Covariance Tapering for Interpolation of Large Spatial Datasets

Reinhard Furrer, Marc G. Genton and Douglas Nychka

Interpolation of a spatially correlated random process is used in many areas. The best unbiased linear predictor, often called kriging predictor in geostatistical science, requires the solution of a large linear system based on the covariance matrix of the observations. In this article, we show that tapering the correct covariance matrix with an appropriate compactly supported covariance function reduces the computational burden significantly and still has an asymptotic optimal mean squared error. The effect of tapering is to create a sparse approximate linear system that can then be solved using sparse matrix algorithms. Extensive Monte Carlo simulations support the theoretical results. An application to a large climatological precipitation dataset is presented as a concrete practical illustration.

Keywords: kriging, sparse matrix, asymptotic optimality, large linear systems, compactly supported covariance.

1 Introduction

Many applications of statistics in the geophysical and environmental sciences depend on estimating the spatial and temporal extent of a physical process based on irregularly spaced observations. In many cases the most interesting spatial problems are large and their analysis overwhelms traditional implementations of spatial statistics. For example, in understanding recent climate change for the US it is useful to infer monthly average temperature or precipitation fields for the past century. These surfaces are estimated from the historical record of meteorological measurements taken at irregular station locations and the complete surfaces facilitate direct comparison with numerical climate models[1] and also can serve as inputs to ecological and vegetation models. Estimating monthly precipitation fields for the US involves more than 5,900 station locations at the peak network size and must be repeated over the more than 1,200 months of the historical record. In addition, each estimated field should be evaluated on a fine grid of size approximately 1,000×1,000 (corresponding to a resolution of roughly 2.4 km latitude and 4 km longitude).

The size of this spatial problem for climate studies is not unusual and, in fact, geophysical datasets several orders of magnitude larger can be expected based on satellite and other modern observing systems. Because of the size of these problems it is well known that a naive implementation of spatial process prediction, such as kriging, is not feasible. In addition, more complex approaches such as Bayesian hierarchical space-time models often have a kriging like step as one of the full conditional distributions in a Gibbs sampling scheme. Thus, these more flexible methods are also limited in their application to large spatial problems without making the spatial prediction step more efficient.

In this work we propose an approximation to the standard linear spatial predictor that can be justified by asymptotic theory and is both accurate and computationally efficient. Our basic idea is to taper the spatial covariance function to zero beyond a certain range. This results in sparse systems of linear equations that can be solved efficiently. Moreover, we show that the tapering can be done to give a linear

Reinhard Furrer is a postdoctoral researcher at Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO, 80307–3000, furrer@ucar.edu. Marc G. Genton is Associate Professor at North Carolina State University, Raleigh, NC, 27695–8203, genton@stat.ncsu.edu. Douglas Nychka is Senior Scientist at Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO, 80307–3000, nychka@ucar.edu.

[1] Coupled atmosphere and ocean general circulation models

predictor that is nearly the same as the exact solution. The effect of tapering can be analyzed using the infill asymptotic theory for a misspecified covariance and we find it interesting that in our case the "misspecification" is deliberate and has computational benefits. In addition, we belive that many large spatial datasets fit the assumptions made by infill asymptotic analysis.

1.1 Spatial Prediction

Assume that a spatial field $Z(\boldsymbol{x})$ is a process with covariance function $K(\boldsymbol{x}, \boldsymbol{x}')$ for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D} \subset \mathbb{R}^d$, and is observed at the n locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. For the application in Section 4, Z is monthly average precipitation for a particular month and $\mathcal{D}$ is the coterminous US. A common problem is to predict $Z(\boldsymbol{x}^*)$ given the n observations for an arbitrary $\boldsymbol{x}^* \in \mathcal{D}$. In geostatistics the standard approach is based on the principle of minimum mean squared error, termed *kriging* (*e.g.* Cressie, 1990, 1993), and as motivation we start with the simplest spatial model. Assume that $Z(\boldsymbol{x})$ has mean zero and is observed without any measurement error. Then the best linear unbiased prediction (BLUP) at an (unobserved) location $\boldsymbol{x}^*$ is then

$$\hat{Z}(\boldsymbol{x}^*) = \boldsymbol{c}^{*\mathsf{T}} \mathbf{C}^{-1} \mathbf{Z}, \tag{1}$$

where $\mathbf{Z} = \big(Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)\big)^{\mathsf{T}}$, $\mathbf{C}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{c}_i^* = K(\boldsymbol{x}_i, \boldsymbol{x}^*)$. Furthermore, if we assume that Z is a Gaussian process then $\hat{Z}(\boldsymbol{x}^*)$ as given by (1) is just a conditional expectation of $Z(\boldsymbol{x}^*)$ given the observations. We will denote the prediction mean squared error by

$$\mathrm{MSE}(\boldsymbol{x}^*, \hat{K}) = \mathrm{E}\big(\hat{Z}(\boldsymbol{x}^*) - Z(\boldsymbol{x}^*)\big)^2 = K(\boldsymbol{x}^*, \boldsymbol{x}^*) - 2\hat{\boldsymbol{c}}^{*\mathsf{T}} \hat{\mathbf{C}}^{-1} \boldsymbol{c}^* + \hat{\boldsymbol{c}}^{*\mathsf{T}} \hat{\mathbf{C}}^{-1} \mathbf{C} \hat{\mathbf{C}}^{-1} \hat{\boldsymbol{c}}^*, \tag{2}$$

where the hat entities are based on $\hat{K}$. Here is it important to note that the covariance in the second argument defines the *estimate* and may not necessary be the actual covariance of the process. This distinction is important if we want to study the performance of the kriging estimator when the covariance is misspecified, or at least deviates from the actual covariance of the process. However, if K is indeed the true covariance the $\mathrm{MSE}(\boldsymbol{x}^*, K)$ from (2) simplifies to

$$\varrho(\boldsymbol{x}^*, K) = K(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{c}^{*\mathsf{T}} \mathbf{C}^{-1} \boldsymbol{c}^* \tag{3}$$

and is the well known expression for the variance of the kriging estimate. Finally, we note that $\varrho(\boldsymbol{x}^*, \hat{K})$ is a naive prediction standard error computed assuming $\hat{K}$ is the true covariance function.

The computation of $\boldsymbol{u} = \mathbf{C}^{-1}\mathbf{Z}$ in (1) involves the solution of a linear system that is the size of the number of observations. Both the operation count for solving a linear system and also the storage increase by order n^2. Moreover, we wish to evaluate the prediction at many grid points and so the practical applications involve finding $\boldsymbol{c}^{*\mathsf{T}}\boldsymbol{u}$ for many vectors $\boldsymbol{c}^*$. These two linear algebra steps effectively limit a straightforward calculation of the spatial prediction to small problems. Note that for our motivating climate application n = 5,906 and $\boldsymbol{c}^{*\mathsf{T}}\boldsymbol{u}$ must be evaluated on the order of 100,000 times. The direct computation of the prediction error variance (3) is even more demanding as this involves either solving a different linear system at each $\boldsymbol{x}^*$ or directly inverting the matrix $\mathbf{C}$ and performing the multiplications explicitly.

1.2 Tapering

The goal of our work is to give an accurate approximation to (1) and (2) but also to propose a method that scales to large spatial problems. The basic idea is simple, we deliberately introduce zeros into the matrix $\mathbf{C}$ in (1) to make it sparse. How the zeros are introduced, however, is crucial. Let K_θ be a covariance function

that is identically zero outside a particular range described by θ. Now consider a tapered covariance that is the direct (or Schur) product of K_θ and K:

$$K_{\mathsf{tap}}(\boldsymbol{x}, \boldsymbol{x}') = K(\boldsymbol{x}, \boldsymbol{x}')K_\theta(\boldsymbol{x}, \boldsymbol{x}').$$

Throughout the paper we suppose that $K_\theta(\boldsymbol{x}, \boldsymbol{x}) = 1$. The approximate estimate is obtained by replacing the covariance matrices in (1) based on K by those defined by K_{tap}. The intuition behind this choice is that product still preserves some of the shape of K but also it is identically zero outside of a fixed interval. Of equal importance, K_{tap} is a valid covariance. This is based on the result that the Schur product of two positive definite matrices is again positive definite (Horn and Johnson, 1994, Theorem 5.2.1).

