Bernoulli **7**(3), 2001, 421–438

On the effect of covariance function estimation on the accuracy of kriging predictors

HEIN PUTTER[1] and G. ALASTAIR YOUNG[2]

[1]*Department of Medical Statistics, University of Leiden, PO Box 9604, 2300 RC Leiden, The Netherlands. E-mail: h.putter@lumc.nl*
[2]*Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. E-mail: G.A.Young@statslab.cam.ac.uk*

The kriging procedure gives an optimal linear predictor of a spatial process at a point x_0, given observations of the process at other locations $x_1, \ldots, x_n$, taking into account the spatial dependence of the observations. The kriging predictor is optimal if the weights are calculated from the correct underlying covariance structure. In practice, this covariance structure is unknown and is estimated from the data. An important, but not very well understood, problem in kriging theory is the effect on the accuracy of the kriging predictor of substituting the optimal weights by weights derived from the estimated covariance structure. We show that the effect of estimation is negligible asymptotically if the joint Gaussian distributions of the process at $x_0, \ldots, x_n$ under the true and the estimated covariance are contiguous almost surely. We consider a number of commonly used parametric covariance models where this can indeed be achieved.

Keywords: contiguity; covariance function estimation; Gaussian process; kriging; spatial prediction; spectral density

1. Introduction

Kriging is a method for spatial prediction, widely used in mining, hydrology, forestry and other fields. Loosely speaking, given a spatial process which, observed at sampling locations $x_1, \ldots, x_n$, gives observations $z_1, \ldots, z_n$, it gives the optimal unbiased linear predictor of the process at a given point x_0 where the process is not observed, taking into account the spatial dependence of the observations. More specifically, in its simplest form, it is assumed that a stationary Gaussian process $Z(\cdot)$ with covariance function $C(t) = \mathrm{cov}(Z(x + t), Z(x))$ is observed at $x_1, \ldots, x_n$ and we wish to find weights $\alpha_{n1}, \ldots, \alpha_{nn}$ with $\sum_{i=1}^{n} \alpha_{ni} = 1$ such that the prediction error

$$\mathrm{E}\left(\sum_{i=1}^{n} \alpha_{ni} Z(x_i) - Z(x_0) \right)^2$$

is minimized. The restriction on the weights ensures that the kriging predictor $\sum_{i=1}^{n} \alpha_{ni} Z(x_i)$

1350–7265

is unbiased, i.e. $\mathrm{E}(\sum_{i=1}^{n} \alpha_{ni} Z(x_i) - Z(x_0)) = 0$. The kriging algorithm gives the optimal weights as a solution to a system of linear equations involving the covariances of $Z(x_i)$, $i = 0, 1, \ldots, n$ (Cressie 1991, p. 123).

In order to carry out this procedure it is therefore necessary to know the underlying covariance function C. The theoretical kriging predictor is optimal among all unbiased linear predictors: if $Z(\cdot)$ is Gaussian among all unbiased predictors. Hence it is clear that any attempt to approximate the optimal weights while retaining unbiasedness will result in a kriging predictor that has a larger prediction error. Common practice is to estimate C and adjust the weights according to the estimated covariance structure. We shall call the resulting kriging predictor the 'estimating kriging predictor'. Thus, the estimating kriging prediction error is at least as great as the theoretical kriging prediction error.

Although an extremely important practical issue, the effect of estimating C is still not all that well understood. There are essentially two approaches to assessing the influence of misspecifying or approximating the covariance function.

The first of these (Diamond and Armstrong 1984; Warnes 1986; Yakowitz and Szidarovsky 1985) is effectively a numerical analysis. Since the kriging weights are determined by solving linear equations involving the covariance matrix of $Z(x_1), \ldots, Z(x_n)$, approximating the covariance function results in a perturbation of that covariance matrix. The effect of approximating the covariance function on the kriging weights and hence on the kriging predictor can then be expressed in terms of the condition number of the covariance matrix. This program is carried out in detail by Diamond and Armstrong (1984) and gives bounds on the relative difference between the two kriging prediction errors, which are valid for every n and can be applied to every configuration of the sampling and prediction locations. Unfortunately, the bounds are not very sharp. In fact, as we carry out more and more measurements and acquire more observations, the bounds become wider instead of smaller as we would expect in most situations.

In the second approach, due to Stein (1988), the true underlying covariance function C_1 is assumed to be misspecified by a second covariance function C_2. The mean-zero Gaussian random fields with covariance functions C_1 and C_2, defined on a bounded region in $\mathbb{R}^d$, induce Gaussian probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. Following Stein (1988), we call C_1 and C_2 equivalent if the induced probability laws $\mathbb{P}_1$ and $\mathbb{P}_2$ are mutually absolutely continuous. If C_1 and C_2 are equivalent, it is shown in Stein (1988) that, under some conditions, the effect of misspecifying C_1 is asymptotically negligible, in the sense that the ratio of the estimating kriging prediction error and the theoretical kriging prediction error tends to unity as the number of observations tends to infinity. Application of this result, however, requires one to keep C_2 fixed as the number of observations changes. What is worse, typically C and an estimated covariance function $\hat{C}_n$ will not be equivalent for any finite n.

Thus, the question remains open how approximating the true covariance function C by a sequence C_n affects the accuracy of the kriging predictor. Since Yakowitz and Szidarovsky (1985, p. 39) remarked that 'we regard the situation as a (perhaps unfillable) lacuna in kriging theory', to the best of our knowledge the problem has not been solved to a satisfactory degree.

The aim of this paper is to study the effect of estimating the covariance function on the efficiency of kriging predictors. Our analysis is closest in nature to Stein's approach. In fact,

we mimic the proof of Theorem 1 of Stein (1988). However, where Stein passes to the limit first and considers absolute continuity of the resulting Gaussian processes, we retain the dependence on n of $\hat{C}_n$ and the two n-dimensional Gaussian vectors defined by C and $\hat{C}_n$ and consider contiguity of the distributions of these vectors. For a definition of contiguity and related concepts, see Sections 2 and 3.

We prove that, under essentially the same conditions as Stein (1988), the estimating kriging predictor is asymptotically efficient with respect to the theoretical kriging predictor, provided the $(n+1)$-dimensional Gaussian probability distributions P_n and $\hat{P}_n$ of $(Z(x_0), Z(x_1), \ldots, Z(x_n))$ under C and $\hat{C}_n$ are contiguous almost surely.

The asymptotic set-up that Stein considers is infill asymptotics, where samples are taken from a fixed bounded region and where the sampling locations become increasingly dense. This has often been contrasted with increasing-domain asymptotics, where the distance between neighbouring sampling locations remains bounded from below and the domain from which sampling takes place necessarily increases. Recently, however, a mixture of these extremes has become popular (Hall and Patil 1994; Lahiri 1997), which combines the merits of both approaches. This mixture carries a tuning parameter that determines the degree of infilling and increasing domain and makes it a very flexible set-up. Typically, this parameter is tuned in such a way that both the size of the sampling region and the number of observations in each fixed subset of that region grow with n.

