Covariance Tapering for Likelihood-based Estimation in Large Spatial Datasets

Ph.D Thesis Proposal

Cari Kaufman
Department of Statistics, Carnegie Mellon University

October 19, 2005

Abstract

Likelihood-based methods such as maximum likelihood, REML, and Bayesian methods are attractive approaches to estimating covariance parameters in spatial models based on Gaussian processes. Finding such estimates can be computationally infeasible for large datasets, however, requiring $O(n^3)$ calculations for each evaluation of the likelihood based on n observations. We propose 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 compactly supported correlation matrix. This produces matrices which can be be manipulated using more efficient sparse matrix algorithms. We present two approximations to the Gaussian likelihood using tapering. The first tapers the model covariance matrix only, whereas the second tapers both the model and sample covariance matrices. Tapering the model covariance matrix can be viewed as changing the underlying model to one in which the spatial covariance function is the direct product of the original covariance function and the tapering function. Focusing on the particular case of the Matérn class of covariance functions, we give conditions under which tapered and untapered covariance functions give equivalent (mutually absolutely continuous) measures for Gaussian processes on bounded domains. This allows us to evaluate the behavior of estimators maximizing our approximations to the likelihood under a bounded domain asymptotic framework. We give conditions under which estimators maximizing our approximations converge almost surely and quantify their efficiency using using the robust information criterion of Heyde (1997). We present results from a simulation study showing concordance between our asymptotic results and what we observe for moderate but increasing sample sizes. Finally, we discuss a potential application of these methods to a large spatial estimation problem, that of making statistical inference about the climatological (long-run mean) temperature difference between two sets of output from a computer model of global climate, run under two different land use scenarios.

1 Introduction

Much recent work has focused on the problem of estimating the autocovariance functions of spatially correlated stochastic processes. Researchers in geology, hydrology, agriculture, epidemiology, and the environmental and atmospheric sciences require models which can capture correlations among observations due to spatial location. For example, early interest in this problem arose when researchers realized that spatial dependence in agricultural field trials could bias their results (see eg. Whittle, 1954). A more recent example of the importance of spatial modeling is the prediction of air pollution levels at locations which are not regularly monitored, in order to judge compliance with environmental regulations (Holland et al., 2003). In this case, the quality of prediction depends heavily on the model for the relationship between pollution at monitored and unmonitored locations.

Traditional approaches to estimating spatial covariance functions, such as fitting a parametric function to the empirical variogram using weighted least squares, make no explicit distributional assumptions (see Cressie, 1993, chap 2). Likelihood-based methods of estimating the spatial covariance function, such as maximum likelihood, restricted maximum likelihood (REML), and Bayesian estimation, have been posed as alternatives (Cook and Pocock, 1983; Kitanidis, 1983; Mardia and Marshall, 1984; Kitanidis, 1986; Handcock and Stein, 1993). In particular, a widely adopted model assumes the existence of an underlying stochastic process $Z = \{Z(s), s \in S \subset \Re^d\}$, where Z is stationary and Gaussian, with specified mean function $\mathrm{E}[Z(s)] = m(s;\beta)$ and isotropic covariance function

$$Cov(Z(s), Z(s')) = K(||s - s'||;\theta).$$

The functions m and K depend on the unknown parameters $\beta \in \Re^p$ and $\theta \in \Re^q$, which are to be estimated based on a finite number of observations $Z_n = (Z(s_1), \ldots, Z(s_n))'$ at locations in $S_n = \{s_1, \ldots, s_n \in S\}$.

It is often the case in recent data-rich applications that n is large. As will be clear from what follows, this creates difficulties primarily in estimating the covariance parameters θ, so we suppose for simplicity that the mean of Z is known to be zero. (We return to the non-zero mean case and methods for joint estimation of β and θ in Section 6.) Then the log-likelihood function for θ based on Z_n is

$$l_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_n(\theta)| - \frac{1}{2}Z_n'\Sigma_n(\theta)^{-1}Z_n, \tag{1}$$

where $\Sigma_n(\theta) = \{K(||s_i - s_j||;\theta)\}$.

Maximizing (1) to find the MLE for θ was first advocated by Cook and Pocock (1983), Kitanidis (1983), and Mardia and Marshall (1984). Controversy arose over the fact that (1) can be multimodal for certain covariance functions, implying that numerical maximization algorithms can easily converge to a local and not the global maximum (Warnes and Ripley, 1987). However, this does not appear to occur for the class of covariance functions we consider (Mardia and Watkins, 1989; Stein, 1999).

Bayesian methods have gained popularity as well, primarily in applications requiring predictive distributions for the process at locations that have not been observed. The predictor of $Z(s^*), s^* \notin S_n$, equal to the conditional mean $\mathrm{E}\left[Z(s^*)|Z(s), s \in S_n\right]$ is called the kriging predictor, after D.G. Krige, who advocated its use in early mining applications (Krige, 1951). When the true covariance function is known, the kriging predictor is the best linear unbiased predictor (BLUP) of $Z(s^*)$. In practice, however, the true covariance function (or θ in the model above) must be estimated, and it is common to plug an estimate into the kriging equations and proceed as if the true covariance were known. However, inference based on this procedure ignores the uncertainty involved in estimating θ. For this reason, many authors have taken a Bayesian approach, deriving a joint posterior distribution for both θ and $Z(s^*)$ (Kitanidis, 1986; Handcock and Stein, 1993; Handcock and Wallis, 1994).

Whether frequentist or Bayesian in nature, likelihood-based covariance estimation has been widely adopted in both the statistics and earth sciences communities. Its benefits include ease of incorporation into larger models, the use of model comparison criteria such as AIC and BIC, and various optimality properties. Although the assumption of normality may seem restrictive, one can often justify its use in spatial applications by Central Limit Theorem arguments, as observations often consist of averages, or by the fact that the normal distribution possesses the maximum entropy property (Pardo-Igúzquiza, 1998). In return, one inherits the many useful analytical results that have been developed for Gaussian processes (see Abrahamsen, 1997, for a review). One notable extension of this approach to non-Gaussian distributions modifies the generalized linear model framework to allow the linear predictor to include an additive error term with spatially correlated Gaussian process distribution (Diggle et al., 1998).

When the number of observations is large, however, the computational burden of evaluating the log-likelihood (1) can make likelihood-based estimation of θ computationally infeasible. Finding the determinant and inverse of $\Sigma_n(\theta)$ each require $O(n^3)$ operations. Moreover, if numerical methods such as numerical maximization or Markov chain Monte Carlo are required, estimation will involve repeated evaluations of the likelihood. Techniques for overcoming this computational hurdle have been developed mainly for datasets in which the sampling locations form a regular lattice, in which case spectral methods can be used (Whittle, 1954; Guyon, 1982; Stein, 1995; Dahlhaus, 2000). However, the computational advantages of working in the spectral domain do not directly apply when the data are irregularly spaced or even when some observations from the lattice are missing. For analyzing large irregularly spaced spatial datasets, therefore, it is desirable to have a method which will reduce computational expense, while also producing results comparable to those that would have been given by exact likelihood-based techniques. To address this need, we propose using the method of covariance tapering, in which covariance matrices are multiplied element-wise by a compactly supported correlation matrix, giving matrices which can be be manipulated using more efficient sparse matrix algorithms. We give two approximations to the Gaussian log-likelihood using tapering and evaluate the behavior of estimators maximizing these approximations, deriving results in a bounded domain asymptotic framework.

The computational difficulty of applying likelihood-based methods to large spatial problems was recognized by some of its earliest advocates (Mardia and Marshall, 1984; Vecchia, 1988). The next section describes existing methods of addressing this problem, several of which involve likelihood approximation. Section 3 describes the method of covariance tapering and proposes two different approximations to the log-likelihood function (1). Section 4 gives some theoretical results regarding the performance of estimators maximizing these approximations and presents results of a simulation study. Section 5 describes a potential application of our methods to the problem of making statistical inference about the climatological (long-run mean) temperature difference between two sets of output from a computer model of global climate, run under two different land use scenarios. In Section 6, we discuss work that remains to be done, including extensions of our theoretical results and analysis of the climate model data. We propose to fit a Bayesian hierarchical model for the long-run temperature differences, conditional on what we observe for a finite number of model iterations. The approximations we have developed will be used to facilitate fitting the model.

2 Current techniques for large spatial datasets

Under certain sampling schemes, evaluating the log-likelihood function (1) can be done efficiently without resorting to approximation. Zimmerman (1989) outlines several of these cases. For example, when the sampling locations S_n form a regular lattice of R rows and C columns, the covariance matrix is block Toeplitz, and matrix inversion algorithms exist which reduce the required number of computations to $O(R^2C^3)$, from $O(R^3C^3)$ for an arbitrary $RC \times RC$ matrix (Akaike, 1973). If the sampling locations form a regular rectangular lattice and the covariance function K is separable (that is, if K can be expressed as a product of functions in each coordinate), then the covariance matrix is a Kronecker product of two symmetric Toeplitz matrices, and the inverse can be found using $O(R^2C^2)$ computations (Zimmerman, 1989).

Another way in which calculations for lattice data can be simplified is via a spectral representation of the process Z. If Z is stationary, then it can be represented using the Fourier-Stieltjes integral

$$Z(s) = \int \exp\{is'\omega\} dY(\omega),$$

where Y are random functions with uncorrelated increments (see eg. Yaglom, 1987). Whittle (1954) proposed an approximation to the likelihood for lattice data based on Y rather than Z. Discussion of its construction is beyond the scope of this document; we mention only that it can be evaluated efficiently using the fast Fourier transform (Press et al., 1992). Fuentes (2004) extended this idea to construct an approximation for irregularly space data, dividing the spatial domain into a lattice of blocks, then working with the process obtained by integrating Z over each of the blocks.

The remaining approaches we discuss share the common theme of approximating the likelihood by imposing various conditional independence assumptions. Vecchia (1988) proposed one such approximation. The likelihood is first factored into a product of conditional densities:

$$L(\theta; z) = \prod_{i=1}^{n} p_\theta(z_i | z_j, 1 \le j \le i-1).$$

Then, the conditioning sets $\{z_j, 1 \le j \le i-1\}$ are replaced with smaller sets z_{im} consisting of the $\min(i,m)$ observations at those locations closest to that of z_i, giving

$$L(\theta; z) \approx L_m(\theta; z) = \prod_{i=1}^{n} p_\theta(z_i | z_{im}).$$

This approach maintains a multivariate normal distribution for z, with the covariance matrix remaining positive definite, as the product of valid conditional densities is a valid joint density. As m increases, L_m approaches the true likelihood L. Vecchia suggested maximizing $L_m(\theta; z)$ to obtain $\hat{\theta}_m$ for a sequence of increasing m, monitoring the behavior of $-2\log L_m(\hat{\theta}_m; z)$, and stopping when this criterion stabilizes. A likelihood approximation of this type was used by Eide et al. (2002) in fitting a hierarchical Bayesian model to predict the porosity of some offshore petroleum reservoirs. Stein et al. (2004) extended Vecchia's idea to apply to REML estimators and noted that because the derivative of the approximate log-likelihood forms an unbiased estimating equation for θ, the efficiency of the resulting estimators can be compared to those of the usual REML estimators using the robust information criterion described by Heyde (1997). Using this measure, the authors concluded that when defining the smaller conditioning sets z_{im}, it is helpful to include some distant observations, not just the nearest neighbors.

Caragea (2003) explored a related approach, which divides the sampling domain into subregions. The likelihood is then approximated using either 1) the likelihood for the means over each of the subregions, 2) the likelihood for the observations, assuming subregions are independent, or 3) the likelihood for the observations, assuming they are conditionally independent given the means of the subregions. Caragea explored the behavior of estimators maximizing these approximations for data forming a time series, where increasing the sample size corresponded to taking observations in an unbounded domain.

3 Covariance tapering

The intuition behind the subsetting approach of Vecchia (1988) is that correlations between pairs of distant locations often are nearly zero. If we have reason to believe that for a given data set these distant observations are truly independent, then we can model this using a compactly supported covariance function (Gaspari and Cohn, 1999; Gneiting, 2002). The covariance matrix Σ_n then contains zeroes corresponding to these distant pairs, and sparse matrix algorithms (see eg. Pissanetzky, 1984) can be used to manipulate it more efficiently. However, if we do not truly believe the underlying process possesses such a covariance function, we can still exploit this idea for computational purposes. The goal is to set to zero certain elements of Σ_n, such that the resulting matrix remains positive definite and retains some of the original properties of Σ_n. To this end, consider taking the direct product of the true covariance function $K_0(x;\theta)$ and a "tapering function" $K_{taper}(x;\gamma)$, an isotropic correlation function which is identically zero outside a range described by γ. Denote this function by

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

Then the covariance matrix $\{K_1(||s_i - s_j||; \theta, \gamma)\}$ for observations in S_n can be written as $\Sigma_n(\theta) \circ T_n(\gamma)$, where $\Sigma_n(\theta)$ is the original covariance matrix, $T_n(\gamma) = \{K_{taper}(||s_i - s_j||; \gamma)\}$, and the "$\circ$" notation refers to the element-wise product, also called the "Schur" or "Hadamard" product. Then $\Sigma_n(\theta) \circ T_n(\gamma)$ is positive definite (Horn and Johnson, 1991, Theorem 5.2.1). Note that we require K_{taper} to be a correlation function, with $K_{taper}(0; \gamma) = 1$. This ensures the marginal variance of Z is the same under K_0 and K_1. We give a stronger reason for this requirement in Theorem 1.

There are a variety of compactly supported correlation functions that can be used for tapering. A well known correlation function used in spatial statistics is the spherical correlation function

$$K_{taper}(x; \gamma) = (1 - x/\gamma)_+^2 \left(1 + x/(2\gamma)\right), \tag{3}$$

where $(y)_+ = yI_{\{y>0\}}$. The function K_{taper} is thus identically zero for $x \geq \gamma$. Tapering a covariance function and the associated covariance matrix are illustrated in Figures 1 and 2, using the spherical correlation function (3). Although the spherical correlation has been widely used in spatial statistics, it is not ideal for our application, as we demonstrate in Section 4.2. In that section, we return to the choice of an appropriate tapering function.

3.1 Likelihood approximation via covariance tapering

We propose two approximations to the log-likelihood (1) using covariance tapering. The most obvious approximation simply replaces the model covariance matrix $\Sigma_n(\theta)$ with $\Sigma_n(\theta) \circ T_n(\gamma)$, giving

$$l_{n,1taper}(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_n(\theta) \circ T_n(\gamma)| - \frac{1}{2} Z_n' \left[\Sigma_n(\theta) \circ T_n(\gamma)\right]^{-1} Z_n. \tag{4}$$

This is the expression for the log density of Z_n under the model that the underlying stochastic process Z is Gaussian with mean zero and covariance function (2). The effects of using a misspecified covariance function have been widely studied with respect to the performance of kriging predictions (see Section 4.3 of Stein, 1999, for a review), but the implications for estimation have not been as well studied.