Limiting the kriging estimate to a local neighborhood is of course not a new idea. Indeed, a very effective use of covariance tapering is well known in the atmospheric science literature for numerical weather prediction (Gaspari and Cohn, 1999). Examining the weight vector $\boldsymbol{\lambda} = \mathbf{C}^{-1}\boldsymbol{c}^*$ from (1) one would expect this to be close to zero for observation locations that are "far" from $\boldsymbol{x}^*$. The contribution of an observation, say $Z(\boldsymbol{x}_i)$ to the prediction of $Z(\boldsymbol{x}^*)$ decreases as the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}^*$ increases and it would be reasonable to consider only nearby locations for the prediction at $\boldsymbol{x}^*$. This restriction also makes sense even when the process has long-range correlations by interpreting the estimate as a conditional expectation. Although $Z(\boldsymbol{x}^*)$ may be highly correlated with distant observations it can be nearly independent of distant observations *conditional* on its neighbors. The distant observations do not give additional information about $Z(\boldsymbol{x}^*)$ given the observed values of the field close by. Such arguments leads to the so-called kriging neighborhood. One simply calculates the spatial estimate based on a small and manageable number of observations that are close to $\boldsymbol{x}^*$. This approach is quite useful when predicting at a limited number of locations (*e.g.* Johns *et al.*, 2003), but has several drawbacks as pointed out in Cressie (1993). We also acknowledge a parallel development in nearest neighbor and local estimates from nonparametric regression (*e.g.* Cleveland *et al.*, 1992). Here, the form of the estimators is justified by asymptotic optimality and usually depends on measurement error being a significant fraction of the variance in the data. For our purposes we are more concerned with the low noise situation where the fitted surface tends to interpolate or nearly interpolate the observations. However, in all of these cases the difficulty of neighborhood methods is that the neighborhood changes for each point for prediction. Although the computation is reduced for an individual point, prediction of the field without artifacts from changing neighborhoods is problematic. Moreover, we will show that the tapering sparse matrix approach from this work has a similar operation count to nearest neighbor estimators without its disadvantages.

1.3 Outline

The effect of the covariance function on linear predictor has a long and extensive literature; some examples include Diamond and Armstrong (1984); Yakowitz and Szidarovszky (1985); Warnes (1986); Stein and Handcock (1989). In a series of papers Stein (Stein, 1988, 1990a, 1997, 1999b) gives a thorough theoretical discussion of the effect of miss-specifying the covariance function. In his approach, "miss-specified" refers to a covariance similar — in some sense — to the true underlying covariance. Although much of that work is motivated by a covariance that is in error, one might adapt these results to consider the effect of deliberately modifying the "true" covariance through a taper. We note that from a theoretical perspective Stein has also suggested that tapering could be effective (Stein, 1999a, page 53) for reducing the computation.

These remarks motivate the following research questions:

Question A. What is the increase in squared prediction error by using the taper approach?

Question B. What are the associated computational gains?

The next section answers Question A by adapting the asymptotic theory in Stein (1990a, 1997, 1999b) to understand the large sample properties. This is paired with some exact calculations in Section 3.2 to investigate the efficiency for finite samples. Question B can be answered by comparing standard and sparse techniques and Section 3.3 illustrates the gain in storage and computational cost when tapering is used. To emphasize the practical benefits of tapering we report timing results for the large climate application based on monthly precipitation fields in Section 4. To limit the scope of this paper we will only consider stationary processes and, in fact, restrict most of our study to the Matérn family. In addition we do not highlight the more practical spatial processes that admit some fixed effects (also known as spatial drift). The last section discusses the logical extension of tapering algorithms to nonstationary covariances and to spatial models with fixed effects.

2 Asymptotic Properties of Tapering

Our goal is to show that under specific conditions the asymptotic mean squared error of the tapered covariance will converge to the optimal error. Following the theory of Stein we phrase these results in terms of a misspecified covariance. Of course, the misspecification here is deliberate and involves tapering.

An important restriction throughout this analysis is that the processes and tapering functions are second order stationary and isotropic. Moreover, we will focus on the Matérn family of covariances. Assume that the process Z is isotropic, stationary and has an underlying Matérn covariance function defined by $K_{\alpha,\nu}(\boldsymbol{x},\boldsymbol{x}') = C_{\alpha,\nu}(h)$, $h = \|\boldsymbol{x}-\boldsymbol{x}'\|$ with

$$C_{\alpha,\nu}(h) = \frac{\pi^{d/2}\phi}{2^{\nu-1}\Gamma(\nu+d/2)\alpha^{2\nu}}(\alpha h)^{\nu}\mathcal{K}_{\nu}(\alpha h), \quad \alpha>0,\ \phi>0,\ \nu>0, \tag{4}$$

Γ is the Gamma function and $\mathcal{K}_\nu$ is the modified Bessel function of the second kind (Abramowitz and Stegun, 1970). The process Z is m times mean square differentiable iff $\nu > m$ and this concept can be extended to fractional differentiability. The parameters α and ϕ are related to the standard range and the sill, respectively. The Matérn family is a prototype for a scale of covariances with different orders of smoothness and has a simple spectral density

$$\frac{\phi}{(\alpha^2+\rho^2)^{\nu+d/2}}\,. \tag{5}$$

In this work without loss of generality, we assume $\phi = 1$ and so it is convenient to let $f_{\alpha,\nu}(\rho)$ denote the Matérn spectral density in (5) with this restriction. If $\nu = 0.5$, $C_{\alpha,\nu}$ is an exponential covariance, $\nu = n + 0.5$, n an integer, (4) is an exponential covariance times a polynomial of order n and the limiting case $\nu \to \infty$ is the Gaussian covariance.

Our results are asymptotic in the context of a fixed domain size and the sequence of observations increasing within the domain. This is known as infill asymptotics.

Infill Condition. *Let $\boldsymbol{x}^* \in \mathcal{D}$ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ be a dense sequence in $\mathcal{D}$ and distinct from $\boldsymbol{x}^*$.*

Taper Condition *Let f_θ be the spectral density of the taper covariance, C_θ, and for some $\epsilon > 0$ and $M(\theta) < \infty$*

$$f_\theta(\rho) < \frac{M(\theta)}{(1+\rho^2)^{\nu+d/2+\epsilon}}.$$

We will motivate the material in this section by the main result:

Theorem 2.1. (Taper Theorem) *Assume that $C_{\alpha,\nu}$ is a Matérn covariance with smoothness parameter ν and the Infill and Taper Conditions hold. Then*

$$\lim_{n\to\infty} \frac{\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu}C_\theta)}{\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})} = 1, \tag{6}$$

$$\lim_{n\to\infty} \frac{\varrho(\boldsymbol{x}^*, C_{\alpha,\nu}C_\theta)}{\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})} = \gamma. \tag{7}$$

The following sections develop a proof for this result with some details contained in the Appendix A.

2.1 Asymptotic Equivalence of Kriging Estimators

Asymptotic equivalence for two covariance functions is easiest to describe based on tail behavior of the spectral densities. Let C_0 and C_1 be two stationary covariance functions with corresponding spectral densities f_0 and f_1.

Tail Condition. *For two spectral densities f_0 and f_1*

$$\lim_{\rho\to\infty} \frac{f_1(\rho)}{f_0(\rho)} = \gamma, \quad 0 < \gamma < \infty. \tag{8}$$

Based on the tail condition we have the following general result for misspecification.

Theorem 2.2. *Let C_0 and C_1 be isotropic covariance functions with corresponding spectral densities f_0 and f_1. Furthermore assume that Z is a mean zero second order stationary process with covariance C_0 and that the Infill Condition holds. If f_0 and f_1 satisfy the Tail Condition then*

$$\lim_{n\to\infty} \frac{\mathrm{MSE}(\boldsymbol{x}^*, C_1)}{\mathrm{MSE}(\boldsymbol{x}^*, C_0)} = 1, \qquad \lim_{n\to\infty} \frac{\varrho(\boldsymbol{x}^*, C_1)}{\mathrm{MSE}(\boldsymbol{x}^*, C_0)} = \gamma.$$

The first limit indicates that the misspecified estimator has the same convergence rate as the optimal one. The second limit indicates that the naive formula (3) for the kriging variance also has the correct convergence rate. Finally, if $\gamma = 1$ then we have asymptotic equivalence for the estimator and the variance using the wrong covariance function. If $\gamma \neq 1$, we can divide the taper by γ to optain asymptotic equivalence.

This theorem cited above does not identify the rate of convergence of the optimal estimate. However, these are well known for equispaced multidimensional grids (Stein, 1999a). In addition, we believe one can apply some classical interpolation theory (*e.g.* Madych and Potter, 1985) to bound the rate for the kriging estimator for irregular sets of points, but this is a subject of future research.

