Biometrika (2013), **100**, 2, *pp.* 473–484
doi: 10.1093/biomet/ass079
Advance Access publication 7 February 2013

The role of the range parameter for estimation and prediction in geostatistics

By C. G. KAUFMAN and B. A. SHABY

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.
cgk@stat.berkeley.edu bshaby@stat.berkeley.edu

Summary

Two canonical problems in geostatistics are estimating the parameters in a specified family of stochastic process models and predicting the process at new locations. We show that asymptotic results for a Gaussian process over a fixed domain with Matérn covariance function, previously proven only in the case of a fixed range parameter, can be extended to the case of jointly estimating the range and the variance of the process. Moreover, we show that intuition and approximations derived from asymptotics using a fixed range parameter can be problematic when applied to finite samples, even for large sample sizes. In contrast, we show via simulation that performance is improved and asymptotic approximations are applicable for smaller sample sizes when the parameters are jointly estimated. These effects are particularly apparent when the process is mean square differentiable or the effective range of spatial correlation is small.

Some key words: Covariance estimation; Gaussian process; Infill asymptotics; Matérn covariance; Spatial statistics.

1. Introduction

The analysis of point-referenced spatial data relies almost exclusively on a single construct: the stationary Gaussian process with a parametric mean and covariance. Given its prominent role, it is perhaps surprising that the theoretical properties of inference under this model remain incompletely understood. Consider a canonical problem in geostatistics, that of predicting the value of a spatial process with unknown model parameters at locations not contained in the dataset. Stein (2010) gives an overview of asymptotic issues for both estimation and prediction.

Stein (1999) makes a compelling case for using the Matérn covariance model for the Gaussian process $\{Z(s), s \in D \subseteq \Re^d\}$, with

$$\mathrm{cov}\{Z(s_i), Z(s_j)\} = \sigma^2 K(s_i - s_j; \rho, \nu) = \frac{\sigma^2(\|s_i - s_j\|/\rho)^\nu}{\Gamma(\nu)2^{\nu-1}} \mathcal{K}_\nu(\|s_i - s_j\|/\rho), \qquad (1)$$

where $\sigma^2, \rho, \nu > 0$, and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν (Abramowitz & Stegun, 1992, § 9.6). The range parameter ρ controls the rate of decay with distance. This model is particularly attractive because of its flexibility in representing the smoothness of the Gaussian process by varying ν (Stein, 1999).

Zhang (2004) showed that for fixed ν and $d \leqslant 3$, σ^2 and ρ cannot be consistently estimated under infill or fixed-domain asymptotics, where the sampling domain is fixed as the number of observations increases to infinity. However, he also showed that if one fixes ρ at an arbitrary value, then the maximum likelihood estimator for $c = \sigma^2/\rho^{2\nu}$ is consistent. This result follows from a more fundamental result in Zhang (2004) concerning equivalence, or mutual absolute

continuity, of Gaussian measures on bounded domains. Stein (1988, 1990, 1993, 1999) provides conditions under which predictions using a misspecified covariance function are asymptotically efficient and associated standard errors converge almost surely to their targets under infill asymptotics. One such condition is that the misspecified Gaussian measure and the true one are equivalent, providing a link to the results in Zhang (2004).

These results have led to a growing tendency in the applied literature to regard ρ as secondarily influential. For example, Zhang & Wang (2010) find that fixing ρ at arbitrary large values has little impact on predictive performance, and Gneiting et al. (2010) argue that specifying a single ρ for all variables in a multivariate model is not restrictive. Sahu et al. (2007) choose from a small number of fixed values of ρ, while Anderes et al. (2012) produce predictions without ever estimating ρ. These authors borrow intuition from the asymptotic results of Stein (1988), Zhang (2004), and others, and present some variation of the conclusion that fixing ρ at an incorrect value is asymptotically just as good as using the true value. However, as we will show, this intuition cannot be transferred so readily to the finite sample case, as ρ can be quite influential even for large samples. Here, we show that the asymptotics can be extended to joint estimation of σ^2 and ρ, and we demonstrate via simulation that methods that estimate rather than fix ρ are superior on a variety of metrics, despite being asymptotically identical.

2. ASYMPTOTIC THEORY FOR ESTIMATION AND PREDICTION

2·1. *Preliminaries*

Let $Z = \{Z(s), s \in D \subset \Re^d\}$ be a stochastic process on a bounded domain D, with $d = 1, 2$, or 3. Let $G(0, \sigma^2 K_\theta)$ denote the mean zero stationary Gaussian measure for Z with marginal variance $\sigma^2 > 0$ and correlation function K_θ, depending on parameters $\theta \in \Theta \subseteq \Re^p$. For a sampling design $S_n = \{s_1, \ldots, s_n\} \subset D$, we observe $Z_n = \{Z(s_1), \ldots Z(s_n)\}^{\mathrm{T}}$. Our tasks are to use Z_n to estimate σ^2 and θ and to predict $Z(s_0)$ for some location $s_0 \in D$, not in S_n. Our results concern the behaviour of these estimators and predictors under infill asymptotics.

Let $G(0, \sigma^2 K_{\rho,\nu})$ denote a mean zero Gaussian measure with the Matérn covariance function and known smoothness parameter ν. Our focus is on the role played by the range parameter ρ in this model, namely to show that several important results that have been provable only in the case of fixing ρ at an arbitrary value can be extended to the case that ρ is estimated.

