Covariance Tapering for Likelihood Based Estimation in Large Spatial Datasets

Cari Kaufman, Mark Schervish, and Douglas Nychka

Cari Kaufman is Postdoctoral Researcher, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO 80305 (email: cgk@ucar.edu), with a joint appointment at the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709; Mark Schervish is Professor and Department Head, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15214 (email: mark@stat.cmu.edu); and Douglas Nychka is Director, Institute for Mathematics Applied to Geosciences and Senior Scientist, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO 80305 (email: nychka@ucar.edu).

Abstract

Maximum likelihood is an attractive method of estimating covariance parameters in spatial models based on Gaussian processes. However, calculating the likelihood can be computationally infeasible for large datasets, requiring $O(n^3)$ calculations for a dataset with n observations. This article proposes the method of covariance tapering to approximate the likelihood in this setting. In this approach, covariance matrices are "tapered," or multiplied element-wise by a sparse correlation matrix. The resulting matrices can then be manipulated using efficient sparse matrix algorithms. We propose two approximations to the Gaussian likelihood using tapering. One simply replaces the model covariance with a tapered version; 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 estimate, 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.

Keywords: Gaussian process, covariance estimation, compactly supported correlation function, estimating equations

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 dataset 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" dataset on a desktop computer with 2 GB of RAM would be about ten thousand 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 via simulation that their sampling distributions can be quite similar to that of the MLE, even when the approximation is severe.

We consider the commonly used model that 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 zero and 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 $\mathbf{Z} = (Z(\mathbf{s_1}), \ldots, Z(\mathbf{s_n}))'$. For example, one commonly used isotropic covariance function is the exponential function $K(x; \sigma^2, \rho) = \sigma^2 \exp\{-x/\rho\}$.

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

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

(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 distribution 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 datasets. However, these authors assumed the covariance parameters were known, while 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 datasets 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). There are fewer techniques 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, later extended by Stein et al. (2004). One partitions $\mathbf{Z}$ into subvectors $\mathbf{Z_1}, \ldots, \mathbf{Z_b}$, then writes the likelihood as a product of conditional densities $p(\mathbf{Z_j}|\mathbf{Z_{(j-1)}}; \boldsymbol{\theta})$, where $\mathbf{Z_{(j)}}' = (\mathbf{Z_1}', \ldots, \mathbf{Z_j}')$. One then replaces the full conditioning sets $\mathbf{Z_{(j-1)}}$ with smaller subsets $\tilde{\mathbf{Z}}_{\mathbf{(j-1)}} \subseteq \mathbf{Z_{(j-1)}}$, so

that the densities in the product are easier to evaluate. Vecchia (1988) chose $\tilde{\mathbf{Z}}_{(\mathbf{j})}$ to consist of nearest neighbors within $\mathbf{Z}_{(\mathbf{j})}$. Stein et al. (2004) extended Vecchia's idea to restricted maximum likelihood estimators 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 there is little information lost in taking them to be conditionally independent given intermediate locations. A similar idea motivates the covariance tapering approach we explore in this paper.

2 Likelihood Approximation Using Tapering

If we have reason to believe that distant pairs of observations are independent, we can model this structure using a compactly supported covariance function (Gneiting, 2002). Then $\mathbf{\Sigma}(\boldsymbol{\theta})$ then contains zeroes 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 the underlying process possesses such a covariance function, we can use this idea for computational purposes. The goal is to set to zero certain elements of the covariance matrix, such that the resulting matrix is positive definite and retains the original properties of $\mathbf{\Sigma}(\boldsymbol{\theta})$ for proximate locations. To 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 which is identically zero whenever $x \geq \gamma$. Denote this tapered covariance function by

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

The tapered covariance matrix is denoted $\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)$, 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 matrices is again a valid covariance matrix. In addition, requiring K_{taper} to be a correlation function ensures 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 (2007) 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 (2006) also used tapering as a regularization technique for the ensemble Kalman filter. However, these two papers differ from the present context in that they are concerned with estimating the covariance matrix, rather than the parameters of a particular covariance function.

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

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

This is equivalent to using a model in which the process Z is Gaussian with mean zero and covariance function (2). The effects of misspecifying the covariance function have been widely studied with respect to kriging (see Section 4.3 of Stein, 1999), 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, $\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{1taper}(\boldsymbol{\theta})\right] \neq \mathbf{0}$. This means there is no guarantee that the estimator maximizing (3) is asymptotically unbiased. Moreover, in our experience, this estimator can sometimes display sizable bias in practice, especially if the taper range is small relative to the correlation range of the process.

