




Statistical Properties of Covariance Tapers

Michael L. Stein


To cite this article: Michael L. Stein (2013) Statistical Properties of Covariance Tapers, Journal of Computational and Graphical Statistics, 22:4, 866-885, DOI: [10.1080/10618600.2012.719844](https://doi.org/10.1080/10618600.2012.719844)

To link to this article: <https://doi.org/10.1080/10618600.2012.719844>

 View supplementary material 

 Published online: 21 Oct 2013.

 Submit your article to this journal 

 Article views: 532

 View related articles 

 Citing articles: 16 View citing articles 



Statistical Properties of Covariance Tapers

Michael L. STEIN

Compactly supported autocovariance functions reduce computations needed for estimation and prediction under Gaussian process models, which are commonly used to model spatial and spatial-temporal data. A critical issue in using such models is the loss in statistical efficiency caused when the true autocovariance function is not compactly supported. Theoretical results indicate the value of specifying the local behavior of the process correctly. One way to obtain a compactly supported autocovariance function that has similar local behavior to an autocovariance function K of interest is to multiply K by some smooth compactly supported autocovariance function, which is called covariance tapering. This work extends previous theoretical results showing that covariance tapering has some asymptotic optimality properties as the number of observations in a fixed and bounded domain increases. However, numerical experiments show that for purposes of parameter estimation, covariance tapering often does not work as well as the simple alternative of breaking the observations into blocks and ignoring dependence across blocks. When covariance tapering is used for spatial prediction, predictions near the boundary of the observation domain are affected most. This article proposes an approach to modifying the taper to ameliorate this edge effect. In addition, a justification for a specific approach to carrying out conditional simulations based on tapered covariances is given. Supplementary materials for this article are available online.

Key Words: Equivalence of Gaussian measures; Fixed-domain asymptotics; Kriging; Sparse matrices; Unbiased estimating equations.

1. INTRODUCTION

Gaussian process models are commonly used for the statistical analysis of large spatial and spatial-temporal datasets. In many applications, prediction of the process at unobserved locations is an important goal. If the covariance structure of the process is known and the mean is known up to some vector of linear parameters, then best linear unbiased prediction, also known as universal kriging, is the method of choice. For simplicity, the focus here is on simple kriging, in which the mean of the process is known to be 0, but similar computational considerations apply to universal kriging. Computing the simple kriging predictor requires solving a system of linear equations with coefficient matrix C given by the covariance matrix

Michael L. Stein is Professor, Department of Statistics, University of Chicago, Chicago, IL 60637 (E-mail: stein@galton.uchicago.edu).

© 2013 *American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America*
Journal of Computational and Graphical Statistics, Volume 22, Number 4, Pages 866–885
DOI: [10.1080/10618600.2012.719844](https://doi.org/10.1080/10618600.2012.719844)

of the vector of observations. For spatial data, nonnegligible correlation is often observed at distances that are a substantial fraction of the size of the observation domain, so that covariance matrices in spatial statistics are often dense. Furthermore, if the observations are not on a grid, then C generally does not have any exploitable structure and kriging requires solving a system of linear equations with a dense, unstructured matrix. If n is the number of observations, an exact solution of this problem requires $O(n^3)$ calculations and $O(n^2)$ memory.

In most applications, the covariance structure of the process is at least partially unknown and must be estimated from the available observations. Assuming that the process is Gaussian with a parametric model for the covariance structure, likelihood methods (e.g., maximum likelihood or Bayesian) are appropriate. The likelihood function can be readily evaluated using the Cholesky decomposition of C . Likelihood methods generally require many evaluations of the likelihood, necessitating a large number of Cholesky decompositions.

There are many approaches to reducing the computational burden of the required linear algebra, but one popular choice is to use a sparse covariance matrix. Suppose that the process under study, Z , defined on \mathbb{R}^d , has covariance function $K(\mathbf{x}, \mathbf{y}) = \text{cov}\{Z(\mathbf{x}), Z(\mathbf{y})\}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. If, as is generally the case in this work, K depends on (\mathbf{x}, \mathbf{y}) only through $\mathbf{x} - \mathbf{y}$, then it is called the autocovariance function, or acf, for Z . Many commonly used acfs are positive everywhere so that covariance matrices under these models will be dense. What are the consequences of replacing K by a covariance function \tilde{K} that is 0 for \mathbf{x} and \mathbf{y} sufficiently distant? Of course, the answer depends on many factors, but certainly the choice of \tilde{K} is one of them. One possibility is to substitute some compactly supported acf \tilde{K} for K . Another approach that may have broader scope is covariance tapering (Furrer, Genton, and Nychka 2006): suppose K_t is a covariance function that is 0 for all \mathbf{x} and \mathbf{y} sufficiently distant and define \tilde{K} by $\tilde{K}(\mathbf{x}, \mathbf{y}) = K(\mathbf{x}, \mathbf{y})K_t(\mathbf{x}, \mathbf{y})$, which is a valid covariance function since it is the product of valid covariance functions (Stein 1999, p. 20). If K_t only depends on (\mathbf{x}, \mathbf{y}) through the interpoint distance $|\mathbf{x} - \mathbf{y}|$, then the taper is isotropic and the largest distance beyond which it is 0 is the range of the taper. Neither K nor K_t is required to be an acf in this formulation and, indeed, as noted by Anderes et al. (2011), one may not want to take K_t to be an acf even when K is.

Section 2 reviews and extends past theoretical results showing that tapering can have good asymptotic properties in terms of prediction and likelihood approximation for a taper with fixed range and an increasing number of observations in a fixed and bounded region. The subsequent sections detail the results of several numerical experiments and a small simulation study to assess how well tapering performs in a broad range of circumstances.

Section 3 considers the effect of tapering on parameter estimation. One important distinction between approaches is whether they are based on unbiased estimating equations. Kaufman, Schervish, and Nychka (2008) described one way to obtain unbiased estimating equations using tapered covariance matrices and Section 3 describes another. Section 3.1 examines the simple setting in which the covariance matrix of the observations is specified up to a multiplicative constant, for which the means and variances of the resulting estimates all have closed forms. The focus is on two commonly used tapers, Bohman and Wendland₁. In the examples considered, the Bohman taper dominates the Wendland₁ taper and the new approach to obtaining unbiased estimating equations works considerably better than that proposed in Kaufman, Schervish, and Nychka (2008). However, even the combination of

the Bohman taper and the new estimating equations is easily outperformed by dividing the observations into blocks and assuming independence across blocks. For the approaches based on unbiased estimating equations, Section 3.2 gives results when, in addition to an unknown multiplicative constant, the covariance function has an unknown range parameter. Exact properties of the estimates are no longer available, but one can compare estimates based on a standard information measure for unbiased estimating equations as well as by simulation. Both the information measure and the simulations show that the Bohman taper combined with the estimating equations proposed by Kaufman, Schervish, and Nychka (2008) is much more competitive with the block method when the range is unknown. In fact, the tapered estimates of the multiplicative constant are much better when the range is unknown than when it is known. This improvement in the estimates of one parameter when another unknown parameter is added to the model is noteworthy and complicates the comparison of estimation methods. Nevertheless, it does appear that, overall, blocking may be preferable to covariance tapering for parameter estimation.

Section 4 examines tapering and spatial prediction. A critical component of any statistical approach to prediction is to give prediction uncertainties. An attractive approach to quantifying uncertainties is through conditional simulations of the process at unobserved locations of interest given the observations. Section 4.1 describes how to approximate conditional simulations efficiently and effectively based on kriging predictors using a tapered covariance function. Tapering generally has a greater impact on the efficiency of predictors near or outside the boundary of the observation domain than in its interior. Section 4.2 describes a method for deforming a taper near the boundary of the observation domain to address this problem. Section 4.3 gives some numerical results on the efficiency of linear predictors based on tapered covariance functions. Here again, the Bohman taper usually but not always outperforms the Wendland₁ taper. Deforming the Bohman taper can lead to dramatic improvements in the efficiency of predictors near the boundary of the observation domain with little effect in the interior.

Section 5 briefly considers the impact of a nugget effect (a discontinuity at the origin in the acf) and of including an unknown constant mean in the model on tapering. Section 6 summarizes the implications of this work on the use of covariance tapering for reducing computation in spatial statistics. R code for the numerical results and proofs are given in the online supplementary materials.

2. THEORETICAL RESULTS

This section reviews past results and gives some new ones on asymptotically optimal (minimum mean squared error, MSE) linear predictions and equivalent Gaussian measures as they relate to covariance tapering. The following result on asymptotically optimal prediction with a misspecified spectral density does not specifically refer to tapering but will be connected to tapered acfs by Theorem 1. See Stein (1999) for further results on optimal linear prediction with misspecified models.