One possible objection to approximation (4) is that taking its derivative with respect to the elements of θ and setting this equal to zero gives a biased estimating equation for θ (see Technical Appendix, Section A.1). To remedy the bias, first note that we can rewrite the quadratic form in (1) as a trace involving the empirical covariance matrix $\hat{\Sigma}_n = Z_n Z_n'$:

$$Z_n' \Sigma_n(\theta)^{-1} Z_n = \mathrm{tr}\left\{Z_n' \Sigma_n(\theta)^{-1} Z_n\right\} = \mathrm{tr}\left\{Z_n Z_n' \Sigma_n(\theta)^{-1}\right\} = \mathrm{tr}\left\{\hat{\Sigma}_n \Sigma_n(\theta)^{-1}\right\}.$$

This suggests replacing $\hat{\Sigma}_n$ with $\hat{\Sigma}_n \circ T_n(\gamma)$ as well, giving

$$\begin{aligned} l_{n,2tapers}(\theta) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_n(\theta) \circ T_n(\gamma)| - \frac{1}{2}\mathrm{tr}\left\{\left[\hat{\Sigma}_n \circ T_n(\gamma)\right]\left[\Sigma_n(\theta) \circ T_n(\gamma)\right]^{-1}\right\} \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_n(\theta) \circ T_n(\gamma)| - \frac{1}{2} Z_n' \left(\left[\Sigma_n(\theta) \circ T_n(\gamma)\right]^{-1} \circ T_n(\gamma)\right) Z_n, \end{aligned} \tag{5}$$

where the last equality follows from the fact that for square matrices A, B, and C, with B symmetric,

$$\mathrm{tr}\left\{(A \circ B)C\right\} = \sum_i \sum_j (A_{i,j} B_{i,j}) C_{j,i} = \sum_i \sum_j A_{i,j}(C_{j,i} B_{j,i}) = \mathrm{tr}\left\{A(C \circ B)\right\}.$$

Maximizing $l_{n,2tapers}(\theta)$ now corresponds to solving an unbiased estimating equation for θ (see Technical Appendix, Section A.1).

The function $l_{n,2taper}$ possesses an interesting property. It is equal to a constant plus the log of a normal density for Z_n with mean zero and covariance matrix $\left(\left[\Sigma_n(\theta) \circ T_n(\gamma)\right]^{-1} \circ T_n(\gamma)\right)^{-1}$. However, these densities for various n and Z_n do not correspond to any infinite dimensional distribution for the process Z on $\Re^d$. They fail to meet the consistency conditions of Kolmogorov's Existence Theorem (see Billingsley, 1986), because it is not true in general that for an invertible matrix A, the first m rows and columns of A^{-1} equal the inverse of the first m rows and columns of A. Therefore, for example, the covariance between $Z(s_i)$ and $Z(s_j), i, j \leq n$ under the density for the vector Z_n observed at locations in S_n is not necessarily the same as the covariance between the same $Z(s_i)$ and $Z(s_j)$ under the density for Z_{n+1} when S_{n+1} consists of S_n and only one additional point. This strange behavior makes $l_{n,2tapers}$ more difficult to analyze than $l_{n,1taper}$, but we prefer to use it for estimation because it appears to perform much better in practice, as we demonstrate in the simulation study of Section 4.4.

In this paper, we consider the MLEs $\hat{\theta}_n$ obtained by maximizing the log-likelihood (1) and the estimators $\hat{\theta}_{n,1taper}$ and $\hat{\theta}_{n,2tapers}$ obtained by maximizing approximations (4) and (5), respectively. We are also interested in the use of these approximations in Bayesian estimation, and this is in an important part of our application and future work, discussed in Section 6.

In most cases it is difficult to write an exact expression for the maximizer of these functions and so we resort to numerical methods. It is worth noting, however, that if K_0 can be written as

$$K_0(x; \theta) = \sigma^2 C_0(x; \phi) \tag{6}$$

for some correlation function C_0 and $\phi \in \Re^{q-1}$, then we can do the maximization with respect to $\theta = (\sigma^2, \phi)$ using profile versions of these expressions. That is, we can find the value of ϕ maximizing $\sup_{\sigma^2} l_n(\sigma^2, \phi)$, then calculate the corresponding value of σ^2 directly. See Technical Appendix, Section A.2 for details. Maximizing over one fewer parameters can improve the speed and convergence of any numerical maximization algorithms that are used. For general-purpose maximization, we use a quasi-Newton method called BFGS (see Nocedal and Wright, 1999), which is available as an option to the `optim()` function for numerical optimization in R.

4 Performance of estimators using tapering methods

In this section, we evaluate the performance of the estimators $\hat{\theta}_{n,1taper}$ and $\hat{\theta}_{n,2tapers}$ obtained by maximizing approximations (4) and (5), respectively. We first examine their asymptotic behavior as n goes to infinity. However, the type of asymptotic analysis required is different from that of the usual case, in which the observations are *iid*. When taking an increasing number of observations from a spatial process, we either can assume that the domain of the sampling region S increases to infinity, called "increasing domain" asymptotics, or we can assume that S is bounded and that observations become increasingly dense within S, called "bounded domain" or "infill" asymptotics (see Cressie, 1993, Section 5.8). The increasing domain asymptotic framework is the spatial analogue of the type of asymptotic analysis usually done in time series. It is typically easier to analyze, because for processes whose correlations decay with distance, taking observations in an increasingly large domain gives roughly independent observations of the same scenario. However, we believe the bounded domain asymptotic framework is more relevant, as it corresponds to the type of increased sampling that is actually feasible for most spatial problems.

The asymptotic behavior of the MLE and REML estimator for θ has been studied primarily in an increasing domain framework (Mardia and Marshall, 1984; Cressie and Lahiri, 1996; Watkins, 1990). The only results we are aware of for the MLE in a bounded domain framework pertain to the Matérn class of covariance functions. We also focus on this class, and so we review its properties and results for the MLEs under this model in the next section. These provide some idea of what results we might hope to hold for the tapering-based estimators as well.

4.1 Matérn covariance functions

The Matérn class of covariance functions (Matérn, 1986) is widely used in practice and has easily interpretable parameters. A Matérn covariance function with parameters σ^2, ρ, and ν is defined by

$$K(x;\sigma^2,\rho,\nu) = \frac{\sigma^2(x/\rho)^\nu}{\Gamma(\nu)2^{\nu-1}}\mathcal{K}_\nu(x/\rho),$$

where $\mathcal{K}_\nu$ is the modified Bessel function of order ν (see Abromowitz and Stegun, 1967, Section 9.6). This covariance function has the form of (6), with σ^2 representing the marginal variance of the process and the rest of the function its correlation structure. The range parameter ρ controls how quickly the correlation decays with distance, while the smoothness parameter ν controls the function's differentiability at the origin. In particular, K is $2m$ times differentiable at the origin, and the process Z is m times mean square differentiable, if and only if $\nu > m$ (Stein, 1999, Section 2.7). This flexibility in parameterizing the smoothness of the process by changing ν is the main reason this family has been advocated as a default covariance model for most spatial applications (Stein, 1999; Banerjee et al., 2004). When $\nu = 1/2$, the expression simplifies to the exponential covariance function

$$K(x;\sigma^2,\rho) = \sigma^2\exp\{-x/\rho\}, \tag{7}$$

which is not differentiable at the origin, whereas letting $\nu \to \infty$ gives the so-called Gaussian covariance function

$$K(x;\sigma^2,\rho) = \sigma^2\exp\{-x^2/\rho\}, \tag{8}$$

which has infinitely many derivatives at the origin.

Zhang (2004) proved several interesting results concerning the Matérn class, which are crucial in evaluating the performance of our estimators. The first concerns the equivalence of two mean-zero Gaussian measures $G(0,K_0)$ and $G(0,K_1)$. (Throughout, we use $G(m,K)$ to denote the measure for the process Z which is Gaussian with mean function m and covariance function K.) Recall that two probability measures P_0 and P_1 on the same measurable space $(\Omega,\mathcal{F})$ are said to be equivalent, or mutually absolutely continuous, if $P_0(A) = 0$ if and only if $P_1(A) = 0$, for all $A \in \mathcal{F}$. We denote this equivalence by $P_0 \equiv P_1$. Zhang (2004) proved that if 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 $G(0,K_0) \equiv G(0,K_1)$ on the paths of $\{Z(s), s \in S\}$ for any bounded infinite subset $S \in \Re^d$ with $d = 1,2,3$, if and only if $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$.

This has immediate consequences for estimation of σ^2 and ρ. Specifically, if the process $\{Z(s), s \in S \subset \Re^d\}$ with $d = 1,2$, or 3 is observed on a bounded sequence of subsets in S, then there cannot exist consistent estimators of σ^2 and ρ. This is easily seen, because under the true probability measure $G(0,K_0)$, a sequence of consistent estimators $\{(\hat\sigma_n^2,\hat\rho_n)\}_{n=1}^\infty$ contains a subsequence that converges almost surely to (σ_0^2,ρ_0). On the other hand, under $G(0,K_1)$, $\{(\hat\sigma_n^2,\hat\rho_n)\}_{n=1}^\infty$ contains a subsequence that converges almost surely to (σ_1^2,ρ_1). But if $\sigma_0^2/\rho_0^{2\nu} = \sigma_1^2/\rho_1^{2\nu}$, then $G(0,K_0) \equiv G(0,K_1)$, so the subsequence converges almost surely to (σ_1^2,ρ_1) under $G(0,K_0)$ as well. The fact that the sequence $\{(\hat\sigma_n^2,\hat\rho_n)\}_{n=1}^\infty$ contains two subsequences converging almost surely to two different values under $G(0,K_0)$ contradicts the fact that it is consistent.

Although the individual parameters σ^2 and ρ cannot be consistently estimated, Zhang (2004) showed the ratio $c = \sigma^2/\rho^{2\nu}$ can be consistently estimated. In particular, he showed that 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 $G(0,K_0)$. In practice, it is typical to estimate both σ^2 and ρ simultaneously by maximizing $L_n(\sigma^2,\rho)$, rather than fixing ρ^*. Indeed, in a simulation study illustrating his results, Zhang (2004) maximized over both parameters without reference to the fact that his theorem does not apply to this procedure. As part of our simulation study in Section 4.4, we show that the value of ρ^* chosen can severely affect the efficiency of the resulting estimators of c, and that maximizing over both parameters appears to work better than choosing a bad value for ρ^* and only slightly worse than fixing ρ^* at the true value. However, it is unclear whether Zhang's asymptotic result can be extended to the case of joint maximization.

The asymptotic distribution of the MLE was also derived in a very special case by Ying (1991). He considered the Gaussian process defined on the real line with mean zero and exponential covariance function (7), commonly known as the Ornstein-Uhlenbeck process. This is a special case of the processes considered by Zhang (2004), so again, there exist no consistent estimators of σ^2 and ρ. However, by exploiting the Markovian structure of the process, Ying showed the method described by Zhang, fixing ρ^* and maximizing $L_n(\sigma^2, \rho^*)$, gives an estimate of $c = \sigma^2/\rho$ which converges almost surely to the true value and is also asymptotically normally distributed. Additionally, Ying showed these results also hold when one fixes σ^2 and estimates ρ, or when one maximizes over both parameters simultaneously. However, the Markovian structure used in proving these results does not exist in general.

4.2 Equivalence under tapered and untapered covariance functions

Recall that in Section 3.1, we stated that approximation (4) corresponds to replacing the covariance function $K_0(x;\theta)$ with $K_1(x;\theta,\gamma) = K_0(x;\theta)K_{taper}(x;\gamma)$ in the model for the underlying stochastic process Z. If the tapering function K_{taper} is such that the measures $G(0,K_0) \equiv G(0,K_1)$ on the paths of Z, this can be used as a tool for showing almost sure convergence of estimators maximizing (4). That is, if we can show convergence occurs with probability one under $G(0,K_1)$, then we will have also shown it occurs with probability one under the true measure $G(0,K_0)$. In this section, we give conditions on K_{taper} under which equivalence between $G(0,K_0)$ and $G(0,K_1)$ holds for a process $Z \in \Re$. We conjecture that this result holds for higher dimensions as well, with a few modifications.

In the results that follow, we find it useful to characterize covariance functions in terms of their representation in the spectral domain. Briefly, the result known as Bochner's theorem (Bochner, 1955) states that a continuous function K in $\Re^d$ is positive definite if and only if it has spectral representation

$$K(x) = \int_{\Re^d} \exp\{i\omega' x\} F(d\omega),$$

where F is a positive bounded symmetric measure. If F has a density with respect to Lebesgue measure, it is called the spectral density and we denote it by f. For reference, the spectral density corresponding to the Matérn covariance $K(x;\sigma^2,\rho,\nu)$ is

$$f(\omega) = \sigma^2 \frac{\Gamma(\nu+1/2)}{\sqrt{\pi}\Gamma(\nu)\rho^{2\nu}} \frac{1}{(\rho^{-2} + ||\omega||^2)^{\nu+d/2}}. \tag{9}$$

Let f_0 be the spectral density of the true covariance function K_0 and f_1 be the spectral density of a "misspecified" covariance function K_1. A key result of Stein (1993) is that the simple condition $\lim_{\omega\to\infty} \frac{f_1(\omega)}{f_0(\omega)} = \gamma$ for some $0 < \gamma < \infty$ and a technical condition on f_0 are sufficient for asymptotically optimal prediction using K_1 instead of K_0. That is, if Stein's condition is satisfied, then the ratio of the mean squared error of the kriging predictor using $G(0,K_1)$ to the mean squared error using the true measure $G(0,K_0)$ goes to one with n. Furrer et al. (2005) used covariance tapering to decrease the computational burden of kriging, which involves inverting $\Sigma_n(\theta)$ for a fixed value of θ. They applied Stein's condition to the case that K_1 corresponds to a tapered version of K_0, assumed to have Matérn structure. They showed that if the spectral density f_{taper} corresponding to K_{taper} satisfies

$$f_{taper}(\omega) < \frac{M_\epsilon}{(1 + ||\omega||^2)^{\nu+d/2+\epsilon}} \tag{10}$$

for some $\epsilon > 0$ and $M_\epsilon < \infty$, then Stein's condition is satisfied for a process in $\Re^d$. We are interested in conditions on f_{taper} that give equivalence of measures $G(0,K_0)$ and $G(0,K_1)$, where K_0 and K_1 are untapered and tapered Matérn covariance functions, respectively. The following theorem gives some conditions extending (10) which force the tail behavior of f_0 and f_1 to satisfy an even stronger condition, which guarantees equivalence.

Theorem 1. *Let Z be a stationary random process on $\Re$ and let $G(0, K_0)$ and $G(0, K_1)$ represent two mean zero Gaussian measures for Z with covariance functions K_0 and K_1, respectively. Specifically, let K_0 be a Matérn covariance function with parameters σ^2, ρ, and ν. Let K_1 be the direct product of K_0 and an isotropic covariance function K_{taper} which is identically zero beyond a range described by γ:*

$$K_1(x; \sigma^2, \rho, \nu, \gamma) = K_0(x; \sigma^2, \rho, \nu) K_{taper}(x; \gamma).$$

Let f_{taper} be the spectral density corresponding to K_{taper}. Suppose there exist $\epsilon > 0$ and $M_\epsilon < \infty$ such that

1. $f_{taper}(\omega) \leq \frac{M_\epsilon}{(1+\omega^2)^{\nu+1/2+\epsilon}}$,
2. $\epsilon > \max\{1/4, 1-\nu\}$, *and*
3. $\int_\Re f_{taper}(\omega) d\omega = 1$.

Then $G(0, K_0)$ and $G(0, K_1)$ are equivalent on the paths of $\{Z(s), s \in S\}$, for any bounded subset $S \subset \Re$.

The proof of this theorem is given in Technical Appendix, Section A.3. Note that condition 3 means that K_{taper} must be a correlation function.

Two practical questions immediately arise regarding Theorem 1. First, what values of ν are plausible for a given dataset? Second, for a given range of plausible values of ν, what tapering functions satisfy the conditions of Theorem 1? With regard to the first problem, the consensus seems to be that for most spatial fields, the Gaussian covariance function (8) is too smooth (Banerjee et al., 2004), but the upper bound for plausible ν will depend on the particular dataset being analyzed. Based on our experience analyzing temperature fields, we see no difficulties arising from constraining ν to be less than three in this case. However, one can always choose a higher upper bound for ν to be more conservative.