To remedy the bias, one can take an estimating equations approach to formulating

a tapered version of this problem. Essentially, we taper both the covariance matrix and the sample covariance matrix. First, note that one can rewrite the quadratic form in (1) as a trace involving the sample covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{Z}\mathbf{Z}'$:

$$\mathbf{Z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z} = \mathrm{tr}\left\{\mathbf{Z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}\right\} = \mathrm{tr}\left\{\mathbf{Z}\mathbf{Z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\right\} = \mathrm{tr}\left\{\hat{\mathbf{\Sigma}}\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\right\}. \tag{4}$$

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

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

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

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

We refer to (3) and (5) as the one taper and two taper approximations, respectively, and to the estimators maximizing them as the one and two taper estimators. The choice of approximation is context dependent. The one taper approximation is computationally more efficient, as it involves solving the sparse system of equations $\left[\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)\right]^{-1}\mathbf{Z}$, whereas the two taper approximation requires the inverse of a sparse matrix to compute $\left(\left[\mathbf{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)\right]^{-1} \circ \mathbf{T}(\gamma)\right)\mathbf{Z}$. In addition, asymptotic results are more straightforward for the one taper approximation, as we discuss in

the following section. However, 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 one can estimate sampling variability using the robust information criterion (Heyde, 1997). Therefore, 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, in which we illustrate the tapering methods on a large spatial dataset.

3 Asymptotic Results

Two commonly used asymptotic frameworks in spatial statistics are "increasing domain" and "fixed domain" asymptotics (see e.g. Cressie, 1993, Section 5.8). Under increasing domain asymptotics, the sampling region increases without bound, while 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 is widely used in practice and has easily interpretable parameters (Matérn, 1986; Stein, 1999). This function is defined by

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

where $\mathcal{K}_\nu$ is the modified Bessel function of order ν (see Abromowitz and Stegun, 1967, Section 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, Section 2.7, for details).

Zhang (2004) proved several important results about the Matérn class. The first concerns the equivalence of two mean-zero Gaussian measures $G(K_0)$ and $G(K_1)$. (Throughout, let $G(K)$ denote the mean zero 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 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 $\{Z(\mathbf{s}), \mathbf{s} \in T\}$ 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 $d \leq 3$, there cannot exist consistent estimators of both σ^2 and ρ. However, 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 which 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 zero 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 $K_1 = K_0 K_{taper}$, where K_{taper} is an isotropic correlation function*

on $\Re^d$. *Suppose the spectral density* $f_{taper} = (2\pi)^{-d} \int_{\Re^d} \exp\{-i\boldsymbol{\omega}'\mathbf{x}\} K_{taper}(\mathbf{x}) d\mathbf{x}$ *exists and 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.

One 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 $||\mathbf{x}|| > 1$, $\phi_{d,k}(||\mathbf{x}||) = 0$. If $f_{d,k}$ is the spectral density corresponding to $\phi_{d,k}$, Wendland (1998) showed that 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

Theorem 1 can now be used to prove almost sure convergence of the one taper estimator, a direct analogue 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* $\ell_{n,1taper}$ *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* $\ell_{n,1taper}(\sigma^2, \rho^*)$. *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, we are here defining "largeness" relative to the taper range, rather than allowing the taper range to shrink with n and defining "largeness" in an absolute sense.

Unlike ℓ_{1taper}, $\ell_{2tapers}$ does not correspond to altering the 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 to use this expression to determine conditions necessary for convergence. However, it is also possible to use this result to prove convergence in the case that *both* σ^2 and ρ of the Matérn covariance function are unknown. This result is given as a corollary directly following 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 $\ell_{n,2tapers}(\sigma^2)$ be the two taper approximation (5) based on observations of Z at locations in S_n. For all n, define the matrix $\mathbf{W_n} = [(\mathbf{\Gamma_n} \circ \mathbf{T_n})^{-1} \circ \mathbf{T_n}]\,\mathbf{\Gamma_n}$, where $(\mathbf{\Gamma_n})_{ij} = C_0(||\mathbf{s_i} - \mathbf{s_j}||)$ 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 either $\sup_n \left(n^{-1}\sum_{i=1}^{n}\lambda_{n,i}^q\right)^{1/q} < \infty$ for some $1 < q \leq \infty$, or $\lim_n(\sup_{i\leq n}\lambda_{n,i})n^{-1}\log n = 0$. 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 $\Re^d, d \leq 3$ with known parameter ν and unknown parameters σ^2 and ρ. Fix $\rho^* > 0$, and let $\hat{\sigma}^2_{n,2taper}$ maximize $\ell_{n,2tapers}(\sigma^2,\rho^*)$. Define $\mathbf{W_n}$ as in Theorem 3, but with $(\mathbf{\Gamma_n})_{ij} = \sigma^{-2}K_0(||\mathbf{s_i} - \mathbf{s_j}||;\sigma^2,\rho^*,\nu)$. Suppose 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 one to ignore the choice of tapering function, using a bound that depends only on the correlation function of the process.