2.2 Tail Condition for Matérn Covariances

In order to apply Theorem 2.2 it is necessary to verify the Tail Condition for the hypotheses in the Taper Theorem. Recall that the convolution of two functions is equivalent to multiplication of their Fourier transforms. Moreover, one can reverse operations and it is also true that the product of two functions is equivalent to a convolution of their Fourier transforms. Let f_{tap} denote the spectral density for C_{tap} and so we have

$$f_{\mathsf{tap}}(\|\boldsymbol{u}\|) = \int_{\mathbb{R}^d} f_{\alpha,\nu}(\|\boldsymbol{u}-\boldsymbol{v}\|) f_\theta(\|\boldsymbol{v}\|)\, \mathsf{d}\boldsymbol{v}.$$

It is reasonable to expect these two spectral densities to satisfy the Tail Condition when f_θ has lighter tails than $f_{\alpha,\nu}$. Consider the following intuitive reasoning. The spectra $f_{\alpha,\nu}$ and f_θ can be considered as the densities of random variables, say X and Y, respectively. Then, being a convolution, f_{tap} is the density of

$X + Y$. The Tail Condition implies that the variables $X + Y$ and X have the same moment properties. This will hold given the initial tail assumptions on the densities for X and Y. As the tail behavior is related to the behavior of the covariance at the origin, the taper needs to be at least as differentiable at the origin as $C_{\alpha,\nu}$. Stein (1988) noted that this necessary condition is not sufficient. The taper has also to be sufficiently derivable away from zero without imposing restrictions on the domain $\mathcal{D}$. If this condition is not met, the taper has oscillating behavior for high frequencies which is unrealistic for many physical processes. Therefore, in what follows, we always use a taper which is one more time differentiable on $(0, \infty)$ than it needs to be at the origin.

The following proposition gives a rigorous result leveraging the simple form for the Matérn family.

Proposition 2.3. *The Taper Condition for f_θ implies the Tail Condition for f_{tap} and $f_{\alpha,\nu}$.*

2.3 Principal Irregular Terms of a Covariance

The analysis has focused on spectral densities because it provides the most accessible theory. However, because the tapering is done in the spatial domain it would be practical to characterize the Taper or Tail Conditions in terms of the taper covariance directly. The concept of principal irregular terms (PIT) is a characterization of a stationary covariance function at the origin. For a stationary, isotropic covariance function, consider the series expansion of $C(h)$ about zero. An operational definition of the PIT of C is the first term as a function of h in this series expansion about zero that is not raised to an even power (Matheron, 1971). For the Matérn covariance function (4) the PIT is easy to identify. Let m denote the integer part of ν, if ν is noninteger then

$$C_{\alpha,\nu}(h) = \sum_{j=0}^{m} b_j h^{2j} - \frac{\Gamma(-\nu)}{\Gamma(\nu)2^{2\nu}} h^{2\nu} + O(h^{2m+2}), \quad \text{as } h \to 0, \tag{9}$$

and if $\nu = m$

$$C_{\alpha,\nu}(h) = \sum_{j=0}^{m} b_j h^{2j} + \frac{2(-1)^m \Gamma(m + 0.5)}{(2m+1)!\Gamma(m)\sqrt{\pi}} \log(h) h^{2m+1} + O(h^{2m+2}), \quad \text{as } h \to 0, \tag{10}$$

in either case the constants b_j depend only on ν and α and the PIT is the middle term on the right side of (9) and (10). The coefficient of the PIT simplifies to π and $\pi/6$ for $\nu = 0.5$ and $\nu = 1.5$ respectively.

The previous equations also imply that the Matérn covariance function is $2m$ times differentiable at the origin if $\nu > m$. From (5), (9) and (10), we confirm that the behavior at the origin of the covariance is related to the high frequency of the spectrum via the parameter ν. Specifically, differentiability of a function is equivalent to the number of finite moments for the Fourier transform.

Stein (1999a) discusses this loose definition of the PIT in more detail and suggests that for all models used in practice, the PIT is of the form bh^β, $b \in \mathbb{R}$, $\beta > 0$ and not an even integer, or $b \log(h) h^\beta$, $b \in \mathbb{R}$, β an even integer.

It is also simple to identify the PIT of the covariance tapers. Wu (1995); Gaspari and Cohn (1999) and Gneiting (2002) give several procedures to construct compactly supported covariance functions with arbitrary differentiability at the origin and at the support length. Most of these tapers are of polynomial type and hence having a PIT of the form Bh^μ. Moreover, under the Taper Condition, it is straightforward to show that the PIT associated with $C_{\mathsf{tap}} = C_\theta C$ and with C coincide.

In the case of a Matérn covariance function we have a one-to-one relationship between the PIT and the tail behaviour, whereas for polynomial tapers we are led to the following conjecture:

Conjecture 2.4. *Assume a polynomial isotropic covariance function C_θ in $\mathbb{R}^d$ that is integrable with* PIT *Bh^μ. Then the* PIT *and the tail behaviour are related by*

$$\lim_{\rho\to\infty} \rho^{\mu+d} f_\theta(\rho) = \left| B \cdot \frac{\mu!}{2} \left(\frac{2}{\pi}\right)^{(d+1)/2} \right|. \tag{11}$$

A rigorous proof of this conjecture and perhaps additional technical conditions would be based on a special case of a Tauberian theorem. In some special cases one can derive the tail behavior analytically from the tapered covariance and the conjecture is true for these cases. Examples of such covariance functions are polynomial tapers similar to the ones considered in Section 3 or wave covariances (Yaglom, 1987, Example 3, page 122).

Theorem 2.5. *Assume that the above conjecture holds. Let b be the* PIT *coefficient of $C_{\alpha,\nu}$. Assume a polynomial taper C_θ with* PIT *Bh^μ, $\mu \geq 2\nu$. Then the Tail Condition holds with*

$$\gamma = \begin{cases} 1\,, & \text{if } \mu > 2\nu\,, \\ (B+b)/b\,, & \text{if } \mu = 2\nu\,. \end{cases} \tag{12}$$

The result holds for any dimension d.

The connection of the PIT with the Tail Condition also suggests a practical scaling of the tapered covariance. As the PIT does not depend on the range parameter α, any Matérn covariance function with parameters ν and $\alpha^\star \neq \alpha$ satisfies the Tail Condition. This concept is equivalent to compatibility (see Stein, 1988, 1990b or Krasnits′kiĭ, 2000) and can be used to optimize the tapering performance by rescaling the range parameter to $\alpha^\star$ for the covariance $C_{\mathsf{tap}} = C_\theta C_{\alpha^\star,\nu}$. The intuition behind this rescaling is that for large ranges α a small taper length might be less efficient than tapering a small range $\alpha^\star$. In Section 3.2 we report results that indicate adjusting the scale in concert with tapering is slightly more efficient than tapering alone.

3 Finite Sample Accuracy and Numerical Efficiency

In this section we investigate the numerical convergence of the ratios (6) and (7) for different sample sizes, covariance function shape and the choice of taper. These results are complemented by timing studies for the sparse matrix implementations.

3.1 Practical Tapers

For the applications in this work we consider the spherical covariance and the two tapers developed in Wu (1995), and parametrized so that they have support in $[0, \theta)$. All three tapers are valid covariances in $\mathbb{R}^3$. The functions are plotted in Figure 1 and summarized in Table 1. Note that the spherical is linear at the origin and once differentiable at θ. Based on the theory from Section 2, relative to the Matérn smoothness parameter we use the spherical to taper for $\nu < 0.5$, Wu_1 for $\nu < 1.5$ and Wu_2 $\nu < 2.5$. If we admit the conjecture we can use the respective tapers for $\nu \leq 0.5$, $\nu \leq 1.5$, $\nu \leq 2.5$. Appendix B gives some additional analytical results.

3.2 Simulations

Throughout this section, we will focus on the stationary Matérn covariance function in $\mathbb{R}^2$. The factors in the simulation related to the covariance are the smoothness ν, the range α and the taper length θ. As data we select n locations within the unit square $\mathcal{D} = [0,1]^2$ sampled randomly and, if n is a perfect

square, locations on a regular grid. The spatial prediction is for the center location $\boldsymbol{x}^* = (0.5, 0.5)$ and the following quantities are calculated: the root mean squared error (MSE) for estimates of $Z(0.5, 0.5)$ using the actual and tapered covariance, *i.e.* $\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})$, $\mathrm{MSE}(\boldsymbol{x}^*, C_{\mathsf{tap}})$, and the naive estimate of the MSE $\varrho(\boldsymbol{x}^*, C_{\mathsf{tap}})$.

The MSE can be computed exactly for a fixed configuration of observation locations and so the only random element in these experiments is due to the locations being sampled from a uniform distribution over $\mathcal{D}$.

The first experiment examines the convergence analogous to infill asymptotics. The sample size n is varied in the range $[49, 784]$, 100 different sets of uniformly distributed locations are generated at each sample size and for sample sizes that are perfect squares a regular grid of locations is also considered. The covariance parameters are $\nu = 0.5, 1, 1.5$, $\theta = 0.4$ and the range α is fixed so that correlation decreases to 0.05 over the distance 0.4. Note that for the Matérn covariance, the value of α that achieves this criterion must be found numerically and will depend on the choice of ν. As a reference, Figure 1 graphs the covariances and taper functions. Outside a disk of radius 0.4 centered at $(0.5, 0.5)$ the field contributes little information for the prediction and this choice also minimizes any edge or boundary effects in the experiment.

