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 *

October 10, 2012

Abstract

A multi-resolution basis is developed to predict two-dimensional spatial fields based on irregularly spaced observations. The basis functions at each level of resolution are constructed as radial basis functions 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 spatial predictions and also the evaluation of the likelihood. Although our framework has similarities to fixed rank Kriging, the model gives a better approximation to situations where the nugget variance is small and the spatial process is close to interpolating the observations.

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

DRAFT

1 Introduction

Statistical methodology for spatial data 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 spatial

*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.

data and indirect measurements of the spatial process. This development in spatial statistics is coincident with emerging challenges in the geosciences involving new types of observations and comparisons of data 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 and the application of standard techniques is often not feasible or at least will take an unacceptably long time given typical computational resources. Moreover, geophysical processes tend to have a multi-scale character over space that requires the need to apply statistical methods that do not assume a simple parametric model for dependence across a region. This work develops a new statistical model that addresses both of these challenges; our model is applicable to large data sets and has a flexible covariance model and so fills a gap in current statistical methodology.

DRAFT

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 surface and $\boldsymbol{\epsilon}_i$ are mean zero measurement errors. The parameters $\boldsymbol{d}$ represent fixed effects in this model and g is assumed to be a realization of a random process with the form

$$g(\boldsymbol{x}) = \sum_{j=1}^{m} \boldsymbol{c}_j \phi_j(\boldsymbol{x}). \tag{2}$$

Here ϕ_j , $1 \leq j \leq m$ is a sequence of fixed basis functions, $\boldsymbol{c}$ is a vector of coefficients distributed multivariate normal with mean zero and covariance matrix, ρP. Here $\rho > 0$ is useful as a leading scaling parameter for a covariance matrix, P, that may also depend on other parameters. Thus the model for g is a sum of fixed basis functions with stochastic coefficients.

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, perhaps in a nearly interactive mode, we seek to balance the complexity of the models and methodology with feasibility for effective data analysis. In this paper, we will focus on maximum likelihood estimates of parameters in the covariance and other components from (1) and (2). For inference on the spatial process we will focus on the conditional distribution of g given the data and other statistical parameters. Although this basic approach to inference can be expanded to a Bayesian hierarchical approach, our model and the computational scheme will still play a central role in the core of any Bayesian computation.

Our approach combines the representation of a field using a multi-resolution 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 (2) as a sum of L independent processes, $g_l(\boldsymbol{x})$ for $1 \leq l \leq L$, each with a different correlation scale and marginal variance, $\{\alpha_l\}$.

DRAFT

$$g(\boldsymbol{x}) = \sum_{l=1}^{L} \sqrt{\alpha}_l g_l(\boldsymbol{x}) \tag{3}$$

In this way the overall spatial dependence of g can be much more complex than each of the individual components. The multi-resolution structure has a similarity with wavelet estimators although we do not assume orthogonality of the basis functions or sparsity in the coefficients. Each component, g_l is defined as a basis function expansion as in (2). The main principle is to expand the two-dimensional spatial field in families of radial basis functions that are organized on regular grids of increasing resolution and correspond to the schematic of a sum of L independent processes. The 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 $Q = P^{-1}$. The benefit of this approach is that Q is sparse even though P itself can be dense. Thus, g 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 a multi-resolution basis 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 distribution of the locations or locations for prediction 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 some normalizations that 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 P matrix such as in Cressie and Johannesson (2008) or large number of basis and a sparse model such a Markov random field for Q (Eidsvik et al. 2010). An insightful approach was suggested

DRAFT

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. This section also gives details of the basis function construction and the particular Markov random field used for the basis coefficients. 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 with 1760 locations for the North American region and is analyzed using a standard Matérn covariance and the multi-resolution model. Section 6 contains conclusions. Much of the computations in this paper can be reproduced using the `LatticeKrig` package in R and this software serves as a supplement for implementing the numerical methods and a ready source for the data set.

DRAFT

2 Spatial model

2.1 Process and observational models

Based on the set up in the introduction g will be a mean zero, Gaussian process with covariance function:

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

With respect to the observation model in (1) we assume that $\boldsymbol{\epsilon}_i$ are uncorrelated,

normally distributed with mean zero and covariance $\sigma^2 W^{-1}$. Here we assume that σ^2 is a free parameter of the measurement error distribution and W is a known but sparse precision matrix. In most applications W is diagonal or just the identity. Let Φ be the regression matrix with columns indexing the basis functions and rows indexing locations. $\Phi_{i,j} = \phi_j(\boldsymbol{x}_i)$. With these definitions one can now reexpress (1) in matrix vector notation as

$$\boldsymbol{y} = Z\boldsymbol{d} + \Phi\boldsymbol{c} + \boldsymbol{e}$$

and collecting the fixed and random components we have

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

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 and ρ in terms of λ and ρ (i.e. $\sigma^2 = \lambda\rho$). Now set $M_\lambda = (\Phi P\Phi^T + \lambda W^{-1})$ and (5) is the same as

$$\boldsymbol{y} \sim MN(Z\boldsymbol{d}, \rho M_\lambda) \tag{6}$$

DRAFT

2.2 Spatial estimate

From (6) we have the log likelihood

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

As a component of a full Bayesian model this represents part of the marginal distribution having integrated over $\boldsymbol{c}$.

This expression is used to find maximum likelihood estimates (MLEs) of the fixed effects and covariance parameters. In finding this estimate it is convenient to first maximize over the fixed effects and then maximize over some of the covariance parameters. For fixed ρ, P and σ the MLEs for $\boldsymbol{d}$ are also the generalized least squares (GLS) estimates

$$\hat{\boldsymbol{d}} = (Z^T M_\lambda^{-1} Z)^{-1} Z^T M_\lambda^{-1} \boldsymbol{y}. \tag{7}$$

Note this estimate only depends on λ and not on ρ. To maximize ℓ on the remaining parameters set $\boldsymbol{r} = \boldsymbol{y} - Z\hat{\boldsymbol{d}}$ and the profile log likelihood is now

$$\ell(\rho, P, \sigma, \hat{\boldsymbol{d}}) = (-1/2)(\boldsymbol{r}^T(\rho M)_\lambda^{-1}\boldsymbol{r}) - (1/2)log|\rho M_\lambda| + (n/2)log(\pi). \tag{8}$$

Finally, the expression given above can be maximized over ρ giving $\hat{\rho} = \boldsymbol{r}^T M_\lambda^{-1}\boldsymbol{r}/n$. This estimate can be substituted back into (8) to give a profile log likelihood in λ and any covariance parameters that contribute to P.

Although in this paper we will focus on maximum likelihood estimation for inference, one could also consider a prior distribution on the covariance parameters and carry out a Bayesian inference for these model components. In doing so and also integrating out the dependence of the posterior on $\boldsymbol{c}$, one obtains a log marginal posterior that is the sum of (8) and a log prior density on the covariance parameters and those of $\boldsymbol{d}$. Thus except for the influence of the priors, (8) comprises the log posterior for inference on the model parameters. Considering a uniform (and improper) prior on $\boldsymbol{d}$, $\hat{\boldsymbol{d}}$ will also be the posterior mode from a Bayesian perspective. A uniform and improper prior on $\boldsymbol{d}$ may appear to be strange, but in many geostatistical applications it is a reasonable approximation. Typically the likelihood will be informative for this small set of regression parameters and will dominate any diffuse prior. Thus considering the limiting case as the prior becomes diffuse is a reasonable assumption.