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

We illustrate the use of this lemma in a simple example. Suppose the process lies in $\Re$, with correlation function $C_0(x) = \exp\{-|x|/\rho\}$ and $\rho > 0$ known. Suppose the sampling locations are equally spaced, with $s_i = i\Delta, i = 1, \ldots, n$. Then $\mathbf{\Gamma_n}$ 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 $\mathbf{\Gamma_n}$ satisfy $ess\inf f \leq \tau_{n,k} \leq ess\sup f$ (applying Gray, 2006, Lemma 4.1). Therefore, by Lemma 1, $\lambda_{max}\{\mathbf{W_n}\} \leq \lambda_{max}\{\mathbf{\Gamma_n}\}/\lambda_{min}\{\mathbf{\Gamma_n}\} \leq ess\sup f/ess\inf f$. Since $f(\lambda)$ has a maximum at 0 of $\coth(\Delta/(2\rho))$ and a minimum at π of $\tanh(\Delta/(2\rho))$,

$$\lambda_{max}\{\mathbf{W_n}\} \leq \frac{\lambda_{max}\{\mathbf{\Gamma_n}\}}{\lambda_{min}\{\mathbf{\Gamma_n}\}} \leq \coth^2\left(\frac{\Delta}{2\rho}\right) < \infty \tag{7}$$

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 that Δ is not fixed but depends on n. In particular, suppose $\Delta_n = \Delta/n^k$ for some k. The case $k = 0$ corresponds to increasing domain sampling. The case $k = 1$ 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

$$\coth^2\left(\frac{\Delta_n}{2\rho}\right)\frac{\log n}{n} = \left(\frac{e^{-\Delta_n/\rho}+1}{e^{-\Delta_n/\rho}-1}\right)^2\frac{\log n}{n} \to 0 \quad \text{as } n\to\infty. \tag{8}$$

We've shown (8) 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 zero whenever $k < 1/2$. Note this does not include fixed domain sampling.

4 Estimating Sampling Variability

Recall that maximizing the two taper approximation (5) corresponds to solving an unbiased estimating equation for $\boldsymbol{\theta}$. This suggests an estimator of sampling variability in the two taper estimator, based on the robust information criterion (Heyde, 1997). Let $\mathbf{U}(\mathbf{Z};\boldsymbol{\theta})$ be an unbiased estimating function for $\boldsymbol{\theta}$; that is, $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{U}(\mathbf{Z};\boldsymbol{\theta})] = \mathbf{0}$ for all possible values of $\boldsymbol{\theta}$. The robust information matrix corresponding to $\mathbf{U}$ is $\mathcal{E}(\mathbf{U}) = \mathrm{E}\left[\dot{\mathbf{U}}\right]' \mathrm{E}\left[\mathbf{U}\mathbf{U}'\right]^{-1} \mathrm{E}\left[\dot{\mathbf{U}}\right]$, where $\dot{\mathbf{U}}$ is the matrix of derivatives of the vector $\mathbf{U}$ with respect to $\boldsymbol{\theta}$ (Heyde, 1997). Under certain conditions, norming by the sample equivalent of $\mathcal{E}(\mathbf{U})^{-1}$ gives asymptotic normality of the estimator $\hat{\boldsymbol{\theta}}_n$ obtained by maximizing $\mathbf{U}(\mathbf{Z_n};\boldsymbol{\theta})$ (Heyde, 1997, Section 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 $\mathcal{E}(\mathbf{U})^{-1}$ can still 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 $\mathbf{U}_{2tapers}$ be the vector whose i^{th} entry is the partial derivative of $\ell_{2tapers}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_i$. The two matrices needed to calculate $\mathcal{E}(\mathbf{U}_{2tapers})$ have entries

$$\mathrm{E}\left[\dot{\mathbf{U}}_{2tapers}\right]_{i,j} = -\frac{1}{2}\mathrm{tr}\left\{\left[\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_i}\circ\mathbf{T}\right][\boldsymbol{\Sigma}\circ\mathbf{T}]^{-1}\left[\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_j}\circ\mathbf{T}\right][\boldsymbol{\Sigma}\circ\mathbf{T}]^{-1}\right\} \quad (9)$$

$$\mathrm{E}\left[\mathbf{U}_{2tapers}\mathbf{U}'_{2tapers}\right]_{ij} = \frac{1}{2}\mathrm{tr}\left\{\left[\left([\boldsymbol{\Sigma}\circ\mathbf{T}]^{-1}\left[\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_i}\circ\mathbf{T}\right][\boldsymbol{\Sigma}\circ\mathbf{T}]^{-1}\right)\circ\mathbf{T}\right]\boldsymbol{\Sigma}\right. \quad (10)$$
$$\left.\left[\left([\boldsymbol{\Sigma}\circ\mathbf{T}]^{-1}\left[\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_j}\circ\mathbf{T}\right][\boldsymbol{\Sigma}\circ\mathbf{T}]^{-1}\right)\circ\mathbf{T}\right]\boldsymbol{\Sigma}\right\}$$

