Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 49

Bayesian Smoothing of Irregularly-spaced Data Using Fourier Basis Functions

Christopher J. Paciorek*

*Harvard School of Public Health, paciorek@hsph.harvard.edu

Bayesian smoothing of irregularly-spaced data using Fourier basis functions

Christopher J. Paciorek
Department of Biostatistics, Harvard School of Public Health

Abstract

The spectral representation of Gaussian processes via the Fourier basis provides a computationally efficient specification of spatial surfaces and nonparametric regression functions in various statistical models. I describe the representation in detail and introduce the **spectralGP** R library for computations. Because of the large number of basis coefficients, some form of shrinkage is necessary; I focus on a natural Bayesian approach via a particular parameterized prior structure that approximates stationary Gaussian processes. I review several alternative parameterizations in the literature, suggest a simple modification suitable for exponential family data, and provide example code demonstrating MCMC sampling using the **spectralGP** library. I note that mixing can be slow in certain situations for reasons I describe, and provide some suggestions for MCMC techniques to improve mixing, also with example code. Note that I do not attempt to exhaustively compare parameterizations or MCMC techniques, but hope to provide a range of alternatives that may be useful for various models as well as some general recommendations grounded in experience.

Keywords: Bayesian statistics, Fourier basis, FFT, geostatistics, generalized linear mixed model, generalized additive model, Markov chain Monte Carlo, spatial statistics, spectral representation.

1. Introduction

Smoothing in the context of spatial modeling and nonparametric regression, often in an additive modelling scenario such as a generalized linear mixed model (GLMM) or generalized additive model (GAM), is a common technique in applied statistical work. A basic general model is

$$\begin{aligned} Y_i &\sim \mathcal{F}(f(\boldsymbol{x}_i, \boldsymbol{s}_i), \phi) \\ h(f(\boldsymbol{x}_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta} + g(\boldsymbol{s}_i; \boldsymbol{\theta}), \end{aligned} \tag{1}$$

where $Y_i,\ i = 1, \ldots n$, is the ith outcome, $\mathcal{F}$ is commonly an exponential family distribution,

ϕ is a dispersion parameter, $h(\cdot)$ is the link function, $\boldsymbol{x}_i$ is a vector of covariates for the ith observation, and $g(\boldsymbol{s}_i;\boldsymbol{\theta})$ is a smooth function, parameterized by $\boldsymbol{\theta}$, evaluated at the location or covariate value of the ith observation, $\boldsymbol{s}_i$, depending on whether the smooth function is in the spatial domain or covariate space. In this work I focus on on settings in which $g(\cdot;\boldsymbol{\theta})$ is a spatial surface, but results hold generally for one dimension and potentially for higher dimensions.

There have been two basic approaches to modelling the smooth function, $g(\cdot;\boldsymbol{\theta})$, each with a variety of parameterizations. One approach considers the function as deterministic within a generalized additive model (GAM) framework (Hastie and Tibshirani 1990; Wood 2006), e.g. using a thin plate spline or radial basis function representation with the function estimated via a penalized approach. The other takes a random effects, or equivalent stochastic process, approach in which the smooth function is treated stochastically, potentially via a Bayesian approach. Within this latter approach, one might consider a collection of correlated random effects, in which case (1) is a generalized linear mixed model (GLMM) (McCulloch and Searle 2001; Ruppert, Wand, and Carroll 2003). Alternatively, stochastic process representations such as kriging (Cressie 1993) or Bayesian versions of kriging (Banerjee, Carlin, and Gelfand 2004) usually take $g(\cdot;\boldsymbol{\theta})$ to be a Gaussian process. The random effects approach can also be considered as a stochastic process representation based on the implied covariance function of the process induced by the covariance structure of the random effects. Note that by considering a prior over functions or equivalently over the coefficients of basis functions, the additive model can be expressed in a Bayesian fashion, and there are connections between the thin plate spline and stochastic process approaches (Cressie 1993; Nychka 2000) and also between thin plate splines and mixed model representations (Ruppert *et al.* 2003). When interest lies in the linear coefficients and the smooth structure/spatial covariance is a nuisance, one approach to fitting such models is via estimating equations (e.g., (Heagerty and Lele 1998; Heagerty and Lumley 2000). My primary interest is in situations in which the smooth function is the outcome of interest, e.g., in predicting exposure to pollutants or spatial surfaces of climate variables, in which case such methods are not useful.

While models of the form (1) have a simple structure, unless the responses are Gaussian and the sample size is limited, fitting them can be difficult for computational reasons. If the response were Gaussian, there are many methods, both classical and Bayesian, for estimating $\boldsymbol{\beta}$, $g(\cdot;\boldsymbol{\theta})$, and $\boldsymbol{\theta}$. Most methods rely on integrating $g(\cdot;\boldsymbol{\theta})$ out of the model to produce a marginal likelihood or posterior, thereby moving the smooth structure out of the mean and into the variance, such that the observations have a simple, mean structure, (in (1) this is linear in a set of covariates), and a variance that is a sum of independent noise and spatially correlated structure. This leaves a small number of parameters to be estimated, often using numerical maximization or MCMC. However, for large n, computations can be burdensome as they involve matrix calculations of $O(n^3)$. In the non-Gaussian case and in hierarchical modeling in which the unknown process lies in the hierarchy of the model, this integration cannot be done analytically, which leads to substantial difficulty in fitting the model because of the high dimensional quantities that need to be estimated, as well as burdensome matrix calculations. One set of approaches to the problem focuses on the integral in the GLMM framework, using EM (McCulloch 1994, 1997; Booth and Hobert 1999) and numerical integration (Hedeker and Gibbons 1994; Gibbons and Hedeker 1997) to maximize the likelihood or approximating the integral to produce a penalized quasi-likelihood that can be maximized by iteratively weighted least squares (IWLS) (Ruppert *et al.* 2003).

Likelihood and covariance approximations can reduce the computational complexity of the matrix calculations (Stein, Chi, and Welty 2004; Furrer, Genton, and Nychka 2006), while the `gam()` function in the mgcv library in R uses the reduced rank thin plate spline approach of Wood (2004) fit by penalized IWLS. Rue and Tjelmeland (2002) exploit computationally efficient methods for fitting Markov random field (MRF) models by approximating stationary GPs using MRFs.

An alternative is to fit a Bayesian version of the model using a computationally efficient basis. The approach introduced by Wikle (2002) approximates a stationary GP structure for $g(\cdot;\boldsymbol{\theta})$ using a spectral representation to decompose the function in an orthogonal basis, in particular using Fourier basis functions and employing the FFT for fast computation (Wikle 2002; Royle and Wikle 2005; Paciorek and Ryan 2005). While the Fourier basis approach has some adherents and is one of the few efficient alternatives within the Bayesian paradigm, the intricacies and bookkeeping involved in working with the complex-valued basis coefficients make it hard to simply apply the methodology and replicate results. My goal here is to present the representation in detail (Section 2), and provide an R library, **spectralGP**, for working with the approach that handles the bookkeeping and sampling of coefficients for use within Markov chain Monte Carlo (MCMC) (Section 3). I describe several parameterizations for exponential family data (Section 4), and discuss detailed MCMC implementation and mixing issues that arise in fitting models, as well as general recommendations on parameterizations and sampling techniques (Section 5). I note that my experience shows slower mixing than one would desire; advances in this area are an open area for research.

2. Fourier basis representation

To simplify the notation I use $\boldsymbol{g_s}$ to denote the vector of values calculated by evaluating $g(\cdot)$ for each of the elements of $\boldsymbol{s}$ (e.g., for each observation location), namely $\boldsymbol{g_s} = (g(\boldsymbol{s}_1), \ldots, g(\boldsymbol{s}_n))^T$, suppressing the dependence on hyperparameters. Also, where necessary, I denote a set of unspecified parameters as $\boldsymbol{\theta}$. Proposal values are denoted with a *, e.g., $\boldsymbol{\theta}^*$, and vectors of augmented quantities with a tilde, e.g., $\tilde{\boldsymbol{Y}}$.

2.1. Basic process model

In many Bayesian models, the unknown functions, be they spatial or regression surfaces, are represented as a Gaussian process or by a basis function representation. Diggle, Tawn, and Moyeed (1998) formalized the idea of generalized geostatistical models, with a latent Gaussian spatial process, as the natural extension of kriging models to exponential family responses. They used Bayesian estimation, suggesting a Metropolis-Hastings implementation, with the spatial function sampled sequentially at each observation location at each MCMC iteration. However, as shown in their examples and discussed elsewhere (Christensen, Møller, and Waagepetersen 2000; Christensen and Waagepetersen 2002; Christensen, Roberts, and Sköld 2006), this implementation is slow to converge and mix, as well as being computationally inefficient because of the covariance matrix involved in calculating the prior for $\boldsymbol{g_s}$.

An alternative approach that avoids large matrix calculations is to express the unknown function in a basis, $\boldsymbol{g_s} = \boldsymbol{\Psi u}$, where $\boldsymbol{\Psi}$ contains the basis function values evaluated at the locations of interest, and estimate the basis coefficients, $\boldsymbol{u}$. These coefficients are taken to have a prior distribution; constraints on the function, such as degrees of smoothness, are imposed

through this prior distribution and the basis choice. When the coefficients are normally distributed, this representation can be viewed as a GP evaluated at a finite set of locations, with $\text{Cov}(\boldsymbol{g_s}) = \boldsymbol{\Psi}\text{Cov}(\boldsymbol{u})\boldsymbol{\Psi}^T$.

Isotropic GPs can be represented in an approximate fashion using their spectral representation as a Fourier basis expansion, which allows one to use the Fast Fourier Transform (FFT) to speed calculations. Here I describe the basic model in two-dimensional space, following Wikle (2002).

The key to the spectral approach is to approximate the function $g(\cdot)$ on a grid, $\boldsymbol{s}^{\#}$, of size $M = M_1 \times M_2$, where M_1 and M_2 are powers of two. Evaluated at the grid points, the vector of function values is represented as

$$\boldsymbol{g_{s^{\#}}} = \boldsymbol{\Psi}\boldsymbol{u}, \tag{2}$$

where $\boldsymbol{\Psi}$ is a matrix of orthogonal spectral basis functions, and $\boldsymbol{u}$ is a vector of complex-valued basis coefficients, $u_m = a_m + b_m i$, $m = 1, \ldots, M$. The spectral basis functions are complex exponential functions, i.e., sinusoidal functions of particular frequencies; constraints on the coefficients ensure that $\boldsymbol{g_{s^{\#}}}$ is real-valued and can be expressed equivalently as a sum of sine and cosine functions. To approximate mean zero stationary GPs, the basis coefficients have the prior distribution,

$$\begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma_\theta}) \tag{3}$$

where the diagonal (asymptotically; see (Shumway and Stoffer 2000, Section T3.12) covariance matrix of the basis coefficients, $\boldsymbol{\Sigma_\theta}$, parameterized by $\boldsymbol{\theta}$, can be expressed in closed form (for certain covariance functions) using the spectral density of the covariance function desired to parameterize the approximated GP.

To make this more explicit, consider the Matérn covariance popular in spatial statistics,

$$C(\tau; \rho, \nu) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\tau}{\rho}\right)^{\nu} \mathcal{K}_\nu \left(\frac{2\sqrt{\nu}\tau}{\rho}\right), \tag{4}$$

where τ is distance, σ^2 is the variance of the process, ρ is the range (correlation decay) parameter, and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind, whose order is the differentiability parameter, $\nu > 0$. This covariance function has the desirable property that sample functions of GPs parameterized with the covariance are $\lfloor \nu - 1 \rfloor$ times differentiable. As $\nu \to \infty$, the Matérn approaches the squared exponential form, with infinitely many sample path derivatives, while for $\nu = 0.5$, the Matérn takes the exponential form with no sample path derivatives.

The spectral density of this covariance, which is used to calculate the elements of $\boldsymbol{\Sigma_\theta}$, evaluated at spectral frequency, $\boldsymbol{\omega}$, is

$$\phi(\boldsymbol{\omega}; \rho, \nu) = \sigma^2 \frac{\Gamma(\nu + \frac{d}{2})(4\nu)^\nu}{\pi^{\frac{d}{2}}\Gamma(\nu)(\pi\rho)^{2\nu}} \cdot \left(\frac{4\nu}{(\pi\rho)^2} + \boldsymbol{\omega}^T\boldsymbol{\omega}\right)^{-\left(\nu+\frac{d}{2}\right)}, \tag{5}$$

where d is the dimension of the space (two in this case) and the parameters are as above. For an appropriate set of spectral frequencies, the diagonal elements of $\boldsymbol{\Sigma_\theta}$ are the values of $\phi(\cdot; \rho, \nu)$ at those frequencies, and the off-diagonals are zero.

To construct real-valued processes with an approximate GP distribution based on the complex-valued coefficients given above, some detailed bookkeeping and constraints are required. The

details that follow draw on Dudgeon and Mersereau (1984), Borgman, Taheri, and Hagan (1984), and Wikle (2002). The basis functions represented in the basis matrix, $\mathbf{\Psi}$, capture behavior at different frequencies, with the most important basis functions for function estimation being the low-frequency basis functions.

