real, $\frac{1}{2}\operatorname{var}\{Z(x) - Z(y)\} = \theta_2|x - y|^{\theta_1}$ for some $\theta_2 > 0$ and $\theta_1 \in (0, 2)$. We have observations at -1, a and 0 with $-1 < a < 0$ and we wish to predict $Z(1)$, but we can only choose two of the observations with which to predict $Z(1)$.

Proposition 2. For any possible value for $\boldsymbol{\theta} = (\theta_1, \theta_2)$, the vector of length 2 that yields the BLUP with the minimum error variance is $(Z(-1), Z(0))$ (unless $\theta_1 = 1$, in which case $(Z(-1), Z(0))$ and $(Z(a), Z(0))$ both minimize the error variance).

Proof. When $\theta_1 = 1$, Z is Brownian motion and the result follows from the fact that the BLUP of $Z(1)$ based on all three observations is just $Z(0)$, so consider $\theta_1 \neq 1$. Let us first show that $(Z(-1), Z(0))$ is always better than $(Z(a), Z(0))$. The BLUP of $Z(1)$ based on $Z(0)$ is just $Z(0)$ and its mean-squared error is $2\theta_2$. If we add an observation at $-r$ with $r > 0$, then the relative reduction in the mean-squared error of the BLUP is $1 - \operatorname{corr}\{Z(1) - Z(0), Z(0) - Z(-r)\}^2$, where corr means correlation. Define

$$\rho_{\theta_1}(r) = \operatorname{corr}\{Z(1) - Z(0), Z(0) - Z(-r)\} = \tfrac{1}{2}(r^{1/2} + r^{-1/2})^{\theta_1} - \tfrac{1}{2}(r^{\theta_1/2} + r^{-\theta_1/2}).$$

Lemma 1. For all $\theta_1 \in (0, 2)$ except $\theta_1 = 1$, $\rho_{\theta_1}(r)^2$ achieves its unique maximum at $r = 1$, is strictly increasing on $(0, 1)$ and strictly decreasing on $(1, \infty)$.

The proof is given in Appendix A.2. Since minimizing the mean-squared error of the BLUP with respect to $r > 0$ is equivalent to maximizing $\rho_{\theta_1}(r)^2$, it follows that observing Z at -1 and 0 yields a smaller mean-squared error for $Z(1)$ than observing at a and 0 for all possible $\boldsymbol{\theta}$ with $\theta_1 \neq 1$.

Finally, let us prove that $(Z(-1), Z(a))$ is also inferior to $(Z(-1), Z(0))$ when $\theta_1 \neq 1$. For any set of observation and predictand locations, if we multiply all the interpoint distances by some positive constant u, the mean-squared error of the BLUP is multiplied by u^{θ_1}. Setting $u = 1/(1-a) < 1$ thus yields that $(Z(-1), Z(a))$ is inferior to $(Z\{-(1+a)/(1-a)\}, Z(0))$ for predicting $Z(1)$. Since $-1 < -(1+a)/(1-a) < 0$, lemma 1 implies that $(Z\{-(1+a)/(1-a)\}, Z(0))$ is in turn inferior to $(Z(-1), Z(0))$ and proposition 2 follows.

Thus, for a fractional Brownian motion Z with unknown mean and covariance parameter $\boldsymbol{\theta}$, if we want to predict $Z(1)$ by using $Z(0)$ and $Z(-r)$ for some $r > 0$, then the best choice for r is 1 irrespective of the value of $\boldsymbol{\theta}$. However, consider estimating $\boldsymbol{\theta}$ on the basis of $Z(1), Z(0)$ and $Z(-r)$ with $r > 0$ by using the restricted likelihood. Then $r = 1$, which gives three equally spaced observations, may not be a good choice since evenly spaced observation networks tend to work poorly for estimating covariance structures (Stein (1999), section 6.6, Pettitt and McBratney (1993), Lark (2002) and Zhu (2002)). This in fact appears to be so. Suppose that we assess the quality of the observation network by $I_r^{11}(\theta_1)$, the value of the first diagonal element of the inverse Fisher information matrix for the restricted likelihood, with smaller values indicating a better estimator. We focus on θ_1 because it is the more interesting parameter, but we would obtain the same result if we considered the determinant of the inverse Fisher information matrix, which equals $\theta_2^2 I_r^{11}(\theta_1)$. Using Stein (1999), equation (8) in chapter 6, and lengthy but straightforward calculations, it is possible to show that

$$I_r^{11}(\theta_1) = 4\{1 - \rho_{\theta_1}(r)^2\}^2 / [\log^2(r) + r^{-\theta_1} B\{B - A\log(r)\}]$$

where $A = (1+r)^{\theta_1} - 1 - r^{\theta_1}$ and $B = (1+r)^{\theta_1}\log(1+r) - r^{\theta_1}\log(r)$. Numerical evaluation of $I_r^{11}(\theta_1)$ viewed as a function of r for various θ_1-values strongly suggests that it is maximized by $r = 1$ for all $\theta_1 \in (0, 2)$, although we do not have a proof. What is easy to show is that for any fixed $\theta_1 \in (0, 2)$, as $r \to 0$ or as $r \to \infty$,

$$\log^2(r)\, I_r^{11}(\theta_1) / I_1^{11}(\theta_1) \to \log^2(2)(1 - 2^{\theta_1 - 2})^{-2},$$

so for any given θ_1 choosing r very small or very large is much better than choosing $r = 1$.

This example, in which we pick one subset of the observations and then estimate the parameters based on their exact restricted likelihood, is not an approximation of the form (2). Let us next consider a simple example in which a design S that is uniformly optimal for prediction is disastrous for estimation when using equation (2) to approximate the restricted likelihood. Suppose that Z is a stationary Gaussian process on the real line with unknown constant mean and $\mathrm{cov}\{Z(x), Z(y)\} = C\, K_\alpha(|x - y|)$, where C and α are both unknown, $C > 0$ and $\alpha \in I$, some open interval. Suppose further that, for every $\alpha \in I$, K_α is a positive and strictly decreasing function on $[0, \infty)$. We observe Z at $x_j = \Delta j$ for $j = 1, \ldots, n$. We wish to approximate the restricted likelihood for C and α by using prediction vectors and conditioning vectors of length 1, except for $\mathbf{Z}_1$, which has length 2. From the assumptions on K_α, it is obvious that, among $Z(\Delta), \ldots, Z\{\Delta(j-1)\}$, the one that gives the minimum mean-squared error for the BLUP of $Z(\Delta j)$ is $Z\{\Delta(j-1)\}$, irrespective of the value of C and α. However, it is apparent that it is not possible to estimate both C and α by using these conditioning and prediction sets. More specifically, we obtain from equation (2) that

$$\mathrm{rl}(\boldsymbol{\theta}; S) = -\frac{n-1}{2}\log(2\pi) - \frac{n-1}{2}\log[2\,C\{K_\alpha(0) - K_\alpha(\Delta)\}] - \frac{1}{2C\{K_\alpha(0) - K_\alpha(\Delta)\}}\sum_{j=2}^{n}[Z\{\Delta(j+1)\} - Z(\Delta j)]^2,$$

so, although it is possible to estimate the product $C\{K_\alpha(0) - K_\alpha(\Delta)\}$, it is not possible to estimate C and $K_\alpha(0) - K_\alpha(\Delta)$ separately. If we do not always use the most recent past observation as the conditioning vector, but sometimes use the second most recent past observation, then it will be possible to estimate both C and α.

4. Numerical results

4.1. Designs and models

There are innumerable models, observation networks, prediction vectors and conditioning vectors that we could choose to investigate the general approach to approximating restricted likelihoods described in Section 2. Here, we shall only consider one observation network: 1000 sites selected randomly out of the 10000 points in the plane (i, j), $i, j = 1, \ldots, 100$, and then each perturbed by adding a random point in $[-0.25, 0.25]^2$ (the exact locations can be found at `http://galton.uchicago.edu/~stein/approx-lik.html`). This sample size was chosen because, although it is fairly large, it is still possible to calculate the Fisher information for the exact restricted likelihood. The observations are irregularly sited but have a fixed minimum interobservation distance of 0.5, which avoids numerical difficulties as well as statistical problems that can arise with observations that are very close to each other when considering models with no measurement error.