Let D be some bounded subset of \mathbb{R}^d ; $\mathbf{x}_1, \mathbf{x}_2, \dots$ a dense sequence of points in D ; and let $D_{-n} = D \setminus \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. Define $e(\mathbf{x}, n)$ to be the error of the optimal linear predictor of $Z(\mathbf{x})$ under spectral density f based on $(Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_n))$. For another spectral density

\tilde{f} , let $\tilde{e}(\mathbf{x}, n)$ be the error of the optimal linear predictor under \tilde{f} . Write E for expectations under f and \tilde{E} for expectations under \tilde{f} . Thus, we always have $E\tilde{e}(\mathbf{x}, n)^2 \geq Ee(\mathbf{x}, n)^2$. Next, let \mathcal{Q}^d be the space of real-valued functions f on \mathbb{R}^d such that $f(\boldsymbol{\omega})/|\zeta(\boldsymbol{\omega})|^2$ is bounded away from 0 and ∞ as $|\boldsymbol{\omega}| \rightarrow \infty$ for some function ζ that is the Fourier transform of a square integrable function with bounded support. For example, if $f(\boldsymbol{\omega})|\boldsymbol{\omega}|^\alpha$ is bounded away from 0 and ∞ as $|\boldsymbol{\omega}| \rightarrow \infty$ for some $\alpha > d$, then $f \in \mathcal{Q}^d$. If $f \in \mathcal{Q}^d$ and

$$\lim_{|\boldsymbol{\omega}| \rightarrow \infty} \frac{\tilde{f}(\boldsymbol{\omega})}{f(\boldsymbol{\omega})} = c \quad (1)$$

for $c \in (0, \infty)$, then (Stein 1999, p. 130)

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x} \in D_{-n}} \frac{E\tilde{e}(\mathbf{x}, n)^2}{Ee(\mathbf{x}, n)^2} = 1 \quad (2)$$

and

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x} \in D_{-n}} \left| \frac{\tilde{E}\tilde{e}(\mathbf{x}, n)^2}{E\tilde{e}(\mathbf{x}, n)^2} - c \right| = 0. \quad (3)$$

When $c = 1$, these results imply that if f is the correct model and \tilde{f} is another model, in terms of the efficiency of linear predictors within the region D or in terms of evaluation of their MSEs, there is asymptotically negligible loss in using \tilde{f} as if it were the truth.

Furrer, Genton, and Nychka (2006) showed that (1) and hence (2) and (3) hold when appropriate tapers are applied to the commonly used Matérn acfs (Stein 1999). For an isotropic process in d dimensions, the spectral density for a Matérn acf is of the form $f(\boldsymbol{\omega}) = \phi(\rho^{-2} + |\boldsymbol{\omega}|^2)^{-\kappa-d/2}$ for positive constants ϕ , ρ , and κ . The corresponding isotropic acf is

$$K(\mathbf{x}) = \frac{\pi^{d/2} \rho^{2\kappa} \phi}{2^{\kappa-1} \Gamma(\kappa + \frac{1}{2}d)} \left(\frac{|\mathbf{x}|}{\rho} \right)^\kappa K_\kappa \left(\frac{|\mathbf{x}|}{\rho} \right),$$

where K_κ is a modified Bessel function. Define $\mathcal{M}_\kappa(r) = r^\kappa K_\kappa(r)$ for $r \geq 0$. The parameter κ controls the smoothness of the Matérn process, which has exactly m mean square derivatives in any direction if and only if $\kappa > m$. The parameter ρ controls the rate of decay of the correlations with distance; call ρ the range parameter of the model even though, unlike the range of a compactly supported acf, it is of course not a distance beyond which the acf equals 0. The spectral density corresponding to the product of two acfs with spectral densities f and f_i is the convolution $f * f_i$. For an integrable function h on \mathbb{R}^d , define $I(h) = \int_{\mathbb{R}^d} h(\boldsymbol{\omega}) d\boldsymbol{\omega}$. Proposition 1 in Furrer, Genton, and Nychka (2006) shows that if f is a Matérn spectral density and $f_i(\boldsymbol{\omega})(1 + |\boldsymbol{\omega}|^2)^{\kappa+d/2+\delta}$ is bounded for some $\delta > 0$, then f and $\tilde{f} = f * f_i$ satisfy (1) for $c = I(f_i)$ (the result mistakenly allows $\delta = 0$).

The basic idea underlying this result is that if $f_i(\boldsymbol{\omega})$ decreases more rapidly than $f(\boldsymbol{\omega})$ as $\boldsymbol{\omega} \rightarrow \infty$, then (1) should hold. It is not essential that f be a Matérn spectral density, as the following theorem, proven in the online supplementary material, shows:

Theorem 1. Suppose f is an integrable, nonnegative function on \mathbb{R}^d and there exists a monotonic function h on $[0, \infty)$ such that $f(\boldsymbol{\omega}) \leq h(|\boldsymbol{\omega}|)$, $h(|\boldsymbol{\omega}|)/f(\boldsymbol{\omega})$ is bounded for all $|\boldsymbol{\omega}|$ sufficiently large and $h(r)/h(2r)$ is bounded for all r sufficiently large. In addition,

suppose, for all finite s ,

$$\lim_{\omega \rightarrow \infty} \sup_{|\mathbf{v}| < s} \left| \frac{f(\omega + \mathbf{v})}{f(\omega)} - 1 \right| = 0 \quad (4)$$

and f_t is a bounded nonnegative function on \mathbb{R}^d satisfying $f_t(\omega)/f(\omega) \rightarrow 0$ as $\omega \rightarrow \infty$. Then

$$\lim_{\omega \rightarrow \infty} \frac{(f * f_t)(\omega)}{f(\omega)} = I(f_t). \quad (5)$$

Thus, if in addition to the conditions of Theorem 1, $f \in \mathcal{Q}^d$, then (2) and (3) hold. Note that (4) also appears in theoretical results in Stein (2012a), where it is argued that spectral densities that do not satisfy this condition should generally not be used in practice. The conditions on f are quite weak; they are satisfied if, for example, for some function g on $[0, \infty)$ that is regularly varying at ∞ (Bingham, Goldie, and Teugels 1987), $f(\omega)/g(|\omega|)$ is bounded away from 0 and ∞ outside some neighborhood of the origin.

If, for $f \in \mathcal{Q}^d$, we have the (in practice) stronger condition than (1) that

$$\int_{|\omega| > R} \left\{ \frac{\tilde{f}(\omega) - f(\omega)}{f(\omega)} \right\}^2 d\omega < \infty \quad (6)$$

for some $R > 0$, then the Gaussian measures for the processes on any bounded domain D under the spectral densities f and \tilde{f} are equivalent. Write $G_D(K)$ or $G_D(f)$ as convenient to refer to the probability measure for the Gaussian process on D with mean 0 and acf K (or spectral density f) and use the symbol \equiv to indicate equivalence of measures. For two joint densities p and \tilde{p} on some set of observations, the Kullback divergence of \tilde{p} from p is $E \log(p/\tilde{p})$, where the expectation is under the density p . One implication of $G_D(f) \equiv G_D(\tilde{f})$ is that the Kullback divergences between the multivariate normal distributions over all finite subsets of points in D induced by the spectral densities f and \tilde{f} are bounded (Ibragimov and Rozanov 1978, sec. 3.2). This result suggests that if $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ and $\tilde{\mathcal{F}} = \{\tilde{f}_\psi : \psi \in \Psi\}$ are two classes of spectral densities such that $f_\theta \in \mathcal{Q}^d$ for all $\theta \in \Theta$ and, for every $\theta \in \Theta$ there exists a $\psi \in \Psi$ such that (6) holds with $f = f_\theta$ and $\tilde{f} = \tilde{f}_\psi$, then for n sufficiently large, it should not affect likelihood calculations much if one uses the model $\tilde{\mathcal{F}}$ rather than the correct \mathcal{F} . Indeed, Kaufman, Schervish, and Nychka (2008); Du, Zhang, and Mandrekar (2009); Kaufman and Shaby (2011); and Wang and Loh (2011) gave limited results in this regard for tapering as applied to Matérn acfs. In particular, Wang and Loh (2011) obtained results in the setting when the range of the taper function is allowed to decrease as the sample size increases.

Kaufman, Schervish, and Nychka (2008, theorem 1) gave the following result proving that covariance tapering applied to Matérn acfs can yield equivalent Gaussian measures:

Theorem 2. Consider a process Z on a bounded domain $D \subset \mathbb{R}^d$ with $d \leq 3$. Let K be a Matérn acf with smoothness parameter κ and let $\tilde{K} = KK_t$, where K_t is the acf on \mathbb{R}^d corresponding to the isotropic spectral density f_t with $I(f_t) = 1$. If there exist constants M and δ such that $f_t(\omega) \leq M(1 + |\omega|^2)^{-\kappa - d/2 - \delta}$ on \mathbb{R}^d with $\delta > \max(\frac{d}{4}, 1 - \kappa)$, then $G_D(K) \equiv G_D(\tilde{K})$.

