CARNEGIE MELLON UNIVERSITY

NONSTATIONARY GAUSSIAN PROCESSES
FOR REGRESSION AND SPATIAL MODELLING

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
for the degree
DOCTOR OF PHILOSOPHY
in
STATISTICS

by
Christopher Joseph Paciorek

Department of Statistics
Carnegie Mellon University
Pittsburgh Pennsylvania 15213

May 2003

Abstract

Recent work in the areas of nonparametric regression and spatial smoothing has focused on modelling functions of inhomogeneous smoothness. In the regression literature, important progress has been made in fitting free-knot spline models in a Bayesian context, with knots automatically being placed more densely in regions of the covariate space in which the function varies more quickly. In the spatial statistics literature, attention has focused on using nonstationary covariance structures to account for inhomogeneity of the spatial field.

In this dissertation, I use nonstationary covariance functions in Gaussian process (GP) prior distributions over functions to perform both nonparametric regression and spatial smoothing in a Bayesian fashion. I extend the kernel convolution method of Higdon et al. (1999) to create a class of nonstationary covariance functions. I prove that the nonstationary covariance functions retain the differentiability properties of the stationary correlation functions on which they are based, provided there is sufficient smoothness in the underlying kernel structure used to generate the nonstationarity. The stationary Matérn covariance function has desirable differentiability properties; the generalized kernel convolution method developed here provides a Matérn-based nonstationary covariance function.

I develop a generalized nonparametric regression model and assess difficulties in identifiability and in fitting of the model using Markov Chain Monte Carlo (MCMC) algorithms. Of particular note, I show how to improve MCMC performance for non-Gaussian data based on an approximate conditional posterior mean. The modelling approach produces a flexible response surface that responds to inhomogeneity while naturally controlling overfitting. For Gaussian errors, on test datasets in one dimension, the GP model performs well, but not as well as the free-knot spline method. However, in two and three dimensions, the nonstationary GP model seems to outperform both free-knot spline models and a stationary GP model. Unfortunately, as implemented the method is not feasible for datasets with more than a few hundred observations because of the computational difficulties involved in fitting the model.

The nonstationary covariance model can also be embedded in a spatial model. In particular, I analyze spatiotemporal climate data, using a nonstationary covariance matrix to model the spatial structure of the residuals. I demonstrate that the nonstationary model fits the covariance structure of the data better than a stationary model, but any improvement in point predictions relative to a stationary model or to the maximum likelihood estimates is minimal, presumably because the data are very smooth to begin with. My comparison of various correlation models for the residuals highlights the difficulty in fitting high-dimensional covariance structures.

Acknowledgements

I would like to acknowledge the advice, suggestions, support, and friendship of a number of people who helped me during the writing of this thesis and the rest of my time as a graduate student in the Department of Statistics. First, I would like to thank my advisor, Mark Schervish, for his ongoing involvement in this work. In his understated way, Mark offered hands-on advice without hands-on management, giving suggestions that often guided me in a better direction. He has always been willing to delve into the details of the work and talk through issues, sometimes long past a reasonable meeting length. I would also like to thank the other members of my committee. James Risbey got me started on the Advanced Data Analysis project that led to this thesis. I thank him for his collaboration, advice, and friendship. I thank Doug Nychka for an enjoyable and productive visit to NCAR as well as for his various suggestions and his assistance with the wavelet model. I thank Valérie Ventura for her support and for her collaboration on two related projects. Larry Wasserman offered an open door, of which I did not take as much advantage as I would have liked toward the close of my graduate work. The statistics department staff and the rest of the faculty in the department have helped me in various ways, and I thank them for making the department a great place to do graduate work. I'd also like to thank Chris Holmes for the Matlab code on his website.

I want to thank the graduate students in the department. In particular, Fang Chen, my longtime officemate, has tolerated with good humor my often foul moods during the writing of this thesis.

Seven years is a long time to be in graduate school; I'm glad Susan Davidson was there for many of them. I'd also like to thank old friends who do not need to be named to know their importance to me. Little did we know in grade school and high school that I would be 31 when I finally finished school. John, how many seconds have I been in school?

Finally, but most importantly, I thank my parents for setting me on the road with what I would need to get to this point.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Definition

This thesis treats the general problem of nonparametric regression, also known as smoothing, curve-fitting, or surface-fitting. Both the statistics and machine learning communities have investigated this problem intensively, with Bayesian methods drawing particular interest recently. Spline-based methods have been very popular among statisticians, while machine learning researchers have approached the issue in a wide variety of ways, including Gaussian process (GP) models, kernel regression, and neural networks. The same problem in the specific context of spatial statistics has been approached via kriging, which is essentially a Gaussian process-based method.

Much recent effort has focused on the problem of inhomogeneous smoothness, namely when the function of interest has different degrees of smoothness in one region of covariate space than another region. Many standard smoothing methods are not designed to handle this situation. Methods that are able to model such functions are described as spatially adaptive. Recent Bayesian spline-based methods have concentrated on adaptively placing knots to account for inhomogeneous smoothness. Spatial statisticians have been aware of this issue for some time now, since inhomogeneous smoothness can be expected in many spatial problems, and have tried several approaches to the problem. One approach, which I use as the stepping-off point for my work, is a Bayesian treatment of kriging in which the covariance model used in the Gaussian process prior distribution for the spatial field is nonstationary, i.e., the covariance structure varies with spatial location. Higdon, Swall, and Kern (1999) pioneered one approach to nonstationarity in the spatial context, while machine learning researchers have implemented the approach in a limited way for nonparametric regression problems.

1.2 Gaussian Processes and Covariance Functions

Gaussian process distributions and the covariance functions used to parameterize these distributions are at the heart of this thesis. Before discussing how Gaussian processes and competing methods are used to perform spatial smoothing and nonparametric regression, I will introduce Gaussian processes and covariance functions.

The Gaussian process distribution is a family of distributions over stochastic processes, also called random fields or random functions (I will generally use 'function' in the regression context and 'process' or 'field' in the context of geographic space). A stochastic process is a collection of random variables, $Z(\boldsymbol{x}, \omega)$, on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ indexed by a variable, $\boldsymbol{x} \in \mathcal{X}$. For the purpose of this thesis, this indexing variable represents space, either geographic space or covariate space (feature space in the language of machine learning), and $\mathcal{X} = \Re^P$. In another common context, the variable represents time. Fixing ω and letting $\boldsymbol{x}$ vary gives sample paths or sample functions of the process, $z(\boldsymbol{x})$. The smoothness properties (continuity and differentiability) of these sample paths is one focus of Chapter 2. More details on stochastic processes can be found in Billingsley (1995) and Abrahamsen (1997), among others.

The expectation or mean function, $\mu(\cdot)$, of a stochastic process is defined by

$$\mu(\boldsymbol{x}) = \mathrm{E}(Z(\boldsymbol{x}, \omega)) = \int_\Omega Z(\boldsymbol{x}, \omega) d\mathcal{P}(\omega).$$

The covariance function, $C(\cdot, \cdot)$ of a stochastic process is defined for any pair $(\boldsymbol{x_i}, \boldsymbol{x_j})$ as

$$\begin{aligned} C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \mathrm{Cov}(Z(\boldsymbol{x_i}, \omega), Z(\boldsymbol{x_j}, \omega)) \\ &= \mathrm{E}((Z(\boldsymbol{x_i}, \omega) - \mu(\boldsymbol{x_i}))(Z(\boldsymbol{x_j}, \omega) - \mu(\boldsymbol{x_j}))) \\ &= \int_\Omega (Z(\boldsymbol{x_i}, \omega) - \mu(\boldsymbol{x_i}))(Z(\boldsymbol{x_j}, \omega) - \mu(\boldsymbol{x_j})) d\mathcal{P}(\omega). \end{aligned}$$

For the rest of this thesis, I will suppress the dependence of $Z(\cdot)$ on $\omega \in \Omega$. Stochastic processes are usually described based on their finite dimensional distributions, namely the probability distributions of finite sets, $\{Z(\boldsymbol{x_1}), Z(\boldsymbol{x_2}), \ldots, Z(\boldsymbol{x_n})\}\,, n = 1, 2, \ldots,$ of the random variables in the collection $Z(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}.$ Unfortunately, the finite dimensional distributions do not completely determine the properties of the process (Billingsley 1995). However, it is possible to establish the existence of a version of the process whose finite dimensional distributions determine the sample path properties of the process (Doob 1953, pp. 51-53; Adler 1981, p. 14), as discussed in Section 2.5.2.

A Gaussian process is a stochastic process whose finite dimensional distributions are multivariate normal for every n and every collection $\{Z(\boldsymbol{x_1}), Z(\boldsymbol{x_2}), \ldots, Z(\boldsymbol{x_n})\}$. Gaussian processes are specified by their mean and covariance functions, just as multivariate Gaussian distributions are specified by their mean vector and covariance matrix. Just as a covariance matrix must be positive definite, a covariance function must also be positive definite; if the function is positive definite, then the finite dimensional distributions are consistent (Stein 1999, p. 16). For a covariance function on $\Re^P \otimes \Re^P$ to be positive definite, it must satisfy

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\boldsymbol{x_i}, \boldsymbol{x_j}) > 0$$

for every n, every collection $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$, and every vector $\boldsymbol{a}$. This condition ensures, among other things, that every linear combination of random variables in the collection will have positive variance. By Bochner's theorem (Bochner 1959; Adler 1981, Theorem 2.1.2), the class of weakly stationary, continuous non-negative definite complex-valued functions is equivalent to the class of bounded non-negative real-valued measures. In particular, a stationary, continuous correlation function is the characteristic function, or Fourier transform, of a distribution function and the inverse Fourier transform of a such a correlation function is a distribution function. See Abrahamsen (1997) or Stein (1999) for more details.

Gaussian processes are widely used in modeling spatial data (Diggle, Tawn, and Moyeed 1998; Holland, Oliveira, Cox, and Smith 2000; Lockwood, Schervish, Gurian, and Small 2001). In particular, the geostatistical method of kriging assumes a Gaussian process structure for the unknown spatial field and focuses on calculating the optimal linear predictor of the field. In most applications, stationary (also known as homogeneous) covariance functions are used for simplicity.

Stationarity in the wide sense (weak stationarity) is defined as

$$\begin{aligned} (i) \quad & \mathrm{E}\,|Z(\boldsymbol{x})|^2 && < \infty \\ (ii) \quad & \mathrm{E}Z(\boldsymbol{x}) && = \mu \\ (iii) \quad & C(\boldsymbol{x_i}, \boldsymbol{x_j}) && = C(\boldsymbol{x_i} - \boldsymbol{x_j}), \end{aligned} \tag{1.1}$$

where μ is a constant mean. The condition (1.1) requires that the covariance be solely a function of the separation vector. In addition, if the covariance is also solely a function of a distance metric, the process is said to be isotropic. In $\Re^P$, an isotropic process is a function only of Euclidean distance, $\tau = \|\boldsymbol{x_i} - \boldsymbol{x_j}\|$. Recent research has focused on modelling nonstationary covariance, as summarized in Section 1.3.

Ensuring positive definiteness involves ensuring the positive definiteness of the correlation function, $R(\cdot, \cdot)$, defined by

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{C(\boldsymbol{x_i}, \boldsymbol{x_j})}{\sigma(\boldsymbol{x_i})\sigma(\boldsymbol{x_j})},$$

where $\sigma^2(\boldsymbol{x_i}) = C(\boldsymbol{x_i}, \boldsymbol{x_i})$ is the variance function. The only restriction on the variance function is that it be positive. Many stationary, isotropic correlation functions have been proposed (Yaglom 1987; Abrahamsen 1997; MacKay 1997). Here I introduce several common stationary, isotropic correlation functions for which I produce nonstationary versions in Section 2.3. The following correlation functions are all positive definite on $\Re^p, p = 1, 2, \ldots$.

1. Power exponential:

$$R(\tau) = \exp\left(-\left(\frac{\tau}{\kappa}\right)^{\nu}\right), \kappa > 0, 0 < \nu \leq 2 \tag{1.2}$$

2. Rational quadratic (Cauchy):

$$R(\tau) = \frac{1}{\left(1 + (\frac{\tau}{\kappa})^2\right)^{\nu}}, \kappa > 0, \nu > 0 \tag{1.3}$$

3. Matérn:

$$R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{2\sqrt{\nu}\tau}{\kappa}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\tau}{\kappa}\right), \kappa > 0, \nu > 0 \tag{1.4}$$

κ and ν are parameters, τ is distance, and K_ν is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun 1965, sec. 9.6). The power exponential form (1.2) includes two commonly used correlation functions as special cases: the exponential ($\nu = 1$) and the squared exponential ($\nu = 2$), also called the Gaussian correlation function. These two correlation functions are also related to the Matérn correlation (1.4). As $\nu \rightarrow \infty$, the Matérn approaches the squared exponential correlation. The use of $\frac{2\sqrt{\nu}\tau}{\kappa}$ rather than simply $\frac{\tau}{\kappa}$ ensures that the Matérn correlation approaches the squared exponential correlation function of the form

$$R(\tau) = \exp\left(-\left(\frac{\tau}{\kappa}\right)^2\right) \tag{1.5}$$

and that the interpretation of κ is minimally affected by the value of ν (Stein 1999, p. 50). For $\nu = 0.5$, the Matérn correlation (1.4) is equivalent to a scaled version of the usual exponential correlation function

$$R(\tau) = \exp\left(-\frac{\sqrt{2}\tau}{\kappa}\right).$$

In general, κ controls how fast the correlation decays with distance, which determines the low-frequency, or coarse-scale, behavior of sample paths generated from stochastic processes with the given correlation function. ν controls the high-frequency, or fine-scale, smoothness properties of the sample paths, namely their continuity and differentiability. An exception is that the smoothness does not change with ν for the rational quadratic function. In Section 2.5.4, I discuss the smoothness characteristics of sample paths based on the correlation functions above.

1.3 Spatial Smoothing Methods

The prototypical spatial smoothing problem involves estimating a smooth field based on noisy data collected at a set of spatial locations. Statisticians have been interested in constructing smoothed maps and in doing prediction at locations for which no data were collected. The standard approach to the problem has been that of kriging, which involves using the data to estimate the spatial covariance structure in an ad hoc way and then calculating the mean and variance of the spatial field at each point conditional on both the data and the estimated spatial covariance structure (Cressie 1993). This approach implicitly uses the conditional posterior mean and variance from a Bayesian model with constant variance Gaussian errors and a Gaussian process prior for the spatial field.

In particular, when performing kriging, researchers have generally assumed a stationary, often isotropic, covariance function, with the covariance of the responses at any two locations assumed to be a function of the separation vector or of the distance between locations, but not a function of the actual locations. Researchers often estimate the parameters of an isotropic covariance function from the semivariogram,

$$\gamma(\boldsymbol{x_i} - \boldsymbol{x_j}) = \frac{\text{Var}\,(Z(\boldsymbol{x_i}) - Z(\boldsymbol{x_j}))}{2},$$

which is estimated based on the squared differences between the responses as a function of the distance between the locations.

Next I develop the basic Gaussian process prior model underlying kriging. (See Cressie (1993, Chapter 3) for the traditional description of kriging.) The model is

$$\begin{aligned} Y_i &\sim \text{N}(f(\boldsymbol{x_i}), \eta^2), \\ f(\cdot) &\sim \text{GP}\left(\mu_f, C_f(\cdot,\cdot;\boldsymbol{\theta}_f)\right), \end{aligned}$$

where each $\boldsymbol{x_i} \in \Re^2, i = 1, \ldots, n$, is a spatial location. $f(\cdot)$ has a Gaussian process prior distribution with covariance function, $C_f(\cdot,\cdot;\boldsymbol{\theta}_f)$, which is a function of hyperparameters, $\boldsymbol{\theta}_f$. I will refer to the entire function as $f(\cdot)$ and to a vector of values found by evaluating the function at a finite set of points as $\boldsymbol{f} = (f(\boldsymbol{x_1}, \ldots, f(\boldsymbol{x_n}))^T$, while $f(\boldsymbol{x})$ will refer to the function evaluated at the single point $\boldsymbol{x}$. If $\boldsymbol{x}$ takes infinitely many different values, then $C_f(\cdot,\cdot)$ is the covariance function, and if a finite set of locations is under consideration, then $C_{\boldsymbol{f}}$ is the covariance matrix calculated by applying the covariance function to all pairs of the locations. Taking $C_{\boldsymbol{Y}} = \eta^2 I_n$ and suppressing the dependence of $C_{\boldsymbol{f}}$ on $\boldsymbol{\theta}_f$, the conditional posterior distribution for $\boldsymbol{f}$, $\Pi(\boldsymbol{f}|\boldsymbol{Y}, \eta, \mu_f, \boldsymbol{\theta}_f)$, is normal with

$$\begin{aligned} \text{E}\left(\boldsymbol{f}|\boldsymbol{Y},\eta,\mu_f,\boldsymbol{\theta}_f\right) &= C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y} + C_{\boldsymbol{Y}}\left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1}\mu_f && (1.6) \\ &= \mu_f + C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1}\left(\boldsymbol{Y} - \mu_f\right) && (1.7) \\ \text{Cov}\left(\boldsymbol{f}|\boldsymbol{Y},\eta,\mu_f,\boldsymbol{\theta}_f\right) &= \left(C_{\boldsymbol{f}}^{-1} + C_{\boldsymbol{Y}}^{-1}\right)^{-1} && (1.8) \\ &= C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1} C_{\boldsymbol{Y}}. && (1.9) \end{aligned}$$

The posterior mean (1.6) is a linear combination of the prior mean, μ_f, and the observations, $\boldsymbol{Y}$, weighted based on the covariance terms. The second form of the posterior mean (1.7) can be seen to be a linear smoother (Section 1.4.3) of the data offset by the function mean. Here and elsewhere in this thesis, as necessary, I take $\mu = \mu\mathbf{1}$, when a vector-valued object is required. The posterior variance (1.8) is the inverse of the sum of the precision matrices.

Prediction at unobserved locations is simply the usual form for a conditional Gaussian distribution. If we take

$$\boldsymbol{f} = \begin{pmatrix} \boldsymbol{f_1} \\ \boldsymbol{f_2} \end{pmatrix}$$

and

$$C_{\boldsymbol{f}} = \begin{pmatrix} C_{\mathbf{11}} & C_{\mathbf{12}} \\ C_{\mathbf{21}} & C_{\mathbf{22}} \end{pmatrix},$$

where $\mathbf{1}$ indicates the set of locations at which data have been observed and $\mathbf{2}$ the set at which one wishes to make predictions, then the conditional posterior for $\boldsymbol{f_2}$, $\Pi(\boldsymbol{f_2}|\boldsymbol{f_1}, \boldsymbol{Y}, \eta, \mu_f, \boldsymbol{\theta}_f)$ is normal with

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{f_2}|\boldsymbol{f_1}, \boldsymbol{Y}, \eta, \mu_f, \boldsymbol{\theta}_f\right) &= \mu_f + C_{\mathbf{21}} C_{\mathbf{11}}^{-1}(\boldsymbol{f_1} - \mu_f) && (1.10) \\ \mathrm{Cov}\left(\boldsymbol{f_2}|\boldsymbol{f_1}, \boldsymbol{Y}, \eta, \mu_f, \boldsymbol{\theta}_f\right) &= C_{\mathbf{22}} - C_{\mathbf{21}} C_{\mathbf{11}}^{-1} C_{\mathbf{12}}, && (1.11) \end{aligned}$$

and the marginal (with respect to $\boldsymbol{f_1}$) posterior for $\boldsymbol{f_2}$, $\Pi(\boldsymbol{f_2}|Y, \eta, \mu_f, \boldsymbol{\theta}_f)$ is normal with

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{f_2}|\boldsymbol{Y}, \eta, \mu_f, \boldsymbol{\theta}_f\right) &= \mu_f + C_{\mathbf{21}} \left(C_{\mathbf{11}} + C_{\boldsymbol{Y}}\right)^{-1} \left(\boldsymbol{Y} - \mu_f\right) && (1.12) \\ \mathrm{Cov}\left(\boldsymbol{f_2}|\boldsymbol{Y}, \eta, \mu_f, \boldsymbol{\theta}_f\right) &= C_{\mathbf{22}} - C_{\mathbf{21}} \left(C_{\mathbf{11}} + C_{\boldsymbol{Y}}\right)^{-1} C_{\mathbf{12}}. && (1.13) \end{aligned}$$

While the Gaussian process model is defined over an infinite dimensional space, the calculations are performed in the finite dimensional space at the locations of interest. One important drawback of Gaussian process models, discussed further in Chapters 3 and 6, is that the computational burden is $O(n^3)$ because of the need to invert matrices of order n (i.e., to solve systems of equations of the form $C\boldsymbol{b} = \boldsymbol{y}$). Unlike some competitors, such as splines, for which there is a simple expression for the function once the parameters are known, for Gaussian process models, prediction involves the matrix operations given above.

The standard kriging approach allows one to flexibly estimate a smooth spatial field, with no pre-specified parametric form, but has several drawbacks. The first is that the true covariance structure may not be stationary. For example, if one is modelling an environmental variable across the

United States, the field is likely to be much more smooth in the topographically-challenged Great Plains than in the Rocky Mountains. This is manifested as different covariance structures in those two regions; the covariance structure changes with location. Assuming a stationary covariance structure will result in oversmoothing the field in the mountains and undersmoothing the field in the Great Plains. A second drawback is that the usual kriging analysis does not account for the uncertainty in the spatial covariance structure, since fixed hyperparameters are often used. A final drawback is that an ad hoc approach to estimating the covariance structure may not give as reliable estimates as a more principled approach.

These issues have been addressed in various ways. Smith (2001, p. 66) suggests using likelihood-based methods for estimating the covariance structure as an alternative to ad hoc estimation. Handcock and Stein (1993) present a Bayesian version of kriging that accounts for uncertainty in the spatial covariance structure. Higdon (1998), Higdon et al. (1999), and Swall (1999) have used Bayesian Gaussian process analogues to kriging that account for uncertainty in the covariance structure, although they have encountered some difficulty in implementing models in which all the covariance hyperparameters are allowed to vary, and they have been forced to fix some hyperparameters in advance. Recently, a number of approaches have been proposed for modelling nonstationarity. (For a review, see Sampson, Damian, and Guttorp (2001).) I will first describe in detail the method of Higdon et al. (1999), since it is their approach that I use as the foundation for my own work, and then I will outline other approaches to the problem.

The approach of Higdon et al. (1999) is to define a nonstationary covariance function based on the convolution of kernels centered at the locations of interest. They propose a nonstationary spatial covariance function, $C(\cdot,\cdot)$, defined by

$$C(\boldsymbol{x_i},\boldsymbol{x_j})=\int_{\Re^2}K_{\boldsymbol{x_i}}(\boldsymbol{u})K_{\boldsymbol{x_j}}(\boldsymbol{u})d\boldsymbol{u}, \tag{1.14}$$

where $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, and $\boldsymbol{u}$ are locations in $\Re^2$ and $K_{\boldsymbol{x}}$ is a kernel function (not necessarily non-negative) centered at $\boldsymbol{x}$. This covariance function is positive definite for spatially-varying kernels of any functional form, as I show in Chapter 2. They motivate this construction as the covariance function of a process, $Z(\cdot)$, constructed by convolving a white noise process, $\psi(\cdot)$, with a spatially-varying kernel, $K_{\boldsymbol{x}}$:

$$Z(\boldsymbol{x})=\int_{\Re^2}K_{\boldsymbol{x}}(\boldsymbol{u})\psi(\boldsymbol{u})d\boldsymbol{u}.$$

The evolution of the kernels in space produces nonstationary covariance, and the kernels are usually parameterized so that they vary smoothly in space, under the assumption that nearby locations will share a similar local covariance structure. Higdon et al. (1999) use Gaussian kernels, which give a closed form for $C(\boldsymbol{x_i}, \boldsymbol{x_j})$, the convolution (1.14), as shown in Section 2.2.

Fuentes and Smith (2001) and Fuentes (2001) have an alternate kernel approach in which the process is taken to be the convolution of a fixed kernel over independent stationary processes, $Z_{\boldsymbol{\theta}(\boldsymbol{u})}(\cdot)$,

$$Z(\boldsymbol{x}) = \int K(\boldsymbol{x} - \boldsymbol{u}) Z_{\boldsymbol{\theta}(\boldsymbol{u})}(\boldsymbol{x}) d\boldsymbol{u}.$$

The resulting covariance, $C(\cdot, \cdot)$ is expressed as

$$C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int K(\boldsymbol{x_i} - \boldsymbol{u}) K(\boldsymbol{x_j} - \boldsymbol{u}) C_{\boldsymbol{\theta}(\boldsymbol{u})}(\boldsymbol{x_i} - \boldsymbol{x_j}) d\boldsymbol{u}.$$

For each $\boldsymbol{u}$, $C_{\boldsymbol{\theta}(\boldsymbol{u})}(\cdot, \cdot)$ is a covariance function with parameters $\boldsymbol{\theta}(\boldsymbol{u})$, where $\boldsymbol{\theta}(\boldsymbol{u})$ is a (multivariate) spatial process that induces nonstationarity in $Z(\cdot)$. This method has the advantage of avoiding the need to parameterize smoothly varying positive-definite matrices, as required in Higdon et al. (1999)'s Gaussian kernel approach. One drawback to the approach is the lack of a general closed form for $C(\boldsymbol{x_i}, \boldsymbol{x_j})$ and the need to compute covariances by Monte Carlo integration; this is of particular concern because of the numerical sensitivity of covariance matrices (Section 3.3). In addition to Bayesian methods, Fuentes and Smith (2001) and Fuentes (2001) describe spectral methods for fitting models when the data are (nearly) on a grid; these may be much faster than likelihood methods.

In a completely different approach, Sampson and Guttorp (1992) have used spatial deformation to model nonstationarity. (See Meiring, Monestiez, Sampson, and Guttorp (1997) for a discussion of computational details.) They map the original Euclidean space to a new Euclidean space in which approximate stationarity is assumed to hold and then use a stationary covariance function to model the covariance in this new space. Schmidt and O'Hagan (2000) and Damian, Sampson, and Guttorp (2001) have presented Bayesian versions of the deformation approach in which the mapping is taken to be a thin-plate spline and a stationary Gaussian process, respectively. Das (2000) has extended the deformation approach to the sphere, modelling nonstationary data collected on the surface of the globe. In the remainder of this thesis, I focus on the Higdon et al. (1999) approach because I find it more easily adaptable to the problems at hand and potentially less computationally burdensome through the closed-form expression for the covariance terms (Section 2.3).

While most attention in the spatial statistics literature has focused on smoothing fields based a single set of spatial observations, in many cases, replicates of the field are available, for example with environmental data collected over time. This sort of data is becoming even more common with the growing availability of remotely-sensed data. In this situation, one has multiple replicates for estimating the spatial covariance structure. The methods that I describe in this thesis allow one to model such replicated data, albeit with certain restrictions, such as modelling only non-negative covariances. Nychka, Wikle, and Royle (2001) have proposed a method for smoothing the empirical covariance structure of replicated data by thresholding the decomposition of the empirical covariance matrix in a wavelet basis. This approach has the advantages of allowing for very general types of covariance structure and of being very fast by virtue of use of the discrete wavelet transform. One potential drawback to the approach is that it is not clear how much or what type of thresholding to do, since there is no explicit model for the data. Given the difficulties involved in modelling high-dimensional covariance structures, it is also not clear how well the resulting smoothed covariance approximates the true covariance in a multivariate sense, although Nychka et al. (2001) have shown in simulations that individual elements of the smoothed covariance matrix can closely approximate the elements of stationary covariance matrices. In modelling storm activity data in Chapter 5, I compare a nonstationary covariance model based on the methods of Higdon et al. (1999) to smoothing the empirical covariance as proposed by Nychka et al. (2001). Both methods may encounter difficulties that reside in their attempt to model or approximate the full covariance structure of many locations, which involves the intricacies of high-dimensional covariance structures.

One advantage of the nonstationary covariance model based on Higdon et al. (1999) is that it fully defines the covariance at unobserved as well as observed locations and does not require a regular grid of locations. This stands in contrast to the approach of Nychka et al. (2001), although they have briefly suggested an iterative approach to deal irregularly-spaced locations. Nott and Dunsmuir (2002) present a method for extending a given covariance at observed locations to unobserved locations in a locally-stationary fashion; this might be used in conjunction with the Nychka et al. (2001) method.

1.4 Nonparametric Regression Methods

1.4.1 Gaussian process methods

Performing Bayesian nonparametric regression using Gaussian process priors for functions is essentially the same as a Bayesian approach to the kriging methodology given above. Much of the work in this area has been done by machine learning researchers, although the general approach was first introduced by O'Hagan (1978), who has subsequently used the methods to analyze computer experiments via surface fitting (Kennedy and O'Hagan 2001). The basic approach is to define the same model as given for the spatial smoothing problem,

$$\begin{aligned} Y_i &\sim \mathrm{N}\left(f\left(\boldsymbol{x_i}\right), \eta^2\right) \\ f(\cdot) &\sim \mathrm{GP}\left(\mu_f, C_f\left(\cdot, \cdot; \boldsymbol{\theta}_f\right)\right), \end{aligned}$$

where each $\boldsymbol{x_i} \in \Re^P, i = 1, \ldots, n$, is a P-dimensional vector of covariates (features in machine learning jargon). If we condition on the hyperparameters, the expressions for the posterior of $\boldsymbol{f}$ are the same as given for the spatial model (1.6-1.13).

Since work by Neal (1996) showing that a certain form of neural network model converges, in the limit of infinite hidden units, to a Gaussian process regression model, Gaussian process approaches have seen an explosion of interest in the machine learning community, with much recent attention focusing on methods for efficient computation (Section 3.7). Machine learning researchers have used many of the covariance functions used in the spatial statistics literature, with the most used seemingly the multivariate squared exponential,

$$C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sigma^2 \exp\left(\sum_{p=1}^{P} \frac{\left(x_{i,p} - x_{j,p}\right)^2}{\kappa_p^2}\right), \tag{1.15}$$

which allows the smoothness of the function to vary with covariate, based on the covariate-specific scale parameter, κ_p (Rasmussen 1996; Neal 1997). One appealing feature of a GP model with this covariance structure is that the number of parameters in the model is the same as the number of covariates and hence grows slowly as the dimensionality increases, in contrast to many multivariate regression models. In a tutorial, MacKay (1997) mentions several covariance functions well-known amongst spatial statisticians, including the power exponential form, while also discussing the notion of 'warping' or 'embedding', which is the deformation approach of Sampson and Guttorp (1992). Vivarelli and Williams (1999) discuss the use of a more general squared exponential covariance of the form

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\frac{1}{2}\left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)^T \Sigma \left(\boldsymbol{x_i} - \boldsymbol{x_j}\right)\right),$$

where Σ is an arbitrary positive definite matrix, rather than the diagonal matrix implicit in (1.15). In the spatial statistics literature, this approach is commonly used to model anisotropy, the situation in which the spatial correlation decays at different rates in different spatial directions.

While stationarity is generally recognized in the spatial statistics literature as an assumption likely to be violated, and research in this area is ongoing and diverse, there has been relatively little work on nonstationarity in the regression context. In this context, nonstationarity exhibits itself as the smoothness of the regression function varying in the covariate space. MacKay (1997) proposed the use of nonstationary covariance functions as a way of dealing with inhomogeneous smoothness. Gibbs (1997) modelled a one-dimensional situation with a mapping approach similar to that of Sampson and Guttorp (1992) as well as a Gaussian process model with a nonstationary covariance function that is equivalent to the nonstationary covariance used by Higdon et al. (1999). Gibbs (1997) showed that dealing with the inhomogeneity gave a qualitatively better estimate of the regression function than using a stationary covariance function.

In this thesis, I extend the nonstationary covariance methods of Gibbs (1997) and Higdon et al. (1999) and describe an implementation in the nonparametric regression setting. My work provides one approach by which the one- (Gibbs) and two-dimensional (Higdon) nonstationary models can be extended to higher dimensions, although in practice, the computational demands of the models limit the dimensionality that can be entertained. I assess the models in one, two, and three dimensions, but do not know how they would perform in higher dimensions.

1.4.2 Other methods

The nonstationary GP regression model that I propose has a number of competitors. In the statistics literature, many researchers have focused on spline-based models with others advocating wavelet bases, while machine learning researchers have investigated many modelling approaches, including neural networks, kernel regression, and regression versions of support vector machines. Various tree methods that divide the covariate space and fit local models in the regions are also popular. In this section, I will describe some of the competing methods and outline connections between the GP model and other models. The methods can be roughly divided into three categories with varying approaches to the bias-variance tradeoff: a) penalized or regularized fitting, which includes some of the spline-based methods and wavelet thresholding, b) model selection approaches that use a small number of parameters to prevent overfitting, which include fixed-knot spline techniques and basis function regression, and c) model averaging approaches such as free-knot splines and the Gaussian process models described above.

Splines are flexible models that take the form of piecewise polynomials joined at locations called knots. Continuity constraints are generally imposed at the knots so that the function is smooth. Once the knots are fixed, estimating a regression function is the same as fitting a linear regression model,

$$\mathrm{E}(Y_i|x_i) = f(x_i) = \sum_{k=1}^{K+2} b_k(x_i)\beta_k,$$

since the function $f(\cdot)$ is linear in the basis functions, $b_k(\cdot)$, determined by the knots, with coefficients β_k. This fitting approach is termed regression splines. Cubic polynomials are most commonly used for the piecewise functions; this makes $f(\cdot)$ a cubic spline. A natural cubic spline forces the function to be linear outside a bounded interval. Splines and natural splines can be represented by many different bases. Among these are the truncated power basis and the B-spline basis; the B-spline basis is generally preferred because it is computationally stable (DiMatteo, Genovese, and Kass 2002).

To address the bias-variance tradeoff at the heart of nonparametric regression, spline researchers have taken several general approaches. A number of researchers have attempted to adaptively place knots based on the observed data; Zhou and Shen (2001) do this in a non-Bayesian iterative fashion, searching for the optimal knot locations. The Bayesian adaptive regression splines (BARS) method of DiMatteo et al. (2002) builds on previous work (Denison, Mallick, and Smith 1998a) to adaptively sample the number and locations of knots in a Bayesian fashion using reversible-jump Markov chain Monte Carlo (RJMCMC). These free-knot spline approaches allow the estimated function to adapt to the characteristics of the data, in particular to variable smoothness of the function over the space $\mathcal{X}$. These approaches allow movement between different basis representations of the data, thereby performing model averaging.