We first consider $\mathbf{Z}_j$ having only one observation for $j > 1$ and $\mathbf{Z}_1$ having length $p+1$, so that $\mathbf{Z}_{(j)}$ has length $n_{(j)} = p + j$. We order the observations by the sum of their co-ordinates, i.e. from the lower left-hand corner to the upper right-hand corner of the observation region. We take conditioning vectors of constant size m (beyond the initial few). More specifically, for $j > m$, $\mathbf{S}_{(j-p-1)}$ (the conditioning set for $\mathbf{Z}_{j-p} = Z_j$) consists of the $m' < m$ nearest neighbours of $\mathbf{x}_j$ among $\{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}\}$ plus $m - m'$ additional points from $\{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}\}$ whose ranked distances to $\mathbf{x}_j$ equal $m + \lfloor l(j-m-1)/(m-m') \rfloor$ for $l = 1, \ldots, m-m'$. Taking $l = m - m'$, we see that we always select the most distant 'past' observation. Finally, for $j \leqslant m$, $\mathbf{S}_{(j-p-1)}$ is the entire past. We shall denote this design as $D(m, m')$. The design $D(m, 0)$ uses just the m nearest neighbours of $\mathbf{x}_j$ in the conditioning vector for $j > m$. Choosing points that are not nearest neighbours by ranks of distances without regard to directions leads to rather haphazard patterns in the locations of the observations in the conditioning vectors. We make no claims that this approach is nearly optimal in any sense. However, when we tried designs in which the distant past points were chosen more systematically, we obtained inferior results.

We also consider a scheme using prediction vectors obtained by dividing the observation region into approximately square cells containing nearly equal numbers of observations. For a positive integer J with $J^2 \leqslant n$, first divide the n observations into J strips based on the first co-ordinate of the observations so that each strip has nearly n/J observations. Specifically, after separately ordering the observations by the values of their first co-ordinate, for $j = 1, \ldots, J$, strip j contains observations from $\lfloor n(j-1)/J + 0.5 \rfloor + 1$ to $\lfloor nj/J + 0.5 \rfloor$, where $\lfloor x \rfloor$ is the largest integer that is no greater than x. Next, divide each of the J strips into J rectangles so that each rectangle has nearly n/J^2 observations. Specifically, after reranking the observations within each strip by the values of their second co-ordinate (so that observations from $\lfloor n(j-1)/J + 0.5 \rfloor + 1$ to $\lfloor nj/J + 0.5 \rfloor$ still make up strip j), the ith rectangle in the jth strip

contains observations from $\lfloor n\{(j-1)J+i-1\}/J^2+0.5\rfloor+1$ to $\lfloor n\{(j-1)J+i\}/J^2+0.5\rfloor$. The observations in the ith rectangle of strip j then make up the prediction vector $\mathbf{Z}_{(J-1)j+i}$. For a given $\mathbf{Z}_j$, define the distances between $\mathbf{Z}_j$ and a past observation as the minimum of the distances between that observation and those in $\mathbf{Z}_j$. For a positive integer $m > m'$, now pick $\mathbf{S}_{(j-1)}$ essentially as in the first scheme: if $n_{(j-1)}$ is less than m, then $\mathbf{S}_{(j-1)} = \mathbf{Z}_{(j-1)}$; otherwise, $\mathbf{S}_{(j-1)}$ is made up of the m' nearest neighbours together with those past points whose ranked distances to $\mathbf{Z}_j$ equal $m + \lfloor l(n_{(j-1)} - m)/(m-m')\rfloor$ for $l = 1, \ldots, m-m'$. Denote this design by $D_J(m, m')$. Here we consider designs with $J = 8$ so that the 1000 observations are partitioned into $b = 64$ subsets of size 15 or 16.

We shall consider two models for the covariance structures for the Gaussian random fields. The first is an exponential model, $\mathrm{cov}\{Z(\mathbf{x}), Z(\mathbf{y})\} = \theta_2 \exp(-\theta_1|\mathbf{x}-\mathbf{y}|/\theta_2)$, and the second a power law variogram model, $\frac{1}{2}\mathrm{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\} = \theta_2|\mathbf{x}-\mathbf{y}|^{\theta_1}$. We mostly take the mean function to be an unknown constant, but we do consider a mean function that is linear in the co-ordinates for the power law model. We also consider a power law model observed with measurement error of variance θ_3: $\frac{1}{2}\mathrm{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\} = \theta_3 \mathbf{1}_{\{|\mathbf{x}-\mathbf{y}|>0\}} + \theta_2|\mathbf{x}-\mathbf{y}|^{\theta_1}$. Except for the power law model, we report only on results for the designs $D(m, m')$.

4.2. Relative efficiencies

Let us first consider the results for the exponential model. The parameterization that is used here, $\theta_2 \exp(-\theta_1|\mathbf{x}-\mathbf{y}|/\theta_2)$, has the property that θ_1 describes the local variations of the process ($\frac{1}{2}\mathrm{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\} \sim \theta_1|\mathbf{x}-\mathbf{y}|$ as $|\mathbf{x}-\mathbf{y}| \to 0$), whereas θ_2 only substantially affects variations on scales that are not small compared with θ_2/θ_1. Table 1 gives the ratios (as percentages) of

Table 1. Relative efficiencies of estimators by using approximate restricted likelihoods compared with the exact REML estimators as measured by the diagonal elements of the inverse information matrices based on 1000 observations as described in Section 4.1†

θ_1	m'/m	*Relative efficiencies (%) for the following components and values of m:*					
		Component 1			*Component 2*		
		$m=8$	$m=16$	$m=32$	$m=8$	$m=16$	$m=32$
0.02	1	87.3	94.8	97.5	24.8	33.6	44.7
	0.75	91.1	97.4	99.2	71.0	83.1	90.5
	0.5	85.7	95.2	98.6	74.7	82.9	90.9
0.1	1	83.1	91.6	95.3	35.3	48.9	61.9
	0.75	89.0	96.6	99.0	58.6	78.4	89.8
	0.5	88.0	95.8	98.9	72.9	85.8	93.7
0.5	1	79.7	89.1	94.2	63.1	78.9	88.4
	0.75	77.6	89.3	94.4	58.2	78.2	88.8
	0.5	81.9	92.2	96.2	69.1	81.6	91.0
2	1	84.2	91.6	95.6	88.6	94.1	97.0
	0.75	81.8	89.8	94.6	83.9	92.1	96.0
	0.5	79.8	88.8	94.1	80.5	90.4	95.2

†The designs that are used for approximate likelihoods are of the form $D(m, m')$ as defined in Section 4.2. The model for the covariance function is $\theta_2 \exp(-\theta_1 d/\theta_2)$, where d is the interpoint distance; $\theta_2 = 1$ in all cases.

the diagonal elements of the inverse Fisher information matrix of the contrasts to the diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ for various designs and values for $\boldsymbol{\theta}$. These ratios are a measure of relative efficiency of the approximate likelihood to the full likelihood and, because the Fisher information matrix gives the maximum information, they must all be less than 1. We use $\theta_2 = 1$ in all cases and consider $\theta_1 = 0.02, 0.1, 0.5, 2$, the smaller values of θ_1 corresponding to more strongly correlated random fields. For θ_1, the efficiency of the various designs does not depend strongly on either the true value of θ_1 or m'/m, the fraction of the conditioning sets made up of nearest neighbours. For $m = 8$, all efficiencies are at least 77% and for $m = 32$ they are all at least 94%. In contrast, the efficiencies of the estimates of θ_2 depend strongly on the true value of θ_1 and m'/m. In particular, the designs with $m' = m$ are vastly inferior to the designs with $m' < m$ for $\theta_1 = 0.02$ or $\theta_1 = 0.1$, are competitive for $\theta_1 = 0.5$ and are slightly superior for $\theta_1 = 2$.