DRAFT

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

$$[\boldsymbol{c}|\boldsymbol{y}, \boldsymbol{d}, \sigma, \rho, P] \sim MN(\hat{\boldsymbol{c}}, \rho P - \rho P \Phi^T (M_\lambda)^{-1} \Phi P) \tag{9}$$

with

$$\hat{\boldsymbol{c}} = P\Phi^T M_\lambda^{-1}(\boldsymbol{y} - Z\boldsymbol{d}) \tag{10}$$

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

This estimate is also well-known as the best linear unbaised (BLUE) estimator. The conditional variance of $g(\boldsymbol{x})$ can be interpreted as the mean squared error from a geostatistical perspective or from a Bayesian perspective quantifying the uncertainty in $g(\boldsymbol{x})$ when the other parameters are known. Although there are closed form expressions for the conditional covariance it is usually more efficient computationally to draw samples from the posterior and quantify the uncertainty in the estimate by Monte Carlo sample statistics. In geostatistics this technique is referred to as conditional sampling and an efficient algorithm for the sampling is given in the next section. Typically this conditional sampling is done at the estimated covariance parameters but under a Bayesian analysis one can sample from the marginal posterior for these parameters. By varying the covariance parameters in this manner in drawing the conditional samples one also includes the uncertainty in the predictions with respect to estimating the covariance model.

DRAFT

2.3 Radial Basis functions (RBF)

Our full model proposes a multi-resolution 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. Here we "unroll" the indices over the two dimensional grid into a single index to make the notation simpler.

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

$$\phi_j^* = \phi(||\boldsymbol{x} - \boldsymbol{u}_j||/\theta) \tag{11}$$

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 in their argument and can be constructed with different degrees of differentiability at the origin. They are also positive definite and this property is

attractive 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:

$$\phi(d) = \begin{cases} (1-d)^6(35d^2+18d+3)/3 & \text{for } 0 \le d \le 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

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 multi-resolution aspect replicates this model at each level. $\boldsymbol{c}$ 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. An alternative is to find the elements of Q directly as inner products of the basis functions. Specifically if $\mathcal{L}$ is the differential operator motivated from a partial differential equation then

$$[Q]_{jk} = \int \mathcal{L}(\Phi_j)\mathcal{L}(\Phi_j)d\boldsymbol{x}$$

DRAFT

will provide an finite approximation to the continuous process. We take a more direct approach to specifying Q and this avoids the integrations to determine the inner products. These integrals may become difficult in more than two dimensions or if the lattice is not on a rectangular grid. However, this strategy does have the advantage that there is an unambiguous limiting spatial model as the basis becomes dense. Finally, we note that the discretization of the SPDE based on a particular choice of basis functions especially for a limited basis size, may not remove all of the non-stationary artifacts and so adjustments such as those described in Section 2.6 may still be needed.

Given an autoregression matrix B and $\boldsymbol{e}$, a random vector distributed $N(0, \rho I)$, we construct the distribution of $\boldsymbol{c}$ according to $\boldsymbol{c} = B^{-1}\boldsymbol{e}$. The autoregressive interpretation is that $B\boldsymbol{c} = \boldsymbol{e}$. That is, B transforms the correlated field to discrete white noise with variance ρ. For our use we will constrain B to be sparse. Each row has a small

number of non-zero elements which we will define based on the nodes of the corresponding basis functions. Although one may generalize the elements to other regular node patterns, here the node points will form a rectangular grid. 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 $B_{j,j} = 4+\kappa^2$ with $\kappa \geq 0$ and the off diagonal elements to be -1. Adjustments made to B on the edges and corners are given in Table 1 and enforce different boundary conditions, in this case either free derivatives at the boundaries or boundary conditions of zero. By linearity $\boldsymbol{c}$ has covariance matrix $\rho B^{-1}B^{-T}$ and precision matrix given by $Q = (1/\rho)B^TB$. Because B is formulated as unconditional weights on the field, any choice of B will lead to a valid covariance and so 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 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 from Table 1 will approximate a Matérn process with scale parameter κ in LR2011 and smoothness $\nu = 1$.

DRAFT

2.5 Multi-resolution

From the previous sections we have developed a basis and covariance for a specific grid. The multi-resolution 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

	Interior			Left edge			Upper left corner		
	.	-1	.	-1/2	.	.	1+ $\kappa^2/4$	-1/2	.
Free BC	-1	$4+\kappa^2$	-1	2+ $\kappa^2/2$	-1		-1/2	.	.
	.	-1	.	-1/2	.	.	.	.	.
	.	-1	.	-1	.	.	4+ κ^2	-1	.
Zero BC	-1	$4+\kappa^2$	-1	4+ κ^2	-1		-1	.	.
	.	-1	.	-1	.	.	.	.	.

Table 1: Autoregressive weights used to define a Markov random field for $\boldsymbol{c}$. Each table is a schematic of the basis coefficients organized in a regular grid according to the locations of the basis function centers. The weights correspond to the different cases of when a coefficient is associated with an edge, corner or interior grid point and whether an edge correction is made.

DRAFT

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 (11). Let L denote the total number of levels then the (unnormalized) multi-resolution basis is

$$\phi^*_{j,l} = \phi(||\boldsymbol{x} - \boldsymbol{u}^l_j||/\theta_l)$$

where $1 \leq l \leq L$ and $1 \leq j \leq m(l)$ and $m(l) = (m_x - 1)(m_y - 1)4^{l-1} + m_x + m_y + 1$

The vectors of coefficients associated with each level will be denoted by $\boldsymbol{c}^l$ and so the multi-resolution representation for g is

$$g(\boldsymbol{x}) = \sum_{j=1}^{m(1)} \boldsymbol{c}^1_j \phi^*_{j,1}(\boldsymbol{x}) + \sum_{j=1}^{m(2)} \boldsymbol{c}^2_j \phi^*_{j,2}(\boldsymbol{x}) + \ldots + \sum_{j=1}^{m(L)} \boldsymbol{c}^L_j \phi^*_{j,L}(\boldsymbol{x}) \tag{12}$$

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 B such that $B\boldsymbol{C}$ is $N(0, \rho I)$. Although B be a general matrix we have found it useful to restrict attention to a block diagonal form. Equation (12) is written in this way and is suggestive of the process being a sum of L processes. We assume that they are also independent. 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,

$(1/\sqrt{\alpha_l})B_l$. Here 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 B and also for the precision matrix:

$$\boldsymbol{Q} = (1/\rho)\begin{bmatrix} (1/\alpha_1)(B_1)^T B_1 & 0 & \dots & 0 \\ 0 & (1/\alpha_2)(B_2)^T B_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & (1/\alpha_L)(B_L)^T B_L \end{bmatrix} \tag{13}$$

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.

DRAFT

For both free and zero boundary edge adjustments there is substantial difference in the variance over the domain. 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, the normalization can take advantage of the overall strategy of using sparse matrix decompositions and we believe reducing edge effects and other artifacts is worth the extra computation.

3 Computation

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 Φ, W and Q to be sparse matrices. As introduction

we use throughout the fact that the sparse Cholesky decomposition has two steps: a symbolic factorization to determine the sparsity pattern of the decomposition and then a numerical factorization to compute the triangular matrix. In several instances the sparsity pattern is fixed and only the values of the entries change. Anticipating this case one stores the symbolic factorization and only recomputes the numerical portion. In our *R* implementation using the `spam` sparse matrix package (Furrer and Sain 2010), we have found that approximately half of the computational time in the decomposition is taken up by the symbolic factorization.

3.1 Normalization to constant marginal variance

Computing the normalization $\omega(\boldsymbol{x})$ breaks into the following steps:

1. Form the sparse precision matrix $\boldsymbol{Q}$.
2. Compute the sparse Cholesky decomposition $\boldsymbol{Q} = AA^T$.
3. Solve the triangular linear system $A\boldsymbol{v} = \boldsymbol{z}$ with $\boldsymbol{z}_j = \phi_j(\boldsymbol{x})$.
4. $\omega(\boldsymbol{x}) = (\sum_{j=1}^{m} \boldsymbol{z}_j^2)^{-1/2}$.