Figure 2 summarizes the results. The convergence is considerably slower for (7). The higher the smoothing parameter, the larger the variation of the MSE ratios from the random locations. Note further that for smaller n the ratio of the regular grid is above the mean of the ratios arising from the random locations. This is not surprising, since for random patterns there are often more locations within the taper range. For the spherical taper with $\nu = 0.5$, $\gamma = 1.5$, *cf.* left lower panel.

The second simulation examines the influence of the taper shape and support on accuracy. The locations are fixed on a 20×20 grid in the unit square and we predict at $\boldsymbol{x}^* = (0.5, 0.5)$. We calculate the ratio of $\mathrm{MSE}(\boldsymbol{x}^*, C_{\mathsf{tap}})$ and $\mathrm{MSE}(\boldsymbol{x}^*, C_{\alpha,\nu})$ for different θ, ν and different tapers. Figure 3 gives the results. We

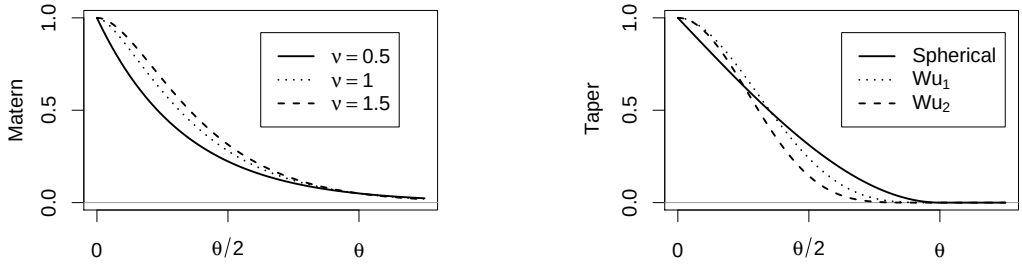

Figure 1: Matérn covariance with effective range θ (*i.e.* $\alpha = \theta/3, \theta/4, 0.21 \cdot \theta$), sill 1 and different smoothing parameters (left). Spherical, Wu_1 and Wu_2 tapers with taper length θ and sill 1 (right).

Table 1: Characteristics of the practical tapers used in the simulations. ($x_+ = \max\{0, x\}$.)

Taper	$C_\theta(h)$ for $h \geq 0$	PIT	Derivative(s) at zero	Valid taper for
Spherical	$\left(1 - \frac{h}{\theta}\right)_+^2 \left(1 + \frac{h}{2\theta}\right)$	$\frac{3h}{2\theta}$	1	$\nu < 0.5$
Wu_1	$\left(1 - \frac{h}{\theta}\right)_+^4 \left(1 + 4\frac{h}{\theta} + 3\frac{h^2}{\theta^2} + \frac{3h^3}{4\theta^3}\right)$	$-\frac{35h^3}{4\theta^3}$	3	$\nu < 1.5$
Wu_2	$\left(1 - \frac{h}{\theta}\right)_+^6 \left(1 + 6\frac{h}{\theta} + \frac{41h^2}{3\theta^2} + 12\frac{h^3}{\theta^3} + 5\frac{h^4}{\theta^4} + \frac{5h^5}{6\theta^5}\right)$	$-\frac{77h^5}{2\theta^5}$	5	$\nu < 2.5$

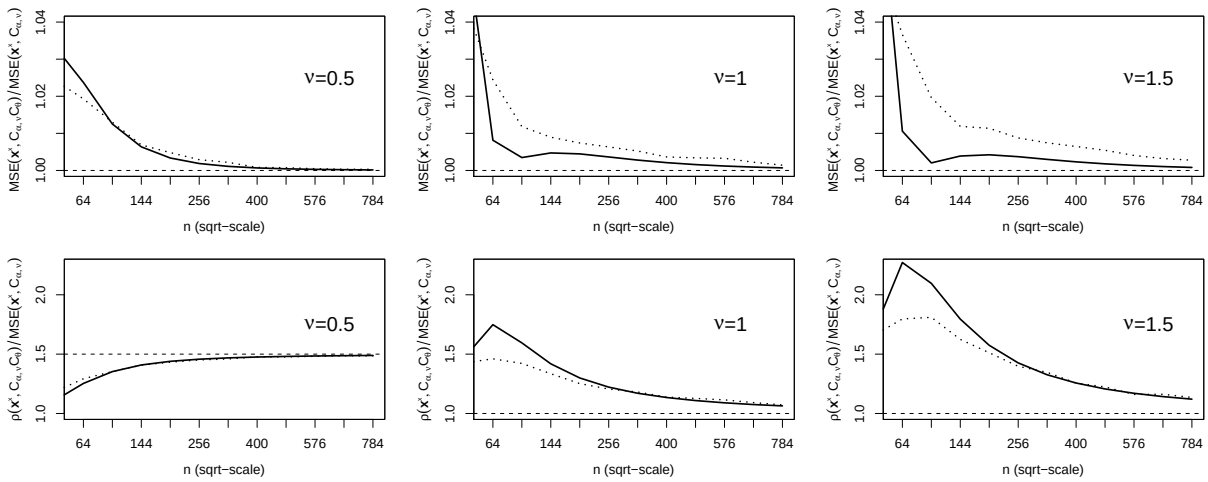

Figure 2: A comparision between taper estimate and the BLUP with respect to different true covariance functions. The ratio of the MSE of the tapered estimate to that of the BLUP is displayed in the top row and the ratio of the naive estimate of the MSE to its actual MSE in the bottom row, *cf.* ratios given by (6) and (7). The smoothness parameter was $\nu = 0.5, 1.0, 1.5$ for left, middle and right column respectively. We used a spherical (left column) and a Wu_2 tapers (middle and right columns). The solid line corresponds to the ratios of a regular grid in the unit square. 100 samples of n random points in the unit square were sampled and the dotted lines show the mean of the MSE ratios.

note that the ratio increases for increasing smoothness. For $\nu = 0.5$ and $\theta > 0.15$, all three tapers perform similarly. Wu_1 is slightly better than Wu_2 for comparable smoothness parameters. For the Wu_2 taper, θ should be chosen slightly bigger. This may be explained by the fact that it decays much faster than the spherical beyond $\theta/3$. The rough behavior for Wu_2 might be explained by numerical instabilities. If our goal is to be within 5% of the optimal MSE then according to Figure 3 a rule of thumb is to require 16 to 24 points within the support of the taper. A few more points should be added for very smooth fields. As a reference we also added the normed MSE of the nearest neighbor kriging with nearest neighbor distance θ to Figure 3. This approach performs very well, even if we include as few as 12 neighbors ($\theta = 0.1$).

We indicated in Section 2 that the original covariance could be scaled and tapered. To illustrate this approach, consider again the same simulation setup with effective range of $C_{\alpha,\nu}(\cdot)$ of 0.4. For a fixed $\theta = 0.15$ of a Wu_2 taper, we used $C_{\mathsf{tap}} = C_\theta C_{\alpha^\star,\nu}$ for different values of $\alpha^\star$. Figure 4 shows that by reducing the range, we can gain approximately one to two percent, *i.e.* with effective range between 0.2 and 0.3. Note that the values observed at the effective range of 0.4 correspond to the values at $\theta = 0.15$ in the corresponding panels of the last column of Figure 3.

Finally, we were curious about what would happen if we simply tapered with a hard threshold, *i.e.* use the "top hat" taper $I_{\{h \le \theta\}}$. This is a naive approach and mimics the idea of nearest neighbors. The resulting matrices $\mathbf{C}$ in (1) are not necessarily positive definite for all θ. Neglecting the statistical paradigm to work with positive definite covariance matrices, top hat tapers often lead to numerical instabilities and the MSE ratios are inferior to those from positive definite tapers.

