A Multi-resolution Gaussian process model for the analysis of large spatial data sets.

Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling,
Finn Lindgren, and Stephan Sain *

August 13, 2013

Abstract

A multi-resolution model is developed to predict two-dimensional spatial fields based on irregularly spaced observations. The radial basis functions at each level of resolution are constructed using a Wendland compactly supported correlation function with the nodes arranged on a rectangular grid. The grid at each finer level increases by a factor of two and the basis functions are scaled to have a constant overlap. The coefficients associated with the basis functions at each level of resolution are distributed according to a Gaussian Markov random field (GMRF) and take advantage of the fact that the basis is organized as a lattice. Several numerical examples and analytical results establish that this scheme gives a good approximation to standard covariance functions such as the Matérn and also has flexibility to fit more complicated shapes. The other important feature of this model is that it can be applied to statistical inference for large spatial datasets because key matrices in the computations are sparse. The computational efficiency applies to both the evaluation of the likelihood and spatial predictions.

Keywords: Spatial estimator, Kriging, Fixed Rank Kriging, Sparse Cholesky Decomposition, Multi-resolution

1 Introduction

Statistical methodology for spatial data is a well developed field and has roots in geostatistics and multivariate analysis. More recently the breakthroughs in Bayesian hierarchical models have added rich new classes of models for handling heterogenous

*Douglas Nychka, is Senior Scientist, National Center for Atmospheric Research, PO Box 3000, Boulder CO 30307-3000 (nychka@ucar.edu), Soutir Bandyopadhyay, is Assistant Professor, Lehigh University, Bethlehem, PA, 18015, Dorit Hammerling is Post Doctoral Scientist, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC, 27709-4006, Finn Lindgren is Lecturer, University of Bath, Bath, BA2 7AY, UK and Stephan Sain is Scientist, National Center for Atmospheric Research, PO Box 3000, Boulder CO 30307-3000.

spatial data and indirect measurements of spatial processes (Banerjee et al. (2003), Cressie and Wikle (2011)). This development in spatial statistics is coincident with emerging challenges in the geosciences involving new types of observations and comparisons of such observations to complex numerical models. For example, as attention in climate science shifts to understanding the regional and local changes in future climate there is a need to analyze high resolution simulations from climate models and to compare them to surface and remotely sensed observations at fine levels of details. These kinds of geoscience applications are characterized by large numbers of spatial locations. The application of standard techniques is often not feasible or at least will take an unacceptably long time given standard algorithms and typical computational resources. Moreover, geophysical processes tend to have a multi-scale character over space that requires statistical methods that allow for potentially complicated spatial dependence beyond a simple parametric model that adjusts for a correlation range and process smoothness. This work develops a new statistical model that addresses both of these challenges; our model is applicable to large data sets and supports a more flexible covariance structure that can be a mixture of more standard covariance functions. Thus our model fills a gap in current statistical methodology.

We assume that spatial observations $\{\boldsymbol{y}_i\}$ are made at unique two-dimensional spatial locations, $\{\boldsymbol{x}_i\}$, for $1 \leq i \leq n$, according to the additive model:

$$\boldsymbol{y}_i = Z_i^T \boldsymbol{d} + g(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i, \tag{1}$$

where Z is a matrix of covariates and $\boldsymbol{d}$ a vector of linear parameters, g is a smooth Gaussian process and $\boldsymbol{\epsilon}_i$ are mean zero measurement errors. The parameters $\boldsymbol{d}$ represent fixed effects in this model.

The statistical problem in this setting is to determine g at locations where observations are not available and quantify the uncertainty of the spatial predictions. Given our main goal to develop an acceptable methodology to handle large data sets, we seek to balance the complexity of the models and methodology with feasibility for effective data analysis. We will focus on maximum likelihood estimates of parameters in the covariance and other model components. For prediction we will adopt the conditional distribution of g given the data and other statistical parameters. Our approach com-

bines the representation of a field using a multi-resolution (MR) basis with statistical models for the coefficients as a process on a lattice. In this sense it is a blending of ideas from fixed rank Kriging (Katzfuss and Cressie 2011, Cressie and Johannesson 2008) and stochastic partial differential equations (SPDE) including the work in Lindgren and Rue (2007), Rue and Held (2005) and Lindgren et al. (2011) (LR2011). It is useful to view the unknown spatial process in (1) as a sum of L independent processes, $g_l(\boldsymbol{x})$, for $1 \leq l \leq L$, marginal variances $\{\rho\alpha_l\}$, and

$$g(\boldsymbol{x}) = \sum_{l=1}^{L} g_l(\boldsymbol{x}), \tag{2}$$

Here the parameter $\rho > 0$ is useful as a leading scaling parameter for the covariance matrix and the elements of $\boldsymbol{\alpha}$ sum to one. In this way the overall spatial dependence of g can be much more complex than the spatial dependence of each of the individual components. Each component, g_l is defined through a basis function expansion as

$$g_l(\boldsymbol{x}) = \sum_{j=1}^{m(l)} c_j^l \phi_{j,l}(\boldsymbol{x}). \tag{3}$$

where ϕ_j^l, $1 \leq j \leq m(l)$, is a sequence of fixed basis functions and $\boldsymbol{c}^l$ is a vector of coefficients distributed multivariate normal with mean zero and covariance matrix, $\rho\boldsymbol{P}_l$. $\boldsymbol{P}_l$ may also depend on additional parameters. Thus the model for g is a sum of fixed basis functions with stochastic coefficients.

Our two main ideas address the basis functions and the covariance model for the coefficients. We use families of radial basis functions that are organized on regular grids of increasing resolution. These radial basis functions have compact support and like wavelet bases give computational efficiencies because of this feature. In our treatment, each increase in resolution will be by a factor of two and the levels associated with finer spatial scales will have more basis functions. Conversely, the representation has a parsimony in that the coarser scales require fewer basis functions to approximate the stochastic processes. The spatial dependence among the coefficients for each level of resolution is modeled using a Gaussian Markov random field (GMRF), specifically a spatial autoregressive (SAR) model. The fact that the basis functions are organized on a lattice gives the SAR a simple form along with its precision matrix, which we

denote as $\boldsymbol{Q}_l = \boldsymbol{P}_l^{-1}$. The benefit of this approach is that $\boldsymbol{Q}_l$ is sparse even though $\boldsymbol{P}_l$ itself can be dense. Thus, g_l can exhibit long range correlations among coefficients widely separated in the lattice even though the precision matrix is sparse.

We have found that this combination of MR bases with companion GMRFs for the coefficients at each level can approximate standard families of covariance functions such as the Matérn, but also provides a rich model for more general spatial dependence. It should be noted that we make no assumption on the observation or prediction locations even though the latent components of our model will exploit regular grids. We are also able to give some analytical results that suggest why this model can approximate a range of spatial processes exhibiting different degrees of smoothness.

Many of the ingredients for this model are not new, however, their particular combination with a view towards efficient computations for large and irregular spatial data sets has not been exploited in previous works. The key is to introduce sparsity into the computations in a way that does not compromise covariance models with long range correlations and models with many degrees of freedom. This is achieved by using compactly supported radial basis functions and computing directly the *precision* matrix of the basis coefficients, not the covariance matrix. In addition we add a normalization of the marginal process variance that can reduce the degree of artifacts from using a discrete basis. The net result is a flexible covariance model that has rank comparable or greater than the number of spatial locations and where spatial prediction, conditional simulation and evaluation of the likelihood can be done on a modest laptop computer.

Recent work on statistical methods for large spatial data sets has used a fixed rank Kriging approach to make computations feasible. This can either take the form of a small number of basis functions and an unstructured and dense $\boldsymbol{P}$ matrix such as in Cressie and Johannesson (2008) or large number of basis and a sparse model such a Markov random field for $\boldsymbol{Q}$ (Eidsvik et al. 2010). An insightful approach was suggested in Stein (2008) and later in Sang and Huang (2011) where a low rank process was combined with a process that has a compactly supported covariance. This superposition of two processes anticipates our model where we consider a mixture of covariances at multiple scales. Reflecting the fact that the likelihood calculation

carries most of the computational cost, there has been work on approximations to the likelihood for spatial models by binning the observations and using spectral methods (Fuentes 2007) or considering a partial likelihood (Michael L. Stein 2004) or pseudo likelihood (Caragea and Smith 2007). Our approach differs from these papers in that we are able to compute the likelihood exactly.

The next section describes the fixed rank Kriging model and its likelihood under a setting where the process and measurement errors have a Gaussian distribution. Section 3 outlines the computational algorithm and gives some timing results. The approximation properties of this basis/lattice model are reported in Section 4 with the proofs of the asymptotic results relegated to the Appendix. Section 5 provides an example for a climate precipitation data set and Section 6 is our conclusions. Much of the computations in this paper can be reproduced using the `LatticeKrig` package in R, which serves as a supplement for implementing the numerical methods and a ready source for the data set from Section 5.

2 The spatial model

2.1 Process and observational models

Although we have introduced g as a MR, to streamline notation in this section it is convenient to view this model as $g(\boldsymbol{x}) = \sum_{j=1}^{m} c_j \phi_j(\boldsymbol{x})$, where we have combined the MR bases into a single basis, the MR coefficients into a single coefficient vector, and m is the total number of basis functions.