When θ_1 is small, there is not much information about θ_2 in the observations, despite their number. Thus, when $\theta_1 = 0.02$, the second diagonal element of the inverse Fisher information matrix is 0.552, so even the exact REML estimator of θ_2 will be highly variable. This lack of information follows from the fact that θ_2 has little effect on fluctuations of the random field at distances that are small compared with θ_2/θ_1, which equals 50 in this case. What little information is available is contained in the dependences at longer distances, so it is not surprising that approximating conditional densities by using only nearest neighbours works poorly. As Stein (1999) described, if we are only interested in spatial interpolation of the random field, the value of θ_2 will have little effect when θ_1 is small, so a severe loss of efficiency in estimating θ_2 may be tolerable. However, for estimating the unknown mean of the process and, more importantly, assessing the variance of an estimate of this mean, the value of θ_2 is critical.

The exponential model is a subclass of the three-parameter Matérn model for covariance functions (Stein, 1999). We have computed relative efficiencies of $D(m, m')$ designs for this model and obtained results that are qualitatively similar to those for the exponential model, with the relative efficiencies being greater for all three parameters for $m'/m = 0.75$ or $m'/m = 0.5$ and $m = 16$ or $m = 32$ in all the cases that we considered.

Results for the power law variogram model $\frac{1}{2}\,\mathrm{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\} = \theta_2|\mathbf{x} - \mathbf{y}|^{\theta_1}$ and two different mean functions are given in Table 2. We first consider the mean an unknown constant. For the smaller values for θ_1 of 0.6 and 1 and $m = 8$, $m' = m$ designs are best, whereas, for $m = 16$ and especially $m = 32$, designs with $m' < m$ are competitive or even superior, especially for $m'/m = 0.75$. For the larger values of θ_1 of 1.4 and 1.8, the results are dramatically different. First, all designs with $m = 8$ have poor relative efficiency, especially for estimating θ_1. However, in all cases, the designs with $m' = m$ are clearly the worst and, overall, designs with $m'/m = 0.75$ are slightly better than designs with $m'/m = 0.5$. For $\theta_1 = 1.8$, there is a striking similarity between the relative efficiencies for θ_1 and θ_2 that does not exist for the other θ_1-values. We do not have a good explanation for these similarities, which also occur in some of the other tables.

Before trying to explain the results for the power law model, it is essential to examine how they change if the mean function is taken to be an unknown linear polynomial in the coordinates. As Table 2 shows, many of the results that were found in the constant mean case are now reversed. In particular, the relative merits of designs with $m' = m$ to designs with $m' < m$ are now quite good for larger θ_1 but quite poor for smaller θ_1, especially regarding the relative efficiency of the estimators of θ_1. The changes in relative efficiency for the same design for the two different mean functions show dramatic differences depending on m'/m. For $m' = m$, the relative efficiencies are always smaller for $\theta_1 = 0.6$ or $\theta_1 = 1$ when the mean function is linear, most notably going from 84.1% to 38.4% for θ_1 when $m' = m = 8$ and $\theta_1 = 0.6$. In contrast, when

Table 2. Relative efficiencies of estimators by using approximate restricted likelihoods for the exact REML estimators†

θ_1	m'/m	*Relative efficiencies (%) for the following components and values of m:*					
		Component 1			*Component 2*		
		$m=8$	$m=16$	$m=32$	$m=8$	$m=16$	$m=32$
0.6	1	84.1/38.4	92.2/60.0	96.2/77.1	99.4/82.1	98.0/90.4	99.1/94.9
	0.75	76.1/64.8	90.1/86.1	96.6/93.8	91.2/87.6	97.2/96.6	99.2/98.8
	0.5	61.0/65.0	85.0/82.8	95.3/92.8	83.2/85.5	95.0/94.8	98.7/98.4
1	1	75.7/45.4	86.3/66.9	94.2/81.5	95.4/91.7	98.4/97.0	99.4/98.6
	0.75	72.0/59.7	86.7/84.4	96.1/93.8	92.4/89.8	97.9/97.8	99.5/99.5
	0.5	64.3/67.4	82.8/80.6	94.8/92.7	85.0/85.8	95.8/95.8	99.1/99.0
1.4	1	43.0/50.4	66.3/72.1	85.7/84.4	68.1/66.0	87.2/85.4	95.7/93.2
	0.75	51.2/51.2	80.5/81.0	94.9/93.4	73.1/68.6	91.9/90.3	97.9/96.9
	0.5	52.9/57.1	76.0/74.7	93.3/91.9	70.7/67.9	88.0/85.9	96.9/96.0
1.8	1	26.7/55.7	48.5/76.9	74.5/87.2	26.7/54.2	48.7/75.8	74.6/86.5
	0.75	35.1/45.1	75.2/78.9	92.8/92.8	35.5/44.2	74.9/78.0	92.6/92.4
	0.5	40.2/44.1	69.6/69.8	91.1/90.8	40.7/43.0	69.1/68.9	90.8/90.3

†See Table 1 for details. The covariance structure is defined by the variogram model $\frac{1}{2}\,\mathrm{var}\{Z(\mathbf{x})-Z(\mathbf{y})\}=\theta_2|\mathbf{x}-\mathbf{y}|^{\theta_1}$; $\theta_2=1$ in all cases. The first number is for a constant mean and the second for a first-order polynomial.

$\theta_1=1.8$ and $m'=m=8$, relative efficiencies for θ_1 and θ_2 both increase from about 25% to 55% when going from a constant to a linear mean function.

To investigate the efficiency of designs with prediction vectors of length greater than 1, we considered $D_8(m,m')$ with $m=16$ and $m=32$ for the power law covariance model and a constant mean. We do not consider $m=8$ since we do not recommend using conditioning vectors that are shorter than the prediction vectors. It is not apparent whether a $D_8(m,m')$ design should perform better or worse than a $D(m,m')$ design. The sizes of the conditioning vectors are the same in each case but, because the $D_8(m,m')$ design uses the correct joint conditional distribution for the observations within each prediction vector, the $D_8(m,m')$ design may have an advantage. In contrast, the conditioning vectors are chosen specially for each observation in $D(m,m')$, whereas they are required to be identical for all 15 or 16 observations in a prediction vector for $D_8(m,m')$, which may favour the $D(m,m')$ design. At least in the present circumstances, $D(m,m')$ and $D_8(m,m')$ perform remarkably similarly, so we do not give a separate table for the results for $D_8(m,m')$. Indeed, for the four values of θ_1 and six values of (m,m') ($m=16$ or $m=32$, $m'/m=1, 0.75, 0.5$), the relative efficiency of the $D_8(m,m')$ design is never more than 0.01 less than that of $D(m,m')$. For larger θ_1, $D_8(m,m')$ is often moderately superior (relative efficiency better by up to 0.1), especially for $m=16$ and $m'=16$ or $m'=8$.

To investigate the effect of measurement errors on the relative efficiencies of designs, we consider the power law model with the power θ_1 equal to 1 or 1.4, $\theta_2=1$ and θ_3, the measurement error variance, equal to 0.2 or 1. When $m=8$, the designs with $m'=8$ are usually superior, sometimes substantially so. When $m=32$, the designs with $m'=32$ are consistently inferior to designs with $m'=16$ or $m'=24$.

These numerical results show some important overall trends. First, the larger the value of m, the better the designs with $m'<m$ tend to perform relative to the designs with $m'=m$. Indeed, for $m=32$, among all the results presented here, the only case in which $D(32,32)$ is superior

to $D(32, 24)$ is for the exponential model with $\theta_1 = 2$. For this model, the correlations die out very quickly, so it is not surprising that conditioning on just nearest neighbours works well. However, even in this case, designs with $m' < m$ are only modestly worse than those with $m' = m$, so the relative efficiency of $D(32, 24)$ is within 0.01 of that for $D(32, 32)$ for both θ_1 and θ_2. In general, we find that models with stronger spatial correlations favour designs $D(m, m')$ with $m' < m$, although the different results for the power law model with the constant and linear mean functions are difficult to explain.