3.3 Numerical Performance

For symmetric, positive definite matrices $\mathbf{C}$, the estimator in (1) is found by first performing a Cholesky factorization on $\mathbf{C} = \mathbf{A}\mathbf{A}^\mathsf{T}$. Then one successively solves the triangular systems $\mathbf{A}\boldsymbol{w} = \mathbf{Z}$ and $\mathbf{A}^\mathsf{T}\boldsymbol{u} = \boldsymbol{w}$ giving $\boldsymbol{u} = \mathbf{C}^{-1}\mathbf{Z}$. The final step is the dot product $\boldsymbol{c}^{*\mathsf{T}}\boldsymbol{u}$. The common and widely used numerical

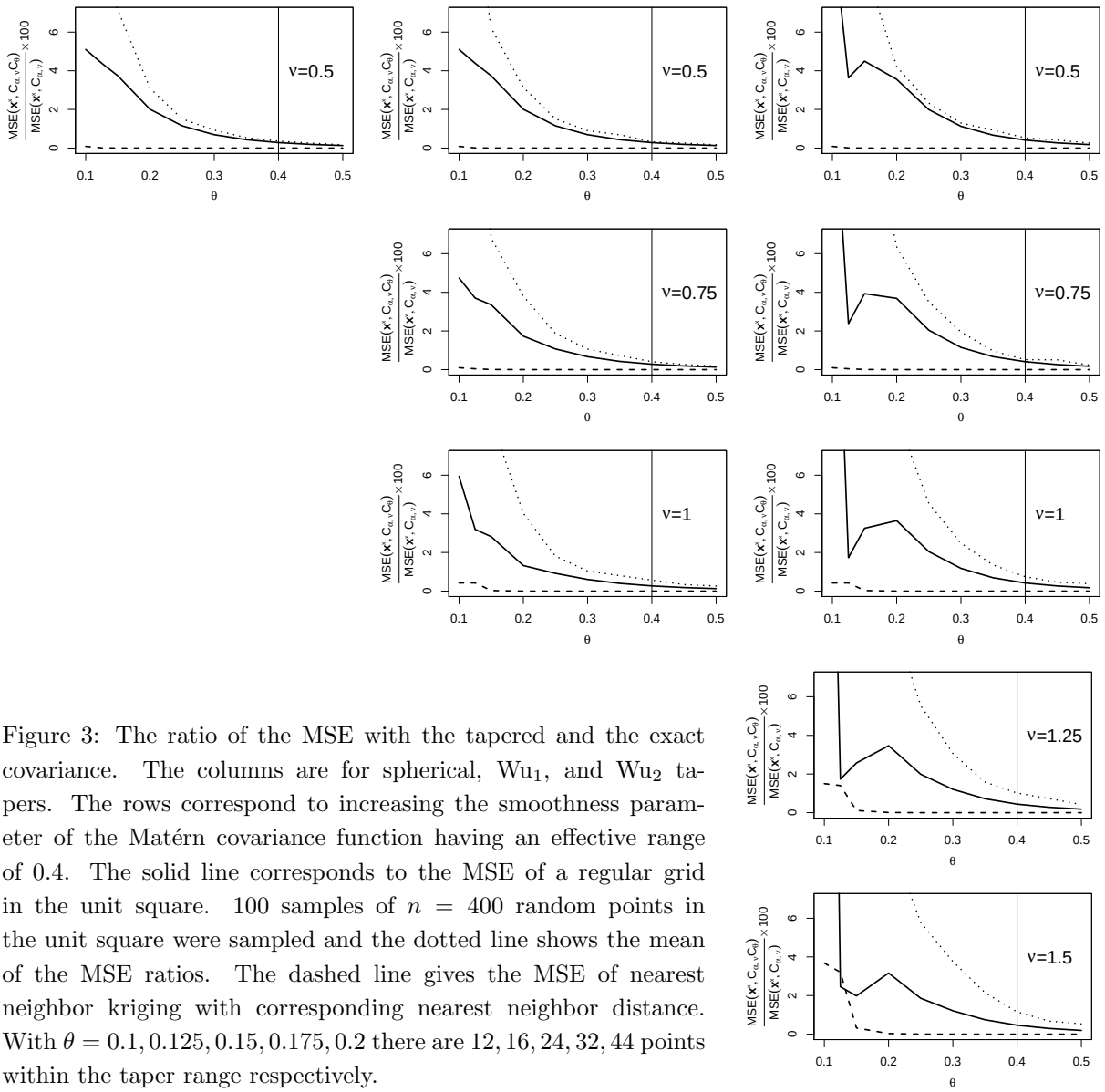

Figure 3: The ratio of the MSE with the tapered and the exact covariance. The columns are for spherical, Wu_1, and Wu_2 tapers. The rows correspond to increasing the smoothness parameter of the Matérn covariance function having an effective range of 0.4. The solid line corresponds to the MSE of a regular grid in the unit square. 100 samples of $n = 400$ random points in the unit square were sampled and the dotted line shows the mean of the MSE ratios. The dashed line gives the MSE of nearest neighbor kriging with corresponding nearest neighbor distance. With $\theta = 0.1, 0.125, 0.15, 0.175, 0.2$ there are $12, 16, 24, 32, 44$ points within the taper range respectively.

software packages MATLAB and R contain a toolbox (Gilbert *et al.*, 1992) and a library SPARSEM (Ihaka and Gentleman, 1996) respectively with sparse matrix techniques functions to perform the Cholesky factorization.

The performance of the factorization depends on the number of non-zero elements of $\mathbf{C}$ and on how the locations are ordered. We first discuss the storage gain of sparse matrices. A sparse matrix is stored as the concatenation of the vectors representing its rows. The non-zero elements are identified by two integer vectors. An $n \times m$ sparse matrix $\mathbf{S}$ with z non-zeros entries requires

$$8z + 4z + 4n + 1 \text{ bytes}, \tag{13}$$

if we have "typical" precision with 8-byte reals and 4-byte integers. For a regular equispaced $n \times m$ grid

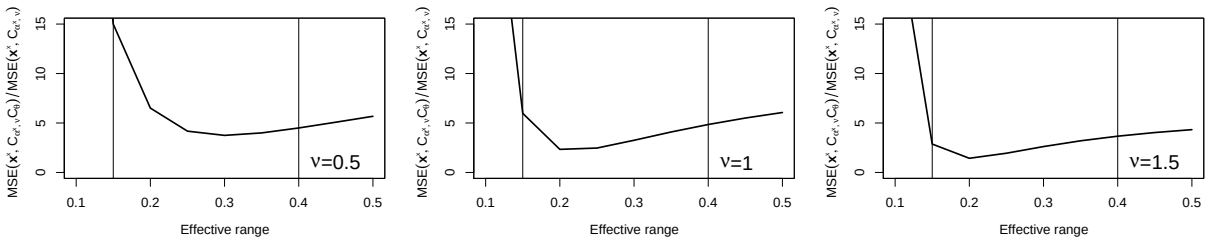

Figure 4: Root mean squared errors ratios with tapered and exact covariance. Similar to the last column of Figure 3 but using $C_{\mathsf{tap}} = C_\theta C_{\alpha^\star,\nu}$. The abscisse denotes the effective range associated with $\alpha^\star$. The true covariance has an effective range of 0.4. The taper length is $\theta = 0.15$.

with spacing h and taper support θ, the number of non-zero elements is given by

$$z = \sum_{l=0}^{n-1}(1 + I_{\{l>0\}})(n-l)\sum_{k=0}^{K_l-1}(1 + I_{\{k>0\}})(m-k), \quad K_l = \min\Big(m, \Big\lceil \big((c/h)^2 - l^2\big)_+^{1/2} \Big\rceil\Big), \qquad (14)$$

with $(x)_+ = \max\{0, x\}$ and $\lceil\cdot\rceil$ the biggest integer function. For irregular grids, we cannot determine directly the number of non-zero elements, but the formula can be used as a fairly good approximation if the locations are uniformely distributed within a rectangle.

The reduction in storage space allows us to work with much bigger problems. We have to distinguish between the limitations due to the physical restrictions (RAM, available access memory) and the limitations due to the software (addressing of arrays). The former determines nowadays still the upper bound of the problem size. To illustrate the latter, MATLAB, for example, can handle matrices with up to $2^{28} - 1$ elements[2], or sparse matrices with up to roughly 2^{29} non-zero entries.

It is not obvious that the Cholesky factor $\mathbf{A}$ of a sparse matrix will also be sparse. Define the semi-bandwidth s of a symmetric matrix $\mathbf{S}$ as the smallest value for which $\mathbf{S}_{i,i+s} = 0$, for all i, then the Cholesky factor $\mathbf{A}$ has a semi-bandwidth of at most s. If the locations are not "ordered", then $\mathbf{A}$ is virtually "full". But by ordering the locations deliberately, sparsity of the factor can be insured. For ordered $n \times m$ grids with the numbering along the smaller dimension first, say n, the semi-bandwidth is

$$(n-1)L + K_L - 1, \quad L = \operatorname*{argmin}_l\{K_l \geq 0\},$$

where K_l is given by (14). Other possible permutations are the Cuthill–McKee or minimum-degree ordering. See Figure 5 for an illustration of the effect of ordering. Although having a much larger semi-bandwidth, the minimum degree ordering[3] performs slightly better in computational cost and storage than the reverse Cuthill–McKee ordering (George and Liu, 1981). In the R library SPARSEM, there exist no explicit permutation functions and the sparse Cholesky decomposition relies on the sparse factorization algorithm by Ng and Peyton (1993).