Based on the set up in the introduction g will be a mean zero Gaussian process with a covariance matrix $\rho \boldsymbol{P}$ and covariance function:

$$COV(g(\boldsymbol{x}), g(\boldsymbol{x}')) = \sum_{j,k=1}^{m} \rho \boldsymbol{P}_{j,k} \phi_j(\boldsymbol{x}), \phi_k(\boldsymbol{x}'). \tag{4}$$

with $\boldsymbol{P}$ having dimension $m \times m$.

With respect to the observation model in (1) we assume that $\boldsymbol{\epsilon}_i$ are uncorrelated, normally distributed with mean zero and covariance $\sigma^2 \boldsymbol{W}^{-1}$. Here we assume that σ^2 is a free parameter of the measurement error distribution and $\boldsymbol{W}$ is a known but

sparse precision matrix. In most applications $\boldsymbol{W}$ is diagonal and we take $\boldsymbol{W}$ to be the identity for our example in Section 5. Let $\boldsymbol{\Phi}$ be the regression matrix with columns indexing the basis functions and rows indexing locations. $\boldsymbol{\Phi}_{i,j} = \phi_j(\boldsymbol{x}_i)$. With these definitions one can now rewrite (1) in matrix vector notation as $\boldsymbol{y} = \boldsymbol{Zd} + \Phi\boldsymbol{c} + \boldsymbol{e}$ and collecting the fixed and random components we have

$$\boldsymbol{y} \sim MN(\boldsymbol{Zd}, \rho\boldsymbol{\Phi P\Phi}^T + \sigma^2\boldsymbol{W}^{-1}). \tag{5}$$

As a last step it is useful to reparametrize this model to better mesh with the computations and in some instances to simplify formulas. Let $\lambda = \sigma^2/\rho$ and we reparametrize σ in terms of λ and ρ (i.e. $\sigma^2 = \lambda\rho$). Now set $\boldsymbol{M}_\lambda = (\boldsymbol{\Phi P\Phi}^T + \lambda\boldsymbol{W}^{-1})$ and (5) is the same as $\boldsymbol{y} \sim MN(\boldsymbol{Zd}, \rho\boldsymbol{M}_\lambda)$.

2.2 Spatial estimate

From (5) we have the log likelihood

$$\ell(\boldsymbol{y}|\rho, \boldsymbol{P}, \lambda, \boldsymbol{d}) = (-1/2)(\boldsymbol{y} - \boldsymbol{Zd})^T(\rho\boldsymbol{M}_\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{Zd}) - (1/2)log|\rho\boldsymbol{M}_\lambda| + (n/2)log(\pi)$$

This expression is used to find maximum likelihood estimates (MLEs) of the fixed effects and covariance parameters. For computation it is often convenient to first maximize over the fixed effects and the covariance parameter ρ analytically to reduce the number of parameters for optimization. For fixed ρ and $\boldsymbol{P}$ the MLEs for $\boldsymbol{d}$ are also the generalized least squares (GLS) estimates

$$\hat{\boldsymbol{d}} = (\boldsymbol{Z}^T\boldsymbol{M}_\lambda^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{M}_\lambda^{-1}\boldsymbol{y}. \tag{6}$$

Note this estimate only depends on λ and not on ρ. Set $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{d}}$ and substitute back in the full log likelihood giving

$$\ell(\boldsymbol{y}|\rho, \boldsymbol{P}, \sigma, \hat{\boldsymbol{d}}) = (-1/2)(\boldsymbol{r}^T(\rho\boldsymbol{M})_\lambda^{-1}\boldsymbol{r}) - (1/2)log|\rho\boldsymbol{M}_\lambda| + (n/2)log(\pi). \tag{7}$$

Finally, the expression given above can be maximized analytically over ρ giving $\hat{\rho} = \boldsymbol{r}^T\boldsymbol{M}_\lambda^{-1}\boldsymbol{r}/n$. This estimate can be substituted back into (7) to give a *profile* log likelihood that only depends on $\lambda = \sigma^2/\rho$ and on any other covariance parameters that determine $\boldsymbol{P}$.

The inference for the basis coefficients depends on the standard results for the conditional normal distribution. Specifically, the conditional distribution of $\boldsymbol{c}$ given $\boldsymbol{y}$ and all other parameters in the model at their true values is a multivariate normal

$$[\boldsymbol{c}|\boldsymbol{y}, \boldsymbol{d}, \sigma, \rho, \boldsymbol{P}] \sim MN(\hat{\boldsymbol{c}}, \rho\boldsymbol{P} - \rho\boldsymbol{P}\boldsymbol{\Phi}^T(\boldsymbol{M}_\lambda)^{-1}\boldsymbol{\Phi}\boldsymbol{P}) \tag{8}$$

with

$$\hat{\boldsymbol{c}} = \boldsymbol{P}\boldsymbol{\Phi}^T\boldsymbol{M}_\lambda^{-1}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{d}) \tag{9}$$

This conditional mean, $\hat{\boldsymbol{c}}$, is taken to be the point estimate (or prediction) of $\boldsymbol{c}$ and by linearity, the spatial prediction for $g(\boldsymbol{x})$ at an arbitrary location is $\hat{g}(\boldsymbol{x}) = \sum_{j=1}^{m} \phi_j(\boldsymbol{x})\hat{\boldsymbol{c}}_j$. Typically a vector of the spatial covariates, $z(\boldsymbol{x})$, is also provided at this location. To reproduce the familiar universal Kriging estimator, $\boldsymbol{d}$ is set at the GLS estimate given above and so the full spatial prediction is: $\hat{y}(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x})^T\hat{\boldsymbol{d}} + \hat{g}(\boldsymbol{x})$.

2.3 Radial Basis functions (RBF)

Our full model proposes a MR basis where each level of resolution takes the same form and so we start with describing a single level of basis functions on a common scale. The basis functions are essentially translations and scalings of a single radial function. Let ϕ be a unimodal, symmetric function in 1-dimension and let $\{\boldsymbol{u}_j\}$, $1 \leq j \leq m$ be a rectangular grid of points in two dimensions. Consistent with radial basis function terminology, we will refer to the grid points as *node points* and let θ be a scale parameter. The basis functions are then

$$\phi_j^* = \phi(||\boldsymbol{x} - \boldsymbol{u}_j||/\theta) \tag{10}$$

Geometrically, the basis will consist of bumps centered at the node points with overlap controlled by the choice of θ. In this work we will take ϕ to be a two-dimensional Wendland covariance (Wendland 1995) that has support on $[0, 1]$. The Wendland functions are polynomials on $[0, 1]$. They are also positive definite, which is an attractive property when the basis is used for interpolation. In this work we use a Wendland function valid up to 3 dimensions and belonging to C^4:

$$\phi(d) = (1-d)^6(35d^2 + 18d + 3)/3 \text{ for } 0 \leq d \leq 1, \text{ and zero otherwise.}$$

In all examples in this work we fix the scale factor to be 2.5 times the grid spacing. Thus in two dimensions and away from edges each RBF overlaps with 68 others.

2.4 Markov Random fields

In parallel with the preceding section we describe the stochastic model for the coefficients of a basis constructed at a single level of resolution. The MR aspect replicates this model at each level. The coefficient vector $\boldsymbol{c}$ at a single level follows a Gaussian Markov random field (GMRF) and is organized by the node points. We will assume the special case that the coefficients follow a spatial autoregression (SAR). The difference with this model for $\boldsymbol{c}$ and that in LR2011 is that we define the SAR independently from the choice of basis.

Given an autoregression matrix $\boldsymbol{B}$ and $\boldsymbol{e}$, a random vector distributed as $N(0, \rho I)$, we construct the distribution of $\boldsymbol{c}$ according to $\boldsymbol{c} = \boldsymbol{B}^{-1}\boldsymbol{e}$. The autoregressive interpretation is that $\boldsymbol{Bc} = \boldsymbol{e}$. That is, $\boldsymbol{B}$ transforms the correlated field to white noise with variance ρ. For our use we will constrain $\boldsymbol{B}$ to be sparse. Let $\mathcal{N}_j$ denote the indices of the nearest neighbors of $\boldsymbol{u}_j$. For an interior point this will be four neighbors, but less for the nodes at edges and corners. Following LR2011 for interior lattice points we take $\boldsymbol{B}_{j,j} = 4+\kappa^2$ with $\kappa \geq 0$ and the off diagonal elements to be -1. Although one can modify the weights at the edges of the lattice to approximate free boundary conditions, we have found that adding a buffer and keeping zero boundary conditions provides an easier solution. The boundary effects are also diminished by the normalization discussed in Section 2.6. By linearity $\boldsymbol{c}$ has covariance matrix $\rho\boldsymbol{B}^{-1}\boldsymbol{B}^{-T}$ and precision matrix given by $\boldsymbol{Q} = (1/\rho)\boldsymbol{B}^T\boldsymbol{B}$. Because $\boldsymbol{B}$ is formulated as unconditional weights on the field, any choice of $\boldsymbol{B}$ will lead to a valid covariance and so $\boldsymbol{Q}$ will be positive definite. It is well known that the SAR weights do not specify the Markov structure directly. For nonzero weights on the four neighbors $\boldsymbol{Q}$ will be a sparse matrix with each row having 12 nonzero elements: the first, second and third order neighbors. Thus, $\boldsymbol{c}$ will be a GMRF conditional on this larger clique of points. The results in LR2011 provide the connection between this GMRF and approximations to the Matérn family of spatial covariances. In this particular case one expects that the SAR described above

will approximate a Matérn process with scale parameter κ in LR2011 and smoothness $\nu = 1$.