4.3. Approximating the information

If the approximation to the restricted likelihood in equation (2) is accurate, we might hope to replace the robust information measure $\mathcal{E}\{\mathbf{G}(S)\}$ with the simpler and much easier to compute $E\{\dot{\mathbf{G}}(S)\}$, whose components are given by equation (4). However, in all the cases that we have examined, this yields overoptimistic values for the information in the sense that the ratios of the diagonal elements of $E\{\dot{\mathbf{G}}(S)\}^{-1}$ to the corresponding diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ are less than 1. These ratios tend to track grossly the patterns that are shown in Tables 1–3 so that, whenever the estimates are badly inefficient, $E\{\dot{\mathbf{G}}(S)\}$ tends to be badly overoptimistic. The degree of overoptimism can sometimes be quite extreme: this ratio is 0.097 for θ_1 in the power law model with constant mean under the design $D(8, 8)$ when $\theta_1 = 1.8$. Thus, we do not recommend using $E\{\dot{\mathbf{G}}(S)\}$ to approximate $\mathcal{E}\{\mathbf{G}(S)\}$.

If a full calculation of $\mathcal{E}\{\mathbf{G}(S)\}$ is not feasible, we recommend using the sampling method that is given by expression (6) to approximate $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$ and hence $\mathcal{E}\{\mathbf{G}(S)\}$. We applied this approach in several cases with the design $D(8, 8)$ and $r = 3$ in expression (6) and found the results to be generally adequate. For example, consider the power law model with $\theta_1 = 1.8$ and a constant mean, for which $E\{\dot{\mathbf{G}}(S)\}^{-1}$ underestimates both diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ by about a factor of 10. We applied expression (6) 10 times with $r = 3$ to approximate $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$

Table 3. Relative efficiencies of estimators by using approximate restricted likelihoods for the exact REML estimators†

(θ_1, θ_3)	m'/m	*Relative efficiencies (%) for the following components and values of m:*								
		Component 1			*Component 2*			*Component 3*		
		$m = 8$	$m = 16$	$m = 32$	$m = 8$	$m = 16$	$m = 32$	$m = 8$	$m = 16$	$m = 32$
(1, 0.2)	1	70.7	80.5	89.0	75.9	86.6	92.6	78.6	87.9	93.3
	0.75	61.0	80.2	93.0	65.2	85.0	95.2	68.9	86.6	95.7
	0.5	46.8	74.5	91.3	49.5	78.6	93.7	54.5	81.0	94.4
(1, 1)	1	65.6	78.5	87.5	68.6	83.9	91.2	73.8	86.7	92.8
	0.75	54.9	77.0	91.8	58.2	81.1	94.1	64.6	84.3	95.2
	0.5	43.1	70.2	89.7	46.3	73.7	91.8	54.1	78.3	93.4
(1.4, 0.2)	1	49.8	61.1	78.2	82.6	90.7	95.4	71.4	78.0	88.3
	0.75	51.1	75.0	91.9	69.2	91.2	98.0	64.0	85.7	96.0
	0.5	42.8	71.4	90.2	46.6	84.5	96.6	48.2	81.2	94.8
(1.4, 1)	1	52.8	61.0	76.3	74.5	85.7	91.8	75.1	81.0	88.6
	0.75	47.7	73.1	90.9	55.4	85.3	96.5	59.8	84.8	95.9
	0.5	36.3	68.1	89.0	35.9	76.1	94.4	44.2	78.7	94.4

†See Table 1 for details. The model for the variogram is $\frac{1}{2}\,\mathrm{var}\{Z(\mathbf{x}) - Z(\mathbf{y})\} = \theta_3 \mathbf{1}\{\mathbf{x} \neq \mathbf{y}\} + \theta_2 |\mathbf{x} - \mathbf{y}|^{\theta_1}$; $\theta_2 = 1$ in all cases.

and hence $\mathcal{E}\{\mathbf{G}(S)\}$, and in all 10 cases the approximate values for the diagonals of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ were within 5% of the truth.

An attractive property of using sampling to approximate $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$ is that it is easy to check in practice whether the approximation is working well. Specifically, viewing the b blocks as the sampling units, we can calculate an empirical covariance matrix for the estimates of $E\{g_k(S)\,g_l(S)\}$ for $k,l=1,\ldots,q$ given by expression (6). Since $E\{\dot{\mathbf{G}}(S)\}$ is known exactly, the estimate of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ is a known linear transformation of the estimate of $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$. Thus, standard errors for the estimates of the elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ can be readily obtained.

4.4. Prediction versus estimation

The examples in Section 3 show that picking the conditioning vector to predict well under the true model is not necessarily a good idea when the goal is estimation. The discrepancy between these goals can also occur for larger observation networks. Consider using the same network of 1000 observations as in Section 4.2 and the designs $D(m,m')$, so that the prediction vectors are of length 1. Since V_j is the variance of the BLUP error given the entire past, $\log\{V_j(S)/V_j\}$ is non-negative and values nearer to 0 indicate predictions whose error variances are closer to the best possible. We shall use the average of these log-ratios as our measure of the quality of the predictions for a particular design and $\boldsymbol{\theta}$-value. Table 4 gives these average log-ratios under the model $\frac{1}{2}\operatorname{var}\{Z(\mathbf{x})-Z(\mathbf{y})\}=|\mathbf{x}-\mathbf{y}|^{1.4}$ and a constant mean. Increasing m decreases the average log-ratio, which is expected since using longer conditioning vectors should yield prediction variances that are closer to those obtained by conditioning on all past observations. More interestingly, these averages are smaller for larger m'/m, so choosing all nearest neighbours in the conditioning vectors yields the best predictions overall. In contrast, Table 2 shows that $m'/m=1$ yields the least efficient designs for estimating the parameters of the covariance function when $\theta_1=1.4$.

We see that, at least qualitatively, these results support those in Section 3. Specifically, including observations in the conditioning vector that give information about the random field over different spatial scales can be a good idea for parameter estimation even when it is a bad idea for prediction. Thus, in developing an intuition for selecting conditioning vectors to approximate likelihoods, it is not generally appropriate to think about what designs would give good predictions.

Table 4. Average log-ratio of error variances of approximate BLUPs to error variances of exact BLUPs by using designs $D(m,m')$ for observations 2–1000 under the model $\frac{1}{2}\operatorname{var}\{Z(\mathbf{x})-Z(\mathbf{y})\}=|\mathbf{x}-\mathbf{y}|^{1.4}$

m'/m	*Log-ratios (×100) for the following values of m:*		
	$m=8$	$m=16$	$m=32$
1	4.99	1.29	0.40
0.75	7.29	1.69	0.41
0.5	13.49	3.53	0.77

5. Computational considerations

This section considers the computational effort that is required to calculate the approximate likelihood in equation (2). We argue that a reasonable choice for the relative sizes of the conditioning sets and the prediction sets is that the prediction sets be approximately half the size of the conditioning sets.

For simplicity, let us assume that all the conditioning sets are of the same size c and all the prediction sets are of the same size $d = n/b$. In approximating the floating point operations that are needed to implement the algorithm, we shall assume that $c + d$ is large but small compared with n. To simplify the considerations further, we shall assume that the mean of Z is known to be 0 so that $p = 0$. When p is in fact positive, the additional computations that are required are negligible as long as p is much smaller than $c + d$.