DRAFT

For finding the normalization at more than one value of $\boldsymbol{x}$ the Cholesky decomposition will not change and so steps 1 and 2 need only be done once. Also for changing values of the GMRF the sparsity pattern does not change and so the symbolic factorization of $\boldsymbol{Q}$ can be reused for different sets of covariance parameters. For some GMRF weights it possible to identify this problem with a discretization of the classical Poisson problem in two dimensions from numerical analysis. There exist very efficient fast Poisson solvers that could eliminate the computational overhead in this step.

3.2 Estimating the basis coefficients

A common calculation to find the concentrated likelihood in (5) and $\hat{\boldsymbol{d}}$ in (7) is to evaluate $M_\lambda^{-1}\boldsymbol{w}$ for an arbitrary vector $\boldsymbol{w}$. Recall that $M_\lambda = \Phi P \Phi^T + \lambda W^{-1}$ and taken at face value M_λ is a dense, potentially large matrix and so difficult to work with directly. The strategy is to transform M_λ using matrix identities to involve the sparse precision matrix. The matrix identities themselves are well known and central to

reducing the dimensions of the calculations in fixed rank Kriging. They work by shifting the linear system size from matrices with dimensions of the number of observations (n) to dimensions of the number of basis functions (m). Usual fixed rank Kriging has m much In our application, however, exploiting these identities has the potential effect to greatly *increase* the nominal dimensions of the problem. When the number of basis functions exceeds that of the observations the matrix sizes will be larger but due to sparsity of the matrices the computations are feasible. In fact as reported in Section 3.5 the computations can be much faster than a direct calculation.

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

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

where $G = \Phi^T W\Phi + \lambda Q$. Using this identity one can now use the Cholesky decomposition for G to solve the linear system $G\boldsymbol{v} = (\Phi^T W)\boldsymbol{w}$ for $\boldsymbol{v}$ and it follows that

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

DRAFT

Because Φ, W and Q are all sparse, G will also be sparse. Moreover it also follows that G is positive definite. Thus the linear system can be solved using a sparse Cholesky decomposition. Note that an important limitation of this computational strategy is that λ can not be identically zero. In practice this is not an issue as λ can be set to small enough to give a result that is close to interpolation and is still numerically stable.

To find $\hat{\boldsymbol{c}}$ as in (10) one can use the Sherman-Morrison-Woodbury formula or the representation of the conditional multivariate normal in terms of precision matrices to derive the identity

$$P\Phi^T M_\lambda^{-1} = P\Phi^T(W - (W\Phi)G^{-1}(\Phi^T W)) = G^{-1}\Phi^T W$$

Thus $\hat{\boldsymbol{c}} = G^{-1}\Phi^T W(\boldsymbol{y} - Z\hat{\boldsymbol{d}})$ is found by taking advantage of sparsity of Φ and W for multiplication and the sparse Cholesky factorization of G. The evaluation of $\hat{g}(\boldsymbol{x})$ can also be computated in an efficient manner if the sum is restricted to basis functions that are nonzero at $\boldsymbol{x}$.

3.3 Evaluating the determinant of the observation covariance matrix

The other intensive computation occurs in the likelihood as the determinant of M_λ.

Using elementary properties of determinants

$$|M_\lambda| = |\Phi P \Phi^t + \lambda W^{-1}| = |\lambda W^{-1}||UU^T + I_n| = \frac{\lambda^n |UU^T + I_n|}{|W|} \tag{14}$$

with $U = (\lambda)^{-1/2} W^{1/2} \Phi P^{1/2}$, adopting symmetric square roots of W and P, and I_n being an $n \times n$ identity matrix.

Now one can use a special case of Sylvester's Theorem. For an $n \times m$ matrix U and identity matrices I_n and I_m. $|UU^T + I_n| = |U^T U + I_m|$.

Applying this identity with the choice of U given above

$$\begin{aligned} |U^T U + I_n| = |U^T U + I_m| &= |(1/\lambda) P^{1/2} (\Phi^T W \Phi) P^{1/2} + I_m| \\ &= \lambda^{-m} |P^{-1/2}||\Phi^T W \Phi + \lambda P^{-1}||P^{-1/2}| \\ &= \lambda^{-m} |\Phi^T W \Phi + \lambda P^{-1}||P^{-1}| \end{aligned} \tag{15}$$

substituting for G and Q we have $|U^T U + I_n| = \lambda^{-m}|G|/|Q|$ and so combining with (14)

$$|M_\lambda| = \frac{\lambda^{n-m}|G|}{|Q||W|}$$

The matrices, W, G and 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.

3.4 Conditional Simulation

In this section we describe a Monte Carlo algorithm to sample from the conditional distribution of $\boldsymbol{d}$ and $\boldsymbol{c}$ (or g) given the observations and the covariance model. From Section 2 the joint distribution of $\boldsymbol{c}$ and $\boldsymbol{d}$ conditioned on $\boldsymbol{y}$ and the covariance parameters is multivariate normal with mean given by $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{c}}$. We use the following algorithm to generate a sample from the conditional multivariate distribution.

1. Generate $\boldsymbol{c}$ as $N(0, \rho P)$ by solving the triangular system $A\boldsymbol{c} = \boldsymbol{w}$ where $AA^T = 1/\rho Q$ and $\boldsymbol{w} \sim N(0, I)$

DRAFT

2. Generate synthetic observations $\boldsymbol{y}^* = \Phi \boldsymbol{c} + \boldsymbol{e}^*$ with $\boldsymbol{e}^* \sim N(0, \sigma^2 W^{-1})$ and compute estimates $\hat{\boldsymbol{d}}^*$ and $\hat{\boldsymbol{c}}^*$ based on $\boldsymbol{y}^*$

3. $\boldsymbol{d}^* = \hat{\boldsymbol{d}} + \hat{\boldsymbol{d}}^*$ and $\boldsymbol{c}^* = \hat{\boldsymbol{c}} + (\hat{\boldsymbol{c}}^* - \boldsymbol{c})$ will be a draw from the draw conditional distribution for $(\boldsymbol{d}, \boldsymbol{c})$ and a draw from the predictive distribution for $\boldsymbol{y}$ is given by

$$y(\boldsymbol{x}) = Z(\boldsymbol{x})^T \boldsymbol{d}^* + \sum_{j=1}^{m} \phi_j(\boldsymbol{x}) \boldsymbol{c}_j^*$$

3.5 Timing

Here we present some timings for the computations with the main comparison being the dense matrix computations associated with Kriging. The number of spatial locations was varied between 500 and 20000 and spatial predictions were found for an exponential covariance model and several choices of the lattice multi-resolution model. Note that the computation time is essentially independent of the spatial data, the distribution of spatial locations, and the 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 multi-resolution basis function model. Times reported are for a Macbook Pro laptop (2.3 Ghz Intel Core i7) and R 2.14 (R Development Core Team 2011). Both of these functions compute the predictions at the observations for a fixed covariance model, evaluate the likelihood, compute the coefficients for predicting the surface at arbitrary points, and an approximate estimate of the generalized cross-validation function. 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 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, 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