This result says that if f is the spectral density for the Matérn acf K and $f_t(\omega)/f(\omega)$ tends to 0 sufficiently quickly as $\omega \rightarrow \infty$ and $I(f_t) = 1$, then tapering induces an equivalent Gaussian measure on any bounded region. Just as Theorem 1 extends Proposition 1 in Furrer, Genton, and Nychka (2006) beyond Matérn models, it is possible to generalize Theorem 2 to models for f beyond the Matérn. By the transitivity of equivalence, if K_1 is some other acf on \mathbb{R}^d for which $G_D(K) \equiv G_D(K_1)$, then $G_D(KK_t) \equiv G_D(K_1)$ for K and K_t as in Theorem 2. If one were using tapering to replace K_1 by a covariance function with compact support, it would be more natural to use K_1K_t rather than KK_t . The following result extends Theorem 2 to cover this case.

Theorem 3. Suppose the conditions of Theorem 2 hold for K and K_t with corresponding spectral densities f and f_t , respectively. In addition, suppose K_1 is an acf with spectral density f_1 and there exists a nonincreasing function g on $[0, \infty)$ and finite R such that $|f_1(\omega) - f(\omega)| \leq g(|\omega|)$ with

$$\int_{|\omega| > R} \left\{ \frac{g(|\omega|)}{f(\omega)} \right\}^2 d\omega < \infty. \quad (7)$$

Then $G_D(K_1) \equiv G_D(K_1K_t)$.

The condition (7) is sufficient to prove $G_D(K) \equiv G_D(K_1)$ and since Theorem 2 implies $G_D(K) \equiv G_D(KK_t)$, it suffices to show $G_D(KK_t) \equiv G_D(K_1K_t)$, which is proven in the online supplementary material. The condition (7) is slightly stronger than what is needed to prove $G_D(K) \equiv G_D(KK_t)$ in that the function g is required to be monotonic, but this restriction is quite mild and it appears possible to choose g to make (7) hold for all models used in practice for which (6) is true.

3. PARAMETER ESTIMATION

Kaufman, Schervish, and Nychka (2008) showed how covariance tapering can be used to approximate maximum likelihood estimates for unknown parameters of the covariance function. They described two possible approximations and this section gives a third. Write $C(\theta)$ for the covariance matrix of a set of n observations \mathbf{Z} as a function of an unknown parameter $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$ and T for a taper matrix, which can be any positive semidefinite matrix with ones on its diagonal. Writing $A \circ B$ for the elementwise product of commensurate matrices A and B , define $\tilde{C}(\theta) = C(\theta) \circ T$. Then the first approximation in Kaufman, Schervish, and Nychka (2008) to the log-likelihood is

$$\ell_1(\theta) = -\frac{1}{2} \log |\tilde{C}(\theta)| - \frac{1}{2} \mathbf{Z}' \tilde{C}(\theta)^{-1} \mathbf{Z}. \quad (8)$$

Kaufman, Schervish, and Nychka (2008) showed that estimating θ by maximizing $\ell_1(\theta)$ can sometimes lead to estimates with substantial bias (Kaufman, Schervish, and Nychka 2008), so they proposed

$$\ell_2(\theta) = -\frac{1}{2} \log |\tilde{C}(\theta)| - \frac{1}{2} \mathbf{Z}' \{\tilde{C}(\theta)^{-1} \circ T\} \mathbf{Z}. \quad (9)$$

Like the exact log-likelihood, setting the gradient of (9) equal to 0 yields a set of unbiased estimating equations in that $E_{\theta} \frac{\partial}{\partial \theta_i} \ell_2(\theta) = 0$ for all θ and all i . One can also generate unbiased estimating equations directly, without reference to optimizing some criterion function. For example, writing $\tilde{C}_i(\theta)$ for $\frac{\partial}{\partial \theta_i} \tilde{C}(\theta)$, we have $\frac{\partial}{\partial \theta_i} \mathbf{Z}' \tilde{C}(\theta)^{-1} \mathbf{Z} = -\mathbf{Z}' \tilde{C}(\theta)^{-1} \tilde{C}_i(\theta) \tilde{C}(\theta)^{-1} \mathbf{Z}$, which suggests the following unbiased estimating equations:

$$\frac{1}{2} \mathbf{Z}' \tilde{C}(\theta)^{-1} \tilde{C}_i(\theta) \tilde{C}(\theta)^{-1} \mathbf{Z} - \frac{1}{2} \text{tr}\{\tilde{C}(\theta)^{-1} \tilde{C}_i(\theta) \tilde{C}(\theta)^{-1} C(\theta)\} = 0 \quad (10)$$

for $i = 1, \dots, p$.

As a comparison, consider just partitioning the observations into some collection of blocks and approximating the log-likelihood by summing the log-likelihood for each block, which acts as if the responses in different blocks are independent. This procedure can be interpreted as a taper where the taper matrix has (i, j) th entry equal to 1 if observations i and j are in the same block and 0 otherwise, which is a valid correlation matrix. Thus, I will call this procedure a block taper. For the block taper, it is not difficult to show that (8) and (9) are the same and that setting the gradient of these approximate likelihoods to $\mathbf{0}$ gives (10), so that if (8) is maximized at a unique extreme point in the interior of the parameter space, there is no difference between the three methods.

Before examining the statistical properties of these estimates, it is worthwhile to compare their computational requirements. Computing $\ell_1(\theta)$ requires a single sparse Cholesky decomposition followed by a single (sparse) backsolve to get the quadratic form and a sum of the logarithms of the diagonal elements of the Cholesky decomposition to get the log determinant. Computing $\ell_2(\theta)$ appears to require explicit evaluation of $\tilde{C}(\theta)^{-1}$ and, hence, somewhat more effort (roughly double) than $\ell_1(\theta)$. Computing the left-hand side of (10) via the Cholesky decomposition of $\tilde{C}(\theta)$ requires $O(n)$ sparse backsolves for each i , and this may involve more effort than the Cholesky decomposition itself depending on the sparseness of the decomposition. However, it is not necessary to use the Cholesky decomposition in this case because (10) does not require $\log |\tilde{C}(\theta)|$. In particular, because (10) only requires solves in the positive definite matrix $\tilde{C}(\theta)$, it is natural to use iterative methods such as preconditioned conjugate gradient (Golub and van Loan 1996, sec. 10.3) to do the calculations. Of course, one could just consider the gradients of $\ell_1(\theta)$ and $\ell_2(\theta)$ and similarly avoid computing $\log |\tilde{C}(\theta)|$, which is the basic idea behind the approach in Anitescu, Chen, and Wang (2012) for finding maximum likelihood estimates. Note, though, that solving a system of nonlinear equations is not the same as maximizing a function and, to the extent that solving the nonlinear equations is more difficult, using (10) to estimate θ may require more iterations than maximizing (8) or (9).

For the block taper, the log-likelihood is just the sum of the log-likelihoods for each block. Thus, unlike general sparse matrices, there is no need to permute the rows and columns of the covariance matrix to get an efficient sparse decomposition (Davis 2006, chap. 4). For very large n and isotropic tapers, it may be difficult to store $\tilde{C}(\theta)$ or its Cholesky decomposition even when the matrices are sparse. For the block taper, because calculations can be done separately for each block, there is no need to store matrices whose dimensions are larger than the block size.

3.1 UNKNOWN SCALE PARAMETER

When $C(\theta) = \theta C$, so that C is known up to the scale parameter θ , there are simple closed forms for all of the resulting estimates and their means and variances. The exact maximum likelihood estimate, given by $\hat{\theta} = \frac{1}{n} \mathbf{Z}' C^{-1} \mathbf{Z}$, has mean θ and variance $\frac{2\theta^2}{n}$. Defining $\tilde{C} = C \circ T$, the maximizer of (8) is

$$\hat{\theta}_1 = \frac{1}{n} \mathbf{Z}' \tilde{C}^{-1} \mathbf{Z}, \quad (11)$$

which has mean $\frac{\theta}{n} \text{tr}(\tilde{C}^{-1} C)$ and variance $\frac{2\theta^2}{n^2} \text{tr}\{(\tilde{C}^{-1} C)^2\}$. The maximizer of (9) is

$$\hat{\theta}_2 = \frac{1}{n} \mathbf{Z}' (\tilde{C}^{-1} \circ T) \mathbf{Z}, \quad (12)$$

which has mean θ and variance $\frac{2\theta^2}{n^2} \text{tr}\{(\tilde{C}^{-1} \circ T) C\}^2$. Finally, the solution of (10) is

$$\hat{\theta}_3 = \frac{\mathbf{Z}' \tilde{C}^{-1} \mathbf{Z}}{\text{tr}(\tilde{C}^{-1} C)}, \quad (13)$$

with mean θ and variance $2\theta^2 \text{tr}\{(\tilde{C}^{-1} C)^2\} / \{\text{tr}(\tilde{C}^{-1} C)\}^2$.