Figure 6 compares the performance of the SPARSEM library of R and the SPARSE toolbox of MATLAB on a Linux powered 2.6 GHz Xeon processor with 4 Gbytes RAM. The sparse and standard approaches are approximately of the order of n and n^3 respectively. We notice that for all grid sizes n MATLAB outperforms R and for small n, the SPARSEM library is not efficient. Gilbert *et al.* (1992) indicate how to improve the computational costs for smaller order. However, most of the functions are built-in such that the user has no ability to manipulate the source code.

[2] `http://www.mathworks.com/support/solutions/data/1103.shtml`

[3] `http://www.mathworks.com/access/helpdesk/help/techdoc/ref/symmmd.shtml`

4 Application

In this section we apply covariance tapering in kriging to a large irregularly spaced dataset. We use aggregated monthly precipitation for April 1948 at 5,909 stations in the US[4]. For a detailed description of

[4]Available at `http://www.cgd.ucar.edu/stats/Data/US.monthly.met.html`.

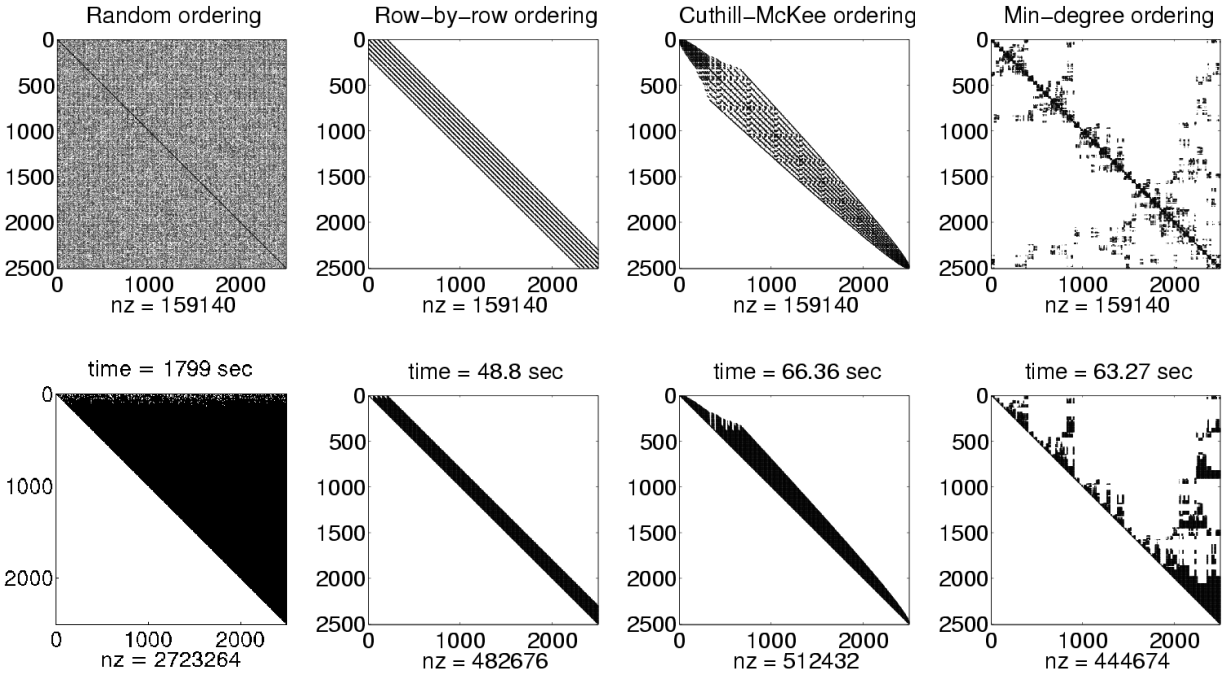

Figure 5: Influence of ordering on the performance. The top row shows the structure of the covariance matrix, the bottom row its upper triangular Cholesky factor. The first column is for an arbitrary numbering, the second for a row-by-row numbering, the third column is after a reverse Cuthill-McKee reordering, the last after a minimum-degree reordering. We considered an equispaced 50×50 grid in the unit square with taper length 0.05. The indicated time is for solving 100 linear systems in MATLAB and nz states the number of nonzero elements in the matrix.

Table 2: Necessary times to create the prediction anomaly field in R with sparse and classical techniques. The result of the sparse approach is depicted in Figure 8. The matrix $\tilde{\mathbf{C}}$ contains as columns the vectors $\boldsymbol{c}^*$ for the different points on the prediction grid. (Linux, 2.6 GHz Xeon processor with 4 Gbytes RAM, SPARSEM, FIELDS and BASE libraries.)

Action	Time (sec)			
	Sparse	Sparse+FFT	Classic+OPT	Classic
1 Reading data, variable setup	0.54	0.54	0.54	0.54
2 Creating the matrix $\mathbf{C}$	6.35	6.35	21.59	41.34
3 Solving $\mathbf{C}\boldsymbol{x} = \mathbf{Z}$ Cholesky	0.28	0.28	169.09	169.09
3 Solving $\mathbf{C}\boldsymbol{x} = \mathbf{Z}$ Backsolve	0.03	0.03	6.13	6.13
4 Multiplying $\tilde{\mathbf{C}}^{\mathsf{T}}$ with $\mathbf{C}^{-1}\mathbf{Z}$	733.82	26.99	1830.86	4638.01
5 Creating the figure	6.19	6.19	6.19	6.19
Total	747.12	40.92	2034.40	4859.81

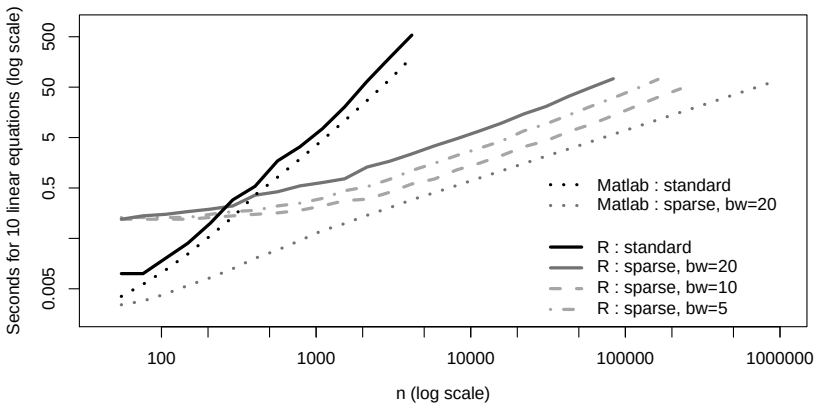

Figure 6: Comparison of the performance between R and MATLAB. A positive definite taper was applied to an equispaced one-dimensional grid of size n in $[0, 1]$. The range of the taper was such that the semi-bandwidth (bw) is 20, 15 or 10. Standard refers to Cholesky decomposition and two Backsolves.

the data we refer to Johns *et al.* (2003). Instead of working with the raw data, we standardize the square root values. The resulting values represent anomalies and are closer to a Gaussian distribution than the raw data (Johns *et al.*, 2003). Further, there is evidence that the anomaly field is closer to being second order stationary compared to the raw scale (Fuentes *et al.*, 1998). For the estimation of the second order structure of the anomalies we refer again to Johns *et al.* (2003). They justify a slight anisotropy of 0.85 in the North–South and East–West direction. The resulting fitted covariance structure is a mixture of two exponential covariances with range parameter α of 40.73 and 523.73 miles with respective sill ϕ of 0.277 and 0.722. We rescale the resulting covariance structure with a factor of 5 as explained in Section 2 and taper with a spherical covariance with a range of 50 miles. The taper range was chosen as small as possible but such that all the locations had at least one point within the taper range. On average, each point has approximately 20 other locations within 50 miles (see Figure 7). The resulting sparse covariance matrix $\mathbf{C}$ has only 0.35% non-zero elements. The prediction is then performed on a regular 0.025×0.05 latitude/longitude grid within the coterminous US. Figure 8 shows the kriged anomaly field consisting of more than 6.6×10^5 predicted points. Table 2 summarizes the required times to construct the predicted field and the displayed figure with sparse and classical techniques. The sparse approach is faster by a factor of over 560 for step 3. Using the FFT approach with the library FIELDS we can speed up the time consuming step 4 considerably (column Sparse+FFT). The Classic+OPT approach consists of classical techniques where costly loops are programmed in Fortran.

Finally, notice that the predicted anomaly field can be back-transformed using predicted or interpolated climatological means and standard deviations.