Note that the derivatives in (9) and (10) are matrix quantities depending on the sampling locations, so in general the entries of $\mathcal{E}(\mathbf{U}_{2tapers})^{-1}$ do not have closed form expressions. However, it is computationally straightforward to calculate both (9) and (10), as all inverses involve sparse matrices. To construct variance estimates for $\hat{\boldsymbol{\theta}}_{2tapers}$, one can plug the estimator into (9) and (10), calculate $\mathcal{E}(\mathbf{U}_{2tapers})$, then take the diagonal elements of the inverse. In the next section, we show via simulation that this procedure gives reasonable 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 to that of the MLE? Second, how should one choose the taper range γ? Finally, are the variance estimators proposed in Section 4 good estimators of sampling variability?

We simulated 1000 datasets, each consisting of a multivariate normal vector of length 300. Each dataset was generated using the same 300 locations, consisting of a random selection of perturbed gridpoints. To obtain the perturbed gridpoints, we first generated a two dimensional grid over $[0,1]^2$ with increments of 0.03. To each gridpoint, we added a random amount of noise in each coordinate, uniformly

distributed on $[-0.01, 0.01]$. Therefore, each perturbed gridpoint is at least 0.01 units distant from any its neighbors. This avoids numerical singularities due to the sampling locations being too close together (Stein et al., 2004).

The covariance function was exponential with $\sigma^2 = 1$ and $\rho = 0.2$. With this choice, pairs of observations have negligible (<0.05) correlation when their locations are more than 0.6 units 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 dataset 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 $\ell_{2tapers}$ 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 most commonly used in practice. Indeed, the simulation study in Zhang (2004) used joint maximization, even though the asymptotic results concerned a fixed ρ. Kaufman (2006) showed that fixing ρ at a value 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 two different values of the taper range, $\gamma = 0.6$ and $\gamma = 0.2$. 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. In this example, when $\gamma = 0.6$, 37% of off-diagonal elements in $\mathbf{\Sigma} \circ \mathbf{T}$ are zero; when $\gamma = 0.2$, the number climbs to 89%.

Figure 1 shows boxplots of the estimates. As γ decreases, the bias in the one taper estimates increases. In contrast, we see negligible bias and only a small increase in the variance of the two taper estimates. While 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 rough estimate of which can be obtained by eye or by subsampling the data), the two taper approximation is more accurate for highly correlated processes.

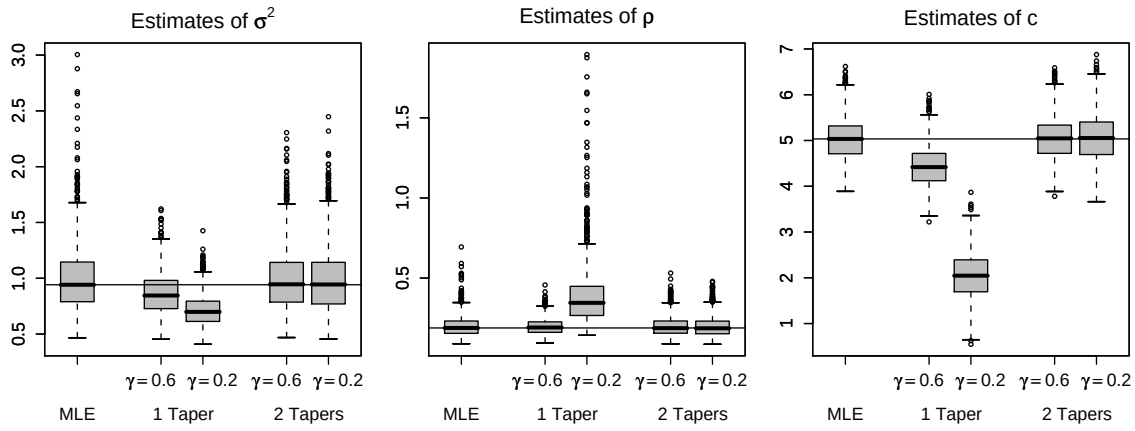

Figure 1: Boxplots of sampled estimates in the simulation study. Horizontal lines indicate the median of the distribution for the MLE in each case.

We have shown the tapering estimators can be comparable to the MLE when the covariance is exponential. In an extended version of this simulation study, Kaufman (2006) showed this result holds across a variety of Matérn covariance functions. We have also 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.

Because the true sampling variances are well-approximated with a sample of 1000, we can use the empirical variances of the estimates in the simulation to assess the accuracy of the information based variance estimators described in Section 4. For each iteration of the simulation study, we calculated variance estimates for the MLE based on the Fisher information matrix, and we 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 to the simulated variances. Results are shown for the more severe taper range of $\gamma = 0.2$. 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 are clearly less appropriate for the skewed distributions of the estimators of σ^2 and ρ.