With regard to the second question, we currently use the compactly supported radial basis functions constructed by Wendland (1995, 1998). These functions are generated using recursive transformations of the truncated power function $\phi_l(r) = (1-r)_+^l$. Specifically, for a given dimension d and nonnegative integer k, the function $\phi_{d,k}$ is defined as $I^k \phi_{\lfloor d/2 \rfloor + k + 1}$, where $(I\phi)(r) = \int_r^\infty t\phi(t)dt$. Then $\phi_{d,k}$ is positive definite on $\Re^d$ and has the form

$$\phi_{d,k}(r) = \begin{cases} p_{d,k}(r) & 0 \leq r \leq 1 \\ 0 & r > 1 \end{cases}$$

with $p_{d,k}$ a polynomial of degree $\lfloor d/2 \rfloor + 3k + 1$, and this function is of minimal degree among the class of polynomials having up to $2k$ continuous derivatives (Wendland, 1998). Note that the functions for $d = 2, 3$ are the same, and that $\phi_{d,k}$ remains positive definite in lower dimensions but is no longer of minimal degree. The first several functions for $d = 1, 2, 3$ are given in Table 1. To adapt $\phi_{d,k}$ to be a correlation function with support over $[0, \gamma)$, we find a positive constant $c_{d,k}$ such that $c_{d,k}\phi_{d,k}(0) = 1$, then form

$$K_{taper}(x, \gamma) = c_{d,k}\phi_{d,k}(x/\gamma). \tag{11}$$

Now, if $f_{d,k}$ is the spectral density corresponding to $\phi_{d,k}$, Wendland (1998, Theorem 2.1) showed that there exists a positive constant M such that $f_{d,k}(\omega) \leq M(1 + ||\omega||^2)^{-d/2-k-1/2}$. Therefore, K_{taper} as defined in (11) satisfies the conditions of Theorem 1 for all $\nu < \nu'$ whenever $k > \max\{1/2, \nu' - 1/4\}$. We conjecture that Theorem 1 holds for $d \geq 1$ if the first two conditions are replaced with 1. $f_{taper}(\omega) \leq \frac{M_\epsilon}{(1+\omega^2)^{\nu+d/2+\epsilon}}$, and 2. $\epsilon > \max\{d/4, 1-\nu\}$. In this case, K_{taper} will satisfy the conditions for all $\nu < \nu'$ whenever $k > \max\{1/2, \nu' - 1/2 + d/4\}$.

4.3 Almost sure convergence under the approximations

Now we return to the asymptotic behavior of estimators maximizing the approximations (4) and (5) under the Matérn model. First, we make use of Theorem 1 to prove a result similar to Zhang's (2004) result concerning the MLE, but instead using approximation (4). Like Zhang, we consider the model $G(0, K_0)$, with K_0 the Matérn covariance function with parameters σ_0^2, ρ_0, and known ν.

Recall that Zhang (2004) showed that fixing ρ^* and maximizing the likelihood $L_n(\sigma^2, \rho^*)$ gives an estimator $\hat{\sigma}_n^2$ satisfying $\hat{\sigma}_n^2/\rho^{*2\nu} \to c_0 = \sigma_0^2/\rho_0^{2\nu}$ almost surely under $G(0, K_0)$. The main tool in the proof of this result is the equivalence of $G(0, K_0)$ and $G(0, K_1)$, where K_1 is Matérn with parameters $\sigma^{2*} = \sigma_0^2(\rho^*/\rho_0)^{2\nu}, \rho^*$, and ν. Then, to show $\hat{\sigma}_n^2/\rho^{*2\nu} \to c_0$ almost surely under $G(0, K_0)$, the fact that $G(0, K_0) \equiv G(0, K_1)$ implies it is sufficient to show $\hat{\sigma}_n^2/\rho^{*2\nu} \to c_0$ almost surely under $G(0, K_1)$. In other words, one needs to show $\hat{\sigma}_n^2 \to \sigma^{2*}$ almost surely under $G(0, K_1)$. Luckily, this is straightfoward.

Additionally, Theorem 1 tells us there is a tapered Matérn covariance function K_2 such that $G(0, K_2) \equiv G(0, K_1)$. Therefore, to show that the estimator $\hat{\sigma}_{n,1taper}^2$ maximizing (4) for fixed ρ^* and known ν converges almost surely, it is sufficient to show that $\hat{\sigma}_{n,1taper}^2 \to \sigma^{2*}$ almost surely under $G(0, K_2)$, giving the following theorem. The details of the proof are in Technical Appendix, Section A.4. This theorem is limited to processes in one dimension by reliance on Theorem 1 in its proof; we note that extension of Theorem 1 to higher dimensions would immediately extend this theorem as well.

Theorem 2. *Let Z be a stationary, mean zero Gaussian process on $\Re$ with Matérn covariance function with parameters σ_0^2, ρ_0, and ν. Suppose that ν is known, but σ_0^2 and ρ_0 are unknown. Let $\{S_n\}_{n=1}^{\infty}$ be an increasing sequence of finite subsets of $\Re$ such that $\bigcup_{n=1}^{\infty} S_n$ is bounded and infinite. Let $l_{n,1taper}(\sigma^2, \rho)$ be the approximation to the log-likelihood given in (4) under this model, with the tapering function also satisfying the conditions of Theorem 1. For any fixed $\rho^* > 0$, let $\hat{\sigma}_{n,1taper}^2$ maximize $l_{n,1taper}(\sigma^2, \rho^*)$. Then $\hat{\sigma}_{n,1taper}^2/\rho^{*2\nu} \to \sigma_0^2/\rho_0^{2\nu}$ almost surely as $n \to \infty$.*

As noted in Section 3.1, the approximation defined in (5) does *not* correspond to altering the infinite-dimensional distribution for Z. Therefore, the equivalence result in Theorem 1 is not applicable in this case. Instead, we directly examine the expression for $\hat{\sigma}_{n,2tapers}^2$ obtained by fixing ρ^* and maximizing (5). This provides a condition, much harder to check, that $\hat{\sigma}_{n,2tapers}^2$ converges almost surely.

Theorem 3. *Let Z be a stationary, mean zero Gaussian process on $\Re^d, d = 1, 2, 3$, with Matérn covariance function $K_0(x; \sigma_0^2, \rho_0, \nu) = \sigma_0^2 C_0(x; \rho_0, \nu)$. Suppose that ν is known, but σ_0^2 and ρ_0 are unknown. Let $\{S_n\}_{n=1}^{\infty}$ be an increasing sequence of finite subsets of $\Re$ such that $\bigcup_{n=1}^{\infty} S_n$ is bounded and infinite. Let $l_{n,2tapers}(\sigma^2, \rho)$ be the approximation to the log-likelihood given in (5) under this model. Let $K_{taper}(x; \gamma)$ be a tapering function. Fix ρ^*, and for all n define the matrix*

$$W_n = \left[(\Gamma_n(\rho^*, \nu) \circ T_n(\gamma))^{-1} \circ T_n(\gamma)\right]^{-1},$$

where $\Gamma_n(\rho^*, \nu) = \{C_0(||s_i - s_j||; \rho^*, \nu)\}$ and $T_n(\gamma) = \{K_{taper}(||s_i - s_j||; \gamma)\}$. Denote by $\{\lambda_{n,i}\}_{i=1}^n$ the eigenvalues of $W_n^{-1}\Gamma_n(\rho^*, \nu)$. Suppose that either of the following two conditions holds:

1. $\sup_n \left(\frac{1}{n}\sum_{i=1}^n \lambda_{n,i}^q\right)^{1/q} < \infty$ *for some* $1 < q \leq \infty$, *or*
2. $\lim_n (\sup_{i \leq n} \lambda_{n,i}) n^{-1} \log n = 0$.

Then the estimator $\hat{\sigma}_{n,2tapers}^2$ maximizing $l_{n,2tapers}(\sigma^2, \rho^*)$ satisfies $\hat{\sigma}_{n,2tapers}^2/\rho^{*2\nu} \to \sigma_0^2/\rho_0^{2\nu}$ almost surely as $n \to \infty$.

The proof of this theorem is in the Technical Appendix, Section A.5. In Section 6, we discuss the role of the eigenvalues in proving this Theorem and how these conditions may be checked in practice.

4.4 Simulation study

To explore whether the convergence described in Theorems 2 and 3 agrees with what we observe for our estimators using finite but increasing sample sizes, and to compare our results to what has been observed in the literature for the Matérn covariance, we carried out a simulation study similar to that in Zhang (2004). For each of 1000 iterations, we simulated observations from a Gaussian process with mean zero and exponential covariance function (7), with $\sigma^2 = 1$ and $\rho = 0.2$. This function is shown in Figure 3 (a). Note that it has negligible correlation (< 0.05) for values of x greater than about 0.6. We generated observations at all 289 locations shown in Figure 3 (b). Then we estimated σ^2 and ρ by maximizing either the log-likelihood (1), the approximation (4), or the approximation (5). When using the tapering approximations, we used the function $c_{2,1}\phi_{2,1}(x/\gamma)$ described in Section 4.2, with either $\gamma = 0.6$ or $\gamma = 0.3$. The tapered versions of K_0 are also shown in Figure 3 (a). We carried out the estimation using three different sample sizes, with $n = 125, 221$, or 289. The sampling locations are shown in Figure 3 (b).

4.4.1 Joint estimation of σ^2 and ρ

Although Zhang (2004) showed for the exponential model only that $c = \sigma^2/\rho$ can be estimated consistently by fixing $\rho = \rho^*$ and maximizing the likelihood $L_n(\sigma^2, \rho^*)$ as a function of σ^2, his simulation maximized $L_n(\sigma^2, \rho)$ over both σ^2 and ρ, then examined the performance of the estimators $\hat{\sigma}^2$, $\hat{\rho}$, and $\hat{c} = \hat{\sigma}/\hat{\rho}$. We note that our results also apply only to estimators that fix ρ^*, but to compare with Zhang's results and because joint estimation of σ^2 and ρ is what is commonly done in practice, we start by estimating both. In the next section we carry out the same procedure with ρ^* fixed at various values.

The distributions of the estimators of σ^2 and ρ are shown in Figures 4 and 5, respectively. First note that the leftmost column in each plot duplicates Zhang's (2004) simulation result, in that the distributions of the MLEs of these individual parameters don't appear to be changing with n. Bias is evident in the distributions of the estimators maximizing $l_{n,1taper}$, which is greater for the more severe taper range of $\gamma = 0.3$. For the parameter ρ, this bias appears to be diminishing with n, although there is no evidence for this with σ^2. The most striking thing about these plots, however, is how similar the distributions of the estimators maximizing $l_{n,2taper}$ look to those of the MLEs.

We compare the distributions of $\hat{c} = \hat{\sigma}^2/\hat{\rho}$ in Figure 6. Again, the leftmost column duplicates Zhang's (2004) simulation results, in that the MLE of this ratio is becoming more concentrated about its true value as n increases. The same holds for the estimators maximizing $l_{n,2tapers}$, whose distributions appear very similar to those of the MLE. What bias exists in the distributions of the estimators maximizing $l_{n,1taper}$ appears to be decreasing with n, although it is unclear from this simulation study whether it would disappear entirely if n were large enough.

4.4.2 Estimating σ^2 for fixed values of ρ

Now we examine the behavior of estimators covered by Zhang's (2004) results and our Theorems 2 and 3. That is, we know that for any fixed ρ^*, maximizing either the log-likelihood $l_n(\sigma^2, \rho^*)$ or the approximations $l_{n,1taper}(\sigma^2, \rho^*)$ or $l_{n,2tapers}(\sigma^2, \rho^*)$ will provide an estimator of σ^2 such that the ratio $\hat{\sigma}^2/\rho^*$ converges almost surely to $c_0 = \sigma_0^2/\rho_0$. However, intuition suggests that the efficiency of these estimators for small samples could vary widely, depending on the particular value of ρ^* that is chosen. To explore this possibility, we carried out the same simulation as above (using the same 1000 simulated data sets) but estimated σ^2 with ρ^* fixed at either 0.1, the true value of 0.2, or 0.4. The results are shown in Figures 7 through 9, respectively.

We begin by comparing the results for ρ^* fixed at the true value of 0.2 (Figure 8) to those of the unconstrained MLE (Figure 6). The distributions for the MLE look relatively similar, although the distributions for fixed ρ^* appear slightly more concentrated around the true value. The bias in the estimators maximizing $l_{n,1taper}$ is also noticeably reduced when ρ^* is fixed at the true value. However, the efficiency of estimators maximizing $l_{n,2tapers}$ actually seems to decrease when ρ^* is fixed at the true value, compared to when it is estimated. It is not obvious why this should be so.

The effect of fixing ρ^* at something other than the true value of $\rho_0 = 0.2$ shows up as bias in the estimators which only gradually decreases with n. In Figure 7, ρ^* is fixed at half the true value. Estimators of $c = \sigma^2/\rho$ are correspondingly biased upward. For the estimators maximizing $l_{n,1taper}$, which are biased downward when ρ is not constrained, this actually appears to improve their performance, so that they appear preferable to estimators maximizing $l_{n,2taper}$. On the other hand, when ρ^* is fixed at twice the true value of ρ_0 (Figure 9), we see the opposite effect: the usual bias in $l_{n,1taper}$ is magnified by the additional effect of choosing too large a ρ^*. Again, the bias is decreasing with n for all the estimators, although quite slowly for $l_{n,1taper}$ under the more severe taper range.

4.5 Estimating uncertainty in the tapering-based estimators

Recall that in Section 3.1, we noted that maximizing approximation (4) corresponds to solving a biased estimating equation for θ, whereas maximizing approximation (5) corresponds to solving an unbiased estimating equation for θ. In this section, we explain a criterion which can be used in practice to compute a rough estimate for the sampling variability in $\hat{\theta}_{n,2tapers}$, based on the theory of unbiased estimating equations.

Consider a function $G(Z_n;\theta)$ with $\mathrm{E}_\theta\left[G(Z_n;\theta)\right] = 0$ for all possible values of θ. G is called an unbiased estimating function for θ; that is, we set $G(Z_n;\theta)$ equal to zero and solve for θ to obtain an estimate. Heyde (1997) defined the robust information matrix corresponding to G as

$$\mathcal{E}(G) = \mathrm{E}\left[\dot{G}\right]' \mathrm{E}\left[GG'\right]^{-1} \mathrm{E}\left[\dot{G}\right], \tag{12}$$

where $\dot{G}$ is the matrix of derivatives of the vector G. When G is the score function, $\mathcal{E}(G)$ is simply the Fisher information matrix. Among a class of estimating functions, a function G is said to be O_F-optimal if it maximizes $\mathcal{E}(G)$; that is, there is no other G^* for which the matrix $\mathcal{E}(G^*) - \mathcal{E}(G)$ is positive definite (Heyde, 1997). The score function is often O_F-optimal, and it is in our particular case. However, if the score function is not among the estimating equations under consideration (for instance, for computational reasons), then we might study the behavior of $\mathcal{E}(G)$ for those G in the class of estimating equations we are considering. For example, Stein et al. (2004) suggested comparing the diagonal elements of the inverse Fisher information matrix to the diagonal elements of the inverse of $\mathcal{E}(G)$, where G was based on his approximation to the spatial likelihood, described in Section 2.

Under certain conditions, norming by the sample equivalent of $\mathcal{E}(G)^{-1}$ gives asymptotic normality of the estimator $\hat{\theta}_n$ obtained by maximizing $G(Z_n;\theta)$ (Heyde, 1997, Section 2.5). For the estimators considered by Stein et al. (2004), as well as for our estimators, there is no proof of asymptotic normality under a fixed domain sampling scheme. Indeed, this cannot be the case for the individual parameters σ^2 and ρ when the covariance is Matérn with smoothness ν, because Zhang (2004) showed these are not even consistently estimable. To explore the appropriateness of the inverse elements of $\mathcal{E}(G)^{-1}$ as rough estimators of sampling variance of our estimators, we calculated $\mathcal{E}(G)$ for each of the sample sizes and estimation methods described in the simulation study of the previous section. We then compared the diagonal elements of the inverse of $\mathcal{E}(G)$ to the empirical variances of the estimates found in the simulation study. (The details of calculating $\mathcal{E}(G)$ for general matrices Σ_n and T_n are given in Technical Appendix, Section A.6.) The results can be seen in Tables 2 through 4. The theoretical and simulated values are roughly the same, although there are some discrepancies. In practice, one might consider using the values of $\mathcal{E}(G)^{-1}$ (evaluated at the maximized value of θ) to give a rough idea of the sampling variability of the estimators maximizing approximation (5). Based on this comparison to simulated variances, we recommend doing the calculation, but treating the interpretation with some skepticism. We note that one fortunate aspect of this calculation is that the $n \times n$ matrix inversions involved in calculating $\mathcal{E}(G)$ are all for matrices that have been tapered to the same degree as in the estimation process (see Technical Appendix, Section A.6). Therefore, this step should be computationally feasible. We also note that the estimation of uncertainty regarding these estimators is somewhat more naturally posed in a Bayesian framework, which we return to in Section 6.

