Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets

Cari G. KAUFMAN, Mark J. SCHERVISH, and Douglas W. NYCHKA

Maximum likelihood is an attractive method of estimating covariance parameters in spatial models based on Gaussian processes. But calculating the likelihood can be computationally infeasible for large data sets, requiring $O(n^3)$ calculations for a data set with n observations. This article proposes the method of covariance tapering to approximate the likelihood in this setting. In this approach, covariance matrixes are "tapered," or multiplied element wise by a sparse correlation matrix. The resulting matrixes can then be manipulated using efficient sparse matrix algorithms. We propose two approximations to the Gaussian likelihood using tapering. One of these approximations simply replaces the model covariance with a tapered version, whereas the other is motivated by the theory of unbiased estimating equations. Focusing on the particular case of the Matérn class of covariance functions, we give conditions under which estimators maximizing the tapering approximations are, like the maximum likelihood estimator, strongly consistent. Moreover, we show in a simulation study that the tapering estimators can have sampling densities quite similar to that of the maximum likelihood estimator, even when the degree of tapering is severe. We illustrate the accuracy and computational gains of the tapering methods in an analysis of yearly total precipitation anomalies at weather stations in the United States.

KEY WORDS: Compactly supported correlation function; Covariance estimation; Estimating equation; Gaussian process.

1. INTRODUCTION

This article addresses the problem of estimating the covariance function of a spatially correlated Gaussian process when the set of observations is large and irregularly spaced. Maximum likelihood estimation has been used for some time by the geostatistical community (Kitanidis 1983; Mardia and Marshall 1984); however, evaluating the likelihood requires order n^3 operations for a data set of size n, making these methods computationally intractable for large n. We introduce two approximations to the likelihood using the method of covariance tapering. These approximations significantly reduce the computation of the likelihood for moderate sample sizes, and they make possible otherwise infeasible calculations for large sample sizes. (The definitions of "moderate" and "large" are system-dependent, but, for example, a "large" data set on a desktop computer with 2 GB of RAM would be about 10,000 data points.) In addition to their computational benefits, the estimators maximizing our approximations share some desirable properties with the maximum likelihood estimator (MLE). We give conditions under which they are, like the MLE, strongly consistent, and we demonstrate through simulation that their sampling distributions can be quite similar to those of the MLE, even when the approximation is severe.

We consider the model in which the data are drawn from an underlying Gaussian process, $Z = \{Z(\mathbf{s}), \mathbf{s} \in S \subset \Re^d\}$. To streamline our development, we assume that the mean of the process is 0 and that the covariance function is stationary and isotropic. Write $K(x; \boldsymbol{\theta})$ to represent the covariance between any two observations whose locations are x units distant from one another. K is assumed known up to the parameter vector $\boldsymbol{\theta} \in \Re^p$, which must be estimated based on a finite number of observations, $\mathbf{Z} = (\mathbf{Z}(\mathbf{s}_1), \ldots, \mathbf{Z}(\mathbf{s}_n))'$. One commonly used isotropic covariance function is the exponential function $K(x; \sigma^2, \rho) = \sigma^2 \exp\{-x/\rho\}$.

Cari G. Kaufman is Assistant Professor, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: *cgk@stat.berkeley.edu*). Mark J. Schervish is Professor and Head, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15214 (E-mail: *mark@stat.cmu.edu*). Douglas W. Nychka is Director, Institute for Mathematics Applied to Geosciences and Senior Scientist, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO 80305 (E-mail: *nychka@ucar.edu*). This work was supported in part by National Science Foundation grant DMS-0355474. The authors thank the referees for their helpful comments.

The vector $\mathbf{Z}$ has a multivariate normal distribution, with log-likelihood function

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{Z}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z} \tag{1}$$

(ignoring a constant), where $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = K(\|\mathbf{s}_i - \mathbf{s}_j\|; \boldsymbol{\theta}), i, j = 1, \ldots, n$.

The advantage of using a Gaussian process model rather than simply specifying the finite set of observations to be multivariate normal and estimating the covariance matrix is that the Gaussian process model implies a joint distribution for the observations $\mathbf{Z}$ and the process at any unobserved location $\mathbf{s}^*$. Deriving predictions for $Z(\mathbf{s}^*)$ according to its conditional expectation given $\mathbf{Z}$ is a canonical problem in geostatistics, called kriging (see, e.g., Stein 1999). This computation is also expensive for large n. Furrer et al. (2006) suggested using covariance tapering to ease the computational burden of kriging large data sets. But these authors assumed that the covariance parameters were known, whereas we focus on their estimation.

The computational difficulty of finding the MLE was recognized by some of its earliest advocates (Mardia and Marshall 1984; Vecchia 1988). Efficient computational techniques have been developed mainly for data sets in which the sampling locations form a regular lattice. In this case the covariance matrix has a special structure that can be exploited computationally (Whittle 1954; Zimmerman 1989). Fewer techniques are available for irregularly spaced data. Fuentes (2007) developed an approximation to the likelihood based on integrating a spatial process over grid cells, so as to obtain a lattice structure that can be modeled in the spectral domain. Vecchia (1988) proposed a likelihood approximation in the spatial domain, which was later extended by Stein et al. (2004). Here $\mathbf{Z}$ is partitioned into subvectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_b$, and then the likelihood is written as a product of conditional densities $p(\mathbf{Z}_j|\mathbf{Z}_{(j-1)}; \boldsymbol{\theta})$, where $\mathbf{Z}'_{(j)} = (\mathbf{Z}'_1, \ldots, \mathbf{Z}'_j)$. Next, the full conditioning sets $\mathbf{Z}_{(j-1)}$ are replaced with smaller subsets, $\tilde{\mathbf{Z}}_{(j-1)} \subseteq \mathbf{Z}_{(j-1)}$, making the densities in the product easier to evaluate. Vecchia (1988) chose $\tilde{\mathbf{Z}}_{(j)}$ to consist of nearest neighbors within $\mathbf{Z}_{(j)}$. Stein et al. (2004) extended Vecchia's idea

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Theory and Methods
DOI 10.1198/016214508000000959

to RMLEs and examined more flexible choices of conditioning sets. The intuition behind this approach is that correlations between pairs of distant locations often are nearly zero, so little information is lost in taking them to be conditionally independent given intermediate locations. A similar idea motivates the covariance tapering approach that we explore in this article.

2. LIKELIHOOD APPROXIMATION USING TAPERING

If we have reason to believe that distant pairs of observations are independent, then we can model this structure using a compactly supported covariance function (Gneiting 2002). Then $\mathbf{\Sigma}(\boldsymbol{\theta})$ contains 0's corresponding to these distant pairs, and sparse matrix algorithms (see, e.g., Pissanetzky 1984) can be used to evaluate the likelihood efficiently. Even if we do not believe that the underlying process has such a covariance function, we can exploit this idea for computational purposes. The goal is to set certain elements of the covariance matrix to 0, such that the resulting matrix is positive definite and retains the original properties of $\mathbf{\Sigma}(\boldsymbol{\theta})$ for proximate locations. Toward this end, consider taking the product of the original covariance function $K_0(x;\boldsymbol{\theta})$ and a tapering function $K_{taper}(x;\gamma)$, an isotropic correlation function that is identically 0 whenever $x \geq \gamma$. Denote this tapered covariance function by

$$K_1(x;\boldsymbol{\theta},\gamma) = K_0(x;\boldsymbol{\theta})K_{taper}(x;\gamma), \qquad x > 0. \tag{2}$$

The tapered covariance matrix is denoted by $\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)$, where $\mathbf{T}(\gamma)_{ij} = K_{taper}(\|\mathbf{s}_i - \mathbf{s}_j\|;\gamma)$. The "$\circ$" notation refers to the element wise matrix product, also called the Schur or Hadamard product. Some relevant properties of the Schur product are listed in Appendix A. Notably, the Schur product of two covariance matrixes is again a valid covariance matrix. In addition, requiring K_{taper} to be a correlation function ensures that the marginal variance of the process Z is the same under K_0 and K_1. It is important to note the equivalence of tapering the covariance matrix and tapering the covariance function; $[\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]_{ij} = K_1(\|\mathbf{s}_i - \mathbf{s}_j\|;\boldsymbol{\theta},\gamma)$.

Bickel and Levina (2008) developed estimators of a covariance matrix by banding the sample covariance matrix and noted that although banding does not guarantee positive definiteness, tapering does. Furrer and Bengtsson (2007) also used tapering as a regularization technique for the ensemble Kalman filter. But these two works differ from the present context in that they are concerned with estimating the covariance matrix rather than the parameters of a particular covariance function. In contrast, we taper for purely computational reasons; the MLE for $\boldsymbol{\theta}$ is still optimal, in a sense that we describe in Section 4, but often is computationally infeasible to calculate.

We propose two approximations to the log-likelihood (1) using covariance tapering. The first approximation simply replaces the model covariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$ with $\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)$, giving

$$l_{1taper}(\boldsymbol{\theta}) = -\frac{1}{2}\log|\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| - \frac{1}{2}\mathbf{Z}'[\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1}\mathbf{Z}. \tag{3}$$

This is equivalent to using a model in which the process Z is Gaussian with mean 0 and covariance function (2). The effects of misspecifying the covariance function have been widely studied with respect to kriging (see, Stein 1999, sec. 4.3), but the implications for estimation have not been as well studied.

One possible objection to this approximation is that the corresponding "score" function is biased, that is, $E[\frac{\partial}{\partial\boldsymbol{\theta}} l_{1taper}(\boldsymbol{\theta})] \neq \mathbf{0}$. This means that there is no guarantee that the estimator maximizing (3) is asymptotically unbiased. Moreover, in our experience this estimator sometimes can display sizeable bias in practice, especially if the taper range γ is small relative to the true correlation range of the process.