The reason that it is justifiable to fix ρ, at least in an asymptotic sense, follows from a result by Zhang (2004) stating that when $d \leqslant 3$, two Gaussian measures with different values of ρ can be equivalent. Specifically, Theorem 2 of Zhang (2004) states that for fixed $\nu > 0$, $G(0, \sigma_0^2 K_{\rho_0,\nu})$ and $G(0, \sigma_1^2 K_{\rho_1,\nu})$ are equivalent on bounded domains if and only if $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$.

The parameter $c = \sigma^2/\rho^{2\nu}$ is what Stein (1999) calls a microergodic parameter. Stein (1999, p. 175) suggests reparameterizing into microergodic and non-microergodic components of the parameter vector, which we here define as c and ρ, respectively. He conjectures that if all model parameters are estimated by maximum likelihood, the asymptotic behaviour of the maximum likelihood estimator for the microergodic parameter is the same as if the non-ergodic component were known. In the next section, we outline existing results that concern the asymptotic behaviour for the maximum likelihood estimator for c when ρ is fixed and we extend them to the case that ρ is estimated, showing that Stein's conjecture is true for the Matérn model.

2·2. *Estimation of covariance parameters*

Theorem 2 of Zhang (2004) has an important corollary for estimation, namely that there do not exist consistent estimators of σ^2 or ρ under infill asymptotics. However, this corollary does

not imply that the data contain no information about σ^2 and ρ individually. Indeed, simulations show that sampling distributions for the maximum likelihood estimators can in many cases be quite concentrated about the true values, even though the estimators are not consistent (Zhang, 2004). Some intuition can be given by appealing to the asymptotic framework of increasing the domain of observations while keeping the density constant. Mardia & Marshall (1984) give regularity conditions, which hold under an increasing-domain framework, under which the maximum likelihood estimators for all model parameters are consistent and asymptotically normal. Any finite set of observation locations could conceivably be a member in a sequence under either fixed-domain or increasing-domain asymptotics. Because the increasing-domain framework can be mimicked by fixing the domain but decreasing the range parameter (Zhang & Zimmerman, 2005), it is not surprising that when the true range parameter is small relative to the sampling domain, it can be well estimated from data.

The likelihood function for σ^2 and ρ under the Matérn model with fixed $\nu > 0$ is

$$\mathcal{L}_n(\sigma^2, \rho) = (2\pi\sigma^2)^{-n/2}|\Gamma_n(\rho)|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} Z_n^{\mathrm{T}} \Gamma_n(\rho)^{-1} Z_n\right\}, \tag{2}$$

where $\Gamma_n(\rho)$ is the matrix with entries $K(s_i - s_j; \rho, \nu)$ $(i, j = 1, \ldots, n)$ for K defined as in (1). We consider two types of estimators obtained by maximizing (2). The first fixes $\hat{\rho}_n = \rho_1$ for all n and maximizes $\mathcal{L}_n(\sigma^2, \rho_1)$. The second maximizes (2) over both σ^2 and ρ. In either case, we may write $\hat{\sigma}_n^2(\hat{\rho}_n) = \arg\max_{\sigma^2} \mathcal{L}_n(\sigma^2, \hat{\rho}_n) = Z_n^{\mathrm{T}} \Gamma_n(\hat{\rho}_n)^{-1} Z_n / n$, where $\hat{\rho}_n$ is either ρ_1 or the value that maximizes the profile likelihood for ρ, when a unique maximizer exists. In most cases the latter estimator must be found numerically. We may likewise express the corresponding estimators of $c = \sigma^2/\rho^{2\nu}$ as a function of $\hat{\rho}_n$, namely

$$\hat{c}_n(\hat{\rho}_n) = \hat{\sigma}_n^2(\hat{\rho}_n)/\hat{\rho}_n^{2\nu} = Z_n^{\mathrm{T}} \Gamma_n(\hat{\rho}_n)^{-1} Z_n / (n\hat{\rho}_n^{2\nu}). \tag{3}$$

The following result defines the asymptotic behaviour of $\hat{c}_n(\rho_1)$ for an arbitrary fixed value $\rho_1 > 0$. It combines Theorem 3 of Zhang (2004) and Theorem 3 of Wang & Loh (2011).

THEOREM 1. *Let S_n be an increasing sequence of subsets of D. Then under $G(0, \sigma_0^2 K_{\rho_0,\nu})$ as $n \to \infty$,*

(a) $\hat{c}_n(\rho_1) \to \sigma_0^2/\rho_0^{2\nu}$ *almost surely, and*
(b) $n^{1/2}\{\hat{c}_n(\rho_1) - \sigma_0^2/\rho_0^{2\nu}\} \to N\{0, 2(\sigma_0^2/\rho_0^{2\nu})^2\}$ *in distribution.*

A key contribution of the current paper is to show that Theorem 1 can be used to prove that the maximum likelihood estimator $\hat{c}_n(\hat{\rho}_n)$ has the same asymptotic behaviour as does $\hat{c}_n(\rho_1)$ for any ρ_1, including the true value ρ_0. We make use of the following lemma, which shows that $\hat{c}_n(\hat{\rho}_n)$ is monotone when viewed as a function of $\hat{\rho}_n$.