Spline models can also be approached from the smoothing spline perspective, which involves minimizing the penalized sum of squares,

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int \left(f^{(m)}(s)\right)^2 ds, \tag{1.16}$$

where (m) denotes the mth derivative of $f(\cdot)$ and λ is a smoothing parameter that penalizes lack of smoothness, as measured by the integrated squared derivative. The solution to this optimization problem turns out to be a natural spline of degree $2m - 1$ with knots at each data point (the cubic spline is of degree 3 and corresponds to $m = 2$) (Wahba 1990). By changing the value of λ one changes the smoothness of the estimated function. A compromise between the smoothing and regression splines approaches is that of penalized splines (Wand 2003), in which an intermediate number of knots is chosen in advance, often based on the distribution of the covariates, and the minimization of (1.16) is done. This approach tries to choose the underlying basis functions more carefully than the smoothing splines approach, which relies heavily on the penalty term, so as to reduce the computational cost involved in placing knots at each data point. The smoothing and penalized spline models are typically fit via classical methods with the smoothing parameter chosen by cross-validation. Approaching the curve-fitting problem from the smoothing spline perspective gives spatially homogeneous functions, since λ acts on the whole space, while nonstationarity can be obtained manually through the placement of knots in the penalized splines approach. For spatially heterogeneous functions, Ruppert and Carroll (2000) suggest a penalized splines approach in which the penalty function varies spatially and is modelled itself as penalized spline with a single penalty parameter.

DiMatteo (2001) shows how both the regression and smoothing/penalized splines approaches can be seen as estimates from Bayesian models with particular prior structures. In particular she focuses on the following Bayesian model,

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{\beta}, B, \sigma^2, K, \boldsymbol{\xi}, \eta &\sim \mathrm{N}_n(B\boldsymbol{\beta}, \eta^2 I) \\ \boldsymbol{\beta} \mid \sigma^2, \delta, K, \boldsymbol{\xi}, \eta &\sim \mathrm{N}_{K+2}(0, \eta^2 D(\delta)), \end{aligned}$$

with additional priors specified for K, the number of knots, and $\boldsymbol{\xi}$, a vector of knot locations. The matrix B is the basis matrix and $\boldsymbol{\beta}$ is a vector of coefficients. $D(\delta)$ is a covariance matrix whose structure varies depending on the type of spline model. $D(\delta) = n(B^T B)^{-1}$ for regression splines, while for smoothing and penalized splines, $D(\delta) = (\lambda\Omega)^{-1}$, where Ω is a matrix whose elements depend on integrated derivatives of the underlying basis functions. Based on this model, we can see that, conditional on the knots, the spline model is a Gaussian process prior model, with $\boldsymbol{f} = B\boldsymbol{\beta}$, so that we have $\boldsymbol{f} \sim \mathrm{N}(0, \eta^2 BD(\delta)B^T)$. The prior covariance matrix for $\boldsymbol{f}$ is a particular function of the error variance, the basis functions, and the covariance matrix D. For the regression spline approach, if the number and locations of the knots change, the prior covariance changes as well, so the free-knot spline model can be thought of as adaptively choosing a nonstationary prior covariance structure. The relationship of smoothing splines to Gaussian process priors can also be seen directly from the formulation of minimizing the penalized sum of squares loss function, $L(\boldsymbol{f})$,

$$\begin{aligned} L(\boldsymbol{f}) &= \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f^{(2)}(x))^2 dx \\ &\propto -\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{2\eta^2} - \frac{1}{2}\lambda' \boldsymbol{f}^T \Omega \boldsymbol{f} \\ &= \log g(\boldsymbol{Y}|\boldsymbol{f}, \eta) + \log \Pi(\boldsymbol{f}|\lambda', \Omega) \\ &\propto \log \Pi(\boldsymbol{f}|\boldsymbol{Y}, \eta, \lambda', \Omega), \end{aligned}$$

where g is the likelihood function and the last term is the log posterior density for $\boldsymbol{f}$. Hence the natural cubic spline that is the solution to minimizing the penalized sum of squares is the posterior mean from a GP model with a particular prior covariance. The prior for $\boldsymbol{f}$, $\mathrm{N}(0, (\lambda'\Omega)^{-1})$, is partially improper because Ω is positive semi-definite, having two zero eigenvalues, corresponding to improper priors for constant and linear functions (Green and Silverman 1994, p. 55).

Thin-plate splines are the generalization of smoothing splines to higher dimensions (Green and Silverman 1994, Ch. 7). Natural thin-plate splines are the solution to minimizing a generalization of (1.16) to higher order partial derivatives of the function (Green and Silverman 1994, p. 142), and hence can be viewed as a GP-based model in the same way that smoothing splines can be. The implicit underlying covariance, which is fixed in advance and not fit to the data, is a generalized covariance (Cressie 1993, p. 303; O'Connell and Wolfinger 1997). A single parameter controls the degree of smoothing, so the thin-plate spline approach yields a spatially homogeneous smoother.

Moving from one-dimensional curve-fitting to surface-fitting in higher dimensions, such as with thin-plate splines, is an important challenge that poses many difficulties, including those of defining an appropriate model, avoiding overfitting, finding structure in spaces with sparse data (the curse of dimensionality), and interpretability. Many authors take the approach of using additive models (Hastie and Tibshirani 1990), including the work of DiMatteo et al. (2002) in extending BARS to higher dimensions; these approaches retain interpretability and attempt to find structure of a particular form by constraining the model. However, in many cases, the underlying function may not be additive in nature and may contain interactions of various sorts. The team of researchers who first worked on reversible-jump MCMC methods for spline-based regression (Denison et al. 1998a) have taken several approaches to nonparametric regression modelling in multivariate spaces (Denison, Holmes, Mallick, and Smith 2002). One approach is a Bayesian version (Denison, Mallick, and Smith 1998b) of the MARS algorithm (Friedman 1991). MARS uses basis functions that are tensor products of univariate splines in the truncated power basis. In the Bayesian formulation, knots are allowed at the data points and their number and locations are sampled via RJMCMC. Holmes and Mallick (2001) use multivariate linear splines, also fit in a Bayesian fashion using RJMCMC. The basis functions are truncated linear planes, which give a surface that is continuous but not differentiable where the planes meet. One reason for the use of the truncated power basis and linear splines in dimensions higher than one is the difficulty in generalizing the B-spline basis to higher dimensions (Bakin, Hegland, and Osborne 2000).

A number of researchers have worked on models in which the covariate space is partitioned and then a separate model is fit in each of the regions. Tree models such as CART divide the space in a recursive fashion and fit local functions at the branches of the tree, with the simplest implementation involving locally constant functions. (See Chipman, George, and McCulloch (1998) for a Bayesian version of CART.) Other partitioning methods use more general approaches to dividing the space. Hansen and Kooperberg (2002) divide two-dimensional spaces into triangles and fit piecewise, continuous linear two-dimensional splines. Denison et al. (2002, Chapter 7) discuss partitioning models based on a Voronoi tessellation. Rasmussen and Ghahramani (2002) use a mixture of Gaussian processes in which each index point belongs to a single Gaussian process and the mixture is modelled as a Dirichlet process prior, but the individual GPs are not tied to disjoint regions from a partition of the space. A generalization of partitioning models allows for overlap between regions, with the regression function at a location being a weighted average of a set of functions. This gives a mixture-of-experts model, i.e., a mixture model in which the weights for the mixture vary with the covariates. Wood, Jiang, and Tanner (2002) take such an approach, using a mixture of smoothing splines, each with its own smoothing parameter, and weights that are multinomial probit functions of the covariates. The number of smoothing splines is chosen based on BIC, but the rest of the model is fit via MCMC. Note that the nonstationary covariance model of Fuentes (2001) and Fuentes and Smith (2001) shares the flavor of these partitioning and mixture models, as it performs locally-weighted averaging of stationary covariance models.

Of late, statisticians have intensively investigated the use of wavelet bases for regression functions, with the original classical approach to coefficient thresholding presented by Donoho and Johnstone (1995). Wavelet basis functions are localized functions, which, combined with nonlinear shrinkage of the coefficients, can model spatially inhomogeneous functions, including sharp jumps. Bayesian estimation of wavelet basis models involves placing priors on the coefficients, thereby incorporating the degree and nature of the thresholding into the Bayesian model (Vidakovic 1999). This has the same flavor as a formulation of the penalized splines model in which the basis function coefficients are shrunk by taking the coefficients to be a random effect (Wand 2003).

Neural networks have received much attention from machine learning researchers, with some work by statisticians as well (Lee 1998; Paige and Butler 2001). In particular, machine learning work on GPs intensified after Neal (1996) showed that a Bayesian formulation of neural network regression, based on a multilayer perceptron with a single hidden layer and particular choice of standard priors, converged to a Gaussian process prior on regression functions in the limit of infinitely many hidden units. A common form of the neural network model specifies the regression function to be

$$f(\boldsymbol{x}) = \beta_0 + \sum_{k=1}^{K} \beta_k g_k \left(\boldsymbol{u_k^T x} \right),$$

where the $g_k(\cdot)$ functions are commonly chosen to be logistic (sigmoid) functions and the $\boldsymbol{u_k}$ parameters determine the position and orientation of the basis functions. This is very similar to the multivariate linear splines model, except that Holmes and Mallick (2001) take $g_k(\cdot)$ to be the identity function. One drawback to fitting neural network models is the multimodality of the likelihood (Lee 1998).

Gaussian process models are closely related to a Bayesian formulation of regression using fixed basis functions. Consider the following Bayesian regression model,

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{k=1}^{K} b_k(\boldsymbol{x})\beta_k \\ \boldsymbol{\beta} &\sim \mathrm{N}\left(0, C_{\boldsymbol{\beta}}\right). \end{aligned}$$

This is equivalent to a Gaussian process prior model in function space with a prior covariance matrix, $C_{\boldsymbol{f}} = BC_{\boldsymbol{\beta}}B^T$ where B is the basis matrix composed of the $b_k(\cdot)$ functions. In other words, the basis chosen and the prior over the coefficients implies a Gaussian process prior for the function with a particular prior covariance. Changing the basis will of course change the prior covariance. Gibbs and MacKay (1997) show that basis function regression using an infinite number of radial basis functions (functions proportional to Gaussian densities) is equivalent to GP regression with a form of the squared exponential covariance (1.15). A Gaussian process regression model is equivalent to a Bayesian regression model with an infinite number of basis functions (Williams 1997). This can be seen by using the Karhunen-Loève expansion (Mercer's theorem) to expand the covariance function as a weighted sum of infinitely many eigenfunctions,

$$C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{k=1}^{\infty} \lambda_k g_k(\boldsymbol{x_i}) g_k(\boldsymbol{x_j}),$$

and taking the eigenfunctions to be the basis and the eigenvalues, λ_k, to be the variances of the coefficients. One can approximate the GP model by truncating the summation; using the eigenfunctions instead of another basis minimizes the MSE when approximating random functions with the truncation. (See Cohen and Jones (1969) for more details.) Machine learning researchers have called the basis function viewpoint the 'weight-space view' to contrast with the 'function-space view' of considering directly the Gaussian process (or other distribution) prior over functions (Williams 1997). Depending on the question, one of the approaches may be more instructive and/or computationally efficient. When the number of basis functions exceeds the number of observations, the GP approach is more computationally efficient, with the basis function approach more efficient in the opposite situation.

As we have seen, many methods can be seen as GP methods in which the covariance structure is implicit in the form of the model. The direct GP-based regression model takes the approach of explicitly modelling the covariance structure. Which approach is preferable will depend on computational convenience, the ease of using explicit covariance models as opposed to other parameterizations in which the covariance is implicit, and the extent to which different parameterizations fit data observed in practice. One of the goals of this thesis is to explore nonstationary covariance functions that may allow GP methods to better compete with nonparametric regression models with an implicit nonstationary covariance structure.

1.4.3 Smoothing as local linear averaging

Smoothing usually involves estimating the regression function as a local average of the observed data. The key issues determining the performance of a method are how the level of smoothing is chosen and whether the degree of smoothing is the same throughout the space. Smoothing involves the usual bias-variance tradeoff: more smoothing results in lower variance but higher bias, while less smoothing results in lower bias but higher variance. Many nonparametric regression methods can be seen as linear smoothers of the data where the estimate of the function at a finite set of values is

$$\hat{\boldsymbol{f}} = S\boldsymbol{y}$$

for some smoothing matrix S (also known as the hat matrix in the standard regression context). See Hastie, Tibshirani, and Friedman (2001) for an overview of various methods. The simplest linear smoothing method is nearest-neighbor averaging. Smoother estimates are produced by having the weights die off smoothly as a function of distance from the focal point, which gives kernel smoothing (local constant fitting), such as the original Nadaraya-Watson estimator,

$$\hat{f}(x_i) = \frac{\sum_{i=1}^{n} K_{\lambda}(x, x_i) y_i}{\sum_{i=1}^{n} K_{\lambda}(x, x_i)}.$$

Locally-weighted linear (e.g., loess) and polynomial regression reduce bias in various respects, but increase the variance of the estimator. For all these methods, the choice of the smoothing parameter is crucial.

The spline and Gaussian process methods can be seen to take the form of a linear smoother when conditioning on the knots or hyperparameters, respectively. For the Bayesian formulation of the regression spline model in DiMatteo (2001), conditional on the knots, the smoothing matrix is $S = n\eta^2 B(B^T B)^{-1} B^T$. Changing the knots changes the basis and therefore the smoothing matrix. For smoothing splines the smoothing matrix, which in the DiMatteo (2001) model is $S = \eta^2 B(B^T B + \lambda\Omega)^{-1} B^T$, is in the form of ridge regression or Bayesian regression with a prior over coefficients. Similarly, for the Gaussian process model, since $\hat{\boldsymbol{f}} = \mu_f + C_{\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}(\boldsymbol{y} - \mu_f)$, we see that if $\mu_f = 0$, we have

$$S = C_{\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}. \tag{1.17}$$

Note that we can always incorporate μ_f as a additive constant in $C_{\boldsymbol{f}}$ by integrating it out of the model. So, conditional on the covariance and noise variance parameters, the GP model is a linear smoother, and changing the covariance changes the smoothing matrix, in the same way that changing the knots changes the spline smoothing matrix. Green and Silverman (1994, p. 47) discuss in detail how spline smoothing can be seen as locally-weighted averaging of the observations. We can recover the implicit weights used in calculating the estimate for a given location from the rows of the smoothing matrix; this is a discrete representation of the smoothing kernel at the location. Nonstationary GP models will have smoothing kernels that change with location, much like adaptive kernel regression techniques (Brockmann, Gasser, and Herrmann 1993; Schucany 1995). The nonstationary GP models defined in this thesis have the advantage of being defined when there is more than one covariate, while work on adaptive kernel regression appears to have concentrated on the single covariate setting.

1.4.4 Modelling non-Gaussian data

Modelling non-Gaussian responses can, in principle, be done in similar fashion to the Gaussian regression problems discussed above. The model is

$$
\begin{aligned}
Y_i &\sim \mathrm{D}\left(g\left(f\left(\boldsymbol{x_i}\right)\right)\right) \\
f(\cdot) &\sim \mathrm{GP}\left(\mu_f, C_f\left(\cdot,\cdot;\boldsymbol{\theta}_f\right)\right),
\end{aligned}
$$

where D is an appropriate distribution function, such as the Poisson for count data or the binomial for binary data, and $g(\cdot)$ is an appropriate link function. This is a nonparametric version of the generalized linear models described in McCullagh and Nelder (1989) and is of essentially the same structure as generalized additive models (Hastie and Tibshirani 1990), except that the hidden function is a Gaussian process and the relationship of this function to the covariates is not additive. Diggle et al. (1998) define this model in the spatial context, while Christensen and Waagepetersen (2002) suggest MCMC sampling via the Langevin approach to speed mixing of the chain. Neal (1996), MacKay (1997), and Williams and Barber (1998), among others, have used such GP-based models to perform classification based on the binomial likelihood, while Neal (1997) used a t distribution likelihood to create a robust regression method. In Section 3.6.2, I discuss methods for improving the MCMC algorithm in cases such as this in which the function cannot be integrated out of the model, and in Section 4.6.4, I give an example of fitting non-Gaussian data. Biller (2000) and DiMatteo et al. (2002) use RJMCMC to fit generalized regression spline models for one-dimensional covariates.

1.5 Thesis Outline

This thesis is organized in the following fashion. I start by developing a general class of nonstationary correlation functions in Chapter 2 and presenting results on the smoothness of functions drawn from Gaussian process priors with these nonstationary correlation functions. In Chapter 3, I present the general methodology that takes advantage of these nonstationary correlation functions to perform smoothing, both in a spatial context and a regression context. This chapter goes into the details of model parameterization and fitting via Markov chain Monte Carlo (MCMC). In particular, I describe previous approaches to parameterization and fitting and provide a new MCMC sampling scheme, which I term posterior mean centering (PMC), that allows for faster mixing when the function cannot be integrated out of the model. I also discuss the numerical and computational issues involved in fitting the model. Chapter 4 presents the nonparametric regression model in more detail and assesses the performance of the nonstationary Gaussian process approach in comparison with spline-based methods on simulated and real datasets. Chapter 5 discusses the use of nonstationary covariance models to account for spatial correlation in analyzing time trends in a spatial dataset replicated in time. Finally, Chapter 6 gives an overview of the results of the thesis, discusses the contributions of the thesis, and presents areas for future work.

1.6 Contributions

This thesis makes the following original contributions:

- A class of closed-form nonstationary correlation functions, of which a special case is a nonstationary form of the Matérn correlation.

- Proof that the new nonstationary correlation functions, when embedded in a Gaussian process distribution, specify sample paths whose smoothness reflects the properties of the underlying stationary correlation function upon which the nonstationary correlation is constructed.

- A method, which I call posterior mean centering, for improving mixing of Markov chain Monte Carlo fitting of Gaussian process models when the unknown function cannot be integrated out of the model.

- A parameterization for a nonstationary Gaussian process nonparametric regression model and demonstration that the model can be used successfully in low-dimensional covariate spaces, albeit using a more computationally-intensive fitting process than competing methods.

- A hierarchical model, based on a nonstationary spatial covariance structure for the residuals, for making inference about linear trends at multiple spatial locations.

- A comparison of methods for, and demonstration of the difficulty involved in, fitting the covariance structure of replicated spatial data.

Chapter 2

Theoretical Development

2.1 Introduction

In this chapter I present a new class of nonstationary correlation functions and determine the smoothness properties of Gaussian processes (GPs) whose correlation functions lie in the class. The new class is a generalization of the kernel convolution covariance of Higdon et al. (1999). Next, I review results on the continuity and differentiability of sample paths from isotropic Gaussian processes based on the characteristics of the correlation functions of the GPs. I apply these results to the generalized kernel convolution correlation functions and show that they retain the smoothness properties of the isotropic correlation functions upon which they are based, provided that the underlying kernel structure is sufficiently smooth. I close by discussing some potential advantages of the generalized kernel convolution correlation.

2.2 Nonstationary Covariance Functions Using Convolutions of Kernels

In this section, I describe in detail the approach of Higdon et al. (1999) (henceforth HSK) for defining nonstationary covariance functions. HSK propose a nonstationary spatial covariance function, $C(\cdot,\cdot)$, defined by

$$C(\boldsymbol{x_i},\boldsymbol{x_j}) = \int_{\Re^2} K_{\boldsymbol{x_i}}(\boldsymbol{u})K_{\boldsymbol{x_j}}(\boldsymbol{u})d\boldsymbol{u}, \tag{2.1}$$

where $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, and $\boldsymbol{u}$ are locations in $\Re^2$, and $K_{\boldsymbol{x}}$ is a kernel function centered at $\boldsymbol{x}$. They motivate this construction as the covariance function of a white noise process, $\psi(\cdot)$, convolved with the kernel function to produce the process, $Z(\cdot)$, defined by

$$Z(\boldsymbol{x}) = \int_{\Re^2} K_{\boldsymbol{x}}(\boldsymbol{u})\psi(\boldsymbol{u})d\boldsymbol{u}.$$

One can avoid the technical details involved in carefully defining such a white noise process by using the definition of positive definiteness to show directly that the covariance function is positive definite in every Euclidean space, $\Re^p, p = 1, 2, \ldots$:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \int_{\Re^P} K_{\boldsymbol{x_i}}(\boldsymbol{u}) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= \int_{\Re^P} \sum_{i=1}^{n}\sum_{j=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) a_j K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= \int_{\Re^P} \sum_{i=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) \sum_{j=1}^{n} a_j K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= \int_{\Re^P} \left(\sum_{i=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) \right)^2 d\boldsymbol{u} \\
&\geq 0. \qquad (2.2)
\end{aligned}$$

The key to achieving positive definiteness is that each kernel is solely a function of its own location. Apart from this restriction, the structure of the kernel is arbitrary. I will return to this proof of positive definiteness when I generalize the HSK approach in Section 2.3.

Next I show the closed form of the HSK covariance for Gaussian kernels based on the equivalence of convolutions of densities with sums of independent random variables:

$$\begin{aligned}
C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \int K_{\boldsymbol{x_i}}(u) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= \int \frac{1}{(2\pi)^{\frac{P}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{u})^T \Sigma_i^{-1} (\boldsymbol{x_i} - \boldsymbol{u}) \right) \\
&\quad \times \frac{1}{(2\pi)^{\frac{P}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x_j} - \boldsymbol{u})^T \Sigma_j^{-1} (\boldsymbol{x_j} - \boldsymbol{u}) \right) d\boldsymbol{u}.
\end{aligned}$$

Recognize the expression as the convolution

$$\int h_{\boldsymbol{A}}(\boldsymbol{u} - \boldsymbol{x_i}) h_{\boldsymbol{U}}(\boldsymbol{u}) d\boldsymbol{u} = \int h_{\boldsymbol{A},\boldsymbol{U}}(\boldsymbol{u} - \boldsymbol{x_i}, \boldsymbol{u}) d\boldsymbol{u},$$

where $h(\cdot)$ is the normal density function, $\boldsymbol{A} \sim \mathrm{N}(\boldsymbol{0}, \Sigma_i)$, $\boldsymbol{U} \sim \mathrm{N}(\boldsymbol{x_j}, \Sigma_j)$, and $\boldsymbol{A}$ and $\boldsymbol{U}$ are independent. Now consider the transformation $\boldsymbol{W} = \boldsymbol{U} - \boldsymbol{A}$ and $\boldsymbol{V} = \boldsymbol{U}$, which has Jacobian of 1. This gives us the following equalities based on the change of variables:

$$\begin{aligned} \int h_{\boldsymbol{A},\boldsymbol{U}}(\boldsymbol{u} - \boldsymbol{x_i}, \boldsymbol{u}) d\boldsymbol{u} &= \int h_{\boldsymbol{W},\boldsymbol{V}}(\boldsymbol{u} - (\boldsymbol{u} - \boldsymbol{x_i}), \boldsymbol{u}) d\boldsymbol{u} \\ &= \int h_{\boldsymbol{W},\boldsymbol{V}}(\boldsymbol{x_i}, \boldsymbol{u}) d\boldsymbol{u} \\ &= h_{\boldsymbol{W}}(\boldsymbol{x_i}). \end{aligned}$$

Since $\boldsymbol{W} = \boldsymbol{U} - \boldsymbol{A}$, $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{x_j}, \Sigma_i + \Sigma_j)$ and therefore

$$\begin{aligned} C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= h_{\boldsymbol{W}}(\boldsymbol{x_i}) \\ &= \frac{1}{(2\pi)^{\frac{P}{2}} \mid \Sigma_i + \Sigma_j \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{x_j})^T (\Sigma_i + \Sigma_j)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}) \right). \end{aligned}$$

Absorbing the necessary constants into the matrices in the quadratic form and dividing by the standard deviation function, $\sigma(\boldsymbol{x_i}) = \frac{1}{2^{\frac{P}{2}} \pi^{\frac{P}{4}} |\Sigma_i|^{\frac{1}{4}}}$, we arrive at the nonstationary correlation function, $R(\cdot, \cdot)$, defined by

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{P}{2}} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2} \right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}) \right). \tag{2.3}$$

Examining the exponential and its quadratic form, we see that this is nothing but a squared exponential stationary correlation, but in place of the squared Mahalanobis distance, $\tau^2 = (\boldsymbol{x_i} - \boldsymbol{x_j})^T \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})$, for arbitrary fixed positive definite matrix Σ, we instead use a quadratic form with the average of the kernel matrices for the two locations. If the kernel matrices are constant, we recover the special case of the squared exponential correlation based on Mahalanobis distance. If they are not constant with respect to $\boldsymbol{x}$, the evolution of the kernel covariance matrices in space produces nonstationary covariance. To construct a covariance function, one merely includes a variance function.

Independently, Gibbs (1997, p. 49, equ. 3.82) derived a special case of the HSK covariance function in which the kernel matrices, Σ_i, are taken to be diagonal positive definite matrices. Gibbs (1997) makes an astute observation about the characteristics of the nonstationary covariance model that applies to the HSK covariance and to my generalization of HSK as well (Section 2.3). When the size of the kernels changes quickly, the resulting correlation structure can be counterintuitive because of the function, which Gibbs calls the 'prefactor', in front of the exponential in (2.3). When the kernels centered at $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are similar in size, the numerator and denominator more or less cancel out, but when one kernel is much larger than the other, the square root of the determinant in the denominator dominates the product of the fourth roots of the determinants in the numerator; this effect causes smaller correlation than achieved based solely on the exponential term. This is most easily seen graphically in a one-dimensional example in Figure 2.1, where I show $R(-0.5, x)$ (the correlation between the point -0.5 and all other points) and $R(0.5, x)$, when the kernel size changes drastically at $x = 0$. We see that the correlation of $x = 0.5$ with the points to its left drops off more quickly than the correlation of $x = -0.5$ with its neighboring points, because of the effect of the prefactor, even though the kernel centered at $x = 0.5$ is large, and the kernel centered at $x = -0.5$ is small. This is counter to intuition and to our goal for the nonstationary function because at certain distances, the correlation between two points whose kernels are relatively small is larger than the correlation between a point whose kernel is small and a point whose kernel is large. For this example, sample functions are least smooth at the x values where the kernel size changes quickly (Figure 2.1d), rather than being least smooth at the x values with the small kernels. This effect seems to be restricted to situations in which the kernel sizes change very quickly, so it may not be material in practice. However, the phenomenon may arise occasionally in sample paths in the regression modelling, as discussed in Section 4.6.1.

2.3 Generalized Kernel Convolution Covariance Functions

One potential drawback to the kernel convolution approach is that the HSK formulation using Gaussian kernels produces a nonstationary covariance with smoothness properties similar to the stationary squared exponential correlation (as shown in Section 2.5.5). In particular, if the kernel matrices vary sufficiently smoothly in the covariate space, then the sample paths based on the nonstationary covariance are infinitely differentiable. Stein (1999) discusses in detail why such highly smooth paths are undesirable and presents an asymptotic argument for using covariance functions in which the smoothness is allowed to vary. In hopes of avoiding such a high degree of smoothness, one might think of extending the HSK approach by using non-Gaussian kernels, but unless the convolution (2.1) can be done in closed form, this would entail numerical integration.

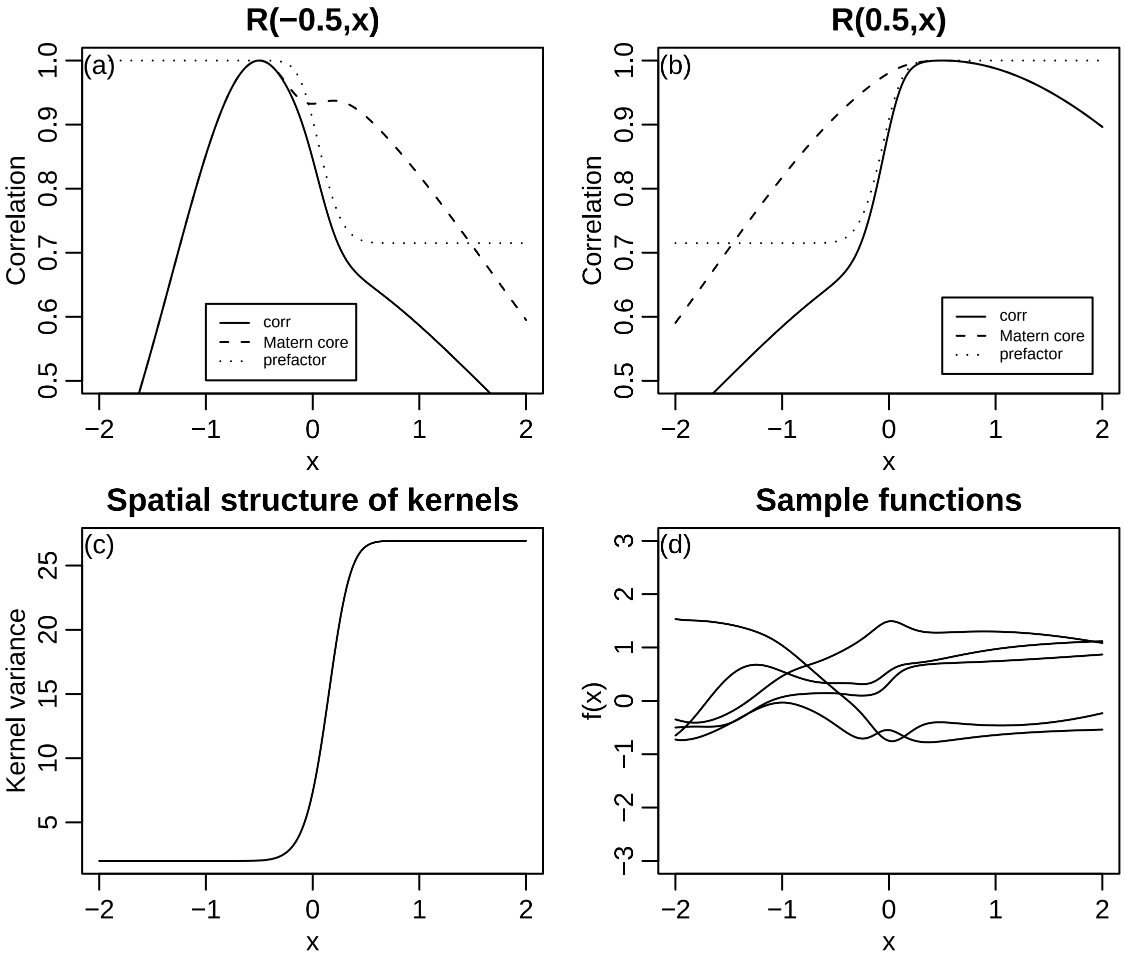

Figure 2.1. (a) Correlation of $f(-0.5)$ with the function at all other points. (b) Correlation of $f(0.5)$ with the function at all other points. (c) Kernel variance as a function of x. (d) Five sample functions drawn from the Gaussian process distribution; notice that the functions are least smooth at the location of the sharp change in the kernel size.

In this section I extend the HSK covariance in a way that provides a closed form correlation function. I produce a class of nonstationary correlation functions that provide more flexibility than the HSK formulation. Consider the quadratic form,

$$Q_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}), \tag{2.4}$$

at the heart of the correlation function (2.3) constructed via the kernel convolution. We have seen that the HSK nonstationary correlation function is nothing but the squared exponential correlation with this new quadratic form in place of a Mahalanobis distance. This relationship raises the possibility of producing a nonstationary version of an isotropic correlation function by using Q_{ij} in place of $\tau^2 = \boldsymbol{\tau}^T\boldsymbol{\tau}$ in the isotropic function. In practice, one uses $\frac{\tau}{\kappa} = \sqrt{Q_{ij}}$, since the scale parameter κ is redundant and can be absorbed into the kernel matrices, Σ_i and Σ_j, which are allowed to vary in size during the modelling anyway. The following general result applies in particular to correlation functions that are positive definite in Euclidean space of every dimension, in particular the power exponential, rational quadratic, and Matérn correlation functions (1.2-1.4).

Theorem 1 *If an isotropic correlation function, $R(\tau)$, is positive definite on $\Re^p$ for every $p = 1, 2, \ldots$, then the function, $R(\cdot,\cdot)$, defined by*

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} R\left(\sqrt{Q_{ij}}\right) \tag{2.5}$$

with $\sqrt{Q_{ij}}$ used in place of τ, is positive definite on $\Re^p$, $p = 1, 2, \ldots$, and is a nonstationary correlation function.

Proof: The proof is a simple application of Theorem 2 of Schoenberg (1938, p. 817), which states that the class of functions positive definite on Hilbert space is identical with the class of functions of the form,

$$R(\tau) = \int_0^\infty \exp\left(-\tau^2 s\right) dH(s), \tag{2.6}$$

where $H(\cdot)$ is non-decreasing and bounded and $s \geq 0$. The class of functions positive definite on Hilbert space is identical to the class of functions that are positive definite on $\Re^p$ for $p = 1, 2, \ldots$ (Schoenberg 1938). We see that the covariance functions in this class are scale mixtures of the squared exponential correlation. The underlying stationary correlation function with argument $\sqrt{Q_{ij}}$ can be expressed as

$$\begin{aligned}
R\left(\sqrt{Q_{ij}}\right) &= \int_0^\infty \exp\left(-Q_{ij}s\right) dH(s) \\
&= \int_0^\infty \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\frac{\Sigma_i}{s} + \frac{\Sigma_j}{s}}{2}\right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})\right) dH(s) \\
&= \int_0^\infty \int_{\Re^P} K_{\boldsymbol{x_i},s}(\boldsymbol{u}) K_{\boldsymbol{x_j},s}(\boldsymbol{u}) d\boldsymbol{u} dH(s).
\end{aligned}$$

Since s is non-negative, it becomes part of the kernel matrices, and the last expression can be seen to be positive definite based on (2.2).
Q.E.D.