In principle, contiguity of $\{P_n\}$ and $\{\hat{P}_n\}$ will depend both on the configuration of the prediction and sampling locations and on the quality of the estimator $\hat{C}_n$. Our result on asymptotic efficiency of the estimating kriging predictor is only useful if it is indeed possible to find an estimator $\hat{C}_n$ of C such that $\{P_n\}$ and $\{\hat{P}_n\}$ are contiguous almost surely. In Section 3 we study the contiguity of Gaussian random vectors in more detail and give conditions, first on the covariance matrices involved, then on $\hat{C}_n$ relative to C that guarantee contiguity of $\{P_n\}$ and $\{\hat{P}_n\}$. The usual way the covariance function is estimated is by using a nonparametric pilot estimate of C and from that fitting (usually by eye!) a class of commonly used parametric covariance functions (Cressie 1991, pp. 61–62). The parameter of that parametric class is then estimated using standard statistical techniques such as least squares, maximum likelihood and minimum norm quadratic estimation. There are a number of reasons for this. First of all, such a nonparametric pilot estimate is notoriously ill behaved away from the origin (Journel and Huijbregts 1978), although this may be true as well for covariance functions estimated directly within parametric classes. More importantly, this covariance function estimator may not be permissible in the sense that a covariance matrix derived from it need not be positive definite. For a detailed account of permissibility issues see Christakos (1984). Hall and Patil (1994) propose permissible kernel-type nonparametric estimators of covariance functions. There is a slightly worrying issue concerning covariance function estimation under infill asymptotics, pointed out by Lahiri (1996). He shows that the most commonly used nonparametric estimator of C is not consistent under infill asymptotics. This makes one very suspicious about the behaviour of any estimator of C which is derived from that nonparametric covariance function estimator along the lines outlined above. It is not yet clear to us whether (permissible) nonparametric covariance function estimators $\hat{C}_n$ may lead to probability distributions P_n and $\hat{P}_n$ which are contiguous almost surely. However, within a number of these parametric covariance function

models, it can indeed be shown that the parameter within that model can be estimated in such a way that the corresponding Gaussian distributions are indeed contiguous almost surely. Section 4 contains a number of stylized examples that illustrate this point and connect our work with that of Stein (1988) and Stein and Handcock (1989). The examples considered here are admittedly limited in scope, with regularly sited spatial data. We do believe that the examples in Section 4 can in fact be extended to moderately irregularly spaced spatial data as well. To prove contiguity in more general (parametric) covariance function models however, such as the Matérn model, as proposed by Stein (1999), may pose formidable problems. Finally, Section 5 discusses the estimation of the estimating kriging prediction error. We conclude this section with a number of remarks.

The fact that, within a parametric family of covariance functions $\{C_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$, the estimating kriging predictor using covariance function $C_{\hat{\theta}}$ is asymptotically efficient with respect to the theoretical kriging predictor using covariance function C_θ, implies that the optimal prediction error is the same, whether we know θ or not. This phenomenon, called *adaptation*, is well known in parametric and semi-parametric estimation (Bickel *et al.*, 1993, Section 2.4), but we are unaware of any previous occurrences of adaptation in the literature in the context of spatial prediction.

In the geostatistical literature the variogram

$$2\gamma(t) = \mathrm{E}(Z(x+t) - Z(x))^2$$

is used more often than the covariance function, one of the reasons being that it requires a weaker assumption than the (second-order) stationarity needed for the covariance function, namely that the process has stationary increments. In case of stationarity, the relation between them is given by

$$\gamma(t) = C(0) - C(t).$$

Cressie (1991, p. 70) argues with some justification that variogram estimation is to be preferred to covariance function estimation. Because conditions for contiguity are most naturally expressed in terms of covariance matrices, we have preferred to state our results in terms of covariance functions rather than variograms. However, it is straightforward to translate those results to variograms as well.

As in Stein (1988), our results, formulated for *ordinary kriging*, where Gaussian processes are assumed to have constant mean, carry over to the more realistic case of *universal kriging*, where Z is a Gaussian process with mean

$$\mathrm{E}\,Z(x) = \sum_{i=0}^{p} \beta_i f_i(x) = \beta^{\mathrm{T}} f(x)$$

and covariance function C, where f_i are specified functions and β_i are regression coefficients. Typically, $f_0 \equiv 1$. In the case of universal kriging, the $(n+1)$-dimensional Gaussian probability distributions P_n and $\hat{P}_n$ of $(Z(x_0), Z(x_1), \ldots, Z(x_n))$, with mean vector $(\beta^{\mathrm{T}} f(x_0), \ldots, \beta^{\mathrm{T}} f(x_n))$ and covariance matrices Σ_n and $\hat{\Sigma}_n$, have to be contiguous almost surely. Since the mean vectors are the same for P_n and $\hat{P}_n$, there is no difference in that respect between ordinary kriging and universal kriging. The presence of the nuisance

parameter β may affect estimation of the (parameters of the) covariance function. However, typically this will not affect the rate of convergence of the estimated covariance function and hence will also not affect contiguity of $\{\hat{P}_n\}$ with respect to $\{P_n\}$. Cressie (1991, Section 3.4.3), discusses estimation of the covariance function in the presence of the nuisance parameter β.

It is possible to include Gaussian measurement error into the model. Suppose we do not observe $Z(x_1), \ldots, Z(x_n)$ exactly but instead we observe $Z_i = Z(x_i) + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed normal random variables with mean 0 and variance τ^2, denoted in this paper by N_{τ^2}. This can be incorporated into the model by adding a term $\tau^2 \mathbf{1}_{\{t=0\}}$ to the covariance function $C(t)$ and a term $\hat{\tau}_n^2 \mathbf{1}_{\{t=0\}}$ to the estimated covariance function $\hat{C}_n(t)$, where $\hat{\tau}_n^2$ is an estimator of τ^2. With these adaptations, the main result in Section 2 goes through unchanged.

2. Asymptotic efficiency of the estimating kriging predictor

Let $x_0, x_1, x_2, \ldots$ be an infinite sequence of distinct points in a (not necessarily bounded) subset $\mathscr{D}$ of $\mathbb{R}^d$. Let $\{Z(x) : x \in \mathbb{R}^d\}$ be a stationary Gaussian process with mean $\mathrm{E}Z(x) \equiv 0$, covariance function

$$C(t) = \mathrm{cov}(Z(x+t),\ Z(x)),$$

and probability law $\mathbb{P}$. We think of C as the true, but typically unknown, underlying covariance function of the process. Define $Z_i = Z(x_i)$, $i = 1, \ldots, n$. We observe $Z_1, \ldots, Z_n$, and we wish to predict $Z_0 = Z(x_0)$ on the basis of these observations.

Let Σ_n be the $(n+1) \times (n+1)$ covariance matrix of $Z_0, Z_1, \ldots, Z_n$, where for convenience we let indices run from 0 to n, i.e.

$$\Sigma_{n,ij} = \mathrm{cov}(Z(x_i),\ Z(x_j)) = C(x_i - x_j), \qquad i, j = 0, 1, \ldots, n. \tag{2.1}$$

We also define the $n \times n$ submatrix Ω_n, the n-vector ω_n and the scalar σ^2 by

$$\begin{aligned} \Omega_{n,ij} &= \Sigma_{n,ij}, \qquad i, j = 1, \ldots, n, \\ \omega_{n,i} &= \Sigma_{n,0i}, \qquad i = 1, \ldots, n, \\ \sigma^2 &= \Sigma_{n,ii} = C(0), \qquad i = 0, \ldots, n. \end{aligned}$$

Then for $\alpha_n = (\alpha_{n1}, \ldots, \alpha_{nn})^{\mathrm{T}}$,

$$\alpha_n = \Omega_n^{-1} \left(\omega_n + \mathbf{1} \frac{1 - \mathbf{1}^{\mathrm{T}} \Omega_n^{-1} \omega_n}{\mathbf{1}^{\mathrm{T}} \Omega_n^{-1} \mathbf{1}} \right) \tag{2.2}$$

defines an n-vector which clearly satisfies $\mathbf{1}^{\mathrm{T}} \alpha_n = \sum_{i=1}^n \alpha_{ni} = 1$. Here $\mathbf{1}$ denotes an n-vector consisting of 1s and a^{T} denotes the transpose of a vector or matrix a.