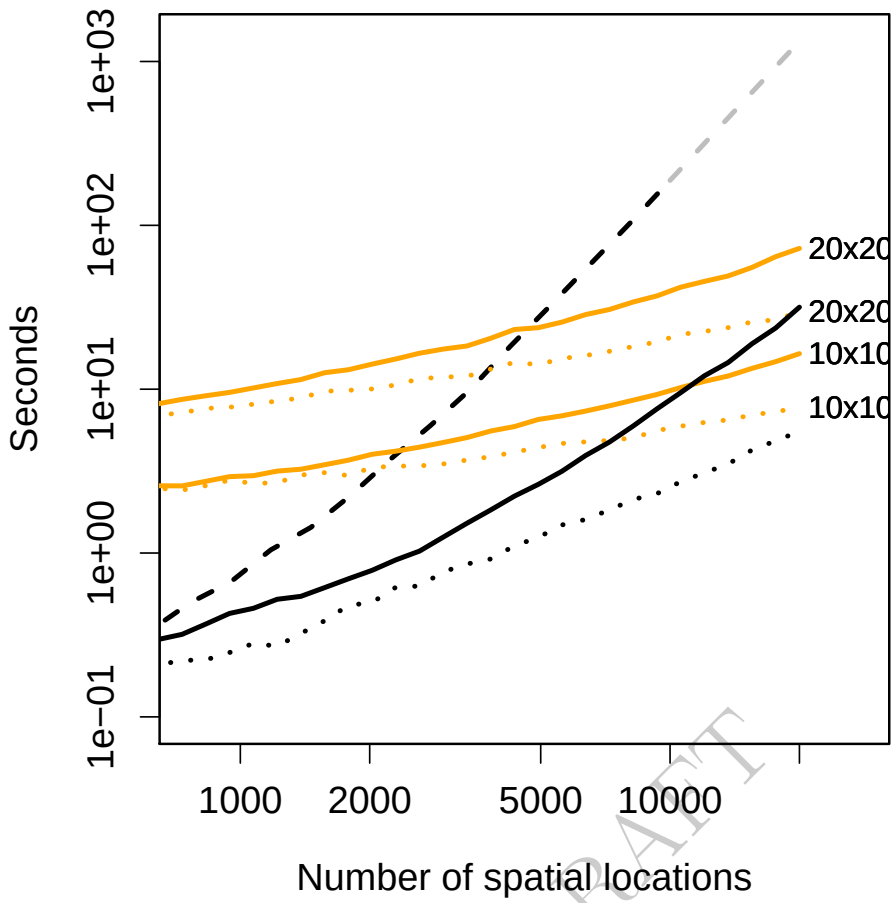
DRAFT

line black line is the same scheme but without normalizing the basis functions. For 20000 spatial locations the times for these models are 31 seconds (normalized) and 5.5 seconds (unnormalized).

The grey lines report timing with the number of basis functions keep fixed. The line labeled 10 has four levels of multi resolution and where the coarsest basis has centers on a 10×10 grid and totaling 7159 basis functions. The line labeled 20 is the same specification but with the coarsest grid being a 20×20 grid with a total of 31,259 basis functions. The solid lines are timings with normalized basis functions and dotted are without normalization.

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 G discussed in Section 3.2. For this work we have not exploited more efficient algorithm to reduce the work in normalization step and this is the reason for the appreciable difference between the normalized and unnormalized cases. As might be expected the two covariance models with fixed degrees of freedom (7,159 and 31,259) are closer to being linear as a function sample size because of the sparse construction of Φ.

DRAFT



DRAFT

Figure 1: Timing results for the lattice/basis model and standard Kriging. Plotted are times in seconds as the number of spatial locations is varied between 500 and 20000. Spatial locations are generated from the uniform distribution on the domain $[-1, 1] \times [-1, 1]$. (However, the particular values spatial locations do not effect these timing results.) 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 multi-resolution 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.

4 Properties of the covariance model

In this section we present some theoretical expansions and some exact computations that outline the properties of the covariance family introduced in Section 2. As a foundation, we first consider a convolution approximation to the sum over radial basis functions.

4.1 Process convolution

It is simplest to 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 given by

$$\int_{\boldsymbol{R}^2}\int_{\boldsymbol{R}^2} \frac{1}{\theta^4}\phi(||\boldsymbol{x} - \boldsymbol{u}||/\theta)C_\nu(||\boldsymbol{u} - \boldsymbol{v}||/\kappa)\phi(||\boldsymbol{x}' - \boldsymbol{v}||/\theta)d\boldsymbol{u}d\boldsymbol{v}. \tag{16}$$

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 (16) 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 multi-resolution 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

DRAFT

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 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 multi-resolution process for different choices of the scale and weight sequences. An interesting result is that the multi-resolution 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.

DRAFT

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

$$A_1 < S(r) r^{2\mu+2} < A_2$$

with $\mu = \beta_1/\beta_2$.

Corollary 4.2. *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

Given the theoretical foundation for the multi-resolution covariance it is interesting to explore the approximation for a finite number of levels and for the lattice process. As an example of this property ,the setup from Theorem 1 was evaluated to approximate members of the Matérn family. Figure 2 gives the resulting approximations for 2, 4, and 6 multi-resolution levels (k_l, for l being 2,4,and 6) and for smoothness $\mu = 0.5$ (exponential) and 1.0 (Whittle). The multi-resolution and the Matérn correlation functions are isotropic and so their are plotted against the distance of separation. The weight formulae are taken from the Corollary to Theorem 1. The Matérn covariance used for the convolution has smoothness 1.0 with a scale of 1.0 and the Wendland function, ϕ, has smoothness degree 2.0 with a scale of 2.5 (i.e. identically zero outside a radius of 2.5). The Matérn covariance function for comparison was determined by finding the scale parameter that minimizes the sum of squares (i.e. a discrete L_2 norm) between the weighted mixture given by the multi-resolution model and the Matérn family. Although one might derive this value from the asymptotic formula in practice this would be estimated and so it seems reasonable to choose a scale as the result of an optimization of the norm. This gave 0.159 for $\mu = .5$ and .131 for $\mu = 1.0$.

DRAFT

The approximation improves with increasing level and the benefit from more levels is most apparent near the origin. Moreover the approximation can be understood by the support of the basis functions from the first row of plots. Here we see that finest level of the multi-resolution correlation function has a half height on the order of .05 and this distance is about the order where the six level approximation breaks down for the exponential. In general we have found the approximation to be accurate provided enough levels of resolution are included. As a rule of thumb 6 levels seems to be a practical upper limit to represent members of the Matérn family.