This approach replaces the kernel at each location with a scale mixture of kernels where a common scale is used for all the locations (See Matérn (1986, pp. 32-33) for some discussion of generating new stationary correlation functions as scale mixtures of stationary correlation functions.) Using different distributions, H, for the scale parameter, S, produces different nonstationary correlation functions. A nonstationary version of the rational quadratic correlation function of the form (2.5) is

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{P}{2}} \left|\Sigma_i\right|^{\frac{1}{4}} \left|\Sigma_j\right|^{\frac{1}{4}}}{\left|\Sigma_i + \Sigma_j\right|^{\frac{1}{2}}} \left(\frac{1}{1+Q_{ij}}\right)^{\nu}.$$

This can be seen to be of the scale mixture form by taking $S \sim \Gamma(\nu, 1)$,

$$\int \exp(-Q_{ij}s)dH(s) = E(\exp(-Q_{ij}s)) = M_S(-Q_{ij}; \nu, 1) = \left(\frac{1}{1+Q_{ij}}\right)^{\nu},$$

where M_S is the moment generating function of S. This makes sense since the rational quadratic correlation function has the form of a t density, which is a mixture of Gaussians with an inverse gamma distribution for the variance of the Gaussian, which is proportional to $\frac{1}{S}$. A nonstationary version of the Matérn correlation function is

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{P}{2}} \left|\Sigma_i\right|^{\frac{1}{4}} \left|\Sigma_j\right|^{\frac{1}{4}}}{\left|\Sigma_i + \Sigma_j\right|^{\frac{1}{2}}} \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu Q_{ij}}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu Q_{ij}}\right). \tag{2.7}$$

Using an integral expression for the Bessel function (Gradshteyn and Ryzhik 1980, p. 340, equ. 9; McLeish 1982), one can easily show that in this case S is distributed inverse-gamma $(\nu, 1/4)$. In Section 2.5.4 (stationary) and Section 2.5.5 (nonstationary), I show that the existence of moments of S is directly related to the existence of mean square and sample path derivatives of processes whose covariance is produced by mixing a squared exponential covariance over the scale parameter. Rather than producing a closed form nonstationary correlation function by substituting the quadratic form (2.4) into an isotropic correlation function, one can instead construct nonstationary correlation functions (possibly without closed form) by choosing a distribution over the scale parameter, with the distribution chosen to produce the desired smoothness properties based on the moments of the distribution.

The Matérn correlation function is proportional to the Bessel density function (McLeish 1982). Based on the fact that convolutions of Bessel densities are also Bessel densities (McLeish 1982; Matérn 1986, pp. 29-30), we might expect that the Matérn nonstationary correlation (2.7) could be derived directly from a convolution of Bessel densities. However, the Bessel distribution is closed under convolution only for fixed scale parameters (this can be seen by multiplying two t densities, since the characteristic function of the Bessel density is proportional to a t density), so it does not directly correspond to a convolution of the type (2.1) for which $\Sigma_i \neq \Sigma_j$.

The quadratic form (2.4) defines a semi-metric space (Schoenberg 1938), in which the distance function is $(\boldsymbol{x_i'} - \boldsymbol{x_j'})^T(\boldsymbol{x_i'} - \boldsymbol{x_j'})$, where $\boldsymbol{x_i'} = \left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-\frac{1}{2}} \cdot \boldsymbol{x_i}$. However, the new location, $\boldsymbol{x_i'}$, varies depending on the other point, $\boldsymbol{x_j}$, through its dependence on Σ_j. The distance function violates the triangle inequality, even if one considers the points as lying in a higher dimensional space, so the space is not an inner-product space. To see this, consider a one-dimensional example with three points on a line, two points equidistant from the central point and on either side, $x_1 = -1, x_2 = 0, x_3 = 1$. Let the Gaussian kernel at the center point decay slowly along the line, $K_{x_2}(x) = \phi(x; -1, 3^2)$ while the two other Gaussian kernels decay more quickly along the line, $K_{x_1}(x) = \phi(x; -1, 1), K_{x_3}(x) = \phi(x; 1, 1)$. The distance between the central point and either side point is then 0.2, which smaller than half the distance, 4, between the two side points.

To construct the nonstationary correlation functions introduced here, we need kernels at all locations in the space $\mathcal{X}$. As described in Section 3.2, the kernels are modelled as functions of stochastic processes that determine the kernel eigenvectors and eigenvalues. This induces stochastic processes for the elements of the kernel matrices. As I will discuss in Section 2.5.5, the smoothness properties of these elements in part determine the smoothness of stochastic processes parameterized by the nonstationary correlation introduced here.

2.4 Nonstationary Covariance on the Sphere

The generalized kernel convolution covariance model can be extended for use on the sphere, S^2, and other non-Euclidean spaces. On the sphere, the equivalence of translation and rotation causes difficulty in defining kernels that produce correlation behavior varying with direction. The following recipe allows one to create a nonstationary model for the sphere. First, define a truncated

Gaussian kernel at each location in a Euclidean projection of the sphere centered at that location. Let the value of the kernel be zero for distances at which the Euclidean approximation to angular distance is poor. As usual, let the kernels vary smoothly in space. Project the kernels from the Euclidean projection back to the spherical domain, thereby defining a set of kernels on S^2, one kernel at each location of interest. Then define the correlation function as the convolution in the spherical domain. The key to showing positive definiteness (2.2) of the kernel convolution covariance is that each kernel is solely a function of the location of the kernel; this approach satisfies that condition. The additional integration over a scale parameter can also be done here in the spherical domain to produce a class of nonstationary correlation functions on the sphere. In practice, as described in Section 5.4.1.4, I have calculated the correlations directly in the Euclidean projections with untruncated kernels so as to be able to use the analytic form for the correlation (2.5). I found that the resulting correlation, although not guaranteed to be positive definite, does not cause numerical problems. Note that smoothness properties of processes on S^2 follow from those of processes on $\Re^3 \supset S^2$ if one chooses a covariance function that is positive definite on $\Re^3$ and sets $\tau_{ij} = 2\sin\left(\frac{\rho_{ij}}{2}\right)$, where ρ_{ij} is the angular distance between locations $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$.

There has been little work on nonstationary covariance modelling on the sphere apart from a nonlinear mapping approach (Das 2000) that extended the work of Sampson and Guttorp (1992). Das (2000) mapped the original sphere to a new sphere in which stationarity is assumed to hold and then used stationary covariance models valid on S^2.

2.5 Smoothness Properties of Covariance Functions

2.5.1 Overview

The functional form and parameter values of the covariance function of a Gaussian process distribution determine the smoothness properties of the process and sample paths drawn from the distribution. Covariance functions can give Gaussian processes whose sample paths range from discontinuous to analytic. While data can inform the choice of covariance function to some degree, this decision is also a philosophical choice based on one's conception of the underlying physical or scientific process.

Two important characteristics of stochastic processes are mean square properties and sample path properties, which I define in Section 2.5.2. All of the Gaussian processes I consider here are both mean square and sample path continuous based on simplifications of the arguments given here. The key difference amongst correlation functions lies in the differentiability properties associated with them. Using results from the stochastic process literature, in Section 2.5.4 I derive the mean square and sample path differentiability properties of stochastic processes parameterized by the stationary correlation functions (1.2-1.4). These results are well-known but are not collected or proven in one place to my knowledge. In particular, sample path properties of familiar isotropic correlation functions are relatively little discussed (but see Abrahamsen (1997)). I present the material here because the smoothness properties associated with the nonstationary kernel convolution correlation functions (Section 2.5.5) follow from those of the underlying isotropic correlation functions on which they are based. I focus particularly on sample path properties, because I believe these are most relevant when selecting a correlation function. The analyst is more likely to be able to make some intuitive judgement about sample path properties of the process at hand than about mean square properties. Mean square properties are easier to derive, being directly related to derivatives of the covariance function and moments of the spectral distribution, and much of the literature concentrates on these (e.g., Stein (1999)). Even if one is not directly interested in mean square properties, they are useful as a first step in determining sample path properties, as we will see. At times hereafter I refer to mean square and sample path properties of a correlation function, by which I mean properties of mean-zero stochastic processes with the given correlation function. For sample path properties, the results hold only for Gaussian processes.

2.5.2 Theoretical framework

First I give brief definitions of mean square and sample path continuity and differentiability, following Adler (1981). In the remaining sections of this chapter, I will use $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{u}$ to indicate locations in the covariate space and x_p to indicate the pth scalar element of $\boldsymbol{x}$. In the stationary case, let $\boldsymbol{\tau} = \|\boldsymbol{x} - \boldsymbol{y}\|$, and in the isotropic case, let $\tau = \|\boldsymbol{\tau}\| = \sqrt{\boldsymbol{\tau}^T\boldsymbol{\tau}}$. This avoids having double subscripts indicating both location and covariate. However as a notational exception for the following definition, let $\boldsymbol{x_i}, i = 1, 2, \ldots$ be a sequence of locations such that $\|\boldsymbol{x_i} - \boldsymbol{x}\| \rightarrow 0$

as $i \rightarrow \infty$. If $Z(\boldsymbol{x_i}) \overset{m.s.}{\rightarrow} Z(\boldsymbol{x})$ as $i \rightarrow \infty$, then $Z(\cdot)$ is continuous in mean square at $\boldsymbol{x}$. If the convergence is almost sure, then $Z(\cdot)$ is almost surely continuous at $\boldsymbol{x}$. If there exists $Z_p^{(1)}(\boldsymbol{x})$ such that

$$\frac{Z(\boldsymbol{x} + \epsilon \boldsymbol{u_p}) - Z(\boldsymbol{x})}{\epsilon} \overset{m.s.}{\rightarrow} Z_p^{(1)}(\boldsymbol{x}), \epsilon \rightarrow 0$$

where $\boldsymbol{u_p}$ is the unit vector in the pth direction, then $Z_p^{(1)}(\boldsymbol{x})$ is the mean square derivative of $Z(\cdot)$ at $\boldsymbol{x}$. Again if the convergence is almost sure, then $Z_p^{(1)}(\boldsymbol{x})$ is the pth-order almost sure partial derivative at $\boldsymbol{x}$. Mth-order partial derivatives, either mean square or almost sure, can be defined in similar fashion. If almost sure continuity or differentiability hold simultaneously with probability one for all $\boldsymbol{x} \in I \subset \Re^P$ then $Z(\cdot)$ is sample path continuous or has a sample path partial derivative, respectively, on I. Sample path differentiability involves the existence and continuity of sample path partial derivatives, as I will discuss shortly.

In general, because the finite dimensional distributions of a stochastic process do not determine the sample path properties of the process, showing sample path continuity or differentiability relies on the notion of separability pioneered by Doob (1953, pp. 51-53) and discussed in detail in Gihman and Skorohod (1974, p. 164) and Adler (1981, p. 14). A separable process is one for which the finite-dimensional distributions determine the sample path properties. By virtue of Theorem 2.4 of Doob (1953, p. 57), for any stochastic process, $\tilde{Z}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}$ with $\mathcal{X}$ a linear space, there exists a version of the process, $Z(\boldsymbol{x})$, that is separable and is stochastically equivalent to the original process. Once one assumes that one is working with the separable version of the stochastic process, almost sure continuity (differentiability) can be extended to sample path continuity (differentiability) because the probability one statement at individual points holds simultaneously on a dense countable set of points, and the sample path properties of the separable process are determined by the properties of the process on the dense countable set. From this point forward, I will assume all processes are separable.

Mean square properties of correlation functions are frequently analyzed, in part because they can be readily determined from the correlation function, or for stationary correlation functions, from the spectral representation of the correlation function. A process is mean square continuous at $\boldsymbol{u}$ if and only if the covariance function $C(\boldsymbol{x}, \boldsymbol{y})$ is continuous at $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{u}$ (Cramér and Leadbetter 1967, p. 83; Loève 1978, p. 136; Adler 1981, p. 26). All the correlation functions considered in this work are continuous, and therefore processes with these correlation functions are mean square continuous. Mean square differentiability is directly related to the existence of derivatives of the correlation function. A random field, $Z(\cdot)$, has mean square partial derivative at $\boldsymbol{u}$, $Z_p^{(1)}(\boldsymbol{u})$ if and only if $\partial^2 C(\boldsymbol{x}, \boldsymbol{y})/(\partial x_p \partial y_p)$ exists and is finite at $(\boldsymbol{u}, \boldsymbol{u})$ (Adler 1981, p. 27). If the derivative exists, the covariance function of the mean square partial derivative process is the partial derivative of the original covariance function. For stationary processes, one need only consider the partial derivative evaluated at $\mathbf{0}$, and the correlation function of the partial derivative process is $\partial^2 C(\boldsymbol{\tau})/\partial \tau_p^2$. The existence of Mth-order mean square partial derivatives is equivalent to the finiteness of the relevant $2M$th-order partial derivatives of the covariance function (Adler 1981, p. 27; Vanmarcke 1983, p. 111),

$$\frac{\partial^{2M} C(\boldsymbol{x}, \boldsymbol{y})}{\partial x_{p_1} \cdots \partial x_{p_M} \partial y_{p_1} \cdots \partial y_{p_M}},$$

evaluated at $(\boldsymbol{u}, \boldsymbol{u})$ for $p_m \in \{1, \dots, P\}, m \in \{1, \dots, M\}$.

Stationary covariance functions can be expressed as the Fourier transform of the spectral distribution, $H(\cdot)$,

$$C(\boldsymbol{\tau}) = \int_{\Re^P} \exp(i\boldsymbol{w}^T \boldsymbol{\tau}) dH(\boldsymbol{w}).$$

If $H(\cdot)$ is absolutely continuous with respect to Lebesgue measure, then the spectral density, $h(\cdot)$, exists and can be expressed as

$$h(\boldsymbol{w}) = \frac{1}{(2\pi)^P} \int_{\Re^P} \exp(-i\boldsymbol{w}^T \boldsymbol{\tau}) C(\boldsymbol{\tau}) d\boldsymbol{\tau}.$$

In other words, the covariance function is the characteristic function of the distribution, $H(\cdot)$. In the isotropic case, $H(\boldsymbol{w})$ depends only on $\|\boldsymbol{w}\|$ (Adler 1981, p. 35). Using the well-known relationship between derivatives of a characteristic function evaluated at the origin and moments of the distribution, one can see that mean square differentiability is equivalent to the existence of moments of the spectral distribution. A process on $\Re^1$ is mean square differentiable if and only if

$$\int w^2 dF(w) < \infty,$$

because the existence of the second moment is equivalent to having two derivatives of the covariance at the origin (Stein 1999, p. 27). As mentioned above, in higher dimensions, the existence of a $2M$th-order partial derivative of the covariance at the origin is equivalent to having the respective Mth-order mean square partial derivative exist. This is also equivalent to the existence of the $2M$th-order spectral moments (Adler 1981, p. 31),

$$(-1)^M \left. \frac{\partial^{2M} C(\boldsymbol{\tau})}{\partial \tau_{p_1}^2 \cdots \partial_{p_M}^2} \right|_{\boldsymbol{\tau}=\mathbf{0}} = \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH(\boldsymbol{w}) < \infty, \tag{2.8}$$

for $p_m \in \{1, \ldots, P\}, m \in \{1, \ldots, M\}$. While the spectral relationship is very useful for assessing mean square differentiability in the stationary case, it does not provide the covariance function of the mean square derivative, only the value of the covariance at $\mathbf{0}$.

Mean square properties are important in part because they are useful for showing sample path properties. In the discussion that follows in this paragraph, except where noted, the cited results are for processes on $\Re^1$. While I have not seen the results shown formally on $\Re^P$, I presume they hold there as well. Cambanis (1973, Theorem 6) has shown that a real, separable, measurable Gaussian process that is not mean square differentiable at any point has with probability one paths that are almost nowhere differentiable. Since mean square differentiability is generally straightforward to determine, the more difficult cases involve showing sample path differentiability for processes that are mean square differentiable. Doob (1953, p. 536) shows that for a separable process, sample functions of the process are absolutely continuous, and hence the functions have derivatives almost everywhere. Furthermore, the mean square derivative process is equal to the sample path derivative process with probability one (Doob 1953, p. 536; Cramér and Leadbetter 1967, p. 85; Yaglom 1987, p. 67). In the Gaussian case, on $\Re^P$, the derivative processes are also Gaussian processes, and the joint distributions of all of these processes are Gaussian (Adler 1981, p. 32). For a function of P variables, if all first-order partial derivatives, $Z_p^{(1)}(\cdot), p = 1, \ldots, P$, exist and are continuous, the function is continuously differentiable and this is sufficient for the function to be first-order differentiable (Olmsted 1961, p. 267; Leithold 1968, pp. 795-796). Since the partial derivatives are themselves functions on $\Re^P$, higher-order derivatives are defined recursively by differentiating the lower-order derivative functions. I demonstrate Mth-order sample path differentiability by showing that all Mth-order partial derivative processes, $Z_{p_1 \cdots p_M}^{(M)}(\cdot)$ for $p_m \in \{1, \ldots, P\}, m \in \{1, \ldots, M\}$ exist and are sample path continuous.

Sample path continuity is difficult to demonstrate in the non-Gaussian case. Adler (1981, p.

48) gives the condition for a stationary process that for $\alpha > 0$ and $\epsilon > \alpha$, if

$$E|Z(\boldsymbol{x}+\boldsymbol{\tau}) - Z(\boldsymbol{x})|^{\alpha} \leq \frac{c\|\boldsymbol{\tau}\|^{2P}}{|\log\|\boldsymbol{\tau}\||^{1+\epsilon}},$$

then $Z(\cdot)$ will be sample path continuous over any compact set in $\Re^{P}$. In the Gaussian case, the conditions are less strict. If $Z(\cdot)$ is a zero-mean Gaussian process with continuous covariance and for some finite $c > 0$ and some $\epsilon > 0$,

$$E|Z(\boldsymbol{x}) - Z(\boldsymbol{y})|^{2} \leq \frac{c}{|\log\|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon}} \tag{2.9}$$

for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I, then the process has continuous sample paths on I (Adler 1981, p. 60). In the stationary case, the condition simplifies to

$$C(\boldsymbol{0}) - C(\boldsymbol{\tau}) \leq \frac{c}{|\log\|\boldsymbol{\tau}\||^{1+\epsilon}}. \tag{2.10}$$

Adler (1981, p. 64) also gives a condition for sample path continuity for stationary, zero-mean Gaussian processes based on the spectral representation of the covariance. If for some $\epsilon > 0$,

$$\int_{\Re^{P}} |\log(1+\|\boldsymbol{w}\|)|^{1+\epsilon} dH(\boldsymbol{w}) < \infty,$$

then the process is sample path continuous. However, I have not been able to use this spectral condition to demonstrate the continuity of derivative processes.

To summarize, the steps involved in proving sample path differentiability for Gaussian processes are as follows. I focus only on correlation functions in the sections that follow, assuming that the variance function is either constant or has sample paths as smooth as those based on the correlation function. First I show that Mth-order mean square differentiability holds (using either the derivative of the correlation or moments of the spectral distribution) and determine the covariance of the Mth-order mean square partial derivative processes (using the derivatives of the correlation). These mean square derivative processes are probabilistically equivalent to the sample path derivative processes. Then I show that all the Mth-order derivative processes are sample path continuous based on their covariance functions and either condition (2.9) or (2.10).

Processes that are infinitely mean square differentiable may also be mean square analytic. Loève (1978, p. 137) and Stein (1999, p. 33) state that processes on $\Re^{1}$ are mean square analytic if the covariance function $C(x,y)$ is analytic at (u,u). Both the squared exponential and rational quadratic correlation functions (despite its having a second parameter) are analytic, as I show in Section 2.5.4.4 by demonstrating complex differentiability of the correlation function. This suggests that the sample paths for processes with these correlation functions will be analytic, although I have not seen this result proven. Stein (1999) argues that analytic sample paths are unrealistic for physical processes, since an analytic function is fully determined by its values in a small interval.

2.5.3 Lemmas for proofs

Before venturing into the details of the smoothness properties of Gaussian processes based on correlation function properties, I provide definitions and lemmas that I will use in the proofs. I state the lemmas here for $x_p - y_p$ and $\|\boldsymbol{x} - \boldsymbol{y}\|$ but in the stationary setting, I use $\tau_p = x_p - y_p$ and $\|\boldsymbol{\tau}\| = \|\boldsymbol{x} - \boldsymbol{y}\|$.

First I introduce some notation involving partial derivatives, which will clarify the arguments involved in the remainder of this chapter. The classes $\mathcal{D}^{(m)}(\cdot)$ are defined recursively as follows. Let

$$\mathcal{D}^{(1)}(\cdot) = \left\{ \frac{\partial(\cdot)}{\partial x_1}, \ldots, \frac{\partial(\cdot)}{\partial x_P}, \frac{\partial(\cdot)}{\partial y_1}, \ldots, \frac{\partial(\cdot)}{\partial y_P} \right\}.$$

I will use $D^{(1)}(\cdot) \in \mathcal{D}^{(1)}(\cdot)$ to represent a term taking the form of an element in the class. Then for $m > 1$ let $\mathcal{D}^{(m)}(\cdot) = \left\{ D^{(1)}(D^{(m-1)}(\cdot)) \right\}$. For $m = 2$ we have

$$\mathcal{D}^{(2)}(\cdot) = \left\{ \frac{\partial^2(\cdot)}{\partial x_1^2}, \frac{\partial^2(\cdot)}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^2(\cdot)}{\partial x_1 \partial y_P}, \ldots, \frac{\partial^2(\cdot)}{\partial y_P^2} \right\}.$$

To denote partial derivatives with respect to $\boldsymbol{x}$ only, I use $D_{\boldsymbol{x}}^{(m)}$ and for particular partial derivatives with respect to a coordinate of $\boldsymbol{x}$ I use $D_{x_i}^{(m)}$.

Definition 2 *I define $g(\boldsymbol{x}, \boldsymbol{y})$ to be $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$, denoted $g(\boldsymbol{x}, \boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$, if there exists $c > 0$ such that for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I,*

$$\frac{g(\boldsymbol{x}, \boldsymbol{y})}{\|\boldsymbol{x} - \boldsymbol{y}\|^a} \leq c.$$

Schervish (1995, p. 394) provides an alternate definition, with more extensive notation and related properties, based on a sequence of numbers, r_n, replacing $\|\boldsymbol{x} - \boldsymbol{y}\|$ and the region I. Some properties that will be useful are

- *If* $g_1(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$ *and* $g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_2})$ *then* $g_1(\boldsymbol{x},\boldsymbol{y}) \cdot g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1+a_2})$*.*

- *If* $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$ *then* $g(\boldsymbol{x},\boldsymbol{y})^{a_2} = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1 a_2})$*.*

- *If* $g_1(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$ *and* $g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_2})$*, with* $a_1 \leq a_2$*, then* $g_1(\boldsymbol{x},\boldsymbol{y}) + g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$*.*

Next I prove a series of lemmas.

Lemma 3 $x_p - y_p = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$

Proof: First, the claim is equivalent to $(x_p - y_p)^2 = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$ by the properties of $O_I(\cdot)$. The following bound holds:

$$\frac{(x_p - y_p)^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^2} = \frac{(x_p - y_p)^2}{\sum_q (x_q - y_q)^2} \leq 1.$$

Q.E.D.

Lemma 4 *If* $D^{(1)}(g(\boldsymbol{x},\boldsymbol{y}))$ *exists, then* $g(\boldsymbol{x},\boldsymbol{x}) - g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$*.*

Proof: Consider

$$\frac{g(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{x},\boldsymbol{x})}{\|\boldsymbol{x}-\boldsymbol{y}\|}. \tag{2.11}$$

By standard results in advanced calculus texts, such as Buck (1965, p. 243) or Leithold (1968, p. 795), if the function $g(\cdot)$ is a differentiable function of $\boldsymbol{y}$, then we can express the numerator as

$$g(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{x},\boldsymbol{x}) = D^{(1)}_{y_1}(g(x,y))(x_1 - y_1) + \cdots + D^{(1)}_{y_P}(g(x,y))(x_P - y_P) + R(x_1 - y_1, \ldots, x_P - y_P),$$

where

$$\lim_{\|\boldsymbol{x}-\boldsymbol{y}\| \to 0} \frac{R(x_1 - y_1, \ldots, x_P - y_P)}{\|\boldsymbol{x}-\boldsymbol{y}\|} = 0.$$

Therefore the expression (2.11) can be bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I based on the above limit and on lemma 3.

Q.E.D.

Lemma 5 *If* $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$ *for* $a > 0$*, then* $g(\boldsymbol{x},\boldsymbol{y}) |\log \|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon}$ *is bounded for* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in a region* I*.*

Proof:

$$\begin{aligned} g(\boldsymbol{x},\boldsymbol{y})|\log\|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon} &= \frac{g(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|^a}\|\boldsymbol{x}-\boldsymbol{y}\|^a|\log\|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon} \\ &= \frac{g(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|^a}\left|\|\boldsymbol{x}-\boldsymbol{y}\|^{\frac{a}{1+\epsilon}}|\log\|\boldsymbol{x}-\boldsymbol{y}\||\right|^{1+\epsilon}. \end{aligned}$$

The fraction is bounded by the assumption that $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$. The term inside the outer absolute value has limit of 0 as $\|\boldsymbol{x}-\boldsymbol{y}\| \to 0$ by Abramowitz and Stegun (1965, p. 68,equ. 4.1.31) and hence is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I, and the constant power, $1+\epsilon$, does not affect the boundedness.

Q.E.D.

Lemma 6 *If the kernel matrix elements are continuous,* $Q_{xy} = (\boldsymbol{x}-\boldsymbol{y})^T\left(\frac{\Sigma_x+\Sigma_y}{2}\right)^{-1}(\boldsymbol{x}-\boldsymbol{y})$ *is* $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$. *If the kernel matrix elements are once sample path differentiable, then* $D^{(1)}(Q_{xy})$ *is* $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. *If the kernel matrix elements are* m *times sample path differentiable, then* $D^{(m)}(Q_{xy}) = O_I(1)$.

Proof: First consider Q_{xy} and absorb the divisor of 2 in the matrix inverse into the kernel matrices. By continuity of the kernel matrix elements, the elements of $\Sigma_x + \Sigma_y$ are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I. Expressing $Q_{xy} = \sum_i\sum_j(x_i-y_i)(x_j-y_j)(\Sigma_x+\Sigma_y)^{-1}_{ij}$, it is clear that $Q_{xy} = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$. Next, without loss of generality, consider the first partial derivative $D^{(1)}_{x_i}(Q_{xy}) = c_1\sum_p(x_p-y_p)(\Sigma_x+\Sigma_y)^{-1}_{ip} + (\boldsymbol{x}-\boldsymbol{y})^T D^{(1)}_{x_i}(\Sigma_x+\Sigma_y)^{-1}(\boldsymbol{x}-\boldsymbol{y})$. The first term is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$, assuming the kernel matrix elements are continuous. For the second term,

$$\frac{\partial(\Sigma_x+\Sigma_y)^{-1}}{\partial x_i} = (\Sigma_x+\Sigma_y)^{-1}\frac{\partial(\Sigma_x+\Sigma_y)}{\partial x_i}(\Sigma_x+\Sigma_y)^{-1},$$

so the second term is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$ provided the kernel matrix elements are once differentiable. Hence the sum is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. Finally consider

$$D^{(m)}(Q_{\boldsymbol{xy}}) = c_1 D^{(m-2)}\left((\Sigma_x+\Sigma_y)^{-1}\right) + \ldots + c_2(\boldsymbol{x}-\boldsymbol{y})^T D^{(m)}\left((\Sigma_x+\Sigma_y)^{-1}\right)(\boldsymbol{x}-\boldsymbol{y}).$$

All the terms are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I if the kernel matrix elements are at least m times sample path differentiable, so the sum is $O_I(1)$.

Q.E.D.

Lemma 7 *If $ES^{2M} < \infty$, one can interchange differentiation and integration of*

$$\frac{\partial}{\partial Q} \int S^{2M-m} \exp(-Qs) dH(s)$$

for $m \in \{1, \ldots, 2M\}$.

Proof: Corollary 2.4.1 of Casella and Berger (1990, p. 71) states that to interchange differentiation and integration, it is sufficient that for some $\epsilon > 0$, there exists a function $g_0(Q, s)$, integrable with respect to s, such that

$$\left| \frac{\partial g(Q, s)}{\partial Q} |_{Q=Q_0} \right| \leq g_0(Q, s), \forall Q_0 \; s.t. \; |Q_0 - Q| \leq \epsilon$$

where the dominating function $g_0(Q, s)$ is integrable with respect to s. In this case, we have

$$\begin{aligned} \left| \frac{\partial}{\partial Q} g(Q, s)|_{Q=Q_0} \right| &= \left| \frac{\partial}{\partial Q} s^{2M-m} \exp(-Qs)|_{Q=Q_0} \right| \\ &= s^{2M-m+1} \exp(-Q_0 s) \\ &\leq s^{2M-m+1} \\ &= g_0(Q, s), \end{aligned}$$

where the function g_0 does not involve Q and hence is integrable for all Q for $m \in \{1, \ldots, 2M\}$ with respect to $H(s)$ by the assumption that $ES^{2M} < \infty$. Note that differentiating with respect to a function, $a(Q)$ merely introduces the multiplicative factor $\partial Q / \partial a(Q)$, which is constant with respect to s.

Q.E.D.

Lemma 8 *Elements of the kernel matrices, $\Sigma_{\boldsymbol{x}}$, are bounded for $\boldsymbol{x}$ in a region I if and only if the eigenvalues of the kernel matrices are bounded in I. Furthermore, if the eigenvalues and the eigenvector elements are sample path differentiable, the matrix elements are sample path differentiable.*

Proof: Consider the spectral decomposition of a kernel matrix $\Sigma_{\boldsymbol{x}} = \Gamma_{\boldsymbol{x}} \Lambda_{\boldsymbol{x}} \Gamma_{\boldsymbol{x}}^T$. Element-wise, and suppressing the dependence on $\boldsymbol{x}$, this gives us

$$\Sigma_{ij} = \sum_k \Gamma_{ik} \Gamma_{jk} \Lambda_{kk}.$$

Since the eigenvectors are normalized vectors, their elements are bounded. Therefore the elements of $\Sigma_{\boldsymbol{x}}$ are bounded if the eigenvalues, Λ_{kk}, are bounded. Equivalently, $\Lambda_{\boldsymbol{x}} = \Gamma_{\boldsymbol{x}}^T \Sigma_{\boldsymbol{x}} \Gamma_{\boldsymbol{x}}$, so the eigenvalues can be expressed as products involving the kernel matrix elements and must be bounded if the matrix elements are bounded. Finally, based on the expansion of $\Sigma_{\boldsymbol{x},ij}$ in terms of sums and products of the eigenvector elements and eigenvalues, it's clear that if the eigenvector elements and eigenvalues are sample path differentiable, the kernel matrix elements will be as well. Q.E.D.

2.5.4 Smoothness properties of isotropic correlation functions

The stationary correlation functions on which I focus in this work are all positive definite on $\Re^p, p = 1, 2, \ldots$, and can therefore be expressed using the Schoenberg (1938) representation (2.6) as a scale mixture of squared exponential correlation functions. In this section I use the representation to show the differentiability properties of the correlation functions and in Section 2.5.5 I do the same for the counterpart nonstationary correlation functions. I show that the smoothness properties of the scale mixture correlation functions are directly related to the existence of moments of the scale parameter.

In the isotropic case, for mean square differentiability, I use the condition on the spectral representation (2.8), to show that if M moments of the scale parameter exist then the Mth-order mean square partial derivatives exist (Section 2.5.4.1). Next, to assess sample path differentiability, I work with the derivatives of the correlation function, as I have not been able to make progress using the spectral representation. Unfortunately, using the derivatives of the correlation function, I am only able to show that the existence of $2M$ moments of the scale parameter is sufficient for Mth-order sample path differentiability (Section 2.5.4.2). In Table 2.1 I give an overview of the smoothness properties of the correlation functions I have been discussing.

2.5.4.1 Mean square differentiability and scale mixtures

Theorem 9 *A stochastic process, $Z(\cdot)$, with isotropic correlation function that can be expressed in the Schoenberg (1938) representation (2.6) has Mth-order mean square partial derivatives if M*

Table 2.1. Smoothness properties of Gaussian processes parameterized by various correlation functions. The asterisk indicates that the sample path part of this statement is a conjecture. In Section 2.5.4.2 I prove only that the Matérn is $\lceil \frac{\nu}{2} - 1 \rceil$ times sample path differentiable

	Smoothness Properties	
	Mean square and sample path	Mean square
Correlation form	differentiability	analyticity
Power exponential, $\nu < 2$	no	no
Matérn	$\lceil \nu - 1 \rceil$ times*	no
Squared exponential	infinitely	yes
Rational quadratic	infinitely	yes

moments of the scale parameter, S, are finite.

Proof: Adler (1981, p. 31) gives the general relationship between the existence of spectral moments and mean square differentiability in the stationary setting. The Mth-order mean square partial derivatives exist if

$$(-1)^M \left. \frac{\partial^{2M} R(\boldsymbol{\tau})}{\partial \tau_{p_1}^2 \cdots \partial \tau_{p_M}^2} \right|_{\boldsymbol{\tau}=\mathbf{0}} = \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) < \infty,$$

where $p_m \in \{1, \ldots, P\}, m \in \{1, \ldots, M\}$. So to show the existence of the Mth-order mean square derivative, we need to consider the $2M$th-order moments of the spectral density. Expressed in terms of the Schoenberg (1938) representation (2.6), these are

$$\begin{aligned} \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) &= \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 h_W(\boldsymbol{w}) d\boldsymbol{w} \\ &= \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 \int_{\Re^P} \exp(-i\boldsymbol{w}^T\boldsymbol{\tau}) R(\boldsymbol{\tau}) d\boldsymbol{\tau} \\ &\propto \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 \int_{\Re^P} \int \exp(-i\boldsymbol{w}^T\boldsymbol{\tau}) \exp(-\boldsymbol{\tau}^T\boldsymbol{\tau} s) dH_S(s) d\boldsymbol{\tau} d\boldsymbol{w}. \end{aligned}$$

First, interchange the order of integration with respect to s and $\boldsymbol{\tau}$ by Fubini's theorem, which is justified because the exponential functions are bounded. Recognize that the integral with respect to $\boldsymbol{\tau}$ is in the form of a normal density with $\boldsymbol{\tau} \sim \mathrm{N}_P\left(-\frac{i\boldsymbol{w}}{2s}, \frac{1}{2s}I\right)$. This gives us

$$\int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) \quad \propto \quad \int_{\Re^P} \int w_{p_1}^2 \cdots w_{p_M}^2 \left(\frac{1}{2s}\right)^{\frac{P}{2}} \exp\left(-\frac{\boldsymbol{w}^T\boldsymbol{w}}{4s}\right) dH_S(s) d\boldsymbol{w}.$$

Once again interchange the order of integration by Fubini's theorem, which is justified because the integrand is non-negative, and the next steps will show that it is integrable. We see that the integral with respect to $\boldsymbol{w}$ takes the form of a product of moments with respect to $\boldsymbol{w} \sim N_P(0, 2sI)$. A straightforward calculation shows that $E\left(W_{p_1}^2 \cdots W_{p_M}^2\right) \propto S^M$, which gives us

$$\begin{aligned} \int_{\Re^p} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) &\propto \int s^M dH_S(s) \\ &= E\left(S^M\right). \end{aligned}$$

So we see that the Mth-order mean square partial derivatives exist if the scale parameter, S, has M moments.