Define the linear predictor

$$\mathscr{Z}_n(x_0) = \sum_{i=1}^n \alpha_{ni} Z(x_i) \tag{2.3}$$

and its error

$$e_n(x_0) = \mathscr{Z}_n(x_0) - Z(x_0). \tag{2.4}$$

The weights α_{ni} defined by (2.2) are such that $\mathscr{Z}_n(x_0)$ is unbiased and that $\mathrm{var}_C(e_n(x_0))$ is minimized among all weights α_{ni} with $\sum_{i=1}^n \alpha_{ni} = 1$. We call $\mathscr{Z}_n(x_0)$ the theoretical kriging predictor and its mean squared error

$$\mathrm{var}_C(e_n(x_0)) = \sigma^2 - \omega_n^{\mathrm{T}} \Omega_n^{-1} \omega_n + \frac{(\mathbf{1}^{\mathrm{T}} \Omega_n^{-1} \omega_n - 1)^2}{\mathbf{1}^{\mathrm{T}} \Omega_n^{-1} \mathbf{1}} \tag{2.5}$$

the theoretical kriging prediction error.

Having observed $Z_1, \ldots, Z_n$, let $\hat{C}_n(t)$ be an estimator of $C(t)$ based on $Z_1, \ldots, Z_n$. Analagous to (2.1)–(2.2), define $\hat{\Sigma}_n$, $\hat{\Omega}_n$, $\hat{\omega}_n$, $\hat{\sigma}_n^2$ and $\hat{\alpha}_n$ by

$$\hat{\Sigma}_{n,ij} = \hat{C}_n(x_i - x_j), \quad i, j = 0, 1, \ldots, n, \tag{2.6}$$

$$\hat{\Omega}_{n,ij} = \hat{\Sigma}_{n,ij}, \quad i, j = 1, \ldots, n, \tag{2.7}$$

$$\hat{\omega}_{n,i} = \hat{\Sigma}_{n,0i}, \quad i = 1, \ldots, n, \tag{2.8}$$

$$\hat{\sigma}_n^2 = \hat{\Sigma}_{n,ii} = \hat{C}_n(0), \quad i = 0, \ldots, n, \tag{2.9}$$

and

$$\hat{\alpha}_n = \hat{\Omega}_n^{-1} \left(\hat{\omega}_n + \mathbf{1} \frac{1 - \mathbf{1}^{\mathrm{T}} \hat{\Omega}_n^{-1} \hat{\omega}_n}{\mathbf{1}^{\mathrm{T}} \hat{\Omega}_n^{-1} \mathbf{1}} \right). \tag{2.10}$$

The resulting linear predictor,

$$\hat{\mathscr{Z}}_n(x_0) = \sum_{i=1}^n \hat{\alpha}_{n,i} Z(x_i), \tag{2.11}$$

is called the estimating kriging predictor. Clearly, this is no longer the optimal unbiased linear predictor of $Z(x_0)$. In fact, it does not necessarily enjoy any of these properties (optimal, unbiased, linear), the latter two failing because $\hat{\alpha}_{n,i}$ now depends on $Z_1, \ldots, Z_n$. It has to be noted, though, that the estimating kriging predictor is in fact often unbiased (Christensen 1991). Let

$$\hat{e}_n(x_0) = \hat{\mathscr{Z}}_n(x_0) - Z(x_0). \tag{2.12}$$

Its variance $\mathrm{var}_C(\hat{e}_n(x_0))$ is the prediction mean squared error of the estimating kriging predictor $\hat{\mathscr{Z}}_n(x_0)$ and is called the estimating kriging prediction error.

To state our result we need to define contiguity first. For every n, let $(\mathscr{X}_n, \mathscr{A}_n)$ be a measurable space and let $\{Q_n\}$ and $\{Q'_n\}$ be two sequences of probability measures on $(\mathscr{X}_n, \mathscr{A}_n)$.

Definition 2.1. *The sequence $\{Q'_n\}$ is contiguous with respect to $\{Q_n\}$ if, for every $A_n \in \mathscr{A}_n$, $Q_n(A_n) \to 0$ implies $Q'_n(A_n) \to 0$.*

For more information on contiguity, see Roussas (1972) or Prakasa Rao (1987).

Let P_n and $\hat{P}_n$ be the $(n+1)$-dimensional Gaussian distributions with mean zero and covariance matrices Σ_n and $\hat{\Sigma}_n$, respectively. We shall call $\hat{\mathscr{Z}}_n(x_0)$ asymptotically efficient with respect to $\mathscr{Z}_n(x_0)$ if

$$\lim_{n\to\infty} \frac{\mathrm{E}_C(\hat{\mathscr{Z}}_n(x_0) - Z(x_0))^2}{\mathrm{E}_C(\mathscr{Z}_n(x_0) - Z(x_0))^2} = 1. \tag{2.13}$$

Thus, when (2.13) is fulfilled, if C is the true underlying covariance function, the estimating kriging predictor $\hat{\mathscr{Z}}_n(x_0)$ based on the estimated $\hat{C}_n$ performs asymptotically equally as well as the theoretical kriging predictor $\mathscr{Z}_n(x_0)$ based on the correct covariance function C. Clearly, the definition of asymptotic efficiency in (2.13) depends on the sampling locations $x_1, x_2, \ldots$. If (2.13) is true for all limit points x_0 of the sampling locations, then it becomes a property of the covariance function C and $\hat{C}_n$ only, and we shall call $\hat{C}_n$ asymptotically efficient with respect to C.

We are now ready to state our result.

Theorem 2.1. *If*

$$\mathrm{var}_C(e_n(x_0)) \to 0, \qquad \text{as } n \to \infty, \tag{2.14}$$

and $\{\hat{P}_n\}$ *and* $\{P_n\}$ *are contiguous,* $\mathbb{P}$*-almost surely, then* $\hat{\mathscr{Z}}_n(x_0)$ *is asymptotically efficient with respect to* $\mathscr{Z}_n(x_0)$ *as in* (2.13).

Proof. We shall start by proving (2.13) for a deterministic sequence of alternative covariance functions $\hat{C}_n$ such that $\{\hat{P}_n\}$ and $\{P_n\}$ are contiguous. Later the requirement that $\hat{C}_n$ be deterministic shall be removed. We follow the proof of Theorem 1 of Stein (1988) quite closely. In the proof we shall use the notation $Z(x)$ to denote a mean-zero Gaussian random field on a subset $\mathscr{D}$ of $\mathbb{R}^d$. The underlying covariance function is either $C(t)$ or $\hat{C}_n(t)$. It will be clear from the context which is the underlying covariance function; in particular, in calculating (co)variances we shall use the name of the covariance function as a subscript.