LEMMA 1. *Let $S_n = \{s_1, \ldots, s_n \in D \subseteq \Re^d\}$ denote any set of locations in any dimension. Fix $\nu > 0$ and define $\Gamma_n(\rho)$ to be the matrix with entries $K(s_i - s_j; \rho, \nu)$ as in (1). Define $\hat{c}_n(\rho) = Z_n^{\mathrm{T}} \Gamma_n(\rho)^{-1} Z_n/(n\rho^{2\nu})$. Then for any $0 < \rho_1 < \rho_2$, $\hat{c}_n(\rho_2) \leqslant \hat{c}_n(\rho_1)$ for any vector Z_n.*

Proof. Let $0 < \rho_1 < \rho_2$. The difference

$$\hat{c}_n(\rho_1) - \hat{c}_n(\rho_2) = Z_n^{\mathrm{T}}\{\rho_1^{-2\nu}\Gamma_n(\rho_1)^{-1} - \rho_2^{-2\nu}\Gamma(\rho_2)^{-1}\}Z_n/n$$

is nonnegative for any Z_n if the matrix $A = \rho_1^{-2\nu}\Gamma_n(\rho_1)^{-1} - \rho_2^{-2\nu}\Gamma_n(\rho_2)^{-1}$ is positive semidefinite. By Corollary 7.7.4(a) of Horn & Johnson (1990, p. 473), A is positive semidefinite

if and only if the matrix $B = \rho_2^{2\nu}\Gamma_n(\rho_2) - \rho_1^{2\nu}\Gamma_n(\rho_1)$ is positive semidefinite. The entries of B may be expressed in terms of a function $K_B : \Re^d \to \Re$, with

$$B_{ij} = K_B(s_i - s_j) = \rho_2^{2\nu} K(\|s_i - s_j\|; \rho_2, \nu) - \rho_1^{2\nu} K(\|s_i - s_j\|; \rho_1, \nu),$$

and B is positive semidefinite if K_B is a positive definite function. Define

$$\begin{aligned} f_B(\omega) &= \frac{1}{(2\pi)^d} \int_{\Re^d} \mathrm{e}^{-i\omega^{\mathrm{T}}x} K_B(x)\,\mathrm{d}x \\ &= \frac{1}{(2\pi)^d} \left\{ \rho_2^{2\nu} \int_{\Re^d} \mathrm{e}^{-i\omega^{\mathrm{T}}x} K(x; \rho_2, \nu)\,\mathrm{d}x - \rho_1^{2\nu} \int_{\Re^d} \mathrm{e}^{-i\omega^{\mathrm{T}}x} K(x; \rho_1, \nu)\,\mathrm{d}x \right\}. \end{aligned} \tag{4}$$

Both integrals in (4) are finite, with

$$\frac{1}{(2\pi)^d} \int_{\Re^d} e^{-i\omega^{\mathrm{T}}x} K(x; \rho, \nu)\,\mathrm{d}x = \frac{\Gamma(\nu + d/2)}{\pi^{d/2}\Gamma(\nu)} \rho^{-2\nu} \left(\rho^{-2} + \|\omega\|^2\right)^{-(\nu+d/2)},$$

the spectral density of the Matérn correlation function. Therefore,

$$f_B(\omega) = \frac{\Gamma(\nu + d/2)}{2^d \pi^{3d/2}\Gamma(\nu)} \left\{ \left(\rho_2^{-2} + \|\omega\|^2\right)^{-(\nu+d/2)} - \left(\rho_1^{-2} + \|\omega\|^2\right)^{-(\nu+d/2)} \right\}.$$

To show that K_B is positive definite it suffices to show that $f_B(\omega)$ is positive for all ω. This is clear because $0 < \rho_1 < \rho_2$. Therefore $\hat{c}_n(\rho_2) \leqslant \hat{c}_n(\rho_1)$ for any vector Z_n. □

We can now use Theorem 1 to prove a more general result for the maximum likelihood estimator when the parameter space for ρ is a bounded interval. This condition was also used by Ying (1991), who proved Theorem 2 when D is the unit interval and $\nu = 1/2$. These bounds are not restrictive in practice, as the interval may be taken to be arbitrarily large.

THEOREM 2. *Let S_n be an increasing sequence of subsets of D. Suppose $(\sigma_0^2, \rho_0)^{\mathrm{T}} \in (0, \infty) \times [\rho_L, \rho_U]$, for any $0 < \rho_L < \rho_U < \infty$. Let $(\hat{\sigma}_n^2, \hat{\rho}_n)^{\mathrm{T}}$ maximize (2) over $(0, \infty) \times [\rho_L, \rho_U]$. Then under $G(0, \sigma_0^2 K_{\rho_0,\nu})$,*

(a) $\hat{\sigma}_n^2/\hat{\rho}_n^{2\nu} \to \sigma_0^2/\rho_0^{2\nu}$ *almost surely, and*
(b) $n^{1/2}\left(\hat{\sigma}_n^2/\hat{\rho}_n^{2\nu} - \sigma_0^2/\rho_0^{2\nu}\right) \to N\left\{0, 2\left(\sigma_0^2/\rho_0^{2\nu}\right)^2\right\}$ *in distribution.*

Proof. By assumption, $\rho_L \leqslant \hat{\rho}_n \leqslant \rho_U$ for every n. Define two sequences, $\hat{c}_n(\rho_L)$ and $\hat{c}_n(\rho_U)$, according to (3). By Lemma 1, $\hat{c}_n(\rho_L) \leqslant \hat{c}_n(\hat{\rho}_n) = \hat{\sigma}_n^2/\hat{\rho}_n^{2\nu} \leqslant \hat{c}_n(\rho_U)$ for all n with probability one. Combining this with Theorem 1 applied to $\hat{c}_n(\rho_L)$ and $\hat{c}_n(\rho_U)$ implies the result. □

Theorem 2 is useful because it applies to the procedure that is most often adopted in practice, of allowing the range parameter to be estimated over a bounded interval. In fact, the method of proof in Theorem 2 works for any bounded sequence $\hat{\rho}_n$, provided that $\hat{\sigma}_n^2$ is defined as in (3). This would include, for example, estimating ρ using the variogram and inserting it into (3), but not joint estimation of ρ and σ^2 using the variogram. In practice, the bounds for numerical optimization of ρ can be chosen to be arbitrarily wide, subject to numerical stability.