5 Discussion

An omnipresent example in spatial statistics is the prediction onto a large grid of a correlated quantity. In this article, we showed that truncating the covariance function to zero with appropriate polynomial tapers preserves asymptotic optimality and results in tremendous efficiencies in computation. For sparse matrices, one can use well established algorithms in numerical analysis to handle the sparse systems. Commonly used software packages such as R or MATLAB contain libraries or toolboxes with the required functions. We showed that for large fields tapering results in a significant gain in storage and computation. In the precipitation dataset we achieve a speedup of more than 560 to solve the linear system. In fact, the

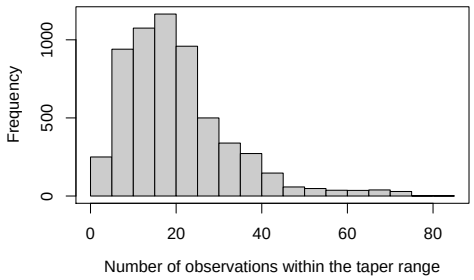

Figure 7: Histogram of the number of observations within the taper range. Three locations had just one observation within the range. The median is 18 observations and the mean is slightly over 20.

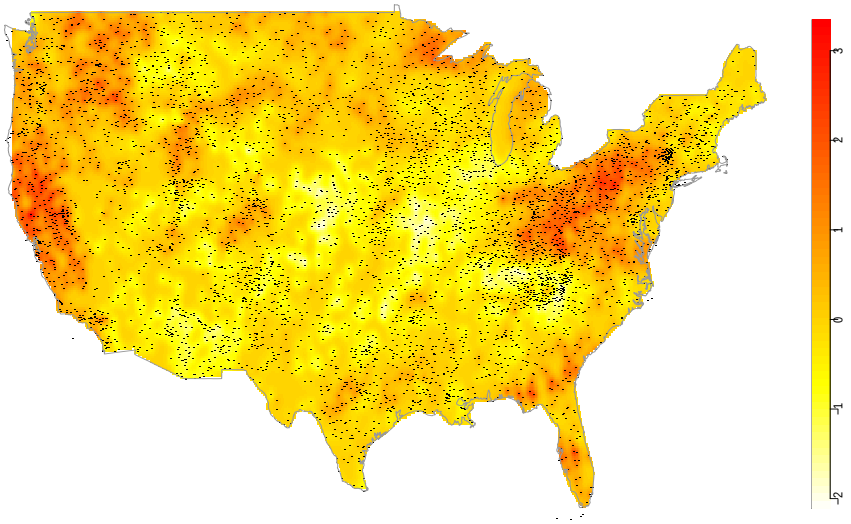

Figure 8: Kriged surface of the precipitation anomaly field of April 1948. The dots represent the 5,906 locations of the observations.

manageable size of the observed and predicted fields can be far bigger than with classical approaches.

Although we developed the theory for zero-mean processes with continuous covariance functions, these assumptions are not restrictive. Suppose a spatial process of the form

$$Y(\boldsymbol{x}) = \boldsymbol{m}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{15}$$

where $\boldsymbol{m}$ is a known function in $\mathbb{R}^p$ and $\boldsymbol{\beta}$ is an unknown parameter in $\mathbb{R}^p$. Similar to equation (1), the BLUP of $Y(\boldsymbol{x}^*)$ is then given by

$$\hat{Y}(\boldsymbol{x}^*) = \boldsymbol{c}^{\mathsf{T}}\mathbf{C}^{-1}\left(\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}}\right) + \boldsymbol{m}(\boldsymbol{x}_0)^{\mathsf{T}}\hat{\boldsymbol{\beta}}, \qquad \text{where } \hat{\boldsymbol{\beta}} = (\mathbf{M}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{Y} \tag{16}$$

with $\mathbf{M} = \big(\boldsymbol{m}(\boldsymbol{x}_1), \ldots, \boldsymbol{m}(\boldsymbol{x}_n)\big)^{\mathsf{T}}$. The sparse approach could be used with an iterative procedure illustrated as follows. We estimate the mean structure, *i.e.* the vector $\boldsymbol{\beta}$ in (15), via ordinary least squares (OLS), then $\mathbf{Y} - \mathbf{M}\hat{\boldsymbol{\beta}}^*$ is kriged yielding $\mathbf{Z}^*$. With OLS on $\mathbf{Y} - \mathbf{Z}^*$ we obtain a second estimate $\hat{\boldsymbol{\beta}}^*$ and so forth. This convenient back-fitting procedure converges to the BLUP and a few iterations usually suffice to obtain precise results. If p is not too big, the BLUP can be also obtained by solving $p + 2$ linear systems as given by equation (16) using the approach delineated in this paper. From the theoretical aspect, Yadrenko (1983, page 138), and Stein (1990b) show that if the difference of the true mean structure and

the presumed mean structure is sufficiently smooth, then Theorem 2.2 still holds. Further, Stein (1999a, Theorem 4.2), gives analogous results for processes with a nugget effect.

Conjecture 2.4 was stated for polynomial tapers only. We believe that a similar statement might be true for a much broader class of covariance functions. It is straightforward to show that the multiplicative constant is a general case of the constant for the Matérn covariance function. Necessary conditions would be that the spectral density has unbounded support and that the PIT exists. The conjecture holds for many covariance functions used in practice.

It remains an open question how accurate the tapering approximation will be for nonstationary problems. However, our numerical results suggest that tapering is effective for different correlation ranges. A possible strategy is to choose a conservative taper that is accurate for the smallest correlation range in the domain. Of course the identification of nonstationary covariances is itself difficult for large datasets but perhaps sparse techniques will also be useful in covariance estimation.

Although there are still many open questions regarding the theoretical properties of tapering and its practical application, we believe that this work is a useful step toward the analysis of large spatial problems that often have substantial scientific importance.

Appendix A: Proofs

Proof. (Theorem 2.2) The spectral density of the Matérn covariance satisfies Condition (2.1) of Stein (1993) and with the Tail Condition, *i.e.* (8), Theorems 1 and 2 of Stein (1993) hold. □

Proof. (Propostion 2.3) Without loss of generality, we suppose that $\alpha = 1$ and so $f_{1,\nu}(\|\boldsymbol{\omega}\|) = M_1/(1 + \|\boldsymbol{\omega}\|^2)^{\nu+d/2}$, $\nu > 0$. We need to prove that the limit

$$\lim_{\|\omega\|\to\infty} \frac{\int_{\mathbb{R}^d} f_{1,\nu}(\|\boldsymbol{x}\|) f_\theta(\|\boldsymbol{x} - \boldsymbol{\omega}\|)\, \mathsf{d}\boldsymbol{x}}{f_{1,\nu}(\|\boldsymbol{x}\|)} \tag{17}$$

exists and is not zero. As the spectral densities are radially symmetric, we choose an arbitrary direction for $\boldsymbol{\omega}$ and we set $\|\boldsymbol{x}\| = r\|\boldsymbol{u}\|$ and $\|\boldsymbol{\omega}\| = \rho\|\boldsymbol{v}\|$, with $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$. The convolution reduces to

$$\int_{\mathbb{R}^d} f_{1,\nu}(\|\boldsymbol{x}\|) f_\theta(\|\boldsymbol{x} - \boldsymbol{\omega}\|)\, \mathsf{d}\boldsymbol{x} = \int_{\partial B_d} \int_0^\infty f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\, \mathsf{d}r\, \mathsf{d}U(\boldsymbol{u}),$$

where ∂B_d is the surface of the unit sphere in $\mathbb{R}^d$ and U is the uniform probability measure on ∂B_d. We integrate over the three annuli A,B,C discribed by the radii $[0, \rho - \Delta], [\rho - \Delta, \rho + \Delta], [\rho + \Delta, \infty)$ (as illustrated in Figure 9 for $d = 2$). We will bound each part under the ansatz of choosing a sufficiently large ρ and $\Delta = \mathcal{O}(\rho^\delta)$, for some well-chosen $0 < \delta < 1$. The basic idea is that we can bound the inner

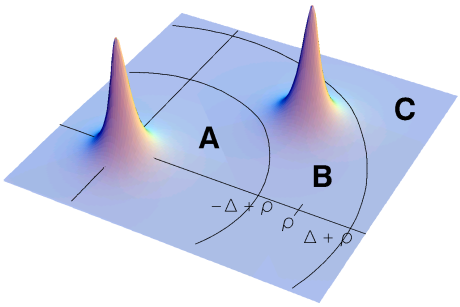

Figure 9: Separation of the convolution integral into three annuli, illustration for two dimensions.

integral independently of $\boldsymbol{u}$ for the respective intervals. Then the outer integrals are just the surface of the hypersphere, *i.e.* $2\pi^{d/2}/\Gamma(n/2)$, times the inner bound. Within the ball A, the Taper Condition implies that $f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|)$ is bounded by $M/(1+(\rho-\Delta)^2)^{\nu+d/2+\epsilon}$. Hence,

$$\int_0^{\rho-\Delta} f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r \leq \frac{M}{(1+(\rho-\Delta)^2)^{\nu+d/2+\epsilon}} \int_0^{\rho-\Delta} f_{1,\nu}(r) r^{d-1}\,\mathsf{d}r.$$

As $f_{1,\nu}$ is a density in $\mathbb{R}^d$ the last integral is finite. Since $f_{1,\nu}$ is monotonically decreasing in ρ, we have for the second part

$$\int_{\rho-\Delta}^{\rho+\Delta} f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r \leq f_{1,\nu}(\rho) \int_{\rho-\Delta}^{\rho+\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r. \tag{18}$$

Again, as f_θ is a density in $\mathbb{R}^d$ the last integral is finite and is positive for all $\Delta > 0$.