2.5 Extension to a MR process

In the previous sections we have developed a basis and a covariance for a specific grid. The MR model extends this idea by successively halving the spacing of the grid points and specifying a GMRF for the coefficients at each level. Between levels we assume coefficients are independent. To make this idea explicit assume that the spatial domain is the rectangle $[a_1, a_2] \times [b_1, b_2]$ and the initial grid $\{\boldsymbol{u}_j^1\}$ is laid out with $m_x \times m_y$ grid points with the spacing $\delta \equiv (a_2 - a_1)/(m_x - 1) = (b_2 - b_1)/(m_y - 1)$. Note here the constraint that the spatial domain and numbers of grid points are matched so that the grid spacing is the same in the x and y dimensions. Subsequent grids are defined with spacings $\delta_l = \delta 2^{-(l-1)}$ and yield a sequence of grids, $\{\boldsymbol{u}_j^l\}$ that increase roughly by a factor of four in size from level l to level $l+1$. To define the basis functions for the l^{th} level we take $\theta_l = \theta/2^{(l-1)}$ and define the radial basis functions as in (10). Let L denote the total number of levels then the (unnormalized) MR basis is $\phi^*_{j,l} = \phi(||\boldsymbol{x} - \boldsymbol{u}_j^l||/\theta_l)$, where $1 \le l \le L$, $1 \le j \le m(l)$, and $m(l) = (m_x - 1)(m_y - 1)4^{l-1} + m_x + m_y + 1$. The total number of basis functions is approximately $(m_x m_y)(4^L)$, (This is not exact because m grid points are subdivided into $2m - 1$ points at the next level.) When buffer nodes are added to reduce edge effects we take these as a *fixed* number of extra points that are added to each edge of the grid. The number of basis functions follows a more complicated expression when buffer nodes are added at each level but is still grows at roughly 4^L.

Recall that the vector of coefficients associated with each level is $\boldsymbol{c}^l$ and the MR representation for g is given by equations (2) and (3) with either the unnormalized MR basis $\{\phi^*_{j,l}\}$ or the normalized basis described in Section 2.6 below. It should be noted that the MR basis by itself does not contribute too much additional computation burden. The main difference in a single level of basis functions and a MR are additional nonzero elements in the inner matrix, $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$, due to coarse resolution basis functions overlapping with finer resolution ones. Although the MR will have more nonzero

elements in the inner product matrix, there are many fewer coarse functions for overlap and so the total number of nonzero elements does not increase substantially. This feature can be seen in the timing results in Section 4.

It is useful to illustrate how the number of basis functions depend on the number of levels. Suppose that an initial grid of 10×10 is chosen for a square spatial domain, $L = 4$ and 5 extra, buffer node points are added on each side to moderate the edge effects. The first level, will comprise $(10 + 10) \times (10 + 10) = 400$ grid points including a buffer region on all four sides of the spatial domain. The second level will decrease the grid spacing by a factor of two giving 19×19 grid points included in the spatial domain and being aligned with the coarser grid. To these are appended 5 buffer points on each edge giving a total of $29 \times 29 = 841$ points. Subsequent levels yield $(37 + 10) \times (37 + 10) = 2209$ and $(73 + 10) \times (73 + 10) = 6889$ grid points. The four levels sum to 10399 grid points/basis functions and of these 7159 have nodes that are included in the spatial domain.

In general we can stack these coefficients as $\boldsymbol{c} = (\boldsymbol{c}^1, \boldsymbol{c}^2, ..., \boldsymbol{c}^L)$ and the natural extension of the SAR model is a sparse matrix $\boldsymbol{B}$ such that $\boldsymbol{Bc}$ is $N(0, \rho I)$. Although $\boldsymbol{B}$ can be a general matrix we have found it useful to restrict attention to a block diagonal form. Let $\alpha_1, \alpha_2, \ldots, \alpha_L$ be a vector of positive weights and for the l^{th} level we assume $\boldsymbol{c}^l$ follow a GMRF with a SAR matrix, $(1/\sqrt{\alpha_l})\boldsymbol{B}_l$. Here $\boldsymbol{B}_l$ has the same form as in the single level but with the κ parameter possibly depending on the level. One can interpret $\rho\alpha_l$ as parameterizing the marginal variance of the l^{th} level process and κ_l is an approximate scale parameter. Thus we are lead to a block diagonal form for $\boldsymbol{B}$ and also for the precision matrix:

$$\boldsymbol{Q} = (1/\rho) \begin{bmatrix} (1/\alpha_1)(\boldsymbol{B}_1)^T\boldsymbol{B}_1 & 0 & \ldots & 0 \\ 0 & (1/\alpha_2)(\boldsymbol{B}_2)^T\boldsymbol{B}_2 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & (1/\alpha_L)(\boldsymbol{B}_L)^T\boldsymbol{B}_L \end{bmatrix} \tag{11}$$

$\boldsymbol{Q}$ will have dimension $m \times m$ equal to the total number of basis functions but of course will be sparse and $\boldsymbol{c}$ will have length m.

2.6 Normalization to approximate stationarity

Based on the specific form for $\boldsymbol{Q}$ we have found it useful to normalize the basis functions to give a better approximation to stationary covariance functions. It is well known that a GMRF on a finite lattice can exhibit edge effects and other artifacts in the covariance model that are not physical. Moreover the radial basis functions having nodes on a discrete set can also contribute to patterns in the implied covariance matrix. One obvious correction for this effect is to weight the basis functions so that when (4) is evaluated one will obtain a constant marginal variance. Accordingly, let $\omega(\boldsymbol{x}) = \sqrt{COV(g(\boldsymbol{x}), g(\boldsymbol{x}))}$ from (4) and normalize the basis functions as $\phi_j(\boldsymbol{x}) = \phi_j^*(\boldsymbol{x})/\omega(\boldsymbol{x})$. Because this normalization is tied to the choice of covariance model it means that the basis is no longer independent of the GMRF and this linkage adds more computational overhead. However, computing $\omega(\boldsymbol{x})$ can take advantage of the sparse precision matrix and we believe reducing edge effects and other artifacts is worth the extra computation.

3 Computational strategy and timing results

The estimators defined in the previous section can be found efficiently by a judicious use of sparse matrix decompositions and matrix identities. Most of these computations depend on the constructions of $\boldsymbol{\Phi}$, $\boldsymbol{W}$ and $\boldsymbol{Q}$ to be sparse matrices. Our basic approach exploits the fact that a sparse and positive definite matrix can be factored into a sparse cholesky decomposition. With this decomposition it is efficient to evaluate inverses and determinants. In this section we outline the key numerical steps and the reader should refer to Nychka et al. (2013) and the commented `LatticeKrig` package source code for details.

3.1 Spatial prediction and evaluating the likelihood

A basic calculation that illustrates the computational strategy is to evaluate $\boldsymbol{M}_\lambda^{-1}\boldsymbol{w}$ for an arbitrary vector $\boldsymbol{w}$. Recall that $\boldsymbol{M}_\lambda = \boldsymbol{\Phi}\boldsymbol{P}\boldsymbol{\Phi}^T + \lambda\boldsymbol{W}^{-1}$ and taken at face value $\boldsymbol{M}_\lambda$ is a dense, potentially large matrix and so difficult to work with directly. The strategy is to transform $\boldsymbol{M}_\lambda$ using matrix identities to involve the sparse precision matrix. The

Sherman-Morrison-Woodbury formula (Henderson and Searle (1981)) can be applied to give

$$\boldsymbol{M}_\lambda^{-1} = \left(\boldsymbol{\Phi P \Phi}^T + \lambda \boldsymbol{W}^{-1}\right)^{-1} = (\boldsymbol{W} - (\boldsymbol{W\Phi})\boldsymbol{G}^{-1}(\boldsymbol{\Phi}^T\boldsymbol{W}))$$

where $\boldsymbol{G} = \boldsymbol{\Phi}^T\boldsymbol{W\Phi} + \lambda\boldsymbol{Q}$. Because $\boldsymbol{\Phi}$, $\boldsymbol{W}$ and $\boldsymbol{Q}$ are all sparse, $\boldsymbol{G}$ will also be sparse and positive definite. Using this identity one can now use the sparse Cholesky decomposition for $\boldsymbol{G}$ to solve the linear system $\boldsymbol{Gv} = (\boldsymbol{\Phi}^T\boldsymbol{W})\boldsymbol{w}$ for $\boldsymbol{v}$ and it follows that

$$\boldsymbol{M}_\lambda^{-1}\boldsymbol{w} = \boldsymbol{W}\boldsymbol{w} - \boldsymbol{W\Phi}\boldsymbol{v}$$

Note that an important limitation of this computational strategy is that λ can not be identically zero. To compute $\hat{\boldsymbol{c}}$ we use the identity $\hat{\boldsymbol{c}} = \boldsymbol{G}^{-1}\boldsymbol{\Phi}^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{d}})$ and exploit the sparsity of $\boldsymbol{\Phi}$ and $\boldsymbol{W}$ for multiplication and the sparse Cholesky factorization of $\boldsymbol{G}$. Finally note that the evaluation of $\hat{g}(\boldsymbol{x})$ can also be computed in an efficient manner if the sum is restricted to basis functions that are nonzero at $\boldsymbol{x}$.