Since $\mathscr{Z}_n(x_0)$ is optimal under C, we must have

$$\frac{\mathrm{var}_C(e_n(x_0))}{\mathrm{var}_C(\hat{e}_n(x_0))} \leqslant 1.$$

Following Stein (1988), we write

$$\frac{\mathrm{var}_C(e_n(x_0))}{\mathrm{var}_C(\hat{e}_n(x_0))} = \frac{\mathrm{var}_C(e_n(x_0))}{\mathrm{var}_{\hat{C}_n}(e_n(x_0))} \cdot \frac{\mathrm{var}_{\hat{C}_n}(e_n(x_0))}{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))} \cdot \frac{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))}{\mathrm{var}_C(\hat{e}_n(x_0))}.$$

Since $\hat{\mathscr{Z}}_n(x_0)$ is optimal under $\hat{C}_n$, we have

$$\frac{\mathrm{var}_{\hat{C}_n}(e_n(x_0))}{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))} \geqslant 1. \tag{2.15}$$

Hence it suffices to show that

$$\liminf_{n\to\infty} \frac{\mathrm{var}_C(e_n(x_0))}{\mathrm{var}_{\hat{C}_n}(e_n(x_0))} \geqslant 1 \tag{2.16}$$

and

$$\liminf_{n\to\infty} \frac{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))}{\mathrm{var}_C(\hat{e}_n(x_0))} \geqslant 1. \tag{2.17}$$

Define

$$Y_n = \frac{e_n(x_0)}{(\mathrm{var}_C(e_n(x_0)))^{1/2}}, \qquad \hat{Y}_n = \frac{\hat{e}_n(x_0)}{(\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0)))^{1/2}},$$

so that both $\mathrm{E}_C Y_n = \mathrm{E}_{\hat{C}_n} \hat{Y}_n = 0$ and $\mathrm{E}_C Y_n^2 = \mathrm{E}_{\hat{C}_n} \hat{Y}_n^2 = 1$. The appropriate lemma of Stein (1988) is now as follows:

Lemma 2.2. *Any subsequence $n_1, n_2, \ldots$ contains a further subsequence $n_{k_1}, n_{k_2}, \ldots$ such that, with $\tilde{Y}_m = Y_{n_{k_m}}$ and $\tilde{P}_m = P_{n_{k_m}}$, we have, for every $\varepsilon > 0$,*

$$\tilde{P}_M \left(\left| \mathrm{M}^{-1} \sum_{m=1}^{M} \tilde{Y}_m^2 - 1 \right| > \varepsilon \right) \to 0, \qquad \text{as } M \to \infty. \tag{2.18}$$

The proof of Lemma 2.2 goes through unchanged, since it uses only properties of normal random variables and optimality of kriging predictors, which remain true under C and $\hat{C}_n$.

For reasons that will become clear later, we proceed by proving a slightly stronger statement than (2.16), namely

$$\lim_{n\to\infty} \frac{\mathrm{var}_C(e_n(x_0))}{\mathrm{var}_{\hat{C}_n}(e_n(x_0))} = 1. \tag{2.19}$$

Supposing that (2.19) is not true, there exists a subsequence $n_1, n_2, \ldots$ satisfying

$$\lim_{k\to\infty} \frac{\mathrm{var}_C(e_{n_k}(x_0))}{\mathrm{var}_{\hat{C}_{n_k}}(e_{n_k}(x_0))} = \lim_{k\to\infty} \frac{\mathrm{var}_C(Y_{n_k})}{\mathrm{var}_{\hat{C}_{n_k}}(Y_{n_k})} = \lim_{k\to\infty} \frac{1}{\tau_{n_k}^2} = c \neq 1, \tag{2.20}$$

where $\tau_n^2 = \mathrm{var}_{\hat{C}_n}(Y_n)$. Note that $c > 0$, since otherwise, with $\mu_n = \mathrm{E}_{\hat{C}_n} Y_n$, we would have, Y_n being normal,

$$P_{n_k}(|Y_{n_k} - \mu_{n_k}| > \tau_{n_k}^{1/2}) \to 0, \qquad \hat{P}_{n_k}(|Y_{n_k} - \mu_{n_k}| > \tau_{n_k}^{1/2}) \to 1,$$

which is in contradiction with the contiguity of $\{P_n\}$ and $\{\hat{P}_n\}$. So let us suppose that, as $k \to \infty$,

$$\mathrm{var}_{\hat{C}_{n_k}}(Y_{n_k}) \to c^{-1} < \infty. \tag{2.21}$$

Pick a further subsequence such that (2.18) holds. For that subsequence we have

$$\lim_{M\to\infty} \mathrm{E}_{\hat{C}_M} \left(M^{-1} \sum_{m=1}^{M} \tilde{Y}_m^2 \right) = c^{-1}. \tag{2.22}$$

By the contiguity of $\{P_n\}$ and $\{\hat{P}_n\}$, we have

$$\hat{P}_M\left(\left|M^{-1}\sum_{m=1}^{M}\tilde{Y}_m^2 - 1\right| > \varepsilon\right) \to 0, \qquad as\ n \to \infty, \tag{2.23}$$

for all $\varepsilon > 0$. Also, since $Z(x)$ is Gaussian, using (2.21), we have, as $M \to \infty$,

$$\begin{aligned}\mathrm{var}_{\hat{C}_M}\left(M^{-1}\sum_{m=1}^{M}\tilde{Y}_m^2\right) &= 2M^{-2}\sum_{l=1}^{M}\sum_{m=1}^{M}(\mathrm{cov}_{\hat{C}_M}(\tilde{Y}_l,\ \tilde{Y}_m))^2 \\ &\leqslant 2M^{-2}\sum_{l=1}^{M}\sum_{m=1}^{M}\mathrm{var}_{\hat{C}_M}(\tilde{Y}_l)\cdot\mathrm{var}_{\hat{C}_M}(\tilde{Y}_m) \to 2c^{-2}.\end{aligned}$$

Applying Theorem 4.5.2 in Chung (1974) to (2.23), we obtain

$$\lim_{M\to\infty}\mathrm{E}_{\hat{C}_M}\left(M^{-1}\sum_{m=1}^{M}\tilde{Y}_m^2\right) = 1,$$

which is in contradiction with (2.22). Thus, we have established (2.19) and a fortiori (2.16). It remains to show (2.17). Now, using the proof above on $\hat{Y}_n$, (2.17) will follow from (2.16) only if the analogue of (2.14) holds for $\hat{C}_n$ and $\hat{e}_n(x_0)$ as well – this is a small gap in the proof of Stein (1988)! – i.e. we require

$$\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0)) \to 0 \quad as\ n \to \infty. \tag{2.24}$$

Note, however, that

$$\frac{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))}{\mathrm{var}_C(e_n(x_0))} = \frac{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))}{\mathrm{var}_{\hat{C}_n}(e_n(x_0))}\cdot\frac{\mathrm{var}_{\hat{C}_n}(e_n(x_0))}{\mathrm{var}_C(e_n(x_0))}. \tag{2.25}$$

The first term on the right-hand side of (2.25) is less than or equal to 1 by (2.15), the latter tends to 1 by (2.19). Hence

$$\limsup_{n\to\infty}\frac{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))}{\mathrm{var}_C(e_n(x_0))} \leqslant 1,$$

which, together with (2.14), establishes (2.24). Hence, Lemma 2.2 can be applied to $\hat{Y}_n$ under $\hat{P}_n$ and the contradiction argument following that, to prove (2.17) and (2.14) for a deterministic sequence of alternative covariance functions $\hat{C}_n$ such that $\{\hat{P}_n\}$ and $\{P_n\}$ are contiguous.