For the last term, we have

$$\int_{\rho+\Delta}^{\infty} f_{1,\nu}(r) f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r \leq f_{1,\nu}(\rho) \int_{\rho+\Delta}^{\infty} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r.$$

As ρ tends to infinity, the integral will tend to zero.

Now, as $\rho, \Delta \to \infty$ with $\Delta/\rho \to 0$, the fraction (17) is bounded by

$$\lim_{\rho\to\infty} \frac{(1+\rho^2)^{\nu+d/2}}{(1+(\rho-\Delta)^2)^{\nu+d/2+\epsilon}} \int_{\partial B_d} \int_{\rho-\Delta}^{\rho-\Delta} f_{1,\nu}(r) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) + \int_{\partial B_d} \int_{\rho-\Delta}^{\infty} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) = 1.$$

To show that the limit is strictly positive, consider annulus B

$$\int_{\mathbb{R}^d} f_{1,\nu}(\|\boldsymbol{x}\|) f_\theta(\|\boldsymbol{x} - \boldsymbol{\omega}\|)\,\mathsf{d}\boldsymbol{x} \geq f_{1,\nu}(\rho+\Delta) \int_{\partial B_d} \int_{\rho-\Delta}^{\rho-\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}),$$

then for all $\rho > \rho_0$, the integral is positive and has a lower bound. Further, the fraction (17) has limit

$$\lim_{\rho\to\infty} \frac{(1+\rho^2)^{\nu+d/2}}{(1+(\rho+\Delta)^2)^{\nu+d/2}} \int_{\partial B_d} \int_{\rho-\Delta}^{\rho+\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) = 1.$$

□

Proof. (Theorem 2.1) The proof of the theorem is a direct consequence of the Theorem 2.2 and Proposition 2.3. □

Proof. (Theorem 2.5) The case of $\mu > 2\nu$ corresponds to Proposition 2.3 and is independent of the conjecture. Consider the case $\mu = 2\nu$. It is straightforward to show that the conjecture holds for the Matérn covariance function. For polynomial tapers the existence of $\lim_{\rho\to\infty} \rho^{\mu+d} f_\theta(\rho)$ implies that there exists two constants $M_1 = M_1(\rho_0)$ and $M_2 = M_2(\rho_0)$ such that $M_1 \leq \rho^{\mu+d} f_\theta(\rho) 2/\mu!(\pi/2)^{(d+1)/2} \leq M_2$ for $\rho > \rho_0$. The proof follows closely the one of Proposition 2.3 for the upper bound with $\epsilon = 0$. For the lower bound, annulus A has to be taken into account to show that

$$\frac{M_1 + b}{b} \leq \lim_{\rho\to\infty} \frac{f_{\alpha,\nu}(\rho+\Delta)}{f_{\alpha,\nu}(\rho)} \int_{\partial B_d} \int_{\rho-\Delta}^{\rho-\Delta} f_\theta(\|r\boldsymbol{u} - \rho\boldsymbol{v}\|) r^{d-1}\,\mathsf{d}r\,\mathsf{d}U(\boldsymbol{u}) \leq \frac{M_2 + b}{b}.$$

As $\rho \to \infty$, we can choose M_1 and M_2 arbitrary close to B. □

Appendix B: Spectral Densities of Taper Functions

Let C be an isotropic covariance function in $\mathbb{R}^d$. The corresponding spectral density can be obtained by

$$f(\rho) = (2\pi)^{-d/2} \int_0^\infty (\rho r)^{-(d-2)/2} \mathcal{J}_{(d-2)/2}(\rho r) r^{d-1} C(r) \, \mathsf{d}r,$$

where $\mathcal{J}$ is the Bessel function of the first kind. For $d = 1$ and $d = 3$, $\mathcal{J}_{(d-2)/2}$ can be written as a function of r, a cosine and a sine function respectively. For polynomial tapers, it is thus straightforward to obtain the spectral densities. As the expressions are rather long, we only give the tail behavior. In one dimension we have:

$$\text{Spherical: } \lim_{\rho\to\infty} \rho^2 f_\theta(\rho) = \frac{3}{2\pi\theta}, \qquad \text{Wu}_1: \ \lim_{\rho\to\infty} \rho^4 f_\theta(\rho) = \frac{105}{2\pi\theta^3}, \qquad \text{Wu}_2: \ \lim_{\rho\to\infty} \rho^6 f_\theta(\rho) = \frac{4620}{\pi\theta^5}.$$

For $d = 3$, we observe that the tail behavior is decreased by ρ^2. For $d = 2$, we can only numerically verify the tail behavior. The tail behavior of these and other invested tapers matches the conjecture.

References

Abramowitz, M. and Stegun, I. A., editors (1970). *Handbook of Mathematical Functions*. Dover, New York.

Cleveland, W. S., Grosse, E., and Shyu, W. (1992). Local regression models. In Chambers, J. and Hastie, T., editors, *Statistical Models in S*, 309–376. Wadsworth and Brooks, Pacific Grove.

Cressie, N. A. C. (1990). The origins of kriging. *Mathematical Geology*, **22**, 239–252.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons Inc., New York, revised reprint.

Diamond, P. and Armstrong, M. (1984). Robustness of variograms and conditioning of kriging matrices. *Journal of the International Association for Mathematical Geology*, **16**, 809–822.

Fuentes, M., Kelly, R., Kittel, T., and Nychka, D. (1998). Spatial prediction of climate fields for ecological models. Technical report, Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, CO.

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

George, A. and Liu, J. W. H. (1981). *Computer solution of large sparse positive definite systems*. Prentice-Hall Inc., Englewood Cliffs, N.J.

Gilbert, J. R., Moler, C., and Schreiber, R. (1992). Sparse matrices in MATLAB: design and implementation. *SIAM Journal on Matrix Analysis and Applications*, **13**, 333–356.

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

Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Johns, C., Nychka, D., Kittel, T., and Daly, C. (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association*, **98**, 796–806.

Krasnits′kiĭ, S. M. (2000). On a spectral condition for the equivalence of Gaussian measures corresponding to homogeneous random fields. *Theory of Probability and Mathematical Statistics*, **60**, 95–104.

Madych, W. R. and Potter, E. H. (1985). An estimate for multivariate interpolation. *Journal of Approximation Theory*, **43**, 132–139.

Matheron, G. (1971). The theory of regionalized variables and its applications. *Cahiers du Centre de Morphologie Mathématiques*, **No. 5**, Fontainebleau, France.

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

Stein, M. (1990a). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *The Annals of Statistics*, **18**, 850–872.

Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, **16**, 55–63.

Stein, M. L. (1990b). Bounds on the efficiency of linear predictions using an incorrect covariance function. *The Annals of Statistics*, **18**, 1116–1138.

Stein, M. L. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters*, **17**, 399–404.

Stein, M. L. (1997). Efficiency of linear predictors for periodic processes using an incorrect covariance function. *Journal of Statistical Planning and Inference*, **58**, 321–331.

Stein, M. L. (1999a). *Interpolation of Spatial Data*. Springer-Verlag, New York.

Stein, M. L. (1999b). Predicting random fields with increasing dense observations. *The Annals of Applied Probability*, **9**, 242–273.

Stein, M. L. and Handcock, M. S. (1989). Some asymptotic properties of kriging when the covariance function is misspecified. *Mathematical Geology*, **21**, 171–190.

Warnes, J. J. (1986). A sensitivity analysis for universal kriging. *Mathematical Geology*, **18**, 653–676.

Wu, Z. M. (1995). Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, **4**, 283–292.

Yadrenko, M. Ĭ. (1983). *Spectral theory of random fields*. Translation Series in Mathematics and Engineering. Optimization Software Inc. Publications Division, New York.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions. Vol. I.* Springer Series in Statistics. Springer-Verlag, New York.

Yakowitz, S. J. and Szidarovszky, F. (1985). A comparison of Kriging with nonparametric regression methods. *Journal of Multivariate Analysis*, **16**, 21–53.