The other intensive computation occurs in the likelihood as the determinant of $\boldsymbol{M}_\lambda$. Here we use a special case of Sylvester's Theorem: For an $n \times m$ matrix $\boldsymbol{U}$ and identity matrices $\boldsymbol{I}_n$ and $\boldsymbol{I}_m$, $|\boldsymbol{UU}^T + \boldsymbol{I}_n| = |\boldsymbol{U}^T\boldsymbol{U} + \boldsymbol{I}_m|$. Using elementary properties of matrices one can derive the identity $|\boldsymbol{M}_\lambda| = \lambda^{n-m}|\boldsymbol{G}|/(|\boldsymbol{Q}||\boldsymbol{W}|)$. The matrices, $\boldsymbol{W}$, $\boldsymbol{G}$ and $\boldsymbol{Q}$ are all positive definite and sparse so the determinants can be found efficiently from the product of the diagonal elements of the Cholesky decompositions.

Based on exploiting matrix sparsity and these classic matrix identities one can evaluate the likelihood in an efficient manner. With this option we just use standard maximum likelihood methods of inference on the covariance parameters.

In this work we suggest finding the prediction errors using the well known Monte Carlo technique of conditional simulation. Under the assumption that the covariance model is known, one generates a sample from the conditional distribution of g and $\boldsymbol{d}$ given the observations. The prediction variance can be approximated from Monte Carlo draws from this conditional distribution. This computation can be done in two steps: simulating an unconditional random process at the prediction and observation locations and then determining the prediction errors based on synthetic/simulated observations for this realization. The first step is an standard application of multivariate simulation

by solving a linear system based on the Cholesky decomposition of the precision matrix and the second step is the same spatial estimator that is applied to actual data.

Here we present some timing results for the computations with the main comparison being with the dense matrix computations associated with Kriging. The spatial locations were uniformly distributed over the domain $[0,1] \times [0,1]$ and the number were varied between 500 and 20000. The likelihood function and spatial predictions were found for an exponential covariance model and several choices of the lattice MR model. For these algorithms the computation time is dominated by basic linear algebra and does not depend on the values of the spatial data, the distribution of spatial locations, and the specific values of the covariance parameters. The timing is done for the function `mKrig` in the R package `fields` (Furrer et al. 2012) implementing standard Kriging and for the function `LKrig` in the R package `LatticeKrig` (Nychka et al. 2012) implementing the MR basis function model. Times reported are on a *single* processor for a Macbook Pro laptop (2.3 Ghz Intel Core i7, 8Gb memory) and R 3.0.1 (R Development Core Team 2011). Both of these functions compute the predictions at the observations for a fixed covariance model, evaluate the likelihood and compute the coefficients for predicting the surface at arbitrary points. Despite this varied output from the functions, the Cholesky decomposition in both `mKrig` and `LKrig` dominate the time for large n.

Figure 1 reports the total time ("wall clock" time) for these functions using the R utility `system.time`. The dashed line is the time for the standard "Kriging" estimate using `mKrig` up to 10,000 observations and with times extrapolated to 20,000. Thus the time for 20,000 observations and standard Kriging is estimated to be about 1,300 seconds (about 21 minutes). The solid black line is the time for the function `LKrig` with a single level with the number of basis functions chosen to be approximately equal to the sample size, and with the basis functions normalized to have unit marginal variance. The dotted black line is the same scheme but without normalizing the basis functions. Note that for 20000 spatial locations the times for this case are 66 seconds (normalized) and 5.4 seconds (unnormalized). The grey lines report timing with the number of basis functions keep fixed and with (solid) and without (dashed) normalization. The lines

labeled 10 have four levels ($L = 4$) of MR and where the coarsest basis has centers on a 10×10 grid ($m_x = m_y = 10$) and giving 7159 basis functions with nodes within the spatial domain and 10339 total. The lines labeled 20 have coarsest grid being a 20×20 grid ($m_x = m_y = 20$) with a total of 31,259 basis functions within the spatial domain and 37,439 total. The memory for this case is dominated by storage of the sparse matrix $\boldsymbol{G}$ comprising 7.4×10^6 nonzero elements taking 60Mb of memory.

These results indicate substantial time savings over the dense matrix computations and evaluations of the likelihood are feasible even for 20000 spatial locations. The unnormalized computation times are particularly striking and are largely dominated by the sparse Cholesky decomposition of the matrix $\boldsymbol{G}$ discussed in Section 3. For this work we have not exploited more efficient algorithms in the normalization step and there is a significant difference between the normalized and unnormalized cases. As might be expected the two covariance models with fixed number of basis functions ("10" and "20" cases) are closer to being linear as a function sample size. At the sample size of approximately 10400 the 10×10, $L = 4$ case and the single level model 103×103, $L = 1$ case have equal numbers of basis functions. However, because of the difference in levels the four level model has a $\boldsymbol{G}$ with 1.88×10^6 nonzero elements compared to $.67 \times 10^6$ for the one level model. This difference in sparsity explains the timing differences for the unnormalized computations. The normalized computations are apparently dominated by the normalization computation.

4 Properties of the covariance model

4.1 Comparison to a convolution process

As a foundation, we first consider a convolution approximation to the sum of radial basis functions. First we define a single convolution process and then extend this to an infinite mixture. Let z be a unit variance, isotropic, two dimensional Matérn process with spatial scale parameter κ, smoothness ν, and $C_\nu(||\boldsymbol{x} - \boldsymbol{x}'||/\kappa) = E(z(\boldsymbol{x})z(\boldsymbol{x}'))$, the corresponding covariance function. Also let ϕ be a compactly supported RBF with

$\phi(0) = 1$. For $\theta > 0$ a scale parameter, define the convolution process

$$g(\boldsymbol{x}) = \int_{\boldsymbol{R}^2} \frac{1}{\theta^2} \phi(||\boldsymbol{x} - \boldsymbol{u}||/\theta) z(\boldsymbol{u}) d\boldsymbol{u}.$$

This type of process for statistical modeling is well-established (see Higdon 1998) and as written will be Gaussian, mean zero, and have an isotropic covariance function. Now consider a sequence of independent Matérn processes, $z_l(\boldsymbol{x})$ with $\{\theta_l\}$ a sequence of scale parameters for the convolution kernel and "hard wire" $\kappa_l = 1/\theta_l$. These define a sequence of convolution processes $g_l(\boldsymbol{x})$ according to (12) with the same marginal variance. Finally, let k_l denote the covariance function for the l^{th} process. Given, non-negative weights $\{\alpha_k\}$ that are summable we are lead to the MR process that is Gaussian, mean zero and covariance given by

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l=1}^{\infty} \alpha_l k_l(\boldsymbol{x}, \boldsymbol{x}').$$

Given this representation, a theoretical question is how the choice of $\{\theta_l\}$ and $\{\alpha_l\}$ influence the properties of k. In particular, is it possible to construct covariances that represent different degrees of smoothness than those implied by the basis functions and Matérn process used in the convolution? Typically the smoothness of an isotropic, stationary Gaussian process is tied to the differentiability of the covariance function at the origin. An alternative measure is to characterize the tail behavior of the spectral density of the process. Under isotropy the spectral density will be radially symmetric and we focus on the decay rate as r increases. In particular, for spectral densities whose tails are bounded by a fixed polynomial decay we will take the polynomial order as a convenient measure of the process smoothness. For the Matérn family a smoothness of ν and dimension 2 the spectral density will have a tail behavior following $r^{-(2\nu+2)}$ as $r \to \infty$. For example the exponential covariance (ν= 1/2) will have a spectral density that decreases at the polynomial rate r^{-3}. A covariance spectrum with tail behavior of the same order might be expected to provide a process model with similar smoothness to the exponential at small spatial scales. The following theorem reports the tail behavior for the MR process for different choices of the scale and weight sequences. An interesting result is that the MR process can reproduce a scale of different decay

rates for the tail of the spectral density and can recover the -3 rate of decay for the exponential covariance.