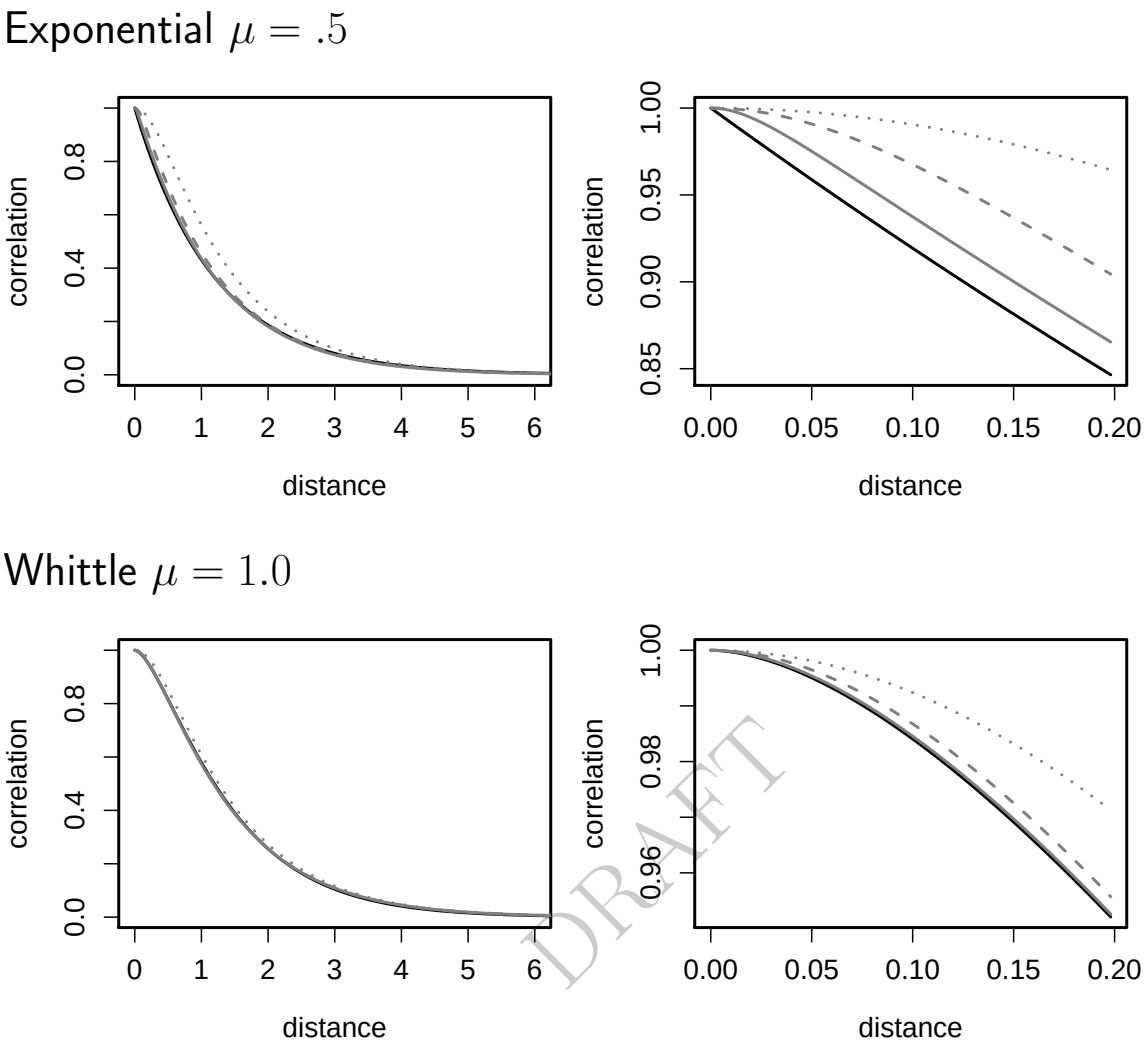


Figure 2: Theoretical approximation of Matérn correlation functions using multi resolution process based on process convolutions. The first row plots the approximations to an exponential function (black solid). The grey lines are the approximations for 2 (dotted), 4 (dashed) and 6 (solid) levels of the multi-resolution. On the left side is the full range of this case and the left is a magnification near the origin and with a log scale on the y axis. The second row is organized the same way as the second row but in this case the approximation is to a Whittle covariance function (a Matérn covariance with a smoothness of 1.0).

The theoretical approximation is based on a continuous convolution of the basis functions with the Matérn covariance and does not match the discrete stochastic model used for data analysis. A more practical comparison is how well a multi-resolution

basis can match members of the Matérn family. Here we indicate 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 first row in Figure 3 shows the approximation for an exponential covariance with range parameters .1, .5 and 1.0. using 3 and 4 levels of multi-resolution basis functions. The multi-resolution 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. The plots in the left hand column 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.

DRAFT

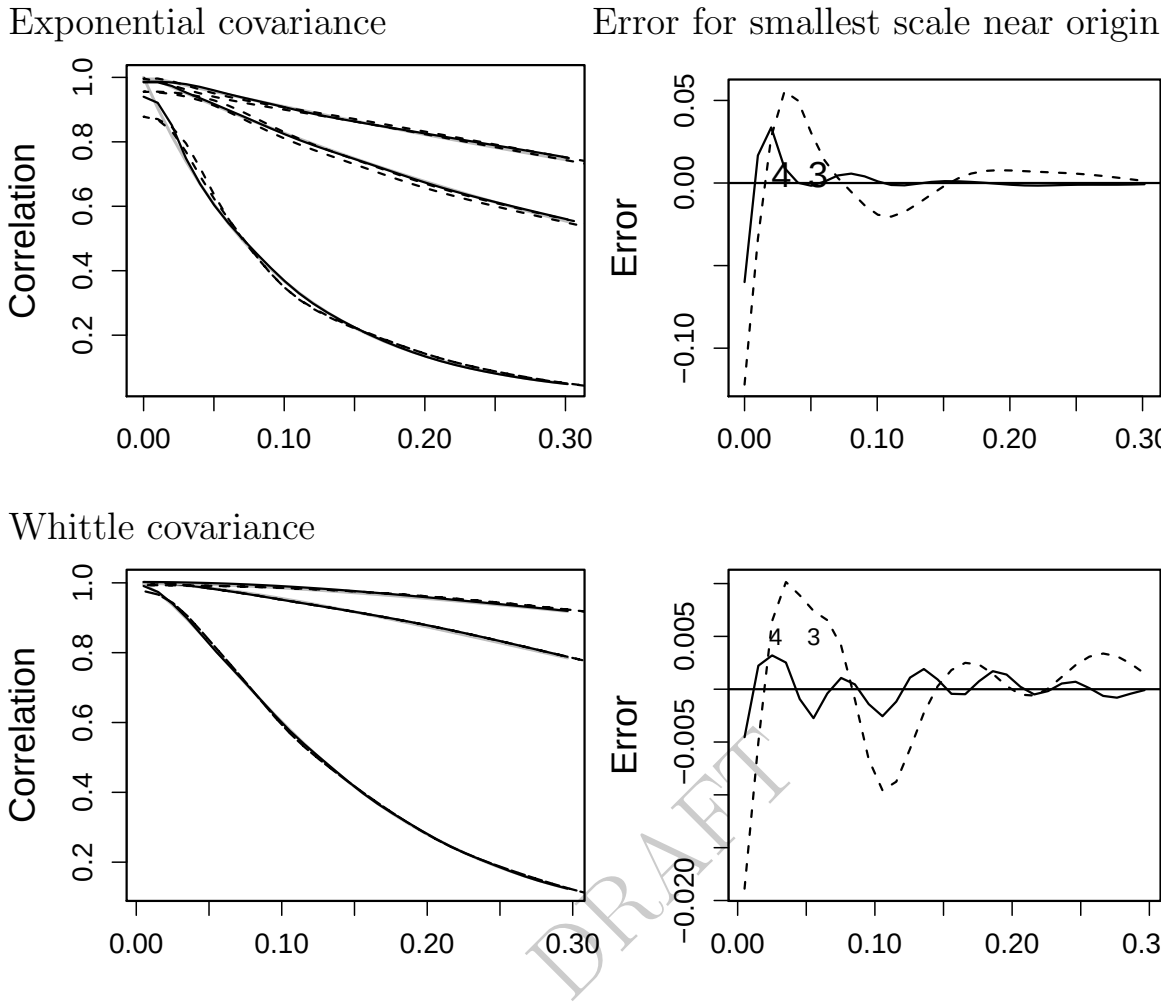


Figure 3: Approximation of Matérn covariances using the lattice/basis model. For the plots on the left hand side columns the solid grey lines are the true correlation functions. First row is an exponential correlation with range parameter (.1, .5 and 1.0), second row is the Whittle correlation with ranges .1,.5 and 1.0 and the third row is a mixture of two exponential correlation functions. Black lines are the approximations to these correlation functions. Approximation are indicated in black and are based on a three level (dashed) or four level (solid) multi-resolution model. The domain for the approximation is a $[-1, 1] \times [-1, 1]$ a square and the initial grid is 10×10 for the range parameters .1, .5. The initial grid for the 1.0 ranges is 5×5. The left columns are the approximations with the true correlations over the distance limits $[0, .3]$ to highlight errors near the origin even though the approximation is found for a spatial domain $[-1, 1] \times [-1, 1]$. The right columns 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.