Q.E.D.

2.5.4.2 Sample path differentiability and scale mixtures

I next show that the existence of $2M$ moments of S is sufficient for Mth-order sample path differentiability.

Theorem 10 *A Gaussian process, $Z(\cdot)$, with isotropic correlation function that can be expressed in the Schoenberg (1938) representation (2.6), is Mth-order sample path differentiable if $2M$ moments of the scale parameter, S, are finite.*

Proof: To evaluate the condition (2.10) for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I and thereby show continuity of the derivative processes, we need to calculate the covariance functions of the Mth-order derivative processes, which by Adler (1981, p. 27) are of the form

$$(-1)^M \frac{\partial^{2M} R(\boldsymbol{\tau})}{\partial\tau_{p_1}^2 \cdots \partial\tau_{p_M}^2}.$$

If $ES^{2M} < \infty$, we can interchange differentiation and integration when the term inside the integral is of order $S^m, m < 2M$ by lemma 7. The $2M$th partial derivative of a correlation function of the form (2.6) is

$$\begin{aligned} & c_M \int s^M \exp\left(-\boldsymbol{\tau}^T \boldsymbol{\tau} s\right) dH(s) \\ & + c_{M+1} \left(\sum_{i,j} \tau_{p_i} \tau_{p_j} \right) \int s^{M+1} \exp\left(-\boldsymbol{\tau}^T \boldsymbol{\tau} s\right) dH(s) \\ & + \cdots + c_{2m} \tau_{p_1}^2 \cdots \tau_{p_M}^2 \int s^{2M} \exp\left(-\boldsymbol{\tau}^T \boldsymbol{\tau} s\right) dH(s) \end{aligned} \qquad (2.12)$$

for $i, j \in \{1, \ldots, M\}$. To evaluate the condition (2.10) we need to consider the boundedness of

$$(-1)^M \left. \frac{\partial^{2M} R(\boldsymbol{\tau})}{\partial \tau_{p_1}^2 \cdots \partial \tau_{p_M}^2} \right|_{\boldsymbol{\tau}=\mathbf{0}} - (-1)^M \left. \frac{\partial^{2M} R(\boldsymbol{\tau})}{\partial \tau_{p_1}^2 \cdots \partial \tau_{p_M}^2} \right|_{\boldsymbol{\tau}=\boldsymbol{\tau}_0} . \qquad (2.13)$$

Consider the difference (2.13) of the terms in (2.12). All but one of the differences are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I when multiplied by $|\log \|\boldsymbol{\tau}\||^{1+\epsilon}$ based on the moment condition, boundedness of the exponential function and by lemma 5 since $\tau_{p_m} = O_I(\|\boldsymbol{\tau}\|)$. The exception is the difference of the first term, which is

$$\begin{aligned} c_M \int s^M \left(1 - \exp\left(-\boldsymbol{\tau}^T \boldsymbol{\tau} s\right)\right) dH(s) &= c_M \int s^M s \cdot c_1 \sum_{i,j} \tau_{p_i} \tau_{p_j} \exp(-\boldsymbol{c}^T \boldsymbol{c} s) dH(s) \\ &\propto \sum_{i,j} \tau_{p_i} \tau_{p_j} \int s^{M+1} \exp(-\boldsymbol{c}^T \boldsymbol{c} s) dH(s), \end{aligned} \qquad (2.14)$$

where the first equation follows by a multivariate Taylor expansion (Schervish 1995, p. 665), and the constant c_1 is a function of $\boldsymbol{c}$, which lies on the line segment joining $\boldsymbol{\tau}$ and $\mathbf{0}$. Again by lemma 5, the moment condition (which is applicable since $M+1 \leq 2M$ for $M \geq 1$) and the boundedness of the exponential function, we have that (2.14), when multiplied by $|\log \|\boldsymbol{\tau}\||^{1+\epsilon}$, is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I. This demonstrates that all of the Mth-order partial derivative processes are sample path continuous based on their covariance function; this gives us sample path differentiability as discussed in Section 2.5.2. Since I was arbitrary, the result holds throughout $\mathcal{X}$.

Q.E.D.

2.5.4.3 Application of results to specific correlation functions

Let's consider how these results on mean square and sample path differentiability apply to the correlation functions considered in this work. The exponential correlation is not mean square differentiable since its scale parameter has an inverse-gamma $\left(\frac{1}{2}, \frac{1}{4}\right)$ distribution (nor is the general power exponential correlation for $\nu < 2$ by seeing that the correlation function is not twice differentiable) and hence is not sample path differentiable by Cambanis (1973, Theorem 6). For the squared exponential and rational quadratic correlation functions, their respective scale parameter distributions are a point mass at 1 and a gamma distribution, both of which have infinitely many moments. Hence Gaussian processes with these correlation functions are infinitely sample path differentiable. The remaining correlation of interest is the Matérn. I have shown that for $\nu > M$, the Matérn gives processes that are at least Mth-order mean square differentiable, while for $\nu > 2M$ the processes are at least Mth-order sample path differentiable.

2.5.4.4 Mean square analyticity

We have seen that both the squared exponential and rational quadratic correlation functions give Gaussian processes with infinitely many mean square and sample path derivatives. I next show that both of these correlation functions are also mean square analytic on $\Re^1$. Loève (1978, p. 137) gives the result that a process is analytic in mean square if and only if the covariance function is analytic at every point (x, x), which for stationary functions simplifies to the point 0. By Churchill (1960, p. 40) a function is analytic at a point if the derivative as a function of a complex argument, $\tau = a + bi$, exists at that point and every point in a neighborhood of the point. Here I show that for processes on $\Re^1$, both the squared exponential and rational quadratic correlation functions are analytic functions at 0. Without loss of generality, let the scale parameter $\kappa = 1$. For the squared exponential correlation function,

$$\frac{\partial}{\partial \tau} \exp(-\tau^2) = -2\tau \exp(-\tau^2)$$

is finite for all complex τ and hence is differentiable in a neighborhood of 0. For the rational quadratic correlation function,

$$\frac{\partial}{\partial \tau} \left(\frac{1}{1+\tau^2} \right)^{\nu} = 2\nu\tau \left(\frac{1}{1+\tau^2} \right)^{\nu - 1}$$

is finite for $\tau \neq \pm i$. Hence in a neighborhood of 0, the derivative exists and so the rational quadratic function is analytic at 0. Hence processes with squared exponential or rational quadratic correlation are mean square analytic. While I have not seen proof of the result that processes with these correlation functions are sample path analytic, this may well be the case, which would provide further evidence that they may not be good choices for modelling data (Stein 1999).

2.5.5 Smoothness properties of kernel convolution covariance functions

Here I prove that Gaussian processes with nonstationary correlation of the generalized kernel convolution form,

$$\begin{aligned} R(\boldsymbol{x},\boldsymbol{y}) &= \frac{2^{\frac{p}{2}}\left|\Sigma_{\boldsymbol{x}}\right|^{\frac{1}{4}}\left|\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{4}}}{\left|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{2}}} R\left(\sqrt{Q_{\boldsymbol{xy}}}\right) \\ &= \frac{2^{\frac{p}{2}}\left|\Sigma_{\boldsymbol{x}}\right|^{\frac{1}{4}}\left|\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{4}}}{\left|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{2}}} \int_0^\infty \exp\left(-Q_{\boldsymbol{xy}}s\right) dH(s), \qquad (2.15) \end{aligned}$$

where

$$Q_{\boldsymbol{xy}} = (\boldsymbol{x}-\boldsymbol{y})^T\left(\frac{\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}}{2}\right)^{-1}(\boldsymbol{x}-\boldsymbol{y}), \qquad (2.16)$$

have smoothness properties similar to those of the isotropic correlation function, $R(\tau)$, on which they are based, provided the underlying kernel matrices vary smoothly. This result makes sense intuitively because if the kernels are smooth, then in a small neighborhood, the covariance is nearly stationary, so the smoothness properties should depend on the properties of the underlying isotropic correlation. The anisotropy in the correlation due to $\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}$ will not play a role in smoothness properties, because this merely rotates and scales the space. This does not affect smoothness properties, which relate to the behavior of a stationary covariance function near the origin.

The nonstationary correlation functions (2.15) are scale mixtures of the HSK nonstationary correlation (2.3). For these nonstationary scale mixtures, just as I demonstrated for the stationary scale mixtures, differentiability properties are directly related to existence of moments of the scale parameter. Here I prove that if $ES^{2M}<\infty$, then stochastic processes with the generalized kernel convolution covariance are Mth-order mean square differentiable (Section 2.5.5.1), and Gaussian processes are Mth-order sample path differentiable (Section 2.5.5.2).

2.5.5.1 Mean square differentiability

Theorem 11 *A nonstationary stochastic process, $Z(\cdot)$, has Mth-order mean square derivatives if it has a correlation function that can be expressed in the form (2.15), the scale parameter, S, has*

$2M$ moments, the elements of the kernel matrices, Σ_x, are M times sample path differentiable, and the kernel matrices are not singular.

Proof:

By Adler (1981, p. 27), the finiteness of

$$\left.\frac{\partial^{2M} R(\boldsymbol{x},\boldsymbol{y})}{\partial x_{p_1}\cdots\partial x_{p_M}\partial y_{p_1}\cdots\partial y_{p_M}}\right|_{\boldsymbol{x}=\boldsymbol{y}=\boldsymbol{u}}$$

for $p_m \in \{1,\ldots,P\}\,, m \in \{1,\ldots,M\}$ is sufficient for the existence of Mth-order mean square partial derivative processes. The $2M$th partial derivatives of (2.15) take the form

$$\begin{aligned}&\sum_{m_1=0}^{M}\sum_{m_2=0}^{M}\sum_{m_3=0}^{2M}\sum_{m_4=0}^{2M} D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_x|^{\frac{1}{4}}\right) D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_y|^{\frac{1}{4}}\right)\\ &\times D^{(m_3)}\left(|\Sigma_x+\Sigma_y|^{-\frac{1}{2}}\right) D^{(m_4)}\left(\int \exp(-Q_{xy}s)dH(s)\right).\end{aligned} \tag{2.17}$$

First consider $D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right)$. The highest order derivative involving only terms in $\boldsymbol{x}$ is of the form $D_{\boldsymbol{x}}^{(M)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right)$. By assumption, the kernel matrices are not singular, so negative powers after differentiation do not cause the expression to be infinite. A determinant can be expressed as a product of the elements in the matrix. By assumption, the elements of $\Sigma_{\boldsymbol{x}}$ are M times sample path differentiable, so the Mth-order derivative of the determinant is finite. The same argument holds by symmetry for $D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right)$. Next consider $D^{(m_3)}\left(|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right)$. Once again, since the matrices are assumed to not be singular, the power does not pose a problem. The determinant can be considered as the product of the sum of elements of $\Sigma_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{y}}$. By assumption the Mth-order derivatives with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ exist, so once again, the derivative is finite.

The final term is $D^{(m_4)}\left(\int \exp(-Q_{xy}s)dH(s)\right)$. By lemma 7, we can interchange differentiation and integration. Since finiteness of the $2M$th order derivatives implies finiteness of lower order derivatives, we need only assess

$$\begin{aligned} D^{(2M)}\left(\int \exp(-Q_{xy}s)dH(s)\right) &= c_1 D^{(1)}_{x_{p_1}}(Q_{\boldsymbol{xy}})\cdots D^{(1)}_{y_{p_M}}(Q_{\boldsymbol{xy}})\int s^{2M}\exp(-Q_{xy}s)dH(s)\\ &\quad +\ldots + c_{2M}D^{2M}Q_{xy}\int s\exp(-Q_{xy}s)dH(s).\end{aligned}$$

First consider the integrals. Using the assumption on the moments of S and the boundedness of the exponential function, the integrals are bounded. By lemma 6 the derivatives of Q_{xy} are $O_I(1)$, so they are bounded.

I have shown that all the terms in the $2M$th-order partial derivative (2.17) of the correlation function are finite, and therefore that the $2M$th-order mean square derivative exists. Since the argument applies to arbitrary $2M$th-order partial derivatives of the correlation and holds for arbitrary $\boldsymbol{u}$, all $2M$-th order mean square derivatives exist on $\mathcal{X}$.
Q.E.D.

2.5.5.2 Sample path differentiability

Theorem 12 *A nonstationary Gaussian process, $Z(\cdot)$, is Mth-order sample path differentiable if its correlation function can be expressed in the form (2.15), the scale parameter, S, has $2M$ moments, the elements of the kernel matrices, Σ_x, are $M+1$ times sample path differentiable, and the kernel matrices are not singular.*

Proof: To assess the condition (2.9) for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I, we need to assess $E|Z(\boldsymbol{x})-Z(\boldsymbol{y})|^2 = (E(Z(\boldsymbol{x})^2) - E(Z(\boldsymbol{x})Z(\boldsymbol{y}))) + (E(Z(\boldsymbol{y})^2 - E(Z(\boldsymbol{x})Z(\boldsymbol{y})))$. By symmetry, we need to consider only one of the two terms:

$$\begin{aligned} E(Z(\boldsymbol{x})^2) - E(Z(\boldsymbol{x})Z(\boldsymbol{y})) &= C(\boldsymbol{x},\boldsymbol{x}) - C(\boldsymbol{x},\boldsymbol{y}) \\ &= D^{(2M)}(R(\boldsymbol{x},\boldsymbol{y}))|_{\boldsymbol{x},\boldsymbol{x}} - D^{(2M)}(R(\boldsymbol{x},\boldsymbol{y}))|_{\boldsymbol{x},\boldsymbol{y}}. \end{aligned} \quad (2.18)$$

The $2M$th-order partial derivatives of the nonstationary correlation (2.15) are of the form

$$\begin{aligned} D^{(2M)}R(x,y) &= \sum_{m_1=0}^{M}\sum_{m_2=0}^{M}\sum_{m_3=0}^{2M}\sum_{m_4=0}^{2M} D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right) D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right) \\ &\quad \times D^{(m_3)}\left(|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}|^{-\frac{1}{2}}\right) D^{(m_4)}\left(\int \exp(-Q_{\boldsymbol{xy}}s)dH(s)\right), \end{aligned} \quad (2.19)$$

with the constraint that $m_1+m_2+m_3+m_4 = 2M$. Next, let

$$\begin{aligned} g_1(\boldsymbol{x},\boldsymbol{y}) &= D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right) \\ g_2(\boldsymbol{x},\boldsymbol{y}) &= D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right) \\ g_3(\boldsymbol{x},\boldsymbol{y}) &= D^{(m_3)}\left(|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}|^{-\frac{1}{2}}\right) \\ g_4(\boldsymbol{x},\boldsymbol{y}) &= D^{(m_4)}\left(\int \exp(-Q_{\boldsymbol{xy}}s)dH(s)\right), \end{aligned}$$

and successively apply the identity,

$$f(\boldsymbol{x},\boldsymbol{x})g(\boldsymbol{x},\boldsymbol{x}) - f(\boldsymbol{x},\boldsymbol{y})g(\boldsymbol{x},\boldsymbol{y}) = (f(\boldsymbol{x},\boldsymbol{x}) - f(\boldsymbol{x},\boldsymbol{y}))g(\boldsymbol{x},\boldsymbol{x}) - (g(\boldsymbol{x},\boldsymbol{x}) - g(\boldsymbol{x},\boldsymbol{y}))g(\boldsymbol{x},\boldsymbol{y}), \tag{2.20}$$

to

$$g_1(\boldsymbol{x},\boldsymbol{x})g_2(\boldsymbol{x},\boldsymbol{x})g_3(\boldsymbol{x},\boldsymbol{x})g_4(\boldsymbol{x},\boldsymbol{x}) - g_1(\boldsymbol{x},\boldsymbol{y})g_2(\boldsymbol{x},\boldsymbol{y})g_3(\boldsymbol{x},\boldsymbol{y})g_4(\boldsymbol{x},\boldsymbol{y}),$$

i.e., one of the terms in the sum (2.19). This gives us that (2.18) can be expressed as

$$\begin{aligned}\sum_{m_1=0}^{M}\sum_{m_2=0}^{M}\sum_{m_3=0}^{2M}\sum_{m_4=0}^{2M} \quad & (g_1(\boldsymbol{x},\boldsymbol{x}) - g_1(\boldsymbol{x},\boldsymbol{y}))\, g_2(\boldsymbol{x},\boldsymbol{x})g_3(\boldsymbol{x},\boldsymbol{x})g_4(\boldsymbol{x},\boldsymbol{x}) \\ & + (g_2(\boldsymbol{x},\boldsymbol{x}) - g_2(\boldsymbol{x},\boldsymbol{y}))\, g_1(\boldsymbol{x},\boldsymbol{y})g_3(\boldsymbol{x},\boldsymbol{x})g_4(\boldsymbol{x},\boldsymbol{x}) \\ & + (g_3(\boldsymbol{x},\boldsymbol{x}) - g_3(\boldsymbol{x},\boldsymbol{y}))\, g_1(\boldsymbol{x},\boldsymbol{y})g_2(\boldsymbol{x},\boldsymbol{y})g_4(\boldsymbol{x},\boldsymbol{x}) \\ & + (g_4(\boldsymbol{x},\boldsymbol{x}) - g_4(\boldsymbol{x},\boldsymbol{y}))\, g_1(\boldsymbol{x},\boldsymbol{y})g_2(\boldsymbol{x},\boldsymbol{y})g_3(\boldsymbol{x},\boldsymbol{y}). \end{aligned} \tag{2.21}$$

To satisfy the condition (2.9) I need only show that for $i \in \{1,2,3,4\}$ $(g_i(\boldsymbol{x},\boldsymbol{x}) - g_i(\boldsymbol{x},\boldsymbol{y}))|\log\|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon}$ is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I. By lemma 5 it is sufficient that $(g_i(\boldsymbol{x},\boldsymbol{x}) - g_i(\boldsymbol{x},\boldsymbol{y})) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$, which is satisfied by lemma 4 if $g_i(\boldsymbol{x},\boldsymbol{y})$ is once differentiable. $g_i(\boldsymbol{x},\boldsymbol{y})$ itself will involve at most M derivatives with respect to each of $\boldsymbol{x}$ and $\boldsymbol{y}$, so satisfying lemma 4 will involve at most the existence of $M+1$ derivatives; I focus on the highest order derivatives, as the other order derivatives are differentiable if the highest order derivatives are differentiable.

For $g_1(\boldsymbol{x},\boldsymbol{y})$ (equivalently for $g_2(\boldsymbol{x},\boldsymbol{y})$) the determinant can be expressed as a product of the elements of $\Sigma_{\boldsymbol{x}}$, so $g_1(\boldsymbol{x},\boldsymbol{y})$ is differentiable by the assumption that the elements of the kernel matrices are $M+1$ times differentiable. Note that raising the determinant to a power has no effect because the kernel matrices are assumed to not be singular. Similarly, for $g_3(\boldsymbol{x},\boldsymbol{y})$, the determinant can be considered as a product of terms of $\Sigma_{\boldsymbol{x}} + \Sigma_{\boldsymbol{y}}$, so $g_3(\boldsymbol{x},\boldsymbol{y})$ is differentiable by the assumption that the elements of the kernel matrices are $M+1$ times differentiable (since the $2M$th-order partial derivative involves at most M partial derivatives with respect to each of $\boldsymbol{x}$ and $\boldsymbol{y}$). Next consider the $2M$th-order partial derivative of $g_4(\boldsymbol{x},\boldsymbol{y})$,

$$\begin{aligned} D^{(2M)}\left(\int \exp(-Q_{\boldsymbol{xy}}s)dH(s)\right) &= D^{(2M)}(Q_{\boldsymbol{xy}})\int s\exp(-Q_{\boldsymbol{xy}}s)dH(s) \\ &\quad + \cdots + D^{(1)}_{x_1}(Q_{\boldsymbol{xy}})\cdots D^{(1)}_{y_M}(Q_{\boldsymbol{xy}}) \\ &\quad \times \int s^{2M}\exp(-Q_{\boldsymbol{xy}}s)dH(s), \end{aligned} \tag{2.22}$$

where I interchange differentiation and integration based on lemma 7. First consider the difference

$$\begin{aligned} & D^{(1)}_{x_1}(Q_{\boldsymbol{xy}}) \cdots D^{(1)}_{y_M}(Q_{\boldsymbol{xy}}) \int s^{2M} \exp(-Q_{\boldsymbol{xy}} s) dH(s) \Big|_{\boldsymbol{x},\boldsymbol{x}} \\ & - D^{(1)}_{x_1}(Q_{\boldsymbol{xy}}) \cdots D^{(1)}_{y_M}(Q_{\boldsymbol{xy}}) \int s^{2M} \exp(-Q_{\boldsymbol{xy}} s) dH(s) \Big|_{\boldsymbol{x},\boldsymbol{y}} . \end{aligned}$$

$D^{(1)}(Q_{\boldsymbol{xy}})\Big|_{\boldsymbol{x},\boldsymbol{x}} = 0$ and in the second term $D^{(1)}(Q_{\boldsymbol{xy}})\Big|_{\boldsymbol{x},\boldsymbol{y}} = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$ by lemma 6, so the whole expression is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. Note that I make use of the existence of the $2M$th moment of S and the boundedness of the exponential term. Hence this difference satisfies (2.9). Next consider the remaining terms in (2.22). Let

$$\begin{aligned} g_5(\boldsymbol{x},\boldsymbol{y}) &= D^{(m_5)}(Q_{\boldsymbol{xy}}) \\ g_6(\boldsymbol{x},\boldsymbol{y}) &= \int s^{m_6} \exp(-Q_{\boldsymbol{xy}} s) dH(s), \end{aligned}$$

where $m_6 \leq 2M-1$ and $D^{(m_5)}(Q_{\boldsymbol{xy}})$ is a product of terms involving derivatives of various orders of $Q_{\boldsymbol{xy}}$ with $m_5 \in \{1,\ldots,2M\}$. Applying the identity (2.20) to $g_5(\boldsymbol{x},\boldsymbol{x})g_6(\boldsymbol{x},\boldsymbol{x}) - g_5(\boldsymbol{x},\boldsymbol{y})g_6(\boldsymbol{x},\boldsymbol{y})$ I need only show that $g_5(\boldsymbol{x},\boldsymbol{y})$ and $g_6(\boldsymbol{x},\boldsymbol{y})$ are once differentiable to satisfy (2.9). First consider differentiating $g_5(\boldsymbol{x},\boldsymbol{y})$. The derivative $D^{(m_1)}(Q_{\boldsymbol{xy}})$ is at most order $2M$ and therefore order M in either $\boldsymbol{x}$ or $\boldsymbol{y}$. Since I have assumed $M+1$ derivatives of the kernel matrix elements, by lemma 4, $g_5(\boldsymbol{x},\boldsymbol{x}) - g_5(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. Next I show that $g_6(\boldsymbol{x},\boldsymbol{x}) - g_6(\boldsymbol{x},\boldsymbol{y})$ is at least $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$ using a multivariate Taylor expansion (Schervish 1995, p. 665) of $\exp(-Q_{\boldsymbol{xy}} s)$ at $\boldsymbol{y} = \boldsymbol{x}$ with first-order remainder:

$$1 - \exp(-Q_{\boldsymbol{xy}} s) \propto \sum_p s D^{(1)}_{y_p}(Q_{\boldsymbol{xy}})|_{\boldsymbol{x},\boldsymbol{c}} \exp(-Q_{\boldsymbol{xc}} s)(x_p - y_p) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|) \cdot s,$$

where $\boldsymbol{c}$ lies on the line segment joining $\boldsymbol{x}$ and $\boldsymbol{y}$. So we have

$$g_6(\boldsymbol{x},\boldsymbol{x}) - g_6(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|) c_1 \int s^{m_2+1} \exp(-Q_{\boldsymbol{xc}} s) dH(s),$$

and since $m_2 + 1 \leq 2M$, the integral is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ on I, and hence the whole expression is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$.

Therefore, all terms in (2.21) are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ on I when divided by $\|\boldsymbol{x}-\boldsymbol{y}\|$ and the condition (2.9) is satisfied by lemma 5. Since I was arbitrary, the result holds throughout $\mathcal{X}$.

Q.E.D.

2.5.5.3 Implications for nonstationary modelling

For the squared exponential and rational quadratic nonstationary correlation functions, which have infinitely many moments of the scale parameter, Gaussian process sample paths are infinitely mean square and sample path differentiable given sufficient smoothness in the kernel matrices. For the Matérn nonstationary correlation function with $\nu > 2M$, M mean square and sample path derivatives are guaranteed.

In constructing models, the elements of the kernels, $\Sigma_{\boldsymbol{x}}$, will themselves be random fields. It is sufficient that these elements (or the eigenvalues and eigenvector elements by lemma 8) be sample path differentiable to the Mth or ($M + 1$)th order to have Mth-order mean square or sample path differentiability, respectively. This suggests that at the highest level in the model hierarchy, one will need a stationary covariance structure to easily guarantee the desired sample path differentiability. In the regression modelling of Chapter 4, this stationarity is imposed on the kernel eigenstructure.

Also note that to use a nonstationary covariance model, as opposed to a correlation model, one introduces a variance function $\sigma^2(\boldsymbol{x})$. In my applications, I take this to be constant and therefore it has no effect on differentiability of the resulting processes. However, one could easily take this to be a random field and if it has $M + 1$ sample path derivatives, it is easy to show that theorems 11 and 12 continue to hold using arguments analogous to those regarding differentiation of the determinants of the kernels.

2.6 Discussion

This chapter introduces a class of nonstationary correlation functions based on familiar stationary correlation functions. By extending the original Higdon et al. (1999) kernel convolution method, the class provides a much broader set of nonstationary correlation functions than previously available. Some of these correlation functions may be better able to model particular datasets than the nonstationary correlation based on the squared exponential form, because they are more flexible in various respects.

In particular, I have provided nonstationary correlation functions with a range of smoothness properties, in contrast to the original HSK nonstationary covariance function (Fuentes and Smith

2001). The HSK approach gives infinitely differentiable sample paths unless a lack of smoothness is enforced through the kernel structure, which would be a rather ad hoc way to reduce smoothness. One of the new functions is a nonstationary version of the attractive Matérn correlation function, which has a parameter that indexes the mean square and sample path differentiability of Gaussian processes with this correlation function. With the new nonstationary correlation functions, one can create a smooth underlying kernel structure and yet retain control over sample path smoothness. In addition, one can create one's own nonstationary correlation function using any distribution for the scale parameter, S, involved in the generalization of HSK.

A related advantage of the generalization involves asymptotic behavior. Stein (1999) has shown that the convergence of kriging methods relies on the behavior of the correlation function near the origin and the compatibility, in a certain technical sense, of the modelled correlation function with the true correlation. He recommends the Matérn correlation because of its ability to adapt to different behavior near the origin, while the power exponential family is restricted to the extremes of non-differentiability and analytic behavior. Hence the nonstationary Matérn correlation function introduced in this chapter may be of particular interest.

The results in this chapter give sufficient, but seemingly not necessary conditions for smoothness properties of stochastic processes. In the stationary case, I have proven that the existence of $2M$ moments of the scale parameter is sufficient for Mth-order sample path differentiability. I suspect the condition can be weakened so that the existence of M moments is equivalent to Mth-order sample path differentiability. In the case of the Matérn, which is a mixture of the squared exponential correlation with a scale parameter distributed as inverse-gamma $\left(\nu, \frac{1}{4}\right)$ ($\lceil \nu - 1 \rceil$ moments), this would give exactly Mth-order sample path differentiability when $M < \nu \leq M + 1$, which is the same condition as for mean square differentiability (Stein 1999, p. 32). Furthermore, based on Cambanis (1973, Theorem 6), we know that the Matérn has no more than M sample path derivatives when $M < \nu \leq M + 1$. For the special case of the Matérn , one may be able to prove the sample path differentiability result by representing the correlation function as an infinite sum based on the properties of Bessel functions, for which Stein (1999, p. 32) and references therein would be a starting point. I do not have any suggestions for the general scale mixture case. For mean square and sample path differentiability of the nonstationary kernel convolution correlation functions, it may be possible to weaken the $2M$ moment condition to the existence of M moments of the scale parameter, although I do not have a suggestion for how to proceed. Also, in the case of sample path differentiability, I have required $M + 1$ sample path derivatives of the kernel matrix elements for ease of argument, but more subtle reasoning may allow this condition to be weakened to M derivatives.

The exact smoothness conditions are an issue only for the results of this chapter with respect to the Matérn correlation function, since the exponential, rational quadratic, and squared exponential lie in the extremes of differentiability. However, even for the Matérn, the key fact is that the differentiability varies with ν, not the exact number of derivatives as a function of ν, so sharpening the results in this chapter is of limited practical import.

Chapter 3

Methodological Development

3.1 Introduction

In this chapter, I present the basic methodology for using stationary and nonstationary covariance models as components in a Bayesian hierarchical model. I open by presenting methods for parameterizing the Gaussian kernels of the nonstationary correlation functions that I present in Chapter 2. Numerical sensitivity is an important issue for Gaussian process (GP) models because of the high correlation between the function values and the resulting numerical singularity of covariance matrices. I give an overview of approaches for dealing with this. I describe parameterizations for GP models, discuss some of the issues involved in particular parameterizations, and present the parameterization I have chosen to use. One difficulty is that the GP model involves an inherent degree of non-identifiability that I will describe. Next I discuss Markov chain Monte Carlo (MCMC) proposal schemes, describing previous approaches and discussing the mixing difficulties involved in sampling GP models. These schemes deal with the numerically-singular matrices in different ways, and I describe the issues involved. I suggest a new type of proposal scheme, which I call posterior mean centering (PMC), as a way to improve mixing and provide evidence that mixing improves with this scheme when the process cannot be integrated out of the model. The PMC scheme is implemented so as to avoid problems with singularity. One of the major drawbacks to GP models, including the models discussed in this thesis, is the $O(n^3)$ computation involved in working with the covariance matrices; I close by giving an overview of some approaches that have been suggested for speeding up the computations.

For clarity, in the sections that follow, with the exception of Section 3.2, I will assume the following simple model for the data and parameters:

$$\begin{aligned} Y_i &\sim \mathrm{N}\left(f\left(x_i\right), \eta^2\right) && (3.1)\\ f(\cdot) &\sim \mathrm{GP}\left(\mu, \sigma^2 R(\kappa)\right) && (3.2)\\ (\mu, \sigma, \kappa, \eta) &\sim \Pi(\mu)\cdot\Pi(\sigma)\cdot\Pi(\kappa)\cdot\Pi(\eta), && (3.3) \end{aligned}$$

with a stationary correlation function, $R(\kappa)$, parameterized by a scalar correlation scale parameter κ. In the nonstationary models considered in Chapters 4 and 5, κ is replaced by the convolution kernels and attendant hyperparameters used to construct the nonstationary correlation function. The methods and discussion given here apply to those more complicated hierarchical models with embedded Gaussian process distributions. Here and elsewhere in this thesis, as necessary, I take $\mu = \mu\mathbf{1}$, when a vector-valued mean is required.

3.2 Parameterizing the Kernels for Nonstationary Covariance Models

In Higdon et al. (1999)'s convolution approach to constructing nonstationary covariance functions, the kernels completely determine the nonstationary covariance structure of the GP. In Section 2.5.5, I show that to retain the smoothness of the underlying stationary correlation function upon which a kernel-based nonstationary correlation function is constructed, it is required that the kernel matrices vary smoothly in space. Using Gaussian kernels, this means that we need positive definite matrices that vary smoothly. The second key challenge is that the prior be uniform over the orientations of the Gaussian kernels.

3.2.1 Two-dimensional foci approach

Higdon et al. (1999) parameterize spatially-varying positive definite matrices based on the foci of the one standard deviation ellipse of the Gaussian distribution on $\Re^2$. Gaussian process priors with stationary squared exponential covariance functions are placed on the x and y coordinates of one focus. They force the areas of the ellipses for different locations to be the same, but not fixed, value. Swall (1999, p. 94) allows this area to vary but finds that the model overfits when sampling the variance and correlation scale parameters for the GP determining the area; she fixes these hyperparameters and suggests reparameterization or more informative prior distributions to solve the problem. It is not clear why this overfitting occurred; I have generally not found this to be a problem with the parameterizations used in the regression modelling in this work. Hilbert and Cohn-Vassen (1952) describe a generalization of the focal construction of an ellipse to ellipsoids in three dimensions. However the construction is complicated and extensions to higher dimensions are not apparent. I have not found other means of intuitively constructing high-dimensional ellipsoids. Hence, while the focal approach is feasible for the storm data of Chapter 5, other approaches are needed for data in higher dimensional spaces, such as the regression modelling of Chapter 4. For both the regression modelling and spatial modelling, I choose to use the eigendecomposition approach described in Section 3.2.3.

3.2.2 Cholesky decomposition

One approach to modelling smoothly-varying positive definite matrices in higher dimensions uses the Cholesky decomposition of the matrix, $\Sigma = LL^T$, where L is lower triangular. If one chooses the square of the diagonal elements, $L_{p,p}^2 \sim \chi^2_{d-p+1}, d > P-1, p = 1, \ldots, P$, and the off-diagonals to be $\mathrm{N}(0,1)$, with all the elements independent, then Σ will be positive definite and distributed Wishart (d, I) (Odell and Feiveson 1966). Now consider matrices, $\Sigma_{\boldsymbol{x}}$, defined at every point in the space. To produce spatially-varying positive definite matrices, let each off-diagonal element, $L_{i,j}(\cdot)$, be a random process distributed spatially according to Gaussian process with mean zero and a correlation function as the covariance function, thereby giving standard normal marginal distributions at each location. Let the GPs for each of the elements be independent. For the diagonal elements, $L_{p,p}(\cdot)$, use independent Gaussian processes with correlation function as the covariance function and have the diagonal elements be the square root of the inverse χ^2_{d-p+1} CDF transformation of the Gaussian process realizations. The result is spatially varying positive definite matrices that are marginally Wishart (d, I), and that vary smoothly in space according the correlation matrices specified for the underlying GPs of the elements of the matrices. Since the scale matrix of the constructed Wishart marginal distribution is the identity matrix, the marginal distribution for the resulting positive definite matrix at each location is rotationally symmetric.