Theorem 4.1 Assume (1) ϕ is a two-dimensional Wendland covariance function of order K. (2) the smoothness of the Matérn processes is fixed at $\nu = 1$. (3) $\alpha_l = e^{-2\beta_1 l}$ and $\theta_l = e^{-\beta_2 l}$ with $\beta_1, \beta_2 > 0$ and $(\beta_1/\beta_2 + 1) < (5 + 2K)$. If $S(r)$ denotes the spectral density of g (or k) with respect to the radial coordinate then there are constants independent of r, $0 < A_1, A_2 < \infty$ such that

$$A_1 < S(r) r^{2\mu+2} < A_2, \text{ with } \mu = \beta_1/\beta_2.$$

Corollary 4.1 Under assumptions (1) and (2) and $\theta_l = 2^{-l}$, $\alpha_l = \theta_l^{2\nu}$ and $(\nu + 1) < (5 + 2K)$, $S(r)$ will have tail behavior with the same polynomial order as a two-dimensional, Matérn process spectrum with smoothness ν.

The proof of this theorem is given in the Appendix.

4.2 Numerical approximation

The theoretical approximation is based on a continuous convolution of the basis functions with the Matérn covariance. We have found that the theoretical sequence of weights gives an accurate approximation when 6 or more levels are considered. However this theoretical comparison does not exactly match the discrete stochastic model used for data analysis. A more practical comparison is how well the discrete MR basis proposed here can match members of the Matérn family. We investigate the quality of the approximation given $\theta_l = 2^{-l}$ but optimizing over $\{\kappa_l\}$ and and $\{\alpha_l\}$. Note that this scheme is slightly different than the theoretical setup because κ_l is allowed to vary independently from θ_l and α_l is not constrained to be a power of θ_l. The most important constraint in choosing an approximation is the initial choice of grid size (m_x and m_y) and the number of levels, L. The spacing of the nodes should be chosen so the coarsest level is comparable to the process correlation range and L such that the finest basis functions have smaller scale than the finest spatial scale of the process. One advantage of this model is that flexibility in choosing the range parameter κ means that the grid spacings need not exactly conform to the correlation scale of the process.

The first column in Figure 2 shows the approximation for an exponential covariance with range parameters .1, .5 and 1.0 using 3 and 4 levels of MR basis functions. The MR parameters κ_l and α_l have been found by minimizing the mean squared error between the approximation and the target covariance function on a grid of 200 distances in the interval $[0, 1]$. The coarsest basis function centers are organized on a 10×10 grid on the square $[-1, 1] \times [-1, 1]$ and so with four levels the approximation has $10^2 + 19^2 + 37^2 + 73^2 = 7159$ two-dimensional basis functions with nodes that are included in the spatial domain. There are 10339 basis functions total considering the buffer regions. The plots in the upper row are the target and approximate covariances as a function of distance from the point $(0, 0)$ along the x-axis. The approximation is close to being stationary and isotropic and so this comparison is representative for distances along other orientations. In the plots the solid curve is the covariance, the dotted line is the approximation with 3 levels, and the dashed line is the approximation at 4 levels.

Not surprisingly the approximation breaks down at small distances that are below the resolution of the finest basis functions This feature is highlighted by the plots in the lower row where the approximation is given for points in a range close to zero. The characters "3" and "4" indicate the smallest scale of the basis functions and thus indicate the limits of the MR for the 3 and 4 level choices. In general it is straightforward to improve this approximation by increasing L beyond 4. A similar approximation is made for the Whittle covariance ($\mu = 1$) except for the largest range parameter the coarsest basis has centers on a 5×5 grid (giving a total of 1484 basis functions). This case is an example where the smoothness of the covariance at zero does require as detailed basis functions and in fact we found empirically that the courser initial grid (5×5) gives a better approximation. Note that in the error plot there is also a small artifact, a rippling feature that is from the discrete spacing of the basis functions. The third column of Figure 2 is an example of the ability of the MR to approximate more general correlation functions. This is perhaps the most strikingly example of the flexibility of this model. Here the target is a mixture of exponentials: $.4 \exp(-d/.1) + .6 \exp(-d/3)$. For reference the individual exponent

correlation functions are plotted as grey solid lines. The approximation is also accurate with the error localized near the origin and being large below the smallest scale of the MR.

5 North American summer precipitation

The MR lattice model was applied to a substantive climate data set in order to test its practical value and compare it to standard Kriging. The goal is to estimate the average summer rainfall on a fine grid for North America based on high quality surface observations (NOAA/NCDC 2011). These types of fields are an important reference in studying the Earth's climate system. GHCN data is quality controlled, curated and served by the US National Climatic Data Center and for this example we use 1720 stations from North America. For each station, a least squares trend line was fit to the summer precipitation totals (June, July, August) for the period 1950-2010 and the trend line was evaluated at the midpoint time (1980.5). Note that with complete observations this is just the sample mean and we will refer to these statistics as the station "mean summer precipitation". However, 75% of the adjusted stations are missing at least 10 values in this period and the least squares analysis will differ from a sample average.

The version of the climate data used is the R data set `NorthAmericanRainfall` in the `LatticeKrig` package and a spatial model was fit using stereographic map coordinates for the station locations. This projection gave spatial coordinates whose euclidean distances were similar to great circle distance (see Figure 3). The spatial model was fit to the log of mean precipitation with the spatial coordinates and elevation included as a linear fixed effects. Three correlation models were considered and we report the MLEs for the relevant parameters and the effective degrees of freedom (EDF).

Matern (2 parameters) A stationary, isotropic Matern with range and smoothness parameters.

$\hat{\sigma} = .1084$, EDF= 943.

Matern-like (2 parameters) A three-level, MR covariance with coarsest level having a lattice of 16×13 included the rectangular spatial domain amounting to approximately 4000 basis functions. A common value for κ was used to control the range at all levels. The first MR model constrains $\{\alpha_1, ..., \alpha_3\}$, $\alpha_k \sim 2^{-2\nu}$ with the additional constraint that $\sum \alpha_k = 1$.

$\hat{\sigma} = .1402$, $\hat{\nu} = .49$, $\hat{\kappa} = .96$, EDF= 489.4.

Multi-resolution (3 parameters) The same three-level structure as the Matern-like model with κ a common parameter with $\alpha_k > 0$ and $\sum \alpha_k = 1$.

$\hat{\sigma} = .1353$, $\hat{\boldsymbol{\alpha}} = (0.91, 0.00, 0.09)$, $\hat{\kappa} = .7071$, EDF= 550.6.

All three covariance functions include the variance parameter, ρ being the marginal variance of the spatial process and the parameter, σ^2, that is the measurement error (or nugget) variance.

The initial grid size for the MR models and the number of levels was identified by trying several sizes and comparing likelihood values when models were nested. We also avoided configurations where κ was large suggesting an uncorrelated model for the GMRF. The covariance parameters were estimated by maximum likelihood and confidence regions for the parameters were derived using the large sample chi-squared approximation to -2 times the log likelihood. Based on a 95% confidence set the range parameter for the Matern model was not constrained from above and so a thin-plate spline model, i.e. a limiting process as the range becomes large, is not ruled out. The smoothness parameter however has an MLE of .64. Figure 4 compares the correlation functions for these three different models based on the confidence sets for the parameters. Here the usual 95% confidence set for the model parameters based on the likelihood was translated into a confidence band for the corresponding correlation functions. The MR models have the flexibility to have long range correlations and it is interesting for these data that their shape is different than the Matern family. Also it is striking that the three level MR has $\alpha_2 \approx 0$ suggesting omitting the middle resolution level. The spatial predictions given by all three models are similar, however, and within the prediction uncertainty measures. The measurement error variance is smaller for the Matern compared to the lattice models and this is consistent with the

Matern representing a slightly rougher process than the MR models. In this case the Matern process captures more of the fine scale variability and so less is represented by the measurement error/nugget term.

Figure 5 is an example of the expected precipitation surface for a subregion over the Rocky Mountains centered on Colorado. The MR covariance with the MLE parameters reported above is used for these estimates, which are evaluated on 200×200 grid. 200 conditional fields were simulated and to increase the accuracy of this sample the realizations were centered so that their mean matched their conditional expected mean, which can be computed exactly. Although the spatial model was estimated on a log scale of precipitation, the conditional samples were transformed to the raw scale of precipitation totals to represent the distribution for unlogged values. Specifically the surface in (a) is mean of the exponentiated conditional fields. Here the elevation covariate explains a large amount of the spatial structure but this component is modified by the smooth nonparametric component based on the location. Plot (b) is the estimated prediction standard error as a percentage of the mean predicted field.

6 Discussions and Conclusions

This work has developed a new model for a spatial process: a lattice/basis model that builds on ideas from fixed rank Kriging and the computational efficiencies that are inherited from Markov random fields. The key contribution is that an independent sum of the processes at different scales can approximate a larger family of processes not limited to the properties of the covariance at each resolution level. One advantage of our model is numerical evidence that it can accurately reproduce the Matérn family of covariances. Also we give some asymptotic results based on a theoretical convolution model that indicate that a range of smoothness properties can be achieved. This result is unexpected given that the lattice/basis process has a fixed smoothness controlled by the choice of basis functions.