A similar method of proof can be used to show consistency and asymptotic normality of the maximum tapered likelihood estimator proposed by Kaufman et al. (2008). The online Supplementary Material contains analogues of Lemma 1 and Theorem 2 for this estimator.

Arguments following from Zhang (2004) would suggest that the range parameter may be fixed in practice. However, as we shall show in § 3, the estimator $\hat{c}_n(\rho_1)$ can often display sizeable bias, making the approximation in Theorem 1 quite inaccurate. Confidence intervals constructed using Theorem 1 can, due to this bias, have empirical coverage probabilities very near to zero in some cases. In contrast, we will show that confidence intervals for c constructed using Theorem 2 have close to nominal coverage even for moderate sample sizes.

2·3. *Prediction at new locations*

We now consider predicting the value of the process at a new location s_0 not in S_n. Stein (1988, 1990, 1993, 1999) has considered this problem when an incorrect model is used. Predictors under the wrong model can be consistent under relatively weak conditions. Our focus is therefore on two other desirable properties, asymptotic efficiency and asymptotically correct estimation of prediction variance. In a seminal paper, Stein (1988) showed that both of these properties hold when the model used is equivalent to the true measure. In the case of the Matérn covariance, Theorem 2 of Zhang (2004) indicates that this holds for a model with the correct ν and microergodic parameter $\sigma^2/\rho^{2\nu}$. This has led to statements in the literature to the effect that the parameter $c = \sigma^2/\rho^{2\nu}$ can be consistently estimated, and this is what matters for prediction. While this statement contains an element of truth, we will argue in this section that it can also be somewhat misleading, both in an asymptotic sense, as well as in guiding choices for applications.

Define

$$\hat{Z}_n(\rho) = \gamma_n(\rho)^{\mathrm{T}}\Gamma_n(\rho)^{-1}Z_n, \tag{5}$$

where $\{\gamma_n(\rho)\}_i = K(s_0 - s_i; \rho, \nu)$ and $\{\Gamma_n(\rho)\}_{ij} = K(s_i - s_j; \rho, \nu)$ $(i, j = 1, \ldots, n)$. The predictor $\hat{Z}_n(\rho)$ is the best linear unbiased predictor for $Z_0 = Z(s_0)$ under a presumed model $G(0, \sigma^2 K_{\rho,\nu})$ for any value of σ^2. This predictor does not depend on σ^2, only on ρ and ν. Therefore, any intuition that one can fix $\rho = \rho_1$, and that plug-in predictions will improve with n due in any way to convergence of $\hat{c}_n(\rho_1)$ with n, is a misunderstanding of asymptotic results. Equivalence, although sufficient for asymptotic efficiency, is not necessary. The way in which c is relevant for prediction concerns estimates of the mean squared error of the predictor. Under model $G(0, \sigma_0^2 K_{\rho_0,\nu})$, this is

$$\begin{aligned}\mathrm{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho) - Z_0\} = \sigma_0^2\{1 - 2\gamma_n(\rho)^{\mathrm{T}}\Gamma_n(\rho)^{-1}\gamma_n(\rho_0) \\ + \gamma_n(\rho)^{\mathrm{T}}\Gamma_n(\rho)^{-1}\Gamma_n(\rho_0)\Gamma_n(\rho)^{-1}\gamma_n(\rho)\},\end{aligned} \tag{6}$$

where $\gamma_n(\rho_0)$ and $\Gamma_n(\rho_0)$ are defined analogously to their counterparts using ρ. In the case that $\rho = \rho_0$, this expression simplifies to

$$\mathrm{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_0) - Z_0\} = \sigma_0^2\{1 - \gamma_n(\rho_0)^{\mathrm{T}}\Gamma_n(\rho_0)^{-1}\gamma_n(\rho_0)\}. \tag{7}$$

In practice, it is common to estimate the model parameters and then insert them into (5) and (7), treating them as known. The asymptotic properties of this procedure, so-called plug-in prediction, are quite difficult to obtain. Instead, most theoretical development has been under a framework in which plug-in parameters are fixed, rather than being estimated from observations at an increasing sequence of locations. We will indicate how these results may be extended to include estimation of the variance parameter σ^2 with a fixed value of ρ, making precise the sense in which the statement regarding c at the beginning of this section should be interpreted.

The following result is an application of Theorems 1 and 2 of Stein (1993).

THEOREM 3. *Suppose $G(0, \sigma_0^2 K_{\rho_0,\nu})$ and $G(0, \sigma_1^2 K_{\rho_1,\nu})$ are two Gaussian process measures on D with the same value of $\nu > 0$.*

(a) *As $n \to \infty$,*

$$\frac{\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_1) - Z_0\}}{\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_0) - Z_0\}} \to 1.$$

(b) *Furthermore, if $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$, then as $n \to \infty$,*

$$\frac{\text{var}_{\sigma_1^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}}{\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_1) - Z_0\}} \to 1. \tag{8}$$