The first step in representing the function is to choose the grid size, M_d, in each dimension, $d = 1, \ldots, D$, to be a power of two. The M_d frequencies in the dth dimension are then $\omega^d \in \{0, 1, \ldots, \frac{M_d}{2}, -\frac{M_d}{2} + 1, \ldots, -1\}$, where the superscript represents the dimension. There is a complex exponential basis function for each distinct vector of frequencies, $\boldsymbol{\omega} = (\omega^1_{m_1}, \ldots, \omega^D_{m_D})$, $m_d \in \{0, \ldots, M_d - 1\}$, with corresponding complex-valued basis coefficient, $u_{m_1,\ldots,m_D}$.

First I show how to construct a random, mean zero, Gaussian process in one dimension from the M spectral coefficients, $u_m = a_m + b_m i$, $m = 0, \ldots, M-1$, and complex exponential basis functions, $\psi_m(s_j) = \exp(i\omega_m s_j)$, whose real and imaginary components have frequency ω_m. The circular domain of the process is $S^1 = (0, 2\pi)$ with the process evaluated only at the discrete grid points, $s_j \in \{0, 2\pi\frac{1}{M}, \ldots, 2\pi\frac{M-1}{M}\}$. To approximate real-valued processes, $u_0, \ldots, u_{M/2}$ are jointly independent, u_0 and $u_{M/2}$ are real-valued ($b_0 = b_{M/2} = 0$), and the remaining coefficients are determined, $u_{M/2+1} = \bar{u}_{M/2-1}, \ldots, u_{M-1} = \bar{u}_1$, where the overbar is the complex conjugate operation. This determinism causes the imaginary components of the basis functions to cancel, leaving a real-valued process,

$$\begin{aligned} g(s_j) &= \sum_{m=0}^{M-1} \psi_m(s_j) u_m = \sum_{m=0}^{\frac{M}{2}} \exp(i\omega_m s_j)(a_m + b_m i) + \sum_{m=\frac{M}{2}+1}^{M-1} \exp(i\omega_m s_j)(a_{M-m} - b_{M-m} i) \\ &= a_0 + 2\sum_{m=1}^{\frac{M}{2}-1} (a_m \cos(\omega_m s_j) - b_m \sin(\omega_m s_j)) + a_{M/2}\cos(\omega_{\frac{M}{2}} s_j). \end{aligned} \tag{6}$$

Hence for a grid of M values, the process is approximated as a linear combination of M spectral basis functions corresponding to M real-valued sinusoidal basis functions, including the constant function ($\omega_0 = 0$). To approximate mean zero Gaussian processes with a particular stationary covariance function, the coefficients have independent, mean zero Gaussian prior distributions with the spectral density for the covariance function, $\phi(\cdot; \boldsymbol{\theta})$, e.g., (5), determining the prior variances of the coefficients:

$$\mathrm{V}(u_m) = \phi(\omega_m; \boldsymbol{\theta}) \Rightarrow \{\mathrm{V}(a_0) = \phi(\omega_0; \boldsymbol{\theta});\ \mathrm{V}(a_{M/2}) = \phi(\omega_{M/2}; \boldsymbol{\theta});\ \mathrm{V}(a_m) = \mathrm{V}(b_m) = \frac{1}{2}\phi(\omega_m; \boldsymbol{\theta}),\ \text{o.w.}\} \tag{7}$$

.

The setup is similar in two dimensions, with a matrix of $M = M_1 M_2$ coefficients, $((u_{m_1,m_2}))$, $m_d \in \{0, \ldots, M_d - 1\}$, and corresponding frequency pairs, $(\omega^1_{m_1}, \omega^2_{m_2})$, and a toroidal domain. As seen in Table 1, many coefficients are again deterministically given by other coefficients to ensure that the process is a linear combination of real-valued sinusoidal basis functions of varying frequencies and orientations in $\Re^2$. The real and imaginary components of each coefficient, $u_{m_1,m_2} = a_{m_1,m_2} + b_{m_1,m_2} i$, are again independent. For $(m_1, m_2) \in \{(0,0), (\frac{M_1}{2}, 0), (0, \frac{M_2}{2}), (\frac{M_1}{2}, \frac{M_2}{2})\}$, $b_{m_1,m_2} = 0$ and $\mathrm{V}(a_{m_1,m_2}) = \phi(\omega^1_{m_1}, \omega^2_{m_2}; \boldsymbol{\theta})$, while for the remaining complex-valued coefficients, $\mathrm{V}(a_{m_1,m_2}) = \mathrm{V}(b_{m_1,m_2}) = \frac{1}{2}\phi(\omega^1_{m_1}, \omega^2_{m_2}; \boldsymbol{\theta})$.

2.2. Periodicity and Euclidean domains

Table 1: Visual display of the spectral coefficients for a two-dimensional process. The frequencies in each dimension are indicated by the row and column labels, with $h_d = \frac{M_d}{2}$ for $d = 1, 2$. The * operation indicates that one takes the matrix or vector, flips it in both the horizontal and vertical directions (just the horizontal or vertical in the case of a vector) and then takes the complex conjugates of the elements.

	0	1	...	h_2	$-h_2+1$	...	-1
0	$u_{0,0}$	$\boldsymbol{u_{0,\cdot}}$		u_{0,h_2}	$\boldsymbol{u^*_{0,\cdot}}$		
1	$\boldsymbol{u_{\cdot,0}}$	$\boldsymbol{u_A}$			$\boldsymbol{u^*_B}$		
$\vdots$							
h_1	$u_{h_1,0}$			u_{h_1,h_2}			
$-h_1+1$	$\boldsymbol{u^*_{\cdot,0}}$	$\boldsymbol{u_B}$		$\boldsymbol{u^*_A}$			
$\vdots$							
-1							

The construction produces periodic functions; in one dimension the process lives on a circular domain, while in two dimensions the process lives on a torus. To work in Euclidean space, we need to map the space onto the periodic domain. The goal is to use the representation on Euclidean domains without inducing anomalous correlations between locations that are far apart in Euclidean space, but close in the periodic domain. To do this, I suggest mapping the Euclidean domain of interest onto a portion of the periodic domain and ignoring the remainder of the periodic domain as follows.

In one dimension, $g(0) = g(2\pi)$, so the correlation function, $C(\tau) = C(g(0), g(\tau))$ of the process at distances $\tau \in (\pi, 2\pi)$ is the mirror image of the correlation function for $\tau \in (\pi, 0)$ with $\text{Cor}(g(0), g(2\pi)) = 1$ (Figure 1). I avoid artifacts from this periodicity by mapping the interval $(0, 2\pi)$ to $(0, 2)$ and mapping the original domain of the observations to $(0, 1)$, thereby computing but not using the process values on $(1, 2)$. Note that the use of $\pi\rho$ rather than ρ in (5) allows us to interpret ρ on the $(0, 1)$ rather than $(0, \pi)$ scale. The modelled process on $(0, 1)$ is a piecewise constant function on an equally-spaced grid of size $M/2 + 1$. This setup ensures that the correlation structure of the approximating process is close to the correlation structure of a GP with the desired stationary correlation function (Figure 1).

As the higher-dimension analogue of the one-dimensional case, I estimate the process on $(0, 1)^D$. To do so, I map the periodic domain $(0, 2\pi)^D$ to $(0, 2)^D$ and then map the observation domain onto the $(0, 1)^D$ portion (maintaining the scale ratio in the different dimensions, unless desired otherwise), thereby calculating but ignoring the process values outside this region. Note that if the original domain is far from square, I unnecessarily estimate the process in large areas of no interest, resulting in some loss of computational efficiency. Wikle (2002) and Royle and Wikle (2005) do not mention the issue of periodicity; it appears that they use a somewhat larger grid than necessary to include all observations (sometimes called padding) and rely on the correlation decaying sufficiently fast that anomalously high correlations between distant observations induced by the periodicity do not occur. For example, notice in Figure 1 that

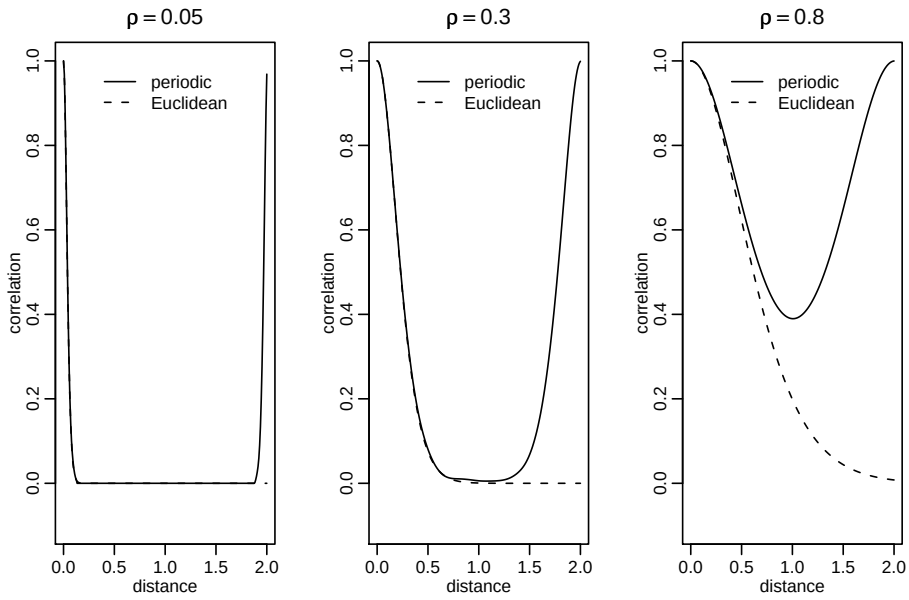

Figure 1: Comparison of correlation structure of GPs based on the standard Matérn covariance on the Euclidean domain, $(0, 2)$ (dashed lines), and approximate GPs based on the Fourier basis for the periodic domain, $(0, 2\pi)$, mapped to $(0, 2)$ (solid lines), for three values of ρ. Note that the Euclidean correlation for $\rho = 0.05$ falls off to zero as rapidly as the periodic case and remains at zero for all remaining distances.

for $\rho = 0.05$ the correlation does not start to rise again until the distance is almost 2.0, so a small amount of padding would suffice).

2.3. Computations and statistical modeling of observations

The process at the observation locations is calculated through an incidence matrix, $\boldsymbol{K}$, which maps each observation location to the nearest grid location on the subset of the periodic domain,

$$\boldsymbol{g_s} = \boldsymbol{K}\boldsymbol{g_{s\#}} = \boldsymbol{K}\boldsymbol{\Psi}\boldsymbol{u}. \tag{8}$$

For a fine grid, the error induced in associating observations with grid locations should be negligible and the piecewise constant representation of the surface tolerable. The computational efficiency comes in the fact that the matrix $\boldsymbol{\Psi}$, which is $M \times M$, need never be explicitly formed, and the operation $\boldsymbol{\Psi}\boldsymbol{u}$ is the inverse FFT, and so can be done very efficiently ($O(M \log_2(M))$). In addition, evaluating the prior for $\boldsymbol{u}$ is fast because the coefficients are independent a priori. This stands in contrast to the standard MCMC setup for GP models, in which the prior on $\boldsymbol{g_s}$ involves an $n \times n$ matrix and therefore $O(n^3)$ operations. Of course the gridding could be done without the Fourier basis approach, but this would only reduce the computations to $O\left((M/2^D)^3\right)$ (the division by 2^D occurs because the padding would not be required). Note that with the gridded approach, the number of observations affects the calculations only through the likelihood, which scales as $O(n)$, because the observations are independent conditional on $\boldsymbol{g_s}$. The complexity of the underlying surface determines the computational efficiency by defining how large M should be; simple surfaces can be estimated

very efficiently even if n is large.

3. R spectralGP library

The **spectralGP** library for R provides an object-oriented representation of the Fourier basis approach to computation with GPs. The library allows one to initialize GP objects, simulate random processes, plot and output the process values, and sample the Fourier basis coefficients using the MCMC sampling schemes described in Section 4. The key functions in **spectralGP** are a constructor function, `gp()`, and a number of S3 methods: two MCMC sampling methods for the coefficients, `Gibbs.sample.coeff.gp()` and `propose.coeff.gp()`; a method for simulating GPs, `simulate.gp()`; a method for changing the GP hyperparameter values, `change.param.gp()`; calculation of the logdensity of the coefficients, `logdensity.gp()`; and prediction and plotting methods. Internal methods not meant for users include calculation of the coefficient variances, `calc.variances.gp()` and updating of the process values after changes in the coefficients, `updateprocess.gp()`. Auxiliary functions allow for copying, `copy.gp()`; determining the basic grid used, `getgrid.gp()`; and extracting the object element names, `names.gp()`. Several functions deal with conversion between coordinate systems: a basic lon/lat to Euclidean x/y projection, `lonlat2xy()`; mapping a Euclidean domain to $(0,1)^D$, `xy2unit()`; and mapping locations in the domain to the closest grid points, `new.mapping()`. Several auxiliary functions are borrowed from the fields library, namely `rdist.earth()` as well as `image.plot()` and its auxiliary functions, `image.plot.info()` and `image.plot.plt()`.