Table 1 gives the variances (relative to that of the mle $\hat{\theta}$) of these three estimates of θ (and the mean for the biased estimate $\hat{\theta}_1$) under two Matérn models with range of 0.25. The observations are taken at the locations $\frac{1}{40}(i - 0.5 + X_{ij}, j - 0.5 + Y_{ij})$ for $i, j \in 1, 2, \dots, 40$ and the X_{ij} 's and Y_{ij} 's iid uniform on $(-0.4, 0.4)$ for a total of 1600 observations. This choice of random perturbation from a grid produces quite a bit of variation in the distances between neighboring observations but maintains a minimum distance between observations of $0.2/40$, which avoids numerical instabilities in the covariance matrices. For observations on a grid, covariance tapering is of less practical interest, since the covariance matrix then has structure that can be exploited to reduce memory and computations even when the covariance matrix is dense. Isotropic tapers depend only on interpoint distance r . The two we examine here are the Wendland₁ (W_1) (Wendland 1995; Furrer, Genton, and Nychka 2006) and Bohman (B) (Bohman 1960; Gneiting 2002) tapers. For

Table 1. Properties of estimates (11)–(13) of θ under exponential ($\theta e^{-4|x|}$, top three rows) and Whittle ($\theta \mathcal{M}_1(4|x|)$, last three rows) acfs with 1600 observations in $[0, 1]^2$ when $\theta = 1$

Taper	$E(\hat{\theta}_1)$	$\text{var}(\hat{\theta}_1)$	$\text{var}(\hat{\theta}_2)$	$\text{var}(\hat{\theta}_3)$	Divergence
Bohman	0.6828	1.113	10.80	2.388	146.1
Wendland ₁	0.5379	0.9998	27.70	3.455	149.0
Block	1	1.130	1.130	1.130	78.53
Bohman	0.2660	1.274	126.2	18.00	409.3
Wendland ₁	0.1373	1.354	212.3	71.81	606.5
Block	1	1.229	1.229	1.229	173.3

NOTE: The block taper assumes independence between $16 \times 10 \times 10$ blocks. Range for Bohman and Wendland₁ tapers is 0.150755, which is chosen to match the number of nonzero elements of covariance matrix in the block approach. Variances are all relative to $\text{var}(\hat{\theta}) = \frac{2}{1600}$. The last column gives the minimized value (over θ) of the Kullback divergence of the tapered covariance matrix $\theta C \circ T$ from the true covariance matrix C .

range 1, the Wendland₁ taper is

$$W_1(r) = (1 - r)^4(1 + 4r) \quad (14)$$

for $0 \leq r \leq 1$ and 0 for $r > 1$. The Bohman taper (Bohman 1960; Gneiting 2002) with range 1 is

$$B(r) = (1 - r) \frac{\sin(2\pi r)}{2\pi r} + \frac{1 - \cos(2\pi r)}{2\pi^2 r} \quad (15)$$

for $0 \leq r \leq 1$ and is 0 for $r > 1$. These tapers are valid acfs in up to three dimensions. To get a taper with range ρ , just divide the argument of the taper by ρ . Both of these tapers are twice but not three times differentiable at the origin, so since the taper should normally be smoother than the acf to which it is applied, they are perhaps most appropriate for acfs K that are not twice differentiable at the origin. There are other tapers with these same properties, but the Bohman taper generally works about as well as any other tapers I have tried in the examples considered in this article.

These isotropic tapers are compared with a block taper made up of $16 \ 10 \times 10$ blocks. Writing $\lfloor \cdot \rfloor$ for the greatest integer function, $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$, this block-taping scheme can be viewed as resulting from the multiplicative taper K_b defined by $K_b(\mathbf{x}, \mathbf{y}) = 1$ if $4\lfloor x_1 \rfloor = 4\lfloor y_1 \rfloor$ and $4\lfloor x_2 \rfloor = 4\lfloor y_2 \rfloor$ and 0 otherwise. In this case, $\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}_3 = \frac{1}{n} \mathbf{Z}' \tilde{C}^{-1} \mathbf{Z}$, which is unbiased for θ ; this estimator for the scale parameter θ was considered in Stein (1986). The range of the isotropic tapers was chosen to equalize the number of nonzero elements in \tilde{C} with that of the block taper, even though block diagonal matrices are easier to work with than more general sparse matrices. Despite giving the isotropic tapers this advantage, the block taper is vastly superior for the exponential acf. The Bohman taper does better than the Wendland₁ taper in all cases and $\hat{\theta}_3$ does much better than $\hat{\theta}_2$, but even the Bohman taper with $\hat{\theta}_3$ has more than twice the variance as $\hat{\theta}_3$ based on the block taper. The advantages of the block taper over the other tapers are larger still for the smoother Whittle acf (the Matérn model with $\kappa = 1$). The downward biases for $\hat{\theta}_1$, especially for the Wendland₁ taper, are severe. The isotropic tapers of course do better if one reduces the range of the true acf, but the block taper also improves in this case and it is hard to find circumstances with just an unknown scale parameter under which the isotropic tapers would be clearly preferred.

The poor performance of the isotropic tapers even in the exponential case is rather disappointing in the light of Theorem 2, which shows that the tapered acfs with $\theta = 1$ yield equivalent Gaussian measures to the true acf, whereas the block taper K_b trivially yields an orthogonal measure, since it implies that the process is discontinuous at the boundaries between the blocks. Thus, in this case, as observations become dense in $[0, 1]^2$, the Kullback divergences for the isotropic tapers stay bounded, whereas the divergences for the block taper tend to ∞ .

3.2 UNKNOWN SCALE AND RANGE

In practice, covariance functions are rarely specified up to an unknown scale parameter. Thus, it is important to consider other settings before drawing general conclusions about the relative merits of various approximations to maximum likelihood estimation. Perhaps the most natural step up in complexity from the models considered in the previous section is

to add unknown range parameters. Exact expressions for means and variances of estimates are no longer available, so we need to resort to simulations or approximations.

Suppose $\mathbf{G}(\boldsymbol{\theta}) = \mathbf{0}$ is a set of unbiased estimating equations for $\boldsymbol{\theta}$. Define $\dot{\mathbf{G}}(\boldsymbol{\theta})$ to be the $p \times p$ matrix whose j th column is the derivative of $\mathbf{G}(\boldsymbol{\theta})$ with respect to θ_j . Then it is natural to judge the efficiency of the estimating equations by the Godambe information matrix (Varin, Reid, and Firth 2011) given by (suppressing the dependence on $\boldsymbol{\theta}$),

$$\mathcal{E}(\mathbf{G}) = (\mathbf{E}\dot{\mathbf{G}})'(\mathbf{E}\mathbf{G}\mathbf{G}')^{-1}(\mathbf{E}\dot{\mathbf{G}}). \quad (16)$$

In some circumstances, $(\mathcal{E}(\mathbf{G}))^{-1}$ gives the approximate covariance matrix of the estimates obtained from these estimating equations (Varin, Reid, and Firth 2011). Calculating $\mathcal{E}(\mathbf{G})$ is straightforward using standard results on derivatives of matrices, details are omitted (see the online supplementary materials for the numerical calculations).

Table 2 gives diagonal elements of $(\mathcal{E}(\mathbf{G}))^{-1}$ under an exponential and a Whittle model for the Bohman and block tapers. The observations are the same as in Table 1. The specific form for the exponential model is $\theta_1\theta_2e^{-|\mathbf{x}|/\theta_2}$, for which the principal irregular term (Stein 1999, p. 28) is $-\theta_1|\mathbf{x}|$ independent of θ_2 . This particular parameterization implies that θ_1 but not θ_2 is consistently estimable under fixed-domain asymptotics (Zhang 2004). The specific form for the Whittle model is $2\theta_1\theta_2^2\mathcal{M}_1(|\mathbf{x}|/\theta_2)$, for which the principal irregular term is $\theta_1|\mathbf{x}|^2 \log |\mathbf{x}|$. Again, θ_1 but not θ_2 is consistently estimable under fixed-domain asymptotics. Table 2 also gives sample variances based on 500 simulations, which are qualitatively similar with the numerical results, despite the lack of consistency of estimates of θ_2 under fixed-domain asymptotics. For all the estimates, given $\hat{\theta}_2$, there exists a closed-form expression for $\hat{\theta}_1$, so maximizing $\ell_1(\boldsymbol{\theta})$ and $\ell_2(\boldsymbol{\theta})$ can be reduced to a one-dimensional optimization and solving (10) to solving a scalar equation. The R routine optimize was used for the optimizations and the R routine uniroot to solve the nonlinear equation (see the online supplementary materials).