Proof. Let f_0 and f_1 be the spectral densities corresponding to $\sigma_0^2 K_{\rho_0,\nu}$ and $\sigma_1^2 K_{\rho_1,\nu}$. The result follows from noting that the function $f_0(\omega)\|\omega\|^{2\nu+d}$ is bounded away from zero and infinity as $\|\omega\| \to \infty$ and that

$$\lim_{\|\omega\|\to\infty} \frac{f_1(\omega)}{f_0(\omega)} = \frac{\sigma_1^2/\rho_1^{2\nu}}{\sigma_0^2/\rho_0^{2\nu}}.$$

These two conditions satisfy those needed for Theorems 1 and 2 of Stein (1993). □

The implication of part (a) of Theorem 3 is that if the correct value of ν is used, any value of ρ will give asymptotic efficiency. The condition $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$ is not necessary for asymptotic efficiency, but it does provide asymptotically correct estimates of mean squared prediction error. The numerator in (8) is the naive mean squared error for $\hat{Z}_n(\sigma_1^2, \rho_1)$, assuming model $G(0, \sigma_1^2 K_{\rho_1,\nu})$, whereas the denominator is the true mean squared error for $\hat{Z}_n(\sigma_1^2, \rho_1)$, under model $G(0, \sigma_0^2 K_{\rho_0,\nu})$. We now show the same convergence happens if ρ is fixed at ρ_1 but σ^2 is estimated via maximum likelihood. This is an extension of part (b) of Theorem 3. Part (a) needs no extension, since the form of the predictor itself does not depend on σ^2.

THEOREM 4. *Suppose $G(0, \sigma_0^2 K_{\rho_0,\nu})$ is a Gaussian process measure on D. Fix $\rho_1 > 0$. For a sequence of observations Z_n on an increasing sequence of subsets S_n of D, define $\hat{\sigma}_n^2 = Z_n^{\mathrm{T}} \Gamma_n(\rho_1)^{-1} Z_n/n$. Then almost surely under $G(0, \sigma_0^2 K_{\rho_0,\nu})$, as $n \to \infty$,*

$$\frac{\text{var}_{\hat{\sigma}_n^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}}{\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_1) - Z_0\}} \to 1.$$

Proof. Define $\sigma_1^2 = \sigma_0^2(\rho_1/\rho_0)^{2\nu}$. Then write

$$\frac{\text{var}_{\hat{\sigma}_n^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}}{\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_1) - Z_0\}} = \frac{\text{var}_{\hat{\sigma}_n^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}}{\text{var}_{\sigma_1^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}} \frac{\text{var}_{\sigma_1^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}}{\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_1) - Z_0\}}.$$

By Theorem 3, $\text{var}_{\sigma_1^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}/\text{var}_{\sigma_0^2,\rho_0}\{\hat{Z}_n(\rho_1) - Z_0\} \to 1$. So we need show only that $\text{var}_{\hat{\sigma}^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}/\text{var}_{\sigma_1^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\} \to 1$ almost surely under $G(0, \sigma_0^2 K_{\rho_0,\nu})$. By (7), $\text{var}_{\hat{\sigma}_n^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\}/\text{var}_{\sigma_1^2,\rho_1}\{\hat{Z}_n(\rho_1) - Z_0\} = \hat{\sigma}_n^2/\sigma_1^2$. Under $G(0, \sigma_1^2 K_{\rho_1,\nu})$, $\hat{\sigma}_n^2$ is equal in

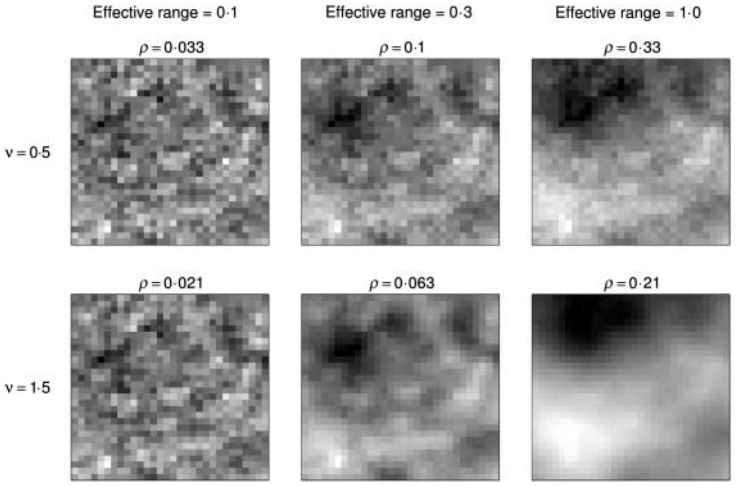

Fig. 1. Simulated random fields on $[0, 1]^2$ under parameter settings used in the simulation study. The value of the range parameter ρ corresponding to each ν and effective range combination is also indicated.

distribution to σ_1^2/n times a χ^2 random variable with n degrees of freedom and hence converges almost surely to σ_1^2 as $n \to \infty$. Because $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$, Theorem 2 of Zhang (2004) gives that $G(0, \sigma_0^2 K_{\rho_0,\nu})$ and $G(0, \sigma_1^2 K_{\rho_1,\nu})$ are equivalent, so that $\hat{\sigma}_n^2 \to \sigma_1^2$ almost surely under $G(0, \sigma_0^2 K_{\rho_0,\nu})$ as well. □

We conjecture that the asymptotic behaviour in part (a) of Theorem 3 and Theorem 4 still holds if ρ_1 is replaced by $\hat{\rho}_n$, the maximum likelihood estimator, although proving this has been elusive in cases of practical interest (Putter & Young, 2001).