Since a GP object will be used repeatedly in a Bayesian MCMC approach, I choose to use a pass-by-reference scheme in the coding, using R environments to mimic object-orientation in traditional languages (E.A. Houseman, pers. comm.). In this way, one can operate on the GP objects and change internal elements without having to pass the entire object back from the function. This is possible because unlike other R objects, environments are not copied when passed to functions. Each instantiation of a gp object is an environment, initialized with a call to `new.env()` and assigned the class "gp". The elements of the object, e.g., `myFun` here, are local variables within the environment, accessed via list-like syntax, e.g., `myFun$value`. S3 methods are used to operate on the gp objects, with the difference from standard R that global changes can be made to the elements of an object within the method by virtue of those elements residing within an environment. For example, the call, `simulate.gp(myFun)`, samples new coefficients and updates the process based on those coefficients without having to pass `myFun` back to the calling environment, yet the changes to `myFun` are effective in the calling environment. Also note that care must be taken when assigning gp objects because environments are not copied when used in assignments; I have created an explicit `copy.gp()` function to make a new copy of a gp object; assignment merely creates an additional name (i.e., pointer) referencing the existing object. I chose not to use S4 methods because my understanding is that their implementation is still somewhat slow and **spectralGP** works with large objects and substantial computation. I have used native R code for the entire library both for simplicity and because the essential computations within the library are already compiled code, namely the functions, `fft()`, `rnorm()`, and `dnorm()`.

Some additional intricacies included in the library are mentioned parenthetically in the next two sections. Note that in the library, the process is scaled by $1/\sqrt{M_1 M_2}$, as described in Section 4.1, relative to the exact process values (6).

Also note that C. Wikle has released Matlab code for the Fourier basis computations on his website (http://www.stat.missouri.edu/~wikle).

4. Basic MCMC sampling schemes for coefficients

In this section, I describe several parameterizations for simple Bayesian exponential family models with associated MCMC sampling schemes for the Fourier basis coefficients. Bayesian estimation of unknown processes represented as GPs relies on shrinkage to estimate the large number of coefficients; coefficients of low-frequency basis functions are strongly informed by the data while those of high-frequency basis functions are shrunk strongly toward their prior distributions. These following basic parameterizations can also be used with relatively straightforward modifications in more complicated hierarchical models, although the added complexity may make the simple blocked sampling scheme the most feasible approach in that case. Note that for simplicity I consider a scalar mean parameter, μ, but this could be replaced by a regression term, e.g., $\boldsymbol{X_i\beta}$, or other additive components.

4.1. Data augmentation Gibbs sampling for normal data

For Gaussian data with mean function based on the latent process, $g(\cdot)$, a missing data scheme allows for Gibbs sampling of the coefficients. This is a simplification of the Gibbs sampling scheme of Wikle (2002).

Take the data model to be

$$\boldsymbol{Y} \sim \mathcal{N}_n(\mu\mathbf{1} + \gamma\boldsymbol{K\Psi u}, \eta^2\boldsymbol{I}), \tag{9}$$

where μ is the process mean and γ is the process standard deviation with σ^2 in (4-5) set to one. Since the prior for the coefficients is normal, we have conjugacy, and the conditional distribution for $\boldsymbol{u}$ is

$$\begin{aligned} \boldsymbol{u}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \mathcal{N}_M(\boldsymbol{V}\frac{\gamma}{\eta^2}(\boldsymbol{K\Psi})^T(\boldsymbol{Y}-\mu\mathbf{1}), \boldsymbol{V}) \\ \boldsymbol{V} &= (\frac{\gamma^2}{\eta^2}\boldsymbol{\Psi}^T\boldsymbol{K}^T\boldsymbol{K\Psi} + \boldsymbol{\Sigma}^{-1})^{-1}. \end{aligned} \tag{10}$$

The sample of $\boldsymbol{u}$ represents precision-weighted shrinkage of the data-driven estimates of the coefficients towards the prior mean of zero.

However, this sampling scheme requires calculation of $\boldsymbol{\Psi}^T\boldsymbol{K}^T\boldsymbol{K\Psi}$, which is not feasible for large number of grid points; note that if $\boldsymbol{K}$ were the identity, since $\boldsymbol{\Psi}$ is an orthogonal matrix, this simplifies to

$$\boldsymbol{V} = (\frac{\gamma^2}{\eta^2}\boldsymbol{I} + \boldsymbol{\Sigma}^{-1})^{-1}, \tag{11}$$

which because $\boldsymbol{\Sigma}$ is diagonal, is easy to calculate. Assuming no more than one observation per grid cell, $\boldsymbol{K} = \boldsymbol{I}$ can be achieved using a missing data scheme by introducing latent pseudo-observations for all grid cells without any associated data, including grid cells in which no data can possibly fall, as described in Section 2.2. Collecting these pseudo observations into a vector, $\tilde{\boldsymbol{Y}}$, they can be sampled within the MCMC using a Gibbs step as

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}_{M-n}(\mu\mathbf{1} + \gamma\tilde{\boldsymbol{K}}\boldsymbol{\Psi u}, \eta^2\boldsymbol{I}), \tag{12}$$

where the $\tilde{\boldsymbol{K}}$ matrix picks out grid cells with no associated data. With this augmentation, $\boldsymbol{Y}$ in (9-10) is a vector of values on the full grid, $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \tilde{\boldsymbol{Y}})$, combining actual observations with pseudo observations, and $\boldsymbol{K} = \boldsymbol{I}$.

Note that this straightforward expression conceals some details required in working with the complex-valued coefficients. In calculating $(\gamma^2/\eta^2 + \boldsymbol{\Sigma}_{ii}^{-1})$, one needs to multiple γ^2/η^2 by one-half for all the elements corresponding to complex-valued coefficients to ensure that the scaling is correct as described in Section 2. Also, the operation $\boldsymbol{\Psi}^T(\boldsymbol{Y} - \mu\boldsymbol{1})$ is the FFT and the correct scaling needs to occur so the result is on the scale of the coefficients. In R, I specify the coefficient variances as $M_1 M_2 \phi(\boldsymbol{\omega}; \boldsymbol{\theta})$ and update the process as $\boldsymbol{\Psi u}/\sqrt{M_1 M_2}$. If I then divide $\boldsymbol{\Psi}^T(\boldsymbol{Y} - \mu\boldsymbol{1})$ by $\sqrt{M_1 M_2}$ when sampling the coefficients (10-11), the desired approximate covariance structure for the process is preserved, namely $\boldsymbol{\Psi u} \dot{\sim} \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{C})$, where the matrix, $\boldsymbol{C}$ is defined by $\boldsymbol{C}_{ij} = C(d(\boldsymbol{s_i}, \boldsymbol{s_j}))$ and $C(\cdot)$ is the covariance function whose spectral density defines $\phi(\cdot; \boldsymbol{\theta})$, e.g., (5). The exact algorithm is given in the `Gibbs.sample.coeff.gp()` function in the **spectralGP** library.

In the appendix I provide template code for fitting this parameterization, denoted as Code A.

If there is more than one observation per grid cell, some possible solutions are to use a finer grid or to take $Y(s_j) = \bar{Y}_j$, namely the average of the observations in the grid cell. Ideally, one would set $\eta_j^2 = \eta^2/n_j$, but this would require calculation of $\boldsymbol{\Psi}^T \boldsymbol{\eta}^{-1} \boldsymbol{\Psi}$, where $\boldsymbol{\eta} = \text{diag}((\eta_1^2, \ldots, \eta_M^2))$, which is computationally infeasible. Instead, I suggest using constant η^2 so long as there are relatively few locations with multiple observations per grid cell. One could also use more extensive data augmentation to supplement the existing observations such that there are n_j pseudo plus true observations per grid cell, with n_j equal to the maximum number of true observations in a cell over all of the grid cells.

Wikle (2002) recommends the uncentered (sensu Gelfand, Sahu, and Carlin (1996)) parameterization for the process variance (9), with γ allowed to vary and $\sigma^2 \equiv 1$ in defining the covariance of the coefficients (5). He notes that moving the parameter closer to the data improves mixing and helps avoid dependence with ρ. Note that I follow this approach in some cases, while in others, I allow σ^2 to vary and fix $\gamma \equiv 1$. In the **spectralGP** library, a value of σ^2 not equal to one is specified with the `variance.param` argument to `gp()` and the `new.variance.param` argument to `change.param.gp()`.

For non-normal data from the exponential family, $Y_i \sim \mathcal{F}(h^{-1}(f_i))$, where $f_i = \mu + \gamma \boldsymbol{K_i \Psi u}$, one might use a Metropolis-Hastings-adjusted version of this Gibbs sampling scheme, again with data augmentation. The proposal makes use of the linearized observations,

$$y_i' \equiv f_i + \frac{\partial h(f_i)}{\partial f_i}(y_i - h^{-1}(f_i)). \tag{13}$$

Ideally, one would use working variances,

$$\eta_i'^2 \equiv \left(\frac{\partial h(f_i)}{\partial f_i}\right)^2 \text{Var}(Y_i), \tag{14}$$

used in fitting GLMs, but a diagonal matrix with $\eta_i'^2$ elements would prevent cancellation in (10). Instead, consider η^2 to be a fixed constant that allows one to tune the proposals for better mixing with reasonable values informed by the working variance expression, perhaps a rough average of the working variances (14) based on an initial fit of the model. Then use a

Metropolis-Hastings sample with the following proposal distribution:

$$\begin{aligned} \boldsymbol{u}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{V}\frac{\gamma}{\eta^2}(K\boldsymbol{\Psi})^T(\boldsymbol{Y'}-\mu\boldsymbol{1}),\boldsymbol{V}) \\ \boldsymbol{V} &= (\frac{\gamma^2}{\eta^2}\boldsymbol{I}+\boldsymbol{\Sigma}^{-1})^{-1}. \end{aligned} \tag{15}$$

I do not pursue this approach for non-normal data further, as I had little success in tuning the proposals to achieve reasonable acceptance, but further research in this area may be worthwhile.

One approach to speeding mixing in the case of normal data is to jointly sample η^2 and $\tilde{\boldsymbol{Y}}$ by first proposing η^{2*} and then, within the same proposal, sampling $\tilde{\boldsymbol{Y}}|\eta^{2*},\ldots$ via a Gibbs sample, conditional on the proposed value, η^{2*}. Because this joint sample is not from the joint conditional of $(\tilde{\boldsymbol{Y}},\eta^2|\ldots)$, we need to use Metropolis-Hastings in determining acceptance based on the ratio of the prior for η^2 and likelihood, $\pi(\eta^{2*})L(\boldsymbol{Y}_{obs}|\eta^{2*},\boldsymbol{\theta})/(\pi(\eta^2)L(\boldsymbol{Y}_{obs}|\eta^2,\boldsymbol{\theta}))$, where $\boldsymbol{Y}_{obs}$ is the actual data, which indicates that acceptance does not depend on the value for the augmented observations, $\tilde{\boldsymbol{Y}}$. So one can propose η^2, decide on acceptance based on the likelihood of the true observations, and then, if accepted, do a Gibbs sample for $\tilde{\boldsymbol{Y}}$ (12). We have effectively integrated $\tilde{\boldsymbol{Y}}$ out of the joint conditional density, $\pi(\eta^2,\tilde{\boldsymbol{Y}}|\boldsymbol{Y}_{obs},\boldsymbol{\theta})$, thereby sampling η^2 without dependence on $\tilde{\boldsymbol{Y}}$ (Rue and Held 2005, pp. 141-143). In the iterations, one may also wish to do a separate Gibbs sample for $\tilde{\boldsymbol{Y}}$ alone, apart from the joint sample with η^2. Template code A in the appendix also includes modifications for this sampling approach.

4.2. Latent layer Gibbs sampling for exponential family data

4.2.1. Parameterizing with two latent layers (the Wikle parameterization)

For non-normal data, rather than losing the Gibbs sampling structure for the coefficients, Wikle (2002) and Royle and Wikle (2005) embed the spectral basis representation in a hierarchical model with additional latent processes and associated variance components. This approach allows one to do Gibbs sampling in various generalized models in which exponential family outcomes are related to a latent spatial process in the mean structure (1).

To take a concrete example, the model for Poisson data is

$$\begin{aligned} Y_i &\sim \mathcal{P}(\exp(\lambda_i)) \\ \lambda_i &\sim \mathcal{N}(\mu+\gamma\boldsymbol{K_i}\boldsymbol{z},\eta^2) \\ \boldsymbol{z} &\sim \mathcal{N}_M(\boldsymbol{\Psi u},\sigma_z^2\boldsymbol{I}), \end{aligned} \tag{16}$$

where $\boldsymbol{\Psi u}$ is the Fourier basis representation with the prior structure (7). One can easily modify the likelihood and link for other exponential family distributions. The model introduces two variance components, η^2 and σ_z^2, and an additional latent process, $\boldsymbol{z}$, defined for each of the grid cells, including those in which no data can fall, as discussed in Section 2.2. Note that the variance components account for overdispersion. In Section 5.1.2, I discuss issues that arise when the data are not overdispersed.