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	0.131	0.148	0.006	0.007	0.202	0.275
Simulated variances	0.100	0.094	0.005	0.004	0.210	0.271
Ratio	1.309	1.571	1.300	1.558	0.960	1.014

6 Data Example

Sizable computational gains can be achieved when applying the tapering methods to large datasets. An example of a large, irregularly spaced spatial dataset is the collection of observations from weather stations in the United States. We consider precipitation data from the National Climatic Data Center (NCDC) for the years 1895 to 1997. This dataset 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 chose to illustrate the tapering methods using the precipitation anomalies from 1962, because this year had one of the most complete data records, with 7352 stations. In addition, it showed no obvious nonstationarity or anisotropy, which would require a more careful choice of tapering function. To calculate the anomalies, we included the full dataset computed by Johns et al. (2003). However, our analysis considers the anomalies from only those stations with a complete observational record. The data and computer code needed to carry out the analysis in the example are available at

http://www.image.ucar.edu/Data/Taper/.

We fit a Gaussian process model to the data, with exponential covariance function. In this case, with 7352 observations, evaluating the likelihood is quite slow, although still possible. We found the maximum likelihood estimates for σ^2 and ρ, as well as the corresponding tapering estimators. When tapering, we used the Wendland function $\phi_{2,1}$, as described in Section 3.1, with a taper range of 50 miles. The resulting matrices are quite sparse, with only 0.33% non-zero off-diagonal entries. For sparse matrix calculations, we used the `spam` package in `R`, available at http://cran.r-project.org/src/contrib/PACKAGES.html.

For any value of ρ, the maximizing value of σ^2 is available in closed form. Therefore, one can minimize the profile versions of the log likelihood and tapering approximations, which are functions only of ρ. These are shown computed over a grid in Figure 2. 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 is not surprising that the one taper estimate is further from the MLE than is the two taper estimate.

We computed variance estimates using the Fisher information matrix and the robust information matrix for the two-taper approximation, as described in Section 4, and we used these to form approximate 95% confidence intervals. For ρ, the MLE was 40.96, with confidence interval (37.15, 44.78); the two taper estimate was 37.60, with confidence interval (33.91, 41.30). These are also indicated in Figure 2. For σ^2, the MLE was 0.723, with confidence interval (0.663, 0.783); the two taper estimate was 0.786, with confidence interval (0.721, 0.851). Note that both sets of confidence intervals overlap.

We now compare the computation time required for each method. All calculations were carried out on a 3.2 GHz dual processor compute node with 4 GB of memory.

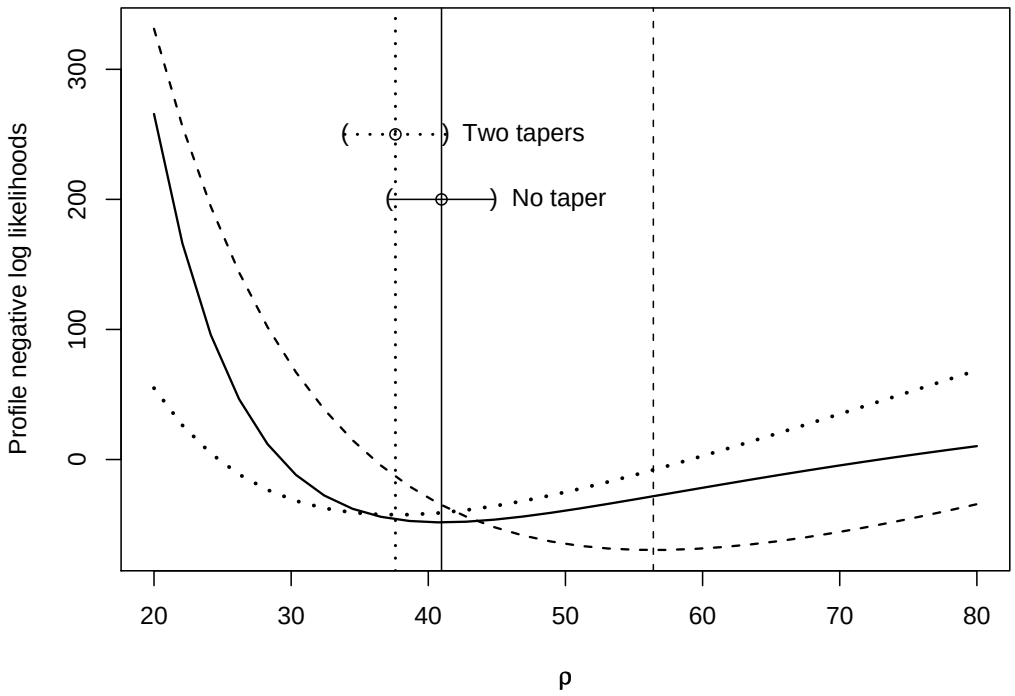

Figure 2: Curves represent the profile negative log likelihood (—), one taper approximation (- - -), and two taper approximation ($\cdots$). The mean of each function has been subtracted to allow comparison of curvature. The corresponding vertical lines indicate the minimum of each function. The MLE and two taper estimate both have information-based confidence intervals, and these are indicated by the horizontal brackets.