The striking result in Table 2 is how much the estimates based on the Bohman taper improve when an unknown range parameter is added to the model. One does not generally expect estimates to improve upon adding further unknown parameters to the model and, indeed, for full maximum likelihood, the diagonal elements of $(\mathcal{E}(\mathbf{G}))^{-1}$, which in this case

Table 2. Properties of estimates (11)–(13) of $\boldsymbol{\theta}$ under exponential ($\theta_1\theta_2e^{-|\mathbf{x}|/\theta_2}$, first two rows) and Whittle ($2\theta_1\theta_2^2\mathcal{M}_1(|\mathbf{x}|/\theta_2)$, last two rows) acfs with 1600 observations in $[0, 1]^2$ when $\boldsymbol{\theta} = (1, 0.25)$

	Bohman ₂		Bohman ₃		Block		Exact MLE	
	K	U	K	U	K	U	K	U
θ_1	13.5	1.39 (1.33)	2.98	1.69 (1.62)	1.41	1.41 (1.34)	1.25	1.29 (1.26)
θ_2	8.13	8.32 (7.16)	7.90	8.15 (7.18)	8.41	8.72 (7.83)	6.44	6.67 (8.22)
θ_1	158	8.32 (7.05)	22.5	2.03 (1.93)	1.54	1.50 (1.44)	1.25	1.30 (1.25)
θ_2	4.95	4.50 (3.54)	4.60	4.71 (3.49)	4.16	4.33 (3.29)	2.95	3.06 (2.84)

NOTE: Bohman and block tapers are as in Table 1; Bohman₂ refers to estimates based on (9) and Bohman₃ refers to estimates based on (10). Columns marked K assume the parameter under consideration is the only unknown parameter and columns marked U treat both parameters as unknown. Entries are diagonal elements of inverse Godambe information matrix times 1000. Numbers in parentheses are sample variances ($\times 1000$) based on 500 simulations. Poor simulation performance of exact MLE of θ_2 for exponential model is due to one outlier.

is given by the inverse Fisher information matrix, cannot decrease as one adds parameters. However, these diagonal elements can decrease for estimates that are not fully efficient and Table 2 shows that they can go down by an order of magnitude or more. For example, for the exponential model, the estimate of θ_1 based on the Bohman taper and maximizing $\ell_2(\theta)$ goes from being by far the worst approximation to the maximum likelihood estimator (MLE) to the best, doing slightly better than the block taper. For the Whittle model, the block approximation is still clearly best, but the gap with the two estimates based on the Bohman taper is much reduced from the θ_2 known case.

These results are very limited and it is certainly possible that there are models, tapers, observation locations, and estimating equations for which approximations based on isotropic tapers substantially outperform block tapers for a similar computational effort. The block taper is itself a crude approximation method and can sometimes be substantially improved upon using the approach in Vecchia (1988), further developed in Stein, Chi, and Welty (2004). However, for the cases considered in Table 2, preliminary explorations have not turned up approaches that do as well statistically as the simple block taper with substantially reduced computational effort.

4. TAPERING AND PREDICTION

Suppose Z is a mean 0 Gaussian process with covariance function K . Let Z_0 be the vector of n_0 observations and Z_1 the vector of n_1 predictands. Writing $\text{cov}(Z)$ for the covariance matrix of a random vector Z ,

$$\text{cov} \begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} = C = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix}$$

is the joint covariance matrix of the observations and predictands partitioned in the obvious way. Then the optimal (minimum MSE) linear predictor of Z_1 in terms of Z_0 is

$$\hat{Z}_1 = C_{10}C_{00}^{-1}Z_0 \quad (17)$$

and

$$\text{cov}(Z_1 - \hat{Z}_1) = C_{11} - C_{10}C_{00}^{-1}C_{01}. \quad (18)$$

Note that computing \hat{Z}_1 only requires a single solve $C_{00}^{-1}Z_0$, whereas computing $\text{cov}(Z_1 - \hat{Z}_1)$ requires n_1 solves. Similarly, using tildes to indicate covariances under \tilde{K} , define

$$\tilde{Z}_1 = \tilde{C}_{10}\tilde{C}_{00}^{-1}Z_0. \quad (19)$$

The covariance matrix of the prediction error is

$$\text{cov}(Z_1 - \tilde{Z}_1) = C_{11} - \tilde{C}_{10}\tilde{C}_{00}^{-1}C_{01} - C_{10}\tilde{C}_{00}^{-1}\tilde{C}_{01} + \tilde{C}_{10}\tilde{C}_{00}^{-1}C_{00}\tilde{C}_{00}^{-1}\tilde{C}_{01}. \quad (20)$$

4.1 CONDITIONAL SIMULATION

In some applications, $n_1 > n_0$, so calculating (20) may not be feasible even when (19) is. An alternative to calculating the exact error covariance matrix is conditional simulations

that obtain random draws from the conditional distribution of \mathbf{Z}_1 given \mathbf{Z}_0 . Such conditional simulations are often more useful than just evaluating the covariance matrix of prediction errors, as they can be used, for example, to approximate the probability of events involving nonlinear functionals of \mathbf{Z} . Matheron showed that this conditional simulation can be done exactly for Gaussian processes with a known covariance structure if one can compute $\hat{\mathbf{Z}}_1$ as in (17) and can unconditionally simulate $(\mathbf{Z}_0, \mathbf{Z}_1)$ (Chilès and Delfiner 1999). Specifically, let $(\mathbf{Z}_0^*, \mathbf{Z}_1^*)$ be a simulation from the unconditional distribution of $(\mathbf{Z}_0, \mathbf{Z}_1) \sim N(\mathbf{0}, C)$ independent of $(\mathbf{Z}_0, \mathbf{Z}_1)$. Then a conditional simulation of \mathbf{Z}_1 given \mathbf{Z}_0 is given by $\mathbf{S}_1 = C_{10}C_{00}^{-1}\mathbf{Z}_0 + \mathbf{Z}_1^* - C_{10}C_{00}^{-1}\mathbf{Z}_0^*$. In some circumstances, the unconditional simulation of $(\mathbf{Z}_0^*, \mathbf{Z}_1^*)$ can be done very efficiently. For example, if \mathbf{Z} is stationary and the locations of $(\mathbf{Z}_0, \mathbf{Z}_1)$ can be viewed as coming from a fine square grid (which might entail a slight distortion of the locations of \mathbf{Z}_0), then unconditional simulations can often be done in $O(n_1 \log n_1)$ calculations and $O(n_1)$ memory using circulant embedding (Dietrich and Newsam 1993; Wood and Chan 1994). The method does not always work, especially for smoother processes with a large range parameters, but there are approaches that sometimes extend the embedding approach to these cases (Stein 2002, 2012b; Gneiting et al. 2006).

When computing $\hat{\mathbf{Z}}_1$ approximately, it is not clear that exactly simulating from the unconditional distribution of \mathbf{Z}_1 is the best choice. To see that this is in fact the right thing to do, consider the (approximate) conditional simulation of \mathbf{Z}_1 given \mathbf{Z}_0 defined by $\tilde{\mathbf{S}}_1 = \tilde{C}_{10}\tilde{C}_{00}^{-1}\mathbf{Z}_0 + \mathbf{Z}_1^* - \tilde{C}_{10}\tilde{C}_{00}^{-1}\mathbf{Z}_0^*$. The conditional covariance matrix of $\tilde{\mathbf{S}}_1$ given \mathbf{Z}_0 is easily shown to equal the right-hand side of (20). Thus, it is appropriate to use the true covariance matrix C in the unconditional simulation step of the conditional simulation algorithm even when \tilde{C} is used to compute the optimal linear predictors.

4.2 DEFORMED TAPERS

Stein (1999, chap. 3) showed that using an incorrect covariance function generally has a larger effect on prediction when extrapolating than interpolating, so tapering should cause more problems when near or outside the boundary of the observation domain. A possible fix is to do less tapering near the edge of the observation domain. Indeed, Anderes et al. (2013) recommended using a taper with variable range to handle situations in which the density of observations varies substantially over the observation domain and developed a sophisticated method based on finding a smooth deformation g of the observation domain and then using the taper $K_t(g(\mathbf{x}) - g(\mathbf{y}))$. Even when the observations are of fairly constant density, deforming the taper near the boundary of the observation domain may yield better predictions near the boundary. When the observation domain is $[0, 1]^2$, a simple way to obtain a deformation that shrinks distances near the boundary of the square but otherwise leaves distances largely unchanged is as follows. For $\mathbf{x} \in [0, 1]^2$, define $g(\mathbf{x}) = (g_1(x_1), g_1(x_2))$, where, for $i = 1, 2$, $g_i(x_i) = G^{-1}(F(x_i))$ for smooth cdfs (cumulative distribution functions) F and G whose densities have support on the whole real line. Limited numerical experiments indicate that taking G to be the cdf for the standard normal and F the cdf for some t distribution works reasonably well. This approach trivially extends to rectangular observation domains (although different F and G for each coordinate may then be appropriate) but does not work more generally.