To finish the proof, we note that, for $\hat{C}_n$ random such that $\{\hat{P}_n\}$ and $\{P_n\}$ are contiguous $\mathbb{P}$-almost surely, the conclusion (2.14) remains valid since a $\mathbb{P}$-null exceptional set where $\{\hat{P}_n\}$ and $\{P_n\}$ are possibly non-contiguous does not contribute to the integrals in (2.13). □

3. Conditions for contiguity

Recall the definition of contiguity given in Definition 2.1 and consider the special case where

$$Q_n = \prod_{i=1}^{n} Q_{ni}, \qquad Q'_n = \prod_{i=1}^{n} Q'_{ni} \tag{3.1}$$

are product measures on a product space $(\mathscr{X}_n, \mathscr{A}_n) = (\prod_{i=1}^n \mathscr{X}_{ni}, \prod_{i=1}^n \mathscr{A}_{ni})$ with marginals Q_{ni} and Q'_{ni} on $(\mathscr{X}_{ni}, \mathscr{A}_{ni})$. Let q_{ni} and q'_{ni} be densities of Q_{ni} and Q'_{ni} with respect to σ-finite measures μ_{ni} on $(\mathscr{X}_{ni}, \mathscr{A}_{ni})$.

Definition 3.1. *The Hellinger distance between Q_{ni} and Q'_{ni} is defined as*

$$H^2(Q_{ni}, Q'_{ni}) = \int (q_{ni}^{1/2} - q_{ni}^{\prime 1/2})^2 \, \mathrm{d}\mu_{ni}. \tag{3.2}$$

The following relation between contiguity of $\{Q_n\}$ and $\{Q'_n\}$ has been proved in Oosterhoff and van Zwet (1979) (see also Prakasa Rao 1987).

Lemma 3.1. *The sequence $\{Q'_n\}$ is contiguous with respect to $\{Q_n\}$ if and only if*

$$\limsup_{n\to\infty} \sum_{i=1}^{n} H^2(Q_{ni}, Q'_{ni}) < \infty \tag{3.3}$$

and

$$\lim_{n\to\infty} \sum_{i=1}^{n} \int_{\{x: q'_{ni}(x) \geqslant c_n q_{ni}(x)\}} q'_{ni}(x) \mathrm{d}\mu_{ni}(x) = 0 \qquad \text{if } c_n \to \infty. \tag{3.4}$$

Supposing that P_n and $\hat{P}_n$ are the distributions of $(n+1)$-dimensional Gaussian mean-zero vectors with covariance matrices Σ_n and $\hat{\Sigma}_n$ respectively, the question remains what conditions on Σ and $\hat{\Sigma}_n$ are needed to guarantee contiguity of $\{\hat{P}_n\}$ with respect to $\{P_n\}$. Let us denote the difference between the covariance matrices by

$$\Delta_n = \hat{\Sigma}_n - \Sigma_n. \tag{3.5}$$

Lemma 3.2. *The sequence $\{\hat{P}_n\}$ is contiguous with respect to $\{P_n\}$ if and only if there exist $0 < K_1 \leqslant K_2 < \infty$ such that*

$$\limsup_{n\to\infty} \sum_{i=0}^{n} \lambda_i^2 \leqslant K_2 \tag{3.6}$$

and

$$\liminf_{n\to\infty} \inf_{0\leqslant i\leqslant n} \lambda_i \geqslant -1 + K_1, \tag{3.7}$$

where $\lambda_0, \ldots, \lambda_n$ are the eigenvalues of $\Sigma_n^{-1}\Delta_n$.

Proof. By Rao (1965, p. 42, (iv, c)), there exists a non-singular $(n+1)\times(n+1)$ matrix B such that

$$B^{\mathrm{T}}\Sigma_n B = I \quad \text{and} \quad B^{\mathrm{T}}\hat{\Sigma}_n B = \Lambda^*, \tag{3.8}$$

where Λ^* is diagonal with elements λ_i^* as the eigenvalues of $\Sigma_n^{-1}\hat{\Sigma}_n = I + \Sigma_n^{-1}\Delta_n$.

It is clear from the definition of contiguity that contiguity is preserved under one-to-one transformations. It is also straightforward to see that

$$\frac{1}{2}H^2(N_{\sigma^2}, N_{\tau^2}) = 1 - \sqrt{\frac{2\sigma\tau}{\sigma^2+\tau^2}} = \frac{(\sigma-\tau)^2}{\sigma^2+\tau^2}\left(1+\sqrt{\frac{2\sigma\tau}{\sigma^2+\tau^2}}\right)^{-1}.$$

Application of Lemma 3.1 then shows that $\{\hat{P}_n\}$ is contiguous with respect to $\{P_n\}$ if and only if

$$\limsup_{n\to\infty}\sum_{i=0}^{n}(\lambda_i^*-1)^2 < \infty \tag{3.9}$$

and

$$\liminf_{n\to\infty}\inf_{0\leqslant i\leqslant n}\lambda_i^* > 0. \tag{3.10}$$

Now let λ_i be an eigenvalue of $\Sigma^{-1}\Delta$ with corresponding eigenvector x_i. Then

$$(I+\Sigma^{-1}\Delta_n)x_i = (1+\lambda_i)x_i, \tag{3.11}$$

so for every eigenvalue λ_i^* of $\Sigma_n^{-1}\hat{\Sigma}_n$ there exists an eigenvalue $\lambda_i = \lambda_i^*$ of $\Sigma_n^{-1}\Delta_n$ with common eigenvector x_i. Replacing $\lambda_i^* - 1$ by λ_i in (3.9) and (3.10) proves the lemma. We note that Ibragimov and Rozanov (1978, pp. 70–77) also contains the essential elements of a proof of this lemma. □

Remark 3.1. The sum of the squares of the eigenvalues of $\Sigma_n^{-1}\Delta_n$ can be calculated by using the trace of the square of $\Sigma_n^{-1}\Delta_n$:

$$\sum_{i=1}^{n}\lambda_i^2 = \mathrm{tr}((\Sigma_n^{-1}\Delta_n)^2). \tag{3.12}$$

Clearly, Σ_n and $\hat{\Sigma}_n$ depend not only on the covariance functions C and $\hat{C}_n$ but also on the spatial configuration of the sampling locations. Via conditions (3.6) and (3.7), the same is true for the Gaussian measures P_n and $\hat{P}_n$. Thus, contiguity of $\{\hat{P}_n\}$ with respect to $\{P_n\}$, and hence most likely asymptotic efficiency of $\hat{\mathscr{Z}}_n(x_0)$ with respect to $\mathscr{Z}_n(x_0)$, will depend on the location of $x_1, x_2, \ldots, x_n$ with respect to x_0. In checking contiguity, one would therefore wish to take that into account. On the other hand, conditions (3.6) and (3.7) might be quite difficult to check in a particular application and it is therefore desirable to give more easily verifiable conditions for contiguity, even if they ignore the configuration of the sampling locations. The following lemma gives conditions, independent of the sampling locations, for contiguity of $\{\hat{P}_n\}$ with respect to $\{P_n\}$. For any covariance function C on $\mathscr{D}\subset\mathbb{R}^d$, let f denote its spectral density (assuming it exists), i.e.

$$C(t) = \int_{\mathbb{R}^d} e^{i\langle t,\nu\rangle} f(\nu) d\nu, \tag{3.13}$$

where, for t and ν in $\mathbb{R}^d$, $\langle t, \nu\rangle = \sum_{j=1}^{d} t_j \nu_j$. The spectral density can be found from the covariance function by the inverse formula of (3.13):

$$f(\nu) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-i\langle t,\nu\rangle} C(t) dt. \tag{3.14}$$

For an extensive treatment of spectral measures in the context of time series, see, for example, Priestley (1981).