Consider the computations that are necessary to find the contribution of the jth block to the approximate likelihood. Define

$$\mathrm{var}\begin{pmatrix} \mathbf{S}_{(j-1)} \\ \mathbf{Z}_j \end{pmatrix} = K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix},$$

suppressing the dependence of the covariance on $\boldsymbol{\theta}$ and j. We need to calculate the vector of errors of the BLUPs $\mathbf{W}_j = \mathbf{Z}_j - K_{21}K_{11}^{-1}\mathbf{S}_{(j-1)}$, its covariance matrix $V_j = K_{22} - K_{21}K_{11}^{-1}K_{12}$, the quadratic form $\mathbf{W}_j' V_j^{-1} \mathbf{W}_j$ and $\det(V_j)$. To obtain these quantities, first calculate the Cholesky decompostion GG' of K_{11}, which requires $\frac{1}{3}c^3$ floating point operations (Golub and van Loan (1996), page 144). Then compute $H = G^{-1}K_{12}$, which requires $c^2 d$ floating point operations (Golub and van Loan (1996), page 89). Calculating $\mathbf{W}_j = \mathbf{Z}_j - H'(G')^{-1}\mathbf{S}_{(j-1)}$ contributes negligibly to the overall effort. Now $K_{21}K_{11}^{-1}K_{12} = H'H$ requires cd^2 floating point operations by exploiting the symmetry of the resulting matrix; computing V_j is then a negligible effort. Next, computing the Cholesky decomposition of V_j requires $\frac{1}{3}d^3$ floating point operations and, from this, obtaining $\mathbf{W}_j' V_j^{-1} \mathbf{W}_j$ and $\det(V_j)$ is a negligible effort. All together, the number of floating point operations required is $\frac{1}{3}c^3 + c^2 d + cd^2 + \frac{1}{3}d^3 = \frac{1}{3}(c+d)^3$ plus lower order terms.

The figure $\frac{1}{3}(c+d)^3$ is the same as is required for the Cholesky decomposition of K. Indeed, it is possible to show that $\mathbf{W}_j' V_j^{-1} \mathbf{W}_j$ and $\det(V_j)$ can be computed in $\frac{1}{3}(c+d)^3$ floating point operations by a procedure in which the optimal predictor of each component of $\mathbf{Z}_j$ is obtained in terms of $\mathbf{S}_{(j-1)}$ and the previous components of $\mathbf{Z}_j$ by sequentially updating Cholesky decompositions.

Since there are $b = n/d$ blocks, the total number of floating point operations that are needed to compute the approximate likelihood in equation (2) is $(n/3d)(c+d)^3$. The results in Section 4.2 suggest that c may be more critical to the accuracy of the approximation than d. Thus, for any given c, it is plausible to choose d to minimize the amount of computation that is required. As a function of d, $(n/3d)(c+d)^3$ is minimized by $d = \frac{1}{2}c$, in which case the total number of floating point operations required is $(9/4)nc^2$. This compares with the $\frac{1}{3}n^3$ floating point operations that are needed to calculate the full likelihood. For the example that was reported on in Section 4 in which $d \approx 16$, $c = 32$ and $n = 1000$, we obtain $d \approx \frac{1}{2}c$, $(9/4)nc^2 \approx 2.3 \times 10^6$ and $\frac{1}{3}n^3 \approx 3.3 \times 10^8$, so the approximate likelihood requires less than 1% of the floating point operations of the full likelihood. For problems with larger n, the relative savings can be even more dramatic.

To compare the efficiency of choosing prediction sets of more than one observation to prediction sets of size 1 as in Vecchia (1988), consider $d = 16$ and $c = 32$ *versus* $d = 1$ and $c = 32$. The approximate floating point operation count for the first design is $2304n$ and for the second is $11\,979n$, so the design with $d = 1$ requires about five times the computation irrespective of n, at least when the computations are done separately for each block (see Section 7).

6. Application to Lake Michigan chlorophyll fluorescence

6.1. Introduction

Phytoplankton, the unicellular algae that are found primarily in oceans and lakes, are important elements of marine ecosystems and of the global carbon cycle. Water samples are often used to measure levels of phytoplankton, but sample sizes are inevitably small in such studies and hence provide limited information about the spatial variation in levels of phytoplankton. In contrast, chlorophyll fluorescence is roughly linearly related to the level of phytoplankton, can be measured *in situ* and may be recorded with high frequency (one observation per second, for instance) over large spatial scales.

The fluorescence profile that is used in this example was obtained in the lower basin of Lake Michigan in mid-March 2000 as part of the Episodic Events–Great Lakes Experiment. A fluorometer was towed in a continuously undulating fashion from the surface to lake bottom and back along a 25-km transect from offshore to near shore at the southern tip of the lake, providing over 13000 *in situ* measurements (Fig. 1). We order the observations by their collection times and when we refer to 'nearest neighbours' we define nearest in terms of this ordering.

6.2. Variogram model

Exploratory analysis suggested no obvious covariates to include in the mean function and that the logarithms of the fluorescence values are more nearly Gaussian than raw fluorescence values. Using h for the distance from the shore (in kilometres) and v for the depth (in metres), we therefore take our process $Z(h, v)$ of log(fluorescence) measurements to be Gaussian with an unknown constant mean and variogram indexed by a parameter $\boldsymbol{\theta}$. Welty and Stein (2003) describe exploratory analyses leading to the variogram model

$$\gamma(h, v) = \theta_0 \mathbf{1}_{\{h^2+v^2>0\}} + \theta_1 h^{\theta_2} + \theta_3[1 - \mathcal{M}_\nu\{(\theta_4^2 h^2 + \theta_5^2 v^2)^{1/2}\}]$$

where $\mathcal{M}_\nu(z) = 2^{1-\nu} z^\nu \mathcal{K}_\nu(z)/\Gamma(\nu)$ is the Matérn correlation function (so that $\mathcal{M}_\nu(0) = 1$) and $\mathcal{K}_\nu$ is a modified Bessel function of order ν (Abramowitz and Stegun, 1965). The term $\theta_1 h^{\theta_2}$ is included to capture the horizontal variations, which are fundamentally different from the vertical variations.

6.3. Conditioning and prediction set designs

The unusual pattern in measurement locations as well as the differences in the variations along the vertical and horizontal directions provide interesting challenges in picking prediction and

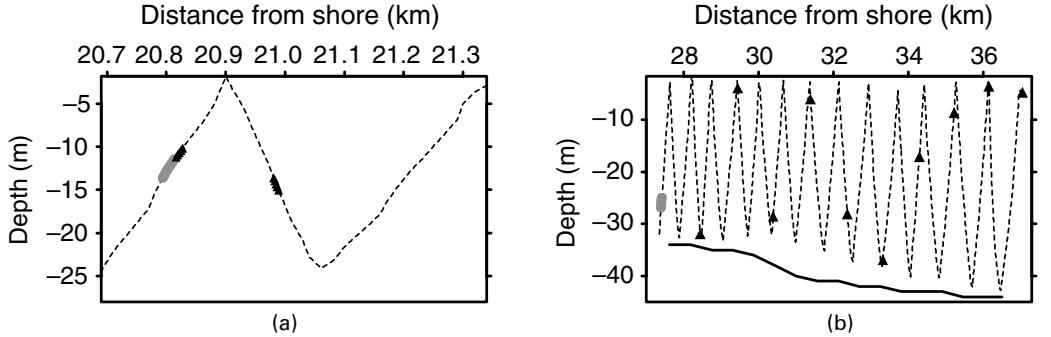

Fig. 1. Saw-tooth-like pattern of the chlorophyll fluorescence measurements: example conditioning set selection (-------; ●, prediction; ▲, conditioning) (a) for nearest neighbours and medium-range points and (b) for far away points (———, bottom of the lake)

conditioning vectors. We divided the data into contiguous blocks of size d for prediction vectors $\mathbf{Z}_j$. On the basis of the results in Section 5, we constructed conditioning vectors $\mathbf{S}_{(j-1)}$ to be size $2d$, twice the length of the prediction vectors.