5 Climate modeling application

There are an increasing number of spatial problems in which covariance estimation plays a role and yet the size of the data overwhelms our computational ability to do standard likelihood estimation, making covariance tapering a useful tool. We focus on the problem of fitting a model to global temperature data, generated by a computer model of climate run under two different land use scenarios. Such computer models numerically solve a system of differential equations describing the evolving behavior of the atmosphere, taking into account such things as radiation, physical dynamics, and surface-atmosphere energy interactions (McGuffie and Henderson-Sellers, 2005). To the extent that the model accurately reflects the true evolution of the climate system, it can be used to carry out "experiments" that would be impossible in reality, for example, determining the effect of doubling the current level of carbon dioxide in the atmosphere. As such, climate models are primary tools for predicting future climate under various scenarios, and they are also useful for studying in what ways observed climate change might be due to a particular type of human behavior.

Statistical analysis of output from climate models is common. For instance, climate modelers frequently carry out tests of mean difference between models run under different input values, performing, for instance, t-tests at each spatial location, sometimes adjusting for multiple testing and correlation between observations using bootstrap-like techniques (Livezey and Chen, 1983; Feddema et al., 2005). However, the fact that the output of most climate models is the result of a deterministic algorithm means that interpretations of certain statistical methods, such as hypothesis testing, are no longer defensible from a frequentist view of probability. That is, there is no chance mechanism generating the output, as in the usual frequentist thought experiment involving a sequence of repeated trials, each resulting in some random outcome (Berk et al., 2001). Rather, if we were to run the climate model repeatedly with the same input values, we would get the same output values each time. On the other hand, the differential equations being solved by the model are nonlinear, so the behavior of the model output is chaotic, behaving as we would expect for a process which was truly random.

There are a number of ways of addressing this problem, most of which take a subjectivist Bayesian view of probability. That is, although quantities associated with the model may not be intrinsically random, we may describe our uncertainty regarding them using the language of prior and posterior probability distributions. For instance, suppose we have uncertainty about the inputs themselves. The technique called "Bayesian melding" (Poole and Raftery, 2000) is designed to combine prior distributions on model inputs and outputs which respect the fact that the outputs are a function of the inputs. We may also have uncertainty about what the model output would be for fixed but unobserved values of the inputs. This is especially relevant if the model is expensive to run, and one would like to find a statistical model to serve as a surrogate predictor (Sacks et al., 1989; O'Hagan, 2004). On the other hand, one may be interested in the degree to which the model reflects reality, in which case one can statistically model both model output and observations in relation to some true underlying state (Fuentes and Raftery, 2005).

However, the case we consider is different from all of these. Namely, we want to make inference about the long-run or "climatological" behavior of the model output, conditional on fixed input values. That is, when climate modelers ask which differences between two sets of model output are "significant," we believe they are referring to their uncertainty about the long-run behavior of the model itself; how the model relates to reality is a separate issue. There are certain aspects of the model which are of interest, but which can only be known with certainty if the model is run forever and at all spatial locations. That is, for fixed input values x, consider the model output as a function $Z_x(s,t)$ of both space and time. Then, let $Z_x(s) = \lim_{t\to\infty} \frac{1}{T}\sum_{t=1}^{T} Z_x(s,t)$, assuming such a limit exists. We observe only $Z_{obs} = \{(Z_x(s,t), Z_{x'}(s,t)); s \in S, t \in T\}$, where S and T are finite sets and x and x' represent two sets of input values. Then we can address, using the posterior distribution for $Z_x(s)$ and $Z_{x'}(s)$ given Z_{obs}, such questions as

- What is the posterior mean surface of $Z_x(s) - Z_{x'}(s)$? For example, x and x' may represent a change in carbon dioxide emissions and Z a temperature field at a specified height.

- For what locations s is the posterior probability that $Z_x(s) - Z_{x'}(s) > T$ greater than 95%? If Z corresponds to temperature, a benchmark value for T in climate change studies is often $1°C$.
- If $f_{x,s}(y)$ is a (perhaps random) function giving a particular impact of $Z_x(s)$ taking value y at location s, what is the posterior expected value of $\int_{S^*}[f_{x,s}(Z_x(s)) - f_{x',s}(Z_{x'}(s))]ds$, where S^* is a region of interest? For instance, f might be a measure of negative health outcomes.

Among the possible anthropogenic causes of climate change, increased emission of greenhouse gases is the most widely studied. However, recent climate models also take into account the effect of surface vegetation, which can be altered due to farming or deforestation (see e.g. Bonan, 1998). The data we consider comes from Feddema et al. (2005) and consists of temperature output from the DOE-PCM climate model (Washington et al., 2000) coupled with the NCAR land surface model (Bonan, 1996). The model was run with preindustrial atmospheric conditions under two different land-use scenarios, simulated by the IMAGE 2.2 model (Alcamo, 1994; Alcamo et al., 1998). One of these scenarios corresponds to modern day land cover, while the other corresponds to an estimate of what modern day land cover would be, had there been no human interference. These are shown in Figure 10. The model was run for 100 years in each case, and the last 40 years are used for analysis. The data consist of temperature output from the model at each of 8192 locations, corresponding to a grid of 128 longitudes and 64 latitudes. The output consists of yearly averages over the winter months (December, January, and February), yearly averages over the summer months (June, July, and August), and yearly averages over the entire year. Figure 11 shows the output from the model under each scenario, averaged over the 40 years. These are virtually indistinguishable, although the differences in the means, shown in Figure 12, reflect that the mean temperature is generally lower under the modern-day land cover classes than for the potential land cover classes. In particular, the locations of the largest temperature changes correspond roughly to those locations in which land cover conversion has taken place; compare Figures 12 and 13.

6 Proposed research

The next logical steps for this research fall roughly into three categories: studying the properties of the tapering functions themselves, extending the results for the behavior of estimators using tapering, and the climate modeling application, to which we will apply tapering techniques. This section describes our proposed work in each of these areas.

6.1 Comparing tapering functions

Theorem 1 suggests one criterion for choosing a tapering function for a real-valued process Z. We believe that the result will also hold for processes Z on $\Re^d, d \geq 1$ if the first two conditions of the theorem are changed to 1. $f_{taper}(\omega) \leq \frac{M_\epsilon}{(1+\omega^2)^{\nu+d/2+\epsilon}}$, and 2. $\epsilon > \max\{d/4, 1-\nu\}$, but we have not yet proven this result, as the required integrals are more difficult to bound in multiple dimensions. We note that proving this result will automatically extend our Theorem 2 as well, which is currently limited to one dimension by reliance on Theorem 1, but which in principle should extend also to $d = 2, 3$ (but not to higher dimensions, due to reliance on Zhang's (2004) result, which is true only for $d \leq 3$.)

Another useful criterion would be a characterization of tapering functions in terms of the spatial, rather than the spectral domain. To do this, we will use the principal irregular term (PIT) of the tapering function K_{taper}, defined to be the first term in the series expansion about 0 of K_{taper} that is not proportional to $|x|$ raised to an even power (Matheron, 1971; Stein, 1999). The behavior of a covariance function at the origin and the tail behavior of its spectral density are related by the Abelian and Tauberian theorems (see Stein, 1999, Section 2.8). We will provide conditions equivalent to those in Theorem 4.2 using the PIT rather than the spectral density, as these would be much easier to check in practice.

Finally, within the class of tapering functions satisfying the conditions of Theorem 1, we would like to determine how the particular form of the tapering function and the value of the taper range γ should be chosen. The tapers constructed by Wendland (1995, 1998) are of minimal degree, but we have not yet considered whether this has any effect on the performance of our estimators, or what a more statistically relevant criterion might be. A few references of note are Gaspari and Cohn (1999), Gneiting (2002), and Ehm et al. (2004), who explore various optimality properties of tapering functions. If we cannot find an appropriate theoretical criterion, we will compare the performance of our estimators using different tapering functions via simulation. We will also develop a rule of thumb for choosing the taper range γ that balances performance and computational efficiency.

6.2 Extending methods and theoretical results for estimators

One interesting unanswered question that is relevant to our research but does not deal with tapering directly is whether Zhang's (2004) result on the convergence of the MLE under the Matérn model can be extended to the case that both ρ and σ^2 are estimated, rather than fixing ρ^* and estimating only σ^2. This result would hold if, for instance, the rate of convergence of $\hat{\sigma}_n^2$ were uniform in ρ^*. If we found a method of proof for the MLE, it would most likely apply to our estimators based on the approximation as well, as the forms of the estimating equations are similar.

Concerning our tapering-based estimators, a critical aspect of this problem that requires further study is determining when the conditions of Theorem 3 will hold. We have explored several approaches to proving convergence under more general conditions, for example using results on M-estimators which are functions of dependent observations, but the behavior of the quadratic form $Z_n([\Gamma_n \circ T_n]^{-1} \circ T_n) Z_n$ as $n \to \infty$ is the crux in all of these approaches. This quadratic form is expressible as $\sum_{i=1}^{n} \lambda_{n,i} \chi_i^2$, where χ_i^2 are *iid* χ_1^2 random variables and $\lambda_{n,i}$ depends on the probability measure for Z. In the statement of Theorem 3, we have used the original probability measure, giving that $\lambda_{n,i}$ is the i^{th} eigenvalue of $([\Gamma_n \circ T_n]^{-1} \circ T_n)\Gamma_n$. On the other hand, if we require the tapering function to satisfy the conditions of Theorem 1, then the true probability measure is equivalent to the probability measure under which $Z_n \sim N_n(0, \Sigma_n \circ T_n)$, which gives that $\lambda_{n,i}$ is the i^{th} eigenvalue of $([\Gamma_n \circ T_n]^{-1} \circ T_n)[\Gamma_n \circ T_n]$, which may be more tractable. The condition on the eigenvalues we require in Theorem 3 can be cast in terms of the condition numbers of the matrices involved, which is a function of the sampling locations S_n as well as the covariance and tapering functions. These conditions should be easier to check when the sampling locations form a regular lattice, so we propose to explore this case first, in hopes of gaining insight into the general requirements.

So far, we have only applied the idea of tapering to provide approximations to the MLE. However, this idea generalizes very naturally to the case of REML or Bayesian estimation. We will give approximations similar to (4) and (5) for REML estimation. Some results exist concerning equivalence of Gaussian measures with different means; we will explore whether these can be used to extend our convergence results to the task of estimating both the mean and variance of the process. To simplify the problem, we will focus on the case that the mean is regression function of a small number of unknown parameters and fixed spatial basis functions, a common assumption in practice. We will also explore the particulars of fitting a Bayesian model incorporating (4) or (5) in place of the exact form of the likelihood. We have done some preliminary work in this direction, estimating σ^2 and ρ in a mean-zero model with exponential covariance function (7), using a conjugate (inverse Gamma) prior for σ^2 and a Gamma prior for ρ. Several issues of interest have arisen. For instance, if one uses a Gibbs sampler, with a Metropolis step for sampling ρ, then the likelihood needs only to be evaluated during the Metropolis step, but the parameters of the full conditional distribution for σ^2 still involve inverting the covariance matrix. Even if covariance tapering is used at each step, this method does not appear optimal from a computational point of view. On the other hand, one might consider sampling ρ and σ^2 jointly using the Metropolis algorithm, or even importance sampling, requiring only one evaluation of the likelihood at each iteration. However, the Metropolis algorithm will potentially take longer to converge to the equilibrium distribution, so there appears to be some tradeoff in computational efficiency between these two options concerning the best sampling algorithm to use. Another issue is the parameterization of the covariance function itself. Would the sampler perform better if the model were reparameterized in terms of Zhang's (2004) estimable parameter $c = \sigma^2/\rho$ and some other non-estimable parameter?

A final aspect of the tapering-based estimators that we have not addressed here is their numerical performance in terms of computation time. We have written functions in R using the SparseM package (Koenker and Ng, 2005), which uses the sparse factorization and permutation algorithm of Ng and Peyton (1993). An initial study of the time required to find the determinant and inverse of $\Sigma_n \circ T_n$ for increasing n suggests that these functions are more efficient than standard matrix algorithms when n is larger than about 200. It has been suggested to us (Reinhard Furrer, personal communication) that the R functions in SparseM are relatively inefficient, but that the underlying source code is efficient and may be called directly from R. We plan to implement this approach, and we will give more detailed results concerning the computation time.

6.3 Climate modeling application

We propose to fit a Bayesian hierarchical model to the temperature output, allowing us to answer the types of questions posed in Section 5. In particular, we will consider a model under which the (prior) distribution of the observed quantities Z_{obs}, conditional on the long-run functions $Z_x(s)$ and $Z_{x'}(s)$, takes them to be observations from an underlying stochastic process in each case with the long-run functions as their means. That is, for observed s and t, take

$$Z_x(s,t) = Z_x(s) + e_x(s,t),$$

where $e_x(s,t)$ is a mean zero Gaussian process. We will model $Z_x(s)$ as a linear combination of spatial functions $\phi_j(s,x)$, representing functions of the input x (for example. indicator functions for the land-cover class at each location s) and possibly also a limited number of spherical harmonic functions, to capture variations in the mean surface unrelated to land cover. Write $Z_x(s) = \sum_{j=1}^{J} \phi_j(s,x)\alpha_j$, where the uncertainty is expressed via a prior distribution on $\{\alpha_j\}$. We expect $e_x(s,t)$ to behave like a mean zero Gaussian process which is stationary in time. A simple model would also assume stationarity and isotropy in space and independence in time. Although we do not anticipate this model will fit well, we will try fitting it first. That is, take $e_x(s,t)$ at locations in S_n and at time t to be normally distributed with mean zero and covariance matrix $\Sigma(\theta_x) = \{K(||s_i - s_j||;\theta_x)\}$. Then we can find posterior distributions for the covariance parameters θ_x, along with $\{\alpha_j\}$, using the approximations to the likelihood we have developed, taking the likelihood for all observations to be the product of the likelihoods at each timepoint. We anticipate fitting even this simple model will require considerable computational resources.

Moving towards a more complicated model for $e_x(s,t)$, one possibility, short of fitting a full spatial-temporal model for $e_x(s,t)$, is to decompose $e_x(s,t)$ into a part that is correlated in time and one that is independent in time. We envision a simple model for the time dependent component, taking it to be a linear combination of spatial functions $\xi_k(s,x)$, representing primary modes of variability; these are multiplied by coefficients $\beta_{k,t}$, following a simple time-series model such as an AR process. The full model would then be

$$Z_x(s,t) = Z_x(s) + e_x(s,t) = \sum_{j=1}^{J} \phi_j(s,x)\alpha_j + \sum_{k=1}^{K} \xi_k(s,x)\beta_{k,t} + \eta_x(s,t).$$

The rationale for decomposing the second-order structure into a sum of finite basis functions whose coefficients vary in time is that in observed climate data, principle component analysis reveals that most of the variability from year to year consists of variations in a few spatial functions. For instance, variability in the first function is typically attributed to the Southern Oscillation, or El Niño. We will consider fixing the $\{\xi_k(s,x)\}$ a priori, using existing climate data as a guide. Then, fitting the model for $e_x(s,t)$ corresponds to fitting a model for a finite number of $\{\beta_{k,t}\}$, as well as the spatial covariance function of the residual process $\eta_x(s,t)$, which we could then model as we originally described modeling $e_x(s,t)$, as a mean zero Gaussian process with stationary and isotropic covariance function, with only a few parameters θ_x to be estimated.