Lemma 3.3. *Suppose that C and $\hat{C}_n$ have spectral measures F and $\hat{F}_n$, respectively, which are absolutely continuous with respect to d-dimensional Lebesgue measure with densities f and $\hat{f}_n$, respectively. Then $\{\hat{P}_n\}$ is contiguous with respect to $\{P_n\}$ if*

$$\limsup_{n\to\infty} n \sup_{\nu\in\mathbb{R}^d} \left(\frac{\hat{f}_n(\nu) - f(\nu)}{f(\nu)} \right)^2 < \infty \tag{3.15}$$

and

$$\liminf_{n\to\infty} \inf_{\nu\in\mathbb{R}^d} \frac{\hat{f}_n(\nu)}{f(\nu)} > 0. \tag{3.16}$$

Proof. Let λ_i be an eigenvalue of $\Sigma_n^{-1}\Delta_n$ with corresponding eigenvector y. Then

$$\lambda_i y^T \Sigma_n y = y^T \Delta_n y, \tag{3.17}$$

and hence

$$|\lambda_i| \leqslant \sup_y \left| \frac{y^T \Delta_n y}{y^T \Sigma_n y} \right|. \tag{3.18}$$

Since

$$\left| \frac{y^T \Delta_n y}{y^T \Sigma_n y} \right| = \left| \frac{\int_{\mathbb{R}^d} \left| \sum_j y_j e^{-ix_j\nu} \right|^2 (\hat{f}_n(\nu) - f(\nu)) d\nu}{\int_{\mathbb{R}^d} \left| \sum_j y_j e^{-ix_j\nu} \right|^2 f(\nu) d\nu} \right| \leqslant \sup_\nu \left| \frac{\hat{f}_n(\nu) - f(\nu)}{f(\nu)} \right|,$$

we have

$$\sum_{i=0}^{n} \lambda_i^2 \leqslant (n+1) \sup_\nu \left(\frac{\hat{f}_n(\nu) - f(\nu)}{f(\nu)} \right)^2.$$

Similarly, if λ_i^* is an eigenvalue of $\Sigma_n^{-1}\hat{\Sigma}_n^{-1}$, then

$$\lambda_i^* \geqslant \inf_y \frac{y^{\mathrm{T}}\hat{\Sigma}_n y}{y^{\mathrm{T}}\Sigma_n y} \geqslant \inf_\nu \frac{\hat{f}_n(\nu)}{f(\nu)}.$$

The lemma then follows on applying Lemma 3.2. □

4. Examples

In this section we shall apply the results of the previous section to a number of examples. The first example was considered in Stein and Handcock (1989).

Example 4.1. Consider equally spaced locations $x_i = i/n$, $i = 0, \ldots, n-1$, in the unit interval in $\mathbb{R}$ and let, for $|t| \leqslant 1$,

$$C(t) = 1 - |t|, \qquad C_n(t) = C(t) + \beta_n \delta(t), \tag{4.1}$$

where β_n is a sequence of bounded real numbers and $\delta(t)$ is a twice continuously differentiable function such that $C_n(t)$ is a permissible covariance function for all n. We shall see later that the behaviour of the derivative of δ at the origin dictates different conditions on β_n for C and C_n for $\{\hat{P}_n\}$ and $\{P_n\}$ to be contiguous. If Σ_n denotes the covariance matrix of $(Z(x_0), Z(x_1), \ldots, Z(x_{n-1}))$, then clearly $\Sigma_{n,ij} = 1 - |i-j|/n$. The matrix Δ_n is defined by $\Delta_{n,ij} = \beta_n \delta(|i-j|/n)$. We use identity (3.12), and study the trace of $(\Sigma_n^{-1}\Delta_n)^2$. Let D_k denote the kth diagonal element of $(\Sigma_n^{-1}\Delta_n)^2$, and define

$$\psi(t) = (\delta(t) + \delta(1-t) + \delta'(t))\delta''(t),$$

$$m_n = \frac{1}{n}\sum_{j=2}^{n-1} \psi\left(\frac{j-1}{n}\right) \approx \int_0^1 \psi(t)\mathrm{d}t.$$

Elementary matrix manipulations show that

$$D_1 = D_n \approx \frac{\beta_n^2}{4}[(\delta(0) + \delta(1) + \delta'(0+))^2 + (\delta(0) + \delta(1) + \delta'(1))^2 - m_n]$$

and, for $2 \leqslant k \leqslant n-1$,

$$D_k \approx \frac{\beta_n^2}{4}\left[4(\delta'(0+))^2 - \frac{\psi\left(\frac{k-1}{n}\right) + \psi\left(\frac{n-k}{n}\right)}{n} + \frac{1}{n^2}\sum_{\substack{j=2 \\ j\neq k}}^{n-1}\left(\delta''\left(\frac{|j-k|}{n}\right)\right)^2\right].$$

If $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $\Sigma_n^{-1}\Delta_n$, we arrive at

$$\sum_{i=1}^n \lambda_i^2 = \sum_{k=1}^n D_k \approx \beta_n^2\Bigg[n(\delta'(0+))^2 + \frac{1}{2}(\delta(0) + \delta(1) + \delta'(0+))^2$$

$$+ \frac{1}{2}(\delta(0) + \delta(1) + \delta'(1))^2 - \int_0^1 \psi(t)\mathrm{d}t\Bigg]. \tag{4.2}$$

It is not difficult to see that the contributions of the remainder terms in (4.2) are indeed negligible. Application of Theorem 2.1 and Lemma 3.2 now tells us that the kriging predictor based on C_n is asymptotically efficient if $\sum_{i=1}^{n}\lambda_i^2$ in (4.2) is bounded and (3.7) holds. The analysis of $\sum_{i=1}^{n}\lambda_i^2$ in (4.2) exhibits two interesting features. First, subject to the permissibility condition, if $\delta'(0+)=0$, then we are in a situation where C_n and C behave similarly at the origin, in the sense that, for every n, $C_n'(0+)=C'(0+)$. As a result, for every n the Gaussian processes determined by C and C_n respectively are mutually absolutely continuous (with some exceptions) and Theorem 1 in Stein (1988) asserts that, for $\beta_n=\beta$ finite, the kriging predictor based on C_n is asymptotically efficient. Our analysis shows that if $\delta'(0+)=0$, $\sum_{i=1}^{n}\lambda_i^2$ remains bounded whenever β_n remains bounded and hence the estimating kriging prediction error is asymptotically efficient if also (3.7) holds. Stein and Handcock (1989, Example 3) discuss this example for $\delta(t)=(1-t^2)/2$ and $0\leqslant\beta_n\leqslant 1$. This indeed makes $C_n(t)$ a permissible covariance function, since

$$C_n(t)=C(t)+\beta_n\frac{1-t^2}{2}=(1-\beta_n)C(t)+\beta_n C_2(t),$$

where, for $|t|\leqslant 1$,

$$C_2(t)=\frac{1}{2}-|t|+\frac{t^2}{2}$$

is a permissible covariance function. This follows because its spectral density

$$f_2(\nu)=(\pi\nu^3)^{-1}(\nu-\sin(\nu))$$

is non-negative. The spectral density of $C(t)$ is given by

$$f_1(\nu)=(\pi\nu^2)^{-1}(1-\cos(\nu)).$$

Now (3.16) of Lemma 3.3 implies that (3.7) is fulfilled if

$$\limsup_{n\to\infty}\beta_n<1, \tag{4.3}$$

since if $f(\nu)\equiv f_1(\nu)$ and $\hat{f}_n(\nu)=f_1(\nu)+\beta_n(f_2(\nu)-f_1(\nu))$ denote the spectral densities of C and C_n respectively, we have

$$\frac{\hat{f}_n(\nu)}{f(\nu)}\geqslant 1+\beta_n\frac{f_2(\nu)-f_1(\nu)}{f_1(\nu)}=1+\beta_n\frac{\cos\nu-\dfrac{\sin\nu}{\nu}}{1-\cos\nu}\geqslant 1-\beta_n,$$

the latter inequality because $(\sin\nu)/\nu\leqslant 1$. Lemmas 3.2 and 3.3, together with Theorem 2.1, imply that for $\delta(t)=(1-t^2)/2$, if (4.3) holds, the estimating kriging predictor is asymptotically efficient with respect to the theoretical kriging predictor. Thus, here our result is in agreement with Example 3 in Stein and Handcock (1989).