To remedy the bias, we can take an estimating equations approach to formulating a tapered version of this problem. Essentially, we taper both the model covariance matrix and the sample covariance matrix. First, note that we can rewrite the quadratic form in (1) as a trace involving the sample covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{Z}\mathbf{Z}'$,

$$\begin{aligned}\mathbf{Z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z} &= \mathrm{tr}\{\mathbf{Z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}\} \\ &= \mathrm{tr}\{\mathbf{Z}\mathbf{Z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\} \\ &= \mathrm{tr}\{\hat{\mathbf{\Sigma}}\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\}. \end{aligned} \tag{4}$$

Replacing both the model and sample covariance matrixes with tapered versions gives

$$\begin{aligned} l_{2tapers}(\boldsymbol{\theta}) &= -\frac{1}{2}\log|\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| \\ &\quad -\frac{1}{2}\mathrm{tr}\big\{[\hat{\mathbf{\Sigma}} \circ \mathbf{T}(\gamma)][\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1}\big\} \\ &= -\frac{1}{2}\log|\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| \\ &\quad -\frac{1}{2}\mathbf{Z}'\big([\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1} \circ \mathbf{T}(\gamma)\big)\mathbf{Z}. \end{aligned} \tag{5}$$

The second form of the expression follows from the trace equality in Appendix A and a reversal of the reasoning in (4). Maximizing $l_{2tapers}(\boldsymbol{\theta})$ now corresponds to solving an unbiased estimating equation for $\boldsymbol{\theta}$, that is, $\mathrm{E}[\frac{\partial}{\partial\boldsymbol{\theta}} l_{2tapers}(\boldsymbol{\theta})] = \mathbf{0}$.

In both approximations, small values of γ correspond to more severe tapering. When $\gamma = 0$, observations are treated as independent, and not all parameters are necessarily estimable, whereas as $\gamma \to \infty$, we recover the full likelihood. But γ can be chosen to be quite small in $l_{2tapers}$ and still give efficient estimators, as we demonstrate in the simulation study presented in Section 5.

We refer to (3) as the one-taper approximation, to (5) as the two-taper approximation, and call the estimators maximizing them the one- and two-taper estimators, or $\hat{\boldsymbol{\theta}}_{1taper}$ and $\hat{\boldsymbol{\theta}}_{2tapers}$. The choice of approximation is context-dependent. The one-taper approximation is computationally more efficient, because it involves solving the sparse system of equations $[\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1}\mathbf{Z}$, whereas the two-taper approximation requires the full inverse of a sparse matrix to compute $([\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1} \circ \mathbf{T}(\gamma))\mathbf{Z}$. In addition, asymptotic results are more straightforward for the one-taper approximation, as we discuss in the next section. But maximizing the two-taper approximation has the advantages of solving an unbiased estimating equation; the bias tends to be smaller in practice, and we describe how sampling variability can be estimated using the robust information criterion (Heyde 1997). Thus we prefer the two-taper approximation in practice, unless the range of the

process is clearly small enough to produce little bias in the one-taper approximation. We return to this question in Section 6, where we illustrate the tapering methods on a large spatial data set.

3. ASYMPTOTIC RESULTS

Two commonly used asymptotic frameworks in spatial statistics are "increasing domain" and "fixed domain" asymptotics (see, e.g., Cressie 1993, sec. 5.8). Under increasing domain asymptotics, the sampling region increases without bound, and the minimum distance between sampled locations is bounded below by a positive constant. Under fixed domain asymptotics, the sampling region is fixed and bounded, and sampling locations become increasingly dense within this region. We focus primarily on fixed domain asymptotics, under which Zhang (2004) recently showed almost sure convergence of the MLE under the Matérn covariance model.

The Matérn covariance function (Matérn 1986) is widely used in practice and has easily interpretable parameters. This function is defined by

$$K(x;\sigma^2,\rho,\nu)=\frac{\sigma^2(x/\rho)^\nu}{\Gamma(\nu)2^{\nu-1}}\mathcal{K}_\nu(x/\rho),\qquad x\geq 0,\sigma^2,\rho,\nu>0, \tag{6}$$

where $\mathcal{K}_\nu$ is the modified Bessel function of order ν (see Abromowitz and Stegun 1967, sec. 9.6). The parameter σ^2 is the marginal variance of the process, ρ controls how quickly the correlation decays with distance, and ν controls the smoothness of the process (see Stein 1999, sec. 2.7 for more details).

Zhang (2004) proved several important results about the Matérn class. The first of these results concerns the equivalence of two mean-0 Gaussian measures, $G(K_0)$ and $G(K_1)$. Throughout, let $G(K)$ denote the mean-0 Gaussian measure,with covariance function K. Recall that two probability measures, P_0 and P_1, on the same measurable space $(\Omega,\mathcal{F})$ are called equivalent if $P_0(A)=0$ if and only if $P_1(A)=0$, for all $A\in\mathcal{F}$. Denote this by $P_0\equiv P_1$. If the true covariance K_0 is Matérn with parameters σ_0^2, ρ_0, and ν, and if K_1 is Matérn with parameters σ_1^2, ρ_1, and ν, then $G(K_0)\equiv G(K_1)$ on the paths of $\{Z(s), s\in T\}$ for any bounded infinite subset $T\in\Re^d$ with $d\leq 3$, if and only if $\sigma_0^2/\rho_0^{2\nu}=\sigma_1^2/\rho_1^{2\nu}$ (Zhang 2004).

This result has immediate consequences for estimation. Under the fixed domain asymptotics with $d\leq 3$, consistent estimators of both σ^2 and ρ cannot exist. But the ratio $c=\sigma^2/\rho^{2\nu}$ is consistently estimable. In particular, for known ν and for any fixed ρ^*, the estimator $\hat{\sigma}_n^2$ obtained by maximizing the likelihood $L_n(\sigma^2,\rho^*)$ is such that $\hat{\sigma}_n^2/\rho^{*2\nu}\to\sigma_0^2/\rho_0^{2\nu}$ almost surely under $G(K_0)$ (Zhang 2004).

3.1 Equivalent Measures Under Tapering

Zhang (2004) used the equivalence of Gaussian measures with different Matérn covariance functions to prove almost sure convergence of his estimator of c. This technique translates a difficult problem under one measure into an easy problem under a different, but equivalent measure. The same principle can be used to develop a consistent estimator of c that maximizes the one-taper approximation (3). The following theorem gives some conditions on the tapering function K_{taper}, under which the tapered and untapered Matérn covariance functions give equivalent mean-0 Gaussian measures. Throughout, Z represents a stochastic process on $\Re^d$.

Theorem 1. Let K_0 be the Matérn covariance function on $\Re^d$, $d\leq 3$, with parameters σ^2, ρ, and ν, and let $K_1=K_0K_{taper}$, where K_{taper} is an isotropic correlation function on $\Re^d$. Suppose that the spectral density

$$f_{taper}(\boldsymbol{\omega})=(2\pi)^{-d}\int_{\Re^d}\exp\{-i\boldsymbol{\omega}'\mathbf{x}\}K_{taper}(\mathbf{x})\,d\mathbf{x}$$

exists and that there exist $\epsilon>0$ and $M_\epsilon<\infty$ such that $f_{taper}(\boldsymbol{\omega})\leq M_\epsilon/(1+\|\boldsymbol{\omega}\|^2)^{\nu+d/2+\epsilon}$, with $\epsilon>\max\{d/4,1-\nu\}$. Then $G(K_0)\equiv G(K_1)$ on the paths of $\{Z(\mathbf{s}),\mathbf{s}\in T\}$, for any bounded subset $T\subset\Re^d$.

Proofs of all results are given in Appendix B.

We can choose a function to satisfy the conditions of Theorem 1 from the family of compactly supported functions constructed by Wendland (1995, 1998) and suggested as tapering functions by Furrer et al. (2006). The Wendland function $\phi_{d,k}(\|\mathbf{x}\|)$ is positive definite on $\Re^d$. For $\|\mathbf{x}\|\in[0,1]$, it is a polynomial of degree $\lfloor d/2\rfloor+3k+1$. For $\|\mathbf{x}\|>1$, $\phi_{d,k}(\|\mathbf{x}\|)=0$. Wendland (1998) showed that if $f_{d,k}$ is the spectral density corresponding to $\phi_{d,k}$, then there exists a positive constant M such that $f_{d,k}(\|\boldsymbol{\omega}\|)\leq M/(1+\|\boldsymbol{\omega}\|^2)^{d/2+k+1/2}$. Therefore, $\phi_{d,k}(\|\mathbf{x}\|/\gamma)$ satisfies the conditions of Theorem 1 for all $\nu\leq\nu'$ whenever $k>\max\{1/2,\nu'+(d-2)/4\}$. For example, when $\nu=1/2$ (the exponential covariance), the Wendland function $\phi_{d,1}(\|\mathbf{x}\|/\gamma)$ is a valid taper for $d\leq 3$.

3.2 Convergence of the Tapering Estimators

We now can use Theorem 1 to prove almost sure convergence of the one-taper estimator, a direct analog of the result for the MLE given by Zhang (2004).

Theorem 2. Let K_0 be the Matérn covariance function on $\Re^d$, $d\leq 3$ with known parameter ν and unknown parameters σ^2 and ρ. Let $\{S_n\}_{n=1}^\infty$ be an increasing sequence of finite subsets of $\Re^d$ such that $\bigcup_{n=1}^\infty S_n$ is bounded and infinite. Let $l_{n,1taper}$ be the one-taper approximation (3) based on observations of Z at locations in S_n, with K_{taper} satisfying the conditions of Theorem 1. Fix $\rho^*>0$, and let $\hat{\sigma}^2_{n,1taper}$ maximize $l_{n,1taper}(\sigma^2,\rho^*)$. Then $\hat{\sigma}^2_{n,1taper}/\rho^{*2\nu}\to\sigma^2/\rho^{2\nu}$ almost surely under $G(K_0)$ as $n\to\infty$.