7 Figures

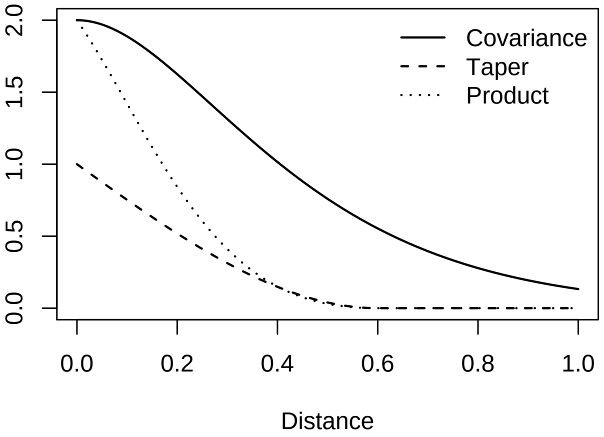

Figure 1: Illustration of tapering a covariance function. In this case, the original covariance function is Matérn (see Section 4.1) with parameters $\sigma^2 = 2, \rho = 0.2$, and $\nu = 2$. The tapering function is the spherical correlation function (3) with $\gamma = 0.6$. Note that the original covariance function is smoother at the origin than the tapering function. This is problematic, as we show in Theorem 1.

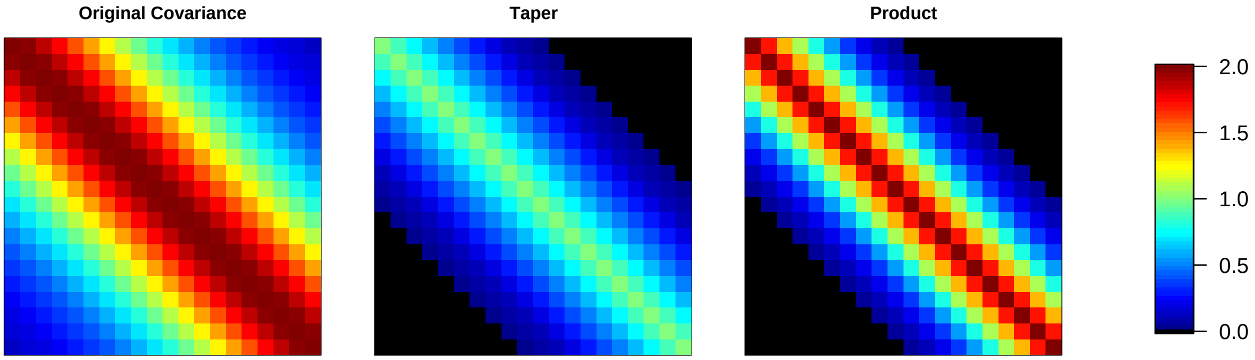

Figure 2: Illustration of tapering a covariance matrix. The sampling locations S_n consist of 20 equally spaced points between 0 and 1, and the covariance function and taper are the same as those in Figure 1.

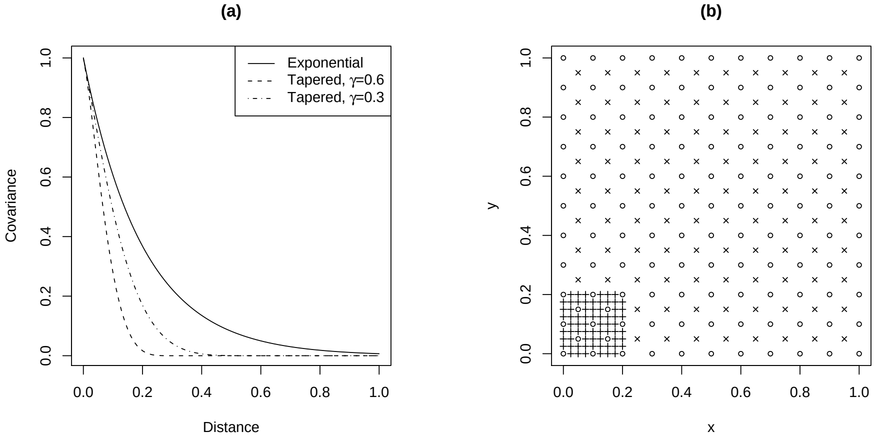

Figure 3: (a) Exponential covariance function used in the simulation and the same function tapered at two different ranges. (b) Locations used in the simulation study: n=125 includes points marked “$\circ$” only, n=221 also includes those marked “×”, and n=289 also includes those marked “+.” The smallest, with $n = 125$, consisted of the set $\{i/10, j/10\}_{i,j\in\{0,\ldots,10\}} \bigcup (x, y)_{x,y\in\{0.05,.15\}}$. The next, with $n = 221$, consisted additionally of the 96 points in $\{(.05 + .1i, .05 + .1j)\}_{i,j\in\{0,\ldots,9\}}$. The largest, with $n = 289$, consisted additionally of the 68 points in $\{(i/40, j/40)\}_{i,j\in\{0,\ldots,8\}}$ which had not been included previously.

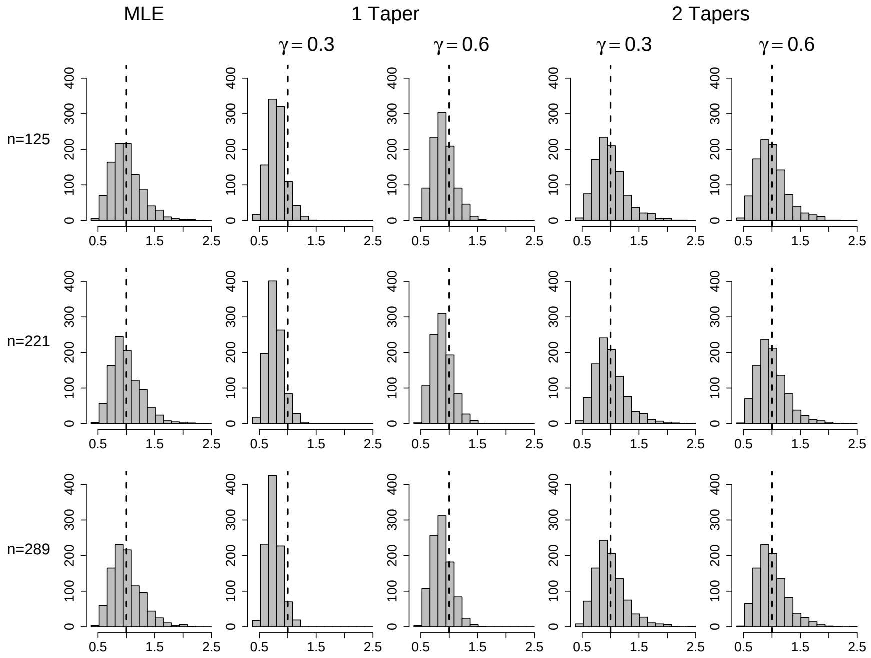

Figure 4: Distributions of estimators of σ^2. True value of $\sigma_0^2 = 1$ is shown as a dotted vertical line.

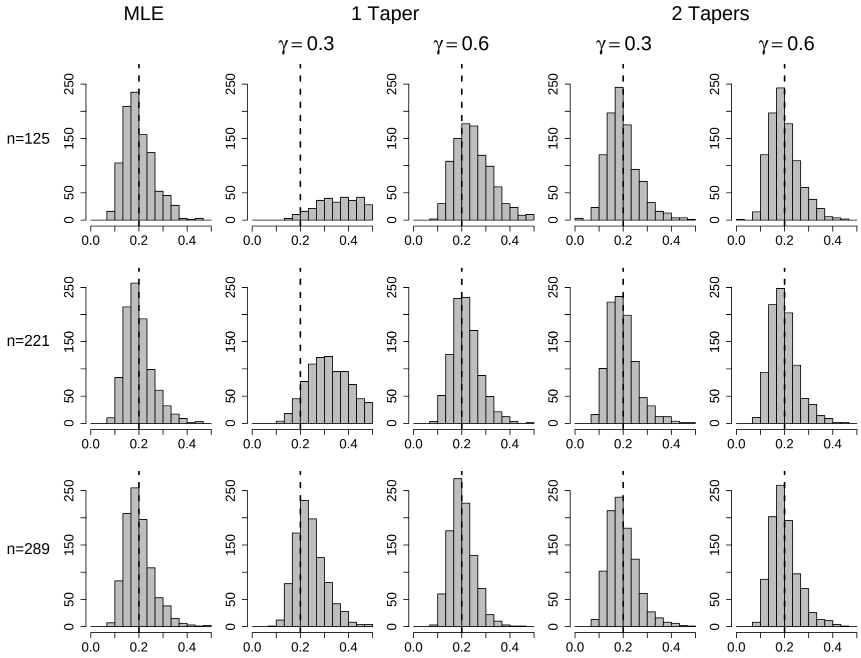

Figure 5: Distributions of estimators of ρ. True value of $\rho_0 = 0.2$ is shown as a dotted vertical line. The histograms for the estimators based on $l_{n,1taper}$ with $\gamma = 0.3$ have been truncated to facilitate comparison without extending the range so much that the defining features of the other plots are lost. Areas represent proportions of total simulations, so it is clear that over half the distribution at $n = 125$ lies above 0.5, while about a quarter of the distribution at $n = 221$ lies above 0.5.

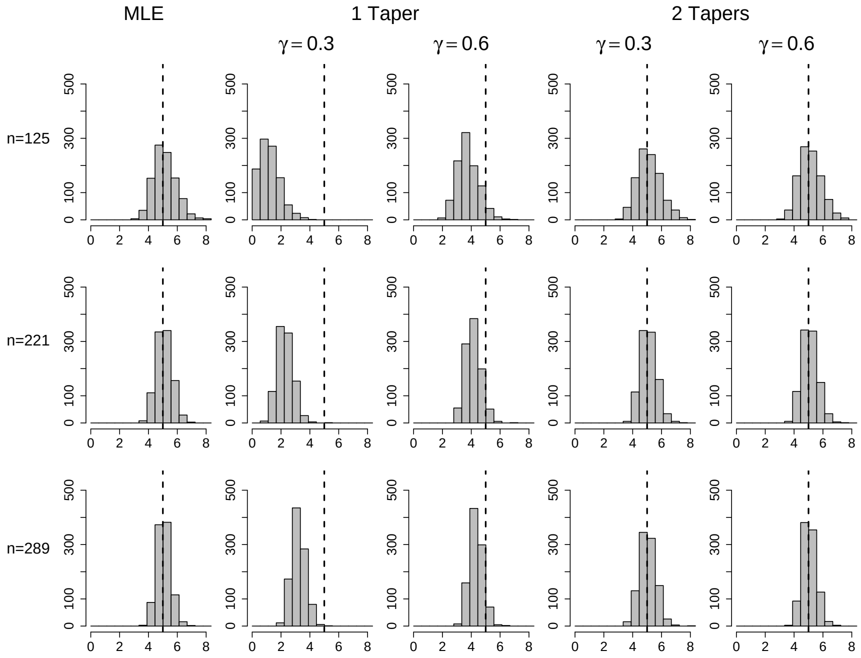

Figure 6: Distributions of estimators of $c = \sigma^2/\rho$. True value of $c_0 = 5$ is shown as a dotted vertical line.

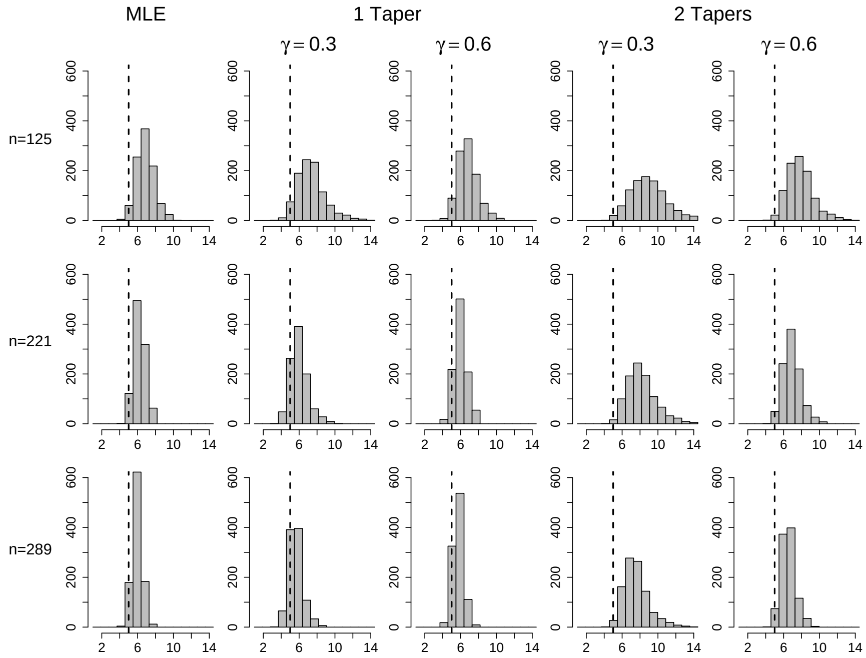

Figure 7: Distributions of estimators of $c = \sigma^2/\rho$ when ρ is fixed at 0.01.

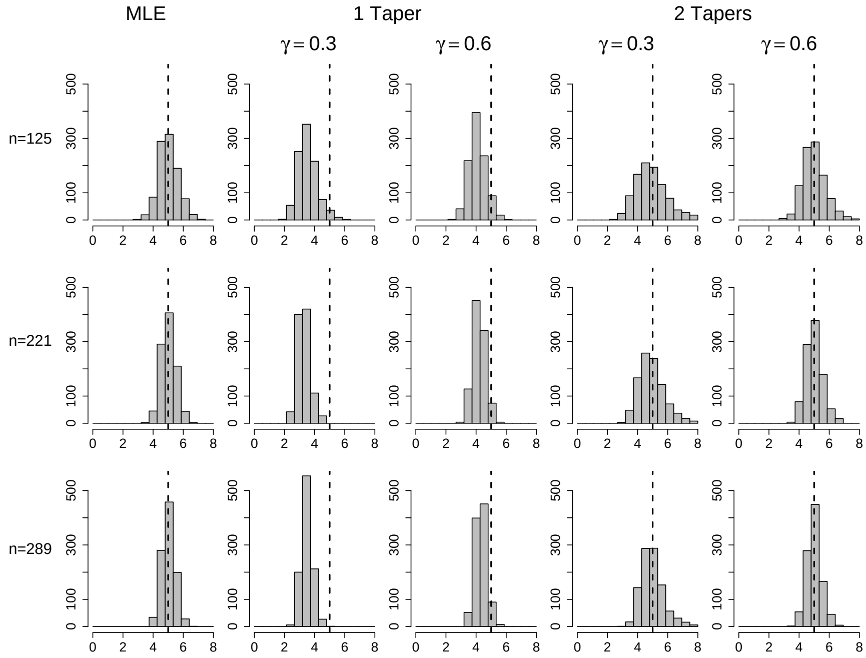

Figure 8: Distributions of estimators of $c = \sigma^2/\rho$ when ρ is fixed at 0.02 (the true value).