Wikle (2002) suggests a Metropolis-Hastings proposal for $\boldsymbol{\lambda}$, with conjugate normal Gibbs sampling for $\boldsymbol{z}$ and $\boldsymbol{u}$:

$$\boldsymbol{z}|\boldsymbol{\lambda},\boldsymbol{u},\boldsymbol{\theta} \sim \mathcal{N}_M\left(\boldsymbol{V}_z\left(\frac{\gamma}{\eta^2}\boldsymbol{K}^T(\boldsymbol{\lambda}-\mu\boldsymbol{1})+\sigma_z^{-2}\boldsymbol{\Psi u}\right),\boldsymbol{V_z}\right)$$

$$\begin{aligned} \boldsymbol{V_z} &= (\frac{\gamma^2}{\eta^2}\boldsymbol{K}^T\boldsymbol{K} + \sigma_z^{-2}\boldsymbol{I})^{-1} \\ \boldsymbol{u}|\boldsymbol{z},\boldsymbol{\theta} &\sim \mathcal{N}_M(\boldsymbol{V_u}\frac{\boldsymbol{\Psi}^T\boldsymbol{z}}{\sigma_z^2}, \boldsymbol{V}_u) \\ \boldsymbol{V_u} &= (\sigma_z^{-2}\boldsymbol{I} + \boldsymbol{\Sigma}^{-1})^{-1} \end{aligned} \tag{17}$$

Similar adjustments as in the previous section are needed in the Gibbs sampling for $\boldsymbol{u}$ to account for the complex-valued coefficients and to scale the proposal correctly. In the **spectralGP** library, $\boldsymbol{u}$ is sampled using the `Gibbs.sample.coeff.gp()` function, with $\boldsymbol{z}$ taking the place of the vector of 'observations'. Template code is given in the appendix as Code B.

Note that sampling can require long chain lengths; Royle and Wikle (2005) used eight chains of length 520,000, retaining every 50th iteration, which suggest slow mixing of the sort I have experienced as well.

4.2.2. A simplified parameterization with a single latent layer (modified Wikle parameterization)

I propose a modification of the model above to eliminate one of the latent layers, thereby moving the coefficients closer to the data in the hierarchy and eliminating $\boldsymbol{z}$ and σ_z^2, which can be difficult to interpret and may not informed by the data (see Section 5.1.2). The simplified model for Poisson data is

$$\begin{aligned} Y_i &\sim \mathcal{P}(\exp(\boldsymbol{K_i}\boldsymbol{\lambda})) \\ \boldsymbol{\lambda} &\sim \mathcal{N}_M(\mu\boldsymbol{1} + \gamma\boldsymbol{\Psi}\boldsymbol{u}, \eta^2\boldsymbol{I}) \end{aligned} \tag{18}$$

where the ith row of $\boldsymbol{K}$ maps the observation to the grid cell in which it falls. One can easily modify the likelihood and link for other exponential family distributions.

One can use Gibbs sampling for the values of $\boldsymbol{\lambda}$ corresponding to the J grid cells with no observations, denoted, $\tilde{\boldsymbol{\lambda}}$, and for $\boldsymbol{u}$:

$$\begin{aligned} \tilde{\boldsymbol{\lambda}}|\boldsymbol{Y},\boldsymbol{u},\boldsymbol{\theta} &\sim \mathcal{N}_J(\mu\boldsymbol{1} + \gamma\tilde{\boldsymbol{K}}\boldsymbol{\Psi}\boldsymbol{u}, \eta^2\boldsymbol{I}) \\ \boldsymbol{u}|\boldsymbol{\lambda},\boldsymbol{\theta} &\sim \mathcal{N}_M(\boldsymbol{V}_u\frac{\gamma}{\eta^2}\boldsymbol{\Psi}^T(\boldsymbol{\lambda} - \mu\boldsymbol{1}), \boldsymbol{V_u}) \\ \boldsymbol{V_u} &= (\frac{\gamma^2}{\eta^2}\boldsymbol{I} + \boldsymbol{\Sigma}^{-1})^{-1}. \end{aligned} \tag{19}$$

For the elements of $\boldsymbol{\lambda}$ corresponding to grid cells in which observations fall, $\boldsymbol{\lambda}_{obs}$, I suggest Metropolis proposals, done individually for each individual element, but computed in an efficient vectorized fashion in R. Some intuition for how the information from the data diffuses to the level of the basis coefficients is that the latent layer, $\boldsymbol{\lambda}$, allows for some fluidity between the process values and the data: individual sampling of $\boldsymbol{\lambda}_{obs}$ for individual grid cells allows the latent layer to accomodate the data based on adjustments to $\boldsymbol{\lambda}_{obs}$ at individual locations, while the Gibbs sample of $\boldsymbol{u}$ translates these adjustments to the coefficients. A single joint sample for the elements of $\boldsymbol{\lambda}_{obs}$ would likely have slower mixing as it would be trying to sample many grid locations at once, with a single acceptance decision, thereby slowing local adjustments. Template code is given in the appendix as Code C.

In similar fashion to joint sampling of $(\eta^2, \tilde{\boldsymbol{Y}})$ in Section 4.1, with the parameterization above, one can improve mixing by jointly sampling η^2 and $\tilde{\boldsymbol{\lambda}}$. First propose η^{2*} and then, within the same proposal, propose $\tilde{\boldsymbol{\lambda}}^*|\eta^{2*}$ from its full conditional based on the proposed value, η^{2*}. Because this joint sample is not from the joint conditional of $(\eta^2, \tilde{\boldsymbol{\lambda}})$, we need a Metropolis-Hastings acceptance decision based on the ratio of the prior for η^2 and likelihood, $\pi(\eta^{2*})L(\boldsymbol{Y}|\eta^{2*}, \boldsymbol{\lambda}_{obs}, ...)/(\pi(\eta^2)L(\boldsymbol{Y}|\eta^2, \boldsymbol{\lambda}_{obs}, \ldots))$, with acceptance not depending on the value for the augmented locations, $\tilde{\boldsymbol{\lambda}}$, thereby effectively integrating $\tilde{\boldsymbol{\lambda}}$ out of the joint conditional density, $\pi(\eta^2, \tilde{\boldsymbol{\lambda}}|\boldsymbol{\lambda}_{obs}, \boldsymbol{Y}, \boldsymbol{\theta})$. So in practice one can propose η^{2*}, decide on acceptance, and then, if accepted, do a Gibbs sample for $\tilde{\boldsymbol{\lambda}}$ (19). In the iterations, one may also wish to do a separate Gibbs sample for $\tilde{\boldsymbol{\lambda}}$ alone. Template code C in the appendix also includes modifications for joint sampling of $(\eta^2, \tilde{\boldsymbol{\lambda}})$.

4.3. Blocked Metropolis sampling for exponential family data (simple parameterization)

An alternative to Gibbs sampling that avoids the use of the additional hierarchical layers and variance components (Section 4.2) is a simple model with straightforward Metropolis sampling for the coefficients described in Paciorek and Ryan (2005). This approach has the advantage of tying the coefficients to the data by involving the coefficients directly in the likelihood, without the intervening levels in Section 4.2. For data that are not overdispersed, the simple model avoids introducing the overdispersion parameter(s), η^2 (and σ_z^2).

The basic approach is to specify the obvious parameterization in which the data are directly dependent on the latent spatial surface, which for Poisson data is

$$Y_i \sim \mathcal{P}(\exp(\mu + \gamma \boldsymbol{K_i} \boldsymbol{\Psi} \boldsymbol{u})). \tag{20}$$

I suggest sampling the coefficients in blocks according to coefficients whose corresponding frequencies have similar magnitudes (Paciorek and Ryan 2005). I use smaller blocks for the low-frequency coefficients, thereby allowing these critical coefficients to move more quickly. The high-frequency coefficients have little effect on the function and are proposed in large blocks. The first block is the scalar, $u_{0,0}$, corresponding to the frequency pair, $(\omega_0^1, \omega_0^2) = (0, 0)$ (but note that in Section 5.2 I suggest not sampling this coefficient). The remaining blocks are specified so that the block size increases as the frequencies increase. For example, the next block might include the coefficients whose largest magnitude frequencies are at most one, i.e., u_{m_1,m_2} s.t. $\max\{|\omega_{m_1}^1|, |\omega_{m_2}^2|\} \leq 1$, but excluding the previous block, giving the block of coefficients, $\{u_{0,1}, u_{1,0}, u_{1,1}, u_{M_1-1,1}\}$. Recall that there are additional coefficients whose largest magnitude frequencies are at most one, e.g., $u_{M_1-1,0}$and u_{M_1-1,M_2-1}, but these are complex conjugates of the sampled coefficients. The next block might be the coefficients whose largest magnitude frequencies are at most two, i.e., u_{m_1,m_2} s.t. $\max\{|\omega_{m_1}^1|, |\omega_{m_2}^2|\} \leq 2$, but excluding the previous block elements. The real and imaginary components of the coefficients in each block are proposed jointly, Metropolis-style, from a multivariate normal distribution with independent elements whose means are the current values. Since the coefficients have widely-varying scales, I take the proposal variance for each coefficient to be the product of a tuneable multiplier (one for each block) and the prior variance of the coefficient, which puts the proposal on the proper scale. In the `add.blocks.gp()` function in **spectralGP** library, the default blocks are set by grouping coefficients based on the frequency thresholds, $0, 1, 2, 4, \ldots, 2^Q$, where $Q = \log_2(\max(M_1, M_2)) - 1$. The coefficients can be proposed in **spec-**

tralGP using the `propose.coeff.gp()` function. Template code for fitting by this approach is given in the appendix as Code D.

4.4. Hyperparameter priors

Assuming exponential family data and the Matérn covariance (4-5), the hyperparameters in the data augmentation and block sampling schemes are $\boldsymbol{\theta} = (\mu, \gamma, \rho, \nu, \eta^2)$, with η^2 not present for some likelihoods. Wikle (2002); Royle and Wikle (2005) include the additional variance components, σ_z^2 and η^2.

Royle and Wikle (2005) use inverse gamma priors for η^2 and σ_z^2 with a diffuse normal prior for μ. For ρ they use the reference prior of Berger, De Oliveira, and Sansó (2001) for the Gaussian likelihood case to avoid the use of an improper diffuse prior that could lead to an improper posterior.

Paciorek and Ryan (2005) specify independent, proper, but non-informative priors for the elements of $\boldsymbol{\theta}$, except for ν, which cannot be estimated for a grid-level process (even for the continuous case, this is difficult to estimate without some observations very close together) and which is fixed in advance ($\nu = 4$ gives smooth processes with a small number of derivatives). Gelman (2006) suggests truncated uniform and folded non-central t distributions on the standard deviation scale for variance components and argues against $\mathcal{IG}(\epsilon, \epsilon)$ priors as these have a sharp peak at small values that can strongly affect inference. Berger *et al.* (2001) argue for reference priors and against proper but non-informative priors, including truncated distributions, in part because they are concerned about the posterior concentrating at extreme values or on the truncation limit. In the setting here, I believe that the exact form of the priors is not critical, except that it is desirable to keep the parameters in a finite interval to prevent them from wandering in extreme parts of the space in which the likelihood is flat. In cases with sufficient data, the prior should play little role in estimation and prediction, whereas the situations that concern Berger *et al.* (2001) with regard to truncation and vague proper priors arise when the data provide little information, in which case their concern about the posterior concentrating on the truncation limit seems little different than having it constrained by the reference prior. I discuss identifiability and priors for σ_z^2 and η^2 in more detail in Section 5.1.

5. MCMC sampling considerations

In Sections 5.1-5.4, I describe some factors that impede mixing and some modifications to the basic sampling schemes discussed above that can help to improve mixing. In Section 5.5 I discuss some general sampling issues and make broad recommendations. Note that for any particular application, these recommendations may not provide the best mixing and consideration of alternatives discussed in this paper may improve matters.

I explored sampling effectiveness using a few simulated datasets, meant to provide a range of function complexity and data intensity. All have Poisson data with the sampling locations sampled uniformly in $(0,1)^2$: Data1 has 225 observations while Data2 has 1000 observations, both with the mean function, $f(x_1, x_2) = 1.9 \cdot (1.35 + \exp(x_1)\sin(13 \cdot (x_1 - 0.6)^2) \cdot \exp(-x_2)\sin(7x_2))$, used by Hwang, Lay, Maechler, Martin, and Schimert (1994) (Fig. 2), a fairly simple function that when fit with a GP has $\hat{\rho} \approx 0.3$. Data3, Data4, and Data5 use the same mean function; a GP with $\rho = 0.05$, $\mu = 0$, $\gamma = \sigma = 1$ (Fig. 2); and 400, 800, and 2500 observations, respectively.