Unfortunately, with this approach the marginal distributions of the ratios of the eigenvalues at a location change in concert with d, the value determining the degrees of freedom of the χ^2 random variables. The larger is d, the smaller is the magnitude of the larger eigenvalues relative to the smaller eigenvalues. In other words, kernels that spread over a large area tend to be spherical in shape. A further difficulty is that since the variances are marginally χ^2_d, it is difficult to express lack of prior information about the size of the variances since a large value for d results in little prior weight on small variances. Because of this difficulty in jointly specifying the prior variance and eigenvalue ratios, I choose instead to use the eigendecompositions of the kernel matrices directly.

3.2.3 Eigendecomposition

The eigendecomposition of a positive definite matrix is $\Sigma = \Gamma\Lambda\Gamma^T$, where Λ is a diagonal matrix of positive eigenvalues, and Γ is a matrix of eigenvectors, an orthonormal matrix.

3.2.3.1 Givens angles

A straightforward, and minimally parameterized, specification of a positive definite matrix is through its eigenvalues and the Givens angles used to construct its eigenvector matrix, which is a rotation matrix. Anderson, Olkin, and Underhill (1987) show that any orthogonal matrix can be expressed as the product of Givens matrices, G_{ij}, and a matrix, D_ϵ, of reflections:

$$\Gamma = (G_{12}G_{13}\cdots G_{1P})(G_{23}\cdots G_{2P})\cdots(G_{P-1,P})D_\epsilon \ ,$$

where

$$G_{ij} = G_{ij}(\rho_{ij}) = \begin{array}{cc} & \begin{matrix} & i & & j & \end{matrix} \\ \begin{matrix} \\ i \\ \\ j \\ \\ \end{matrix} & \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \cos\rho_{ij} & 0 & -\sin\rho_{ij} & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sin\rho_{ij} & 0 & \cos\rho_{ij} & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix} \end{array}$$

and $\rho_{ij} \in (-\frac{\pi}{2}, \frac{\pi}{2})$ is the angle of rotation in the i, j plane. Anderson et al. (1987) further show how to produce a random orthogonal matrix with Haar measure over the orthogonal group, where the distribution over Γ is the same as the distribution over $\Upsilon\Gamma$ for all orthogonal Υ. Take the elements of D_ϵ to be ± 1 independently with probability $\frac{1}{2}$. Let ρ_{ij} be independent with density proportional to:

$$\left(\prod_{j=2}^{P} \cos^{j-2} \rho_{1j}\right) \left(\prod_{j=3}^{P} \cos^{j-3} \rho_{2j}\right) \cdots \left(\prod_{j=P}^{P} \cos^{j-P} \rho_{P-1,j}\right). \tag{3.4}$$

If the marginal prior distribution for a single matrix Γ is specified in this way, then Σ is rotationally invariant since Γ has Haar measure.

This approach includes the matrix of reflections, which cannot be parameterized to vary smoothly in space. Fortunately, this matrix can be omitted without changing the resulting kernel matrices. Each kernel matrix, Σ, can be represented as

$$\Sigma = (G_{12} \cdots G_{1P})(G_{23} \cdots G_{2P}) \cdots (G_{P-1,P}) D_\epsilon \Lambda D_\epsilon (G_{P-1,P})^T (G_{2P}^T \cdots G_{23}^T)(G_{1P}^T \cdots G_{12}^T),$$

and since $D_\epsilon \Lambda D_\epsilon = \Lambda$, we see that D_ϵ is unnecessary. The attractive features of this parameterization are that it is simple to specify prior distributions such that each positive definite matrix is uniform over rotations and that the minimum number of parameters, $P + \frac{P(P-1)}{2}$ are used.

The primary drawback to this approach is the difficulty of parameterizing angles, ρ_{ij}, that vary smoothly in space. One can model non-normal random variables whose domain is monotonic as stochastic processes using the appropriate inverse CDF transformation of an underlying Gaussian process, but the lack of monotonicity of angular-valued random variables prevents that approach here. In the next section, I present an overparameterized alternative that works around this problem.

There are cases in which I need to parameterize individual correlation matrices and am not concerned with the spatial correlation structure of multiple matrices; in these cases I use the eigendecomposition with the Givens angle parameterization of the eigenvector matrix given above. To facilitate mixing, at the expense of identifiability, I expand the support of ρ_{ij} to $(-\pi, \pi)$. This allows smooth movements around the parameter space, without artificial boundaries at $\rho_{ij} = \pm\frac{\pi}{2}$. Furthermore, when a value of ρ_{ij} is proposed outside of $(-\pi, \pi)$, I replace it with $\rho_{ij} + c2\pi$, which gives the same eigenvector matrix, but keeps ρ_{ij} in its support when the integer c is chosen appropriately. The new prior is the same as before (3.4), but with $|\cos \rho_{ij}|$ in place of $\cos \rho_{ij}$. This can be seen to preserve the uniform prior over rotations because (3.4) arises in the proof in Anderson et al. (1987) as the Jacobian of a transformation; for $\rho_{ij} \in \left(-\pi, -\frac{\pi}{2}\right) \bigcup \left(\frac{\pi}{2}, \pi\right)$, we need to use the absolute value of the Jacobian to produce the density. Alternatives to the Givens angle parameterization for individual correlation matrices are discussed in Lockwood (2001, p. 58), including the uniform distribution over correlation matrices.

3.2.3.2 Overparameterized eigenvectors

To avoid the difficulties in working with smoothly-varying angles, I overparameterize the eigenvector matrix. For now, let's consider a single eigenvector matrix at a location. Instead of working with $P + \frac{P(P-1)}{2}$ parameters per positive definite matrix, I work with $P + \frac{P(P-1)}{2} + P - 1$ parameters. I construct the first eigenvector from P independent random variables by normalizing the variables collected into a vector. Then, in each successively smaller-dimensional subspace, I construct eigenvectors by collecting successively one fewer random variable into a normalized vector. I construct the final orthogonal matrix as the product of the orthogonal matrices in the subspaces, imposing constraints as necessary.

An example in three dimensions will clarify the setup. I construct the eigenvector matrix, Γ, as $\Gamma = \Gamma_2 \Gamma_1$. The elements of Γ_2 are determined from the realizations, x, y, z, of three random variables, X, Y, Z, in the following fashion

$$\Gamma_2 = \begin{pmatrix} \frac{x}{l_{xyz}} & \frac{-y}{l_{xy}} & \frac{-xz}{l_{xy} l_{xyz}} \\ \frac{y}{l_{xyz}} & \frac{x}{l_{xy}} & \frac{-yz}{l_{xy} l_{xyz}} \\ \frac{z}{l_{xyz}} & 0 & \frac{l_{xy}}{l_{xyz}} \end{pmatrix},$$

where $l_{xyz} = \sqrt{x^2 + y^2 + z^2}$ and $l_{xy} = \sqrt{x^2 + y^2}$. In this setup $(\Gamma_2)_{32} = 0$ is the required additional constraint on Γ_2, and the remaining elements of the matrix are fully determined. (Note that if the first eigenvector is extremely close to $(0, 0, 1)$, I need an additional constraint, so I also set $(\Gamma_2)_{22} = 0$.) I now descend to the two dimensional subspace orthogonal to Γ_2. Let

$$\Gamma_1 = \begin{pmatrix} \frac{u}{l_{uv}} & -\frac{v}{l_{uv}} \\ \frac{v}{l_{uv}} & \frac{u}{l_{uv}} \end{pmatrix}.$$

Here, the elements of the matrix are fully determined by two random variables, U and V, using a simplification of the construction of Γ_2 above. The final eigenvector matrix is the product, $\Gamma = \Gamma_2\Gamma_1$.

To ensure that the eigenvector matrices and hence the kernel matrices vary smoothly in space, I place GP priors, with mean 0, variance 1 and a correlation function, on the random variables used to construct the eigenvectors (in three dimensions, these are X, Y, Z, U and V). In doing this conversion from angular space to Euclidean space, I have used one extra parameter for each dimension (save the last), relative to the minimally-parameterized Givens angle approach. The kernel matrix parameterization is completed by taking Gaussian process priors for the log of each of the eigenvalue processes, with the mean and variance as hyperparameters. Note that this construction achieves uniformity over rotations, since the eigenvectors in the subspaces are uniformly distributed over rotations. It also achieves smoothness in space, provided that X, Y, Z, U, and V vary smoothly. Note that the elements of Γ_2 and Γ_1, and hence Γ as well, are infinitely differentiable as functions of X, Y, Z, U, and V, so the elements of the kernel matrices will be differentiable to the same degree as the stochastic processes used to construct the elements. If I take X, Y, Z, U, and V to be highly differentiable stochastic processes, then using the result in Section 2.5.5, the smoothness of nonstationary stochastic processes with kernel convolution covariance will depend on the smoothness properties of the underlying stationary correlation, which is my goal.

The parameterization in two dimensions involves only Γ_1. In higher dimensions, it is possible to parameterize the eigenvector matrix in similar fashion to that described above for three dimensions, although working out the correct constraints for Γ_3 and larger matrices is tedious, and it remains an open question as to whether a model with so many processes for the eigenvectors will mix well. There may be representations of the angles/eigenvectors that are simpler than that proposed above.

3.2.3.3 Simplified eigendecomposition model

If one were to use the construction just described to model nonstationary covariance in P-dimensional covariate spaces, one would need $P + \frac{P(P-1)}{2} + P - 1$ processes, with that many prior correlation functions. If these prior correlation functions were, for example, fully anisotropic stationary squared exponential correlations, each would have a matrix Δ in the Mahalanobis distance calculation in its correlation function, $R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{x_j})^T \Delta^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})\right)$, which would require $P + \frac{P(P-1)}{2}$ parameters, making for $\left(2P - 1 + \frac{P(P-1)}{2}\right) \cdot \left(P + \frac{P(P-1)}{2}\right)$ parameters just for the correlation structure of the eigenprocesses. Even in three dimensions, this starts to be unwieldy, with 48 parameters, and could lead to overfitting, so we might think about using simpler prior structures for the eigenvalues and eigenvectors. One possibility, which I employ in the regression modelling, builds on the observation that the primary goal is to have the kernels vary smoothly, and the eigenprocesses are only a means to that end. Instead of having $2P - 1 + \frac{P(P-1)}{2}$ different prior correlation functions, use the same prior correlation function for all the eigenvalue and eigenvector processes. This says that all of the processes used to construct the kernels have the same correlation structure. Note that the different eigenvalues will still have their own individual mean and variance hyperparameters, so that the scales of the eigenvalues may differ.

Other possibilities include requiring that the kernels be axis-oriented, so that the eigenvectors are not modelled and there are only P eigenvalue processes to model, having the eigenvector matrix be the same for all locations so that we need only $\frac{P(P-1)}{2}$ Givens angles to parameterize the eigenvector matrix, or having the eigenprocesses have simple one-parameter correlation functions (i.e., $\Delta = \frac{1}{\kappa} I$).

3.2.4 Basis kernel model

Modelling the eigenvectors and eigenvalues as Gaussian processes requires sampling Gaussian processes in the hierarchy of the model. In addition to the general difficulties involved in sampling from GPs described later in this chapter, when the GPs are not directly involved in the likelihood, the sampling may be particularly difficult. For problems in which the correlation scale of the data is unknown, such as devising a generic regression methodology, using GPs for the kernels may be the best approach. However, for problems in which the approximate correlation scales are known in advance, it may be possible to specify a much simpler model for the kernel matrices that is easier to fit, in particular, easier to sample from via MCMC. Higdon (1998) proposed a basis kernel approach in which a small number of basis kernels are parameterized, and kernels at any location of interest are weighted averages of the basis kernels using a simply-parameterized weight decay function. This reduces the number of parameters required to specify the nonstationary covariance structure and allows for easier sampling because one does not need to deal with the covariance structure of the kernels except via a parameter that determines the degree of locality in the weighted averaging of the basis kernels. The basis kernels should be located closely enough together that they can capture the nonstationary in the data, but far enough apart that they can be assumed independent a priori and so that there are relatively few of them. Since we do not need to explicitly specify correlation between the basis kernels, we can use the Givens angle approach to minimally parameterize the positive definite basis kernel matrices, giving P parameters for the eigenvalues and $\frac{P(P-1)}{2}$ parameters for the angles of each covariance matrix. In modelling the storm data in Chapter 5, I use nine Gaussian basis kernels spread over one third of the Northern Hemisphere, giving me $9 \cdot \left(P + \frac{P(P-1)}{2}\right) = 27$ parameters for the correlation structure. Higdon (1998) modelled ocean temperature data in a portion of the Atlantic Ocean using a set of 8 basis kernels. In practice the basis kernels appear to pick up some nonstationary features of the data, and the basis kernel and weight parameters seem to mix adequately. For the weight decay function, I use the squared exponential function (i.e., a Gaussian density function), but it does not have to be a positive definite function. However note that sample path smoothness will depend on whether the kernel matrices produced by weighted averaging of the basis kernels are sufficiently smooth, so the form of the function does matter. As the dimensionality of the space increases, more basis kernels are needed to cover the space to the same degree of density, so this approach is probably not feasible in high dimensions.

3.3 Numerical Sensitivity of GP Models

One difficulty in using GP models involves the numerical calculation of matrix square roots, solutions of systems of linear equations, and, depending on the fitting methods, matrix inverses. In particular, as the correlation scale increases, correlation matrices can approach numerical singularity. The smallest eigenvalues get so close to zero that the limitations of finite-precision arithmetic come into play, and numerically negative eigenvalues occur. This particularly affects the squared exponential correlation function, but also occurs with the Matérn even for relatively small values of ν. For example, using the R statistical software, which I believe has a relatively accurate Bessel function, and generating 100 points randomly on (0,1), the correlation matrix for the points was numerically singular with $\kappa = 0.25$ and $\nu = 4$ in 7 of 10 draws. Note that with $\nu = 4$ the sample paths are not quite three times differentiable. For $\kappa = 1.0$ and $\nu = 2$, in 7 of 100 draws the correlation matrix was singular; here the sample paths are not quite twice differentiable. The problem is acute with the Matérn and squared exponential correlation functions, because they specify that at small distances, locations are very highly correlated, with the derivative of the correlation function at 0 being 0, whereas the derivative of the exponential correlation function at 0 is negative.

To understand the problem in detail, let $\boldsymbol{f} = L\boldsymbol{\omega}$ where $LL^T = C = \text{Cov}(\boldsymbol{f})$ and $\boldsymbol{\omega} \sim \text{N}(0, I)$. The specific limitation can be seen in the calculation of the lower triangular square root matrix, or Cholesky factor, L, of C. One approach to calculating the Cholesky proceeds column by column. Consider the calculated diagonal element for the ith column, $L_{i,i} = \sqrt{C_{i,i} - \sum_{j=1}^{i-1} L_{i,j}^2}$, which can be interpreted as the residual variance of f_i, the ith element of $\boldsymbol{f}$, after regressing on $f_1, \ldots, f_{i-1}$. When f_i is very highly correlated with $f_1, \ldots, f_{i-1}$, $\sum L_{i,j}^2$ can be affected by round-off error in the calculations and $L_{i,i}^2$ can be smaller than machine precision (usually O $\left(10^{-16}\right)$ for double precision floating point values). When this happens, it means that the f_i is nearly a linear combination of $f_1, \ldots, f_{i-1}$. Several solutions have been proposed for this problem.

A straightforward approach when one only needs the Cholesky and not the inverse of the Cholesky, is to acknowledge that f_i is essentially a linear combination of $f_1, \ldots, f_{i-1}$ by setting $L_{i,i} = 0$ and doing the same for all the elements of that column of the Cholesky, $L_{i+1,i}, \ldots, L_{n,i}$ (Lockwood et al. 2001). In practice one sets a threshold, ϵ, and if the calculated value, $L_{i,i}^2 < \epsilon$, then the column is zeroed out. This states that the random variable is exactly a linear combination of the other random variables. In doing so, one makes the process smoother than it actually is, since one ignores the small amount of residual variability in the random variable. Using a different approach, Swall (1999) solved this problem in an elegant fashion by adaptively reordering the columns of the matrix so that the diagonal elements of L decreased monotonically. Upon reaching a predetermined tolerance level, she declared the remaining elements of $\boldsymbol{f}$ to be linear combinations of the previous elements. This adaptive approach seems likely to be more accurate than that of Lockwood et al. (2001). However, in part to minimize computation, I have employed the Lockwood et al. (2001) approach, which seems sufficiently accurate, provided the columns are not ordered such that many nearby locations with high correlation are closely grouped within the matrix (see below for more details). While both these approaches allow one to calculate an approximation to the Cholesky, if one needs the inverse of the Cholesky to solve a set of linear equations, then the result is values of infinity in the inverse matrix, which prevent further calculation, such as finding $\boldsymbol{\omega} = L^{-1}\boldsymbol{f}$, which is needed to calculate the prior density of $\boldsymbol{f}$. If the ith column is zeroed out, the value of the ith element of $\boldsymbol{\omega}$, ω_i, is unknown since it is not used in calculating $\boldsymbol{f} = L\boldsymbol{\omega}$. One possibility if one needs the inverse would be to set $\omega_i \sim \mathrm{N}(0,1)$, namely a random deviate from the prior on $\boldsymbol{\omega}$, although one would need to think through the implications of this before proceeding. In addition, with this generalized Cholesky algorithm, one cannot calculate the determinant of the matrix, which is also needed to calculate the prior for $\boldsymbol{f}$. Swall (1999) integrated the spatial mean process out of the model and only outside of the main Markov chain did she sample the process in a Gibbs step conditional on the hyperparameters, thereby avoiding the inverse and determinant calculations. However in sampling the spatial foci processes and their hyperparameters (Section 3.2.1), it is not clear how she avoided inverse and determinant calculations with ill-conditioned matrices, since these Metropolis steps would have required calculation of the GP prior for the focal processes.

In the models used in chapters 4 and 5, I set the tolerance of the generalized Cholesky algorithm to either $\epsilon = 10^{-10}$ or $\epsilon = 10^{-12}$. This was based on calculations with various tolerances and parameter values and with distances consistent with those used in the regression and spatial models. In the calculations, I reconstructed $C' = LL^T$ and compared the approximation, C', to the original C, although it is not clear which loss function to use in comparing the original and approximate covariance matrices. Setting the tolerance to a smaller value increases the error by keeping inaccurate values in L, while setting the tolerance to larger values increases error by zeroing out more columns. Based on some additional ad hoc experimentation, for future work, I would suggest a tolerance on the order of $\epsilon = 10^{-9}$.

One might think that if the matrix is numerically singular that calculating a generalized inverse would be a good approach. However, all the generalized inverse provides is one of multiple solutions to the system of linear equations. Also, it doesn't give you the determinant. It does not resolve the problem with numerical singularity, which is that we can't know what some of the elements of $\boldsymbol{a} = C^{-1}\boldsymbol{b}$ are. For the proposal schemes I describe in Section 3.6 one needs a square root of the covariance matrix, so the generalized Cholesky algorithm described above is sufficient, and a generalized inverse is not. An alternative would be to use the eigendecomposition of C and set all the eigenvalues less than some tolerance to zero, with $C^{\frac{1}{2}} = \Gamma\Lambda^{\frac{1}{2}}$. I have not investigated whether this approach gives a more accurate approximation to C than does the generalized Cholesky, but this may be worth more assessment. In the parameterizations and MCMC sampling schemes described in Sections 3.4 and 3.6, one prominent issue will be the need or lack thereof to invert the Cholesky of the covariance. I will present a parameterization and sampling scheme that work with only with the generalized Cholesky and not its non-existent inverse or determinant.

One effect of this numerical challenge is that the correlation function needs to be extremely accurate numerically. This is particularly an issue in calculating the Bessel function at the heart of the Matérn correlation function. An example of the sensitivity is that in running the same code on two different Linux PCs, one with an Intel Pentium processor and the other with an AMD Athlon processor, a difference of O $\left(10^{-15}\right)$ in the calculation of a covariance matrix element resulted in eventually changing the quantitative (though not qualitative) results of an MCMC run. Another example is that during MCMC, I have found that for certain sets of parameter values, I am not able to calculate the covariance matrix sufficiently accurately, and the resulting 'Cholesky' decomposition does not in fact come very close to reconstructing the original covariance such that $LL^T \approx C$. This occurs infrequently, but the result when it happens and the proposal is accepted is an erroneous sample path with local jaggedness that is not consistent with high local correlations specified by the hyperparameters. To illustrate the problem, using the built-in Bessel function in the statistical software R (besselK), I generated 5000 Matérn correlation matrices using $\log\kappa \sim \text{U}(\log(.03), \log(2))$ and $\nu \sim \text{U}(0.5, 30)$ and random permutations of a vector of 100 values of $x \sim \text{U}(0, 1)$. For a small number of correlation matrices (on the order of several dozen), the resulting generalized Cholesky was not a close approximation to the true Cholesky factor. There was no clear pattern relating the values of κ and ν to the inaccuracy, although the worst cases seemed to occur with relatively low values of κ and with values of ν greater than 10. Setting the tolerance at a larger value ameliorates the problem in the cases in which it arises, but at the expense of less accurate generalized Cholesky factors for most other parameter settings. Slightly changing the values of κ and ν removes the problem for an individual matrix. In Figure 3.1 I show an example of a sample function produced with an inaccurate generalized Cholesky factor.

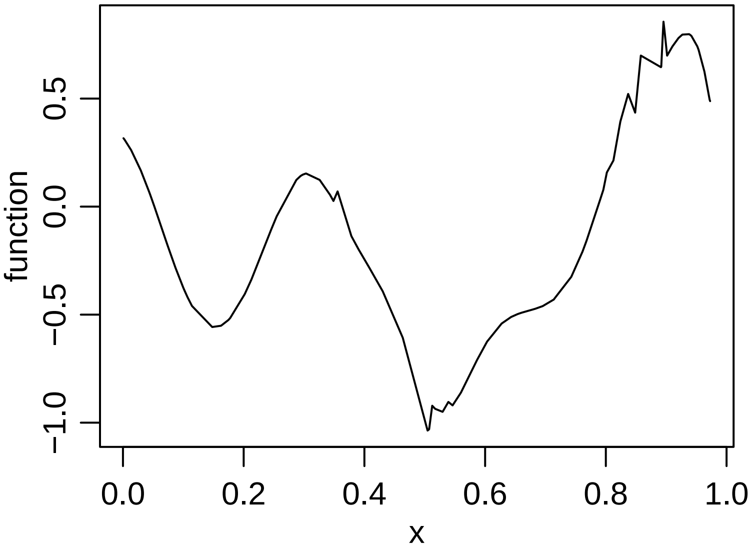

Figure 3.1. Sample function drawn from $f(\cdot) \sim GP(0, R(\kappa = 0.17, \nu = 25))$ *with an inaccurate generalized Cholesky factor. Note the jaggedness in the function at several points.*

In the machine learning literature, the standard solution to numerical singularity is to introduce a small amount of jitter on the diagonal by adding some small value δ to each diagonal element (Neal 1997). Provided the jitter is large enough, the matrix is no longer singular and can be inverted. If the jitter is not too large, then the result will hopefully be similar to the result that would be obtained under infinite-precision arithmetic. However, adding jitter raises some obvious questions of interpretation. A Gaussian process model specifies that nearby locations are highly correlated and that the resulting process is therefore smooth. Introducing jitter makes the process discontinuous, regardless of the correlation function used, because the covariance function after including the effect of jitter is not a continuous function at the origin. In practice the effect of jittering may not materially change the resulting function estimates, but this solution seems troubling. Machine learning researchers tend to set the mean of the GP to zero, which makes the covariance matrix even more ill-conditioned when the observations suggest that the realized function is far from zero, resulting in fitted correlations very close to one.

A final numerical issue involving the covariance matrix is the order of the locations in the covariance matrix. Because the zeroing out of a column in the generalized Cholesky algorithm is dependent on the correlation of the location corresponding to that column with the locations whose columns came previously, the accuracy of the approximation depends on the ordering of the columns. When the correlations are high, one should avoid having many neighboring locations as the early columns in the covariance, as this results in zeroing out columns early in the Cholesky, which can drastically lower the accuracy of the approximation. In practice, I use a random permutation of the columns to avoid placing many nearby locations in nearby columns of the covariance matrix. Another sensitivity with respect to the Cholesky arises because of the key role played by the first column and its location. The values of the random process are all conditional on the value of the process at this first location, so the ordering can affect the mixing of the chain. In Figure 3.2, I show an example of this in which the order affects the mixing of the correlation scale parameter for the kernel eigenvalue process, but seemingly not the other hyperparameters, in a one-dimensional regression problem with $f(x) = \sin \frac{1}{x}$.

3.4 GP Parameterizations

3.4.1 Centered vs. noncentered parameterizations

There are two straightforward parameterizations of a Gaussian process. The first parameterization is simply that for a finite set of values, $\boldsymbol{f}$, from the process

$$\boldsymbol{f} \sim \mathrm{N}(\mu, \sigma^2 R(\kappa)).$$

In this parameterization, which I will call the ‘centered’ parameterization for reasons that will become apparent, the function $\boldsymbol{f}$ has hyperparameters $\mu, \sigma,$ and κ, which reside at a level one higher in the model hierarchy than does $\boldsymbol{f}$. In the second, ‘uncentered’, parameterization, we have

$$\begin{aligned} \boldsymbol{f} &= \mu + \sigma L(\kappa)\boldsymbol{\omega} &(3.5)\\ \boldsymbol{\omega} &\sim \mathrm{N}(0, I), &(3.6)\end{aligned}$$

where $L(\kappa)$ is the Cholesky factor of $R(\kappa)$. Here, $\boldsymbol{f}$ is a deterministic function of the parameters $\boldsymbol{\theta} = \{\mu, \sigma, \kappa, \boldsymbol{\omega}\}$, which together reside at the same level of the model hierarchy.

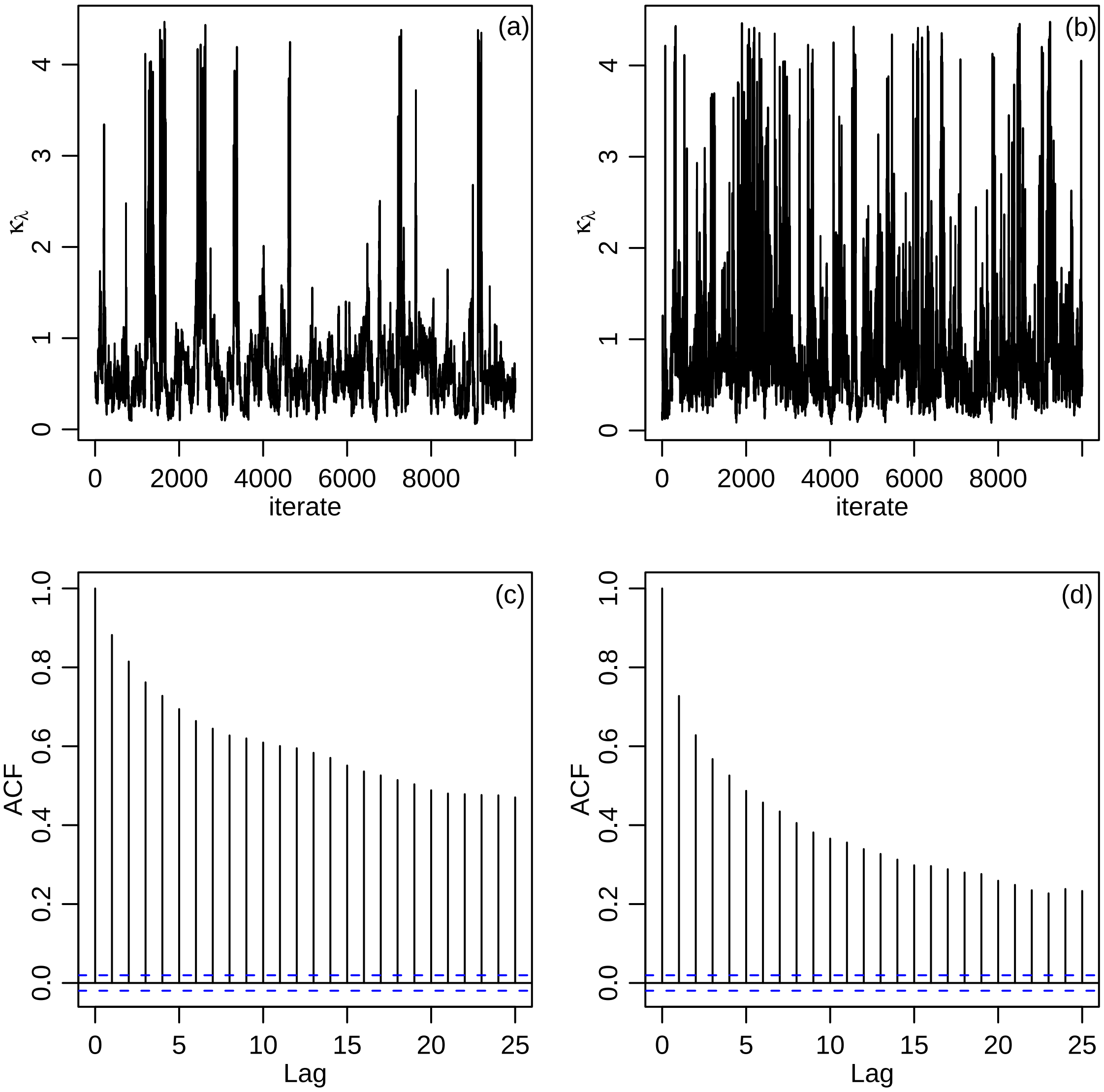

Figure 3.2. Effect of covariate order on κ_λ mixing in a regression problem with $f(x) = \sin \frac{1}{x}$, $x \in [0.1, 0.7]$; the plots are based on subsampling every 10th iteration from chains of length 100,000. I ran two runs, the first with $x_1 = 0.7$, $x_{100} = 0.1$, and proposal standard deviation of 0.24 (acceptance rate of 34%), and the second with $x_1 = 0.1$, $x_{100} = 0.7$, and proposal standard deviation of 0.28 (acceptance rate of 41%). (a) time series plot with $x_1 = 0.7$, (b) time series plot with $x_1 = 0.1$, (c) ACF with $x_1 = 0.7$, (d) ACF with $x_1 = 0.1$.

I use the terms centered and uncentered (Papaspiliopoulos, Roberts, and Sköld (2003) use the term non-centered) to follow the terminology of Gelfand, Sahu, and Carlin (1996), who discuss parameterizations for (generalized) linear mixed models. Their work suggests that when the data are relatively informative about a set of parameters and the priors are uninformative, the model should be reparameterized so that the parameters at the level of the hierarchy directly above the observations are identifiable. In other words, this level should contain parameters that are stochastically (or possibly deterministically) centered about their means rather than containing linear combinations of parameters from which the data will have a hard time distinguishing the parameters. In the alternatives above, we can see that the uncentered parameterization (3.5-3.6) takes $\mathrm{E}\boldsymbol{Y} = \mu + \sigma L(\kappa)\boldsymbol{\omega}$ as a linear combination of parameters, and the data are not able to inform the individual parameters. Only when the prior is informative and the likelihood relatively uninformative is the uncentered parameterization preferred (Gelfand et al. 1996). Papaspiliopoulos et al. (2003) find contrasting situations in which the centered and uncentered are preferred, with a similar conclusion that the centered is better when the likelihood is informative and the uncentered better when the likelihood is relatively uninformative. For a spatial model, they find that the centered outperforms the uncentered in MCMC sampling as the correlation of the spatial process increases.

In my experience, both the centered and uncentered parameterizations lead to slow mixing. In the regression modelling, straightforward sampling of the centered parameterization has the drawback of requiring jitter (Section 3.3), and mixing is even worse than the uncentered parameterization in the example in Section 3.6.2.3. For the uncentered parameterization, one difficulty is that with the generalized Cholesky algorithm, elements of $\boldsymbol{\omega}$ are not used in calculating $\boldsymbol{f}$ when their columns are zeroed out, and the identity of the elements that are unused changes during an MCMC, which may contribute to slow mixing. Note that the discretized parameterization discussed in Section 3.4.2 is similar in structure to the uncentered parameterization and so might be expected to have similar mixing problems. A more important difficulty is that different combinations of $\mu, \sigma, \kappa, \boldsymbol{\omega}$ can give very similar values of $\boldsymbol{f}$. The posterior is probably highly multimodal and moving between these parts of the parameter space is difficult. I concentrate on joint proposals for the function and hyperparameters (as discussed in Section 3.6), because, in my experience, hyperparameter mixing is slow unless the function is adjusted to be consistent with both the proposed hyperparameter(s) and the likelihood. The proposal scheme becomes even more important in the nonstationary case in which the scalar κ is replaced by the kernels and their attendant spatial processes and hyperparameters. In contrast, in work with spatial random effects in generalized linear mixed models, Christensen and co-workers concentrate on devising proposals for the function (the random effects in their terminology), commenting that the dependence amongst the function values is more of an obstacle than dependence between the function values and the other parameters. Christensen, Møller, and Waagepetersen (2000, 2001) and Christensen and Waagepetersen (2002) report successful mixing of both the function values and the hyperparameters with the uncentered parameterization using Metropolis-adjusted Langevin (MALA) updates for $\boldsymbol{\omega}$, despite using standard Metropolis-Hastings updates for σ and κ. They show that mixing is faster than with the centered parameterization, but their comparison is restricted to a naive implementation of the centered parameterization. This leaves open the question of how Langevin-style updates compare to the PMC proposals (Section 3.6) that I use; in Sections 3.6.2.3 and 4.6.4, I investigate this issue empirically.