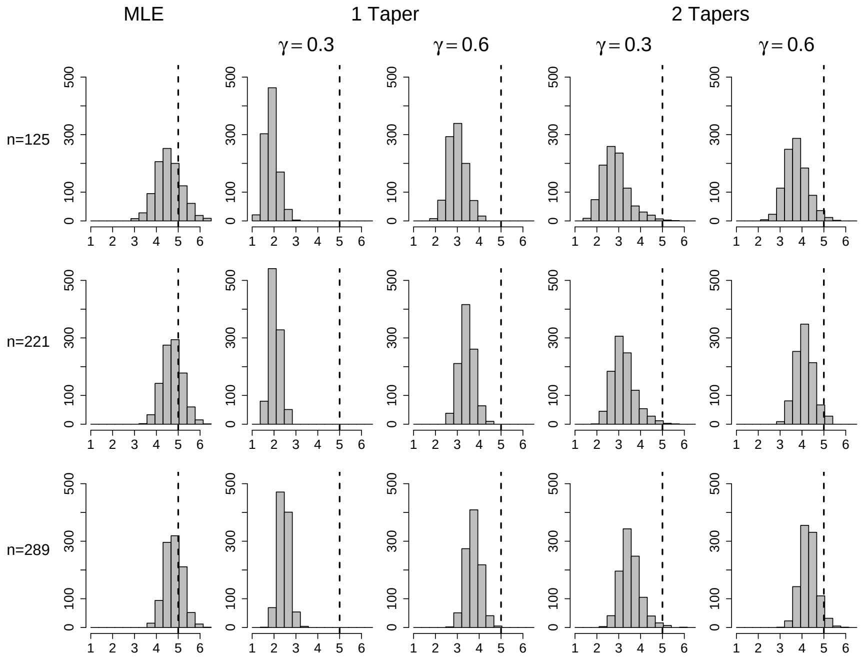

Figure 9: Distributions of estimators of $c = \sigma^2/\rho$ when ρ is fixed at 0.04.

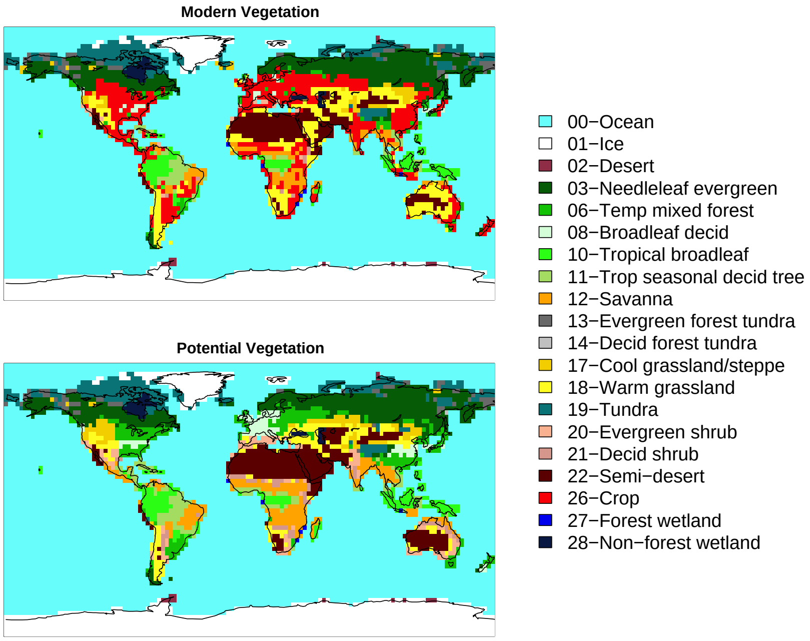

Figure 10: Land cover classes under which the model was run.

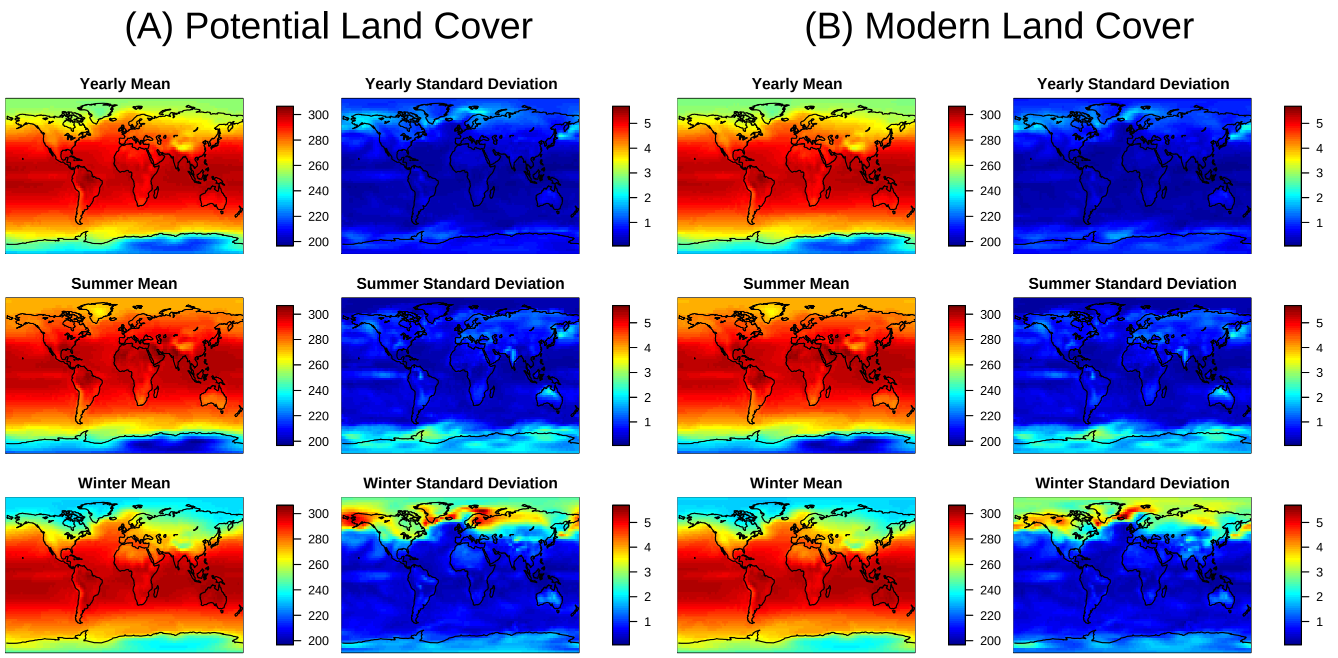

Figure 11: Temperature output under the modern day land cover classes, averaged over 40 years, and standard deviations of these values. The scale is degrees Kelvin.

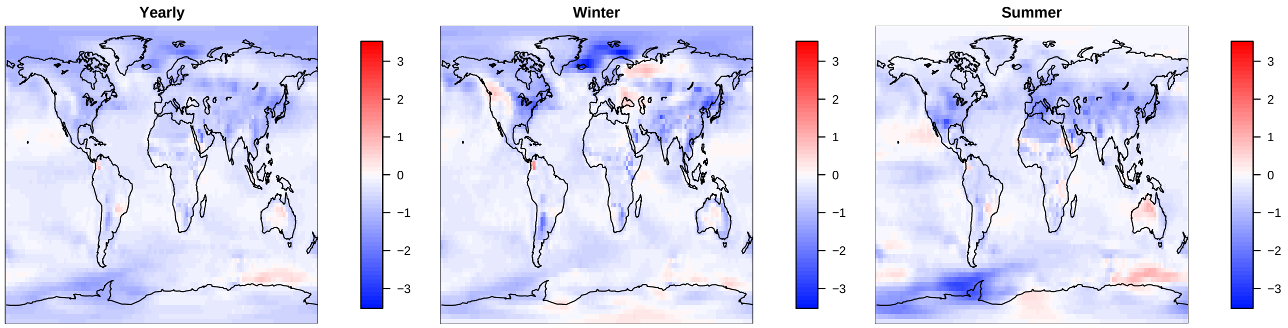

Figure 12: Differences between mean temperature output (modern minus potential). The scale is degrees Kelvin.

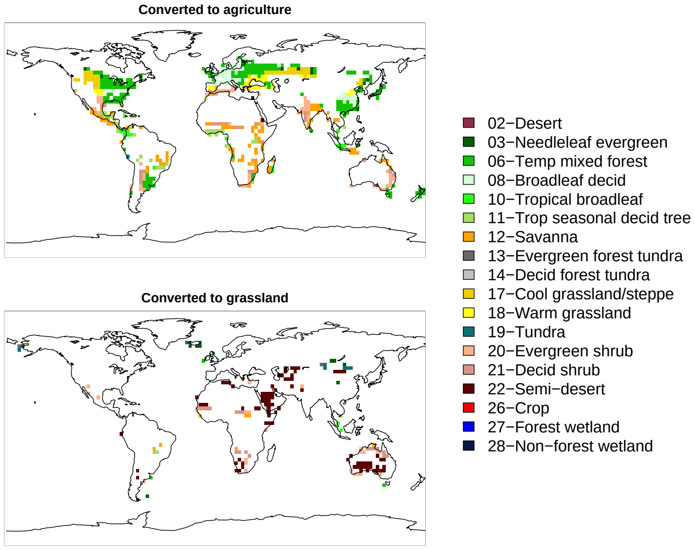

Figure 13: Land cover classes in the potential land cover dataset which were converted to agriculture or grassland in the modern day land cover dataset.

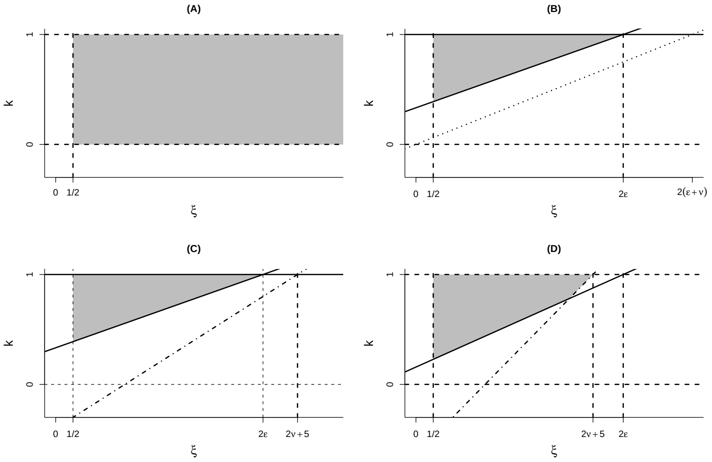

Figure 14: Illustration of how ξ and k are chosen in the proof of Theorem 1. **(A)** First, we require that $\xi > 1/2$. This is the desired rate of convergence. Also, we require that k be strictly between 0 and 1, so that $\Delta = \omega^k \rightarrow \infty$ as $\omega \rightarrow \infty$, but $\omega - \Delta > 0$. **(B)** The solid diagonal line corresponds to the requirement that $k > \frac{2\nu+1+\xi}{2\nu+1+2\epsilon}$ in (19). Because this line intersects $k = 1$ at 2ϵ and we require $k < 1$, this adds an additional requirement that $\xi < 2\epsilon$. Note also that the intercept $\frac{2\nu+1}{2\nu+1+2\epsilon}$ lies strictly between 0 and 1. The dotted diagonal line corresponds to the requirement that $k > \frac{\xi}{2(\nu+\epsilon)}$ in (22). But we've already specified that $\xi < 2\epsilon$, so this bound is redundant, because the function has a zero intercept and is 1 for $2(\epsilon+\nu) > 2\epsilon$. **(C)** The solid diagonal line is the same as in (B). The last requirement is that $k > \frac{\xi-2}{2\nu+3}$ in (24). The right hand side is 1 when $\xi = 2\nu+5$. If $2\nu+5 > 2\epsilon$, as is shown here, this bound (dashed/dotted diagonal line) is also redundant. **(D)** But if $2\nu+5 \leq 2\epsilon$, then there are values of ξ for which we can't find a k both larger than this bound and strictly less than 1. So we add a requirement that $\xi < 2\nu+5$. Clearly $2\nu+5 > 1/2$, so this does not contradict our lower bound for ξ. Putting it all together, this means we first choose $\xi \in (1/2, min\{2\epsilon, 2\nu+5\})$. Then choose $k \in \left(\max\left\{\frac{2\nu+1+\xi}{2\nu+1+2\epsilon}, \frac{\xi-2}{2\nu+3}\right\}\right)$.

8 Tables

$d = 1$	$k = 0$	$\phi_{1,0}(r) = (1 - r)_+$
	$k = 1$	$\phi_{1,1}(r) \propto (1 - r)_+^3(3r + 1)$
	$k = 2$	$\phi_{1,2}(r) \propto (1 - r)_+^5(8r^2 5r + 1)$
$d = 2, 3$	$k = 0$	$\phi_{d,0}(r) = (1 - r)_+^2$
	$k = 1$	$\phi_{d,1}(r) \propto (1 - r)_+^4(4r + 1)$
	$k = 2$	$\phi_{d,2}(r) \propto (1 - r)_+^6(35r^2/3 + 6r + 1)$

Table 1:

	MLE		Tapered, $\gamma = 0.3$		Tapered, $\gamma = 0.6$	
	Theoretical	Simulated	Theoretical	Simulated	Theoretical	Simulated
$n = 125$	0.0768	0.0756	0.0846	0.0768	0.0805	0.0728
$n = 221$	0.0752	0.0726	0.0849	0.0761	0.0792	0.0714
$n = 289$	0.0748	0.0754	0.0843	0.0775	0.0787	0.0732

Table 2: Comparison of theoretical and simulated variances for estimators of σ^2, under the same conditions as in the simulation study.

	MLE		Tapered, $\gamma = 0.3$		Tapered, $\gamma = 0.6$	
	Theoretical	Simulated	Theoretical	Simulated	Theoretical	Simulated
$n = 125$	0.0042	0.0041	0.0047	0.0044	0.0044	0.0040
$n = 221$	0.0037	0.0035	0.0041	0.0036	0.0039	0.0035
$n = 289$	0.0035	0.0035	0.0041	0.0040	0.0037	0.0035

Table 3: Comparison of theoretical and simulated variances for estimators of ρ.

	MLE		Tapered, $\gamma = 0.3$		Tapered, $\gamma = 0.6$	
	Theoretical	Simulated	Theoretical	Simulated	Theoretical	Simulated
$n = 125$	0.5484	0.6195	0.6307	0.6925	0.5584	0.6167
$n = 221$	0.2799	0.2950	0.2960	0.3224	0.2843	0.3076
$n = 289$	0.2041	0.2208	0.2980	0.3310	0.2104	0.2348

Table 4: Comparison of theoretical and simulated variances for estimators of $c = \sigma^2/\rho$.

A Technical Appendix

A.1 Estimating functions

A function $G(Z_n;\theta)$ of the data vector Z_n and parameters θ is called an unbiased estimating function for θ if $\mathrm{E}_\theta[G(Z_n;\theta)] = 0$ for all possible values of θ. In this section we give expressions for the score function for our model, obtained by differentiating the log-likelihood (1) with respect to θ, as well as the estimating functions obtained by differentiating approximations (4) and (5). In particular, we show that the estimating function corresponding to (4) is biased, while the estimating function corresponding to (5) is, like the score function, unbiased. For ease of notation, in what follows we suppress the dependence of any quantity on n and the dependence of the covariance matrices $\Sigma_n(\theta)$ and $T_n(\gamma)$ on θ and γ. We make use of the following three facts from matrix calculus:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\log|\Sigma| &= \mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right\} \\
\frac{\partial}{\partial\theta_i}\Sigma^{-1} &= -\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1} \\
\frac{\partial}{\partial\theta_i}[\Sigma\circ T] &= \frac{\partial\Sigma}{\partial\theta_i}\circ T \quad \text{when T does not depend on } \theta.
\end{aligned}$$

First, let G represent the score function, the vector-valued function whose i^{th} element is

$$\begin{aligned}
G_i &= \frac{\partial}{\partial\theta_i} l(\theta) \\
&= -\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right\} + \frac{1}{2}Z'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}Z. \qquad (13)
\end{aligned}$$

We can verify the well known fact that $\mathrm{E}[G] = 0$ in this particular case. The two terms of (13) will cancel when the expected value is taken, because

$$\begin{aligned}
\mathrm{E}\left[Z'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}Z\right] &= \mathrm{E}\left[\mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}ZZ'\right\}\right] \\
&= \mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\mathrm{E}[ZZ']\right\} \\
&= \mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right\}.
\end{aligned}$$

Now let G_{1taper} be the estimating equation corresponding to (4). That is, G_{1taper} is the vector-valued function with i^{th} element

$$\begin{aligned}
G_{1taper,i} &= \frac{\partial}{\partial\theta_i} l_{1taper}(\theta) \\
&= -\frac{1}{2}\mathrm{tr}\left\{[\Sigma\circ T]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\right\} + \frac{1}{2}Z'[\Sigma\circ T]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right][\Sigma\circ T]^{-1}Z.
\end{aligned}$$

Then cancellation of the two terms under the expected value does not occur, and so $\mathrm{E}[G_{1taper}] \neq 0$.