4.3 NUMERICAL RESULTS

The results in this section use the same 1600 observations on a perturbed 40×40 grid as in Section 3. The processes are then predicted at all locations $\frac{1}{40}(i, j)$ for $i, j \in 0, 1, \dots, 40$ for a total of 1681 predictands, including locations just barely outside the convex hull of the observations. The true acf is a Whittle model. The Wendland₁ (14) and Bohman (15) tapers were applied to the resulting joint covariance matrix of the observations and predictands. The top panel of Figure 1 compares the MSEs for these two tapers when each has a range of 0.25. For the vast majority of predictands, the Bohman taper yields a smaller MSE. The relative differences are often small, but can be substantial, especially for predictands along the boundary of the domain and particularly so for corner predictands.

The middle panel of Figure 1 shows the effect of deforming the Bohman taper near the boundary of the domain as described in Section 5.2 with F the cdf of a t distribution on 3 degrees of freedom and G a standard normal cdf. Figure 2 shows that the resulting mapping of the predictand locations shrinks distances quite a bit near the borders of the region. There is nothing optimal about these choices of F and G , and there is certainly scope for exploring how best to choose these functions depending on K and the density of observations. For example, as the number of observations increases, one might want to increase the degrees of freedom in the t distribution to make it more similar to the normal to limit the deformation's impact to a narrower band near the border.

The taper range on the deformed region is chosen so that the number of nonzero elements in the covariance matrix of the observations is the same as for the undeformed or stationary taper with a range of 0.25. For observations not on the boundary, the MSEs are quite similar for the deformed and stationary tapers, whereas for the predictions on the border, the deformed taper is clearly superior, especially in the corners. Thus, the deformed taper would not make sense in this setting if only predictions well inside the observation domain were of interest. The deformation does not make a substantial difference when applied to the Wendland₁ taper (see Table 3), so it is not always clear when deformation will be beneficial. Finally, the lower panel of Figure 1 shows the loss of efficiency of the predictions based on the deformed Bohman taper relative to the optimal predictor. Predictions not on the border tend to be fairly close to optimal, whereas those on the border are generally

Table 3. Average over 1681 prediction locations of $100 \times$ log ratio mean squared prediction error using tapering compared with optimal mean squared error

Taper/range	ρ					
	0.1		0.2		0.3	
	W	B	W	B	W	B
Stationary 0.1	8.21	9.86	17.00	16.55	25.22	22.19
Deformed	8.38	7.98	16.26	12.60	23.47	16.73
Stationary 0.25	1.09	0.68	2.52	1.64	3.72	2.59
Deformed	1.15	0.48	3.11	1.09	5.15	1.89

NOTE: True acf is Whittle $\mathcal{M}_1(|x|/\rho)$ with $\rho = 0.1, 0.2$, and 0.3 . Wendland₁ (W) and Bohman (B) tapers; "ranges" for deformed tapers chosen to equalize number of nonzero elements of covariance matrix of observations to stationary taper in the preceding line of the table.

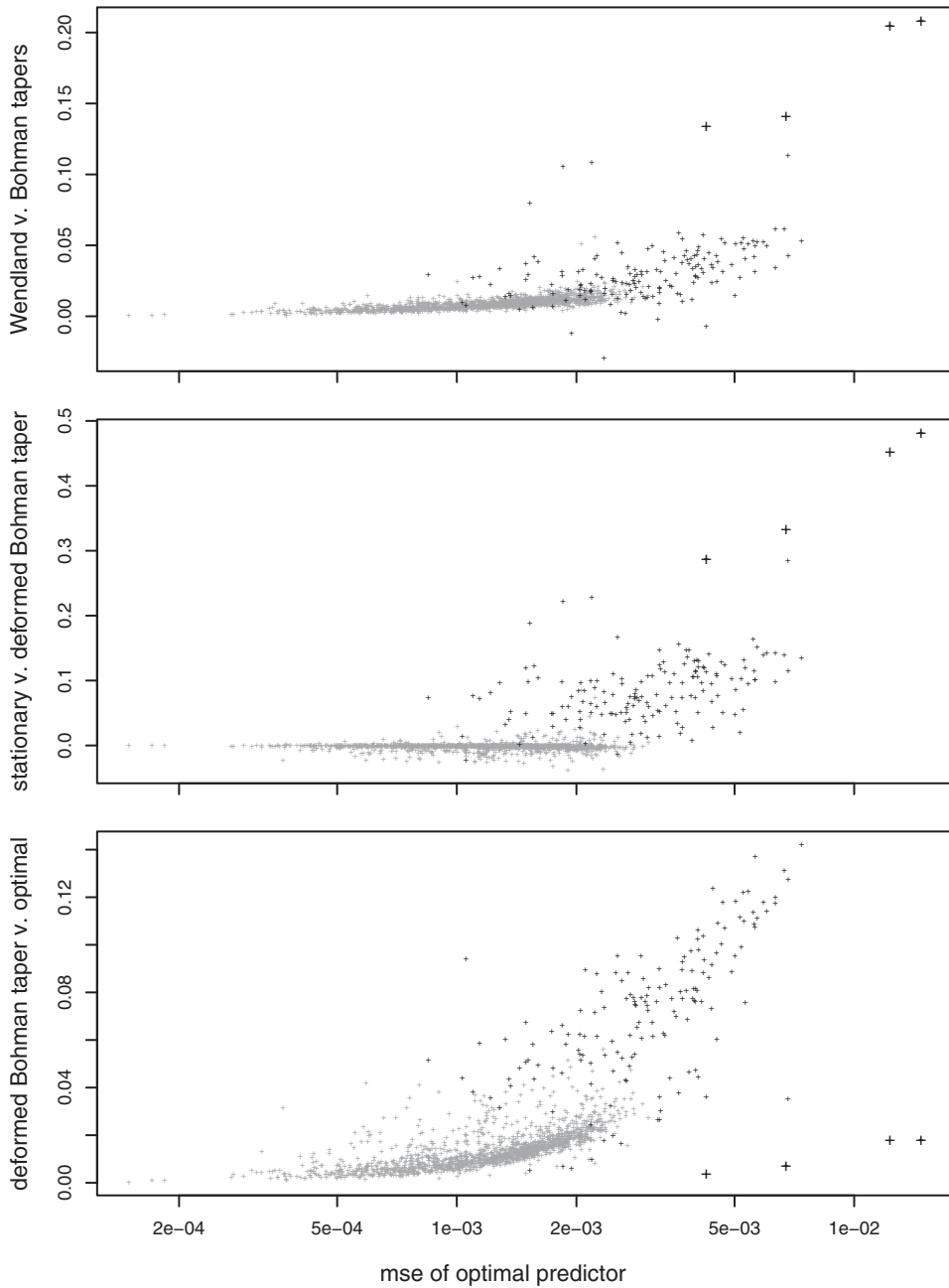


Figure 1. Log ratios of mean squared prediction errors for various tapers applied to the acf $\mathcal{M}_1(|x|/0.3)$. Observations on perturbed 40×40 grid. Top plot compares Wendland₁ taper with Bohman taper. Middle plot compares Bohman taper without and with deformation. Bottom plot compares deformed Bohman taper to optimal predictors. Gray +’s indicate predictands away from boundary, black +’s predictands on the boundary, and large black +’s the 4 corner predictands.

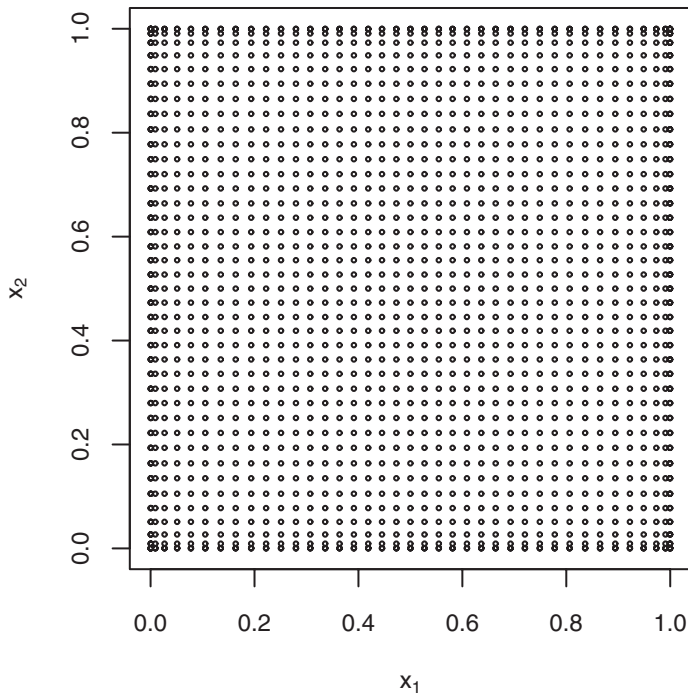


Figure 2. Locations of predictands after deformation.

worse, although the predictions at the corners, which are quite poor if one does not deform, are close to optimal.