The advantage of the uncentered parameterization is that it reparameterizes to a vector of independent variables, $\boldsymbol{\omega}$, and does not require calculation of $L(\kappa)^{-1}$. While the vector $\boldsymbol{\omega}$ has independent components a priori, this is not the case a posteriori, which may explain the difficulties with the uncentered parameterization. As proposed in Christensen, Roberts, and Sköld (2003), to improve mixing, one ideally wants to orthogonalize the parameters with respect to the posterior, adapting to the relative amounts of information in the prior and the likelihood. Papaspiliopoulos et al. (2003) propose a partially non-centered parameterization (PNCP) and Christensen et al. (2003) develop this approach in detail for spatial models, describing a scheme in which a data-dependent reparameterization is done at each MCMC step based on a local transformation that approximately orthogonalizes the parameters with respect to the posterior. They present results that suggest the PNCP works better than either the uncentered or centered alternatives for both the function values and the other parameters. Although this approach seems promising, I do not investigate the PNCP further, in part because implementation requires one to be able to invert $L(\kappa)$ to calculate the data-dependent reparameterization. Modification of the Christensen et al. (2003) approach, in a way that deals with the numerical singularity of my prior covariance matrices, and application to my model either on its own or in addition to the PMC scheme may help to improve mixing, albeit at some computational cost. My spatial model tends to show slow convergence from poorly chosen starting values, which may be because the chain is not geometrically ergodic (Papaspiliopoulos et al. 2003); the PNCP may do a better job of moving quickly out of the tails of the posterior distribution, which would allow for better assessment of MCMC convergence using multiple chains with disparate starting values.

While the centered and uncentered parameterizations are obviously equivalent probabilistically, there are some interesting distinctions that arise in the context of sampling the parameters via MCMC. In the centered parameterization, one samples the hyperparameters, (μ, σ, κ), and the function $\boldsymbol{f}$. In the uncentered, one samples $\boldsymbol{f}$ only as a function of the other parameters $(\mu, \sigma, \kappa, \boldsymbol{\omega})$, which are the actual parameters in the model. To consider further the relationship between the two parameterizations in the context of MCMC sampling, let's consider the basic model (3.1-3.3) and calculate the likelihood and prior. For simplicity, I assume the error variance, η^2, is fixed.

$$
\begin{aligned}
\Pi_{\text{centered}}(\boldsymbol{f}, \mu, \sigma, \kappa|\boldsymbol{y}) &\propto L(\boldsymbol{Y}|\boldsymbol{f}, \mu, \sigma, \kappa)\Pi\left(\boldsymbol{f}|\mu, \sigma, \kappa\right)\Pi(\mu)\Pi(\sigma)\Pi(\kappa) \\
&= \frac{1}{\eta^n}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{f})^T\left(\eta^2 I\right)^{-1}(\boldsymbol{y}-\boldsymbol{f})\right) \\
&\quad \times \frac{1}{\sigma^n|L(\kappa)|}\exp\left(-\frac{1}{2}(\boldsymbol{f}-\mu)^T\left(\sigma^2 R(\kappa)\right)^{-1}(\boldsymbol{f}-\mu)\right) \\
&\quad \times \Pi(\mu)\Pi(\sigma)\Pi(\kappa) \\
\Pi_{\text{uncentered}}(\boldsymbol{\omega}, \mu, \sigma, \kappa|\boldsymbol{y}) &= L(\boldsymbol{Y}|\boldsymbol{\omega}, \mu, \sigma, \kappa)\Pi(\boldsymbol{\omega})\Pi(\mu)\Pi(\sigma)\Pi(\kappa) \\
&= \frac{1}{\eta^n}\exp\left(-\frac{1}{2}(\boldsymbol{y}-(\mu+\sigma L(\kappa)\boldsymbol{\omega}))^T\left(\eta^2 I\right)^{-1}\right. \\
&\quad \left. \times\, (\boldsymbol{y}-(\mu+\sigma L(\kappa)\boldsymbol{\omega}))\right)\cdot\exp\left(-\frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}\right)\Pi(\mu)\Pi(\sigma)\Pi(\kappa).
\end{aligned}
$$

The key difference between these two parameterizations is the presence in the centered model of the normalizing constant, $(\sigma^n|L(\kappa)|)^{-1}$, of the prior on $\boldsymbol{f}$, which involves the determinant of the prior covariance for $\boldsymbol{f}$. This determinant favors less flexible functions $\boldsymbol{f}$, because when the correlation matrix $R(\kappa)$ has high correlations, the inverse of its determinant is large. This determinant is the way in which Occam's razor (see Section 3.5.1.1) plays a role in the model. Provided a less flexible function is as consistent with the data, based on the likelihood, as a more flexible function, the posterior will favor the less flexible function. The uncentered model does not have this determinant and seemingly does not favor less flexible functions in the same way that the centered parameterization does. Yet they are the same model; one is a simple reparameterization of the other. The answer to this seeming paradox can be seen in how one might implement a sampling scheme for the centered model. Suppose one were to sample from the centered parameterization, and consider the proposal for κ. If we propose a change to κ while keeping $\boldsymbol{f}$ the same, anything but a small change in κ will be rejected because the current $\boldsymbol{f}$ is correlated in a fashion that is consistent with κ and not the proposal κ^*. This sampling difficulty is precisely why the uncentered parameterization might seem to be a better sampling scheme than the centered with respect to mixing. However, instead of proposing κ alone, let's consider a joint proposal for $(\kappa, \boldsymbol{f})$. First propose a value κ^*, then propose

$$\boldsymbol{f}^* \sim \mathrm{N}\left(\mu + \sigma L\left(\kappa^*\right)\left(\sigma L(\kappa)\right)^{-1}\left(\boldsymbol{f} - \mu\right), vR\left(\kappa^*\right)\right) \tag{3.7}$$

with v the proposal standard deviation. This seemingly strange construction may make more sense if we consider what happens if $\kappa^* = \kappa$. In that case $\boldsymbol{f}^* \sim \mathrm{N}(f, vR(\kappa))$ namely we are proposing to perturb $\boldsymbol{f}$ in such a way that the new $\boldsymbol{f}^*$ is consistent with the prior correlation matrix. The more complicated joint proposal with κ^* accomplishes the same goal, but with a new value for κ. In (3.7) the inverse of the Cholesky decorrelates the original $\boldsymbol{f}$ and then the decorrelated and demeaned vector $\boldsymbol{\omega} = \left(\sigma L(\kappa)\right)^{-1}\left(\boldsymbol{f} - \mu\right)$ is recorrelated according to $L(\kappa^*)$. How does this bear on the original question of the difference between the centered and uncentered parameterizations? The new joint proposal is a Metropolis-Hastings scheme instead of a Metropolis scheme. Because the proposal is not symmetric, the acceptance ratio now involves the ratio of proposal densities, which I term the Hastings ratio. The Hastings ratio is:

$$\frac{\frac{1}{|L(\kappa)|}\exp\left(-\frac{1}{2}\left(\boldsymbol{f} - \mu - \sigma L(\kappa)\boldsymbol{\omega}^*\right)^T \left(vR(\kappa)\right)^{-1}\left(\boldsymbol{f} - \mu - \sigma L(\kappa)\boldsymbol{\omega}^*\right)\right)}{\frac{1}{|L(\kappa^*)|}\exp\left(-\frac{1}{2}\left(\boldsymbol{f}^* - \mu - \sigma L\left(\kappa^*\right)\boldsymbol{\omega}\right)^T \left(vR\left(\kappa^*\right)\right)^{-1}\left(\boldsymbol{f}^* - \mu - \sigma L\left(\kappa^*\right)\boldsymbol{\omega}\right)\right)} = \frac{\frac{1}{|L(\kappa)|}}{\frac{1}{|L(\kappa^*)|}}.$$

This Hastings ratio exactly cancels the ratio of determinants in the acceptance ratio that comes from the ratio of the posterior terms,

$$\frac{\frac{1}{|L(\kappa^*)|}}{\frac{1}{|L(\kappa)|}},$$

and the result is that we have precisely the acceptance ratio we would have had if we had used the uncentered parameterization, namely an acceptance ratio that does not include the determinant of the prior covariance matrix. In MCMC, Hastings ratios are required when the proposal makes moves to one part of the space more likely than moves to another part of the space. The preference is corrected in the acceptance ratio by downweighting the portions of the space that are preferred in the proposal. So the smoke clears over the parameterization question, and it is apparent that in the centered parameterization, as previously stated, less flexible functions are favored by the determinant in the prior for the function. In the uncentered parameterization, which can be viewed as the centered parameterization but with a certain type of joint proposal, less flexible functions are favored by biased moves in the proposals of the Markov chain. It is this movement preference that results in a model preference for less flexible functions, provided the data do not argue strongly enough in favor of flexible functions through the likelihood. How does the proposal bias toward less flexible functions come about? Any particular less flexible function is more likely to be sampled than a particular flexible function, because when a κ^* that gives high correlations is proposed, we conditionally sample $\boldsymbol{f}^*$ from a smaller space than the $\boldsymbol{f}^*$ sampled when proposing a κ^* that gives smaller correlations.

Hence it seems clear that sampling from either parameterization is equivalent and furthermore that in evaluating mixing in the uncentered parameterization, one need not be concerned with mixing of the $\boldsymbol{\omega}$ parameters but can safely focus on the values of $\boldsymbol{f}$ and its hyperparameters, since the uncentered parameterization is equivalent to the centered parameterization with joint proposals of $\boldsymbol{f}$ and its hyperparameters. This is reassuring, because my experience suggests that the $\boldsymbol{\omega}$ values mix very slowly. The reason for this may be related to the fact that values of $(\boldsymbol{\omega}, \mu, \sigma, \kappa)$ in disparate parts of the parameter space can give very similar values of $\boldsymbol{f}$. Moving between these disparate regions can be very difficult. While slow mixing of $\boldsymbol{\omega}$ may not be of concern, slow mixing of (μ, σ, κ) is still a concern. In particular, σ and κ are the parameters in the model that control the degree of smoothing.

3.4.2 Discretized parameterization

One important drawback to both parameterizations in the previous section is the need to calculate the Cholesky decomposition of the covariance matrix, which is an O $\left(n^3\right)$ operation. In much of his work with the nonstationary kernel convolution covariance, Higdon avoids working with the covariance matrix and its Cholesky by working with a discretized version of the model (Higdon 1998, 2002). The easiest way to understand this approach is to go back to Higdon et al. (1999)'s construction of the kernel convolution nonstationary covariance, in which the nonstationary correlation function,

$$C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int_{\Re^2} K_{\boldsymbol{x_i}}(\boldsymbol{u}) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u},$$

is seen as the correlation function of a process constructed as

$$Z(\boldsymbol{x}) = \int K_{\boldsymbol{x}}(\boldsymbol{u}) \psi(\boldsymbol{u}) d\boldsymbol{u},$$

where $\psi(\boldsymbol{u})$ is a Gaussian white noise process. To construct a process with approximately the kernel convolution covariance, consider a discrete grid of independent standard normal random variables as an approximation to the white noise process. Using the spatially-varying kernels, calculate the value of the process $Z(\boldsymbol{x}) = K\boldsymbol{\psi}$ where K is a matrix in which the values in the ith row are the values of the kernel centered at $\boldsymbol{x_i}$ evaluated at the locations of the white noise grid. This constructs the process, $Z(\boldsymbol{x})$, as a weighted average of a discrete set of white noise values, $\boldsymbol{\psi}$. Changing the kernels changes the covariance structure of the process, while different values of $\boldsymbol{\psi}$ give different realizations. This formulation looks identical to basis function regression, with the white noise values being the basis coefficients and the columns of K being the basis functions evaluated at the locations of interest. The advantage of working with this construction of the process is that calculating a realization of the process is O(nm) operations, where m is the number of locations in the white noise grid.

There are several potential disadvantages of the approach. First, one needs to specify the white noise grid. If the grid is too sparse, one will not be able to adequately model the fine-scale covariance structure in the data. If the grid is sufficiently fine, it seems possible that instead of the covariance structure being modelled by the kernels, the covariance structure might be absorbed into $\boldsymbol{\psi}$ during the fitting process. There is no penalty in the multivariate Gaussian prior density for the presence of correlation between elements when independence is specified, only a penalty when correlation is specified and the elements appear to be independent in reality. This could result in incorrect interpretation of the covariance in the data based on the fitted kernels. Another difficulty is one I observed in practice; it can be difficult to achieve reasonable mixing of the white noise random variables, presumably because many different vectors of white noise values can produce very similar values of the process. This concern might be alleviated by an argument along the lines made for disregarding mixing of the elements of $\boldsymbol{\omega}$ in the previous section, but it is a question that should be addressed. Finally, in high dimensions, the number of grid locations needed to cover the space grows quickly and the computational advantages of the discretization may diminish. For the regression model of Chapter 4, the optimal white noise grid would change with the type of data, so I do not view the approach as being sufficiently general. For the spatial model of Chapter 5 I did not use the discretization because hyperparameter mixing in the discretized model was slow and with replicated observations, the discrete approach requires a separate process (but the same kernels) for each replicate, diminishing the computational advantage.

A general discretized kernel convolution can be done by constructing a process through the convolution of a white noise (or other) process with kernels of arbitrary functional form in place of the Gaussian kernels. Mirroring the construction of the generalized kernel convolution covariance in Chapter 2, one might construct a process with each kernel being a scale mixture of kernels. For example, the Matérn correlation function has the same functional form as the Bessel density (McLeish 1982), which is a scale mixture of Gaussian densities. The difficulty in producing a process with a discretized version of the nonstationary Matérn correlation given in Chapter 2 is that the generalized kernel convolution covariance is based on mixing the product of two Gaussian kernels over the same scale parameter, S,

$$\begin{aligned} C(\boldsymbol{x_i},\boldsymbol{x_j}) \quad &\propto \quad \int\int \exp\left(-\frac{1}{2}\left(\boldsymbol{x_i}-\boldsymbol{u}\right)^T\left(\frac{\Sigma_i}{s}\right)^{-1}\left(\boldsymbol{x_i}-\boldsymbol{u}\right)\right) \\ &\qquad \times \exp\left(-\left(\boldsymbol{x_j}-\boldsymbol{u}\right)^T\left(\frac{\Sigma_j}{s}\right)^{-1}\left(\boldsymbol{x_j}-\boldsymbol{u}\right)\right) d\boldsymbol{u}dH(s), \end{aligned} \tag{3.8}$$

while the covariance function generated by convolving two Bessel densities involves two independent scale parameters:

$$\begin{aligned} C(\boldsymbol{x_i},\boldsymbol{x_j}) \quad &\propto \quad \int\int\int \exp\left(-\frac{1}{2}(\boldsymbol{x_i}-\boldsymbol{u})^T\left(\frac{\Sigma_i}{s_1}\right)^{-1}(\boldsymbol{x_i}-\boldsymbol{u})\right) \\ &\qquad \times \exp\left(-(\boldsymbol{x_j}-\boldsymbol{u})^T\left(\frac{\Sigma_j}{s_2}\right)^{-1}(\boldsymbol{x_j}-\boldsymbol{u})\right) d\boldsymbol{u}dH_1(s_1)dH_2(s_2) \\ &= \quad \int B((\boldsymbol{x_i}-\boldsymbol{u})^T(\Sigma_i)^{-1}(\boldsymbol{x_i}-\boldsymbol{u}))\cdot B((\boldsymbol{x_j}-\boldsymbol{u})^T(\Sigma_j)^{-1}(\boldsymbol{x_j}-\boldsymbol{u}))d\boldsymbol{u}, \end{aligned} \tag{3.9}$$

where B is the Bessel density function. Since (3.8) and (3.9) are not equivalent, it is not clear how one would construct a discretized version of the nonstationary Matérn covariance given in Chapter 2.

3.5 Model Dimensionality and Parameter Identifiability

In this section, I discuss how functions of disparate flexibility are embedded in the Gaussian process prior without an explicit change in model dimensionality. The Gaussian process model moves continuously between less flexible and more flexible functions; by comparison free-knot spline models change dimension in discrete jumps as knots are added and deleted. I describe how the GP regression model implicitly favors less flexible functions, provided they are consistent with the data. While the model has constant nominal dimension, I present one way of estimating the implicit dimension for a given set of hyperparameters. This allows me to put an explicit prior over function flexibility. Finally, while embedding functions with different degrees of flexibility in a single structure is an attractive notion, it has implications for interpreting, parameterizing, and sampling the parameters of the model.

3.5.1 Smoothing and dimensionality in the GP model

3.5.1.1 Occam's razor

As I alluded to earlier in the chapter, many Bayesian nonparametric regression models implicitly penalize more complicated models, which produce more flexible functions. Simpler models that are consistent with the data receive higher prior density because the simpler models spread their probability mass over a smaller function space of less flexible functions, and a correspondingly smaller data space, than do complicated models (Rasmussen and Ghahramani 2001; Denison et al. 2002). In the free-knot spline models, models with more knots spread their probability mass over larger spaces than models with fewer knots. The general effect, called Occam's razor (for the notion that simpler models should be preferred, all else being equal), is most easily understood by considering the marginal likelihood of the data, having integrated the appropriate parameters out of the model. To investigate this in the Gaussian process model, let's integrate the function out of the model, set $\mu = 0$ and $\sigma = 1$ for simplicity, and consider the posterior for κ:

$$\begin{aligned} \boldsymbol{Y} &\sim \mathrm{N}\left(0, R_{\boldsymbol{f}}(\kappa) + C_{\boldsymbol{Y}}\right) \\ \Pi(\kappa|\boldsymbol{Y}) &\propto \Pi(\kappa)L(\boldsymbol{Y}|\kappa) \\ &= \Pi(\kappa)\frac{1}{|R_{\boldsymbol{f}}(\kappa) + C_{\boldsymbol{Y}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T \left(R_{\boldsymbol{f}}(\kappa) + C_{\boldsymbol{Y}}\right)^{-1} \boldsymbol{y}\right). \end{aligned}$$

Now consider two scenarios: one with a value of κ that gives high prior correlation in $R_f(\kappa)$ and the second a value of κ giving low prior correlation. If the data truly do exhibit correlation based on their covariate values, then the quadratic form in the exponential of the likelihood will not be much worse under a model with high prior correlation than a model with low prior correlation, but the inverse of the determinant in the normalizing constant will favor the simpler model. So long as the prior, $\Pi(\kappa)$, is relatively uninformative, the posterior for κ will favor values of κ that induce high prior correlation. As we see in Chapter 4, this Occam's razor effect does seem to happen in practice; even when the GP model lacks the prior over implicit model dimensionality discussed in the next section, it does not appear to overfit even though the priors over the hyperparameters do little to differentiate between more and less flexible functions. This result is mirrored by the findings of Biller (2000) and DiMatteo et al. (2002) for free knot spline models in which the prior on the number of knots has little effect on the inference for the regression function, suggesting that the Bayesian framework allows the data to select the appropriate level of smoothing, with limited effect of the user's prior on the amount of smoothing.

3.5.1.2 Model dimension

In linear regression models, it is simple to estimate the dimension of the model. For polynomial regression, there are $k + 1$ parameters, where k is the degree of the highest polynomial, i.e., two for linear regression, three for quadratic regression, etc. The dimension tells us something about the flexibility of the model and the resulting function estimate. For the cubic regression spline model the dimension is taken to be $k + 4$ where k is the number of knots. This is based on the function estimate being $k + 1$ cubic polynomials, with three constraints at each knot ($k + 4 = (k+1)\cdot(3+1) - 3k$). In these models, the dimension changes discretely as the polynomial degree or number of knots change. In the GP model, the flexibility of the function depends on the degree of correlation in the correlation matrix, R_f, and the value of σ^2. The dimension of the model changes continuously rather than in discrete steps.

A linear smoother is a smoother in which the estimated function can be represented as

$$\tilde{\boldsymbol{f}} = S\boldsymbol{y}.$$

One standard way to estimate the effective number of parameters, or degrees of freedom (df), of a linear smoother is the trace of the smoothing matrix, S (Hastie and Tibshirani 1990, p. 52). This approach is motivated by the special case of polynomial regression, which has smoothing matrix, $S = B\left(B^T B\right)B^T$ where B is the design matrix constructed from the polynomial basis functions. The trace of this matrix is $k+1$, with k the degree of the highest order polynomial in the basis, so for this special case we recover the correct number of parameters in the model. To apply this to the GP regression model, remember that

$$\mathrm{E}(\boldsymbol{f}|\boldsymbol{y}, \theta_f, \eta) = \tilde{\boldsymbol{f}} = \mu + \sigma^2 R_{\boldsymbol{f}}(\sigma^2 R_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}(\boldsymbol{y} - \mu). \tag{3.10}$$

Therefore, we have

$$df = \mathrm{tr}\left(\sigma^2 R_{\boldsymbol{f}}\left(\sigma^2 R_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1}\right) + 1. \tag{3.11}$$

I add one to account for estimating μ, which is clearly the correct thing to do because in the limit of $\sigma \rightarrow 0$ we estimate $\tilde{f} = \mu$, using one parameter for the function and (3.11) gives $df = 1$. For a given set of hyperparameter values, I can estimate the flexibility of the conditional posterior mean function with df. There are alternatives for estimating df (Hastie and Tibshirani 1990), but this approach is widely used and seems adequate for my purposes. It also provides a way to impose one's prior belief about the flexibility of the function on the model. First, place a prior on the parameters of the model. Because of the high dimensionality of the model, it is difficult to know exactly how to choose this prior structure so as to encapsulate one's beliefs about flexibility. Instead of trying to do this painstakingly through the parameters, impose an additional prior constraint that is simply a distribution on df. If the additional prior on df is bounded, the new prior is guaranteed to be proper. While sampling from the Markov chain, for each set of parameter values, you can include in the prior calculation the contribution from the distribution over df. In practice, in one dimension, the model does a good job of choosing the flexibility without a prior over df. However, in higher dimensions, the model does have a tendency to undersmooth, at least with relatively few observations, so the imposition of the additional df prior helps somewhat (Section 4.6.2).

There are some drawbacks to the approach. It gives information about the function only through the conditional posterior mean, rather than the current sample of the function in the Markov chain. Also, when the smoothness parameter of the Matérn approaches $\frac{1}{2}$ (the exponential correlation function), the sample paths can be very unsmooth locally because of the lack of differentiability, but this is not well-reflected in the value of df; the calculation of df seems to primarily reflect the level of smoothness at coarser scales (what I have been referring to as flexibility).

For non-Gaussian likelihoods, this approach is not directly applicable, but we might apply the df calculation to an approximation to the posterior mean of the same form as (3.10),

$$\mu + \sigma^2 R_{\boldsymbol{f}}(\sigma^2 R_{\boldsymbol{f}} + C'_{\boldsymbol{Y}})(\boldsymbol{y}' - \mu),$$

where $C'_{\boldsymbol{Y}}$ and $\boldsymbol{y}'$ are approximations described in Section 3.6.2.3.

3.5.1.3 Covariate selection

One desirable feature of a regression model is that the model ignore covariates that are unrelated to the response. We want the function to be flat in the direction of the unimportant covariate(s). In the GP model, this corresponds to having the modelled covariance between locations be unaffected by the distance between locations with respect to the unimportant covariate(s). In the kernel-based nonstationary covariance, this is achieved by the kernel size being large in the direction of the unimportant covariate(s), with the appropriate eigenvalues being large. Unfortunately, allowing such large eigenvalues in the model can result in poor mixing during MCMC simulation because the chain tends to wander in that part of the parameter space as the likelihood and prior are relatively flat there. Once the eigenvalues are large, changes to them do not drastically affect the resulting correlations produced by the kernel convolution. Part of the issue here is that I propose vectors of parameters to speed the MCMC and use a single proposal variance; proposing elements of the vectors with different proposal variances would allow for better mixing, particularly if large proposal variances were used when the kernels were large in certain directions. To improve mixing, I limit the size of the eigenvalues to prevent them from wandering off to such extremely large values. However, if we are in high dimensions and there are one or more unimportant covariates, this limitation on the eigenvalue size means that some influence of the unimportant covariate(s) on the covariance structure remains. This can result in undersmoothing, which I have observed for some simulated datasets in which one or more covariates have no effect on the response variable. One possible way to avoid this problem would be explicitly include covariate selection in the model with a discrete part of the parameter space that represents very large eigenvalues. This effectively makes the function flat in the direction without requiring the eigenvalue to become extremely large. In addition to having to explicitly change model dimension during MCMC, the implementation of this idea may not be straightforward because of the need to have the eigenvalues at different locations be spatially correlated.

3.5.2 Parameter identifiability and interpretability

The continuously changing model dimension that I discuss in Section 3.5.1.2 has advantages and disadvantages. The primary advantage is that the model can move smoothly between structures of different implicit dimension and these structures are part of the same overall model with the same number of explicit parameters, so methods for dealing with models of different dimension, such as RJMCMC or model selection techniques are not required. Two potential disadvantages are that the meanings of the hyperparameters change as the implicit dimension changes and that in certain situations, parameters in the model become unidentifiable, particularly as the implicit dimension decreases. This lack of identifiability occurs in both the centered and uncentered parameterizations, albeit in somewhat different ways. For simplicity I will focus on the simple stationary Gaussian process prior for $f(\cdot)$ with one covariate, but the ideas apply in higher dimensional covariate spaces and to nonstationary priors.

3.5.2.1 Uncentered parameterization

As discussed in Gelfand et al. (1996), there is an inherent lack of identifiability in the uncentered parameterization, which takes

$$\boldsymbol{f} = \mu + \sigma L(\kappa)\boldsymbol{\omega}. \tag{3.12}$$

Different sets of parameter values $(\mu, \sigma, \kappa, \boldsymbol{\omega})$ can give the same value for $\boldsymbol{f}$. This makes it difficult to interpret the parameters and can lead to slow mixing in an MCMC simulation. The problem is particularly acute in the limit as the correlation approaches 1, namely as $\kappa \to \infty$. In this situation, $f(\boldsymbol{x}) = \mu + \sigma\omega_1$, and the data are unable to distinguish between the three parameters that determine the constant function $\boldsymbol{f}$. Also, there is a second way for $\boldsymbol{f}$ to be constant: $\sigma = 0$. In other words, both σ and κ can cause the covariate to have no impact on the response. They approach this limit in different ways because σ causes global smoothing while κ causes local smoothing that becomes global in its effect as κ gets very large.

The lack of identifiability can also make it very difficult to interpret the parameters. Consider σ and $\boldsymbol{\omega}$ in (3.12). We can obtain the same value for $\boldsymbol{f}$ by multiplying σ by a constant and $\boldsymbol{\omega}$ by the inverse of the constant. Since the $\mathrm{N}(0, I)$ prior on $\boldsymbol{\omega}$ strongly rewards small values of the vector-valued $\boldsymbol{\omega}$, σ will be driven up and $\|\boldsymbol{\omega}\|$ down, except to the extent that the prior on σ prevents this. As the number of elements in the vector increases, the magnitude of this effect increases as well. In situations such as this where the only information about a parameter is in the prior, there tends to be parameter drift during MCMC (Gelfand et al. 1996).

These identifiability issues suggest that the uncentered parameterization may be a poor choice, particularly for σ, but there are also issues of identifiability in the centered parameterization.

3.5.2.2 Centered parameterization

To assess identifiability in this parameterization, let's again consider extreme values of σ and κ. As $\kappa \to \infty$, the function becomes constant and we are in the situation of estimating a single value f based on the observations. This is fine for estimating f, but we are also trying to estimate the two hyperparameters, μ and σ, a task for which we have very little information (essentially just the information in $\bar{\boldsymbol{Y}}$). In practice, for large but not extreme values of κ that produce high correlation, we can get samples of σ that are very large. When this happens, there is little restriction on μ, since the values of $\boldsymbol{f}$ can be far from μ, and therefore the samples of μ tend to be extreme. So one effect of the parameterization is that the variability in μ changes with the value of σ, as can be seen clearly in time series plots of the MCMC samples of the two hyperparameters from the Bernoulli data example in Figure 3.4.

Next, let's consider the relationship between σ and κ. As in the uncentered parameterization, both $\sigma \rightarrow 0$ and $\kappa \rightarrow \infty$ can force the function to be constant. This aspect of the model seems undesirable and in practice, I limit how large κ can become, which helps to improve mixing during MCMC simulations. As I mentioned previously, this restriction on κ (which becomes a restriction on the eigenvalues of the kernels in the nonstationary model) does cause difficulties in higher dimensions with respect to covariate selection and undersmoothing.

The relationship between σ and κ, and more generally between the variance and correlation components of a covariance model is a complicated one. In both stationary and nonstationary settings, I have found nonintuitive results and tight correlations between variance and correlation parameters (also noted in Christensen et al. (2003)) that can slow mixing and confuse interpretation. I have not investigated the issue in the nonstationary model, but presumably the issue arises there in a more complicated fashion involving the kernel matrix eigenvalues. For the stationary GPs that I have used, in both the spatial and stationary regression models, I have found very high correlation between the σ and κ parameters, which leads to very slow mixing in the joint space of the two parameters. The correlation seems to occur because similar function values are reasonable according to the prior on $\boldsymbol{f}$ when both σ and κ are large and also when the two are both small. Large values of σ allow the data to still be consistent with a model of high correlations (large κ). I have addressed this issue to some extent in the sampling of the spatial model by jointly sampling the two parameters with $\log \sigma$ a linear function of $\log \kappa$ plus some added noise, and then also sampling $\log \sigma$ on its own. This corresponds quite closely to reparameterizing the two parameters as their sum and difference and to a reparameterization suggested in Christensen et al. (2003). With an inverse gamma prior on σ^2 it is possible to integrate σ out of the model, which might be expected to improve mixing as found in Huerta, Sansó, and Guenni (2001) and Sansó and Guenni (2002). However, this is not the case for my spatial model, as mixing of κ is still slow. This seems to occur because σ is acting in part as an auxiliary variable that allows the parameters to move more easily about the space. In particular, there seems to be a large part of the space with relatively low prior density that corresponds to large values of κ. This large low-density region is sampled poorly when σ is integrated out of the model. There are also unresolved issues in understanding the decomposition of covariance into variance and correlation in high-dimensional modelling of joint normality that I discuss in Section 5.7.2. Inadequacies in the correlation model appear to force high variance values. This effect seems similar to the correlation between σ and κ that I describe above.

In the regression modelling, I have dealt in part with these issues by fixing σ in the stationary GP priors for the eigenvalues used to construct the kernels, since I know the approximate range of reasonable eigenvalues. Neither kernels that are small relative to the smallest inter-observation distance nor kernels that are much larger than the largest inter-observation distance are desirable. For the eigenvector processes, I take $\sigma = 1$ without loss of generality, because I only use the processes to determine the directions of the eigenvectors and ignore the magnitude information.

While I have been able to use MCMC to sample from the GP models described in this thesis and achieve reasonably good mixing of the hyperparameters and their processes in some cases, slow mixing is an important limitation, and MCMC methods for Gaussian process models are not developed to the point that they can be automated and used widely without extensive user assessment and intervention.

3.6 MCMC Sampling Schemes for GP Models

In this section I discuss possible sampling schemes for MCMC fitting of GP models. After discussing a variety of approaches, I describe a particular joint proposal of process and hyperparameters, which I call posterior mean centering (PMC), that achieves faster mixing of the hyperparameters of the Gaussian process and avoids inversion of the generalized Cholesky factor of the covariance matrix. I use the centered parameterization (Section 3.4.1), but the PMC scheme gives the model the flavor of the uncentered parameterization. For simplicity, I base the discussion here on the simple model (3.1-3.3).

3.6.1 Integrating the process out of the model

If the likelihood is Gaussian or if the Gaussian process is embedded in a model such that there is conjugacy, one can integrate the process out of the model. In the simple model (3.1-3.3), it is straightforward to see that the marginal likelihood (with respect to the process) is

$$\boldsymbol{Y} \sim \mathrm{N}(\mu, \eta^2 I + \sigma^2 R_{\boldsymbol{f}}(\kappa)).$$

The hyperparameters are now directly involved in the likelihood and can be sampled in this lower dimensional parameter space. This is particularly advantageous if the number of locations is large. To estimate or sample from the process, one uses the conditional distribution of the process given the data and hyperparameters, which for the observed locations is normal with

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{f}|\boldsymbol{y},\eta,\mu,\theta\right) &= C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}\right)^{-1}\boldsymbol{y}+C_{\boldsymbol{Y}}\left(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}\right)^{-1}\mu \\
&= \mu+C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}\right)^{-1}\left(\boldsymbol{y}-\mu\right) \\
\mathrm{Cov}\left(\boldsymbol{f}|\boldsymbol{y},\eta,\mu,\theta\right) &= \left(C_{\boldsymbol{f}}^{-1}+C_{\boldsymbol{Y}}^{-1}\right)^{-1} \\
&= C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}\right)^{-1}C_{\boldsymbol{Y}}. \qquad (3.13)
\end{aligned}$$

In an MCMC simulation, one can conditionally sample the process at any chosen iteration of the chain based on the values of the hyperparameters in that iteration. Unfortunately, if there are many locations, computing the conditional variance (3.13) can be computationally intensive, since it involves applying the matrix inverse to $m+n$ vectors of length $m+n$, where n is the number of observed locations and m the number of locations at which one wishes to predict, whereas computation of the conditional mean, which is all that is required in the PMC proposal of Section 3.6.2.2 only requires applying the inverse to one vector, $\boldsymbol{y}-\mu$. If one also wishes to predict at m additional locations, the amount of computation increases even further. One particular advantage of integrating out the process is that instead of having to invert $C_{\boldsymbol{f}}$, computation is done with $(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}})^{-1}$, which for even relatively small error variance, is numerically non-singular. Gelfand, Kim, Sirmans, and Banerjee (2003) take the approach of integrating out the process from within a hierarchical model for housing prices as a function of geographic location and fit the hyperparameters by slice sampling to better handle parameter correlation. Paulo (2002) also integrates out the process and uses a multivariate t proposal density, with scale based on the observed Fisher information, for the hyperparameters. It is also possible to integrate some of the scalar parameters out of the model, in particular, μ, if it has a normal prior distribution, and σ^2, if it has an inverse gamma prior. In sampling spatiotemporal models, Huerta et al. (2001) and Sansó and Guenni (2002) integrate out both the process and σ^2 in order to sample from the marginal distribution of κ. As I discussed in Section 3.5.2, the posteriors for σ and κ can be highly correlated in certain models so integrating out σ may allow for faster parameter mixing, although for the spatial model, I have found that a joint proposal for the two parameters does a better job than the integration approach.