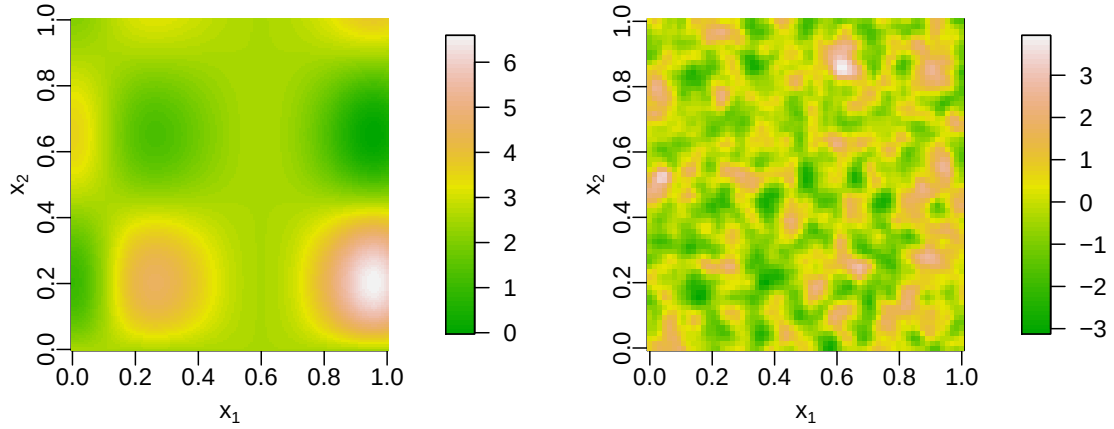

Figure 2: Mean functions used in simulated datasets: simple function (left) with $\hat{\rho} \approx 0.3$ and more complicated GP (right) with $\rho = 0.05$.

I fit the data with the various parameterizations and sampling schemes using MCMC with a burn-in of 10,000 iterations and runs of 100,000 additional iterations. To assess mixing speed, I considered the trace plots, autocorrelations, and effective sample sizes (ESS) (Neal 1993, p. 105),

$$\text{ESS} = \frac{T}{1 + 2\sum_{d=1}^{\infty} \text{Cor}_d(\theta)}, \tag{21}$$

where $\text{Cor}_d(\theta)$ is the autocorrelation at lag d for a given posterior quantity, θ, truncating the summation at the lesser of $d = 10000$ or the largest d such that $\text{Cor}_d(\theta) > 0.05$. I focus on ESS for 1.) the overall log posterior density, $\pi(f(\theta|\boldsymbol{y}))$ (as suggested in Cowles and Carlin (1996) and calculated up to the normalizing constant), 2.) the critical smoothing parameter, ρ, and 3.) a random subset of 200 function estimates.

5.1. Variance component magnitudes and mixing speed

Here I discuss how the magnitude of a key variance component influences mixing. I start with the simple case of normal data.

5.1.1. Normal model and error variance

Under the normal model, as $\eta^2 \to 0$, we have an interpolating surface that passes through the observations. In spatial statistics, such interpolation may arise relatively frequently when measurements are made with little measurement error. In this case, non-zero η^2 is interpreted as fine-scale heterogeneity (Cressie 1993, p. 59). Depending on the dataset, it can be the case that the estimate of η^2 is quite small, but with gaps in the observations, there may quite a bit of uncertainty in prediction at locations between the observations. In this case of interpolation, we have essentially specified a parametric multivariate normal model; if this is not a good interpolation model for the underlying spatial surface, the true uncertainty may be greater than indicated in the posterior because of model uncertainty.

The key sampling consideration arising from small values of η^2 is that the size of MCMC

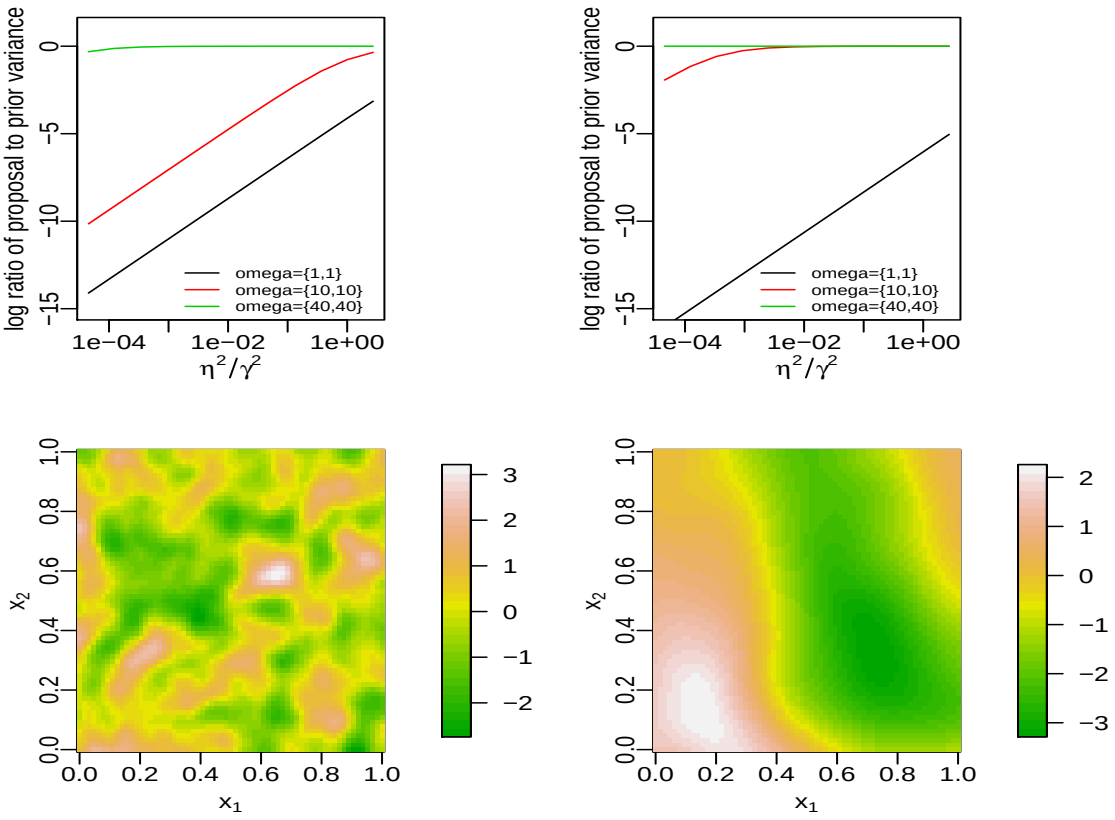

Figure 3: Decay of proposal variances (scaled relative to prior variance) for three representative coefficients as a function of the relative error variance (η^2/γ^2) for (left column) GPs with $\rho = 0.1$ and (right column) GPs with $\rho = 0.4$. Example process realizations are shown in the lower row.

moves for the basis coefficients is quite small. As $\eta^2 \to 0$,

$$V(u_i|\boldsymbol{Y}, \boldsymbol{\theta}) = (\frac{\gamma^2}{\eta^2} + \Sigma_{ii}^{-1})^{-1} \to 0 \tag{22}$$

for the Gibbs sample proposal variance (15-19). For coefficients of low frequency basis functions, as η^2 get small, the proposal variance is a small fraction of the magnitude of the coefficient (for high frequency basis functions, this is not the case, but these have little impact on the process estimate) (Figure 3). This occurs because the process estimates are specified exactly (when $\eta^2 = 0$) at the observation locations, and any proposal at those locations is constrained by the observations. This constrains the proposal for the entire spatial process. Mixing can be challenging in such problems even if uncertainty away from the observations is substantial and of real interest. While this issue seems likely to arise in other GP representations, except when the process values can be integrated out of the model, the issue is particularly clear with the spectral representation.

5.1.2. Latent layer model for exponential family data and dispersion parameter

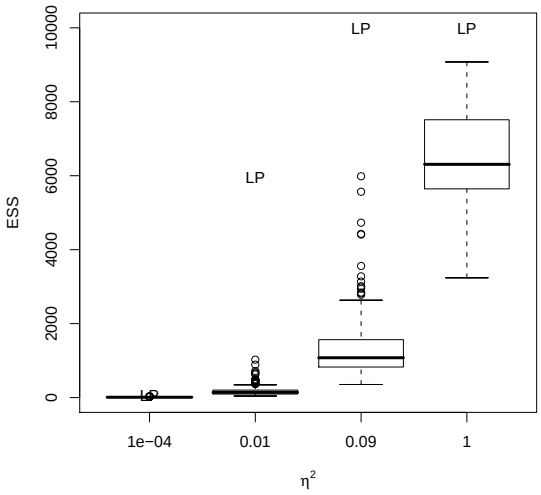

Figure 4: ESS for a sample of 200 function values (boxplots) and for the log posterior density ('LP') as a function of fixed η^2.

The approaches described in Section 4.2 introduce variance component(s) that allow for Gibbs sampling of the coefficients, at the expensive of making the model more complicated and moving the coefficients further away from the data in the model hierarchy. The introduction of the variance components, η^2 and σ_z^2, has important implications for MCMC sampling that relate to the discussion in Section 5.1.1 that the size of η^2 affects the proposal variances for the coefficients.

Let's first consider model (18) with only a single variance component, η^2. If the data are overdispersed, the parameter η^2 can account for this additional dispersion, with $\mu\mathbf{1} + \gamma\boldsymbol{\Psi}\boldsymbol{u}$ representing the unknown smooth function and $\boldsymbol{\lambda}$ a more heterogeneous process in which $\eta^2 > 0$ introduces overdispersion. Note that inference about the unknown smooth function should likely be based on $\mu\mathbf{1} + \gamma\boldsymbol{\Psi}\boldsymbol{u}$ rather than $\boldsymbol{\lambda}$, with simulations showing that inference based on $\boldsymbol{\lambda}$ has larger posterior variances and is overly conservative for the unknown mean function. When there is not overdispersion, the posterior for η^2 should concentrate near zero, indicating that the data are from the exponential family distribution. While this might be the correct inference, if the value of η^2 does approach zero in the MCMC sampling, the chain will mix very slowly, as in the case of normal data, because $V(u_i|\boldsymbol{Y}, \boldsymbol{\lambda}, \boldsymbol{\theta}) \to 0$ as $\eta^2 \to 0$ as in (22). Smaller values of η^2 result in small proposal variances and slow movement of the coefficients. Figure 4 shows the ESS for the log posterior density and for a sample of function values as a function of fixed η^2 for Data3, Poisson data generated without overdispersion.

In the case of overdispersion, the data can inform η^2, and prior distributions such as the inverse gamma priors of Wikle (2002); Royle and Wikle (2005) may suffice. When there is little overdispersion, these priors are more problematic. Note that the inverse gamma prior has a rapidly-decaying left tail, dropping off as $\exp(-1/x)$, so the inverse gamma prior assigns no mass to small values of η^2, preventing the posterior from having mass in this area. For example, the $\mathcal{IG}(0.5, 2)$ prior has very little mass at values less than 0.05 while the $\mathcal{IG}(1, 10)$ has very little mass at values less than 0.006. Fitting the model to Data2 with the $\mathcal{IG}(0.5, 2)$

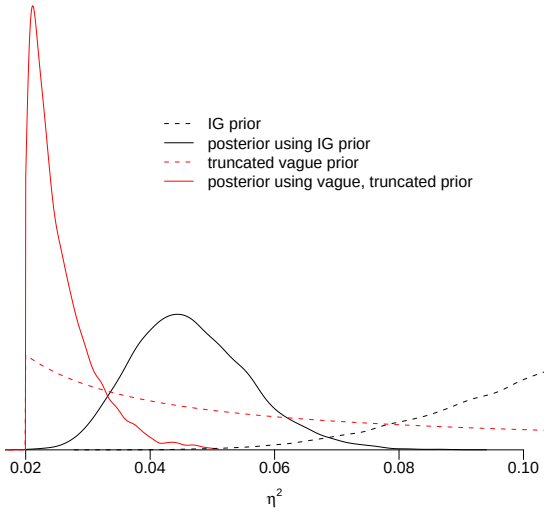

Figure 5: Posterior densities for η^2 under an $\mathcal{IG}(0.5, 2)$ prior and a lognormal prior truncated at 0.02

prior shows that the constraints imposed by the inverse gamma prior cause the posterior to have its mass in the extreme lower range of the prior (Figure 5). Changing the prior to a truncated lognormal distribution results in most of the posterior mass lying at very small values of η^2 near the truncation point, suggesting the sensitivity to the prior.

This suggests that when the observations are not overdispersed ($\eta^2 = 0$) that the prior has substantial impact on the posterior for η^2. A prior that weights the model away from small values of η^2 has the desirable impact of improving mixing at the cost of forcing overdispersion, while use of a prior that allows for small values of η^2 carries the risk of very slow mixing. This suggests that we might choose a prior or even fix η^2 in advance to achieve optimal mixing, treating η^2 as an MCMC tuning parameter. The danger of using large values of η^2 is that while mixing will improve, introduction of overdispersion causes the posterior variances of key quantities such as the function estimates to increase, giving overly conservative inference and less precise point estimates. In Table 2, we see average interval lengths, posterior coverage, and predictive ability based on R^2 for function estimates from Data3 (for which mixing is shown in Figure 4) as a function of fixed values of η^2. For the smallest value of η^2 coverage is too low and intervals are too small because of poor mixing, while for the larger values of η^2 interval lengths increase, with coverage becoming overly conservative and predictive ability declining. A compromise value of η^2 (say 0.3^2 in this case) appears to do a good job of trading off between mixing and precision and interval properties.

In general, to obtain reliable inference about overdispersion, one would want to initially allow sufficient freedom in one's prior for η^2 to allow small values of η^2 to have substantial posterior mass. However, if the data appear to not be overdispersed and one wants to achieve reasonable mixing, one may want to run a version of the model with a fixed, larger value of η^2 and report inference for the other aspects of the model based on that MCMC. One can examine interval length as a function of η^2 in comparison with mixing properties to determine a good value of η^2. Cross-validation may be helpful for assessing coverage.