A critical property of the kriging approach to spatial interpolation is that it provides estimates of mean squared prediction error. If K were somehow known and tapering were being used only to reduce computations, then one should use the exact formula (20) to evaluate these MSEs if this computation were feasible. An alternative, as noted in Section 5.1, is to use conditional simulations based on the untapered acf for the unconditional simulation step. However, these considerations tacitly assume that the true acf is known, which is rarely the case in practice. Thus, one might proceed by acting as if the tapered model were the truth and use

$$\text{cov}(\mathbf{Z}_1 - \tilde{\mathbf{Z}}_1) = \tilde{C}_{11} - \tilde{C}_{10}\tilde{C}_{00}^{-1}\tilde{C}_{01} \quad (21)$$

to approximate the mean squared prediction errors. Note that (21) and (20) both require calculating $\tilde{C}_{00}^{-1}\tilde{C}_{01}$, so unless the number of predictands n_1 is quite small and/or \tilde{C}_{00} is exceptionally sparse relative C_{00} , (21) will not be much easier to compute than (20). For the same setting as in Figure 1, Figure 3 plots the log ratios of the presumed mean squared prediction errors given by the diagonal entries of (21) to the actual MSEs under (20) if the untapered acf were correct. The bulk of the presumed MSEs are substantially off for all three tapers, although the deformed Bohman taper is best overall and does much better for predictions on the boundary. The fact that treating the tapered model as if it were the truth

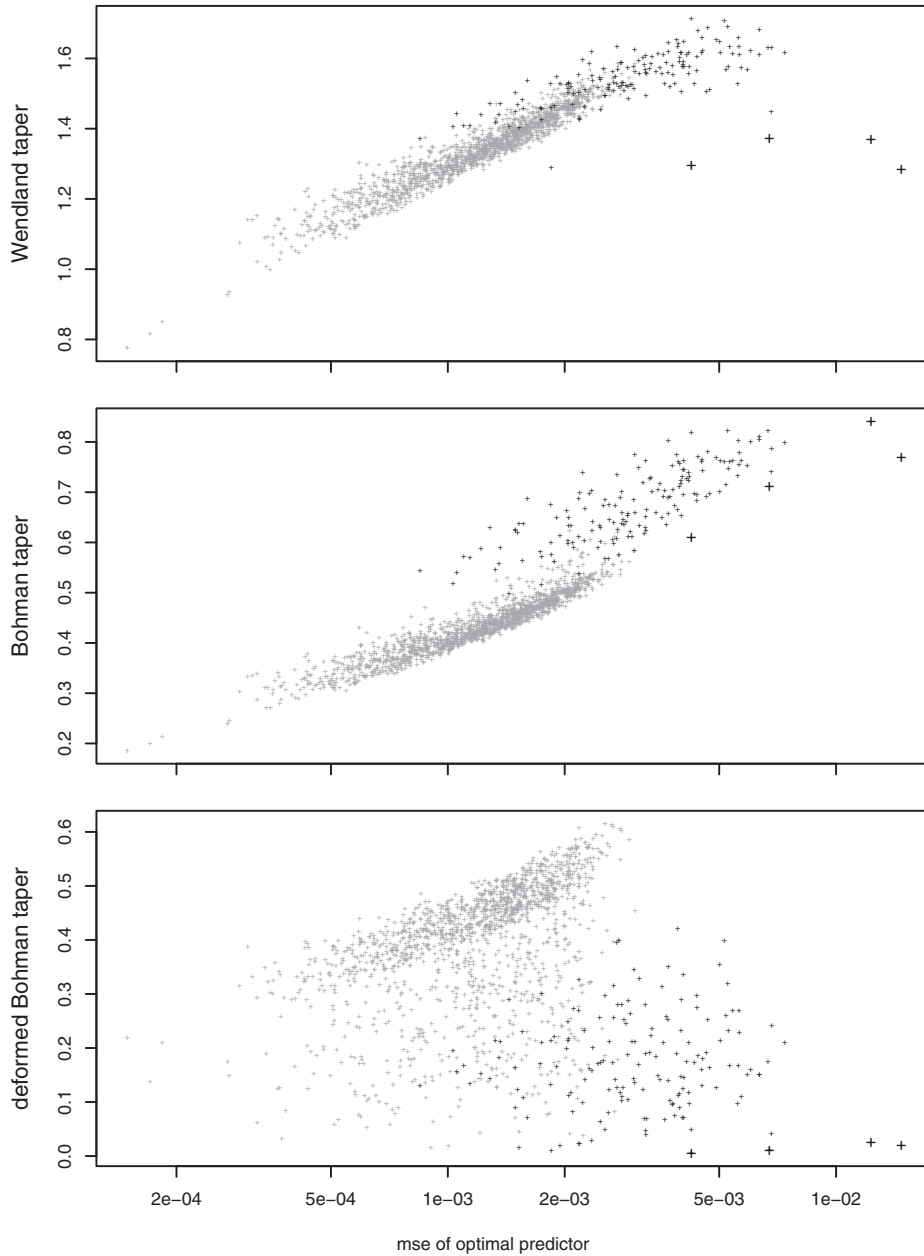


Figure 3. Under same setting as Figure 1, log ratios of presumed to actual mean squared prediction errors. Presumed covariance structures are of form $\theta \tilde{K}$, where θ is chosen to minimize the Kullback divergence from the truth for the observations.

has a larger impact on the evaluation of MSEs than on the actual MSEs of predictions is not surprising in the light of the extensive results in Stein (1999) along these lines on the impact of using an incorrect covariance function on linear predictions. Arguably \tilde{C} should be normalized in some way, but the asymptotic theory of Section 2 indicates that should not be necessary if the observations are sufficiently dense. Furthermore, replacing \tilde{C} by $\theta \tilde{C}$ for

Table 4. Average of $(100 \times)$ log ratio mean squared prediction error using tapering compared with optimal mean squared error

Taper/range	κ					
	0.5		0.75		1.25	
	W	B	W	B	W	B
Stationary 0.1	7.62	8.40	10.40	12.04	33.08	25.60
Deformed	6.71	6.89	9.62	9.28	33.18	19.90
Stationary 0.25	1.02	0.58	1.82	0.98	3.05	2.25
Deformed	0.84	0.40	1.60	0.61	6.36	2.30

NOTE: True acf is Matérn $\theta \mathcal{M}_\kappa(|x|/0.2)$ with $\kappa = 0.5, 0.75$, and 1.25 . Observations, predictands and tapers are as in Table 3.

some $\theta \neq 1$ merely shifts the vertical axes of the plots in Figure 3 and does not change the spread in these log ratios. Based on these results and similar findings in other cases, it is difficult to recommend using (21) instead of (20) when using a taper that produces a high degree of sparseness in the covariance matrix.

Next consider the average log ratio of mean squared prediction errors for the same general setup as in Figure 1 but for a range of acfs. Table 3 considers Whittle processes with various range parameters. Comparing the four combinations of Wendland₁ and Bohman tapers with and without deformation, Table 3 shows that the deformed Bohman taper performs best for all three values of the range parameter of the Whittle model and both values of the range of the tapers considered. The relative advantage of the deformed Bohman taper increases with either of these two ranges. Table 4 considers Matérn models with different values of the smoothness parameter κ . Again, the deformed Bohman taper is generally (but not always) best. For $\kappa = 1.25$, the Bohman and Wendland₁ tapers are only modestly smoother than the underlying Matérn acf (in the notation of Theorem 2, $\delta = \frac{1}{4}$), so perhaps a smoother taper would be in order. The Wendland₂ taper (Furrer, Genton, and Nychka 2006) is a smoother taper (it is four times differentiable as opposed to twice for Bohman and Wendland₁), but in these settings, it always performs worse than the Bohman taper.

All of the examples for the true acf in these tables are Matérn models. However, similar results hold for other models that match the local behavior of the Matérn model. For example, consider the acf $K(\mathbf{x}) = \{1 + (|\mathbf{x}|/0.2)^{1.5}\}^{-1}$ (Gneiting and Schlather 2004), which has the same degree of smoothness at the origin as the Matérn model with $\kappa = 0.75$ (and the same nominal range of 0.2, although it is unclear how to compare ranges for different models). The four values corresponding to the third column of numbers in Table 4 (the Wendland₁ taper) are now, in order, 11.20, 10.50, 1.96, and 1.82 and to the fourth column (the Bohman taper) are 12.81, 9.94, 1.05, and 0.68, which are all just slightly larger than the results for the corresponding Matérn model.