However, let $G_{2tapers}$ be the estimating equation corresponding to (5). Its i^{th} element is

$$\begin{aligned}
G_{2tapers.i} &= \frac{\partial}{\partial\theta_i} l_{2tapers}(\theta) \\
&= -\frac{1}{2}\mathrm{tr}\left\{[\Sigma\circ T]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\right\} + \frac{1}{2}\mathrm{tr}\left\{\left[\hat{\Sigma}\circ T\right][\Sigma\circ T]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right][\Sigma\circ T]^{-1}\right\}. \qquad (14)
\end{aligned}$$

Now the cancellation does occur, and so $\mathrm{E}[G_{2tapers}] = 0$.

A.2 Profile likelihoods

For any fixed value of ϕ, the value of σ^2 which maximizes (1) is $\hat{\sigma}_n^2(\phi) = Z_n'\Gamma_n(\phi)^{-1}Z_n/n$, where $\Gamma_n(\phi) = \Sigma_n(\theta)/\sigma^2 = \{C_0(||s_i - s_j||;\phi)\}$, the correlation matrix. If we plug this value into the log-likelihood, we obtain the profile log-likelihood

$$\begin{aligned} pl_n(\phi) &= \sup_{\sigma^2} l_n(\sigma^2, \phi) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{Z_n'\Gamma_n(\phi)^{-1}Z_n}{n} - \frac{1}{2}\log|\Gamma_n(\phi)| - \frac{n}{2}. \end{aligned} \tag{15}$$

Therefore, maximizing the log-likelihood (1) over both σ^2 and ϕ is equivalent to maximizing (15) over ϕ to obtain $\hat{\phi}_n$, then calculating $\hat{\sigma}_n^2(\hat{\phi}_n)$. Profile versions of the approximations (4) and (5) can also be used. These are

$$pl_{n,1taper}(\phi) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{Z_n'\left[\Gamma_n(\phi)\circ T_n(\gamma)\right]^{-1}Z_n}{n} - \frac{1}{2}\log|\Gamma_n(\phi)\circ T_n(\gamma)| - \frac{n}{2}, \tag{16}$$

with $\hat{\sigma}_{n,1taper}^2 = Z_n'\left[\Gamma_n(\hat{\phi}_{n,1taper})\circ T_n(\gamma)\right]^{-1}Z_n/n$, and

$$pl_{n,2tapers}(\phi) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{Z_n'\left(\left[\Gamma_n(\phi)\circ T_n(\gamma)\right]^{-1}\circ T_n(\gamma)\right)Z_n}{n} - \frac{1}{2}\log|\Gamma_n(\phi)\circ T_n(\gamma)| - \frac{n}{2}, \tag{17}$$

with $\hat{\sigma}_{n,2tapers}^2 = Z_n'\left(\left[\Gamma_n(\hat{\phi}_{n,2tapers})\circ T_n(\gamma)\right]^{-1}\circ T_n(\gamma)\right)Z_n/n$.

A.3 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; hence,

$$f_1(\omega) = \int_{\Re} f_0(x)f_{taper}(\omega - x)dx. \tag{18}$$

Stein (2004, Theorem A.1) provides the following two conditions, which are sufficient to give $G(0, K_0) \equiv G(0, K_1)$ on the paths of $\{Z(s), s \in S\}$ for any bounded subset $S \subset \Re^d$:

A. $\exists\eta > d$ such that $f_0(\omega)|\omega|^\eta$ is bounded away from 0 and ∞ as $|\omega| \to \infty$,

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

In our case, condition A is satisfied for $\eta = 2\nu + 1$, with $f_0(\omega)$ being the Matérn spectral density (9) with $d = 1$. In addition, both f_0 and f_{taper} are isotropic, so $f_0(\omega) = f_0(-\omega)$ and $f_1(\omega) = f_1(-\omega)$. Therefore, we need only show there exists a $c < \infty$ such that $\int_c^\infty \left\{\frac{f_1(\omega)-f_0(\omega)}{f_0(\omega)}\right\}^2 d\omega < \infty$.

Consider $\omega > 0$. Throughout, we use the notation $f(\omega) = O(g(\omega))$ to indicate that $f(\omega) \geq 0$ and that there exist $0 < L < \infty$ and $0 < c < \infty$ such that $f(\omega) \leq Lg(\omega)$ for all $\omega \geq c$. We will show that $\left|\frac{f_1(\omega)}{f_0(\omega)} - 1\right| = O(w^{-\xi})$, where $\xi > 1/2$. This implies that there exist $0 < L < \infty$ and $0 < c < \infty$ such that

$$\int_c^\infty \left\{\frac{f_1(\omega) - f_0(\omega)}{f_0(\omega)}\right\}^2 d\omega \leq \int_c^\infty \frac{L^2}{\omega^{2\xi}}d\omega < \infty.$$

We begin by dividing the range of integration in (18) into three intervals: $(-\infty, \omega - \Delta]$, $(\omega - \Delta, \omega + \Delta]$, and $(\omega + \Delta, \infty)$. The intuition is that the ratio $f_1(\omega)/f_0(\omega)$ is going to one when evaluated over the middle interval, but it is going to zero when evaluated over the outer intervals. We choose Δ as follows. As $\epsilon > 1/4$, we can choose $\xi \in (1/2, \min\{2\epsilon, 2\nu+5\})$. Then we can choose $k \in \left(\max\left\{\frac{2\nu+1+\xi}{2\nu+1+2\epsilon}, \frac{\xi-2}{2\nu+3}\right\}, 1\right)$ and let $\Delta = \omega^k$. The rationale for this choice is that it forces certain inequalities to hold which are needed in the remainder of the proof; these are illustrated in Figure 14. Addressing each interval separately, we have

- When $x \in (-\infty, \omega-\Delta], \omega - x \geq \Delta$, so condition 1 of the theorem implies
$$f_{taper}(\omega - x) \leq \frac{M_\epsilon}{(1+(\omega-x)^2)^{\nu+1/2+\epsilon}} \leq \frac{M_\epsilon}{(1+\Delta^2)^{\nu+1/2+\epsilon}} = \frac{M_\epsilon}{(1+\omega^{2k})^{\nu+1/2+\epsilon}}.$$
Write $f_0(x) = \sigma^2 \frac{M_0}{(\rho^{-2}+x^2)^{\nu+1/2}}$. Note that $\int_\Re f_0(x)dx = \sigma^2$. Therefore,
$$\begin{aligned} 0 < \frac{\int_{-\infty}^{\omega-\Delta} f_0(x) f_{taper}(\omega-x)dx}{f_0(\omega)} &\leq \frac{\frac{M_\epsilon}{(1+\omega^{2k})^{\nu+1/2+\epsilon}} \int_\Re f_0(x)dx}{f_0(\omega)} \\ &\leq \frac{M_\epsilon}{M_0} \frac{(\rho^{-1}+\omega^2)^{\nu+1/2}}{(1+\omega^{2k})^{\nu+1/2+\epsilon}} = O(\omega^{-\xi}), \quad (19) \end{aligned}$$
because $k > \frac{2\nu+1+\xi}{2\nu+1+2\epsilon}$.
- When $x \in (\omega-\Delta, \omega+\Delta]$, we expand $f_0(x)$ about ω. Specifically, for some $\omega^* \in (\omega-\Delta, \omega+\Delta]$,
$$\begin{aligned} \frac{\int_{\omega-\Delta}^{\omega+\Delta} f_0(x) f_{taper}(\omega-x)dx}{f_0(\omega)} &= \frac{\int_{\omega-\Delta}^{\omega+\Delta} [f_0(\omega) + f_0'(\omega)(\omega-x) + f_0''(\omega^*)\frac{(\omega-x)^2}{2}] f_{taper}(\omega-x)dx}{f_0(\omega)} \\ &= \int_{\omega-\Delta}^{\omega+\Delta} f_{taper}(\omega-x)dx + \quad (20) \\ &\quad \frac{f_0''(\omega^*)}{2f_0(\omega)} \int_{\omega-\Delta}^{\omega+\Delta} (\omega-x)^2 f_{taper}(\omega-x)dx, \quad (21) \end{aligned}$$
as the integral of the term corresponding to the first derivative is zero.

  The integral in (20) is clearly less than one by condition 3 of the theorem; we will show it is bounded below by $1 - O(\omega^{-\xi})$. Likewise, calculating f_0'' shows (21) is greater than zero whenever $\omega > \sqrt{\frac{1}{2\rho^2(\nu+1)}}$; we will show it is bounded above by $O(\omega^{-\xi})$. Addressing the integral in (20) first, we have
$$\begin{aligned} \int_{\omega-\Delta}^{\omega+\Delta} f_{taper}(\omega-x)dx &= 1 - \int_{|x|>\Delta} f_{taper}(x)dx \\ &\geq 1 - \int_{|x|>\Delta} \frac{M_\epsilon}{(1+x^2)^{\nu+1/2+\epsilon}} dx \\ &\geq 1 - \int_{|x|>\Delta} \frac{M_\epsilon}{x^{2(\nu+1/2+\epsilon)}} dx \\ &= 1 - \frac{M_\epsilon}{(\nu+\epsilon)\Delta^{2(\nu+\epsilon)}} = 1 - \frac{M_\epsilon}{(\nu+\epsilon)\omega^{2k(\nu+\epsilon)}} \\ &= 1 - O(\omega^{-\xi}) \quad (22) \end{aligned}$$
because $k > \frac{2\nu+1+\xi}{2\nu+1+2\epsilon}$ and $\xi < 2\epsilon$ imply that $k > \frac{\xi}{2(\nu+\epsilon)}$.

  Now we evaluate (21). Let M_{bnd} be the normalizing constant required for $\frac{M_{bnd}}{(1+x^2)^{\nu+1/2+\epsilon}}$ to integrate to 1. We recognize this as the density of $t_{2(\nu+\epsilon)}/\sqrt{2(\nu+\epsilon)}$, where $t_{2\nu+\epsilon}$ denotes a random variable with t distribution and (possibly fractional) degrees of freedom $2(\nu+\epsilon)$. Then,

$$\begin{aligned}\int_{\omega-\Delta}^{\omega+\Delta}(\omega-x)^2 f_{taper}(\omega-x)dx &= \int_{-\Delta}^{\Delta}x^2 f_{taper}(x)dx \\ &\leq \int_{-\Delta}^{\Delta}x^2\frac{M_\epsilon}{(1+x^2)^{\nu+1/2+\epsilon}}dx \\ &\leq \frac{M_\epsilon}{M_{bnd}}\int_{-\infty}^{\infty}x^2\frac{M_{bnd}}{(1+x^2)^{\nu+1/2+\epsilon}}dx \\ &= \frac{M_\epsilon}{M_{bnd}}\mathrm{Var}\left(\frac{t_{2(\nu+\epsilon)}}{\sqrt{2(\nu+\epsilon)}}\right) \\ &\equiv L<\infty,\end{aligned}$$

because $\nu+\epsilon>1$ by condition 2 of the theorem. Now,

$$f_0''(\omega^*)=\frac{\sigma^2 M_0(2\nu+1)}{(\rho^{-2}+\omega^{*2})^{\nu+3/2}}\left[\frac{(2\nu+3)\omega^{*2}}{\rho^{-2}+\omega^{*2}}-1\right],$$

which is decreasing whenever $\omega^*>\sqrt{\frac{3}{2\rho^2(\nu+1)}}$. Therefore, eventually $\omega-\Delta$ will be large enough so that $\sup_{\omega^*\in(\omega-\Delta,\omega+\Delta)}f_0''(\omega^*)=f_0''(\omega-\Delta)=f_0''(\omega-\omega^k)$, and so we can write (21) as

$$\begin{aligned}\frac{f_0''(\omega^*)}{2f_0(\omega)}\int_{\omega-\Delta}^{\omega+\Delta}(\omega-x)^2 f_{taper}(\omega-x)dx &\leq \frac{L}{2f_0(\omega)}\sup_{\omega^*\in(\omega-\Delta,\omega+\Delta)}f_0''(\omega^*) \qquad (23)\\ &\leq \frac{L(2\nu+1)}{2}\frac{(\rho^{-2}+\omega^2)^{\nu+1/2}}{(\rho^{-2}+(\omega-\omega^k)^2)^{\nu+3/2}}\left[\frac{(2\nu+3)(\omega-\omega^k)^2}{\rho^{-2}+(\omega-\omega^k)^2}-1\right] \\ &= O(\omega^{-\xi}), \qquad (24)\end{aligned}$$

because $k>\frac{\xi-2}{2\nu+3}$.

- When $x\in(\omega+\Delta,\infty)$, $f_0(x)\leq f_0(\omega)$, so

$$\begin{aligned}0<\frac{\int_{\omega+\Delta}^{\infty}f_0(x)f_{taper}(\omega-x)dx}{f_0(\omega)} &\leq \int_{\omega+\Delta}^{\infty}f_{taper}(\omega-x)dx \\ &= \int_{-\infty}^{-\Delta}f_{taper}(x)dx \\ &= \frac{1}{2}\int_{|x|>\Delta}f_{taper}(x)dx=O(\omega^{-\xi}) \qquad (25)\end{aligned}$$

by the same reasoning as in (22).

As $\frac{f_1(\omega)}{f_0\omega}$ is the sum of the terms in (19), (22), (24), and (25), we have shown that $\left|\frac{f_1(\omega)}{f_0(\omega)}-1\right|=O(w^{-\xi})$, which completes the proof.

A.4 Proof of Theorem 2

This result is an easy consequence of our Theorem 1, as well as Theorem 2 of Zhang (2004). Let $G(0,K_0)$ be the true probability measure for Z (Gaussian with mean zero and Matérn covariance function with parameters σ_0,ρ_0, and ν). According to Theorem 2 of Zhang (2004), we may find for any fixed $\rho^*>0$ a $\sigma^{2*}>0$ such that $G(0,K_0)$ and $G(0,K_1)$ are equivalent, where $G(0,K_1)$ is the mean zero Gaussian measure for Z with Matérn covariance function with parameters σ^{2*},ρ^*, and ν. That is, let

$\sigma^{2*} = \sigma_0^2(\rho_0/\rho^*)^{2\nu}$. By Theorem 1, we also know that $G(0,K_1)$ is equivalent to $G(0,K_2)$, the mean zero Gaussian measure with covariance function equal to the direct product of K_1 and and K_{taper}, a tapering function satisfying the conditions of Theorem 1. Therefore, to show $\hat{\sigma}^2_{n,1taper}/\rho^{*2\nu} \to \sigma_0^2/\rho_0^{2\nu}$ a.s. $[G(0,K_0)]$, it is sufficient to show $\hat{\sigma}^2_{n,1taper} \to \sigma^{2*}$ a.s. $[G(0,K_2)]$.