Besides the value of the lattice/basis formulation as a new covariance model there is an equally important contribution in computational efficiency for large data sets. In fact it is our perspective that more complex covariance models can only be exploited

when large number of observation locations allow for accurate estimation of covariance parameters. Thus efficient computation is intrinsic to entertaining new spatial models. We have been successful in identifying algorithms that allow for computing the likelihood to estimate covariance parameters and the prediction of the spatial field using large data sets.

Because of the description of the stochastic spatial elements in terms of a SAR, it is straightforward to propose a non stationary extension to the lattice basis model. One would allow both the κ_l and α_l to vary over the lattice at each level. An additional refinement would allow the SAR weights between the neighboring lattice points to be directionally dependent. In particular extending the SAR weights to the 8 first and second order neighbors can allow for a model that has directional or anisotropic dependence. The spatial variation in these parameters could be modeled by a set of covariates and fixed effects or one could include a spatial process prior on these parameter fields. The advantage of our approach and also of the related SPDE and process convolution models is that one will always obtain a valid covariance function because the model focuses on a process level description.

We conjecture that the choice of the Wendland family of RBFs is not crucial and other compacted supported, positive definite functions will work. Moreover by modifying the distance metric to one of chordal distance one can also extend these ideas to the sphere. The one hurdle in an extension to a spherical process, however, is to devise non-rectangular grids for the nodes and to formulate a SAR on these points.

Finally, we note that the lattice/basis model can be implemented using a collection of simple numerical algorithms and readily available software. An R implementation is available with documented and commented source code and uses the general sparse matrix R package `spam`. The `LatticeKrig` source code is largely written in the R language with limited use of lower level C or FORTRAN functions and hence is easy to modify.

References

Banerjee, Sudipto, Alan E Gelfand, and Bradley P Carlin (2003), *Hierarchical modeling and analysis for spatial data*. Crc Press.

Caragea, P. C. and R. L. Smith (2007), "Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models." *Journal of Multivariate Analysis*, 98, 1417–1440.

Cressie, Noel and Christopher K Wikle (2011), *Statistics for spatio-temporal data*. Wiley. com.

Cressie, Noel A. C. and Gardar Johannesson (2008), "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 209–226.

Eidsvik, Jo, Andrew O Finley, Sudipto Banerjee, and Havard Rue (2010), "Approximate bayesian inference for large spatial datasets using predictive process models." *Computational Statistics & Data Analysis*, 56, 1362–1380.

Fuentes, Montserrat (2007), "Approximate likelihood for large irregularly spaced spatial data." *Journal of the American Statistical Association*, 102, 321.

Furrer, Reinhard, Douglas Nychka, and Stephen Sain (2012), *fields: Tools for spatial data*. URL `http://www.image.ucar.edu/Software/Fields`. R package version 6.6.4.

Henderson, H.V. and S. R. Searle (1981), "On deriving the inverse of a sum of matrices." *SIAM Review*, 23, 53–60.

Higdon, David M. (1998), "A process-convolution approach to modelling temperatures in the north atlantic ocean." *Environmental and Ecological Statistics*, 5, 173–190.

Katzfuss, Matthias and Noel Cressie (2011), "Spatio-temporal smoothing and em estimation for massive remote-sensing data sets." *Journal of Time Series Analysis*, 32, 430–446.

Lindgren, Finn and Håvard Rue (2007), "Explicit construction of gmrf approximations to generalized matern fields on irregular grids." Technical report, Lund Institute of Technology.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011), "An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423–498.

Michael L. Stein, Leah J. Welty, Zhiyi Chi (2004), "Approximating likelihoods for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 275296.

NOAA/NCDC (2011). URL `http://www.ncdc.noaa.gov/ghcnm`.

Nychka, Douglas, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain (2013), *A Multi-resolution Gaussian process model for the analysis of large spatial data sets*. URL `http://www.ucar.edu/library/collections/technotes`. NCAR Tech Note.

Nychka, Douglas, Dorit Hammerling, Stephen Sain, and Tia Lerud (2012), *LatticeKrig: Multiresolution Kriging based on Markov random fields*. URL `http://www.image.ucar.edu/Software/MRKriging`. R package version 2.3.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Rue, Håvard and Leonhard Held (2005), *Gaussian Markov random fields : theory and applications*, volume 104. Chapman & Hall/CRC, Boca Raton.

Sang, H. and J.Z. Huang (2011), "A full scale approximation of covariance functions for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Stein, Michael L. (2008), "A modeling approach for large spatial datasets." *Journal of the Korean Statistical Society*, 37, 3.

Wendland, H. (1998), "Error estimates for interpolation by compactly supported radial basis functions of minimal degree." *Journal of Approximation Theory*, 93, 258–272.

Wendland, Holger (1995), "Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree." *AICM*, 4, 389–396.

Acknowledgements

This work supported in part by National Science Foundation grant DMS-0707069 and the National Center for Atmospheric Research. S. Bandyopadhyay is partially supported by the Reidler Foundation of Lehigh University.

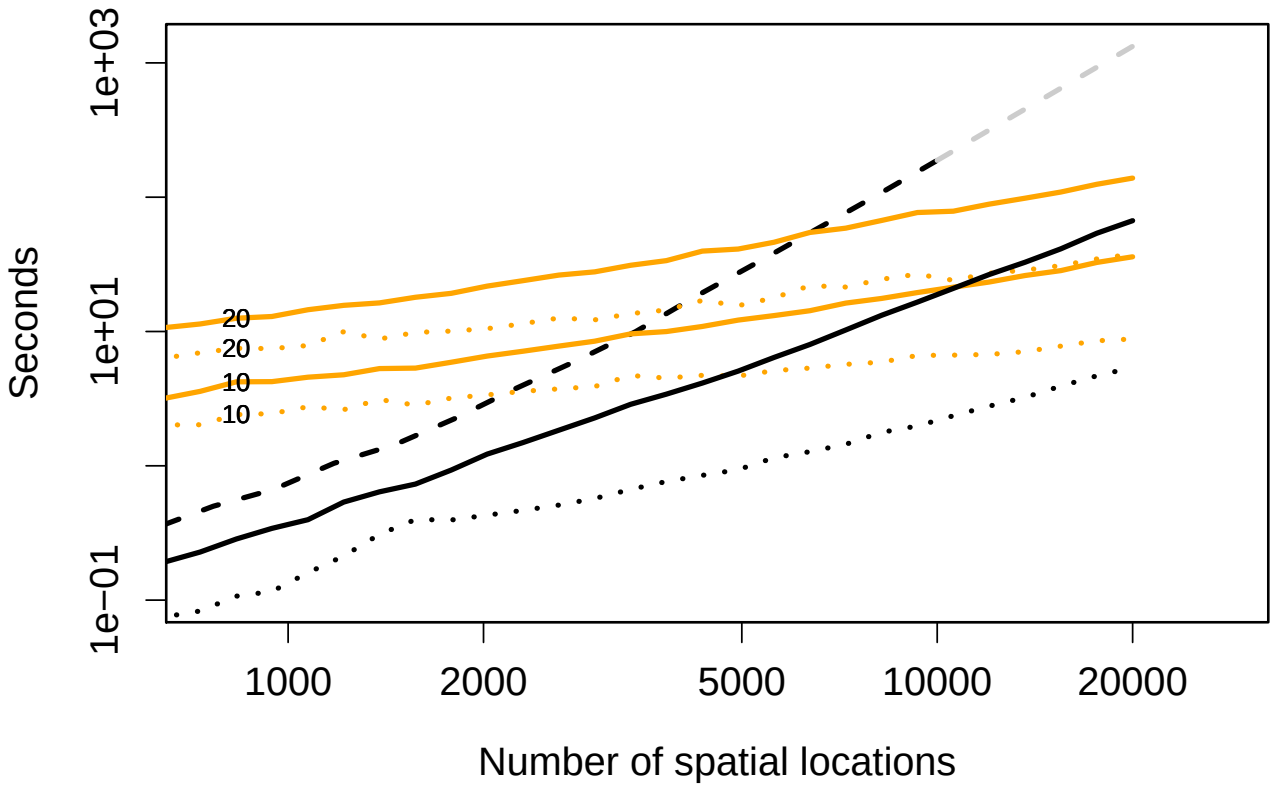

Figure 1: Timing results for the lattice/basis model and standard Kriging in seconds for several different numbers of basis functions and for the standard evaluation of the likelihood based on a dense covariance matrix. The dashed line is the time for the `mKrig` function from the `fields` R package that computes the likelihood and related statistics for an exponential covariance model with a fixed set of covariance parameters using a standard dense matrix Cholesky decomposition. Solid and dotted lines are times for the `LKrig` function from the `LatticeKrig` R package that compute the likelihood and related statistics for a MR lattice covariance with fixed parameters. Solid lines are times with normalization to a constant marginal variance and dotted lines are times without normalization. Among these cases the black lines are for a single level model where the basis functions are chosen to be roughly equal to the number of spatial locations. The orange lines use a fixed number of basis functions comprising four levels and with the coarsest level being either 10×10 or 20×20. Text labels identify these cases.