Each method requires first calculating the distance matrix, which for this dataset took about 40 seconds. In addition, the tapering methods require pre-computing the taper matrix, which took about 20 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 ten repetitions of each function evaluation. There is some initial savings in calculating $\mathbf{\Gamma} \circ \mathbf{T}$ rather than $\mathbf{\Gamma}$, because one needs only to compute the correlation for those distances less than the taper range. However, the most sizable savings come in calculating the Cholesky decomposition, which is

three orders of magnitude faster for the tapered correlation matrix than it is 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 42 seconds.

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

	No taper	One taper	Two tapers
$\mathbf{\Gamma}$ or $\mathbf{\Gamma} \circ \mathbf{T}$	3.35	0.05	0.05
Cholesky decomposition	578.32	0.70	0.70
Log determinant	0.27	0.00	0.00
Backsolve	1.08	0.02	—
Full solve	—	—	41.96
Second taper	—	—	0.02
Quadratic form	0.00	0.00	0.13
Total	583.02	0.77	42.86

7 Discussion

We have proposed two approximations to the likelihood for use in large spatial datasets. The one taper approximation (3) replaces the model covariance matrix by a tapered version, while the two taper approximation (5) tapers both the model and sample covariance matrices. 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 covariance. 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. One 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 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, although both tend to overestimate the variance in the two taper estimators of σ^2 and ρ, at least for the small sample size we examined in the simulation study.

The one taper estimators displayed sizable 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 paper, but when the dataset is large enough to warrant tapering in the estimation of model parameters, it is also typically large enough to warrant it for interpolation. Some 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 we have developed may be extended in several ways.

First consider the case that the mean is not zero but is a linear function, so that $\mathbf{Z} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\mathbf{X}$ is an $n \times p$ fixed regression matrix and $\boldsymbol{\beta}$ is a vector of p unknown coefficients. Then one may approximate the log-likelihood by

$$\begin{aligned} \ell_{2tapers}(\boldsymbol{\theta}, \boldsymbol{\beta}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| \\ &\quad -\frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})' \left([\boldsymbol{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1} \circ \mathbf{T}(\gamma)\right)(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

This choice gives unbiased estimating equations in both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Another extension of these ideas would be to non-isotropic or non-stationary covariance functions. One possible approach to this problem would be to consider the class of random fields which are stationary and isotropic subject to some transformation of the underlying space on which the process is defined, as in 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 Schur Product

This section collects some relevant results on the Schur product. The interested reader should refer to Horn and Johnson (1991, Chapter 5) for more details.

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

2. **The Schur Product Theorem** If $\mathbf{A}$ and $\mathbf{B}$ are positive semidefinite $n \times n$ matrices, 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 $\mathbf{A} \circ \mathbf{B}$.

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

4. **Eigenvalue Inequalities** If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ positive semi-definite matrices, then any eigenvalue $\lambda(\mathbf{A} \circ \mathbf{B})$ of $\mathbf{A} \circ \mathbf{B}$ satisfies

$$\left[\min_{1 \le i \le n} a_{ii}\right] \lambda_{min}(\mathbf{B}) \le \lambda(\mathbf{A} \circ \mathbf{B}) \le \left[\max_{1 \le i \le n} a_{ii}\right] \lambda_{max}(\mathbf{B}), \tag{11}$$

   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 $\mathbf{B}$.

5. **Trace** For square matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, with $\mathbf{B}$ symmetric, $\mathrm{tr}\left\{(\mathbf{A} \circ \mathbf{B})\mathbf{C}\right\} = \mathrm{tr}\left\{\mathbf{A}(\mathbf{B} \circ \mathbf{C})\right\}$.

Appendix B: Proofs

Proof of Theorem 1

Let f_1 be the spectral density corresponding to K_1. The Fourier transform of the product of two functions is the convolution of their Fourier transforms, so we may write $f_1(\boldsymbol{\omega}) = \int_{\Re^d} f_0(\mathbf{x}) f_{taper}(\boldsymbol{\omega} - \mathbf{x})\, d\mathbf{x}$, where $f_0(\mathbf{x}) = \sigma^2 M_0(\rho^{-2} + ||\boldsymbol{\omega}||^2)^{-\nu - d/2}$ is the spectral density corresponding to the Matérn covariance function K_0.