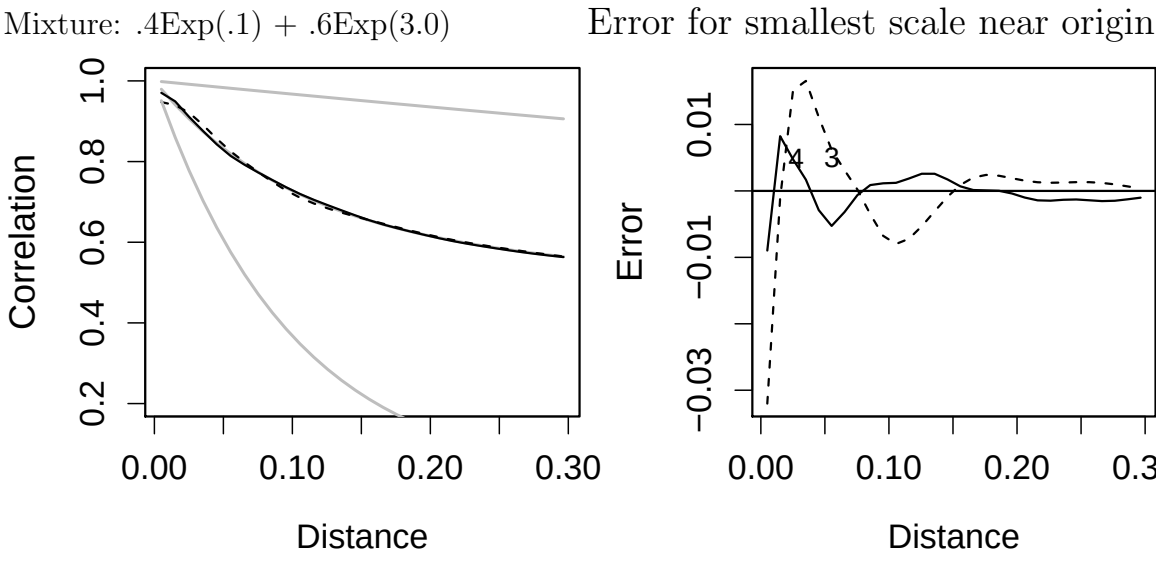


Figure 4: Approximation of a mixture of exponential covariances. The figure is organized the same as a row in the previous but with addition grey lines in the left plot indicating the two separate exponential covariances comprising the mixture.

DRAFT

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 right column 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 multi-resolution for the 3 and 4 level choices. 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). 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. Fiugre 4 is an example of the ability of the multi-resolution 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 dotted lines. The approximation is also accurate with the error localized near the origin and being large below the smallest scale of the multi-resolution.

5 North American summer precipitation

The multi-resolution lattice model was applied to a substantive climate data set in order to test its practical value and compare 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. These type of fields are an important reference in studying the Earth's climate system. This observational climate data can be used to assess the validity of global numerical models that attempt to simulate climate. Typically one requires that irregularly spaced station data be transformed to a regular grid that is comparable to numerical climate model output. Some standard gridded climate datasets are those from the Hadley Center (Jones et al. 2012) and the climate group at the University of Delaware (Willmott and Robeson 1995). However, we suggest that these data products can be improved using spatial methods to quantify the uncertainty and also to facilitate the use of orography (e.g. elevation) for spatial prediction.

The monthly data used to construct this example are freely available as the product Global Historical Climate Network (GHCN) Version 2 Precipitation (NOAA/NCDC 2011). GHCN data is quality controlled, curated and served by the US National Climatic Data Center. The *adjusted* data from this archive has been modified from its raw form to make the record more homogenous. Heterogeneities can come from a variety of sources such as moving the station a short distance or changes in instruments (see `http://www.ncdc.noaa.gov/ghcnm`). The data used here consists of 1720 stations 1595 from the adjusted subset and 125 unadjusted stations in Northern Canada, Alaska and Mexico to fill out the edges of the study domain. For each station, a least squares, straight 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). The trend line for each station at this time is taken to represent the mean climate for this reference period. Note that with complete observations this is just the sample mean. Denote this statistic as R_j for the j^{th} station with a corresponding standard error SE_j. R_j is close to the mean summer precipitation but makes adjustments for stations that do not have complete data records and exhibit linear trends in rainfall. We consider $log(R_j)$ as the observations in (1) to preserve positivity and using the delta method take

DRAFT

the precision of the log transformed rainfall as $W_{i,i} = R_j^2/SE_i^2$. Finally elevation was incorporated as a linear fixed effect in the spatial model and is based on a gridded field at 4km resolution derived from GTOPO30, a global digital elevation model created by the U.S. Geological Survey (see `http://eros.usgs.gov`). For reference the version of the data used here is the R data set `NorthAmericanRainfall` in the `LatticeKrig` package. (Much of the basic analysis is included as source code in the help file for this data set.) For this example we consider only a stationary isotropic covariance functions in stereographic map coordinates (the option `stereographic` from the R `mapproj` package.) This projection gave spatial coordinates whose euclidean distances were similar to great circle distance (see Figure 5). Thus this transformation of coordinates avoided specializing the basis for the sphere. With larger regions this will not be effective and one would have to construct a different set of basis functions and a GMRF better suited to spherical geometry.

A four-level multi-resolution covariance with coarsest level being 10×10 was fit with maximum likelihood estimates (MLEs) for the parameters σ, ρ, $\{\alpha_1, ..., \alpha_4\}$ and a single parameter κ controlling the range. Adjusting for a rectangular domain, i.e. fewer centers in the vertical coordinate, the total number of basis functions is 5782. The variance components,$\{\alpha_k\}$ were constrained to be positive and sum to one and the actual optimization used a mixture model parametrization on a log scale. Specifically, for mixture parameters $\gamma_1, \gamma_2, \gamma_3$

$$\alpha_1 = 1/\boldsymbol{S}, \quad \alpha_2 = e^{\gamma_1}/\boldsymbol{S}, \quad \alpha_3 = e^{\gamma_2}/\boldsymbol{S}, \quad \alpha_4 = e^{\gamma_3}/\boldsymbol{S} \quad \text{with } \boldsymbol{S} = 1 + e^{\gamma_1} + e^{\gamma_2} + e^{\gamma_3}$$

The maximization was done using techniques from the design and analysis of computer experiments (DACE) to be parsimonious about the number of likelihood evaluations. An initial search was done with a *maximin* Latin hypercube design with 50 levels and 50 points (using the `lhs` R package and `dup=10`) on the parameters $\{\alpha_j\} and \kappa$. For each vector of these parameters the likelihood was maximized over σ and ρ. The search was refined using a second design of 25 points about the parameters with the largest likelihood. With this refinement a quadratic approximation was fit to the largest 15 points and this quadratic function was used to sample parameter settings for confidence sets. The uncertainty in the estimated correlation function is summarized by the curve

DRAFT

envelopes in Figure 6 and uses the standard device that -2 times the log likelihood is approximately distributed as a chi squared random variable. For comparison a Matérn covariance was fit and has MLEs for the scale parameter .20 and shape parameter .92. In addition a covariance model using one level of basis functions with the same number and spacing of the 4 level cases was also fit to these data. The estimated correlations for these three cases are compared in Figure 6.

The estimated correlation function based on the Matérn MLEs (dashed line) is distinctly different from the Matérn and this may be due to the flexibility of the multi-resolution model being able to fit the shape near the origin differently than at the longer range correlations. We conjecture that the correlation functions may be tuned largely to the smaller spatial scales of the data and this may explain lack of agreement at larger distances. An intriguing result is that the four level and Matérn models have very similar MLEs for the measurement error (nugget) and marginal process (sill) variances. For the multi-resolution model σ, ρ and the effective number of parameters for the estimated surface are (.0774, .7400, 748.3) and for the Matérn model these parameters are (.0784, 1.263, 772.2). Thus the overall separation into spatial "signal" and "noise" is similar for these models.