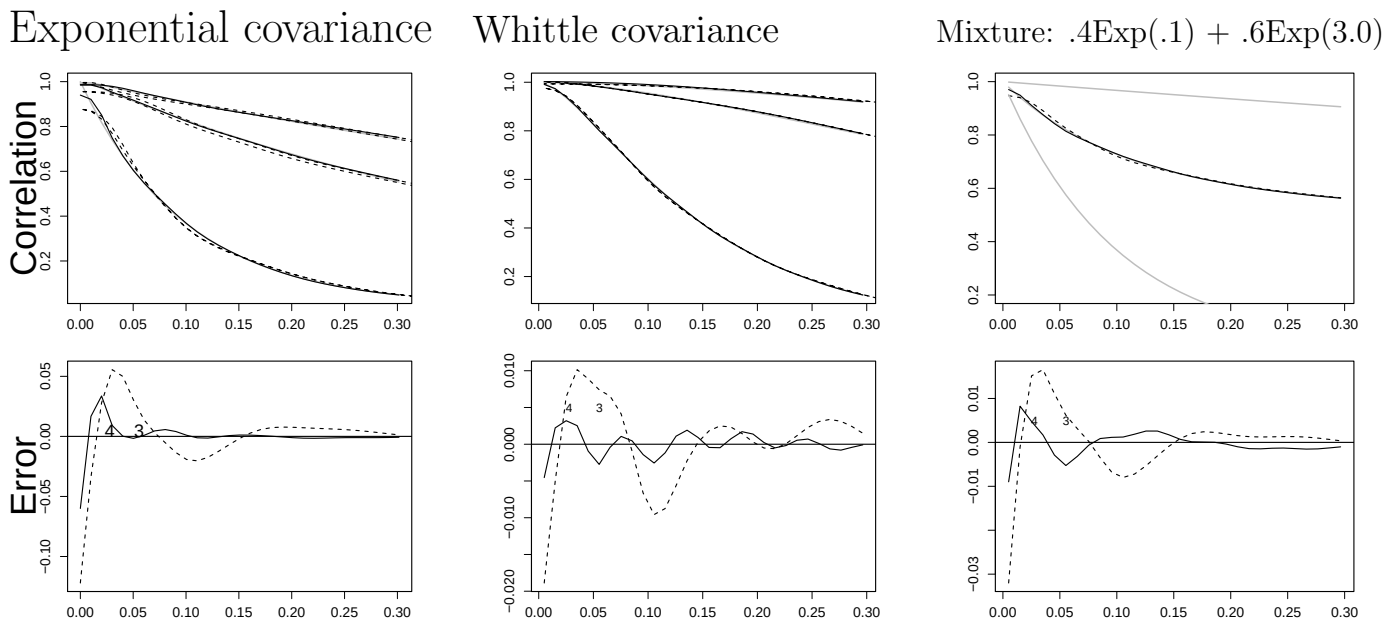

Figure 2: Approximation of Matérn covariances using the lattice/basis model. For the plots on thetop row the solid grey lines are the true correlation functions. First column is an exponential correlation with range parameter (.1, .5 and 1.0), second column is the Whittle correlation with ranges .1,.5 and 1.0 and the third column is a mixture of two exponential correlation functions. Black lines are the approximations to these correlation functions. Approximations are indicated in black with $L = 3$ (dashed) or $L = 4$ (solid). The upper row is the approximations with the true correlations over the distance limits $[0, .3]$. The lower row are the differences between the approximation and the true correlation function for the cases when the range is .1 or for the mixture model. The characters `3` and `4` indicate the support for the basis functions at the third and fourth levels of resolution.

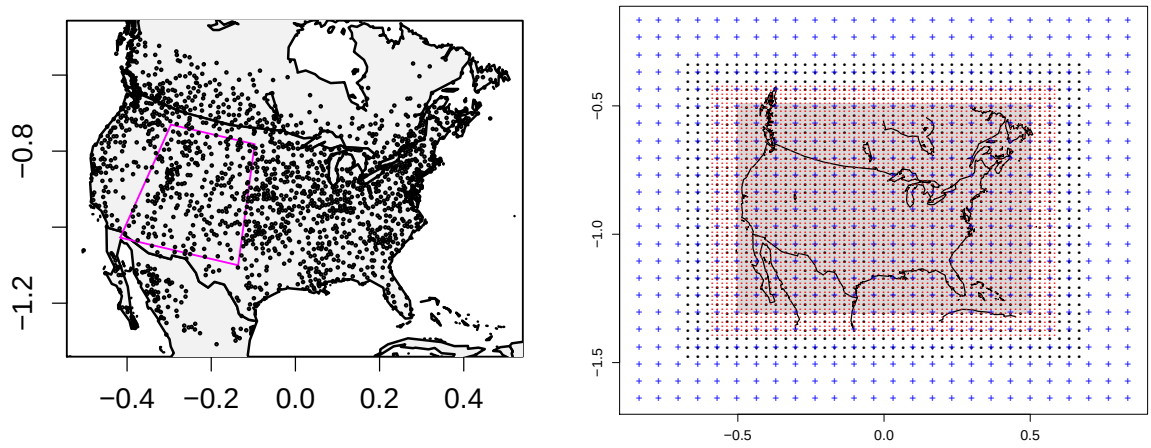

Figure 3: Illustration of the spatial domain and basis grid for the precipitation example. Left plot is a stereographic projection of precipitation location observation locations indicating the the the subregion in figure 5. The right plot shows the three different grids ("+" – coarse, large dot – middle and small dot – fine) defining the nodes for the MR basis including the buffer regions of 5 extra nodes on each side to minimize edges effects. Shading indicates the rectangular spatial domain.

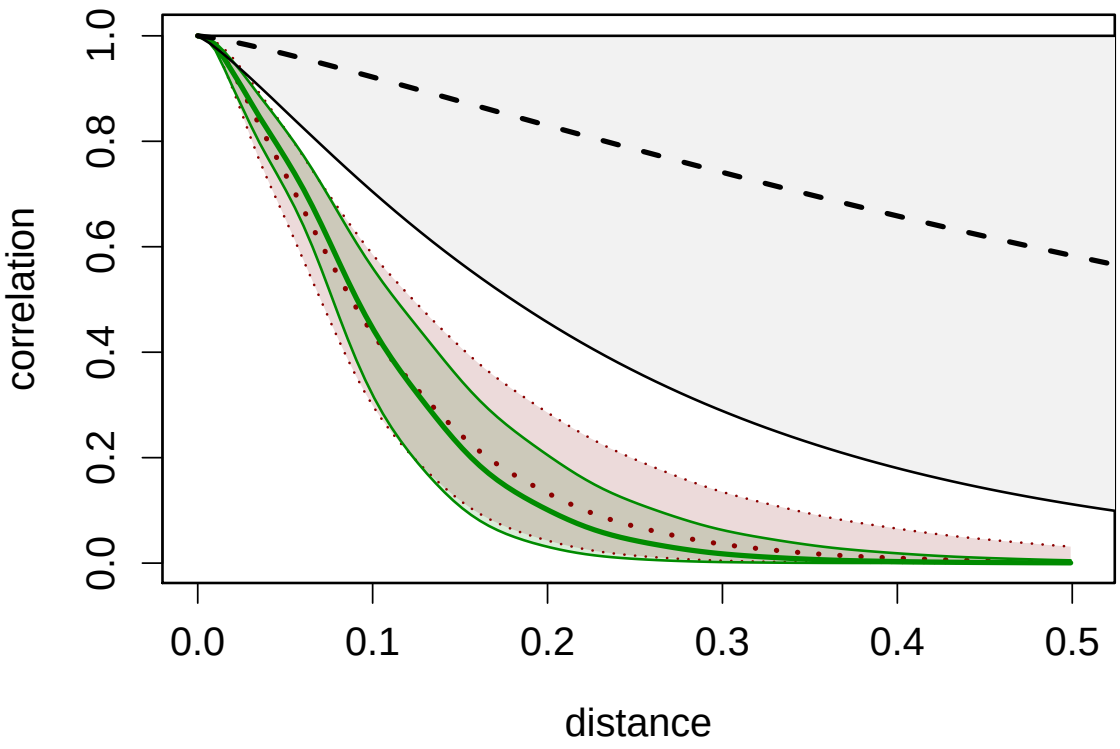

Figure 4: Correlation models fit to the precipitation data. Dashed line is the Matérn correlation function found by maximum likelihood and the light grey shading is an approximate 95% uncertainty region based on a confidence set for the range and smoothness parameters. Dotted line is the estimated correlation and uncertainty (dotted envelope) for the Matern-like covariance model. Solid line with darker shading is a similar summary for the three level MR model.