Table 2: Average 95% credible interval coverage and length, as well as test R^2 of posterior mean, for 200 function estimates, as a function of fixed η^2 for Data3.

η^2	coverage	interval length	test R^2
0.01^2	0.73	1.60	0.33
0.1^2	0.92	2.45	0.57
0.3^2	0.93	2.54	0.57
0.5^2	0.97	2.75	0.53
0.75^2	0.97	2.89	0.49
1^2	0.95	2.93	0.50
2^2	0.94	3.43	0.04

Turning to the model (16), there are two variance components, η^2 and σ_z^2. Interpretation of these parameters is difficult as the parameterization divides any inherent overdispersion into components that are not identifiable if there is no replication within cells. Royle and Wikle (2005) note that in their model η^2 accounts for overdispersion, in particular observer variability in counting birds, while σ_z^2 represents uncorrelated variability across grid cells, inducing a lack of spatial smoothness beyond that induced by the discretization. With replication in the cells, Royle and Wikle (2005) claim that both η^2 and σ_z^2 are identifiable, but that with few replicates posterior correlation of these variance components may be high. Wikle (2002) and Royle and Wikle (2005) sample both components and find reasonable mixing, perhaps because their inverse gamma distributions prevent small values of the components and perhaps because with their real count data there is real overdispersion that provides information about a functional of the variance components ($\eta^2 + \gamma^2\sigma_z^2$), while replication provides information about η^2. In contrast, I have had difficulty achieving reasonable mixing for the two variance components in simulations with no overdispersion ($\sigma_z^2 = \eta^2 = 0$), presumably because of the lack of identifiability and lack of overdispersion. Figure 6 shows example trace plots and superimposed prior and posterior densities for η^2 and σ_z^2 for an MCMC run with Data2. Note how the posteriors concentrate on the smallest values allowed by the priors and how the likelihood mixes well, indicating that the process values, $\boldsymbol{\lambda}$, are well-identified by the data and mix well.

5.2. Non-identifiability of $u_{1,1}$ and μ

The Fourier basis function corresponding to the coefficient, $u_{0,0}$, is a constant function. As such, it is not identifiable with respect to an overall mean parameter, μ, specified outside of the Fourier basis representation of the Gaussian process (16,18,20). One might choose to omit μ from the model, but this would generally be a mistake as the covariance structure (7) imposes a restrictive prior on $u_{0,0}$. A large value of γ or σ would allow for a process mean far from zero, but this would also allow the function to have high variability. An example of where the problem arises is a process with large mean, say 100, but whose variability places the function entirely in $(99, 101)$. Such a process would require a large value of $u_{0,0}$ but if γ or σ is large enough to allow this, each would be so large as favor process estimates that vary widely around 100. Instead, a separate mean parameter is a better choice that will help to avoid slow mixing because of nonidentifiability. One can fix $u_{0,0} = 0$, without otherwise constraining the model. The **spectralGP** library can fix the coefficient and ignore it

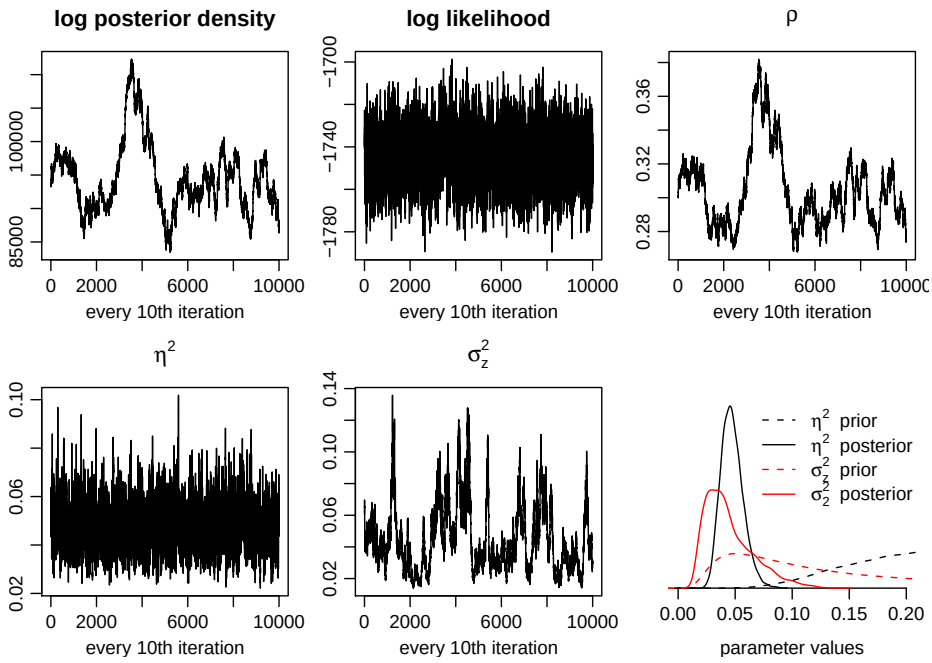

Figure 6: Trace plots for Wikle approach applied to Data2.

when calculating the prior density of the coefficients; this is done based on the `const.fixed` argument to `gp()`. However, to generate GPs using the Fourier basis approximation, one should not fix this constant, in order to retain the desired approximate covariance structure.

5.3. Joint sampling of hyperparameters and process

In parameterizations in which the coefficients are not directly involved in the likelihood, poor mixing may be an issue. In particular, poor mixing can occur for the covariance hyperparameters of the coefficients, namely ρ and, in the centered parameterization, σ^2, or ψ_1 and ψ_2 in the reparameterization described in Section 5.4.

For illustration, consider ρ. The difficulty in sampling ρ is that a simple Metropolis-Hastings proposal for ρ results in a new set of variances for the coefficients, $\boldsymbol{u}$. Since these coefficients are not part of the proposal, proposing ρ^* can easily produce a low prior density for $\boldsymbol{u}|\rho^*$ because the new prior is inconsistent with the current $\boldsymbol{u}$. For example, with a process generated based on $\rho = 0.1$, the prior logdensity of $\boldsymbol{u}|\rho = 0.1$ is 50411 while the prior logdensity of $\boldsymbol{u}|\rho^* = 0.101$ is 50387, a change of 24 in the logdensity, despite the fact that the surfaces generated based on $\rho = 0.1$ compared to $\rho = 0.101$ are indistinguishable even with massive amounts of data, assuming non-negligible error variance. Note that the issue here is not a matter of whether we can sample from the full conditional for ρ; the primary obstacle in sampling ρ is the strong dependence of ρ and $\boldsymbol{u}$ (Rue, Steinsland, and Erland 2004). A better proposal would account for the strong dependence between ρ and $\boldsymbol{u}$ by proposing them jointly, allowing the lower levels of the model hierarchy close to the data to arbitrate between different values of ρ. In the context of Markov random field models, (Rue and Held 2005, pp. 142-143) suggest a similar strategy of jointly sampling process values and their hyperparameters.

My strategy is to tie the covariance hyperparameters more closely to the coefficients and hence to the data by having the effects of proposing new hyperparameter values ripple down through the hierarchy of the model. I do this by jointly proposing a hyperparameter, generically denoted $\theta \in \{\rho, \sigma^2, \psi_1, \psi_2\}$, and, then conditional on that hyperparameter and within the same Metropolis-Hastings proposal, proposing the process coefficients, $\boldsymbol{u}|\theta, \ldots$. This provides a joint proposal for $(\theta, \boldsymbol{u})$ that adjusts $\boldsymbol{u}$ in such a way that it is more consistent with the proposed value, θ^*. Parameterizations that permit proposing $\boldsymbol{u}|\theta, \ldots$ from its full conditional are likely to particularly benefit from this approach, with $\boldsymbol{u}$ effectively integrated out of the joint density. Provided the correct Hastings adjustment (ratio of proposal densities) is made, this joint proposal is a standard, valid Metropolis-Hastings sampling scheme, implemented as a marginal proposal for θ and a conditional proposal for $\boldsymbol{u}|\theta, \ldots$, with a single acceptance decision, as discussed in (Rue and Held 2005, p. 142). I now detail these joint proposals in the three basic sampling schemes described in Section 4.

For the data augmentation sampling approach for normal data, one proposes θ^*, and then uses the full conditional to propose $\boldsymbol{u}|\boldsymbol{y}, \theta^*$ (10). Acceptance is then determined based on the ratio of the proposed and current posterior densities for $\boldsymbol{u}, \theta|\boldsymbol{y}$ divided by the Hastings ratio for $\boldsymbol{u}$ (and θ as well if not proposed symmetrically). Note that the proposal for $\boldsymbol{u}$ is conditional on the proposed θ^*, so a Metropolis-Hastings acceptance decision is needed because we are not doing a joint Gibbs sample for $(\boldsymbol{u}, \theta)$. The Hastings ratio is based on the proposal mean (10) and variance (11), with the variances for complex-valued coefficients scaled by two (7) and not including the coefficients that are complex conjugates of sampled coefficients. This is calculated in the **spectralGP** library using the `Hastings.coeff.gp()` function. Template code is provided in the appendix as Code E.

In the modified Wikle approach, the presence of the latent $\tilde{\boldsymbol{\lambda}}$ values (18) distances the coefficients from the data. However, one can mimic the proposal just described by sampling θ and $\boldsymbol{u}|\theta, \boldsymbol{\lambda}$, conditioning on $\boldsymbol{\lambda}$ rather than $\boldsymbol{y}$, and being satisfied with a proposal for $\boldsymbol{u}$ that is consistent with the new proposed θ^* and the current $\boldsymbol{\lambda}$, albeit without any direct influence of the data. Again a Hastings correction is needed, and can be calculated using the `Hastings.coeff.gp()` function, but with $\boldsymbol{\lambda}$ taking the place of $\boldsymbol{y}$. Template code is given as Code F. For the original Wikle approach, $\boldsymbol{z}$ takes the place of $\boldsymbol{\lambda}$ above.

However, in neither the modified nor original Wikle parameterizations does the sampling directly link θ to the observations, causing there to be no influence of the likelihood on the acceptance. An alternative that carries the changes through to the level of the data is to avoid sampling from the conditionals as described above and instead propose to move $\boldsymbol{u}$ and $\boldsymbol{\lambda}$ (and $\boldsymbol{z}$ in the Wikle parameterization) in such a way that their prior densities remain constant. First propose θ^*. Then, deterministically propose,

$$u_i^* = u_i \cdot \frac{\sqrt{(\Sigma_{\theta^*})_{i,i}}}{\sqrt{(\Sigma_\theta)_{i,i}}}, \quad i = 1, \ldots, M. \tag{23}$$

Modifying u_i based on its prior variance, $(\boldsymbol{\Sigma}_\theta)_{i,i}$, allows the hyperparameters to mix more quickly by avoiding proposals for which the original coefficients are no longer probable based on their new prior variances. In the modified Wikle parameterization, one next proposes

$$\lambda_i^* = \lambda_i - \gamma(\boldsymbol{\Psi}\boldsymbol{u})_i + \gamma(\boldsymbol{\Psi}\boldsymbol{u}^*)_i, \tag{24}$$

while in the Wikle parameterization, one proposes $z_i^* = z_i - (\boldsymbol{\Psi}\boldsymbol{u})_i + (\boldsymbol{\Psi}\boldsymbol{u}^*)_i$ and finally $\lambda_i^* = \lambda_i - \gamma \boldsymbol{K}_i \boldsymbol{z} + \gamma \boldsymbol{K}_i \boldsymbol{z}$. This approach propagates the changes through the model in a way

Table 3: ESS for log posterior density, ρ, and median (range) of 200 sample function values by dataset for three sampling approaches: 1.) sampling γ and ρ using simple Metropolis-Hastings, 2.) jointly sampling each of σ^2 and ρ with $\boldsymbol{u}$ based on conditional Gibbs sample for $\boldsymbol{u}$, and 3.) jointly sampling each of σ^2 and ρ with $\boldsymbol{u}$ based on deterministic modification of $\boldsymbol{u}$ (23). η^2 is fixed at 0.3^2. 'NM' indicates that the chain has not burned in or is mixing so slowly as to make calculation of ESS uninformative.