Because ρ^* and ν are fixed, the exact expression for $\hat{\sigma}^2_{n,1taper}$ is

$$\hat{\sigma}^2_{n,1taper} = Z_n \left[\Gamma_n(\rho*,\nu) \circ T_n(\gamma)\right]^{-1} Z_n/n,$$

where $\Gamma_n(\rho^*,\nu) = \frac{1}{\sigma^{2*}}\left\{K_1(||s_i - s_j||;\sigma^{2*},\rho^*,\nu)\right\}$ and $T_n(\gamma) = \left\{K_{taper}(||s_i - s_j||;\gamma)\right\}.$

But under $G(0,K_2), Z_n \sim N_n(0, \sigma^{2*}\Gamma_n(\rho^*,\nu) \circ T_n(\gamma))$, 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. $[G(0,K_2)]$ by the Strong Law of Large Numbers.

A.5 Proof of Theorem 3

Begin as in the proof of Theorem 2, letting $G(0,K_0)$ denote the true probability measure and $G(0,K_1)$ the probability measure with K_1 equal to a Matérn covariance with parameters $\sigma^{2*} = \sigma_0^2(\rho_0/\rho^*)^{2\nu}, \rho^*$, and ν. Then $G(0,K_0)$ and $G(0,K_1)$ are equivalent by Theorem 2 of Zhang (2004), so it is sufficient to show $\hat{\sigma}^2_{n,2tapers} \to \sigma^{2*}$ a.s. $[G(0,K_1)]$.

The exact expression for $\hat{\sigma}^2_{n,2tapers}$ is

$$\begin{aligned} \hat{\sigma}^2_{n,2tapers} &= Z_n'\left(\left[\Gamma_n(\rho^*,\nu) \circ T_n(\gamma)\right]^{-1} \circ T_n(\gamma)\right) Z_n/n \\ &= Z_n' W_n^{-1} Z_n/n. \end{aligned}$$

Under $G(0,K_1), Z_n \sim N_n(0, \Sigma(\sigma^{2*},\rho^*,\nu))$. Write $\Gamma_n(\rho^*,\nu) = RR'$. Then $\frac{1}{\sigma^*}R^{-1}Z_n \sim N_n(0,I_n)$, so

$$\begin{aligned} \hat{\sigma}^2_{n,2tapers} &= Z_n' W_n^{-1} Z_n/n \\ &= \frac{1}{n}\left(\frac{1}{\sigma^*}R^{-1}Z_n\right)' (\sigma^* R)' W_n^{-1} (\sigma^* R) \left(\frac{1}{\sigma^*}R^{-1}Z_n\right) \\ &= \frac{1}{n} X_n' \left[(\sigma^* R)' W_n^{-1} (\sigma^* R)\right] X_n, \quad \text{where} X_n \sim N_n(0,I_n) \\ &= \frac{\sigma^{2*}}{n} \sum_{i=1}^{n} \lambda_{n,i}\chi_i^2, \end{aligned} \tag{26}$$

where χ_i^2 are *iid* χ_1^2 random variables and λ_{n_i} is the i^{th} eigenvalue of $R'W_n^{-1}R$, which is the same as the i^{th} eigenvalue of $W_n^{-1}\Gamma_n$.

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 $\limsup_{i \leq n} |a_{n,i}|n^{-1}\log n$. We finish the proof by applying these results to (26), with $X_i = \chi_i^2 - 1$ and $a_{n,i} = \lambda_{n,i}$.

A.6 Information matrices

The following fact, using the moment properties of the multivariate normal distribution, is useful in the derivation of the information matrices in this section.

Lemma 1. *Suppose matrices A and B are symmetric and Z has multivariate normal distribution with mean zero and covariance matrix Σ. Then*

$$Cov(Z'AZ, Z'BZ) = 2tr\{A\Sigma B\Sigma\}.$$

Proof.

$$\begin{aligned}
Cov\left(Z'AZ, Z'BZ\right) &= Cov\left(\mathrm{tr}\left\{\hat{\Sigma}A\right\}, \mathrm{tr}\left\{\hat{\Sigma}B\right\}\right) \\
&= \mathrm{E}\left[\mathrm{tr}\left\{(\hat{\Sigma}-\Sigma)A\right\}\mathrm{tr}\left\{(\hat{\Sigma}-\Sigma)B\right\}\right] \\
&= \mathrm{E}\left[\left(\sum_i\sum_j(\hat{\Sigma}-\Sigma)_{ij}A_{ji}\right)\left(\sum_k\sum_l(\hat{\Sigma}-\Sigma)_{kl}B_{lk}\right)\right] \\
&= \sum_i\sum_j\sum_k\sum_l A_{ji}\mathrm{E}\left[(\hat{\Sigma}-\Sigma)_{ij}(\hat{\Sigma}-\Sigma)_{kl}\right]B_{lk} \\
&= \sum_i\sum_j\sum_k\sum_l A_{ji}Cov(Z_iZ_j, Z_kZ_l)B_{lk} \\
&= \sum_i\sum_j\sum_k\sum_l A_{ji}(\mathrm{E}\left[Z_iZ_jZ_kZ_l\right] - \mathrm{E}\left[Z_iZ_j\right]\mathrm{E}\left[Z_kZ_l\right])B_{lk} \\
&= \sum_i\sum_j\sum_k\sum_l A_{ji}(\Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk} - \Sigma_{ij}\Sigma_{kl})B_{lk} \\
&= \sum_i\sum_j\sum_k\sum_l A_{ij}\Sigma_{jl}B_{lk}\Sigma_{ki} + \sum_i\sum_j\sum_k\sum_l A_{ij}\Sigma_{jk}B_{kl}\Sigma_{li} \\
&= 2\mathrm{tr}\left\{\Sigma A\Sigma B\right\},
\end{aligned}$$

□

The Fisher information matrix $\mathcal{E}(U) = \mathrm{E}\left[UU'\right]$ has i, j^{th} element

$$\begin{aligned}
\mathrm{E}\left[UU'\right]_{i,j} &= \frac{1}{4}\mathrm{E}\left[\left(Z'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}Z - \mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\right\}\right)\left(Z'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}Z - \mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right\}\right)\right] \\
&= \frac{1}{4}\mathrm{Cov}\left(Z'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}Z, Z'\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}Z\right) \\
&= \frac{1}{2}\mathrm{tr}\left\{\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\right)\Sigma\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}\right)\Sigma\right\} \\
&= \frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right\}. \qquad (27)
\end{aligned}$$

Here we have applied Lemma 1 with $A = \Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}$ and $B = \Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}\Sigma^{-1}$.

Now we calculate $\mathcal{E}(G) = \mathrm{E}\left[\dot{G}\right]'\mathrm{E}\left[GG'\right]^{-1}\mathrm{E}\left[\dot{G}\right]$. We start with $\mathrm{E}\left[\dot{G}\right]$, differentiating the first and second terms of (14) separately. First,

$$\frac{\partial}{\partial\theta_j}\mathrm{tr}\left\{\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\right\} = -\mathrm{tr}\left\{\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_j}\circ T\right]\right\} + \mathrm{tr}\left\{\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial^2\Sigma}{\partial\theta_i\partial\theta_j}\circ T\right]\right\}.$$

Note this expression is a constant with respect to the data. Next,

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}\mathrm{tr}\left\{\left[\hat{\Sigma}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\left[\Sigma\circ T\right]^{-1}\right\} &= -\mathrm{tr}\left\{\left[\hat{\Sigma}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_j}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\left[\Sigma\circ T\right]^{-1}\right\} \\
&\quad +\mathrm{tr}\left\{\left[\hat{\Sigma}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial^2\Sigma}{\partial\theta_i\partial\theta_j}\circ T\right]\left[\Sigma\circ T\right]^{-1}\right\} \\
&\quad -\mathrm{tr}\left\{\left[\hat{\Sigma}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_i}\circ T\right]\left[\Sigma\circ T\right]^{-1}\left[\frac{\partial\Sigma}{\partial\theta_j}\circ T\right]\left[\Sigma\circ T\right]^{-1}\right\},
\end{aligned}$$

which has expected value

$$\mathrm{tr}\left\{[\Sigma \circ T]^{-1}\left[\frac{\partial^2 \Sigma}{\partial\theta_i \partial\theta_j} \circ T\right]\right\} - 2\mathrm{tr}\left\{[\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right][\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_j} \circ T\right]\right\}.$$

Putting the summands back together, we have

$$\mathrm{E}\left[\dot{G}_{i,j}\right] = -\frac{1}{2}\mathrm{tr}\left\{\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right][\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_j} \circ T\right][\Sigma \circ T]^{-1}\right\}. \tag{28}$$

Next we calculate $\mathrm{E}\left[GG'\right]$. Here, we note that for any conformable matrix C, $\mathrm{tr}\left\{\left[\hat{\Sigma} \circ T\right]C\right\} = \mathrm{tr}\left\{\hat{\Sigma}[C \circ T]\right\} = Z'[C \circ T]Z$, so Lemma 1 also implies that $Cov\left(\mathrm{tr}\left\{[\hat{\Sigma} \circ T]A\right\}, \mathrm{tr}\left\{[\hat{\Sigma} \circ T]B\right\}\right) = 2\mathrm{tr}\left\{[A \circ T]\Sigma[B \circ T]\Sigma\right\}$. Thus, the i, j^{th} entry of $\mathrm{E}\left[GG'\right]$ is

$$\begin{aligned}
\mathrm{E}\left[GG'\right]_{i,j} &= \frac{1}{4}\mathrm{E}\left[\left(\mathrm{tr}\left\{\left[\hat{\Sigma} \circ T\right][\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right][\Sigma \circ T]^{-1}\right\} - \mathrm{tr}\left\{[\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right]\right\}\right)\right. \\
&\quad \left.\left(\mathrm{tr}\left\{\left[\hat{\Sigma} \circ T\right][\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_j} \circ T\right][\Sigma \circ T]^{-1}\right\} - \mathrm{tr}\left\{[\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_j} \circ T\right]\right\}\right)\right] \\
&= \frac{1}{4}\mathrm{Cov}\left(\mathrm{tr}\left\{\left[\hat{\Sigma} \circ T\right][\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right][\Sigma \circ T]^{-1}\right\}, \mathrm{tr}\left\{\left[\hat{\Sigma} \circ T\right][\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_j} \circ T\right][\Sigma \circ T]^{-1}\right\}\right) \\
&= \frac{1}{2}\mathrm{tr}\left\{\left[\left([\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right][\Sigma \circ T]^{-1}\right) \circ T\right]\Sigma\left[\left([\Sigma \circ T]^{-1}\left[\frac{\partial \Sigma}{\partial\theta_i} \circ T\right][\Sigma \circ T]^{-1}\right) \circ T\right]\Sigma\right\}.
\end{aligned}$$

References

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation. Technical Report 917, Norwegian Computing Center, Oslo, Norway.

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

Akaike, H. (1973). Block Toeplitz matrix inversion. *SIAM Journal of Applied Mathematics*, 14:234–241.

Alcamo, J., editor (1994). *IMAGE 2.0: Integrated modeling of global climate change*. Kluwer.

Alcamo, J., Leemans, R., and Kreileman, E., editors (1998). *Global comate change scenarios of the* 21^{st} *century. Results from the IMAGE 2.1 model*. Pergamon.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall.

Berk, R., Fovell, R., Schoenberg, F., and Weiss, R. (2001). The use of statistical tools for evaluating computer simulations – An editorial essay. *Climatic Change*, 51(2):119–130.

Billingsley, P. (1986). *Probability and Measure*. John Wiley & Sons.

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. University of California Press.

Bonan, G. (1996). A land surface model (LSM version 1.0) for ecological, hydrological, and atmospheric studies: Technical description and user's guide. Technical Report NCAR/TN-417+STR, National Center for Atmospheric Research.

Bonan, G. (1998). The land surface climatology of the NCAR land surface model coupled to the NCAR Community Climate Model. *Journal of Climate*, 11:1307–1326.

Caragea, P. C. (2003). *Approximate likelihoods for spatial processes*. PhD thesis, University of North Carolina, Chapel Hill.

Cook, D. and Pocock, S. (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*, 39:361–371.

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

Cressie, N. and Lahiri, S. (1996). Asymptotics for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference*, 50:327–341.

Cuzick, J. (1995). A strong law for weighted sums of i.i.d. random variables. *Journal of Theoretical Probability*, 8:625–641.

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. *Annals of Statisics*, 28(6):1762–1794.

Diggle, P., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Applied Statistics*, 47:299–326.

Ehm, W., Gneiting, T., and Richards, D. (2004). Convolution roots of radial positive definite functions with compact support. *Transactions of the American Mathematical Society*, 356:4655–4685.

Eide, A., Omre, H., and Ursin, B. (2002). Prediction of reservoir variables based on seismic data and well observations. *Journal of the American Statistical Association*, 97:18–28.

Feddema, J., Oleson, K., Bonan, G., Mearns, L., Washington, W., Meehl, G., and Nychka, D. (2005). A comparison of a GCM response to historical anthropogenic land cover change and model sensitivity to uncertainty in present-day land cover representations. *Climate Dynamics*. Under revision.

Fuentes, M. (2004). Spectral methods to approximate the likelihood for irregularly spaced spatial data. Technical Report 2568, North Carolina State University, Institute of Statistics.

Fuentes, M. and Raftery, A. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61:36–45.

Furrer, R., Genton, M. G., and Nychka, D. (2005). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*. Under revision.

Gaspari, G. and Cohn, S. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125:723–757.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83:493–508.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69:95–105.

Handcock, M. and Stein, M. (1993). A Bayesian analysis of kriging. *Technometrics*, 36:403–410.

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89:368–378.

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

Holland, D., Cox, W. M., Scheffe, R., Cimorelli, A. J., Nychka, D., and Hopke, P. K. (2003). Spatial prediction of air quality data. *Environmental Manager*, August 2003:31–35.

Horn, R. and Johnson, C. (1991). *Topics in matrix analysis*. Cambridge University Press.

Kitanidis, P. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, 19:909–921.

Kitanidis, P. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, 22:499–507.

Koenker, R. and Ng, P. (2005). *The SparseM Package*. Available from http://www.econ.uiuc.edu/ roger/research/sparse/sparse.html.

Krige, B. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52:119–139.

Livezey, R. and Chen, W. (1983). Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review*, 111:46–59.

Mardia, K. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146.

Mardia, K. and Watkins, A. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76:289–295.

Matérn, B. (1986). *Spatial Variation*. Springer-Verlag, second edition.

Matheron, G. (1971). *The Theory of Regionalized Variables and its Applications*. Ecole des Mines, Fontainebleau.

McGuffie, K. and Henderson-Sellers, A. (2005). *A Climate Modelling Primer*. Wiley, third edition.

Ng, E. and Peyton, B. (1993). Block sparse cholesky algorithms on advanced uniprocessor computers. *SIAM Journal on Scientific Computing*, 14:1034–1056.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization.* Springer.

O'Hagan, A. (2004). Bayesian analysis of computer code outputs: A tutorial. Technical Report 543/04, Department of Probability and Statistics, University of Sheffield.

Pardo-Igúzquiza, E. (1998). Maximum likelihood estimation of spatial covariance parameters. *Mathematical Geology*, 30(1):95–108.

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

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1992). *Numerical Recipes.* Cambridge University Press, 2 edition.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4:409–435.

Stein, M. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters*, 17.

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. (1995). Fixed-domain asymptotics for spatial periodograms. *Journal of the American Statistical Association*, 90:1277–1288.

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.

Warnes, J. and Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial gaussian processes. *Biometrika*, 74:640–642.

Washington, W., Weatherly, J., Meehl, G., Semtner, A., Bettge, T., Craig, A., Strand, W., Arblaster, J., Wayland, V., James, R., and Zhang, Y. (2000). Parallel climate model (PCM) control and transient simulations. *Climate Dynamics*, 16:755–774.

Watkins, A. (1990). On polynomial expansions for likelihoods of spatial data with finite range correlation function. *Biometrika*, 77:404–408.

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.

Yaglom, A. (1987). *Correlation theory of stationary and related random functions.* Springer-Verlag.

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

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.