DRAFT

Some summaries of the estimated surface for summer precipitation are given in figure 7. The precipitation estimate is evaluated on a 250×250 longitude/latitude grid at the four level covariance model using the MLE estimates and (10) to get $\hat{\boldsymbol{c}}$. Note that the predicted precipitation field from the Gaussian spatial process model are on a log scale and for illustration we transformed the results to the original scale to quantify the uncertainty. Conditional simulation was used to compute an approximate prediction error. Besides being convenient, this Monte Carlo method allows one to adjust for the log transformation. Specifically, 100 fields were sampled from the conditional distribution at (9) using the algorithm from Section 3 and these fields were then centered around the posterior mean/BLUE for log precipitation found exactly from the first step. That is, we sample from the conditional distribution but constrain the first sample moment to be equal to the expected value. Each of the 100 fields were exponentiated and the reported estimates are the mean and coefficient of variation for

this sample. Figure 7 plots the mean estimate for the rainfall surface for the study region and also for a smaller region plots the prediction standard error divided by the mean estimate. Much of the detail in the predicted surface is due to the inclusion of the elevation fixed effect but given the effective number of parameters in the surface is large (748.2) there is also spatial structure from the nonparametric component. The coefficient of variation helps to highlight how uncertainty in estimate depends on data density and mean level. This is particularly apparent in the Western US.

DRAFT

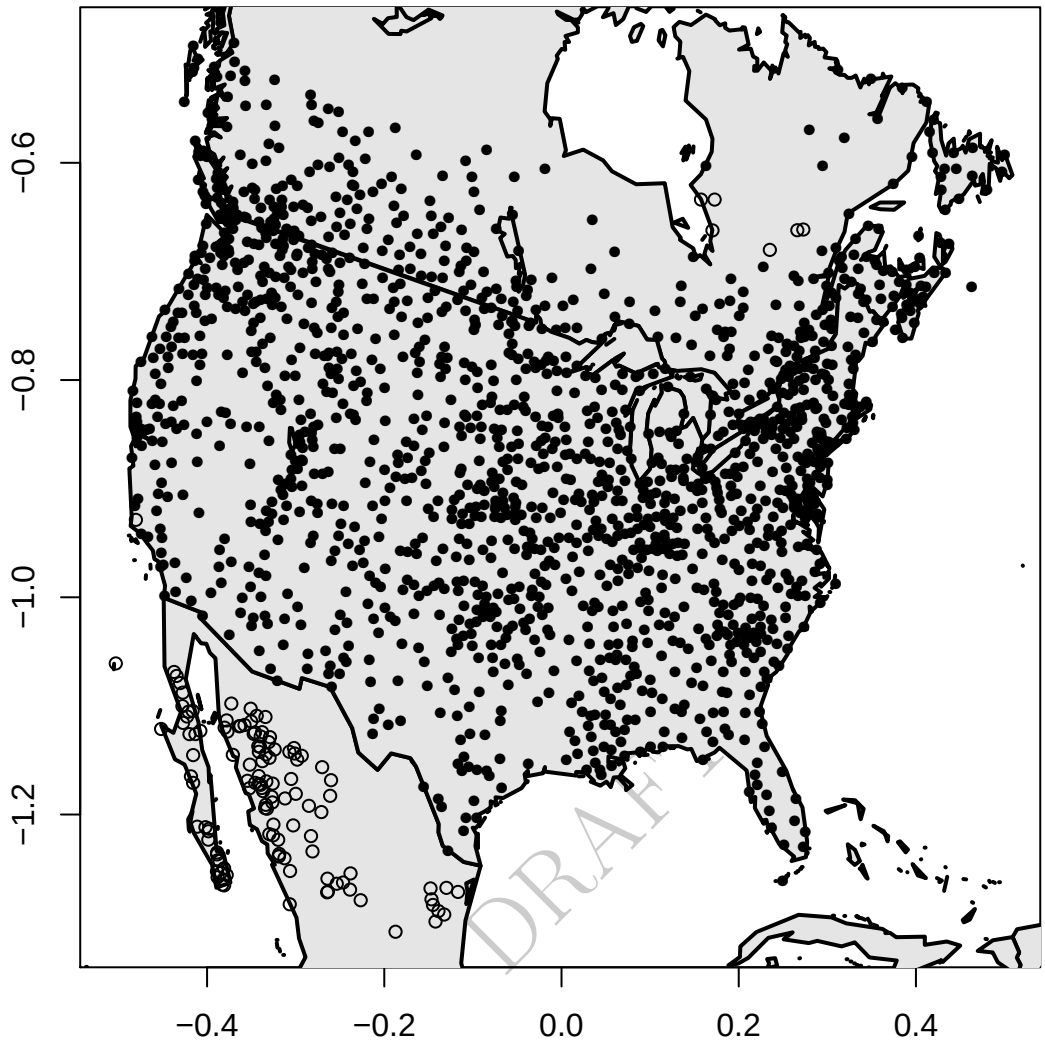


Figure 5: Stereographic projection of precipitation observation locations. Solid points are records that have been adjusted and the open dots are unadjusted station data.

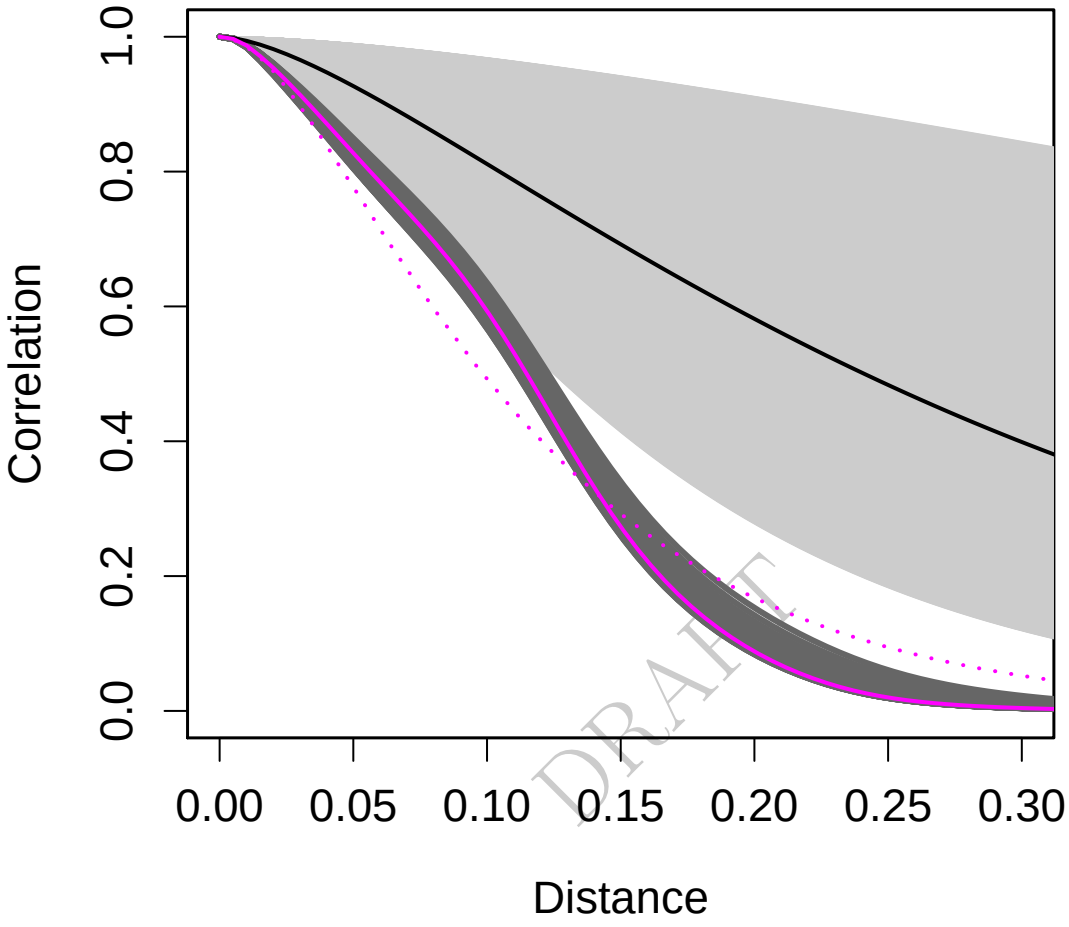


Figure 6: Correlation models fit to the precipitation data. Solid line is the Matérn correlation function found by maximum likelihood and the grey shading is an approximate 95% uncertainty region based on a confidence set for the range and smoothness parameters. Magenta line with grey shading is a similar summary of the correlation covariance for a four level multi resolution model with a initial grid of 10×10 and fit by maximum likelihood. Dotted line is the maximum likelihood estimate for the correlation using a single level but with the fine grid (73×59) matching the fourth level of the multi-resolution model.