Note that the specification of this theorem takes the taper function to be constant with n. This allows the number of pairs of observations within the taper range to go to infinity. If we view the role of an asymptotic result as providing some intuition about estimators when n is "large" in some sense, then here we are defining "largeness" relative to the taper range, rather than allowing the taper range to shrink with n and defining "largeness" in an absolute sense.

On a related note, we should point out that theory for the one-taper approximation is also available under the increasing domain asymptotics. The results of Watkins and Al-Boutiahi (1990) can be applied to show that the limiting distribution of $\hat{\boldsymbol{\theta}}_{1taper}$ is asymptotically normal. The mean of this distribution, call it $\tilde{\boldsymbol{\theta}}_{1taper}$, is the limit of a sequence minimizing

the Kullback–Liebler distance between the (finite-dimensional) distributions with tapered and untapered covariance matrixes. The limiting variance also may be calculated for a given $\tilde{\boldsymbol{\theta}}_{1taper}$ and true value $\boldsymbol{\theta}_0$. We do not pursue this theory any further here.

Turning now to a discussion of the two-taper estimator, recall that unlike l_{1taper}, $l_{2tapers}$ does not correspond to an altered distribution for the process Z. Therefore, the equivalence result in Theorem 1 is not applicable. Instead, the next theorem considers the case in which the covariance function $K_0(x;\sigma^2) = \sigma^2 C_0(x)$, where $C_0(x)$ is a known correlation function. In this case it is possible to solve for $\hat{\sigma}^2_{n,2tapers}$ explicitly and use this expression to determine the conditions necessary for convergence. But it also is possible to use this result to prove convergence in the case where *both* σ^2 and ρ of the Matérn covariance function are unknown. This result is given as a corollary directly after Theorem 3.

Theorem 3. Let $K_0(x;\sigma^2) = \sigma^2 C_0(x)$, where $C_0(x)$ is a known correlation function on $\Re^d$ and σ^2 is unknown. Let $\{S_n\}_{n=1}^{\infty}$ be a sequence of finite subsets of $\Re^d$. Let $l_{n,2tapers}(\sigma^2)$ be the two-taper approximation (5) based on observations of Z at locations in S_n. For all n, define the matrix $\mathbf{W}_n = [(\boldsymbol{\Gamma}_n \circ \mathbf{T}_n)^{-1} \circ \mathbf{T}_n]\boldsymbol{\Gamma}_n$, where $(\boldsymbol{\Gamma}_n)_{ij} = C_0(\|\mathbf{s}_i - \mathbf{s}_j\|)$ and $(\mathbf{T}_n)_{ij} = K_{taper}(\|\mathbf{s}_i - \mathbf{s}_j\|;\gamma)$, $i, j = 1,\ldots,n$. Denote by $\{\lambda_{n,i}\}_{i=1}^n$ the eigenvalues of $\mathbf{W}_n$. Suppose that either $\sup_n(n^{-1} \times \sum_{i=1}^n \lambda_{n,i}^q)^{1/q} < \infty$ for some $1 < q \le \infty$ or $\lim_n(\sup_{i\le n} \lambda_{n,i}) \times n^{-1}\log n = 0$. Then $\hat{\sigma}^2_{n,2tapers} \to \sigma^2$ almost surely under $G(K_0)$ as $n \to \infty$.

Corollary 1. Let K_0 be the Matérn covariance function on $\Re^d$, $d \le 3$ with known parameter ν and unknown parameters σ^2 and ρ. Fix $\rho^* > 0$, and let $\hat{\sigma}^2_{n,2taper}$ maximize $l_{n,2tapers}(\sigma^2, \rho^*)$. Define $\mathbf{W}_n$ as in Theorem 3, but with $(\boldsymbol{\Gamma}_n)_{ij} = \sigma^{-2}K_0(\|\mathbf{s}_i - \mathbf{s}_j\|;\sigma^2,\rho^*,\nu)$. Suppose that the eigenvalues of $\mathbf{W}_n$ satisfy one of the conditions in Theorem 3. Then $\hat{\sigma}^2_{n,2tapers}/\rho^{*2\nu} \to \sigma^2/\rho^{2\nu}$ almost surely under $G(K_0)$ as $n \to \infty$.

3.3 Example

The conditions in Theorem 3 depend on the correlation function, the tapering function, and the sampling locations. These conditions can be difficult to check in practice. The following lemma allows us to ignore the choice of tapering function, using a bound that depends only on the correlation function of the process.

Lemma 1. Let $\boldsymbol{\Gamma}$ and $\mathbf{T}$ be correlation matrixes, and let $\mathbf{W} = [(\boldsymbol{\Gamma} \circ \mathbf{T})^{-1} \circ \mathbf{T}]\boldsymbol{\Gamma}$. Then $\lambda_{max}\{\mathbf{W}\} \le \lambda_{max}\{\boldsymbol{\Gamma}\}/\lambda_{min}\{\boldsymbol{\Gamma}\}$, where λ_{min} and λ_{max} refer to minimum and maximum eigenvalues.

We illustrate the use of this lemma in a simple example. Suppose that the process lies in $\Re$, with correlation function $C_0(x) = \exp\{-|x|/\rho\}$, and that $\rho > 0$ is known. Suppose that the sampling locations are spaced equally, with $s_i = i\Delta$, $i = 1,\ldots,n$. Then $\boldsymbol{\Gamma}_n$ has symmetric Toeplitz form. Define $f(\lambda) = 1 + 2\sum_{k=1}^{\infty} e^{-k\Delta/\rho}\cos(k\lambda) = \sinh(\Delta/\rho)/[\cosh(\Delta/\rho) - \cos(\lambda)]$ for $\lambda \in [0, 2\pi]$. Then, for all n, the eigenvalues $\tau_{n,k}$ of $\boldsymbol{\Gamma}_n$ satisfy $ess\inf f \le \tau_{n,k} \le ess\sup f$ (applying Lemma 4.1 of Gray 2006). Therefore, by Lemma 1, $\lambda_{max}\{\mathbf{W}_n\} \le \lambda_{max}\{\boldsymbol{\Gamma}_n\}/\lambda_{min}\{\boldsymbol{\Gamma}_n\} \le ess\sup f/ess\inf f$. Because $f(\lambda)$ has a maximum at 0 of $\coth(\Delta/(2\rho))$ and a minimum at π of $\tanh(\Delta/(2\rho))$,

$$\lambda_{max}\{\mathbf{W}_n\} \le \frac{\lambda_{max}\{\boldsymbol{\Gamma}_n\}}{\lambda_{min}\{\boldsymbol{\Gamma}_n\}} \le \coth^2\left(\frac{\Delta}{2\rho}\right) < \infty \tag{7}$$

whenever $\rho < \infty$ and $\Delta > 0$. Because $\lambda_{max}\{\mathbf{W}_n\}$ is bounded for all n, the second condition of Theorem 3 is satisfied, so $\hat{\sigma}^2_{n,2tapers}$ converges almost surely.

Now consider the case where Δ is not fixed but depends on n. In particular, suppose that $\Delta_n = \Delta/n^k$ for some k. The case where $k = 0$ corresponds to increasing domain sampling. The case where $k = 1$ gives sampling locations $\{0, \Delta/n, \ldots, (n-1)\Delta/n\}$ always contained within $[0, \Delta)$, an instance of fixed domain sampling. For k between 0 and 1, we have a type of sampling intermediate between the usual fixed domain and increasing domain cases. The derivation of the bound in (7) still holds replacing Δ by Δ_n, and the second condition of Theorem 3 is satisfied if

$$\coth^2\left(\frac{\Delta_n}{2\rho}\right)\frac{\log n}{n} = \left(\frac{e^{-\Delta_n/\rho}+1}{e^{-\Delta_n/\rho}-1}\right)^2\frac{\log n}{n}$$

goes to 0 as $n \to \infty$. We have shown that this holds when $k = 0$. Now consider $k \in (0, 1]$. Because $k > 0$, $e^{-\Delta/\rho n^k} \to 1$ as $n \to \infty$. Also, writing $(e^{-\Delta/\rho n^k} - 1)^2 = \frac{\Delta^2}{\rho^2 n^{2k}}(1 + o(1))$,

$$\lim_{n\to\infty}\left(\frac{e^{-\Delta/\rho n^k}+1}{e^{-\Delta/\rho n^k}-1}\right)^2\frac{\log n}{n} = \lim_{n\to\infty}\frac{4\rho^2 n^{2k}}{\Delta^2(1+o(1))}\frac{\log n}{n} = \lim_{n\to\infty}\frac{4\rho^2}{\Delta^2}\frac{\log n}{n^{1-2k}},$$

which is 0 whenever $k < 1/2$. Note that this does not include fixed domain sampling.

4. ESTIMATING SAMPLING VARIABILITY AND CHOOSING THE TAPER RANGE

Recall that maximizing the two-taper approximation (5) corresponds to solving an unbiased estimating equation for $\boldsymbol{\theta}$. This suggests an estimator of the sample variance of $\hat{\boldsymbol{\theta}}_{2tapers}$ based on the robust information criterion of Heyde (1997).