5. OTHER ISSUES

In many applications of spatial statistics, it is common to include a discontinuity at the origin in the acf, or nugget effect. This nugget effect might reflect microstructure in the underlying process or random measurement error. It is not difficult to show that the

Kullback divergence of a $N(\mathbf{0}, \tilde{C} + \theta I)$ distribution from a $N(\mathbf{0}, C + \theta I)$ distribution is decreasing in θ , so that the presence of a nugget effect makes it more difficult to distinguish different models for the continuous part of the process. Not surprisingly, as the nugget effect grows, numerical experiments indicate that the differences in results for prediction between different taper functions tend to decrease. Thus, for comparing different tapers, the no nugget case is of the greatest interest.

For simplicity, this work has assumed that the mean of the process is known to be 0. If the process has an unknown constant mean, then, given some covariance structure, ordinary kriging, which is just best linear unbiased prediction, is generally used for prediction. Numerical experiments indicate that the results in, for example, [Table 3](#), hardly change when using ordinary kriging. However, I would expect that the estimation of the range parameter, considered in [Section 3.2](#), would be substantially affected by the inclusion of an unknown mean parameter when nearby observations are strongly correlated, since neither the range nor the mean is microergodic (Stein [1999](#), sec. 6.2) under fixed-domain asymptotics.

6. DISCUSSION

Tapering can be an effective tool for approximating optimal linear predictors in that one can often get nearly optimal linear predictors with tapers having a quite small range, yielding a substantial computational saving over computing the exact optimal predictors. If unconditional simulations can be efficiently done under the untapered covariance, then conditional simulations based on combining the tapered predictions with the exact unconditional simulations should be a useful tool. Although further exploration is warranted, the results here suggest caution using isotropic tapers, at least in conjunction with [\(8\)–\(10\)](#), for likelihood approximations. Block tapers, or perhaps more sophisticated related approaches such as those described in Vecchia ([1988](#)); Stein, Chi, and Welty ([2004](#)); and Caragea and Smith ([2007](#)) may hold more promise for approximating maximum likelihood estimates for irregularly sited observations. Inference based on approximated likelihoods that goes beyond using the Godambe information matrix is addressed in, for example, Chandler and Bate ([2007](#)) and Pace, Salvan, and Sartori ([2011](#)).

Predictions even slightly beyond the observation domain are strongly affected by tapering, but this problem can be addressed to some extent by using deformed tapers. There can be substantial differences between tapers with the same degree of smoothness at the origin, with, for example, the Bohman taper generally performing better than the Wendland₁ taper, especially when combined with deformation. There is scope for exploring other tapers, deformations, underlying models for the process, and arrangements of the observations and predictands. Finally, all of the examples considered here are in two dimensions and assume the underlying process is isotropic. Neither of these assumptions is appropriate in most applications of Gaussian processes to computer experiments; see Kaufman et al. ([2011](#)) for an application of covariance tapering in this setting.

SUPPLEMENTARY MATERIALS

There are five supplementary files:

1. Restimate.article: R commands to generate results in [Table 1](#).

2. Rsimulate.article: R commands to generate results in [Table 2](#).
3. Rpredict.article: R commands to generate results in [Tables 3](#) and [4](#).
4. perturbed.grid: Coordinates of 1600 points in perturbed grid.
5. taper-proofs.pdf: Proofs of Theorems 1 and 3.

The first three files contain some additional numerical results not directly reported on in the article.

ACKNOWLEDGMENTS

This research was supported by U.S. Department of Energy grant no. DE-SC0002557. The author thanks an anonymous reviewer for suggesting a simulation study along the lines described in [Section 3.2](#).

[Received November 2011. Revised August 2012]

REFERENCES

- Anderes, E., Huser, R., Nychka, D., and Coram, M. (2013), “Nonstationary Positive Definite Tapering on the Plane,” *Journal of Computational and Graphical Statistics*, 22, 848–865. [[867,877](#)]
- Anitescu, M., Chen, J., and Wang, L. (2012), “A Matrix-Free Approach for Solving the Parametric Gaussian Process Maximum Likelihood Problem,” *SIAM Journal on Scientific Computing*, 34, A240–A262. [[872](#)]
- Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987), *Regular Variation*, New York: Cambridge University Press. [[870](#)]
- Bohman, H. (1960), “Approximate Fourier Analysis of Distribution Functions,” *Arkiv för Matematik*, 4, 99–157. [[873,874](#)]
- Caragea, P. C., and Smith, R. L. (2007), “Asymptotic Properties of Computationally Efficient Alternative Estimators for a Class of Multivariate Normal Models,” *Journal of Multivariate Analysis*, 98, 1417–1440. [[883](#)]
- Chandler, R. E., and Bate, S. (2007), “Inference for Clustered Data Using the Independence Log-Likelihood,” *Biometrika*, 94, 167–183. [[883](#)]
- Chilès, J., and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: John Wiley. [[877](#)]
- Davis, T. A. (2006), *Direct Methods for Sparse Linear Systems*, Philadelphia, PA: SIAM. [[872](#)]
- Dietrich, C., and Newsam, G. (1993), “A Fast and Exact Method for Multidimensional Gaussian Stochastic Simulations,” *Water Resources Research*, 29, 2861–2869. [[877](#)]
- Du, J., Zhang, H., and Mandrekar, V. S. (2009), “Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators,” *The Annals of Statistics*, 37, 3330–3361. [[870](#)]
- Furrer, R., Genton, M. G., and Nychka, D. (2006), “Covariance Tapering for Interpolation of Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 15, 502–523. [[867,869,871,873,882](#)]
- Gneiting, T. (2002), “Compactly Supported Correlation Functions,” *Journal of Multivariate Analysis*, 83, 493–508. [[873,874](#)]
- Gneiting, T., and Schlather, M. (2004), “Stochastic Models That Separate Fractal Dimension and the Hurst Effect,” *SIAM Review*, 46, 269–282. [[882](#)]
- Gneiting, T., Ševčíková, H., Percival, D. B., Schlather, M., and Jiang, Y. (2006), “Fast and Exact Simulation of Large Gaussian Lattice Systems in \mathbb{R}^2 : Exploring the Limits,” *Journal of Computational and Graphical Statistics*, 15, 483–501. [[877](#)]
- Golub, G. H., and van Loan, C. F. (1996), *Matrix Computations* (3rd ed.), Baltimore MD: The Johns Hopkins University Press. [[872](#)]

- Ibragimov, I. A., and Rozanov, Y. A. (1978), *Gaussian Random Processes*, trans. A. B. Aries, New York: Springer-Verlag. [870]
- Kaufman, C., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. (2011), "Efficient Emulators of Computer Experiments Using Compactly Supported Correlation Functions, With an Application to Cosmology," *Annals of Applied Statistics*, 5, 2470–2492. [883]
- Kaufman, C., and Shaby, B. (2011), "The Importance of the Range Parameter for Estimation and Prediction in Geostatistics," arXiv:1108.1851v1. [870]
- Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [867,870,871]
- Pace, L., Salvan, A., and Sartori, N. (2011), "Adjusting Composite Likelihood Ratio Statistics," *Statistica Sinica*, 21, 129–248. [883]
- Stein, M. L. (1986), "A Modification of Minimum Norm Quadratic Estimation of a Generalized Covariance Function for Use With Large Data Sets," *Mathematical Geology*, 18, 625–633. [874]
- (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer. [867,868,869,875,877,881,883]
- (2002), "Fast and Exact Simulation of Fractional Brownian Surfaces," *Journal of Computational and Graphical Statistics*, 11, 587–599. [877]
- (2012a), "When Does the Screening Effect Hold?" *The Annals of Statistics*, 39, 2795–2819. [870]
- (2012b), "Simulation of Gaussian Random Fields With One Derivative," *Journal of Computational and Graphical Statistics*, 21, 155–173. [877]
- Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Datasets," *Journal of the Royal Statistical Society, Series B*, 66, 275–296. [876,883]
- Varin, C., Reid, N., and Firth, D. (2011), "An Overview of Composite Likelihood Methods," *Statistica Sinica*, 21, 5–42. [875]
- Vecchia, A. V. (1988), "Estimation and Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society, Series B*, 50, 297–312. [876,883]
- Wang, D., and Loh, W.-L. (2011), "On Fixed-Domain Asymptotics and Covariance Tapering in Gaussian Random Field Models," *Electronic Journal of Statistics*, 5, 238–269. [870]
- Wendland, H. (1995), "Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree," *Advances in Computational Mathematics*, 4, 389–396. [873]
- Wood, A. T. A., and Chan, G. (1994), "Simulation of Stationary Gaussian Processes in $[0, 1]^d$," *Journal of Computational and Graphical Statistics*, 3, 409–432. [877]
- Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [875]