For $d = 10$ and $d = 20$ we considered three schemes for picking conditioning sets. We let conditioning sets consist entirely of nearest neighbours, a mix of nearest neighbours and far away points, and a mix of nearest neighbours, far away points and medium range points. We let $D_{\mathrm{N}}(d)$, $D_{\mathrm{NF}}(d)$ and $D_{\mathrm{NMF}}(d)$ represent these respective schemes. For example, the conditioning set corresponding to prediction set $\mathbf{Z}_j$ for $D_{\mathrm{N}}(10)$ consists of $\mathbf{Z}_{j-1}$ and $\mathbf{Z}_{j-2}$. That for $D_{\mathrm{NF}}(10)$ consists of $\mathbf{Z}_{j-1}$ as well as one point each from 10 roughly equally spaced blocks from $\mathbf{Z}_{j-1}, \ldots, \mathbf{Z}_1$. That for $D_{\mathrm{NMF}}(10)$ consists of five points from $\mathbf{Z}_{j-1}$, five points from a nearby block of similar depth (see Fig. 1) and one point each from 10 roughly equally spaced blocks from the past. The data and more detailed descriptions of conditioning and prediction set designs are available at `http://galton.uchicago.edu/~stein/approx-lik.html`.

6.4. Parameter estimates

We maximized the approximate restricted log-likelihood analytically for a scale parameter $\alpha = \theta_0$ and then numerically maximized the remaining profiled log-likelihood over $(\theta_1/\alpha, \theta_2, \theta_3/\alpha, \theta_4, \theta_5)$ by using a conjugate gradient algorithm (Press *et al.*, 1992). Although there is no guarantee that the approximate likelihood has only a single mode, we have plotted the surface on grids over several pairs of parameters (with the remaining three parameters fixed) and found no signs of multiple modes. To simplify our initial approach, we did not maximize over ν, the smoothness parameter in the Matérn covariance function. After considering values 0.5, 1.0 and 1.5, we concluded that $\nu = 1.0$ appeared best on the basis of comparisons of the restricted log-likelihood and the shape of the empirical vertical variogram near $v = 0$. As in expression (6), stratified sampling was used to estimate $E\{\mathbf{G}(S)\,\mathbf{G}(S)'\}$, and standard errors were obtained for the elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$.

We found that prediction and conditioning vector designs did affect parameter estimates. Given the increased computation time for longer vectors, we recommend obtaining parameter estimates from designs with smaller values of d and then using these estimates as starting values for maximization with larger prediction sets. We used this strategy first to obtain estimates under $D_{\mathrm{NF}}(10)$, yielding $\hat{\boldsymbol{\theta}}_{\mathrm{NF}}(10) = (3.49 \times 10^{-4}, 3.00 \times 10^{-3}, 1.22, 1.47 \times 10^{-3}, 305, 3.28)$, and then using $\hat{\boldsymbol{\theta}}_{\mathrm{NF}}(10)$ as a starting value found the estimate under $D_{\mathrm{NMF}}(20)$ of $\hat{\boldsymbol{\theta}}_{\mathrm{NMF}}(20) = (3.50 \times 10^{-4}, 2.34 \times 10^{-3}, 1.06, 1.48 \times 10^{-3}, 306, 3.23)$. The discrepancies in estimates for θ_1 and θ_2 under the two designs are not unexpected given that these parameters describe behaviour in the horizontal direction and that the prediction–conditioning sets in the larger design contain significantly more pairs of points that are at nearly the same depth.

Designs that include some far away points in the conditioning vector and that have longer prediction–conditioning vectors yield estimates for $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ with noticeably smaller diagonal elements (Table 5), which suggests less variability in parameter estimates for these designs. The largest decreases in the square roots of the diagonal elements appear when schemes that use only nearest neighbours are modified to contain more distant points as well. There is a dramatic decrease in diagonal elements corresponding to the parameters that describe the long-range horizontal behaviour of the process, θ_1 and θ_2, when we consider $D_{\mathrm{NF}}(d)$ or $D_{\mathrm{NMF}}(d)$ over $D_{\mathrm{N}}(d)$, even though these designs have conditioning vectors of equal length. Schemes that include medium-range points as well as far away points yield smaller diagonal elements for some parameters, though these gains may be offset by slightly larger values for other parameters. $D_{\mathrm{NMF}}(10)$ yields smaller diagonal elements than $D_{\mathrm{NF}}(10)$, but $D_{\mathrm{NMF}}(20)$ yields a smaller value for θ_1, a larger value for θ_2 and similar values for the remaining parameters when compared with $D_{\mathrm{NF}}(20)$.

Table 5. Square roots of the diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ expressed as percentages of parameter values†

	Square roots (%) for the following parameters:					
	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5
$D_{\mathrm{N}}(10)$	8.54	322.91	120.56	4.17	12.63	13.97
$D_{\mathrm{NF}}(10)$	7.61	30.29	19.73	2.98	12.38	12.25
$D_{\mathrm{NMF}}(10)$	7.44	25.72	16.14	2.69	12.29	11.84
$D_{\mathrm{N}}(20)$	7.63	78.48	36.90	2.97	12.14	12.22
$D_{\mathrm{NF}}(20)$	7.39	26.18	12.30	2.68	11.96	11.73
$D_{\mathrm{NMF}}(20)$	7.39	23.22	13.09	2.69	11.96	11.73

†For each design, $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ was computed for $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathrm{NMF}}(20)$, with $E\{\mathbf{G}(S)\ \mathbf{G}(S)'\}$ estimated by stratified random sampling and $r=5$ as in expression (6).

Sampling standard errors for the diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ can be significant for small r. For example, for $r=5$ and design $D_{\mathrm{NMF}}(20)$, we calculated the standard errors for the diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ using basic sampling theory for stratified random samples. For the parameters other than θ_1 and θ_2, these standard errors are all less than 5% of the diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$, but they are about 12% and 14% for θ_1 and θ_2 respectively. The differences in the diagonal elements for $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ from one design to another (when both designs use the same sampling scheme) tend to be much less variable than the individual elements themselves, which is apparent from the nearly identical values for diagonal elements of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$ for parameters other than θ_1 and θ_2 and designs $D_{\mathrm{NF}}(20)$ and $D_{\mathrm{NMF}}(20)$.

We thus recommend using small values of r for comparing designs (with the same off-diagonal sampling $j_1(i),\ldots,j_r(i)$) based on differences in values for $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$. Once a good design has been identified (there may be many), we recommend increasing r to obtain the desired level of certainty in the estimates of $\mathcal{E}\{\mathbf{G}(S)\}^{-1}$. The most important aspect of design appears to be including far away points from the past; once some far away points have been included, improvements may be made by increasing the sizes of the prediction–conditioning sets and possibly to a smaller extent by adjusting point selection schemes to include medium-range points.

7. Discussion

We have just scratched the surface in terms of looking at the relationship between models, observation locations and good choices of prediction and conditioning sets. Certainly we would like to find less haphazard rules for picking the distant locations in the conditioning sets. Another critical issue is to understand better the effect of measurement errors on good designs. For example, we would expect that closely spaced observations make it easier to estimate the measurement error variance, which suggests that there may be an advantage in choosing the distant observations in the conditioning set in clumps, whereas such clumping would probably be inefficient when the measurement error variance is negligible.

When the observations are on a regular lattice, it should be possible to obtain analytic expressions for the asymptotic efficiencies of sufficiently simple designs. We recommend a fixed domain asymptotic approach (Stein, 1999) where feasible, to reflect the situation of strongly dependent neighbouring observations for which we expect using distant observations in conditioning sets to be most helpful. However, we would not advocate the use of our approach when observing

a stationary process on a lattice, but we would instead suggest the use of exact methods that exploit the block Toeplitz structure of the resulting covariance matrices or spectral approximations such as the Whittle likelihood (Whittle, 1954). Furthermore, the results would generally be quite messy for conditioning and prediction sets that are larger than a few observations and results for very small conditioning and prediction sets may not provide much insight.