Following Heyde (1997), let $\mathbf{G}(\mathbf{Z};\boldsymbol{\theta})$ be an unbiased estimating function for $\boldsymbol{\theta}$, that is, $\mathrm{E}[\mathbf{G}(\mathbf{Z};\boldsymbol{\theta})] = \mathbf{0}$ for all possible values of $\boldsymbol{\theta}$. The robust information criterion corresponding to $\mathbf{G}$ is $\mathcal{E}(\mathbf{G}) = \mathrm{E}[\dot{\mathbf{G}}]'\mathrm{E}[\mathbf{G}\mathbf{G}']^{-1}\mathrm{E}[\dot{\mathbf{G}}]$, where $\dot{\mathbf{G}}$ is the matrix of derivatives of $\mathbf{G}$ with respect to $\boldsymbol{\theta}$. An estimating function $\mathbf{G}^*$ within a class $\mathcal{H}$ of unbiased estimating functions is said to be O_F-optimal if it maximizes this information criterion, that is, $\mathcal{E}(\mathbf{G}^*) - \mathcal{E}(\mathbf{G})$ is nonnegative definite for all $\mathbf{G} \in \mathcal{H}$. In particular, the score function $\mathbf{U}$ is O_F-optimal whenever it is in $\mathcal{H}$, and is in this case $\mathcal{E}(\mathbf{U})$ simply the Fisher information matrix. In our context, this means that no tapering is always "better" than tapering by this criterion. But we may restrict $\mathcal{H}$ to be the class of *computable* estimating functions for a particular problem or, more specifically, the class of computable estimating functions derived from the tapering approximation (5) for a specific class of taper functions, such as the Wendland functions $\phi_{d,k}(\|\mathbf{x}\|/\gamma)$ for a specific degree k. The functions in this class are then indexed by the taper range γ, and $\mathcal{E}(\mathbf{G})$ can help us choose an appropriate γ.

Under certain conditions, norming by the sample equivalent of $\mathcal{E}(\mathbf{G})^{-1}$ gives asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}_n$ obtained by maximizing $\mathbf{G}(\mathbf{Z}_n; \boldsymbol{\theta})$ (Heyde 1997, sec. 2.5). Although these conditions do not hold in the case of irregularly spaced observations under the fixed domain sampling scheme, the diagonal elements of $\mathcal{E}(\mathbf{G})^{-1}$ still can give reasonable estimates of sampling variability; for example, Stein et al. (2004) suggested this use of the robust information matrix for estimators maximizing their subsetting approximations. Let $\mathbf{G}_{2tapers}$ be the vector whose ith entry is the partial derivative of $l_{2tapers}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_i$. The two matrixes needed to calculate $\mathcal{E}(\mathbf{G}_{2tapers})$ have entries

$$E[\dot{\mathbf{G}}_{2tapers}]_{i,j} = -\frac{1}{2}\mathrm{tr}\left\{\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i} \circ \mathbf{T}\right][\boldsymbol{\Sigma} \circ \mathbf{T}]^{-1}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j} \circ \mathbf{T}\right][\boldsymbol{\Sigma} \circ \mathbf{T}]^{-1}\right\} \quad (8)$$

and

$$E[\mathbf{G}_{2tapers}\mathbf{G}'_{2tapers}]_{ij} = \frac{1}{2}\mathrm{tr}\left\{\left[\left([\boldsymbol{\Sigma} \circ \mathbf{T}]^{-1}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i} \circ \mathbf{T}\right][\boldsymbol{\Sigma} \circ \mathbf{T}]^{-1}\right) \circ \mathbf{T}\right] \times \boldsymbol{\Sigma}\left[\left([\boldsymbol{\Sigma} \circ \mathbf{T}]^{-1}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j} \circ \mathbf{T}\right][\boldsymbol{\Sigma} \circ \mathbf{T}]^{-1}\right) \circ \mathbf{T}\right]\boldsymbol{\Sigma}\right\}. \quad (9)$$

Note that the derivatives in (8) and (9) are matrix quantities depending on the sampling locations, so in general the entries of $\mathcal{E}(\mathbf{G}_{2tapers})^{-1}$ do not have closed-form expressions. But it is computationally straightforward to calculate both (8) and (9), because all inverses involve sparse matrixes.

Now we consider how these expressions may be used in practice to guide the choice of γ. We suggest first obtaining a pilot estimate, $\hat{\boldsymbol{\theta}}_{pilot}$, through a computationally inexpensive method by, for example, calculating the maximum likelihood estimate for a small, randomly chosen subset of the data. Then calculating $\mathcal{E}(\mathbf{G}_{2tapers}(\mathbf{Z}_n; \hat{\boldsymbol{\theta}}_{pilot}, \gamma))^{-1}$ for a sequence of increasing γ values, starting with γ quite small. As γ increases, we expect the variance estimates along the diagonal to decrease; however, we also expect the computation time to increase. By examining plots of both quantities as a function of γ, we can choose γ to give a reasonable trade-off between them. We hope that a small variance can be obtained within the available computing time. We are confident that this often will be the case, even for data sets with high levels of spatial correlation. In particular, the simulation study described in the next section indicates that γ can be small relative to the range of the underlying process and $\hat{\boldsymbol{\theta}}_{2taper}$ can still have a sampling distribution similar to that of the MLE.

Once a specific value of γ is chosen and $\hat{\boldsymbol{\theta}}_{2tapers}$ is obtained, a variance estimate for this quantity can be obtained by taking the diagonal elements of $\mathcal{E}(\mathbf{G}_{2tapers}(\mathbf{Z}_n; \hat{\boldsymbol{\theta}}_{2tapers}, \gamma))^{-1}$. In the next section we show through simulation that this procedure gives accurate variance estimates in practice.

5. SIMULATION STUDY

We used Monte Carlo simulation to investigate three issues concerning the tapering estimators. First, how does their performance compare with that of the MLE? Second, how does the choice of the taper range γ affect the distributions of the tapering estimators? Finally, are the variance estimators proposed in Section 4 good estimators of sampling variability?

We simulated 1,000 data sets, each consisting of a multivariate normal vector of length 300. Note that here n is small enough to easily calculate the MLE for comparison, which will not generally be the case. Each data set was generated using the same 300 locations, consisting of a random selection of perturbed gridpoints. We first generated a two-dimensional grid over $[0, 1]^2$ with increments of .03. To each gridpoint, we added a random amount of noise in each coordinate, uniformly distributed on $[-.01, .01]$. Therefore, each perturbed gridpoint is at least .01 units distant from any of its neighbors. This avoids numerical singularities from the sampling locations being too close together.

The covariance function was exponential with $\sigma^2 = 1$ and $\rho = .2$. With this choice, pairs of observations have negligible ($<.05$) correlation when their locations are more than .6 unit distant from each other. We call this the effective range of the process, and it provides a point of comparison for the choice of taper range.

We maximized the likelihood (1) for each data set to obtain $\hat{\sigma}_n^2$ and $\hat{\rho}_n$, and we formed $\hat{c}_n = \hat{\sigma}_n^2/\hat{\rho}_n$. Likewise, we estimated σ^2, ρ, and c by maximizing l_{1taper} and $l_{2tapers}$ over σ^2 and ρ. Although existing results under the fixed domain asymptotics consider fixing ρ and maximizing only over σ^2, this type of joint maximization is used most commonly in practice. Indeed, the simulation study of Zhang (2004) used joint maximization, even though the asymptotic results concerned a fixed ρ. Kaufman (2006) showed through simulation that fixing ρ far from its true value can significantly bias estimates of c, whereas joint maximization does not have this drawback.

We used the Wendland tapering function with $k = 1$ and set γ to .6, .2, or .1. Thus we are able to compare the tapering estimates when the taper range is equal to the effective range of the process or only a fraction of it. The sparsity of the matrixes, measured by the percentage of off-diagonal elements in $\boldsymbol{\Sigma} \circ \mathbf{T}$ which are 0, is 37% when $\gamma = .6$, 89% when $\gamma = .2$, and 97% when $\gamma = .1$.

Figure 1 shows boxplots of the estimates. As γ decreases, the biases in the one-taper estimates increase. Indeed, for the one-taper estimates, we chose to show the results only for $\gamma = .6$ and $\gamma = .2$, so that we can compare all distributions on the same scale. In contrast, we see negligible bias in the two-taper estimates and a noticeable increase in variance only when $\gamma = .1$. From this, we conclude that whereas the one-taper approximation is appropriate whenever the taper range can be chosen to be at least as large as the effective range of the process (a pilot estimate of which can again be obtained by subsampling the data), the two-taper approximation is clearly more accurate for highly correlated processes.