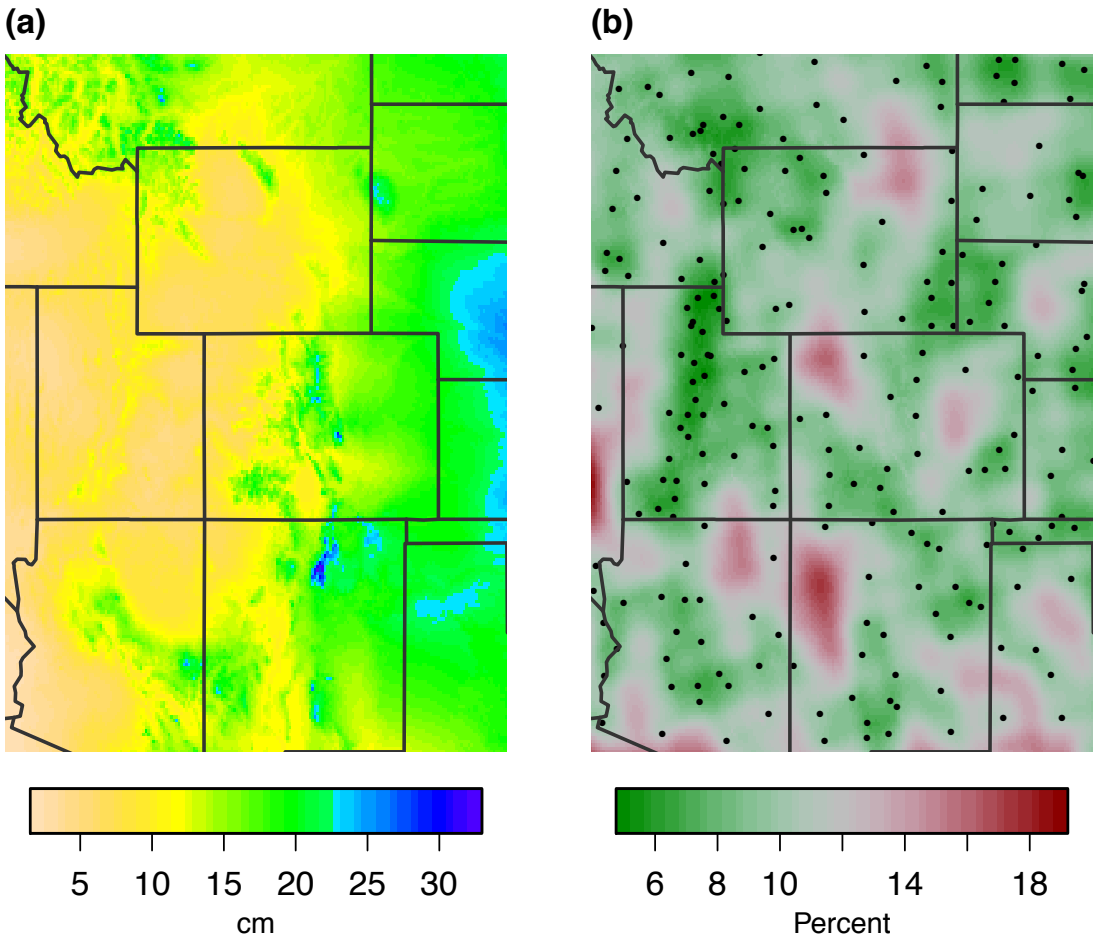

Figure 5: Plot (a) reports the spatial predictions for mean summer (June July and August) precipitation in centimeters and includes elevation as a fixed linear covariate over the Rocky Mountain region of the US. This subregion is outlined in Figure 3. The spatial covariance function is the three level MR model described in Figure 4. Plot (b) reports approximate prediction standard errors for this surface as a percentage of the predicted mean field. Solid points show observation stations.

Appendix

Note that the convolution process has a covariance function given by

$$\int_{\boldsymbol{R}^2}\int_{\boldsymbol{R}^2}\frac{1}{\theta^4}\phi(||\boldsymbol{x}-\boldsymbol{u}||/\theta)C_\nu(||\boldsymbol{u}-\boldsymbol{v}||/\kappa)\phi(||\boldsymbol{x}'-\boldsymbol{v}||/\theta)d\boldsymbol{u}d\boldsymbol{v}. \tag{12}$$

Outline of proof

Let $\widetilde{\phi}_k$ be the spectral density for ϕ and $\widetilde{C}_\nu$ the spectral density of a Matérn field with $\nu = 1$, unit variance and unit spatial scale parameter. Including the scale parameter for the radial basis function kernel and using elementary properties of convolution.

$$\widetilde{S}(r) = \sum_{l=1}^{\infty}\alpha_l\left[\theta_l^2\widetilde{C_\nu}(\theta_l r)\right]\left[\widetilde{\phi_k}(\theta_l r)\right]^2$$

The Matérn spectral density is

$$\widetilde{C_\nu}(r) = \frac{1}{(2\pi)}\frac{1}{(1+r^2)^2}.$$

For the Wendland spectral density there are constants C_1 and C_2 depending only on K such that for all ω

$$C_1 \leq \widetilde{\phi_k}(\omega)(1+\|\omega\|^2)^{3/2+K} \leq C_2,$$

(Wendland (1998)). Using the upper bound on $\widetilde{\phi}$, substituting the expressions for θ_l and α_l and finally combining terms gives the upper bound

$$\widetilde{S}(r) < C'\sum_{l=1}^{\infty}\alpha_l\frac{\theta_l^2}{(1+(r\theta_l)^2)^\eta} = C'\sum_{l=1}^{\infty}\frac{e^{-2\beta_1 l}e^{-2\beta_2 l}}{(1+(re^{-\beta_2 l})^2)^\eta} = C'\sum_{l=1}^{\infty}\frac{e^{-(2\beta_1+2\beta_2)l}}{(1+r^2e^{-2\beta_2 l})^\eta},$$

with $\eta = 2+2(3/2+K) = 5+2K$.

Now apply the useful lemma given below with the identifications $a = 2\beta_1+2\beta_2$, $b = 2\beta_2$ and $c = \eta$ and $s = r^2$. We have the rate given by $r^{-2(a/b)}$ and with $2a/b = 2(2\beta_1+2\beta_2)/2\beta_2 = 2\beta_2/\beta_1+2$. The result for the upper bound now follows and the rate for the lower bound is proved in a similar manner.

Two Useful Lemmas

Lemma 6.1. *Let H be a continuous and integrable function on $[1,\infty]$. Also assume that H is positive and unimodal with maximum at u^*.*

$$\left|\sum_{l=1}^{\infty} H(l) - \int_1^{\infty} H(u)du\right| < H(u^*)$$

Proof Let L be the integer so that $H(L) = max_l H(l)$ also let $I_l = \int_l^{l+1} H(u)du$ then by elementary properties of the integral and the unimodality of H

$$\begin{aligned} I_l &> H(l), \qquad 1 \le l \le (L-1) \\ I_{l-1} &> H(l), \quad (L+1) \le l \le \infty \end{aligned} \tag{13}$$

summing over l gives

$$\sum_{l=1}^{\infty} I_l > \sum_{l \neq L} H(l)$$

Simplifying and rearranging terms

$$\int_1^{\infty} H(u)du - \sum_{l=1}^{\infty} H(l) > -H(L)$$

Again by properties of the integral and H

$$\begin{aligned} I_{l-1} &< H(l), \quad 2 \le l \le L \\ I_l &< H(l), \qquad (L+1) \le l \le \infty \end{aligned} \tag{14}$$

summing over l gives

$$\int_1^{\infty} H(u)du < \sum_{l \neq L} H(l)$$

or

$$\int_1^{\infty} H(u)du - \sum_{l=1}^{\infty} H(l) < H(L)$$

Noting that $H(L) < H(u^*)$ the result now follows.

Lemma 6.2. *For $a, b, c, s > 0$ and for $(a/b) - c < 0$ there are constants $0 < C_1, C_2 < \infty$*

$$C_1 s^{-a/b} < \sum_{l=1}^{\infty} \frac{e^{-al}}{(1 + se^{-bl})^c} < C_2 s^{-a/b}$$

Proof Based on Lemma 1 let $H(u) = e^{-au}/(1 + se^{-bu})^c$. H is unimodal. From basic calculus the maximum of H is $H(u^*) = Cs^{-a/b}$ for $0 < C < \infty$ and C depending only on a, b, c. We now evaluate the approximating integral from Lemma 1 as a function of s.

$$\int_1^\infty H(u)du = \int_1^\infty \frac{e^{-au}du}{(1+se^{-bu})^c} = \int_1^\infty \frac{(e^{-bu})^{a/b}du}{(1+se^{-bu})^c}$$

Now make the substitution $q = e^{-bu}$ giving $dq = -b(e^{bu})du$ or $du = \frac{-dq}{bq}$ and with limits of integration, e^{-b} and 0. One obtains

$$b\int_0^{e^{-b}} \frac{q^{(a/b)-1}dq}{(1+sq)^c} \tag{15}$$

Since $(a/b) > 0$ the pole at zero is integrable and the integral is finite. Now make the substitution $p = sq$ giving $dp = sdq$ and

$$b\int_0^{se^{-b}} \frac{(p/s)^{(a/b)-1}dp}{s(1+p)^c} = bs^{-a/b}\int_0^{se^{-b}} \frac{p^{(a/b)-1}dp}{(1+p)^c} \tag{16}$$

Under the assumption that $a/b - c < 0$ the integral will be finite in the limit as $s \to \infty$. Thus $\int_1^\infty H(u)du$ and $H(u^*)$ converge to zero at the polynomial rate $s^{-a/b}$ and the result follows from application of Lemma 1.