Other researchers avoid having to sample the hyperparameters by fixing them at reasonable values. In kriging, κ is usually chosen based on the empirical semivariogram, using a fitting method such as maximum likelihood or a more ad hoc procedure (Cressie 1993). In the machine learning literature, many researchers integrate the function out of the posterior, find the hyperparameter values that maximize the posterior, plug these fixed hyperparameters into the model, and then perform inference on the function (Gibbs and MacKay 1997; Vivarelli and Williams 1999). This approach tends to rely on using derivative information to perform the maximization, a subject I will address in Section 3.6.2.1. Even in fully-defined Bayesian models, many researchers have chosen to fix hyperparameters because of mixing problems or the computational intensity of fitting the full model (Higdon 1998; Swall 1999). This approach of fixing the hyperparameters may be satisfactory in many situations, but will underestimate the uncertainty in the resulting estimates. If the values are chosen poorly, the resulting estimates may be strongly biased relative to those in which the hyperparameters are fully sampled.

3.6.2 Methods for sampling the process values

Integrating the process out of the model is very common and successful and is part of the reason for the focus on Gaussian likelihood GP models, since model fitting eases considerably. However, there are many situations in which a Gaussian likelihood is not reasonable or one wishes to embed a Gaussian process prior in a complicated model out of which the process cannot be integrated. In recent years, since the introduction of generalized linear models (McCullagh and Nelder 1989), much attention has been paid to generalizing Gaussian likelihood methods to other situations, with attention often focusing on count data (Poisson likelihoods) and binary data (binomial likelihoods). One example of this is the use of spline models for non-Gaussian likelihoods (Biller 2000; DiMatteo et al. 2002). Diggle et al. (1998) discuss the same generalization for kriging methods for spatial data. This situation also arises in generalized linear mixed models (Breslow and Clayton 1993) that include a spatial random effect (Christensen and Waagepetersen 2002). In this thesis, non-conjugacy arises with the GP priors for the eigenprocesses that determine the kernels in the nonstationary GP regression model and in the residual variance process for the spatial model. In the machine learning literature, attention has focused on using Gaussian processes for classification (Neal 1996; MacKay 1997; Williams and Barber 1998).

Sampling these various GP-based models requires sampling process values that cannot be integrated out of the model, and this remains an ongoing research challenge because of parameter correlation (Diggle et al. 1998; Gelfand et al. 2003). Approaches that use gradient information, such as the Langevin algorithm, can help in the mixing of the process values, but hyperparameter mixing can still be slow. More exotic versions of MCMC may help in getting the process and hyperparameters to both mix, but the root of the problem is that different regions of the hyperparameter space can produce reasonable process values yet these areas can be hard to find and move between. The crux of the matter lies in being able to move about in hyperparameter space while sampling process values that are consistent with both the hyperparameters and the data. In Section 3.6.2.2 I outline the approach of posterior mean centering to move in hyperparameter space while automatically proposing process values that are more consistent with both the prior and the likelihood. Another challenge is computational speed. Full sampling of GP models is slow, particularly with many observations or many locations at which one wants to predict, because calculations with the covariance matrix are O(n^3). In Section 3.7 I review some of the work on computation in GP models; most of the research on this topic is being done in the machine learning community.

3.6.2.1 Derivative-based methods

The usual Metropolis-Hastings algorithm does not make use of information from the posterior in choosing proposals. In particular, the gradient of the posterior at the current parameter values may contain valuable information about the direction in which one should sample to move to regions of the parameter space with higher density. The Metropolis-adjusted Langevin Algorithm (MALA), also known as Langevin-Hastings, uses the gradient of the posterior in making proposals and includes in the Hastings ratio the appropriate correction needed because the Langevin diffusion is discretized (Robert and Casella 1999, Section 6.5.2; Christensen et al. 2001). In the context of GP models, Langevin-style proposals are most helpful in proposing the process values, rather than scalar parameters (Roberts and Rosenthal 1998; Christensen and Waagepetersen 2002).

However, Christensen et al. (2000) show that the hyperparameters mix well when the process is proposed using the Langevin approach and the scalar hyperparameters are proposed using standard Metropolis, albeit still requiring at least 100,000 MCMC iterations for only 70 spatial locations. The Langevin approach tends to speed the movement of the chain toward the modes of the process (Robert and Casella 1999). Christensen et al. (2003) use a partially non-centered proposal (PNCP) with Langevin updates and report greatly improved mixing for both the hyperparameters and the process values. Christensen and co-workers (prior to Christensen et al. (2003)) use the uncentered parameterization (3.6) with parameter $\boldsymbol{\omega}$. In the generalized linear mixed model framework, the Langevin proposal for $\boldsymbol{\omega}$ is

$$\boldsymbol{\omega}^{*} \sim \mathrm{N}\left(\boldsymbol{\omega} + \frac{v^2}{2}\nabla(\boldsymbol{\omega}), v^2 I\right), \tag{3.14}$$

where

$$\begin{aligned}\nabla(\boldsymbol{\omega}) &= \frac{\partial}{\partial\boldsymbol{\omega}}\log\Pi(\boldsymbol{\omega} \mid \boldsymbol{y}, \mu, \sigma, \kappa, \eta) \\ &= -\boldsymbol{\omega} + \frac{1}{\eta^2}\sigma L_{\boldsymbol{f}}(\kappa)^T(\boldsymbol{y} - \boldsymbol{g})\end{aligned}$$

and $\boldsymbol{g}$ is the mean function (simply $\boldsymbol{f}$ in the normal likelihood case). The Hastings ratio is:

$$\frac{\exp\left(-\frac{1}{2v^2}\left(\boldsymbol{\omega} - \left(\boldsymbol{\omega}^{*} + \frac{v^2}{2}\nabla\left(\boldsymbol{\omega}^{*}\right)\right)\right)^T\left(\boldsymbol{\omega} - \left(\boldsymbol{\omega}^{*} + \frac{v^2}{2}\nabla\left(\boldsymbol{\omega}^{*}\right)\right)\right)\right)}{\exp\left(-\frac{1}{2v^2}\left(\boldsymbol{\omega}^{*} - \left(\boldsymbol{\omega} + \frac{v^2}{2}\nabla(\boldsymbol{\omega})\right)\right)^T\left(\boldsymbol{\omega}^{*} - \left(\boldsymbol{\omega} + \frac{v^2}{2}\nabla(\boldsymbol{\omega})\right)\right)\right)}.$$

To compare mixing based on the Langevin algorithm to that with the PMC approach (Sections 3.6.2.3 and 4.6.4), I modify this algorithm for the regression and spatial models in this work. I use the centered parameterization, so $\boldsymbol{f}$ is the parameter, but I follow Christensen and co-workers in making use of the derivative of $\boldsymbol{\omega}$ rather than the derivative of $\boldsymbol{f}$. I do this for two reasons; first, Møller, Syversveen, and Waagepetersen (1998) report that the Langevin algorithm performs better when based on $\boldsymbol{\omega}$, and second, the derivative of $\boldsymbol{f}$ involves the inverse of the Cholesky of the covariance, which I cannot generally calculate because of the numerical singularity of the covariance matrix. For the regression model (the proposal for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the spatial model would be similar), the Langevin proposal for $\boldsymbol{f}$ based on (3.14) is

$$\boldsymbol{f}^{*} \sim \mathrm{N}\left(\boldsymbol{f} - \frac{v^2}{2}(\boldsymbol{f} - \mu) + \frac{v^2}{2}\frac{\sigma^2 R_{\boldsymbol{f}}(\boldsymbol{y} - \boldsymbol{g})}{\eta^2}, v^2 R_{\boldsymbol{f}}\right),$$

and the Hastings ratio, expressed partly in terms of $\boldsymbol{\omega}$ for simplicity, is

$$\frac{\exp\left(-\frac{1}{2v^2}\left(\boldsymbol{\omega}-\boldsymbol{\omega}^*+\frac{v^2}{2}\boldsymbol{\omega}^*-\frac{v^2}{2}\frac{\sigma L_{\boldsymbol{f}}^T(\boldsymbol{y}-\boldsymbol{g}^*)}{\eta^2}\right)^T\left(\boldsymbol{\omega}-\boldsymbol{\omega}^*+\frac{v^2}{2}\boldsymbol{\omega}^*-\frac{v^2}{2}\frac{\sigma L_{\boldsymbol{f}}^T(\boldsymbol{y}-\boldsymbol{g}^*)}{\eta^2}\right)\right)}{\exp\left(-\frac{1}{2v^2}\left(\boldsymbol{\omega}^*-\boldsymbol{\omega}+\frac{v^2}{2}\boldsymbol{\omega}-\frac{v^2}{2}\frac{\sigma L_{\boldsymbol{f}}^T(\boldsymbol{y}-\boldsymbol{g})}{\eta^2}\right)^T\left(\boldsymbol{\omega}^*-\boldsymbol{\omega}+\frac{v^2}{2}\boldsymbol{\omega}-\frac{v^2}{2}\frac{\sigma L_{\boldsymbol{f}}^T(\boldsymbol{y}-\boldsymbol{g})}{\eta^2}\right)\right)}.$$

Following Christensen et al. (2001) and Christensen and Waagepetersen (2002) I truncate the gradient. In the Poisson and binomial cases, we have $\eta = 1$ and $\boldsymbol{f}$ is replaced by the mean functions, $\boldsymbol{g} = \exp(\boldsymbol{f})$ (Poisson) or $\boldsymbol{g} = \boldsymbol{m}\frac{\exp(\boldsymbol{f})}{1+\exp(\boldsymbol{f})}$ (binomial), where $\boldsymbol{m}$ is a vector of the number of trials.

The Langevin algorithm is a special case of an MCMC algorithm called Hybrid Monte Carlo (HMC), which is popular among machine learning researchers. The HMC algorithm (Duane, Kennedy, Pendleton, and Roweth 1987) endows Metropolis-Hastings with not only position information (the current parameter values), but also momentum information, which causes the chain to avoid random walks by favoring movement in the same direction on successive steps through the inertia of the parameters. The method was originally devised in statistical physics and is usually described by analogy to a physical system in which a physical particle is moving through a region of variable potential energy (probability). A number of authors have used the HMC algorithm for sampling from GP-based models and claim success in sampling the model parameters (for Gaussian likelihoods, they apply HMC to the hyperparameters only) (Rasmussen 1996; Williams and Rasmussen 1967; Neal 1997; Williams and Barber 1998; Rasmussen and Ghahramani 2002), although they provide little evidence of mixing to which to compare alternate sampling schemes. The Langevin algorithm is HMC using a single leapfrog step of the discretized position and momentum differential equations.

One drawback to derivative-based methods is that one cannot always obtain a closed form for the derivative of the process values. For example, this is the case for the kernel eigenprocesses in the nonstationary GP regression model (Chapter 4) and for the residual variance process in the spatial model (Chapter 5). It would be possible to calculate numerical derivatives for the process values, but this would be very computationally intensive. Because of this limitation of derivative-based methods and because I have not found that these methods particularly improve hyperparameter mixing, I turn to a different approach in the next section. In assessing the performance of the posterior mean centering proposal scheme, I compare it to both Langevin-style proposals for the process values and to joint proposals for the centered parameterization that do not include likelihood information (Sections 3.6.2.3 and 4.6.4).

3.6.2.2 Posterior mean centering

The mean centering approach builds on the uncentered parameterization (3.6). Recall that in this parameterization, the parameters (μ, σ, κ) deterministically change $\boldsymbol{f}$ and therefore are directly involved in the likelihood. Because of the deterministic relationship between the parameters and $\boldsymbol{f}$, the process is automatically consistent with the parameter values, in particular with κ, which determines the correlation of $\boldsymbol{f}$. This is because $\boldsymbol{f}$ is produced by filtering the independent (a priori) random variables in $\boldsymbol{\omega}$ through the generalized Cholesky factor, $L_{\boldsymbol{f}}$, which is a function of the current value of κ. The main drawback to the approach and the reason for slow mixing is that proposals for $\boldsymbol{f}$ are not necessarily consistent with the likelihood, except insofar as $\boldsymbol{f}^*$ is similar to $\boldsymbol{f}$ and $\boldsymbol{f}$ is consistent with the likelihood. The PMC approach uses information in the likelihood as well as the prior in making proposals. In the development of the posterior mean centering approach that follows, I will focus on joint proposals for κ and $\boldsymbol{f}$, but the methodology and discussion carry over to μ and σ straightforwardly (as well as to ν in a Matérn parameterization and to the parameters involved in the nonstationary correlation structure).

First I outline a joint sampling scheme for the centered parameterization that is equivalent to the uncentered parameterization, but in which $\boldsymbol{\omega}$ is merely an implicit parameter that is carried along in the calculation. This scheme is the basis for the PMC scheme. Consider a joint proposal for $(\kappa, \boldsymbol{f})$. First propose a value κ^*, then propose

$$\boldsymbol{f}^* \sim \mathrm{N}\left(\mu + \sigma L_{\boldsymbol{f}}\left(\kappa^*\right)\left(\sigma L_{\boldsymbol{f}}(\kappa)\right)^{-1}(\boldsymbol{f} - \mu), v^2 R_{\boldsymbol{f}}\left(\kappa^*\right)\right). \tag{3.15}$$

As described in Section 3.4.1, the Hastings ratio cancels with the ratio of determinants from the priors in the Metropolis acceptance ratio, and we are left with the acceptance ratio we would have had if we had used the uncentered parameterization and jointly proposed $(\kappa, \boldsymbol{\omega})$. Now let's consider what this proposal for $\boldsymbol{f}$ implies about the proposal for the implicit parameter $\boldsymbol{\omega} = (\sigma L_{\boldsymbol{f}}(\kappa))^{-1}(\boldsymbol{f} - \mu)$.

$$\boldsymbol{\omega}^* \;=\; (\sigma L_{\boldsymbol{f}}(\kappa^*))^{-1}(\boldsymbol{f}^* - \mu)$$

$$\begin{aligned}
&= (\sigma L_{\boldsymbol{f}}(\kappa^*))^{-1}(\mu + \sigma L_{\boldsymbol{f}}(\kappa^*)\boldsymbol{\omega} + vL_{\boldsymbol{f}}(\kappa^*)\boldsymbol{\psi} - \mu) \\
&= \boldsymbol{\omega} + v\boldsymbol{\psi}, \qquad (3.16)
\end{aligned}$$

where $\boldsymbol{\psi}$ is the vector of standard normal values used in generating $\boldsymbol{f}^*$. We see that we can update the implicit parameter and calculate the prior for $\boldsymbol{f}$ in the acceptance ratio without using the inverse of the Cholesky:

$$\begin{aligned}
\Pi\left(\boldsymbol{f}^*|\mu,\sigma,\kappa^*\right) &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{f}^*-\mu\right)^T\left(\left(\sigma^2 R_{\boldsymbol{f}}\left(\kappa^*\right)\right)^{-1}\left(\boldsymbol{f}^*-\mu\right)\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\left(\sigma L_{\boldsymbol{f}}\left(\kappa^*\right)\right)^{-1}\left(\boldsymbol{f}^*-\mu\right)\right)^T\left(\left(\sigma L\left(\kappa^*\right)\right)^{-1}\left(\boldsymbol{f}^*-\mu\right)\right)\right)^T \\
&= \exp\left(-\frac{1}{2}\boldsymbol{\omega}^{*T}\boldsymbol{\omega}^*\right).
\end{aligned}$$

This sampling scheme is equivalent to a straightforward sampling scheme for the uncentered parameterization and is therefore the Metropolis-Hastings approach to which Christensen et al. (2000) and Christensen and Waagepetersen (2002) compare their Langevin approach. To allow one to move $\boldsymbol{f}$ separately from κ, I suggest having a separate proposal for $\boldsymbol{f}$ and making v in (3.16) small. In fact, using the development of reversible jump MCMC (Green 1995), one can show that setting $v = 0$ is allowable so long as one includes in the Hastings ratio the Jacobian of the deterministic mapping $\boldsymbol{f}^* = \mu + \sigma L_{\boldsymbol{f}}\left(\kappa^*\right)\left(\sigma L_{\boldsymbol{f}}(\kappa)\right)^{-1}\left(\boldsymbol{f} - \mu\right)$; this Jacobian cancels with the ratios of determinants from the priors, and the result is the same as if $v > 0$.

To begin the development of the posterior mean centering approach, let's consider sampling the simple model (3.1-3.3) with Gaussian likelihood. Conditional on a proposal κ^* and on the data $\boldsymbol{y}$, a Gibbs sample for $\boldsymbol{f}$ is

$$\boldsymbol{f}^* \sim \mathrm{N}\left(\mu + C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1}(\boldsymbol{y}-\mu), C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1}C_{\boldsymbol{Y}}\right). \qquad (3.17)$$

This suggests that if I want a joint proposal for $(\kappa, \boldsymbol{f})$ that gives $\boldsymbol{f}^*$ consistent with both κ^* and the data, I should first sample κ^* and conditional on κ^*, sample $\boldsymbol{f}^*$ from the conditional distribution (3.17). This should allow for large movements in κ space. However, even knowing the conditional distribution (3.17), I may not want to draw a proposal for $\boldsymbol{f}$ from it because it requires calculating the conditional variance and its Cholesky, which is much more computationally intensive than calculating the conditional mean in (3.17). Instead, let's use the conditional mean, but rather than using the conditional variance, draw the sample so that the variance part of the proposal is consistent with the prior covariance $C_{\boldsymbol{f}}(\kappa^*)$, but not necessarily with the likelihood. The suggested proposal for $\boldsymbol{f}$ uses the conditional mean from (3.17) in place of μ in the proposal (3.15):

$$\boldsymbol{f}^* \sim \mathrm{N}\left(\tilde{\boldsymbol{f}}^* + \sigma L_{\boldsymbol{f}}\left(\kappa^*\right)\left((\sigma L_{\boldsymbol{f}}(\kappa))^{-1}(\boldsymbol{f}-\tilde{\boldsymbol{f}})\right), v^2 R_{\boldsymbol{f}}\left(\kappa^*\right)\right), \tag{3.18}$$

where $\tilde{\boldsymbol{f}}$ is the conditional mean based on κ, and $\tilde{\boldsymbol{f}}^*$ is the conditional mean based on κ^*. The use of $L_{\boldsymbol{f}}(\kappa)^{-1}$ decorrelates the deviation of the current sample from its conditional mean, and $L_{\boldsymbol{f}}(\kappa^*)$ recorrelates the deviation based on the current correlation parameter so that the proposal is consistent with the current prior covariance. Adding $\tilde{\boldsymbol{f}}^*$ then ensures that the proposal is centered on the new conditional mean for $\boldsymbol{f}$. In Figure 3.3 I give an example of how the sample of $\boldsymbol{f}$ changes when one proposes to move from a large value of κ to a much smaller value.

Assuming that κ is proposed using a Metropolis step, the Hastings ratio for this joint proposal is

$$\begin{aligned}
&\frac{\frac{1}{|L_{\boldsymbol{f}}(\kappa)|}\exp\left(-\frac{1}{2v^2}\left(\boldsymbol{f}-\tilde{\boldsymbol{f}}-\sigma L_{\boldsymbol{f}}(\kappa)\boldsymbol{\chi}^*\right)^T\left(L_{\boldsymbol{f}}(\kappa)L_{\boldsymbol{f}}(\kappa)^T\right)^{-1}\left(\boldsymbol{f}-\tilde{\boldsymbol{f}}-\sigma L_{\boldsymbol{f}}(\kappa)\boldsymbol{\chi}^*\right)\right)}{\frac{1}{|L_{\boldsymbol{f}}(\kappa^*)|}\exp\left(-\frac{1}{2v^2}\left(\boldsymbol{f}^*-\tilde{\boldsymbol{f}}^*-\sigma L_{\boldsymbol{f}}(\kappa^*)\boldsymbol{\chi}\right)^T\left(L_{\boldsymbol{f}}\left(\kappa^*\right)L_{\boldsymbol{f}}\left(\kappa^*\right)^T\right)^{-1}\left(\boldsymbol{f}^*-\tilde{\boldsymbol{f}}^*-\sigma L_{\boldsymbol{f}}\left(\kappa^*\right)\boldsymbol{\chi}\right)\right)} \\
&\qquad = \frac{\frac{1}{|L_{\boldsymbol{f}}(\kappa)|}}{\frac{1}{|L_{\boldsymbol{f}}(\kappa^*)|}}
\end{aligned} \tag{3.19}$$

where $\boldsymbol{\chi} = (\sigma L_{\boldsymbol{f}}(\kappa))^{-1}(\boldsymbol{f}-\tilde{\boldsymbol{f}})$ and $\boldsymbol{\chi}^* = (\sigma L_{\boldsymbol{f}}(\kappa^*))^{-1}(\boldsymbol{f}^*-\tilde{\boldsymbol{f}}^*)$. As in the simpler joint proposal (3.15), the Hastings ratio is just the ratio of the determinants of the Cholesky factors and accounts for the biased movement in the $\boldsymbol{f}$ space caused by the changing size of the $\boldsymbol{f}$ space conditional on κ. Once again, it is allowable to set $v = 0$, namely to use a deterministic proposal for $\boldsymbol{f}^*$ conditional on κ^*, provided that we include the Jacobian of the deterministic mapping, which is the same as the Hastings ratio above (3.19).

A nice feature of the PMC approach is that I can avoid ever calculating $L_{\boldsymbol{f}}(\kappa)^{-1}$ or $|L_{\boldsymbol{f}}(\kappa)|$, which are not possible to calculate with the generalized Cholesky that zeroes out columns. Why is this possible? First, the determinant is not needed because the ratio of determinants in the Hastings ratio (3.19) cancels the ratio of determinants from the prior distribution for $\boldsymbol{f}$. Second, calculation of the inverse is not needed even though it appears the expression for $\boldsymbol{f}^*$. To see why, consider

$$\boldsymbol{\chi} \;=\; (\sigma L_{\boldsymbol{f}}(\kappa))^{-1}(\boldsymbol{f}-\tilde{\boldsymbol{f}})$$

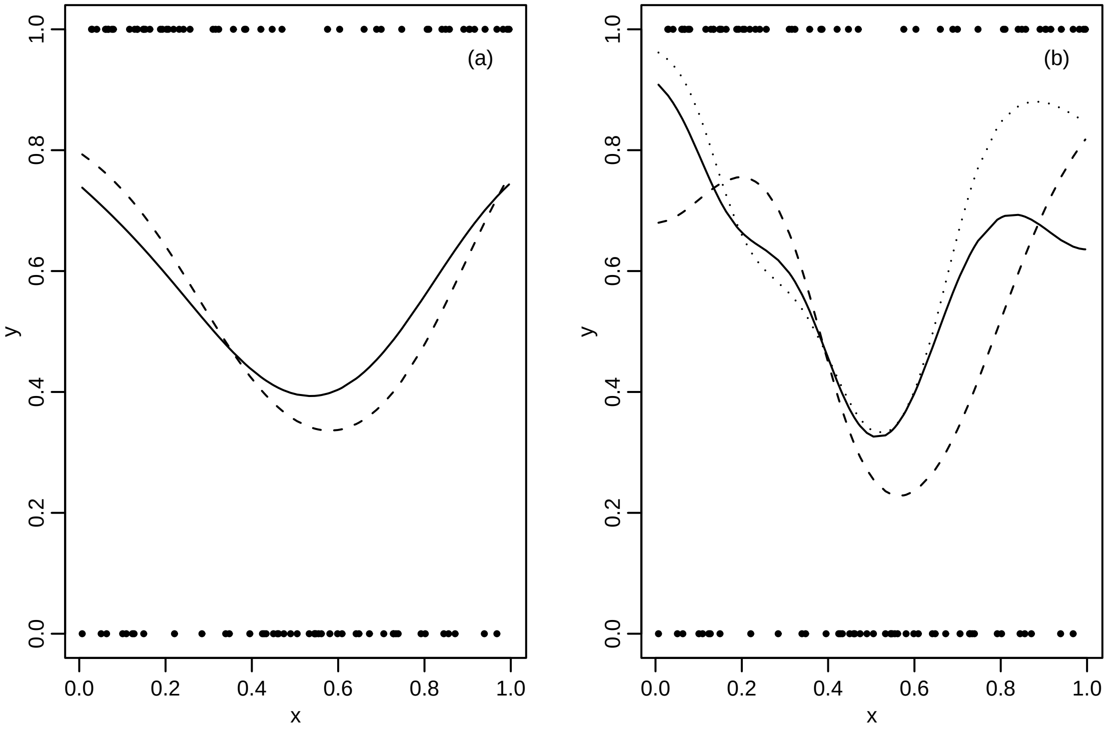

Figure 3.3. Sample function values from an MCMC in a Bernoulli data example with two different values of κ: (a) Sample function (solid line) and conditional posterior mean (dashed line) with $\kappa = 0.70$. (b) Proposing $\kappa^* = 0.30$ and $\boldsymbol{f}^*$ conditional on κ^* using the PMC proposal induces the PMC sample function proposal (solid line) and conditional posterior mean (dashed line). The dotted line is the sample function that would be proposed based on a joint proposal for $(\kappa, \boldsymbol{f})$ without posterior mean centering. Notice that the function proposed without PMC is more extreme than the PMC proposal. Also notice that the conditional posterior mean and sample function proposal are less smooth in (b), but the deviations of the sample function in (a) and the PMC sample function proposal in (b) about their conditional means have similar structure.

$$
\begin{aligned}
&= (\sigma L_{\boldsymbol{f}}(\kappa))^{-1}\left(\boldsymbol{f}-\mu-C_{\boldsymbol{f}}\left(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}\right)^{-1}(\boldsymbol{y}-\mu)\right) \\
&= \boldsymbol{\omega}+\sigma L(\kappa)^{T}\left(C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}\right)^{-1}(\boldsymbol{y}-\mu).
\end{aligned}
$$

We see first that $\boldsymbol{\chi}$ is a function of the implicit parameter $\boldsymbol{\omega}$, which we carry along and keep current in the MCMC scheme so as to avoid calculating $L_{\boldsymbol{f}}(\kappa)^{-1}$. Second, the only inversion involves not $C_{\boldsymbol{f}}$ alone, which we often cannot do, but $C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}$, for which even relatively small noise variance, η^2, makes $C_{\boldsymbol{f}}+C_{\boldsymbol{Y}}$ numerically non-singular. Using this relationship between $\boldsymbol{\omega}$ and $\boldsymbol{\chi}$ we can move through the MCMC iterations, updating $\boldsymbol{f}$, $\boldsymbol{\omega}$, and $\boldsymbol{\chi}$ without ever needing the inverse of the Cholesky. Prediction at unobserved locations can also be done without inverting $C_{\boldsymbol{f}}$. Let $\boldsymbol{f_1}$ be the function at observed locations and $\boldsymbol{f_2}$ at unobserved locations, with $L'_{\boldsymbol{f}}(\kappa)$ the generalized Cholesky factor of the full correlation of both these sets of locations. We can sample $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ as

$$
\begin{pmatrix} \boldsymbol{f_1} \\ \boldsymbol{f_2} \end{pmatrix} = \mu + \sigma L'_{\boldsymbol{f}}(\kappa) \begin{pmatrix} \boldsymbol{\omega} \\ v\boldsymbol{\psi} \end{pmatrix}, \tag{3.20}
$$

where $\boldsymbol{\psi}$ is sampled from a multivariate standard normal.

3.6.2.3 Application of posterior mean centering to non-Gaussian data

The posterior mean centering scheme may be of interest in certain cases with Gaussian likelihoods where one does not want to integrate the process out of the model for some reason. But the more important application of the scheme is when the process cannot be integrated out of the model. MCMC mixing in such models can be quite slow (Christensen and Waagepetersen 2002). In that case, we do not know the exact conditional posterior mean, but in certain situations we may be able to approximate it in a way that gives good results. Remember that this is a proposal, and there is no reason we have to use the exact posterior mean, just as in the scheme above we do not use the exact conditional variance for $\boldsymbol{f}$ even though it is available to us. In Chapter 5 I describe such an approximation for the log residual variances in the model. Here I implement the scheme in a generalized regression framework.

For generalized nonparametric regression models in which the error distribution is not assumed Gaussian, the standard GP model takes the following form,

$$
Y_i \sim \mathrm{D}\left(g\left(f\left(\boldsymbol{x_i}\right)\right)\right)
$$

$$f(\cdot) \quad \sim \quad \text{GP}\left(\mu, \sigma^2 R_{\boldsymbol{f}}(\cdot,\cdot)\right),$$

where D is an appropriate distribution, such as the Poisson for count data or the binomial for binary data, and g is an appropriate inverse link function. In the non-Gaussian case, not only are the observations not Gaussian, but the observations and the process are on different scales because of the link function. In order to propose a new set of values for the process using an approximation to the conditional posterior mean of the form,

$$\mu + C_{\boldsymbol{f}} \left(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1} (\boldsymbol{y} - \mu), \tag{3.21}$$

we need all the quantities on the same scale as the process values. Ideally we could apply the link function to the observations, but this is not feasible, as can be seen with the log link for Poisson data when a zero is observed or the logit link for binomial data whenever all successes or all failures are observed. The iteratively-reweighted least squares algorithm (IRLS) (Hastie and Tibshirani 1990) provides a strategy for solving the problem. IRLS expresses each observation as the first two terms in a Taylor expansion of the observation about its mean and uses the variance of the approximation in fitting a weighted least squares regression. I suggest the same approach in the Gaussian process case as a way to devise a PMC algorithm for non-Gaussian data. First express the linearized observation, y_i', through the Taylor expansion,

$$y_i' = g^{-1}(y_i) \approx f(\boldsymbol{x_i}) + \frac{\partial g(\boldsymbol{x_i})}{\partial f(\boldsymbol{x_i})}(y_i - g(\boldsymbol{x_i})).$$

Take the matrix $C_{\boldsymbol{Y}}'$ to be a diagonal matrix, with diagonal elements,

$$\left(C_{\boldsymbol{Y}}'\right)_{ii} = \text{Var}(Y_i') \approx \text{diag}\left(\left(\frac{\partial g(\boldsymbol{x_i})}{\partial f(\boldsymbol{x_i})}\right)^2 \text{Var}(Y_i)\right),$$

and substitute $C_{\boldsymbol{Y}}'$ and $\boldsymbol{y}'$ into (3.21). In the Poisson case, with log link, we have

$$y_i' = f(\boldsymbol{x_i}) + \frac{y_i - g(\boldsymbol{x_i})}{g(\boldsymbol{x_i})},$$

where $g(\boldsymbol{x_i}) = \exp(f(\boldsymbol{x_i}))$, with the diagonal elements of $C_{\boldsymbol{Y}}'$ being $\frac{1}{g(\boldsymbol{x_i})}$. In the binomial setting, with m_i trials, we have

$$y_i' = f(\boldsymbol{x_i}) + \frac{y_i - m_i g(\boldsymbol{x_i})}{m_i g(\boldsymbol{x_i})(1 - g(\boldsymbol{x_i}))}$$

with the diagonal elements of C'_Y being $\frac{1}{m_i g(\boldsymbol{x_i})(1-g(\boldsymbol{x_i}))}$. Using the logit link, we would have $g(\boldsymbol{x_i}) = \frac{\exp(f(\boldsymbol{x_i})}{1+\exp(f(\boldsymbol{x_i}))}$.

The Hastings ratio for this proposal is slightly complicated due to the fact that the value of $\boldsymbol{y}'$ at the current step and the value of $\boldsymbol{y}'$ one would have if one were at the proposed value $\boldsymbol{f}^*$ are different. (We can't use the $\boldsymbol{y}'$ based on $\boldsymbol{f}^*$ because we are still in the process of calculating $\boldsymbol{f}^*$ and that would introduce circularity into the setup.). As a result the Hastings ratio is

$$\frac{\exp\left(-\frac{1}{2}(\boldsymbol{\chi}^{*0} - \boldsymbol{\chi}^{**})^T(\boldsymbol{\chi}^{*0} - \boldsymbol{\chi}^{**})\right)}{\exp\left(-\frac{1}{2}(\boldsymbol{\chi}^{0*} - \boldsymbol{\chi}^{00})^T(\boldsymbol{\chi}^{0*} - \boldsymbol{\chi}^{00})\right)},$$

where

$$\begin{aligned}
\boldsymbol{\chi}^{00} &= \boldsymbol{\omega} + \sigma L_{\boldsymbol{f}}(\kappa)^T (C_{\boldsymbol{f}} + C'_{\boldsymbol{Y}})(\boldsymbol{y}' - \mu) \\
\boldsymbol{\chi}^{0*} &= \boldsymbol{\omega}^* + \sigma^* L_{\boldsymbol{f}}(\kappa^*)^T (C^*_{\boldsymbol{f}} + C'_{\boldsymbol{Y}})^{-1}(\boldsymbol{y}' - \mu^*) && (3.22) \\
\boldsymbol{\chi}^{*0} &= \boldsymbol{\omega} + \sigma L_{\boldsymbol{f}}(\kappa)^T (C_{\boldsymbol{f}} + C'^*_{\boldsymbol{Y}})^{-1}(\boldsymbol{y}'^* - \mu) && (3.23) \\
\boldsymbol{\chi}^{**} &= \boldsymbol{\omega}^* + \sigma^* L_{\boldsymbol{f}}(\kappa^*)^T (C^*_{\boldsymbol{f}} + C'^*_{\boldsymbol{Y}})^{-1}(\boldsymbol{y}'^* - \mu^*). && (3.24)
\end{aligned}$$

The quantities marked by * are calculated based on the proposed values for $\boldsymbol{f}$ and the hyperparameters. Note that in this scheme we once again do not need to use the inverse of the generalized Cholesky. The joint proposal for a new hyperparameter and for $\boldsymbol{f}$ conditional on the hyperparameter involves the following steps. First propose the hyperparameter, either μ^*, σ^*, or κ^*. Next take

$$\boldsymbol{f}^* = \tilde{\boldsymbol{f}}^* + \sigma^* L_{\boldsymbol{f}}(\kappa^*)\left(\boldsymbol{\chi}^{0*}\right),$$

where $\boldsymbol{\chi}^{0*} \sim N(\boldsymbol{\chi}^{00}, v^2 I)$. $\boldsymbol{\omega}^*$ is then calculated based on $\boldsymbol{\chi}^{0*}$ (3.22), and in turn, $\boldsymbol{\chi}^{*0}$ and $\boldsymbol{\chi}^{**}$ are calculated using (3.23) and (3.24). If the proposal is accepted, the new value of $\boldsymbol{\chi}^{00}$ is $\boldsymbol{\chi}^{**}$, not $\boldsymbol{\chi}^{0*}$, since we now know the value $\boldsymbol{f}^*$. Prediction at unobserved locations is done as in the normal likelihood model (3.20).