We have shown that the tapering estimators can be comparable to the MLE when the covariance is exponential. In an extended version of this simulation study, Kaufman (2006)

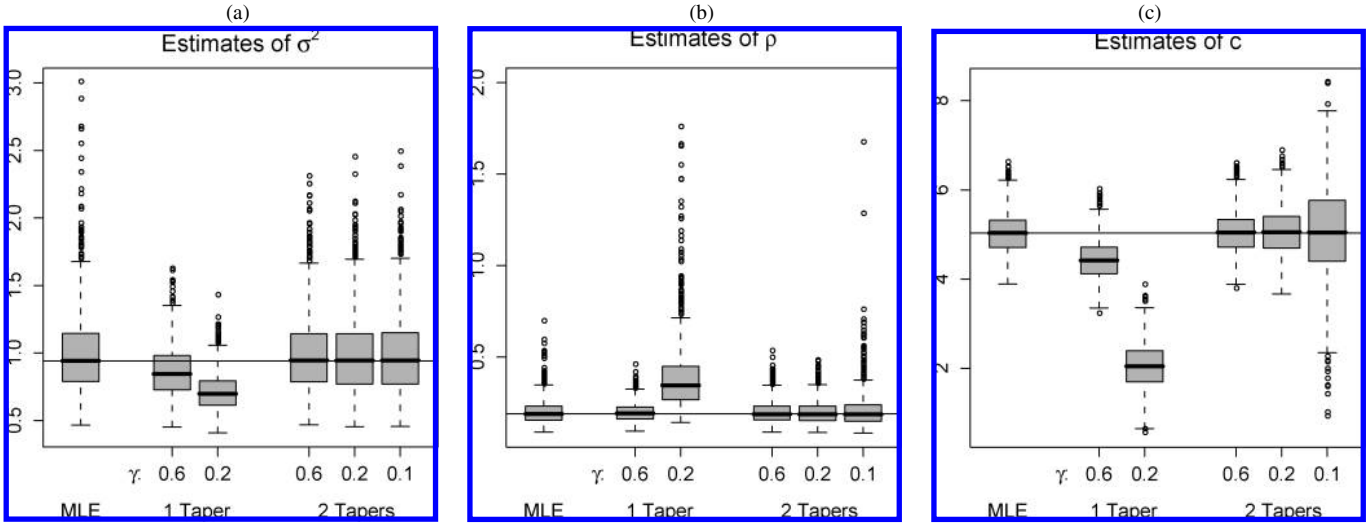

Figure 1. Boxplots of sampled estimates [(a) b^2; (b) ρ; (c) c] in the simulation study. Horizontal lines indicate the median of the distribution for the MLE in each case.

showed that this result holds across a variety of Matérn covariance functions. We also have shown that larger values of the taper range γ produce smaller bias and variance; therefore, it is advisable to choose the largest value of γ for which calculations are computationally feasible. As described in Section 4, variance estimates based on the robust information criterion can help determine whether a particular choice of γ provides an acceptable sampling variance.

We now examine the accuracy of these information-based variance estimators. With 1,000 simulated data sets, the true sampling variances are well approximated by the empirical variances of the estimates in the simulation. For each iteration, we calculated variance estimates for the MLE based on the Fisher information matrix and calculated variance estimates for the two-taper estimator based on the robust information matrix, plugging in the corresponding estimates from that iteration. Table 1 compares the means of these variance estimates with the empirical variances found in the simulation. Results are shown for $\gamma = .2$. For σ^2 and ρ, the estimated variances tend to be higher than the simulated variances, although the two-taper versions are inflated more. For c, the estimated variances are much closer to the simulated variances. This is not surprising, because the variance estimates are based on normal approximations that clearly are less appropriate for the skewed distributions of the estimators of σ^2 and ρ.

6. DATA EXAMPLE

Sizeable computational gains can be achieved by applying the tapering methods to large, irregularly spaced spatial data sets. An example of such a data set is a collection of observations from weather stations in the United States. We consider precipitation data from the National Climatic Data Center for the years 1895 to 1997. This data set was examined in detail by Johns et al. (2003), who focused on imputing missing observations. In this analysis we consider yearly total precipitation anomalies, that is, yearly totals standardized by the long-run mean and standard deviation for each station. We illustrate the tapering methods using the precipitation anomalies from 1962, because this year had one of the most complete data records, with 7,352 stations. In addition, it showed no obvious nonstationarity or anisotropy, which would require a more careful choice of tapering function than is considered here. To calculate the anomalies, we included the full data set computed by Johns et al. (2003); however, our analysis considers the anomalies from only those stations with a complete observational record for this year. The data and computer code needed to carry out the analysis in the example are available at *http://www.image.ucar.edu/Data/precip_tapering/*. For sparse matrix calculations, we used the `spam` package in `R`, available at *http://cran.r-project.org/* under Packages. All calculations were carried out on a 3.2-GHz dual processor compute node with 4 GB of memory.

This analysis illustrates three aspects of working with the tapering approximations in practice, focusing on the more accurate (and computationally demanding) two-taper approximation. First, we address the choice of the taper range γ. We calculate pilot estimates for the covariance parameters using a subset of the data. This allows us to compute the information-based variance estimates discussed in Section 4 for various values

Table 1. Estimated and simulated sample variances for estimators

	$\hat{\sigma}^2$	$\hat{\sigma}^2_{2taper}$	$\hat{\rho}$	$\hat{\rho}_{2taper}$	$\hat{c}$	$\hat{c}_{2taper}$
Mean of estimated variances	.131	.148	.006	.007	.202	.275
Simulated variances	.100	.094	.005	.004	.210	.271
Ratio	1.309	1.571	1.300	1.558	.960	1.014

of γ, as well a variance estimate for the MLE based on the Fisher information. We also examine the time needed to evaluate the two-taper approximation for various values of γ. A comparison of the estimated variance and timing allows us to make a judicious choice of γ for our particular situation. Next, we maximize the tapering approximations for this choice of gamma and compare the results with the MLE. Note that with 7,352 observations, evaluating the full likelihood is quite slow, although still possible. For even larger data sets, it often will not be computable. Finally, we return to the timing question in more detail, breaking down the steps in evaluating the likelihood and the tapering approximations to show where the computational gains are made.

We used a Gaussian process model for this data set with an exponential covariance function. We first maximized the likelihood for a subsample of size 3,000. This gave pilot estimates $\hat{\sigma}^2_{pilot} = .745$, $\hat{\rho}_{pilot} = 52.5$, and $\hat{c}_{pilot} = .745/52.5 = .0142$. We chose the taper function to be the Wendland function $\phi_{2,1}$, as described in Section 3.1. Because the simulation study showed that the estimates of sampling variability for the parameter c are the most accurate, we compared this variance as a function of γ, using the information-based estimates of Section 4. We also computed the variance estimate for the MLE $\hat{c}$, using the Fisher information. Note that all estimates are still based on the subsample of the data.

Figure 2 shows the ratios of the estimated standard errors for $\hat{c}_{2taper}$ and the MLE $\hat{c}$, as well as the time to compute the two-taper approximation for each choice of γ. (For reference, computing the full likelihood took 583 seconds.) Weighing this trade-off, the choice of an appropriate γ is somewhat subjective. Here we choose γ such that the ratio of the standard errors is < 1.5, which gives $\gamma = 70$ miles. The resulting matrixes have only .6% nonzero off-diagonal entries.

For any value of ρ, the maximizing value of σ^2 is available in closed form; therefore, we can minimize the profile versions of the negative log-likelihood and tapering approximations, which are functions only of ρ. These are shown computed over a grid in Figure 3. Vertical lines indicate the minimizing values, which we found using the `optimize` function in R. The taper range is small relative to the correlation range of the process, so it

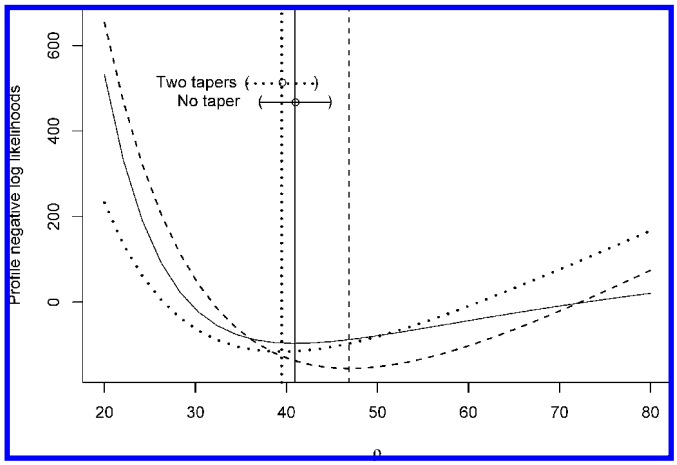

Figure 3. The profile negative log-likelihood (—), one-taper approximation (- - -), and two-taper approximation ($\cdots$). The mean of each function has been subtracted to allow comparison of the curvatures. The corresponding vertical lines indicate the minimum of each function. The MLE and two-taper estimate both have information-based confidence intervals, indicated by the horizontal brackets.

is not surprising that the one-taper estimate is further from the MLE than the two-taper estimate is.

Plugging in the estimates, we computed variance estimates using the Fisher information matrix for the MLE and the robust information matrix for the two-taper approximation, as described in Section 4, and used these to form approximate 95% confidence intervals. For ρ, the MLE was 40.96, with confidence interval (37.15, 44.78), and the two-taper estimate was 39.48, with confidence interval (35.68, 43.27). These also are shown in Figure 3. For σ^2, the MLE was .723, with confidence interval (.663, .783), and the two-taper estimate was .787, with confidence interval (.721, .853). Note that the two sets of confidence intervals overlap. For c, the MLE was .0176, with confidence interval (.0170, .0183), and the two-taper estimate was .0199, with confidence interval (.0191, .0208).

We now compare the computation time required for each method. Evaluating the likelihood requires precomputing the distance matrix, which for this data set took about 40 seconds. In addition, the tapering methods require precomput-

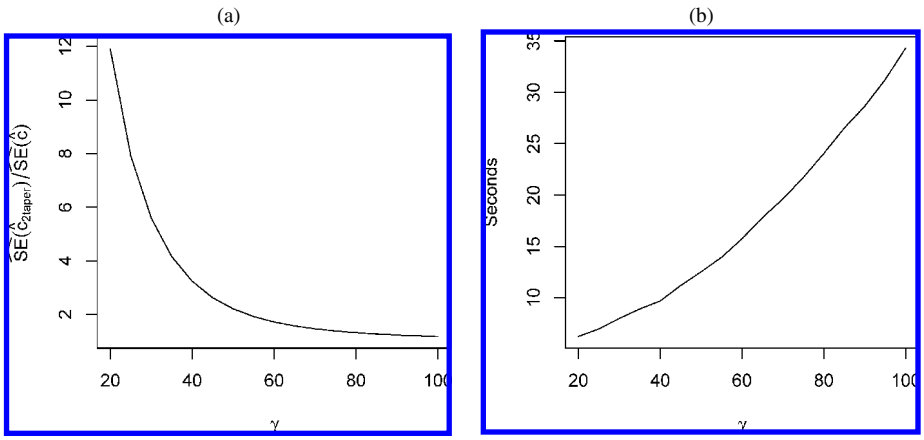

Figure 2. The ratio of estimated standard errors for $\hat{c}_{2taper}$ (a function of γ) and the MLE $\hat{c}$, and the time required for one evaluation of l_{2taper} for various γ.