For the case $\delta'(0+)\neq 0$, Lemma 3.2 gives valuable additional information. Now for fixed n, unless $\beta_n=0$, the Gaussian processes governed by C and C_n are not absolutely continuous, so Theorem 1 of Stein (1988) cannot be applied. It is easy to see that $\sum_{i=1}^{n}\lambda_i^2$ remains bounded if $\beta_n=O(n^{-1/2})$, so for $\{\hat{P}_n\}$ to be contiguous with respect to $\{P_n\}$ we need β_n to be of the order $O(n^{-1/2})$.

Example 4.2. *Exponential covariance functions.* Consider a Gaussian process on $\mathbb{R}$ with covariance function

$$C_{\theta,\sigma^2}(t) = \sigma^2 \mathrm{e}^{-\theta|t|}, \qquad \sigma^2, \theta > 0. \tag{4.4}$$

Suppose the process is observed at $x_i = i/a_n$, $i = 1, \ldots, n-1$ and is to be predicted at $x_0 = 0$. Here a_n determines the degree of infilling and increasing domain. In particular, $a_n = O(n)$ corresponds to infill asymptotics and $a_n = O(1)$ corresponds to increasing-domain asymptotics. The sequence a_n is assumed to be inside the boundaries described above. Let θ_n and σ_n^2 be bounded sequences of numbers. Summing up, we assume that there exist constants $0 < K_1 \leqslant K_2 < \infty$ such that

$$K_1 \leqslant a_n \leqslant K_2 n, \qquad |\theta_n - \theta| \leqslant K_2, \qquad |\sigma_n^2 - \sigma^2| \leqslant K_2. \tag{4.5}$$

As before, let P_n be the joint distribution of the process at prediction and sampling locations under C_{θ,σ^2} and let $\hat{P}_n$ be the joint distribution of the process at the same locations under C_{θ_n,σ_n^2}.

We shall investigate, under various degrees of infilling, what rates are necessary to obtain contiguity of $\{P_n\}$ and $\{\hat{P}_n\}$. In this set-up we have

$$\Sigma_{n,ij} = \sigma^2 \rho^{|i-j|}, \qquad \rho = \exp(-\theta/a_n), \tag{4.6}$$

the inverse of which is given in Cressie (1991, p. 133). Similarly, define $\hat{\Sigma}_n$ by substituting θ_n, σ_n^2 and $\rho_n = \exp(-\theta_n/a_n)$ for θ, σ^2 and ρ in (4.6). Finally, let

$$c_n = \frac{\sigma_n^2}{\sigma^2} \frac{1 - \rho\rho_n}{1 - \rho^2}.$$

As in Example 4.1, we use (3.12) and study the sum of the diagonal elements D_k of $(\Sigma_n^{-1}\Delta_n)^2$. Straightforward calculations yield

$$\sum_{i=1}^{n} \lambda_i^2 = \sum_{k=1}^{n} D_k = S_1 + S_2 + S_3,$$

where

$$S_1 = (2c_n - 1)^2 + (n-2)\left(c_n \frac{1 - 2\rho\rho_n + \rho^2}{1 - \rho\rho_n} - 1\right)^2 + \frac{2c_n^2(\rho_n - \rho)^2}{(1 - \rho\rho_n)^2} \rho_n^{2(n-1)},$$

$$S_2 = 4 \frac{\rho_n c_n^2 (\rho_n - \rho)^2}{1 - \rho\rho_n} \frac{1 - \rho_n^{2(n-2)}}{1 - \rho_n^2},$$

$$S_3 = \frac{2c_n^2(\rho_n - \rho)^2}{1 - \rho_n^2} \left[(n-2) - \frac{1 - \rho_n^{2(n-2)}}{1 - \rho_n^2}\right].$$

The order of magnitude of these terms can be analysed using (4.5) and noting that

$$c_n = O(1), \qquad \rho_n - \rho = O(a_n^{-1}(\theta_n - \theta)), \qquad 1 - \rho_n^2 = O(a_n^{-1}), \qquad 1 - \rho_n^{2n} = O(1),$$

which yields

$$S_1 = O\left(n\left(\frac{\sigma_n^2\theta_n}{\sigma^2\theta} - 1\right)^2\right), \qquad S_2 = O((\theta_n - \theta)^2), \qquad S_3 = O\left(\frac{n}{a_n}(\theta_n - \theta)^2\right).$$

Under infill asymptotics (i.e. $a_n = O(n)$), it is well known (Ying 1991) that θ and σ^2 cannot be estimated consistently. In fact, θ and σ^2 are not identifiable from observing the whole path of the processes because, for $\tilde{\theta}\tilde{\sigma}^2 = \theta\sigma^2$, the Gaussian processes with mean zero and covariance functions C_{θ,σ^2} and $C_{\tilde{\theta},\tilde{\sigma}^2}$ are mutually absolutely continuous. Ying (1991) shows that the maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\sigma}_n^2$ of θ and σ^2 are such that $\sqrt{n}(\hat{\theta}_n\hat{\sigma}_n^2 - \theta\sigma^2)$ converges to a normal distribution and $\hat{\theta}_n - \theta$ and $\hat{\sigma}_n^2 - \sigma^2$ are bounded almost surely. Thus, for the maximum likelihood estimators, $\sum_{i=1}^n \lambda_i^2$ is bounded almost surely.

It is instructive to see how far Lemma 3.3 can bring us in checking the conditions of Lemma 3.2. Let f and $\hat{f}_n$ denote the spectral density corresponding to C_{θ,σ^2} and C_{θ_n,σ_n^2} respectively. Then (Priestley 1981, p. 236)

$$f(\nu) = \frac{\sigma^2\theta}{\pi(\theta^2 + \nu^2)}, \qquad \hat{f}_n(\nu) = \frac{\sigma_n^2\theta_n}{\pi(\theta_n^2 + \nu^2)}.$$

It is easily seen that the supremum of $[\{\hat{f}_n(\nu) - f(\nu)\}/f(\nu)]^2$ is attained at either $\nu = 0$ or $\nu = \infty$, depending on the relative signs of $\sigma_n^2\theta_n - \sigma^2\theta$ and $\theta_n - \theta$. Thus, (3.15) is fulfilled if both $n^{1/2}(\theta_n - \theta)$ and $n^{1/2}(\sigma_n^2 - \sigma^2)$ are bounded. This corresponds to the worst-case scenario of increasing-domain asymptotics above, where $a_n = O(1)$. We cannot expect a condition which guarantees (3.6) for all configurations of sampling locations to give a better result than this.