3. Simulation study

3·1. *Set-up*

Fixing the range parameter is supported by asymptotic results, and it is computationally efficient in practice. However, it is unclear to what degree asymptotic results are appropriate in guiding our choices for applied problems with finite sample sizes. To systematically explore this, we simulate data under a mean zero Gaussian process model for a variety of settings chosen to mimic the range of behaviour we might observe in practice, and we compare the performance of inference procedures that either fix or estimate the range parameter.

We simulate data in the unit square with $\nu = 0{\cdot}5$ or $1{\cdot}5$ and $\sigma^2 = 1$. We also use three effective ranges, choosing values of ρ such that the correlation decays to $0{\cdot}05$ at distances of $0{\cdot}1$, $0{\cdot}3$, or 1. Figure 1 illustrates the effect of these parameter settings. As we shall see, whether a particular sample size is large enough such that finite sample properties are well approximated by asymptotic results depends both on ν and on the effective range of the process.

We also vary the sample size in the simulation, taking $n = 400$, 900, and 1600. To avoid numerical issues, sampling locations are constructed using a perturbed grid. We construct a 67×67 regular grid with coordinates from $0{\cdot}005$ to $0{\cdot}995$ in increments of $0{\cdot}015$ in each dimension. To each gridpoint, we add a uniform $[-0{\cdot}005, 0{\cdot}005]^2$ perturbation. Each of the resulting locations is at least $0{\cdot}005$ units from its nearest neighbour. We then choose random subsets of these locations

to be our observation locations, with each sample size containing the points from smaller sample sizes. We predict over a 50×50 regular grid of locations over $[0, 1]^2$.

For each parameter setting, we simulate 1000 realizations of the Gaussian process observed at the union of $n = 1600$ observation and $m = 2500$ prediction locations. For each dataset and sample size, we estimate σ^2 and ρ by numerically maximizing the profile likelihood for ρ and inserting the result into the corresponding closed-form estimator for σ^2.

We also calculate $\hat{\sigma}_n^2(\rho_1) = Z_n^{\mathrm{T}}\Gamma_n(\rho_1)^{-1}Z_n/n$ for ρ_1 equal to 0·2, 0·5, 1, 2, and 5 times the true value of ρ. Corresponding to each of these parameter estimates, we also construct 95% confidence intervals for $c = \sigma^2/\rho^{2\nu}$ using the normal approximation provided by Theorem 1 when ρ is fixed and Theorem 2 when ρ is estimated. Finally, we construct kriging predictors and estimated standard errors for each of the $m = 2500$ prediction locations by inserting parameter estimates into (5) and (7).

Optimization was carried out using the R (R Development Core Team, 2013) function optim with the L-BFGS-B option, which we restricted to the interval $\rho \in [\varepsilon, 15\rho_0]$, where ε is defined by machine precision, about 10^{-16} on our machine. Neither endpoint was ever returned.

Many of the results show a similar pattern, which can be summarized as follows. The performance of the maximum likelihood estimator, maximizing over both σ^2 and ρ, is generally very good, especially for $n = 1600$. Procedures using a fixed ρ are almost always worse, although the differences are minimal under certain settings. These tend to be for $\nu = 1/2$, corresponding to processes that are not mean square differentiable, and a large effective range. In these cases, particularly when ρ is fixed at something larger than its true value, the estimators and predictors can still perform well. This agrees with some examples in the literature, for which $\nu = 1/2$ and large effective ranges were used (Zhang & Wang, 2010; Wang & Loh, 2011). When the process is smooth, with $\nu = 1{\cdot}5$, and/or the true range of spatial correlation is small, estimation and prediction are markedly improved by estimating ρ via maximum likelihood.

3·2. *Parameter estimation*

Given the asymptotic results in Zhang (2004) and Wang & Loh (2011) for $\hat{c}_n(\rho_1)$ for fixed ρ_1, it is tempting to believe that this estimator can adapt to incorrectly specified values of ρ. While this is true asymptotically, our simulation results show that in many cases this adaptation requires a very large value of n; sampling distributions can be highly biased and can approach the truth very slowly as n increases. Figure 2 illustrates this when $\nu = 1{\cdot}5$ and the effective range is 0·3. Sampling distributions for $\hat{c}_n(\rho_1)$ are noticeably biased. As we expect from Theorem 1, these biases decrease with n, although even when $n = 1600$ the true value of c lies far in the tail of the sampling distribution. In contrast, the sampling distributions for the maximum likelihood estimator $\hat{c}_n(\hat{\rho}_n)$ have negligible bias. Indeed, they behave very similarly to those for the estimator of c that fixes ρ at the truth. Similar effects can be seen for other values of ν and effective range. See Tables S-1 and S-3 in the Supplementary Material for the relative bias of different estimators of c.

If Theorem 1 is used to construct confidence intervals and n is not large enough for the normal approximation to be appropriate, the coverage can be disastrously low. Table 1 shows empirical coverage rates for confidence intervals constructed as $\hat{c}_n(\hat{\rho}_n) \pm 1{\cdot}96\{2\hat{c}_n(\hat{\rho}_n)^2/n\}^{1/2}$ for $\hat{\rho}_n$ equal to the maximum likelihood estimator or a fixed ρ_1. Theorems 1 and 2 imply that these intervals are asymptotically valid. Not surprisingly, however, given the large biases observed when ρ is fixed, the differences in the empirical coverage rates between fixed and estimated ρ are striking, even when n is large. In many cases the coverage for intervals constructed using $\hat{c}_n(\rho_1)$ was 0%, to within Monte Carlo sampling error. Coverage is best when ν is small and effective range is large. For fixed ρ_1, it also appears to be better to choose $\rho_1 > \rho_0$ than $\rho_1 < \rho_0$.