Table 2. Seconds required for each step in evaluating the log-likelihood and tapering approximations

	No taper	One taper	Two tapers
$\mathbf{\Gamma}$ or $\mathbf{\Gamma} \circ \mathbf{T}$	4.11	.32	.32
Cholesky decomposition	578.84	.24	.24
Log determinant	.22	0	0
Backsolving	.88	.01	—
Full solving	—	—	18.96
Quadratic form	0	0	.12
Total	584.05	.56	19.63

ing the pairs of observations with distance less than the taper range, which took about 10 seconds. Because the number of evaluations needed to then minimize each function will vary in practice, Table 2 reports the computation times for a single evaluation of each function, broken down into its component steps. These were calculated by averaging over 10 repetitions of each function evaluation. Some initial savings is realized by calculating $\mathbf{\Gamma} \circ \mathbf{T}$ rather than $\mathbf{\Gamma}$, because we need to compute only the correlation for those distances less than the taper range; however, the most sizeable savings come in calculating the Cholesky decomposition, which is three orders of magnitude faster for the tapered correlation matrix than for the full matrix. The two-taper approximation requires that we compute $(\mathbf{\Gamma} \circ \mathbf{T})^{-1} \circ \mathbf{T}$ rather than simply backsolving using the Cholesky decomposition, which in this case adds an additional 19 seconds.

7. DISCUSSION

We have proposed two approximations to the likelihood for use in large spatial data sets. The one-taper approximation (3) replaces the model covariance matrix with a tapered version, whereas the two-taper approximation (5) tapers both the model and sample covariance matrixes. Both approximations provide significant computational gains over the full likelihood. The one-taper approximation is more computationally efficient than the two-taper approximation, but it suffers from bias when the taper range is small relative to the correlation range of the process. In contrast, the two-taper approximation shows little bias and only slightly increased variance.

We have given conditions for almost sure convergence of the tapering estimators of the Matérn covariance. The conditions on the one-taper estimator are straightforward, relying on the equivalence of the Gaussian measures with tapered and untapered covariances. The conditions on the two-taper estimator are less straightforward, and it would be worthwhile to study whether a simpler set of conditions might be sufficient. We might follow the estimating equations approach (as in Heyde 1997), but we have not been able to make progress along these lines, because an important assumption—that the sequence of estimating functions is a martingale—does not hold in this case.

In finite samples, we showed that the two-taper estimators can have sampling distributions close to those of the MLEs. Both the bias and variance remain comparable to that of the MLE, even when γ is small. The estimator of sampling variability for the two-taper estimators based on the robust information matrix performs comparably to the Fisher information–based estimate for the MLE.

The one-taper estimators displayed sizeable bias in our simulation study when γ was small relative to the correlation range of the process. However, one instance in which we anticipate the one-taper approximation to perform well for a variety of taper ranges is in plug-in prediction. The interpolation of a random field is beyond the scope of this article, but when the data set is large enough to warrant tapering in the estimation of model parameters, it also is typically large enough to warrant tapering for interpolation. Our preliminary work suggests that when tapering is used in the kriging procedure, it is better to plug in the one-taper estimators rather than the two-taper estimators. This is intuitively plausible, because it uses the same covariance model for both estimation and prediction; however, the two-taper approximation does give more efficient estimates of the covariance parameters under the original model, without the large bias observed in the one-taper estimates.

The tapering estimators that we have developed may be extended in several ways. First consider the case that the mean is not 0 but is a linear combination of regression functions, so that $\mathbf{Z} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma}(\boldsymbol{\theta}))$, where $\mathbf{X}$ is an $n \times p$ fixed regression matrix and $\boldsymbol{\beta}$ is a vector of p unknown coefficients. Then we may approximate the log-likelihood by

$$l_{2tapers}(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\frac{1}{2}\log|\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}| - \frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'([\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}]^{-1} \circ \mathbf{T})(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).$$

This choice gives unbiased estimating equations in both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Another extension of these ideas would be to nonisotropic or nonstationary covariance functions. One possible approach to this problem would be to consider the class of random fields that are stationary and isotropic subject to some transformation of the underlying space on which the process is defined, as done by Sampson and Guttorp (1992). In this case tapering with a stationary and isotropic correlation function in the transformed space would have the effect of differentially tapering in the original space. We anticipate that tapering may be used to simplify the computation in a wide class of spatial models, perhaps guided by some of the theoretical concerns presented here.

APPENDIX A: PROPERTIES OF THE SHUR PRODUCT

This appendix collects some relevant results on the Schur product. (The interested reader should see Horn and Johnson 1991, chap. 5 for more details.)

Definition. Two $m \times n$ matrixes, $\mathbf{A}$ and $\mathbf{B}$, have Schur product $\mathbf{A} \circ \mathbf{B} = \{a_{ij}b_{ij}\}$.

The Schur product theorem. If $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite $n \times n$ matrixes, then so is $\mathbf{A} \circ \mathbf{B}$. If in addition, $\mathbf{B}$ is positive definite and $\mathbf{A}$ has no diagonal entry equal to 0, then $\mathbf{A} \circ \mathbf{B}$ is positive definite. In particular, if both $\mathbf{A}$ and $\mathbf{B}$ are positive definite, then so is $\mathbf{A} \circ \mathbf{B}$.

Commutivity. Unlike the standard matrix product, $\mathbf{A} \circ \mathbf{B} = \mathbf{B} \circ \mathbf{A}$.

Eigenvalue inequalities. If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ positive semidefinite matrixes, then any eigenvalue $\lambda(\mathbf{A} \circ \mathbf{B})$ of $\mathbf{A} \circ \mathbf{B}$ satisfies

$$\left[\min_{1\le i\le n} a_{ii}\right]\lambda_{min}(\mathbf{B}) \le \lambda(\mathbf{A} \circ \mathbf{B}) \le \left[\max_{1\le i\le n} a_{ii}\right]\lambda_{max}(\mathbf{B}), \tag{A.1}$$

where $\{a_{ii}\}$ are the diagonal entries of $\mathbf{A}$ and $\lambda_{min}(\mathbf{B})$ and $\lambda_{max}(\mathbf{B})$ are the minimum and maximum eigenvalues of $\mathbf{B}$.

Trace. For square matrixes $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, with $\mathbf{B}$ symmetric, $\mathrm{tr}\{(\mathbf{A} \circ \mathbf{B})\mathbf{C}\} = \mathrm{tr}\{\mathbf{A}(\mathbf{B} \circ \mathbf{C})\}$.

APPENDIX B: PROOFS

Proof of Theorem 1

Let f_1 be the spectral density corresponding to K_1. Because the Fourier transform of the product of two functions is the convolution of their Fourier transforms, we can write

$$f_1(\boldsymbol{\omega}) = \int_{\Re^d} f_0(\mathbf{x}) f_{taper}(\boldsymbol{\omega} - \mathbf{x})\, d\mathbf{x},$$

where $f_0(\mathbf{x}) = \sigma^2 M_0(\rho^{-2} + \|\mathbf{x}\|^2)^{-\nu-d/2}$ is the spectral density corresponding to the Matérn covariance function K_0.

Stein (2004, thm. A.1) provided the following two conditions for the equivalence of $G(K_0)$ and $G(K_1)$ on the paths of Z for bounded subsets: First, that there exists $\eta > d$ such that $f_0(\boldsymbol{\omega})\|\boldsymbol{\omega}\|^\eta$ is bounded away from 0 and ∞ as $\|\boldsymbol{\omega}\| \to \infty$, and second, that there exists $c < \infty$ such that

$$\int_{\|\boldsymbol{\omega}\|>c} \left\{\frac{f_1(\boldsymbol{\omega}) - f_0(\boldsymbol{\omega})}{f_0(\boldsymbol{\omega})}\right\}^2 d\boldsymbol{\omega} < \infty. \tag{B.1}$$

The Matérn spectral density f_0 satisfies the first condition when $\eta = 2\nu + d$. Rewriting the integral in (B.1) using polar coordinates gives

$$\int_{S^d}\int_c^\infty \left\{\frac{f_1(r\mathbf{u}) - f_0(r\mathbf{u})}{f_0(r\mathbf{u})}\right\}^2 r^{d-1}\, dr\, dU(\mathbf{u}),$$

where S^d is the surface of the unit sphere in $\Re^d$ and U is the uniform probability measure on the sphere. To show that (B.1) holds, it thus is sufficient to show that

$$\left|\frac{f_1(r\mathbf{u})}{f_0(r\mathbf{u})} - 1\right| = O(r^{-\xi}), \quad \text{for some } \xi > d/2, \tag{B.2}$$

for all $\mathbf{u} \in S^d$. Throughout, let $f(r) = O(g(r))$ indicate that $f(r) \ge 0$ and there exist positive finite constants L and c such that $f(r) \le Lg(r)$ for all $r \ge c$.