		Sampling Method		
Quantity	Dataset	simple: γ, ρ	joint, Gibbs: σ^2, ρ	joint, deterministic: σ^2, ρ
LP	Data1	NM	42	193
	Data2	NM	87	205
	Data3	NM	17	447
	Data5	NM	143	646
ρ	Data1	NM	37	146
	Data2	NM	70	145
	Data3	NM	15	414
	Data5	NM	125	289
f	Data1	549 (28-1898)	510 (34-1338)	597 (90-1808)
	Data2	1806 (22-3496)	1805 (132-3763)	2137 (229-4231)
	Data3	236 (5-2154)	611 (200-3062)	657 (307-3232)
	Data5	1563 (17-6503)	1964 (519-8958)	2021 (467-8971)

that ties θ directly to the likelihood. Such deterministic proposals are valid MCMC proposals so long as the Jacobian of the transformation is included in the acceptance ratio, based on a modification of the argument in Green (1995). The Jacobian of the transformation for $\boldsymbol{u}$ cancels with the ratio of the prior distributions for $\boldsymbol{u}$, $\pi(\boldsymbol{u}^*|\theta^*)/\pi(\boldsymbol{u}|\theta)$, to give the final Metropolis-Hastings acceptance for the entire joint proposal of $(\theta^*, \boldsymbol{u}^*, \boldsymbol{\lambda}^*)$ or $(\theta^*, \boldsymbol{u}^*, \boldsymbol{z}^*, \boldsymbol{\lambda}^*)$ based only on the ratio of the proposed and current prior densities for θ, the proposed and current likelihoods, and any required Hastings ratio to account for non-symmetric proposals for θ. Note that the transformations for $\boldsymbol{z}$ and $\boldsymbol{\lambda}$ have Jacobian of one. The validity of the deterministic proposal can be seen intuitively by considering Metropolis proposals in place of the transformation (23) with the mean (23) and very small proposal variances, $\zeta^2 \approx 0$, e.g., $u_i^* \sim \mathcal{N}(u_i((\boldsymbol{\Sigma}_{\theta^*})_{i,i})^{1/2}((\boldsymbol{\Sigma}_{\theta})_{i,i})^{-1/2}, \zeta^2)$, and calculating the acceptance ratio of such a proposal. Template code for the modified Wikle parameterization is given in the appendix as Code G. Note that a similar joint proposal could be made for $\theta = \sigma_z^2$ to tie this hyperparameter more closely to the data.

For the coefficient block sampling scheme, no Gibbs scheme is available. Instead, one can carry out a joint sample in a similar manner to that just described, by sampling θ^* and then $\boldsymbol{u}^*|\theta^*$ based on (23). Acceptance is determined by the ratio of the proposed and current prior densities for θ and proposed and current likelihoods, and any required Hastings ratio to account for non-symmetric proposals for θ. Template code is given in the appendix as Code H.

In Table 3, I show a comparison of mixing for the modified Wikle parameterization (18) with 1.) straightforward sampling of γ and ρ, 2.) joint sampling of $\{\sigma, \boldsymbol{u}\}$ and of $\{\rho, \boldsymbol{u}\}$ based on Gibbs samples for $\boldsymbol{u}|\sigma^2, \ldots$ and $\boldsymbol{u}|\rho, \ldots$, and 3.) joint sampling via (23). Both joint sampling approaches appear to mix much more quickly than the simple Metropolis-Hastings

proposals for the hyperparameters. Surprisingly, the joint sampling with the full conditional sampling for $\boldsymbol{u}|\theta,\ldots$ does not mix as well as the deterministic shift of $\boldsymbol{u}|\theta$, perhaps because the conditional Gibbs sample does not modify $\boldsymbol{\lambda}$ and therefore does not involve the likelihood in the determination of proposal acceptance, whereas the deterministic approach shifts $\boldsymbol{\lambda}$ as well as $\boldsymbol{u}$. Note that for $\eta^2 \equiv 0.2^2$, the improved mixing of the deterministic shift proposal compared to the conditional Gibbs is even more marked (not shown).

5.4. Covariance parameterization

In a GP model part of the difficulty in estimating the covariance parameters occurs because of limitations on identifiability. The data cannot readily distinguish the overall variability in the function, captured by γ or σ, from the decay in the spatial correlation, captured by ρ. In Bayesian models, these parameters tend to have high posterior correlation, while Zhang (2004) has shown that these two parameters cannot both be estimated consistently under infill asymptotics, but that a functional of the two can be estimated consistently. Note that in thin plate spline models and in the mixed model representation suggested by Kammann and Wand (2003) and Ruppert *et al.* (2003), there is only one parameter in place of the two covariance parameters here. However, as discussed further in Section 6, comparisons of estimates using the Fourier basis approach here suggest that ρ cannot be fixed in advance without seriously affecting the function estimates because the function heterogeneity is not adequately represented.

Given the results of Zhang (2004), in which the ratio, $\sigma^2/\rho^{2\nu}$, can be estimated consistently, consider reparameterizing on the log scale as $\psi_1 = \log\sigma + \log\rho$ and $\psi_2 = \log\sigma - \log\rho$. This approach uses the centered parameterization, fixing $\gamma \equiv 1$. The reparameterization will tend to reduce posterior correlation and allow each parameter to move more freely. Joint sampling as described in Section 5.3 can also be employed with this reparameterization. Template code for sampling based on the reparameterization and joint sampling of the parameters and process values using deterministic conditional proposals for $\boldsymbol{u}$ (23-24) is given in the appendix as Code I under the modified Wikle parameterization and code J under the block sampling approach.

Since the joint sampling of θ with $\boldsymbol{u}$ based on deterministic proposals for $\boldsymbol{u}$ and $\boldsymbol{\lambda}$ appeared to be the best of the options in Section 5.3, in Table 4, I compare mixing for that approach with the (σ^2, ρ) parameterization and the same joint approach using deterministic proposals with the (ψ_1, ψ_2) parameterization. There is little difference in mixing between the two parameterizations. Table 5 shows posterior correlations of (σ, ρ) and of (ψ_1, ψ_2) based on sampling under the original and the new parameterizations. For Data1, Data2, and Data3, ψ_1 and ψ_2 have little posterior correlation, suggesting that in principle, sampling using the new parameterization would mix more quickly, although this is not the case in Table 4. The minimal difference in mixing was also seen when using the joint sampling with the full conditional samples of $\boldsymbol{u}|\theta,\ldots$ and when fixing $\eta^2 = 0.2^2$ and $\eta^2 = 0.5^2$ and for an alternative reparameterization, $\psi_1 = \log\sigma$ and $\psi_2 = \log\sigma - \log\rho$. In practice, the minimal difference in mixing suggests that the posterior correlation between σ^2 and ρ is not materially hurting mixing, in sharp contrast to the importance of jointly sampling θ and $\boldsymbol{u}$.

5.5. Empirical comparison of sampling methods and recommendations

Based on the evidence provided in Sections 5.3 and 5.4, it appears that joint sampling of θ

Table 4: ESS for log posterior density, ρ, and median (range) of 200 sample function values by dataset when sampling is done using the parameterizations: 1.) $\{\rho, \sigma^2\}$ and 2.) $\{\psi_1, \psi_2\}$. η^2 is fixed at 0.3^2.

		Parameterization	
Quantity	Dataset	original: σ^2, ρ	Zhang: ψ_1, ψ_2
LP	Data1	193	105
	Data2	205	244
	Data3	447	443
	Data5	646	777
ρ	Data1	146	56
	Data2	145	157
	Data3	414	431
	Data5	289	426
f	Data1	597 (90-1808)	591 (89-1427)
	Data2	2137 (229-4231)	2154 (235-4274)
	Data3	657 (307-3232)	656 (305-3169)
	Data5	2021 (467-8971)	1972 (469-8879)

Table 5: Posterior correlations for (σ, ρ) and (ψ_1, ψ_2) when sampling is done using the parameterizations: 1.) $\{\log\sigma, \log\rho\}$ and 2.) $\{\psi_1, \psi_2\}$. η^2 is fixed at 0.3^2.

		Parameterization	
Dataset	Posterior correlation	original: $\log\sigma, \log\rho$	Zhang: ψ_1, ψ_2
Data1	$\text{Cor}(\log\sigma, \log\rho)$	0.70	0.79
	$\text{Cor}(\psi_1, \psi_2)$	0.15	0.24
Data2	$\text{Cor}(\log\sigma, \log\rho)$	0.81	0.84
	$\text{Cor}(\psi_1, \psi_2)$	0.08	0.13
Data4	$\text{Cor}(\log\sigma, \log\rho)$	0.20	0.21
	$\text{Cor}(\psi_1, \psi_2)$	0.26	0.26
Data5	$\text{Cor}(\log\sigma, \log\rho)$	0.56	0.56
	$\text{Cor}(\psi_1, \psi_2)$	0.11	0.12

Table 6: ESS for log posterior density, ρ, and median (range) of 200 sample function values by dataset for the three parameterizations: 1.) modified Wikle with η^2 fixed at 0.3^2, 2.) original Wikle parameterization and sampling approach, and 3.) block sampling.

		Sampling Method		
Quantity	Dataset	modified Wikle	original Wikle	block sampling
LP	Data1	193	NM	54
	Data2	205	NM	25
	Data3	447	NM	53
	Data5	646	NM	NM
ρ	Data1	146	NM	107
	Data2	145	NM	40
	Data3	414	NM	NM
	Data5	289	22	NM
f	Data1	597 (90-1808)	125 (13-321)	551 (247-1166)
	Data2	2137 (229-4231)	56 (9-166)	473 (198-873)
	Data3	657 (307-3232)	298 (106-737)	33 (7-139)
	Data5	2021 (467-8971)	467 (127-1100)	25 (6-83)

and $\boldsymbol{u}$ in the modified Wikle parameterization greatly improves mixing, with deterministic sampling of $\boldsymbol{u}$ better than full conditional sampling for $\boldsymbol{u}$. Also, there is little improvement from using the parameterization with ψ_1 and ψ_2.

Here I compare mixing for the three parameterizations in Section 4: the modified Wikle approach with joint sampling of hyperparameters and coefficients, block sampling with joint sampling of hyperparameters and coefficients, and the original approach of Wikle. Since the latter is essentially the same as the modified Wikle approach with one extra layer, I do not devise a joint sampling scheme for it, but rather consider mixing under the sampling approach proposed by Wikle (2002) and Royle and Wikle (2005). In general, the modified Wikle approach outperforms block sampling and the original Wikle approach. Table 6 shows that for the simple function, block sampling is worse than the modified Wikle approach but shows some degree of mixing, while for the more complicated function, the block sampling approach does not appear to have burned in by 100,000 iterations. The original Wikle approach also has not burned in, as judged by the log posterior density and ρ although the sample function values appear to be mixing reasonably. Note that while the increase in sample size (from Data1 to Data2 and from Data3 to Data5) seems to result in somewhat improved mixing, the effect is not substantial.

A key question is how fine a resolution to use for the grid. While one does not want to oversmooth by virtue of using too coarse a resolution, finer resolution estimation takes longer to run and can exhibit slower mixing, because of the higher-dimensionality of the coefficients that are fit in the MCMC. My suggestion is to use a grid that is fine enough for reasonable prediction with the expected heterogeneity of the surface, but to make use of sensitivity analyses to choose the grid resolution in light of mixing performance and computational speed. For the simple simulated data with an effective value of $\rho \approx 0.3$ (Data1 and Data2), a resolution of $k = 128$ is probably more than sufficient for good prediction (even coarser resolution might be sufficient), and runs with $k = 256$ and $k = 512$ showed slower mixing. For the simulated data with $\rho = 0.05$ (Data3, Data4, and Data5), $k = 128$ also seemed to be

sufficient. Mixing with $k = 256$ was not substantially degraded relative to $k = 128$, but for $k = 512$, mixing was substantially worse.

These results suggest that mixing using the block sampling approach is substantially slower than the modified Wikle approach, particularly with a more variable underlying process. However, results may depend significantly on the form of the model and the exact data used. In Paciorek and Ryan (2005), with a coarse grid, simple spatial functions, and binary observations, mixing was reasonable using the block sampling approach. In a multivariate setting within a complicated hierarchical model (Paciorek and McLachlan, in prep.), with a compound Dirichlet-multinomial likelihood for 10 categories and a coarse 32 by 32 grid, mixing was reasonable, albeit slow, with the block sampling approach, and the modified Wikle approach provided no improvement and was slower to compute.

5.6. Starting values

Good starting values for the coefficients can be difficult to determine because of the high dimensionality of the coefficients and lack of a maximum likelihood based estimate due to the need for shrinkage in estimating the coefficients. In addition as described in Section 2.2, a portion of the domain contains no observations. For the grid points not used to represent the domain of interest $\left(((0,1)^2)^C \cap (0,2)^2\right)$, it is helpful to initiate values for these buffering grid points that keep the variability and spatial range features of the data similar across the whole domain. This can be achieved by 'mirroring' the initial values from the portion of the domain in which the observations lie, as follows, in one dimension,

$$\hat{g}(s_M), \ldots, \hat{g}(s_{M/2+2}) \equiv \hat{g}(s_{M/2}), \ldots, \hat{g}(s_2). \tag{25}$$

In two dimensions, the mirroring occurs first across the the line $s_1 = 1$ (for $s_2 < 1$) and then across the line $s_2 = 1$, such that $\hat{g}(s_{m_1,m_2})$ is defined, for $m_1 > M_1/2+1$ and $m_2 <= M_2/2+1$ as $\hat{g}(s_{m_1,m_2}) \equiv \hat{g}(s_{M_1-m_1+2,m_2})$. For $m_2 > M_2/2+1$, take $\hat{g}(s_{m_1,m_2}) \equiv \hat{g}(s_{m_1,M_2-m_2+2})$.