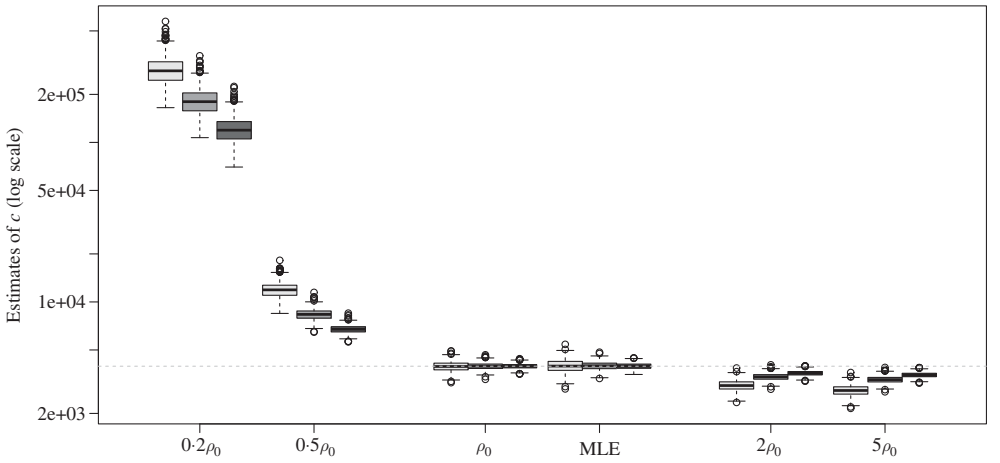

Fig. 2. Sampling distributions for $\hat{c}_n$ when $\nu = 1{\cdot}5$ and the effective range is 0·3. The range parameter is either fixed at the true value (ρ_0), estimated via maximum likelihood, or fixed at a multiple of the truth ($0{\cdot}2\rho_0, \ldots, 5\rho_0$). The four boxplots in each group correspond to sample sizes of $n = 400$, 900, and 1600, reading from left to right. The dashed line indicates the true c.

Table 1. *Empirical coverage rates of nominal* 95% *confidence intervals for* $c = \sigma^2/\rho^{2\nu}$*, expressed as percentages. Intervals are constructed using either the maximum likelihood estimator or estimates of* c *that fix* $\hat{\rho}_n$ *at a multiple of the true value of* ρ*, rounded to the nearest* 1%

		$\nu = 0{\cdot}5$			$\nu = 1{\cdot}5$		
ER		0·1	0·3	1	0·1	0·3	1
$\hat{\rho}_n$	n						
MLE	400	81	92	94	64	87	94
	900	89	94	94	74	91	94
	1600	90	94	94	81	92	95
0·2 ρ	400	0	0	0	0	0	0
	900	0	0	1	0	0	0
	1600	0	0	2	0	0	0
0·5 ρ	400	0	4	88	0	0	4
	900	0	7	90	0	0	9
	1600	0	13	92	0	0	18
2 ρ	400	3	75	93	0	1	83
	900	3	82	93	0	9	89
	1600	5	84	94	0	17	93
5 ρ	400	0	63	92	0	0	77
	900	0	75	93	0	2	86
	1600	0	79	93	0	5	90

MLE, maximum likelihood estimator; ER, empirical coverage rate.

3·3. *Prediction*

The mean squared error of predictor $\hat{Z}_n(\rho_1)$ may be calculated in closed form using (6). When the plug-in predictor $\hat{Z}_n(\hat{\rho}_n)$ is used, we need to integrate over the sampling distribution for $\hat{\rho}_n$, which we approximate by averaging over the simulation results from § 3·2. For both fixed and estimated ρ, we calculate the average mean squared prediction error, averaging over the $m = 2500$ prediction points. Because prediction varies in difficulty according to ν, n, and effective range,

Table 2. *Percent increase in mean squared prediction error relative to the optimal mean squared prediction error using the true value of ρ, rounded to the nearest* 0·1%

		$\nu = 0{\cdot}5$			$\nu = 1{\cdot}5$		
ER		0·1	0·3	1	0·1	0·3	1
$\hat{\rho}_n$	n						
MLE	400	0·2	0·1	0·0	0·2	0·1	0·1
	900	0·1	0·0	0·0	0·1	0·0	0·0
	1600	0·0	0·0	0·0	0·0	0·0	0·0
0·2 ρ	400	36·6	60·4	6·5	103·1	487·0	165·5
	900	56·4	37·5	2·3	218·2	474·1	83·8
	1600	66·2	19·2	0·9	351·4	321·5	41·5
0·5 ρ	400	8·7	2·8	0·2	26·9	20·0	2·9
	900	7·9	1·1	0·1	32·9	10·2	1·3
	1600	5·5	0·4	0·0	29·2	4·7	0·7
2 ρ	400	2·8	0·3	0·0	12·0	2·1	0·3
	900	1·3	0·1	0·0	6·8	1·0	0·1
	1600	0·6	0·0	0·0	3·4	0·4	0·1
5 ρ	400	5·6	0·6	0·1	27·2	4·2	0·6
	900	2·4	0·2	0·0	13·7	1·9	0·2
	1600	1·1	0·1	0·0	6·6	0·9	0·1

we report the percent increase in mean squared prediction error relative to the optimal mean squared prediction error using the true value of ρ, which is calculated from (7).