Stein (2004, Theorem A.1) provides 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

$$\int_{||\boldsymbol{\omega}|| > c} \left\{ \frac{f_1(\boldsymbol{\omega}) - f_0(\boldsymbol{\omega})}{f_0(\boldsymbol{\omega})} \right\}^2 d\boldsymbol{\omega} < \infty. \tag{12}$$

The Matérn spectral density f_0 satisfies the first condition when $\eta = 2\nu + d$. Rewriting the integral in (12) 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 (12) holds, it is therefore sufficient to show that

$$\left| \frac{f_1(r\mathbf{u})}{f_0(r\mathbf{u})} - 1 \right| = O(r^{-\xi}), \text{ for some } \xi > d/2 \tag{13}$$

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

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

$$\left| \frac{f_1(r\mathbf{u})}{f_0(r\mathbf{u})} - 1 \right| \leq \left| \frac{\int_{N_r^c} f_0(\mathbf{x}) f_{taper}(r\mathbf{u} - \mathbf{x}) d\mathbf{x}}{f_0(r\mathbf{u})} \right| + \tag{14}$$

$$\left| \frac{\int_{N_r} f_0(\mathbf{x}) f_{taper}(r\mathbf{u} - \mathbf{x}) d\mathbf{x}}{f_0(r\mathbf{u})} - 1 \right|. \tag{15}$$

Because $d \leq 3$ and $\epsilon > d/4$, we may choose $\xi \in (d/2, \min\{2, 2\epsilon\})$. Then choose $k \in ((d + 2\nu + \xi)/(d + 2\nu + 2\epsilon), 1)$. The remainder of the proof will show that with this choice of k, both (14) and (15) are $O(r^{-\xi})$, so (13) holds.

First consider (14). When $\mathbf{x} \in N_r^c, ||r\mathbf{u} - \mathbf{x}|| > r^k$, so using the bound in the theorem, we have

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

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

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

and (14) is $O(r^{-\xi})$ because we chose $k > \frac{d+2\nu+\xi}{d+2\nu+2\epsilon}$.

Now consider (15). Expanding $f_0(\mathbf{x})$ about $r\mathbf{u}$, we have

$$\left|\frac{\int_{N_r} f_0(\mathbf{x}) f_{taper}(r\mathbf{u}-\mathbf{x})d\mathbf{x}}{f_0(r\mathbf{u})} - 1\right| \leq$$

$$\left|1 - \int_{N_r} f_{taper}(r\mathbf{u}-\mathbf{x})d\mathbf{x}\right| + \tag{16}$$

$$\left|\frac{1}{f_0(r\mathbf{u})}\int_{N_r} (\mathbf{x}-r\mathbf{u})'[\nabla f_0(r\mathbf{u})] f_{taper}(r\mathbf{u}-\mathbf{x})d\mathbf{x}\right| + \tag{17}$$

$$\left|\frac{1}{2f_0(r\mathbf{u})}\int_{N_r} (\mathbf{x}-r\mathbf{u})'[\nabla^2 f_0(\mathbf{m}_{\mathbf{x},\mathbf{r}})](\mathbf{x}-r\mathbf{u}) f_{taper}(r\mathbf{u}-\mathbf{x})d\mathbf{x}\right| \tag{18}$$

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

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

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

Therefore, (16) is $O(r^{-\xi})$ because $\xi < 2\epsilon$ and $\nu > 0$ imply $k > \frac{\xi}{2(\nu+\epsilon)}$.

The integral in (17) is equal to zero 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 (18), first note that for each $r > 0$ and $\mathbf{x} \in N_r$,

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

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

$$\nabla^2 f_0(\mathbf{m}) = \frac{1}{||\mathbf{m}||^2}\left[g''(||\mathbf{m}||) - \frac{g'(||\mathbf{m}||)}{||\mathbf{m}||}\right]\mathbf{m}\mathbf{m}' + \frac{g'(||\mathbf{m}||)}{||\mathbf{m}||}\mathbf{I_d}, \tag{19}$$

where $g(r) = \sigma^2 M_0(\rho^{-2} + r^2)^{-(\nu+d/2)}$. The two matrices in (19) are symmetric, so the maximum eigenvalue of their sum is less than or equal to the sum of their maximum eigenvalues (Horn and Johnson, 1991, 3.4.11a). We have

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

This function is eventually decreasing with $||\mathbf{m}||$. Also, note $||\mathbf{m}_{\mathbf{x},\mathbf{r}}|| > r - r^k$, since $(r - r^k)\mathbf{u}$ is the point on the boundary of N_r which is closest to the origin and $\mathbf{m}_{\mathbf{x},\mathbf{r}}$ was defined to be in N_r. Because $k < 1$, $r - r^k \to \infty$, and so eventually

$g''(||\mathbf{m}_{\mathbf{x},\mathbf{r}}||) \leq g''(r - r^k)$ for all $\mathbf{x} \in N_r$. Therefore, for sufficiently large r,