Lemma 3.3 can also help us to check (3.7) of Lemma 3.2. It is easy to see that (3.16) is fulfilled if σ_n^2 and θ_n are bounded away from 0 and ∞. Thus, a combination of Lemma 3.2 and Lemma 3.3 shows that $\{P_n\}$ and $\{\hat{P}_n\}$ are contiguous under (4.5) if

$$\theta_n - \theta = O\left(\left(\frac{a_n}{n}\right)^{-1/2}\right), \qquad \sigma_n^2\theta_n - \sigma^2\theta = O(n^{-1/2}). \tag{4.7}$$

Under infill asymptotics, these rates are achieved by the maximum likelihood estimators, as proved by Ying (1991). Under increasing-domain asymptotics restricted maximum likelihood estimators exist achieving (4.7) (Cressie and Lahiri 1996). We conjecture that also for $a_n \to \infty$ but $a_n = o(n)$ these rates can be achieved.

The exponential covariance function model in $\mathbb{R}$, given by (4.4), can be extended to more dimensions in two ways. One is retaining isotropy, leading to

$$C_{\theta,\sigma^2}(t) = \sigma^2 \exp(-\theta\|t\|), \qquad t \in \mathbb{R}^d, \tag{4.8}$$

where $\|t\| = (\sum_{j=1}^d t_j^2)^{1/2}$ is the Euclidean length of the vector t. In this model the identifiability problem remains for $d \leqslant 3$, i.e. θ and σ^2 are not identifiable and only the product $\theta\sigma^2$ can be estimated at $\sqrt{n}$-rate.

A second extension,

$$C_{\theta_1,\ldots,\theta_d,\sigma^2}(t) = \sigma^2 \exp\left(-\sum_{j=1}^{d} \theta_j |t_j| \right), \tag{4.9}$$

is mathematically more tractable. It has generally been motivated by applications other than those involving spatial data, such as the modelling of computer experiments. In such a context, the distance between two points may be defined in terms of the number of nodes between them on some grid, leading naturally to consideration of this form of covariance function. In Ying (1993) it is shown that $\theta_1, \ldots, \theta_d$ and σ^2 are identifiable, and asymptotic normality of the maximum likelihood estimators $\hat{\theta}_1, \ldots, \hat{\theta}_d$ and $\hat{\sigma}_n^2$ is proved. The same model was considered by van der Vaart (1996), who proved local asymptotic normality for this model under a two-dimensional regularly spaced grid. It is well known that local asymptotic normality implies contiguity (Bickel *et al.*, 1993, pp. 16–17), so the local asymptotic normality proved by van der Vaart (1996) guarantees that the maximum likelihood estimators are such that $\{P_n\}$ and $\{\hat{P}_n\}$ are contiguous almost surely.

5. Estimation of prediction error

Estimation of the kriging error in the case where the underlying covariance function C is known is straightforward; if C is known then formula (2.5) gives the result. An obvious thing to do when C is unknown is to use (2.5), replacing unknown quantities by their estimated counterparts. Using the notation of (2.6)–(2.9), we then estimate $\mathrm{var}_C(\hat{e}_n(x_0))$ by

$$\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0)) = \hat{\sigma}_n^2 - \hat{\omega}_n^{\mathrm{T}} \hat{\Omega}_n^{-1} \hat{\omega}_n + \frac{(\mathbf{1}^{\mathrm{T}} \hat{\Omega}_n^{-1} \hat{\omega}_n - 1)^2}{\mathbf{1}^{\mathrm{T}} \hat{\Omega}_n^{-1} \mathbf{1}}. \tag{5.1}$$

Computation of (5.1) requires hardly any additional effort, since most of the quantities are needed to compute $\hat{\alpha}_n$ in the first place. It can be shown as a by-product of Theorem 2.1 that $\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))$ is a consistent estimator of $\mathrm{var}_C(\hat{e}_n(x_0))$.

Lemma 5.1. *Under the conditions of Theorem 2.1,*

$$\frac{\mathrm{var}_{\hat{C}_n}(\hat{e}_n(x_0))}{\mathrm{var}_C(\hat{e}_n(x_0))} \to 1 \quad \mathbb{P}\text{-}a.s. \tag{5.2}$$

Proof. The result follows immediately from (2.17). □

Acknowledgements

Research was funded by the EU TMR network on Computational and Statistical Aspects of the Analysis of Spatial Data (ERB-FMRX-CT96-0095). The authors are grateful to W.R. van Zwet and two referees for a number of helpful remarks.

References

Bickel, P.J., Klaasen, C.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, MD: Johns Hopkins University Press.

Christakos, G. (1984) On the problem of permissible covariance and variogram models. *Water Resources Res.*, **20**, 251–265.

Christensen, R. (1991) *Linear Models for Multivariate, Time Series and Spatial Data.* Berlin: Springer-Verlag.

Chung, K.L. (1974) *A Course in Probability Theory*, 2nd edn. New York: Academic Press.

Cressie, N.A.C. (1991) *Statistics for Spatial Data.* New York: Wiley.

Cressie, N. and Lahiri, S.N. (1996) Asymptotics for REML estimation of spatial covariance parameters. *J. Statist. Plann. Inference*, **50**, 327–241.

Diamond P. and Armstrong, M. (1984) Robustness of variograms and conditioning of kriging matrices. *Math. Geol.*, **16**, 809–822.

Hall, P. and Patil, P. (1994) Properties of nonparametric estimators of autocovariance for stationary random fields. *Probab. Theory Related Fields*, **99**, 399–424.

Ibragimov, I.A. and Rozanov, Y.A. (1978) *Gaussian Random Processes.* Berlin: Springer-Verlag.

Journel, A.G. and Huijbregts, C.J. (1978) *Mining Geostatistics*. New York: Academic Press.

Lahiri, S.N. (1996) On inconsistency of estimators based on spatial data under infill asymptotics. *Sankhyā Ser. A*, **58**, 403–417.

Lahiri, S.N. (1997) Asymptotic distribution of the empirical spatial cumulative distribution function predictor and prediction bands based on a subsampling method. Technical report, Department of Statistics, Iowa State University.

Oosterhoff, J. and van Zwet, W.R. (1979) A note on contiguity and Hellinger distance. In J. Jurečková (ed.), *Contributions to Statistics: Jaroslav Hájek Memorial Volume*, pp. 157–166. Dordrecht: Reidel.

Prakasa Rao, B.L.S. (1987) *Asymptotic Theory of Statistical Inference*. New York: Wiley.

Priestley, M.B. (1981) *Spectral Analysis and Time Series, Volume 1: Univariate Series.* London: Academic Press.

Rao, C.R. (1965) *Linear Statistical Inference and Its Applications*, 2nd edn. New York: Wiley.

Roussas, G.G. (1972) *Contiguity of Probability Measures: Some Applications in Statistics.* Cambridge: Cambridge University Press.

Stein, M.L. (1988) Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.*, **16**, 55–63.

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

Stein, M.L. and Handcock, M.S. (1989) Some asymptotic properties of kriging when the covariance function is misspecified. *Math. Geol.*, **21**, 171–190.

van der Vaart, A. (1996) Maximum likelihood estimation under a spatial sampling scheme. *Ann. Statist.*, **24**, 2049–2057.

Warnes, J.J. (1986) A sensitivity analysis for universal kriging. *Math. Geol.*, **18**, 653–676.

Yakowitz, S.J. and Szidarovsky, F. (1985) A comparison of kriging with nonparametric regression methods. *J. Multivariate Anal.*, **16**, 21–53.

Ying, Z. (1991) Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *J. Multivariate Anal.*, **36**, 280–296.

Ying, Z. (1993) Maximum likelihood estimation of parameters under a spatial sampling scheme. *Ann. Statist.*, **21**, 1567–1590.

Received May 1998 and revised April 2000