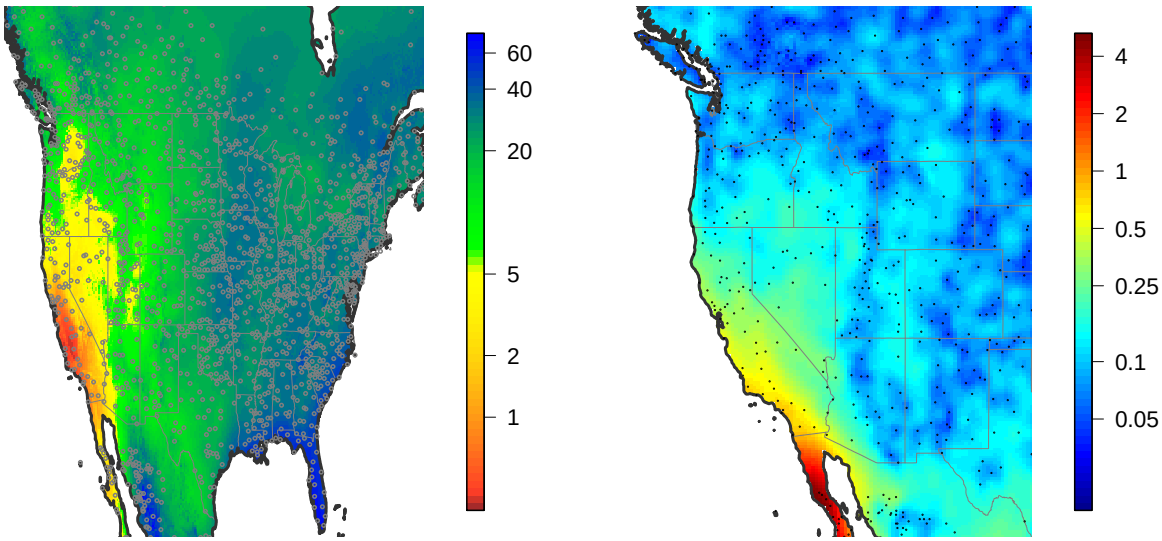


Figure 7: Image on the left reports the spatial predictions for mean summer (JJA) precipitation including elevation as a fixed linear covariate. Units are in millimeters of total rainfall. The spatial covariance function is the four level multi-resolution model described in Figure 6. The uncertainty in the estimated spatial predictions for a Western US subregion of the spatial domain is given in the right image. Plotted is the standard error for the spatial prediction divided by the mean prediction.

DRAFT

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 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. It is still a open task to determine the precise connection

between the convolution model and the discrete sum suggested in Section 4 and so to this extent the theoretical results are at best suggestive of the lattice/basis model properties. The benefits of the lattice/basis model come at some computational costs. In particular we have found it advantageous to normalize the basis functions to give constant marginal variances for a given GMRF. The obvious advantage here is that in fairly heavy-handed way the discrete model and the continuous convolution model agree exactly in terms of the process marginal variance.

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 the computational aspects are intrinsic to applying new spatial models. We have been successful in identifying algorithms that allow for the prediction of the spatial field and computing the restricted likelihood to estimate covariance parameters. In addition, conditional simulation of the spatial fields is a simple method to incorporate uncertainty from both the covariance parameters and also the spatial interpolation or smoothing. Although we have not developed a fully Bayesian algorithm for inference we should emphasize that the main computational ingredients are present to apply integrated nested Laplace approximations (INLA) Rue et al. (2009) or to generate complete Monte Carlo samples from the full posterior distribution.

DRAFT

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. 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. It is still an open question as to how much non-stationary structure can be estimated from data sets currently

encountered in geophysical applications. However, we believe that the lattice/basis formulation will provide competitive computational efficiency and model parsimony compared to other approaches.

Finally, we note that the lattice/basis model can be implemented using a collection of simple numerical algorithms and readily available software. We avoid the triangularization of the spatial locations or computation of basis function inner products and the sequence of regular lattices for the basis function centers simplifies coding the algorithms. An R implementation is available with amply documented and commented source code and uses the generic sparse matrix R package `spam`. The `LatticeKrig` source code is largely written in the R language avoiding lower level C or FORTRAN functions and so is easy to modify. The maximization of the likelihood (or more generally a posterior) can be done using packages and tools in R which makes this model easy to apply for spatial data analysis and inference.

DRAFT

References

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 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. Additional Info: Blackwell Publishing; 20080201; DOI: 10.1111/j.1467-9868.2007.00633.x.

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, URL http://dx.doi.org/10.1198/016214506000000852. M3: doi:10.1198/016214506000000852.

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.

Furrer, Reinhard and Stephan R. Sain (2010), "spam: A sparse matrix r package with emphasis on mcmc methods for gaussian markov random fields." *Journal of Statistical Software*, 36, 1–25, URL http://www.jstatsoft.org/v36/i10.

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, URL http://dx.doi.org/10.1023/A:1009666805688. M3: 10.1023/A:1009666805688.

Jones, P. D., D.H. Lister, T.J. Osborn, C. Harpham, M. Salmon, and C.P. Morice (2012), "Hemispheric and large-scale land surface air temperature variations: an extensive revision and an update to 2010." *Journal of Geophysical Research*.

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.

DRAFT

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, 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, H., S. Martino, and N. Chopin (2009), "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations." *Journal of the royal statistical society: Series b (statistical methodology)*, 71, 319–392.

Rue, Håvard and Leonhard Held (2005), *Gaussian Markov random fields : theory and applications*, volume 104. Chapman & Hall/CRC, Boca Raton. Håvard Rue, Leonhard Held.; : ill. ; 24 cm; Includes bibliographical references (p. 237-253) and indexes.

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.

DRAFT

Willmott, C.J. and S.M. Robeson (1995), "Climatologically aided interpolation (cai) of terrestrial air temperature." *International Journal of Climatology*, 15, 221–229.

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.

Appendix

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.

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

DRAFT

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

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

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

$$\begin{aligned} I_l &> H(l), \qquad 1 \leq l \leq (L-1) \\ I_{l-1} &> H(l), \quad (L+1) \leq l \leq \infty \end{aligned} \tag{17}$$

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

$$\begin{aligned} I_{l-1} &< H(l), \quad 2 \leq l \leq L \\ I_l &< H(l), \qquad (L+1) \leq l \leq \infty \end{aligned} \tag{18}$$

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.

DRAFT

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

$$b \int_0^{e^{-b}} \frac{q^{(a/b)-1}dq}{(1 + sq)^c} \tag{19}$$

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

$$b \int_0^{se^{-b}} \frac{(p/s)^{(a/b)-1}dp}{s(1 + p)^c} = bs^{-a/b} \int_0^{se^{-b}} \frac{p^{(a/b)-1}dp}{(1 + p)^c} \tag{20}$$

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.

DRAFT