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

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

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

where $\mathbf{Y} \sim t_{d,2(\nu+\epsilon)}/\sqrt{2(\nu + \epsilon)}$. But $(\nu + \epsilon) > 1$, so this term is finite. Therefore, we only need to 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 $\hat{\sigma}^2_{n,1taper} \to \sigma^{2*}$ a.s. $[G(K_1^*)]$. Because ρ^* and ν are fixed, $\hat{\sigma}^2_{n,1taper} = \mathbf{Z_n} \left[\mathbf{\Gamma_n^*} \circ \mathbf{T_n}\right]^{-1} \mathbf{Z_n}/n$, where $\mathbf{\Gamma_n^*} = \left\{K_0^*(||\mathbf{s_i} - \mathbf{s_j}||; \sigma^{2*}, \rho^*, \nu)\right\}/\sigma^{2*}$ and $\mathbf{T} = \left\{K_{taper}(||\mathbf{s_i} - \mathbf{s_j}||)\right\}$. Under $G(K_1^*)$, $\mathbf{Z} \sim MVN(0, \sigma^{2*}\mathbf{\Gamma_n^*} \circ \mathbf{T_n})$, so $\hat{\sigma}^2_{n,1taper}$ is distributed as σ^{2*}/n times a χ^2 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 $\frac{1}{\sigma}\mathbf{R_n}^{-1}\mathbf{Z_n} \sim MVN(\mathbf{0}, \mathbf{I_n})$, so

$$\begin{aligned}
\hat{\sigma}^2_{n,2tapers} &= \mathbf{Z_n}'\left(\left[\mathbf{\Gamma_n} \circ \mathbf{T_n}\right]^{-1} \circ \mathbf{T_n}\right)\mathbf{Z_n}/n \\
&= \frac{1}{n}\mathbf{X_n}'\left[(\sigma\mathbf{R_n})'\left[\left(\mathbf{\Gamma_n} \circ \mathbf{T_n}\right)^{-1} \circ \mathbf{T_n}\right](\sigma\mathbf{R_n})\right]\mathbf{X_n}, \quad \text{where } \mathbf{X_n} \sim MVN(\mathbf{0}, \mathbf{I_n}) \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{n}\lambda_{n,i}\chi^2_i, \qquad (20)
\end{aligned}$$

where χ^2_i are $iid\,\chi^2_1$ random variables and λ_{n_i} is the i^{th} eigenvalue of $\mathbf{R_n}'\left[\left(\mathbf{\Gamma_n} \circ \mathbf{T_n}\right)^{-1} \circ \mathbf{T_n}\right]\mathbf{R_n}$, which is the same as the i^{th} eigenvalue of $\mathbf{W_n} = \left[\left(\mathbf{\Gamma_n} \circ \mathbf{T_n}\right)^{-1} \circ \mathbf{T_n}\right]\mathbf{\Gamma_n}$.

Cuzick (1995) gave conditions for the almost sure convergence of weighted sums of iid random variables. Specifically, let $Y_n = \sum_{i=1}^{n} a_{n,i}X_i$, where X_i are iid with mean zero and $\{a_{n,i}\}$ is an array of constants. Then if $\sup_n \left(n^{-1}\sum_{i=1}^{n}|a_{n,i}|^q\right)^{1/q} < \infty$ for some $1 < q \leq \infty$, and $E|X|^p < \infty, p^{-1} + q^{-1} = 1, Y_n/n \to 0$ almost surely. (The case $q = 0$ is interpreted to mean the a_n, i are uniformly bounded.) The result also holds when $q = 1$ under the additional assumption that $\limsup_{i \leq n}|a_{n,i}|n^{-1}\log n$. Finish the proof by applying these results to (20), with $X_i = \chi^2_i - 1$ and $a_{n,i} = \lambda_{n,i}$.

Proof of Corollary 1

By Theorem 2 of Zhang (2004), one may 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 $\hat{\sigma}^{2*}_{n,2tapers}/\rho^{*2\nu} \to \sigma_0^2/\rho_0^{2\nu} a.s.[G(K_0^*)]$. This follows directly from the conditions on $\mathbf{W_n}$ and Theorem 3.

Proof of Lemma 1

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

Here we have used (11) in the second and last lines.

References

Abromowitz, M. and Stegun, I., editors (1967). *Handbook of Mathematical Functions*. U.S. Government Printing Office.

Bickel, P. and Levina, E. (2007). Regularized estimation of large covariance matrices. *The Annals of Statistics*. To appear.

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

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. (2006). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*. To appear.

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*. Springer.

Horn, R. and Johnson, C. (1991). *Topics in matrix analysis*. 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*. PhD 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*. Springer-Verlag, second edition.

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

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

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

Stein, M. (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, Series B*, 66:275–296.

Vecchia, A. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B*, 50:297–312.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4:389–396.

Wendland, H. (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.