Whether a design is efficient will depend somewhat on how the computations are done. For example, the calculations that are described here assume that the computations for different blocks are done separately. When there is substantial overlap between the conditioning sets of different blocks, this could be exploited to reduce the computations by, for example, calculating the Cholesky decomposition of the common observations in two or more conditioning sets just once. If algorithms were developed that exploited these overlaps, then designs with greater overlap between conditioning sets would presumably be favoured. Approximation (2) could be easily implemented on parallel processors. One could just parcel out the prediction and conditioning sets to the different processors, although there would be interesting issues in terms of balancing arithmetic effort against the need to pass information across processors. The details of the implementation could affect the relative computational efficiencies of various designs.

Finally, we mention an interesting inferential problem that is raised by the use of approximate likelihoods in Bayesian analyses. Where the approximation is very good, we could act as if the approximate likelihood were the actual likelihood without much harm. However, especially if the likelihood calculation is just a part of a single step in a Markov chain Monte Carlo algorithm, we may not be able to afford a highly accurate approximation, in which case the status of the resulting approximate posterior as a tool for inference would be dubious. It might be possible to use the robust information measure to adjust the approximate likelihood in some manner, but the details of such a method and demonstrating its effectiveness pose considerable challenges.

Acknowledgements

MLS was supported by National Science Foundation grant DMS 99-71127. The chlorophyll fluorescence measurements were provided by Henry Vanderploeg, Great Lakes Environmental Research Laboratory. The authors thank the referees and the Joint Editor for many helpful suggestions on the exposition of the paper.

Appendix A

A.1. Proof of proposition 1

We can choose an $(n-p)\times n$ matrix C of rank $n-p$ such that $CF=O$ and C is of the following block lower triangular form:

$$\begin{pmatrix} C_1 & O & & \\ & C_2 & O & \\ & & C_3 & O \\ & & \ddots & \ddots \\ & & & \\ & & C_b & \end{pmatrix},$$

where C_1 is $(n_1-p)\times n_1$ with rank n_1-p and, for $j>1$, C_j is $n_j\times n_{(j)}$ with its last n_j columns equal to I_{n_j}. Then $C\mathbf{Z}$ is a linearly independent basis for the space of contrasts of $\mathbf{Z}$ and, for $j>1$, $C_j\mathbf{Z}_{(j)}$ is the error of a linear unbiased predictor of $\mathbf{Z}_j$ based on $\mathbf{Z}_{(j-1)}$. Define $n'_{(j)}=n_{(j)}-p$.

We claim that $B(\boldsymbol{\theta})$ can be written in the form $D(\boldsymbol{\theta})C$, where $D(\boldsymbol{\theta})$ is an $(n-p)\times(n-p)$ matrix whose first n_1-p rows are $(I_{n_1-p}\ O)$ and, for $j>1$, rows from $n'_{(j-1)}+1$ to $n'_{(j)}$ are of the form $(D_j(\boldsymbol{\theta})\ O)$, where $D_j(\boldsymbol{\theta})$ is $n_j\times n'_{(j)}$ with last n_j columns equal to I_{n_j}. That $D_1(\boldsymbol{\theta})$ can be set to $(I_{n_1-p}\ O)$ is trivial since we can take $B_1(\boldsymbol{\theta})\equiv(C_1\ O)$. To prove that $B_j(\boldsymbol{\theta})$ can be written in the form $(D_j(\boldsymbol{\theta})C\ O)$ for $j>1$, we use the fact that the BLUP of $\mathbf{Z}_j$ can be expressed as a sum of any given linear unbiased predictor of $\mathbf{Z}_j$ plus a linear unbiased predictor of $\mathbf{0}$ (i.e. a contrast). Since $C_j\mathbf{Z}_{(j)}$ is the error of a linear unbiased predictor of $\mathbf{Z}_j$ based on $\mathbf{Z}_{(j-1)}$ and the first $n'_{(j)}$ rows of $C\mathbf{Z}$ form a basis for the contrasts of $\mathbf{Z}_{(j-1)}$, for given $\boldsymbol{\theta}$, the error of the BLUP of $\mathbf{Z}_j$ based on $\mathbf{Z}_{(j-1)}$ can be written in the form

$$\sum_{i=1}^{j-1}\Lambda_{ij}(\boldsymbol{\theta})C_i\mathbf{Z}_{(i)}+C_j\mathbf{Z}_{(j)},$$

where $\Lambda_{ij}(\boldsymbol{\theta})$ has dimension $n_j\times n_i$ for $i>1$ and dimension $n_j\times(n_1-p)$ for $i=1$. Setting $D_j(\boldsymbol{\theta})=(\Lambda_{1j}(\boldsymbol{\theta})\ldots\Lambda_{j-1,j}(\boldsymbol{\theta})\,I_{n_j})$ yields the results desired. Let us write $\mathrm{rl}(\boldsymbol{\theta};\mathbf{Z})$ for the log-likelihood of the contrasts, $C\mathbf{Z}$, which is independent of the choice of C (Christensen (1996), page 276). We have

$$\mathrm{rl}(\boldsymbol{\theta};\mathbf{Z})=-\frac{n-p}{2}\log(2\pi)-\frac{1}{2}\log[\det\{C\,K(\boldsymbol{\theta})C'\}]-\frac{1}{2}(C\mathbf{Z})'\{C\,K(\boldsymbol{\theta})C'\}^{-1}C\mathbf{Z}.$$

Using $B(\boldsymbol{\theta})=D(\boldsymbol{\theta})C$ and $\det\{D(\boldsymbol{\theta})\}\equiv 1$, it is easy to show that the joint density of $\mathbf{W}(\boldsymbol{\theta})$, viewed as a function of $\boldsymbol{\theta}$, also gives $\mathrm{rl}(\boldsymbol{\theta};\mathbf{Z})$ despite the fact that this transformation of $\mathbf{Z}$ depends on $\boldsymbol{\theta}$. Proposition 1 follows from the independence of $\mathbf{W}_1(\boldsymbol{\theta}),\ldots,\mathbf{W}_b(\boldsymbol{\theta})$, which is a consequence of the multivariate normality and the fact that the error of a BLUP is uncorrelated with all contrasts of the observations on which it is based.

A.2. Proof of lemma 1

First, because $\rho_\alpha(r)=\rho_\alpha(1/r)$, it suffices to show that $\rho_\alpha(r)^2$ achieves its unique maximum on $[1,\infty)$ at $r=1$, so only consider $r\geqslant 1$ from now on. A straightforward analysis shows that $\rho_\alpha(r)\to 0$ as $r\to\infty$ for all $\alpha\in(0,1)$. Furthermore, $\rho_\alpha(1)=2^{\alpha-1}-1$, which is negative for $\alpha\in(0,1)$ and positive for $\alpha\in(1,2)$. Thus, it suffices to show that ρ_α is strictly increasing on $[1,\infty)$ for $\alpha\in(0,1)$ and is strictly decreasing on $[1,\infty)$ for $\alpha\in(1,2)$. Let $2s=r^{1/2}+r^{-1/2}$, so that $r^{1/2}=s+(s^2-1)^{1/2}$ and s is a strictly increasing function of r from $[1,\infty)$ onto $[1,\infty)$. Define

$$F_\alpha(s)=2\,\rho_\alpha(r)=2^\alpha s^\alpha-\{s+(s^2-1)^{1/2}\}^\alpha-\{s-(s^2-1)^{1/2}\}^\alpha.$$