In the data augmentation scheme for normal data, we suggest using a `gam()` fit to estimate the process values, predicting $\tilde{\boldsymbol{Y}}$ values at unobserved locations using the fitted model, mirroring the values, and then doing a Gibbs sample for the coefficients. In the Wikle approach, one can estimate the spatial process at the grid points based on a `gam()` fit, assign these values to $\boldsymbol{z}$ ($\boldsymbol{\lambda}$ in the modified Wikle approach) and initialize $\boldsymbol{u}$ via a Gibbs sample. For the block sampling scheme, one might use `gam()` to estimate the process on the grid, $\hat{\boldsymbol{g}}_{\boldsymbol{s}\#}$, add error and mirror the values, and then estimate $\boldsymbol{u} = (\frac{\gamma^2}{\eta^2}\boldsymbol{I} + \boldsymbol{\Sigma}^{-1})^{-1}\frac{\gamma}{\eta^2}\boldsymbol{\Psi}^T(\hat{\boldsymbol{g}}_{\boldsymbol{s}\#} - \mu\boldsymbol{1})$, mimicing (19).

Some basic experiments with simulated datasets Data1, Data2, Data3, and Data5 suggest little difference between starting the coefficients based on a Gibbs sample and starting at values simulated from the prior conditional on the hyperparameter starting values. Reasonably rapid burn-in occurred when the coefficients were simulated from their prior, although mixing for Data1 was slightly better for the Gibbs sample starting values. For the coefficients corresponding to low frequencies, the long-run estimates are comparable for the different starting values. However, it may be the case that the Gibbs sample starting values are useful in some circumstances.

6. Discussion

This paper introduces an R library for the Fourier basis representation of Gaussian processes, pioneered by Wikle (2002), and provides template code for fitting Bayesian models for exponential family data. The code can be readily adapted for more complicated hierarchical models as well. I discuss several possible parameterizations, including models allowing for overdispersion, and describe potential nonidentifiability in the hierarchical model of Wikle (2002) that may impact mixing. I document some of the critical issues affecting MCMC mixing in these models, in particular, the difficulty in mixing for ρ in particular and the dependence of mixing speed on the dispersion parameter, η^2. In models with little noise (interpolating models) or non-Gaussian data cases with little overdispersion, a small value of η^2 can substantially impede mixing. Based on a series of experiments with simulated Poisson data, I recommend use of a modified version of the parameterization of Wikle (2002), with an approach for joint sampling of the hyperparameters and the basis coefficients to more efficiently sample the hyperparameters by tying them more closely to the data. In contrast, while the block sampling approach of Paciorek and Ryan (2005) works only somewhat less well for a relatively smooth spatial function, it mixes very poorly for a very unsmooth spatial function. However, the block sampling approach has the virtue of avoiding the overdispersion parameter that, if small, can hurt mixing and of simplicity, which may be helpful in more complicated hierarchical models. I could not achieve reasonable mixing of the parameterization and sampling approach suggested in Wikle (2002), presumably because of dataset-dependent differences in mixing, but also possibly because of the difficulty in replicating Bayesian MCMC schemes. Note that these recommendations and conclusions are based on qualitative rather than exhaustive testing.

The critical smoothing parameters (ρ and either σ or γ) appear to be the parameters that mix most slowly in the Fourier basis representation, as they are in many spatial models. In particular, ρ changes the amount of smoothing, by changing the prior weights on the basis functions, which vary in their frequency. Changing this parameter changes the form of the model, analogous to adding or subtracting basis functions in a free-knot spline model. Achieving reasonable mixing across model spaces is generally difficult.

Some alternative spatial models, such as thin plate splines and radial basis function models with fixed basis functions (Kammann and Wand 2003; Ruppert *et al.* 2003) have modeled spatial functions without estimating a spatial correlation parameter, relying solely on variance components (in the radial basis model) to achieve smoothing. O'Connell and Wolfinger (1997) relate the ratio of σ^2 and η^2 in a Gaussian setting to the smoothing parameter in a thin plate spline model, and Nychka (2000) speculates that this ratio may be more important than the spatial correlation parameter in smoothing noisy data. Zhang (2004) found that ρ and σ^2 cannot both be estimated consistently under infill asymptotics. I experimented with fixing ρ and forcing σ^2 to perform the smoothing role, but found that the model did not estimate the right amount of smoothing and predictive performance was poor. It may be that in this and perhaps other spatial models, models with estimated values of ρ are more efficient. This issue appears not to have been addressed thoroughly in the literature (but see Laslett (1994) and invited comments) and deserves more attention.

One might explore more sophisticated MCMC algorithms to improve mixing. For example, Christensen *et al.* (2006) develop a data-dependent reparameterization scheme for improved MCMC performance and apply the approach with Langevin updates that use gradient information; while promising, the approach is computationally intensive, again involving $n \times n$ matrix computations at each iteration, and software is not available. For the Fourier represen-

tation the high-dimensionality and complex values of the basis coefficients pose an impediment to such an approach. Based on the results here, I believe that proposals that jointly consider the key hyperparameters and the basis coefficients are critical in achieving adequate mixing.

One drawback to the GP model presented here is its restriction to stationary GPs. Future work on this model structure to allow for nonstationarity in the spatial process will consider wavelet bases in place of the Fourier basis used here, in particular the two-dimensional wavelet basis used by Matsuo, Paul, and Nychka (2006) to fit irregular Gaussian data in a non-Bayesian fashion. However, mixing may be more challenging in a more complicated model with additional hyperparameters. An alternative relates to the work of Pintore and Holmes (2006), who have extended the Higdon/Paciorek/Stein (Higdon, Swall, and Kern 1999; Stein 2005; Paciorek and Schervish 2006) nonstationary covariance model based on kernel convolutions to the spectral domain. This allows one to build nonstationarity based on a latent process representing spatially-varying ρ or ν. Given the widespread interest in nonstationary and space-time representations, fast computation for such models is of obvious interest, but it is not clear how these covariance structures would be represented in the type of basis function approach developed here.

7. Appendix: Template code

I provide R template code for various parameterizations and sampling approaches described in Sections 4 and 5. The code makes use of the **spectralGP** library. The code uses easily modifiable R functions for the loglikelihood, prior distributions, and Gibbs sampling; the names of these will be obvious in the code. Also note that parameters take the form of R lists, with components that will be obvious from the code. The code does not save iterations, report acceptance rates, or adapt the proposal variances based on acceptance rates, but these features could be readily added.

References

Banerjee S, Carlin B, Gelfand A (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall.

Berger JO, De Oliveira V, Sansó B (2001). "Objective Bayesian analysis of spatially correlated data." *Journal of the American Statistical Association*, **96**(456), 1361–1374.

Booth JG, Hobert JP (1999). "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm." *Journal of the Royal Statistical Society, Series B*, **61**, 265–285.

Borgman L, Taheri M, Hagan R (1984). "Three-dimensional, frequency-domain simulations of geological variables." In ea Verly G (ed.), "Geostatistics for Natural Resources Characterization, Part 1," pp. 517–541. D. Reidel Publishing Company.

Christensen O, Møller J, Waagepetersen R (2000). "Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo." *Technical Report R-002009*, Department of Mathematics, Aalborg University. URL www.math.auc.dk/~rw/publications.html.

Christensen O, Roberts G, Sköld M (2006). "Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models." *Journal of Computational and Graphical Statistics*, **15**, 1–17.

Christensen OF, Waagepetersen R (2002). "Bayesian prediction of spatial count data using generalized linear mixed models." *Biometrics*, **58**, 280–286.

Cowles MK, Carlin BP (1996). "Markov chain Monte Carlo convergence diagnostics: A comparative review." *Journal of the American Statistical Association*, **91**, 883–904.

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

Diggle PJ, Tawn JA, Moyeed RA (1998). "Model-based geostatistics." *Applied Statistics*, **47**, 299–326.

Dudgeon D, Mersereau R (1984). *Multidimensional Digital Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey.

Furrer R, Genton MG, Nychka D (2006). "Covariance Tapering for Interpolation of Large Spatial Datasets." *Journal of Computational and Graphical Statistics*, **in press**.

Gelfand A, Sahu S, Carlin B (1996). "Efficient parametrizations for generalized linear mixed models." In J Bernardo, J Berger, A Dawid, A Smith (eds.), "Bayesian Statistics 5," pp. 165–180.

Gelman A (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, **1**(3), 515–534.

Gibbons RD, Hedeker D (1997). "Random effects probit and logistic regression models for three-level data." *Biometrics*, **53**, 1527–1537.

Green P (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, **82**, 711–732.

Hastie TJ, Tibshirani RJ (1990). *Generalized Additive Models*. Chapman & Hall Ltd, London. ISBN 0412343908.

Heagerty PJ, Lele SR (1998). "A Composite Likelihood Approach to Binary Spatial Data." *Journal of the American Statistical Association*, **93**, 1099–1111.

Heagerty PJ, Lumley T (2000). "Window Subsampling of Estimating Functions with Application to Regression Models." *Journal of the American Statistical Association*, **95**(449), 197–211.

Hedeker D, Gibbons RD (1994). "A random-effects ordinal regression model for multilevel analysis." *Biometrics*, **50**, 933–944.

Higdon D, Swall J, Kern J (1999). "Non-stationary spatial modeling." In J Bernardo, J Berger, A Dawid, A Smith (eds.), "Bayesian Statistics 6," pp. 761–768. Oxford University Press, Oxford, U.K.

Hwang JN, Lay SR, Maechler M, Martin D, Schimert J (1994). "Regression modeling in back-propagation and projection pursuit learning." *IEEE Transactions on Neural Networks*, **5**, 342–353.

Kammann E, Wand M (2003). "Geoadditive models." *Applied Statistics*, **52**, 1–18.

Laslett GM (1994). "Kriging and Splines: An Empirical Comparison of Their Predictive Performance in Some Applications (Disc: P 401-409)." *Journal of the American Statistical Association*, **89**, 391–400.

Matsuo T, Paul D, Nychka D (2006). "Nonstationary covariance modeling for incomplete data: smoothed Monte Carlo EM approach." *in prep.*

McCulloch CE (1994). "Maximum likelihood variance components estimation for binary data." *Journal of the American Statistical Association*, **89**, 330–335.

McCulloch CE (1997). "Maximum likelihood algorithms for generalized linear mixed models." *Journal of the American Statistical Association*, **92**, 162–170.

McCulloch CE, Searle SR (2001). *Generalized, linear, and mixed models*. John Wiley & Sons.

Neal R (1993). "Probabilistic Inference Using Markov Chain Monte Carlo Methods." *Technical Report CRG-TR-93-1*, Department of Computer Science, University of Toronto. URL www.cs.toronto.edu/~radford/papers-online.html.

Nychka DW (2000). "Spatial-process Estimates As Smoothers." In MGa Schimek (ed.), "Smoothing and regression: approaches, computation, and application," pp. 393–424. John Wiley & Sons.

O'Connell M, Wolfinger R (1997). "Spatial regression models, response surfaces, and process optimization." *Journal of Computational and Graphical Statistics*, **6**, 224–241.

Paciorek C, Ryan L (2005). "Computational techniques for spatial logistic regression with large datasets." *Technical Report 32*, Harvard University Biostatistics.

Paciorek C, Schervish M (2006). "Spatial modelling using a new class of nonstationary covariance functions." *Environmetrics*, **17**, 483–506.

Pintore A, Holmes C (2006). "Non-stationary covariance functions via spatially adaptive spectra." *Journal of the American Statistical Association*, **in review**.

Royle JA, Wikle CK (2005). "Efficient Statistical Mapping of Avian Count Data." *Environmental and Ecological Statistics*, **12**(2), 225–243.

Rue H, Held L (2005). *Gaussian Markov random fields: Theory and applications*. Chapman & Hall, Boca Raton.

Rue H, Steinsland I, Erland S (2004). "Approximating Hidden Gaussian Markov Random Fields." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **66**(4), 877–892.

Rue H, Tjelmeland H (2002). "Fitting Gaussian Markov Random Fields to Gaussian Fields." *Scandinavian Journal of Statistics*, **29**(1), 31–49.

Ruppert D, Wand M, Carroll R (2003). *Semiparametric regression*. Cambridge University Press, Cambridge, U.K.

Shumway R, Stoffer D (2000). *Time Series Analysis and its Applications*. Springer-Verlag, New York.

Stein M (2005). "Nonstationary spatial covariance functions." *Technical Report 21*, University of Chicago.

Stein ML, Chi Z, Welty LJ (2004). "Approximating Likelihoods for Large Spatial Data Sets." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **66**(2), 275–296.

Wikle C (2002). "Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains." In A Lawson, D Denison (eds.), "Spatial Cluster Modelling," pp. 199–209. Chapman & Hall.

Wood S (2006). *Generalized additive models: An introduction with R*. Chapman & Hall, Boca Raton.

Wood SN (2004). "Stable and efficient multiple smoothing parameter estimation for generalized additive models." *Journal of the American Statistical Association*, **99**, 673–686.

Zhang H (2004). "Inconsistent estimation and asymptotically equal interpolation in model-based geostatistics." *Journal of the American Statistical Association*, **99**, 250–261.

Affiliation:

Christopher J. Paciorek
Department of Biostatistics
Harvard School of Public Health
655 Huntington Avenue
Boston, MA 02115, USA
E-mail: paciorek@alumni.cmu.edu
URL: http://www.biostat.harvard.edu/~paciorek/