Let $\mathbf{u}$ be an arbitrary unit vector. Then for all $r > 0$, define $N_r = \{\mathbf{x} \in \Re^d : \|r\mathbf{u} - \mathbf{x}\| \le r^k\}$, where k is specified later. Then

$$\left|\frac{f_1(r\mathbf{u})}{f_0(r\mathbf{u})} - 1\right| \le \left|\frac{\int_{N_r^c} f_0(\mathbf{x}) f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}}{f_0(r\mathbf{u})}\right| \tag{B.3}$$

$$+ \left|\frac{\int_{N_r} f_0(\mathbf{x}) f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}}{f_0(r\mathbf{u})} - 1\right|. \tag{B.4}$$

Because $d \le 3$ and $\epsilon > d/4$, we may choose $\xi \in (d/2, \min\{2, 2\epsilon\})$. Then choose $k \in ((d + 2\nu + \xi)/(d + 2\nu + 2\epsilon), 1)$. In the remainder of the proof we show that with this choice of k, both (B.3) and (B.4) are $O(r^{-\xi})$, so that (B.2) holds.

First, consider (B.3). When $\mathbf{x} \in N_r^c$, $\|r\mathbf{u} - \mathbf{x}\| > r^k$; thus, using the bound in the theorem, we have

$$\begin{aligned} f_{taper}(r\mathbf{u} - \mathbf{x}) &\le \frac{M_\epsilon}{(1 + \|r\mathbf{u} - \mathbf{x}\|^2)^{\nu+d/2+\epsilon}} \\ &\le \frac{M_\epsilon}{(1 + r^{2k})^{\nu+d/2+\epsilon}}. \end{aligned}$$

Also note that $\int_{\Re^d} f_0(\mathbf{x})\, d\mathbf{x} = \sigma^2$, so

$$\frac{\int_{N_r^c} f_0(\mathbf{x}) f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}}{f_0(r\mathbf{u})} \le \frac{M_\epsilon}{M_0}\frac{(\rho^{-2} + r^2)^{\nu+d/2}}{(1 + r^{2k})^{\nu+d/2+\epsilon}},$$

and (B.3) is $O(r^{-\xi})$ because we chose $k > \frac{d+2\nu+\xi}{d+2\nu+2\epsilon}$.

Now consider (B.4). Expanding $f_0(\mathbf{x})$ about $r\mathbf{u}$, we have

$$\begin{aligned} &\left|\frac{\int_{N_r} f_0(\mathbf{x}) f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}}{f_0(r\mathbf{u})} - 1\right| \\ &\quad\le \left|1 - \int_{N_r} f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}\right| \qquad \text{(B.5)} \\ &\qquad + \left|\frac{1}{f_0(r\mathbf{u})}\int_{N_r} (\mathbf{x} - r\mathbf{u})'[\nabla f_0(r\mathbf{u})] \right. \\ &\qquad\quad \left. \times f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}\right| \qquad \text{(B.6)} \\ &\qquad + \left|\frac{1}{2f_0(r\mathbf{u})}\int_{N_r} (\mathbf{x} - r\mathbf{u})'[\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})](\mathbf{x} - r\mathbf{u}) \right. \\ &\qquad\quad \left. \times f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x}\right|, \qquad \text{(B.7)} \end{aligned}$$

where $\nabla f_0(r\mathbf{u})$ is the vector of derivatives of f_0 evaluated at $r\mathbf{u}$ and $\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})$ is the matrix of second derivatives evaluated at a point $\mathbf{m}_{\mathbf{x},r}$ lying between $\mathbf{x}$ and $r\mathbf{u}$, thus in N_r.

Because K_{taper} is a correlation function, f_{taper} is a probability density, and so

$$\begin{aligned} 1 - \int_{N_r} f_{taper}(r\mathbf{u} - \mathbf{x})\, d\mathbf{x} &= \int_{\|\mathbf{y}\|>r^k} f_{taper}(\mathbf{y})\, d\mathbf{y} \\ &\le \int_{\|\mathbf{y}\|>r^k} M_\epsilon(1 + \|\mathbf{y}\|^2)^{-\nu-d/2-\epsilon}\, d\mathbf{y} \\ &\le \int_{\|\mathbf{y}\|>r^k} M_\epsilon \|\mathbf{y}\|^{-2(\nu+d/2+\epsilon)}\, d\mathbf{y} \\ &= \frac{2M_\epsilon \pi^{d/2}}{\Gamma(d/2)}\int_{r^k}^\infty s^{-2(\nu+d/2+\epsilon)} s^{d-1}\, ds \\ &= \frac{M_\epsilon \pi^{d/2}}{\Gamma(d/2)(\nu + \epsilon)} r^{-2k(\nu+\epsilon)}. \end{aligned}$$

Therefore, (B.5) is $O(r^{-\xi})$ because $\xi < 2\epsilon$ and $\nu > 0$ imply that $k > \frac{\xi}{2(\nu+\epsilon)}$. The integral in (B.6) is equal to 0 because f_{taper} is isotropic. That is, for each $\mathbf{x} \in N_r$, $\exists \mathbf{y} \in N_r$ such that $\mathbf{x} - r\mathbf{u} = -(\mathbf{y} - r\mathbf{u})$, but $f_{taper}(r\mathbf{u} - \mathbf{x}) = f_{taper}(r\mathbf{u} - \mathbf{y})$. Therefore, we can divide N_r into two regions, whose integrals have opposite sign.

Finally, considering (B.7), first note that for each $r > 0$ and $\mathbf{x} \in N_r$,

$$\begin{aligned} &(\mathbf{x} - r\mathbf{u})'[\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})](\mathbf{x} - r\mathbf{u}) \\ &\quad = \|\mathbf{x} - r\mathbf{u}\|^2 \mathbf{v}'[\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})]\mathbf{v}, \quad \text{where } \|\mathbf{v}\| = 1 \\ &\quad \le \|\mathbf{x} - r\mathbf{u}\|^2 \sup_{\|\mathbf{v}\|=1} \mathbf{v}'[\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})]\mathbf{v} \\ &\quad = \|\mathbf{x} - r\mathbf{u}\|^2 \lambda_{max}\{\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})\}, \end{aligned}$$

where $\lambda_{max}\{\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})\}$ represents the maximum eigenvalue of $\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})$. Because f_0 is isotropic, we can show that

$$\nabla^2 f_0(\mathbf{m}) = \frac{1}{\|\mathbf{m}\|^2}\left[g''(\|\mathbf{m}\|) - \frac{g'(\|\mathbf{m}\|)}{\|\mathbf{m}\|}\right]\mathbf{m}\mathbf{m}' + \frac{g'(\|\mathbf{m}\|)}{\|\mathbf{m}\|}\mathbf{I}_d,$$

where $g(r) = \sigma^2 M_0(\rho^{-2} + r^2)^{-(\nu+d/2)}$. Because these two matrixes are symmetric, the maximum eigenvalue of their sum is less than or equal to the sum of their maximum eigenvalues (Horn and Johnson 1991, equation 3.4.11a). We have

$$\begin{aligned}
&\lambda_{max}\{\nabla^2 f_0(\mathbf{m})\} \\
&\quad\le \lambda_{max}\left\{\frac{1}{\|\mathbf{m}\|^2}\left[g''(\|\mathbf{m}\|) - \frac{g'(\|\mathbf{m}\|)}{\|\mathbf{m}\|}\right]\mathbf{m}\mathbf{m}'\right\} \\
&\qquad + \lambda_{max}\left\{\frac{g'(\|\mathbf{m}\|)}{\|\mathbf{m}\|}\mathbf{I}_d\right\} \\
&\quad= g''(\|\mathbf{m}\|) - \frac{g'(\|\mathbf{m}\|)}{\|\mathbf{m}\|} + \frac{g'(\|\mathbf{m}\|)}{\|\mathbf{m}\|} \\
&\quad= g''(\|\mathbf{m}\|) \\
&\quad= \frac{\sigma^2 M_0(2\nu+d)}{(\rho^{-2}+\|\mathbf{m}\|^2)^{\nu+d/2+1}}\left[\frac{(2\nu+d+2)\|\mathbf{m}\|^2}{\rho^{-2}+\|\mathbf{m}\|^2} - 1\right].
\end{aligned}$$

This function is eventually decreasing with $\|\mathbf{m}\|$. Also, note that $\|\mathbf{m}_{\mathbf{x},r}\| > r - r^k$, because $(r - r^k)\mathbf{u}$ is the point on the boundary of N_r that is closest to the origin, and $\mathbf{m}_{\mathbf{x},r}$ is defined to be in N_r. Because $k < 1$, $r - r^k \to \infty$, and so eventually $g''(\|\mathbf{m}_{\mathbf{x},r}\|) \le g''(r - r^k)$ for all $\mathbf{x} \in N_r$. Therefore, for sufficiently large r,

$$\begin{aligned}
&\frac{1}{2f_0(r\mathbf{u})}\int_{N_r} (\mathbf{x}-r\mathbf{u})'[\nabla^2 f_0(\mathbf{m}_{\mathbf{x},r})](\mathbf{x}-r\mathbf{u})\, f_{taper}(r\mathbf{u}-\mathbf{x})\, d\mathbf{x} \\
&\qquad\le \frac{g''(r-r^k)}{2g(r)}\int_{N_r} \|\mathbf{x}-r\mathbf{u}\|^2 f_{taper}(r\mathbf{u}-\mathbf{x})\, d\mathbf{x}.
\end{aligned}$$

Using the bound on f_{taper} given in the theorem, we have

$$\begin{aligned}
&\int_{N_r} \|\mathbf{x}-r\mathbf{u}\|^2 f_{taper}(r\mathbf{u}-\mathbf{x})\, d\mathbf{x} \\
&\quad= \int_{\|\mathbf{y}\|\le r^k} \|\mathbf{y}\|^2 f_{taper}(\mathbf{y})\, d\mathbf{y} \\
&\quad\le \int_{\|\mathbf{y}\|\le r^k} \|\mathbf{y}\|^2 \frac{M_\epsilon}{(1+\|\mathbf{y}\|^2)^{\nu+d/2+\epsilon}}\, d\mathbf{y} \\
&\quad\le \int_{\Re^d} \|\mathbf{y}\|^2 \frac{M_\epsilon}{(1+\|\mathbf{y}\|^2)^{\nu+d/2+\epsilon}}\, d\mathbf{y} \\
&\quad\propto E(\|\mathbf{Y}\|^2),
\end{aligned}$$

where $\mathbf{Y} \sim t_{d,2(\nu+\epsilon)}/\sqrt{2(\nu+\epsilon)}$. But $(\nu+\epsilon) > 1$, so this term is finite. Therefore, we need only consider $g''(r-r^k)/g(r)$. But this is $O(r^{-\xi})$, because $k \in (0,1)$ and $\xi < 2$.