Since F_α is continuous on $[1,\infty)$, our result follows if, on $(1,\infty)$, $F'_\alpha(s)<0$ for $\alpha<1$ and $F'_\alpha(s)>0$ for $\alpha>1$. Now

$$\begin{aligned}F'_\alpha(s)&=2^\alpha\alpha s^{\alpha-1}-\alpha\{s+(s^2-1)^{1/2}\}^{\alpha-1}\{1+s(s^2-1)^{-1/2}\}-\alpha\{s-(s^2-1)^{1/2}\}^{\alpha-1}\{1-s(s^2-1)^{-1/2}\}\\&=\frac{2^\alpha\alpha s^\alpha}{(s^2-1)^{1/2}}\left[\left\{\frac{1}{2}+\frac{(s^2-1)^{1/2}}{2s}\right\}^\alpha-\left\{\frac{1}{2}-\frac{(s^2-1)^{1/2}}{2s}\right\}^\alpha-2\frac{(s^2-1)^{1/2}}{2s}\right].\end{aligned}$$

Defining $\Delta=(s^2-1)^{1/2}/2s$ and $G_\alpha(\Delta)=(\frac{1}{2}+\Delta)^\alpha-(\frac{1}{2}-\Delta)^\alpha-2\Delta$, we see that $F'_\alpha(s)$ and $G_\alpha(\Delta)$ always have the same sign. Now Δ is a strictly increasing function of s from $[1,\infty)$ onto $[0,\frac{1}{2})$, so it suffices to show that $G_\alpha(\Delta)>0$ on $(0,\frac{1}{2})$ for $\alpha<1$ and $G_\alpha(\Delta)<0$ on $(0,\frac{1}{2})$ for $\alpha>1$. To prove this, first note that G_α is continuous on $[0,\frac{1}{2}]$ and, on $[0,\frac{1}{2})$,

$$G''_\alpha(\Delta)=\alpha(\alpha-1)\{(\tfrac{1}{2}+\Delta)^{\alpha-2}-(\tfrac{1}{2}-\Delta)^{\alpha-2}\}.$$

Since $(\frac{1}{2}+\Delta)^{\alpha-2}-(\frac{1}{2}-\Delta)^{\alpha-2}<0$ for all $\Delta\in(0,\frac{1}{2})$, $\alpha(\alpha-1)<0$ for $\alpha\in(0,1)$ and $\alpha(\alpha-1)>0$ for $\alpha\in(1,2)$, we see that, on $[0,\frac{1}{2}]$, G_α is convex for $\alpha\in(0,1)$ and concave for $\alpha\in(1,2)$. The result then follows by noting that $G_\alpha(0)=G_\alpha(\frac{1}{2})=0$ for all $\alpha\in(0,2)$.

Appendix B

We give more explicit expressions for $\partial_l\mathbf{W}_j$ and $\partial_l V_j$. Suppose that $\mathbf{S}_{(j-1)}$ has length m and

$$\begin{pmatrix}\mathbf{S}_{(j-1)}\\ \mathbf{Z}_j\end{pmatrix}\sim N\left\{\begin{pmatrix}F_1\boldsymbol{\beta}\\ F_2\boldsymbol{\beta}\end{pmatrix},\begin{pmatrix}K_{11}&K_{12}\\ K_{21}&K_{22}\end{pmatrix}\right\}.$$

Define T to be the first m rows of

$$\begin{pmatrix} K_{11} & F_1 \\ F_1' & O \end{pmatrix}^{-1} \begin{pmatrix} K_{12} \\ F_2' \end{pmatrix},$$

so that $T'\mathbf{S}_{(j-1)}$ is the BLUP of $\mathbf{Z}_j$ in terms of $\mathbf{S}_{(j-1)}$ (Stein (1999), page 8). Using standard results on differentiation of matrix-valued functions, it follows that $\partial_l T$ is given by the first m rows of

$$\begin{pmatrix} K_{11} & F_1 \\ F_1' & O \end{pmatrix}^{-1} \begin{pmatrix} \partial_l K_{12} - (\partial_l K_{11})T \\ O \end{pmatrix}.$$

Then $\partial_l \mathbf{W}_j = -(\partial_l T)'\mathbf{S}_{(j-1)}$ and, using $\mathrm{cov}(\mathbf{W}_j, \partial_l \mathbf{W}_j') = O$, $\partial_l V_j = (-T' I_{n_j})\partial_l K(-T' I_{n_j})'$, where, here, K is the covariance matrix of $(\mathbf{S}_{(j-1)}' \mathbf{Z}_j')'$.

References

Abramowitz, M. and Stegun, I. (1965) *Handbook of Mathematical Functions*, 9th edn. New York: Dover Publications.

Chilès, J. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Christensen, R. (1996) *Plane Answers to Complex Questions: the Theory of Linear Models*, 2nd edn. New York: Springer.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.

Eide, A. L., Omre, H. and Ursin, B. (2002) Prediction of reservoir variables based on seismic data and well observations. *J. Am. Statist. Ass.*, **97**, 18–28.

Genton, M. G. and Gorsich, D. J. (2002) Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices. *Comput. Statist. Data Anal.*, **41**, 47–57.

Golub, G. H. and van Loan, C. F. (1996) *Matrix Computations*, 3rd edn. Baltimore: Johns Hopkins University Press.

Heyde, C. C. (1997) *Quasi-likelihood and Its Application: a General Approach to Optimal Parameter Estimation*. New York: Springer.

Jones, R. H. and Zhang, Y. (1997) Models for continuous stationary space-time processes. In *Modelling Longitudinal and Spatially Correlated Data* (eds T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren and R. D. Wolfinger), pp. 289–298. New York: Springer.

Kang, W., Wan Shin, D. and Lee, Y. (2003) Biases of the restricted maximum likelihood estimators for ARMA processes with polynomial time trend. *J. Statist. Planng Inf.*, **116**, 163–176.

Kitanidis, P. K. (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Wat. Resour. Res.*, **19**, 909–921.

Lark, R. M. (2000) Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *Eur. J. Soil Sci.*, **51**, 717–728.

Lark, R. M. (2002) Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, **105**, 49–80.

Mardia, K. V. and Marshall, R. J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **73**, 135–146.

McGilchrist, C. A. (1989) Bias of ML and REML estimators in regression models with ARMA errors. *J. Statist. Computn Simuln*, **32**, 127–136.

Pardo-Igúzquiza, E. (1998) Maximum likelihood estimation of spatial covariance parameters. *Math. Geol.*, **30**, 95–108.

Pardo-Igúzquiza, E. and Dowd, P. A. (1997) AMLE3D: a computer program for the inference of spatial covariance parameters by approximate maximum likelihood estimation. *Comput. Geosci.*, **23**, 793–805.

Pardo-Igúzquiza, E. and Dowd, P. A. (1998) Maximum likelihood inference of spatial covariance parameters of soil properties. *Soil Sci.*, **163**, 212–219.

Park, J. S. and Baek, J. S. (2001) Efficient computation of maximum likelihood estimators in a spatial linear model with power exponential covariogram. *Comput. Geosci.*, **27**, 1–7.

Pettitt, A. N. and McBratney, A. B. (1993) Sampling designs for estimating spatial variance components. *Appl. Statist.*, **42**, 185–209.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992) *Numerical Recipes*, 2nd edn. Cambridge: Cambridge University Press.

Shapiro, A. and Botha, J. D. (1991) Variogram fitting with a general class of conditionally non-negative definite function. *Comput. Statist. Data Anal.*, **11**, 87–96.

Stein, M. L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Tunnicliffe Wilson, G. (1989) On the use of marginal likelihood in time series model estimation. *J. R. Statist. Soc.* B, **51**, 15–27.

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.

Welty, L. J. and Stein, M. L. (2003) Modeling phytoplankton: covariance and variogram model specification for phytoplankton levels in Lake Michigan. In *Geostatistics for Environmental Applications*. Boston: Kluwer. To be published. (Available from `http://galton.uchicago.edu/~cises/research/cises-tr1.pdf`.)

Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **49**, 305–314.

Wilson, P. D. (1988) Maximum likelihood estimation using differences in an autoregressive-1 process. *Communs Statist Theory Meth.*, **17**, 17–26.

Zhu, Z. (2002) Optimal sampling design and parameter estimation of Gaussian random fields. *PhD Dissertation*. Department of Statistics, University of Chicago, Chicago.