Table 2 shows that plug-in prediction using the maximum likelihood estimator $\hat{\rho}_n$ performs quite well relative to predicting with the true value of ρ. For $n = 900$ and 1600, the increase in mean squared error is less than 0·1 percent in all cases. It is also clear that there are cases in which it makes little difference if ρ is fixed at an incorrect value. This is true when the effective range is large and ρ_1 is fixed at something larger than the true value. However, there are also cases in which fixing ρ can lead to quite a large loss of efficiency. These effects are magnified when we move from $\nu = 0{\cdot}5$ to $\nu = 1{\cdot}5$, suggesting that a misspecified value of ρ is more problematic for smoother processes. This agrees with some earlier cases in the literature in which predictions with a fixed ρ were still quite accurate. For example, Zhang & Wang (2010) examined precipitation data using a predictive process model (Banerjee et al., 2008) and concluded that a variety of prediction metrics did not change when ρ was fixed at values larger than the maximum likelihood estimator. However, the underlying covariance model for the predictive process was Matérn with $\nu = 0{\cdot}5$, corresponding to a process that is not mean square differentiable.

In a similar pattern to what we observe for mean squared error in Table 2, using the maximum likelihood estimator produces intervals with nominal coverage in nearly all cases, and the estimators fixing ρ at something larger than the true value achieve this rate for $n = 900$ and 1600 when the effective range is large. However, the intervals tend to be too conservative when the effective range is large and ρ_1 is too small, and they tend to be not conservative enough when the effective range is small and ρ_1 is too big. See the Supplementary Material for full results.

4. DISCUSSION

We have made a number of simplifying assumptions. Considering the ways in which they may be relaxed provides a rich set of questions for future research. For example, our results concern only mean zero Gaussian processes, which is equivalent to assuming that the mean of the process

is known. Results on equivalence of mean zero Gaussian measures such as Theorem 2 of Zhang (2004) can be used in proving equivalence of Gaussian process measures with different means (Stein, 1999, Ch. 4, Corollary 5). However, the primary difficulty is in extending estimation results. Zhang (2004) indicates that his method of proof is not easily extended to the case of an unknown mean. Asymptotic results for the case $\nu = 1/2$ and $d = 1$ are given in Theorem 3 of Ying (1991), and it seems plausible that similar results might hold for $d = 2$ and 3. With an unknown mean, it might be preferable to use restricted maximum likelihood (Stein, 1999), for which improved infill asymptotic results should also be sought.

We have also not considered what happens when the observations are not of the process Z itself, but of Z observed with measurement error. Again, results for equivalence and prediction can be extended in a relatively straightforward way. We expect something like Theorem 2 should hold for the case that Z is observed with measurement error. However, in a restricted version of this problem, the introduction of the error term reduces the rate of convergence of the maximum likelihood estimator for c from the usual order $n^{-1/2}$ to order $n^{-1/4}$ (Chen et al., 2000).

Perhaps the most important restriction, both here and in previous work, is that ν is assumed to be known. Estimating ν provides desirable flexibility, as this parameter controls the mean square differentiability of the process. However, we know of no results concerning the maximum likelihood estimator in this case. Stein (1999, § 6.7) examines a periodic version of the Matérn model and argues that $\hat{\sigma}_n^2$ and $\hat{\nu}_n$ should have a joint asymptotic normal distribution, but it is an open question whether a similar result holds for nonperiodic fields.

Acknowledgement

This work was supported by the Center for Science of Information, an NSF Science and Technology Center, and the Statistical and Applied Mathematical Sciences Institute. The authors thank the editor, associate editor, and two reviewers for their useful suggestions.

Supplementary material

Supplementary material available at *Biometrika* online includes asymptotic results for maximum tapered likelihood estimation, as well as additional simulation results.

References

Abramowitz, M. & Stegun, I. A., Ed. (1992). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications Inc. Reprint of the 1972 edition.

Anderes, E., Huser, R., Nychka, D. & Coram, M. (2012). Nonstationary positive definite tapering on the plane. *J. Comp. Graph. Statist.* doi: 10.1080/10618600.2012.729982.

Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B **70**, 825–48.

Chen, H.-S., Simpson, D. G. & Ying, Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statist. Sinica* **10**, 141–56.

Gneiting, T., Kleiber, W. & Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Am. Statist. Assoc.* **105**, 1167–77.

Horn, R. A. & Johnson, C. R. (1990). *Matrix Analysis.* Cambridge: Cambridge University Press. Corrected reprint of the 1985 original.

Kaufman, C. G., Schervish, M. J. & Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Am. Statist. Assoc.* **103**, 1545–55.

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

Putter, H. & Young, G. A. (2001). On the effect of covariance function estimation on the accuracy of kriging predictors. *Bernoulli* **7**, 421–38.

R DEVELOPMENT CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

SAHU, S. K., GELFAND, A. E. & HOLLAND, D. M. (2007). High-resolution space-time ozone modeling for assessing trends. *J. Am. Statist. Assoc.* **102**, 1221–34.

STEIN, M. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *Ann. Statist.* **18**, 850–72.

STEIN, M. (2010). Asymptotics for spatial processes. In *Handbook of Spatial Statistics*, Ed. A. E. Gelfand, P. J. Diggle, P. Guttorp & M. Fuentes, 79–88. Boca Raton: CRC 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. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statist. Probab. Lett.* **17**, 399–404.

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

WANG, D. & LOH, W.-L. (2011). On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electron. J. Statist.* **5**, 238–69.

YING, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *J. Mult. Anal.* **36**, 280–96.

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Statist. Assoc.* **99**, 250–61.

ZHANG, H. & WANG, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics* **21**, 290–304.

ZHANG, H. & ZIMMERMAN, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92**, 921–36.

[*Received December* 2011. *Revised October* 2012]