Proof of Theorem 2

By theorem 2 of Zhang (2004), we may find a $\sigma^{2*} > 0$ such that $G(K_0) \equiv G(K_0^*)$, where K_0^* is Matérn with parameters σ^{2*}, ρ^*, and ν. By Theorem 1, $G(K_0^*) \equiv G(K_1^*)$, where $K_1^* = K_0^* K_{taper}$. Therefore, to show $\hat{\sigma}^2_{n,1taper}/\rho^{*2\nu} \to \sigma^2/\rho^{2\nu}$ a.s. $[G(K_0)]$, it is sufficient to show that $\hat{\sigma}^2_{n,1taper} \to \sigma^{2*}$ a.s. $[G(K_1^*)]$. Because ρ^* and ν are fixed, $\hat{\sigma}^2_{n,1taper} = \mathbf{Z}_n[\mathbf{\Gamma}_n^* \circ \mathbf{T}_n]^{-1}\mathbf{Z}_n/n$, where $\mathbf{\Gamma}_n^* = \{K_0^*(\|\mathbf{s}_i - \mathbf{s}_j\|; \sigma^{2*}, \rho^*, \nu)\}/\sigma^{2*}$ and $\mathbf{T} = \{K_{taper}(\|\mathbf{s}_i - \mathbf{s}_j\|)\}$. Under $G(K_1^*)$, $\mathbf{Z} \sim \mathrm{MVN}(0, \sigma^{2*}\mathbf{\Gamma}_n^* \circ \mathbf{T}_n)$, so $\hat{\sigma}^2_{n,1taper}$ is distributed as σ^{2*}/n times a chi-squared random variable with n degrees of freedom; therefore, $\hat{\sigma}^2_{n,1taper} \to \sigma^{2*}$ a.s. $[G(K_1^*)]$ by the strong law of large numbers.

Proof of Theorem 3

Write $\mathbf{\Gamma}_n = \mathbf{R}_n\mathbf{R}_n'$. Then $\frac{1}{\sigma}\mathbf{R}_n^{-1}\mathbf{Z}_n \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I_n})$, so

$$\begin{aligned}
\hat{\sigma}^2_{n,2tapers} &= \mathbf{Z}_n'([\mathbf{\Gamma}_n \circ \mathbf{T}_n]^{-1} \circ \mathbf{T}_n)\mathbf{Z}_n/n \\
&= \frac{1}{n}\mathbf{X}_n'\big[(\sigma\mathbf{R}_n)'[(\mathbf{\Gamma}_n \circ \mathbf{T}_n)^{-1} \circ \mathbf{T}_n](\sigma\mathbf{R}_n)\big]\mathbf{X}_n, \\
&\quad \text{where } \mathbf{X}_n \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}_n) \\
&= \frac{\sigma^2}{n}\sum_{i=1}^n \lambda_{n,i}\chi_i^2, \qquad \text{(B.8)}
\end{aligned}$$

where χ_i^2 are iid χ_1^2 random variables and λ_{n_i} is the ith eigenvalue of $\mathbf{R}_n'[(\mathbf{\Gamma}_n \circ \mathbf{T}_n)^{-1} \circ \mathbf{T}_n]\mathbf{R}_n$, which is the same as the ith eigenvalue of $\mathbf{W}_n = [(\mathbf{\Gamma}_n \circ \mathbf{T}_n)^{-1} \circ \mathbf{T}_n]\mathbf{\Gamma}_n$.

Cuzick (1995) gave conditions for the a.s. convergence of weighted sums of iid random variables. Specifically, let $Y_n = \sum_{i=1}^n a_{n,i}X_i$, where X_i is iid with mean 0 and $\{a_{n,i}\}$ is an array of constants. Then if $\sup_n(n^{-1}\sum_{i=1}^n |a_{n,i}|^q)^{1/q} < \infty$ for some $1 < q \le \infty$, and $E|X|^p < \infty$, $p^{-1} + q^{-1} = 1$, then $Y_n/n \to 0$ almost surely. (The case where $q = 0$ is interpreted to mean that the $a_{n,i}$ are bounded uniformly.) This result also holds when $q = 1$ under the additional assumption that $\limsup_{i\le n} |a_{n,i}|n^{-1}\log n$. Finish the proof by applying these results to (B.8), with $X_i = \chi_i^2 - 1$ and $a_{n,i} = \lambda_{n,i}$.

Proof of Corollary 1

By theorem 2 of Zhang (2004), we can find $\sigma^{2*} > 0$ such that $G(K_0) \equiv G(K_0^*)$, where K_0^* is Matérn with parameters σ^{2*}, ρ^*, and ν. That is, let $\sigma^{2*} = \sigma_0^2(\rho_0/\rho^*)^{2\nu}$. Now it is sufficient to show that $\hat{\sigma}^{2*}_{n,2tapers}/\rho^{*2\nu} \to \sigma_0^2/\rho_0^{2\nu}$ a.s. $[G(K_0^*)]$. This follows directly from the conditions on $\mathbf{W}_n$ and Theorem 3.

Proof of Lemma 1

$$\begin{aligned}
\lambda_{max}\big\{[(\mathbf{\Gamma} \circ \mathbf{T})^{-1} \circ \mathbf{T}]\mathbf{\Gamma}\big\} &\le \lambda_{max}\big\{[(\mathbf{\Gamma} \circ \mathbf{T})^{-1} \circ \mathbf{T}]\big\}\lambda_{max}\{\mathbf{\Gamma}\} \\
&\le \lambda_{max}\{(\mathbf{\Gamma} \circ \mathbf{T})^{-1}\}\lambda_{max}\{\mathbf{\Gamma}\} \\
&= \frac{\lambda_{max}\{\mathbf{\Gamma}\}}{\lambda_{min}\{(\mathbf{\Gamma} \circ \mathbf{T})\}} \\
&\le \frac{\lambda_{max}\{\mathbf{\Gamma}\}}{\lambda_{min}\{\mathbf{\Gamma}\}}.
\end{aligned}$$

Here we have used (A.1) in the second and last lines.

[Received October 2007. Revised August 2008.]

REFERENCES

Abromowitz, M., and Stegun, I. (1967), *Handbook of Mathematical Functions*, New York: Dover.

Bickel, P., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley-Interscience.

Cuzick, J. (1995), "A Strong Law for Weighted Sums of i.i.d. Random Variables," *Journal of Theoretical Probability*, 8, 625–641.

Fuentes, M. (2007), "Approximate Likelihood for Large Irregularly Spaced Spatial Data," *Journal of the American Statistical Association*, 102, 321–331.

Furrer, R., and Bengtsson, T. (2007), "Estimation of High-Dimensional Prior and Posterior Covariance Matrices in Kalman Filter Variants," *Journal of Multivariate Analysis*, 98, 227–255.

Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523.

Gneiting, T. (2002), "Compactly Supported Correlation Functions," *Journal of Multivariate Analysis*, 83, 493–508.

Gray, R. (2006), "Toeplitz and Circulant Matrices: A Review," *Foundations and Trends in Communications and Information Theory*, 2, 155–239.

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

Horn, R., and Johnson, C. (1991), *Topics in Matrix Analysis*, Cambridge, U.K.: Cambridge University Press.

Johns, C., Nychka, D., Kittel, T., and Daly, C. (2003), "Infilling Sparse Records of Spatial Fields," *Journal of the American Statistical Association*, 98, 796–806.

Kaufman, C. G. (2006), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Datasets," unpublished doctoral thesis, Carnegie Mellon University.

Kitanidis, P. (1983), "Statistical Estimation of Polynomial Generalized Covariance Functions and Hydrologic Applications," *Water Resources Research*, 19, 909–921.

Mardia, K., and Marshall, R. (1984), "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression," *Biometrika*, 71, 135–146.

Matérn, B. (1986), *Spatial Variation*, Berlin: Springer-Verlag.

Pissanetzky, S. (1984), *Sparse Matrix Technology*, London: Academic Press.

Sampson, P. D., and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure," *Journal of the American Statistical Association*, 87, 108–119.

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

——— (2004), "Equivalence of Gaussian Measures for Some Nonstationary Random Fields," *Journal of Statistical Planning and Inference*, 123, 1–11.

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Ser. B, 66, 275–296.

Vecchia, A. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society*, Ser. B, 50, 297–312.

Watkins, A., and Al-Boutiahi, F. (1990), "On Maximum Likelihood Estimation of Parameters in Incorrectly Specified Models of Covariance for Spatial Data," *Mathematical Geology*, 22, 151–173.

Wendland, H. (1995), "Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree," *Advances in Computational Mathematics*, 4, 389–396.

——— (1998), "Error Estimates for Interpolation by Compactly Supported Radial Basis Functions of Minimal Degree," *Journal of Approximation Theory*, 93, 258–272.

Whittle, P. (1954), "On Stationary Processes in the Plane," *Biometrika*, 41, 434–449.

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

Zimmerman, D. (1989), "Computationally Exploitable Structure of Covariance Matrices and Generalized Covariance Matrices in Spatial Models," *Journal of Statistical Computation and Simulation*, 32, 1–15.