I compare mixing using the PMC scheme to various other possible sampling schemes on a toy example. I sample $x_i \sim \mathrm{U}(0,1)$ for $i = 1, \ldots, 100$. The response variable is

$$Y_i \sim \text{Bernoulli}\left(p(x_i) = \frac{\exp(f(x_i))}{1+\exp(f(x_i))}\right),$$

where $f(x_i) = \sin(8x_i)$. I take

$$f(\cdot) \sim \text{GP}(\mu, \sigma^2 R(\cdot; \kappa, \nu)),$$

where $R(\cdot; \kappa, \nu)$ is the Matérn correlation function. I compare five sampling schemes:

1. Discrete: I use the parameterization of Higdon (1998) where $\boldsymbol{f} = K\boldsymbol{\omega}$ with $\boldsymbol{\omega} \sim \text{N}(\mu, \sigma^2)$. I take an equally-spaced grid of 30 values for $\boldsymbol{\omega}$ and use the Matérn correlation function as the weight function that determines the elements of K.

2. Uncentered: I take $\boldsymbol{f} = \mu + \sigma L_{\boldsymbol{f}}(\kappa, \nu)\boldsymbol{\omega}$ with $\boldsymbol{\omega} \sim \text{N}(0, I)$ and sample $\boldsymbol{\omega}$ directly rather than $\boldsymbol{f}$.

3. Centered with jitter: I take $\boldsymbol{f} \sim \text{N}(\mu, \sigma^2 R_f(\kappa, \nu))$ and sample in the usual fashion from the hierarchical model, with jitter added to the prior covariance matrix to avoid numerical singularity.

4. Centered with joint sampling: I take $\boldsymbol{f} \sim \text{N}(\mu, \sigma^2 R_f(\kappa, \nu))$ but in sampling any of the hyperparameters, I also sample $\boldsymbol{f}$ conditional on the proposed hyperparameter (as given specifically for κ in (3.15)). In a separate sampling step, I sample $\boldsymbol{f}^* \sim \text{N}\left(\boldsymbol{f}, v^2 R_{\boldsymbol{f}}(\kappa, \nu)\right)$.

5. PMC: I take $\boldsymbol{f} \sim \text{N}(\mu, \sigma^2 R_f(\kappa, \nu))$ but in sampling any of the hyperparameters, I also sample $\boldsymbol{f}$ conditional on the proposed hyperparameter, as given specifically for κ in (3.18) with the modifications to the non-normal likelihood case given at the beginning of this subsection. In a separate sampling step, I sample $\boldsymbol{f}^* \sim \text{N}(\boldsymbol{f}, v^2 R_{\boldsymbol{f}}(\kappa, \nu))$.

For the uncentered, centered-joint and PMC schemes I also ran the MCMC using Langevin sampling applied to the proposals for $\boldsymbol{f}$ ($\boldsymbol{\omega}$ in the case of the uncentered scheme) in the step in which $\boldsymbol{f}$ is sampled separately from the hyperparameters. I had trouble getting the discrete and centered-jitter schemes to perform reasonably when including the Langevin sampling step, so I do not report those results here. The Langevin sampling follows the description given in Section 3.6.2.1, modified as necessary for the sampling schemes above. In adjusting the proposal variances during burn-in, I attempted to achieve the acceptance rates recommended in Roberts and Rosenthal

(2001), namely, 0.44 for scalar parameters, 0.23 for vector parameters, and 0.57 for Langevin updates. I generally came quite close to these rates. Priors are taken to be relatively noninformative, but proper.

I ran the chains for 26000 iterations and retained the last 20000 for assessment of mixing. For many of the schemes, this number of iterations is not even close to sufficient to explore the posterior for some parameters, which I will discuss further in presenting the results. In Table 3.1, I report the relative computational cost for the schemes; the table shows the number of iterations that can be completed for a given scheme relative to completing one iteration of the uncentered scheme, as implemented in the statistical software R.

Table 3.1. Number of iterations that can be completed in the same time as a single iteration of the uncentered scheme.

scheme	number of iterations
discrete	6.79
uncentered	1.00
centered-jittered	0.47
centered-joint	1.01
PMC	0.68

I assess mixing of $\theta \in \{\mu, \sigma, \kappa, \nu, f(0.1), f(0.3), f(0.6), f(0.9)\}$. To evaluate the mixing, I consider time series and autocorrelation function plots, as well as the effective sample size (ESS) approach (Neal 1993, p. 105). In this approach one tries to estimate the effective number of iterations of the chain after accounting for the autocorrelation of the samples. The ESS is defined as

$$ESS \equiv \frac{K}{1 + 2\sum_{k=1}^{\infty} \rho_k(\theta)}, \tag{3.25}$$

where $\rho_k(\theta)$ is the autocorrelation at lag k for θ. In practice, I truncate the summation at the lesser of $k = 1000$ or the first k such that $\rho_k(\theta) < 0.1$, which means that some of the values for ESS are optimistic. I also considered evaluating the sample precision of the estimates, following the approach in Brockwell and Kadane (2002). However, many of the sampling schemes had not fully explored the posterior after the 20000 iterations, leading to misleading estimates of the precision relative to schemes that more fully explored the posterior. In Table 3.2, I present the ESS values (the maximum possible is $K = 20000$) and in Figure 3.4 I show time series plots for μ, σ, and κ for the five schemes without the Langevin algorithm.

Table 3.2. Effective sample size (ESS) by sampling scheme for key model parameters. $\bar{f}$ is the mean ESS for the function values, averaging over $f(x)$ for all 100 values of x.

scheme	μ	$\log\sigma$	$\log\kappa$	ν	$f(0.1)$	$f(0.3)$	$f(0.6)$	$f(0.9)$	$\bar{f}$
discrete	1984	13	34	1378	496	348	149	456	415
uncentered	29	73	220	659	303	418	116	239	239
uncentered (Lang.)	33	116	111	971	1134	506	426	446	589
centered-jitter	125	12	13	20	129	87	22	152	94
centered-joint	33	47	174	980	197	434	102	220	233
cen.-joint (Lang.)	41	125	201	993	670	493	382	538	477
PMC	2225	288	646	1132	466	390	254	378	356
PMC (Lang.)	1715	422	874	1097	771	846	501	737	674

These indicate that when considering the parameters as a whole the PMC scheme is clearly the best. In particular, for $\log\sigma$ and $\log\kappa$ PMC handily outperforms all of the other methods. In comparing mixing of the function values, without the Langevin approach, many of the methods are relatively similar. The Langevin approach improves the mixing of $\boldsymbol{f}$ for the uncentered, centered-joint, and PMC schemes and seems to somewhat improve mixing for some of the hyperparameters. Table 3.2 does not account for the differing computational efficiencies of the methods. If we adjust by replacing K in (3.25) with the number of samples one could draw for each scheme in the time that it would take to draw 20000 for the PMC scheme, we get the adjusted ESS values in Table 3.3. We see that the results have changed qualitatively in two regards. First, the centered-joint and uncentered parameterizations are competitive with PMC with respect to the function values, but still not with respect to the hyperparameters. Second, because the discrete scheme is much faster than any other scheme, for many of the parameters, the discrete scheme now seems to mix

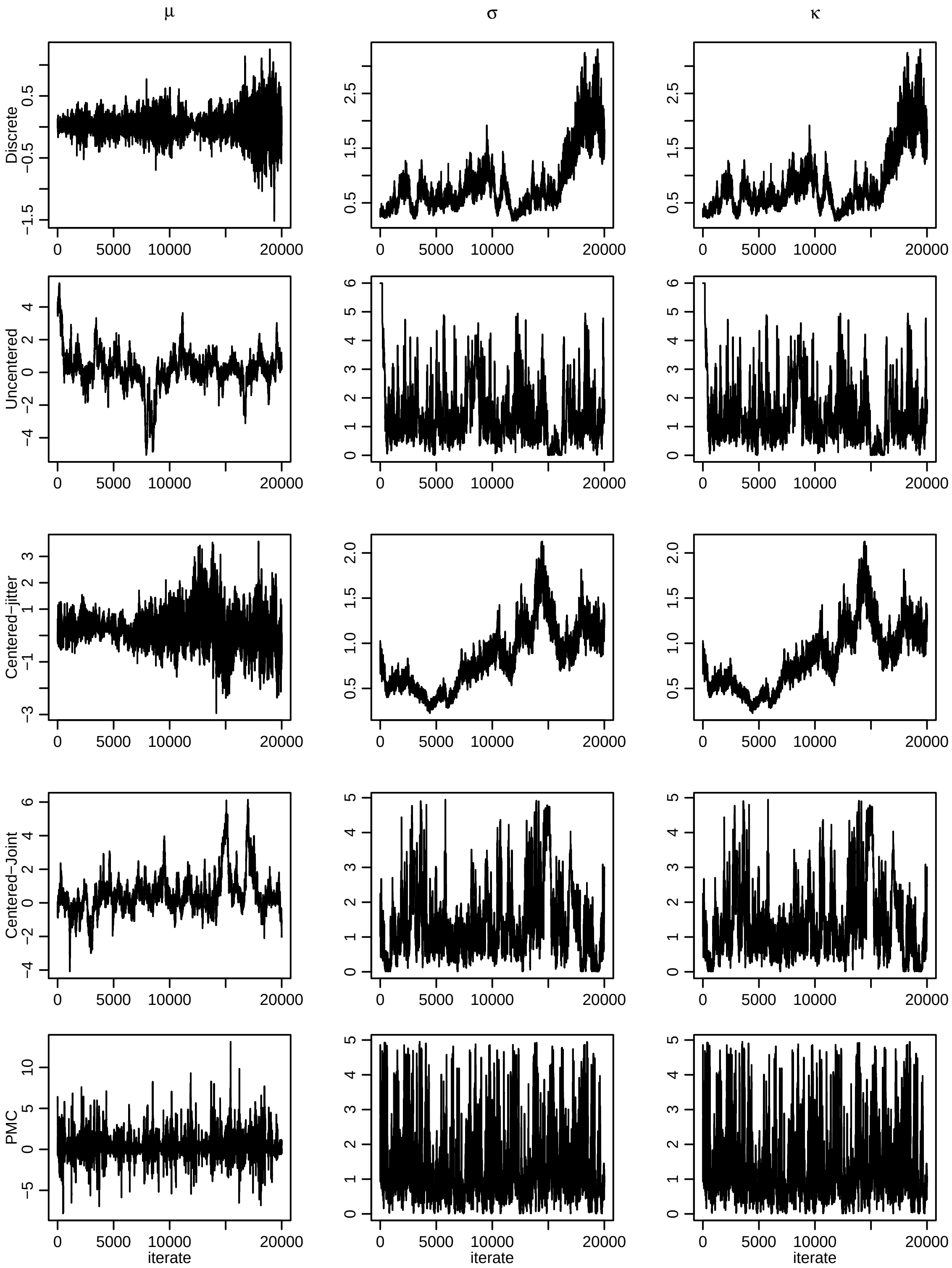

Figure 3.4. Time series plots of μ, σ, and κ for the five basic sampling schemes..

Table 3.3. ESS by sampling scheme for key parameters, adjusted for computational speed. $\bar{f}$ is the mean ESS for the function values, averaging over $f(x)$ at all 100 values of x.

scheme	μ	$\log \sigma$	$\log \kappa$	ν	$f(0.1)$	$f(0.3)$	$f(0.6)$	$f(0.9)$	$\bar{f}$
discrete	19924	133	343	13836	4979	3490	1498	4583	4144
uncentered	42	108	325	974	448	617	171	353	351
uncen. (Lang.)	48	171	163	1428	1667	744	627	655	866
centered-jitter	87	8	9	14	90	61	16	106	65
centered-joint	48	70	258	1458	293	645	152	328	346
cen.-joint (Lang.)	61	186	299	1475	996	732	567	799	708
PMC	2225	288	646	1132	466	390	254	378	372
PMC (Lang.)	1715	422	874	1097	771	845	501	737	674

best. However, this is only the case if we are comfortable with the slow mixing of $\log \sigma$ and $\log \kappa$. In fact, for the discrete scheme the adjusted ESS values for these parameters exaggerate the competitiveness of the scheme. Even the low ESS values in Table 3.3 are optimistic because I cut off the infinite sum at lag 1000. Cutting off the sum at lag 5000 would change the ESS values for the discrete scheme for $\log \sigma$ to 74 and for $\log \kappa$ to 218, which suggests the extremely slow mixing of these parameters is difficult to overcome even with extremely long chains. These parameters are crucial in that they control the flexibility of the regression function, and we seem to be exploring only a portion of the posterior for these parameters in the discrete sampling. Because of this, I would be very uncomfortable using the discrete approach in place of PMC. In sum, the evidence suggests that the PMC sampling scheme dramatically improves the mixing of the hyperparameters relative to the alternative approaches, and that use of the Langevin algorithm in addition to PMC improves the mixing of the function values.

In Section 4.6.4, I demonstrate the success of posterior mean centering with an approximate posterior mean for a binomial data example used in Biller (2000) with similar mixing results to those shown here. In Chapter 5, I use a different PMC scheme in the sampling of the residual variance field and see much faster, albeit still slow, mixing of the hyperparameters for the variance process. In that case, I use the same approximate posterior mean for the centering at every iteration, because a reasonable approximate mean for each iteration is complicated and makes the Hastings ratio difficult to calculate.

3.7 Computational Challenges of GP Models

Fitting GP models presents computational challenges because of the O(n^3) computations involved in calculations involving the covariance matrices. If one is not fitting the model via MCMC, then analysis of relatively large $n,$ say $n \in (1000, 10000)$, or possibly even larger, is feasible, but when the GP is fit within an MCMC, one needs to perform the matrix calculations repeatedly, so feasible sample sizes fall into the hundreds rather than the thousands. This can be partially alleviated if the process can be integrated out of the model, in which case one can just sample the process conditional on the iterations of the MCMC for the remaining parameters, hopefully a relatively small number of times by subsampling the process at long enough intervals so that the hyperparameter draws are approximately independent. However, even in this situation, the likelihood still involves the marginal covariance matrix of the n observations. Unfortunately, there is no simple closed form for a sample from the process conditional on the observations and the current hyperparameter values. This contrasts unfavorably with the free-knot spline model in which sampling the process does not involve matrix inversion and generally involves a basis function matrix with many fewer basis functions than observations.

In practice, fitting spline-based methods using MCMC is much quicker than fitting the nonstationary GP model via MCMC. However, mixing and convergence are an issue for such competing methods, just as they are for the GP method. (See a discussion of this in Biller (2000) with respect to spline models embedded in generalized linear models.) In particular, the movement between model spaces of differing dimension can make convergence assessment difficult.

3.7.1 Local methods

The primary problem with the GP calculations is that they are global in the sense that the prediction for each observation involves calculation with all the other observations, even those far from the focal observation that have practically no influence upon the prediction. Even if many correlations were exactly zero, prediction at the focal observation still involves the n by n matrix inversion problem. Being able to turn the GP model into one involving local calculations would offer a great computational benefit.

One approach is to divide the covariate space and use local models. Several attempts have been made to use local GPs with their own stationary covariance, and then knit the GPs together, which achieves nonstationarity by using different stationary covariances in different regions. Holmes, Mallick, and Kim (2002) use a Voronoi tessellation of the space and fit stationary GPs in the regions. Rasmussen and Ghahramani (2002) assign individual locations to one of a set of stationary GPs, with the assignment governed by a Dirichlet process and not restricted by location. The key problem here is to choose how to divide the space, which entails its own challenges.

3.7.2 Sparse methods

The local models just described are sparse in the sense that prediction at a focal observation is based only a subset of the observations. Sparsity can also be achieved in other ways. A promising approach currently under investigation in the machine learning community is the use of reduced rank approximations to the covariance matrix, working under the notion that not all of the information in the matrix is required to capture the essence of the dependence structure. There are various methods whose details differ (Smola and Bartlett 2001; Williams and Seeger 2001; Williams, Rasmussen, Schwaighofer, and Tresp 2002; Seeger and Williams 2003). The general approach is to choose a submatrix, C_{mm}, of size m by m from the prior covariance matrix and perform the matrix inversion on the submatrix. Following the development in Seeger and Williams (2003), the reduced rank approximation to the covariance is $C \approx \tilde{C}_{nn} = C_{nm}C_{mm}^{-1}C_{nm}^{T}$ where the subscripts indicate the size of the matrices, based on n observed data points and $m < n$. This corresponds to selecting a subset of the covariates, and the methods differ in how they approach this optimization problem, with tradeoffs between computational speed and optimality. From the basis function perspective, using the reduced rank covariance corresponds to using fewer basis functions to represent the function, and the approach can be thought of as a sparse representation in the weight-space. The approach inherently deals with the numerical singularities that arise in GP calculations by explicitly working with the reduced rank covariance.

Some of the reduced rank approaches deal primarily with estimating the posterior mean function conditional on fixed hyperparameters (e.g., Smola and Bartlett (2001)), while Seeger and Williams (2003) suggest optimizing the hyperparameters based on the reduced rank approximation to the marginal likelihood (with respect to the function values), including an approximation of the determinant involved. This is essentially an empirical Bayes approach, but without a prior over the hyperparameters. The approach appears to be generally successful, greatly reducing the computational cost at a limited cost in terms of error, provided the eigenvalues of the covariance decay sufficiently rapidly relative to the error variance and the rank is not reduced too drastically (Williams and Seeger 2001; Williams et al. 2002). At this point, it's not clear how useful the approach would be for MCMC sampling from the full Bayesian model since any approximation to the posterior or marginal posterior will change the stationary distribution, although perhaps not substantively. For the highly-parameterized nonstationary covariance that I employ, it's even less clear how one would proceed, because of the large number of parameters and the need to sample the processes that construct the kernel matrices that determine the nonstationary covariance.

Sparsity in the function-space view involves the use of sparse covariance matrices, namely matrices with many zeroes. Calculations such as the Cholesky decomposition and solutions to linear equations can be done more quickly with sparse matrices than with non-sparse ones, and there are many algorithms and established computer code for performing these calculations. For the covariance matrices in GP priors, if the correlation falls off quickly enough, many elements may be nearly zero. Unfortunately, we cannot just set all the elements below some threshold to zero and still ensure positive definiteness, loss of which would entail being unable to calculate the Cholesky factor or possibly even a reasonable generalized Cholesky factor. An alternative is to enforce sparsity in some other way.

One way is to use kernels that are zero beyond a certain distance (compactly-supported kernels) and calculate the covariance as the convolution of kernels $C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int_{\Re^P} K_{\boldsymbol{x_i}}(\boldsymbol{u}) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u}$.

The covariance will be positive definite by construction if the kernels are solely a function of their location. The drawback to this approach is that there is likely not a closed form for the covariance, so the integration would have to be done numerically, which would increase the computations in calculating the covariance matrix, possibly more than is gained in using sparse matrix calculations.

An alternative is to use compactly-supported correlation functions (Gaspari and Cohn 1999; Gneiting 1999); one way to create such functions is to multiply a base correlation function by another correlation function that is zero beyond a fixed distance (Gneiting 2001). Nonstationarity is obtained if the base correlation function is nonstationary. However, lack of sample path smoothness based on either correlation function will carry over to the product correlation function since mean square differentiability depends on being able to differentiate the correlation function, hence both terms in the product. The simple cases of correlation functions identically zero beyond a fixed distance, such as the spherical and cubic correlation functions (Abrahamsen 1997), do not give sample paths that are mean square or sample path differentiable. Gneiting (2001) gives the compactly-supported correlation function:

$$R(\tau) = (1-\tau)\frac{\sin(2\pi\tau)}{2\pi\tau} + \frac{1}{\pi}\frac{1-\cos(2\pi\tau)}{2\pi\tau}\ , 0 \leq \tau \leq 1$$

which minimizes $R''(0)$ amongst the functions positive definite on $\Re^3$. If this function is positive definite in higher dimensions, it may also be useful in such cases. I investigated this approach in the spatial modelling but found that the compactifying correlation function not only affects the modelled correlation at long distance scales, but also at short distance scales, where I would like the original base correlation function to be the primary influence. Furthermore, there are limits to the computational gain that can be achieved with sparse covariance matrices, since one is still performing matrix computations with large, albeit sparse, matrices.

3.7.3 Approximate matrix calculations

Much research has focused on efficiently computing with large matrices, in particular approximately solving large systems of linear equations. One method for doing this is the conjugate gradient algorithm (Golub and van Loan 1996, Section 10.2). Such methods may be useful for GP models, although the use of an approximation in the calculation of the prior will in principle change the stationary distribution of the chain and in practice may therefore give results too far from the real distribution. An important issue in employing such methods is to decide how accurate the solution needs to be and whether this requires so many iterations of the iterative methods used for approximating the solution that the computational savings are meager relative to finding the exact solution. Also, in addition to doing calculations of the form $C^{-1}\boldsymbol{b}$, I need to calculate the determinant of C. Skilling (1989, 1993) discusses methods for doing this, but they seem to be less well-developed than the $C^{-1}\boldsymbol{b}$ approximation. Williams and Barber (1998) tangentially note poor performance with these approximate methods for classification problems. Gibbs and MacKay (1997) find the values of the hyperparameters that maximize the posterior after integrating the process out of the model. Based on the work of Skilling (1989, 1993), they use the conjugate gradient method and an approximation to the trace of the covariance (which is part of the derivative of the determinant) to do the maximization efficiently.

In the fully Bayesian context, treating the hyperparameters as uncertain and approximating both $C^{-1}\boldsymbol{b}$ and the determinant of C leads one to the centered parameterization and the necessity of employing jitter. As I have demonstrated in Section 3.6.2.3, this sampling scheme does not mix well, so computational savings gained in the approximations may be lost in having to run the sampler for more iterations. The other parameterizations and sampling schemes I have outlined make explicit use of the Cholesky of the covariance, but I do not know of any fast approximations for creating the Cholesky factor or calculating $L\boldsymbol{b}$ based only on knowledge of C.

3.7.4 Parallel processing

For problems that justify the extra programming effort, parallel processing offers another alternative to speed the calculations. The most straightforward way to make use of parallel processing is to run multiple chains and combine the results at the end. The one drawback to this is that burn-in must be ensured at the start of each chain, so if the burn-in time is long, much of the parallel processing time can be spent on burn-in iterations rather than on the useable iterations. Also, this requires effort by the user to ensure that burn-in has occurred or to automate the determination. A promising alternative is the technique of Brockwell and Kadane (2002), which splits one chain amongst multiple processors in a clever way. The main issue for employing this technique is adapting it so it works in the high-dimensional spaces involved when the process itself or the parameters of the nonstationary covariance must be sampled. Finally, there are parallel versions of the Cholesky decomposition (Golub and van Loan 1996, Section 6.6), the construction of which is the primary computational cost of the model. I do not know how large the matrix has to be to make the parallel version more efficient than a non-parallel version, since the communication cost of parallelizing must be amortized.

3.7.5 Non-Bayesian approaches

For large problems in which the computations become an serious impediment, it may be useful to think of non-Bayesian or approximately Bayesian ways of fitting GP models. Maximizing the hyperparameters based on the marginal likelihood and using the conditional distribution of the process is one such approach. It may also be possible to fit the model using classical techniques such as restricted maximum likelihood, described in Higdon (2002), but doing this in the high-dimensional nonstationary model is likely to be difficult.

3.7.6 Fast Fourier transform

Christensen et al. (2000) discuss the use of the FFT to efficiently calculate a matrix square root by embedding the locations under analysis in a larger rectangular grid. This approach may be more applicable for stationary GPs than nonstationary ones, since the nonstationary parameterization I use requires eigenprocesses that determine the kernels and would therefore involve the computational cost of sampling these eigenprocesses and calculating the kernel convolution covariance on the larger grid.

3.7.7 Overview

At this point, I know of no well-established method for efficiently fitting the GP regression model, particularly in a fully Bayesian framework. However, much work is ongoing in this area in the machine learning community. Low-rank approximations to the covariance may make the GP model feasible for large datasets. Whether these approximations or the other approaches described in this section will be feasible for some representation of the nonstationary covariance model that I have developed is an open question.

Chapter 4

Regression Model Results

4.1 Introduction

In this chapter, I describe the Bayesian Gaussian process (GP) nonparametric regression model in detail. I discuss the model implementation via Markov chain Monte Carlo (MCMC). I then evaluate the nonstationary GP model by comparing it to competing Bayesian nonparametric methods in the literature as well as to a stationary GP nonparametric regression model. The competing methods assessed here use free-knot splines to estimate the regression function. I compare the methods on datasets whose covariates range from one- to three-dimensional and for which a normal likelihood is assumed. Some of the datasets are simulated, while others are real. I compare the methods using mean squared error to evaluate point predictions and predictive density calculations to evaluate overall fit. I also fit the model to a dataset for which a binomial likelihood is appropriate, assessing the success of the posterior mean centering sampling scheme on non-Gaussian data.

4.2 Model Structure

The simplest form of the nonstationary GP nonparametric regression model (see Figure 4.1 for a directed acyclic graph of the model) is based on a normal likelihood with a nonstationary GP prior for the regression function,

$$Y_i \sim \mathrm{N}(f(\boldsymbol{x_i}), \eta^2)$$

$$f(\cdot) \quad \sim \quad \text{GP}\left(\mu_f, \sigma_f^2 R_f^{NS}(\cdot,\cdot)\right),$$

where $R_f^{NS}(\cdot,\cdot)$ is the Matérn-based nonstationary correlation function defined by

$$R_f^{NS}(\boldsymbol{x_i},\boldsymbol{x_j}) = \frac{2^{\frac{P}{2}} \mid \Sigma_i \mid^{\frac{1}{4}} \mid \Sigma_j \mid^{\frac{1}{4}}}{\mid \Sigma_i + \Sigma_j \mid^{\frac{1}{2}}} \frac{1}{\Gamma(\nu_f)2^{\nu_f - 1}} (2\sqrt{\nu_f Q_{ij}})^{\nu} K_{\nu_f}(2\sqrt{\nu_f Q_{ij}}).$$

ν_f is the smoothness parameter of the Matérn function, and Q_{ij} is the quadratic form,

$$Q_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}).$$

For the smoothness parameter, ν_f, I place a uniform prior on $(0.5, 30)$, which allows the smoothness to vary between non-differentiable (0.5) and very smooth. Recall that as $\nu_f \to \infty$, one recovers the original Higdon kernel-based nonstationary covariance function. The results in this chapter suggest that the data do not contain much information about this parameter, apart from some indication that smoothness approaching the exponential correlation function is less likely than more smooth functions. I use vague but proper Gaussian priors: $\mu_f \sim \text{N}(0, 10^2)$, $\log \sigma_f \sim \text{N}(0, 9^2)$, and $\log \eta \sim \text{N}(0, 9^2)$. The kernel matrices, Σ_i, are modelled using the eigendecomposition described in Section 3.2.3. Each location (training or test), $\boldsymbol{x_i}$, has a Gaussian kernel with mean $\boldsymbol{x_i}$ and covariance matrix, Σ_i, whose eigendecomposition is

$$\Sigma_i = \Gamma(\gamma_1(\boldsymbol{x_i}), \ldots, \gamma_Q(\boldsymbol{x_i}))D(\lambda_1(\boldsymbol{x_i}), \ldots, \lambda_P(\boldsymbol{x_i}))\Gamma(\gamma_1(\boldsymbol{x_i}), \ldots, \gamma_Q(\boldsymbol{x_i}))^T,$$

where D is a diagonal matrix of eigenvalues and Γ is an eigenvector matrix constructed as described in Section 3.2.3.2. Using that parameterization for the eigenvector matrix, there are $Q = \frac{P(P-1)}{2} + P - 1$ spatial processes that determine the eigenvector matrices. $\gamma_q(\cdot), q = 1, \ldots, Q$, and $\lambda_p(\cdot), p = 1, \ldots, P$, are spatial processes, since implicitly there are eigenvalues and eigenvectors at all points in the covariate space. I will refer to these as the eigenvalue and eigenvector processes, or to them collectively as the eigenprocesses.

The next level of the hierarchy gives the prior distribution for the kernel matrices in terms of the eigenprocesses, $\phi(\cdot) \in \{\log(\lambda_1(\cdot)), \ldots, \log(\lambda_P(\cdot)), \gamma_1(\cdot), \ldots, \gamma_Q(\cdot)\}$. In order for the kernels to vary smoothly in covariate space, I model the eigenvalues and eigenvectors using Gaussian process priors as well, but use stationary covariance functions for simplicity. In one dimension, I need only parameterize one eigenvalue process. I do this by taking the prior for the eigenvalue process to be

$$\log(\lambda(\cdot)) \sim \text{GP}(\mu_\lambda, \sigma_\lambda^2 R_\lambda^S(\cdot)).$$

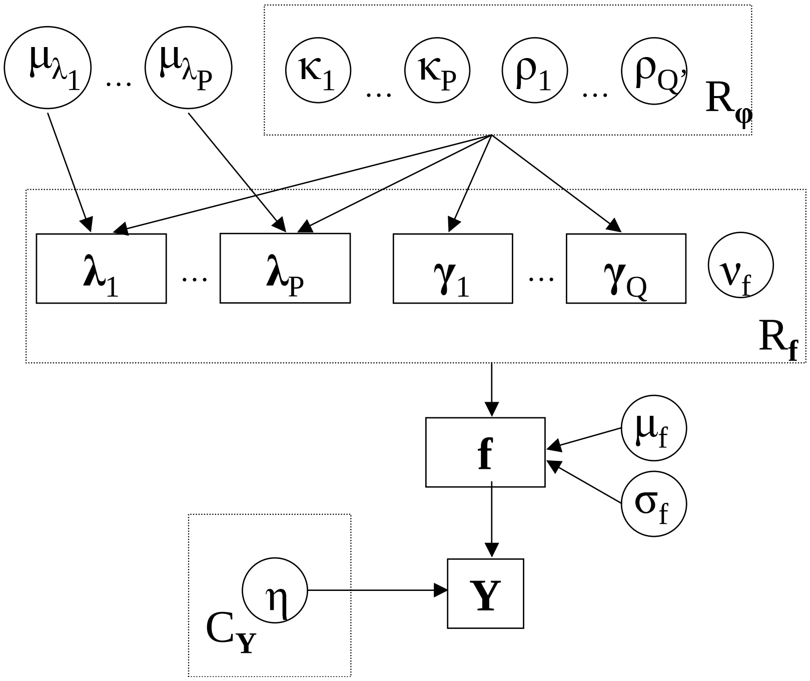

Figure 4.1. Directed acyclic graph for the normal likelihood nonstationary Gaussian process regression model. Bold letters indicate vectors.

In higher dimensions, it is difficult to model the eigenvectors simply and effectively. In part, this is simply the curse of dimensionality; the number of eigenvalue processes is the same as the dimension of the covariate space, P, while the number of eigenvector processes grows more rapidly, with a minimum of $\frac{P(P-1)}{2}$ processes (if one were able to use the Givens angle parameterization). For each of the eigenprocesses, the number of hyperparameters grows with P as well. I focus on the simpler parameterization described in detail in Section 3.2.3.3. I model each of the eigenprocesses with a stationary GP prior,

$$\phi(\cdot) \sim \mathrm{GP}(\mu_\phi, \sigma_\phi^2 R_\phi^S(\cdot)).$$

$R_\phi^S(\tau_{ij})$ is a Matérn stationary correlation function, common to all the processes and defined by the Mahalanobis distance,

$$\tau_{ij} = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_j})\Sigma_\phi(\boldsymbol{x_i} - \boldsymbol{x_j})},$$

where

$$\Sigma_{\boldsymbol{\phi}} = \Gamma(\rho_1, \ldots, \rho_{Q'})D(\kappa_1^2, \ldots, \kappa_P^2)\Gamma(\rho_1, \ldots, \rho_{Q'})^T,$$

Γ is an eigenvector matrix parameterized by the $Q' = \frac{P(P-1)}{2}$ Givens angles, $\rho_q, q = 1, \ldots, Q'$, and D is a diagonal matrix of P eigenvalues, $\kappa_p^2, p = 1, \ldots, P$. In addition to the $P + \frac{P(P-1)}{2}$ hyperparameters for this shared stationary correlation, there are separate hyperparameters, μ_ϕ and σ_ϕ, for each process. I fix the smoothness parameter of the eigenprocesses, $\nu_\phi = 30$, so that the processes are very smooth. Following the results in Section 2.5.5, this ensures that the smoothness of the regression function will be determined by ν_f and not influenced by the smoothness of the kernels. I choose not to let ν_ϕ vary because this parameter should have minimal impact on the estimate for the regression function, and the data are likely to only weakly, if at all, inform the parameter. The priors for the Givens angles follow the development in Section 3.2.3.1.

To simplify the prior specification, I assume without loss of generality, $\boldsymbol{x_i} \in [0,1]^P$, the P-dimensional unit cube. For each eigenvalue process, $\log \lambda_p(\cdot)$, I let μ_{λ_p} be a free parameter, with a vague but proper Gaussian prior, $\mathrm{N}(-4, 2.5^2)$, informed by my knowledge of the reasonable range of values for the size of the kernels. I fix $\sigma^2_{\lambda_p} = 3^2$ (2.5^2 for one-dimensional covariates), because the coarseness and range of the covariates determine a reasonable range for the size of the eigenvalues, since it is the eigenvalues that determine the size of the kernel matrices and therefore the correlation scale. For the eigenvector processes, $\gamma_q(\cdot)$, I take $\mu_{\gamma_q} = 0$ and $\sigma_{\gamma_q} = 1$ without loss of generality, because these processes are used only to determine the directions of the eigenvectors. The κ_p values parameterize the scale of the correlation. I use a Beta$(1,3)$ prior,

$$\Pi(\log(\kappa_p)) \propto (1 - g(\log(\kappa_p)))^2,$$

where the function $g(\cdot)$ maps the range I chose for $\log \kappa_p$, (-3.5,1.5), to (0,1). A uniform prior would favor small values of κ_p and unsmooth eigenprocesses, which I do not think is reasonable, nor necessary, to model the data. The upper limit on $\log \kappa_p$ prevents κ_p from wandering into a part of the parameter space where large changes in κ_p have little effect on the correlation of $\phi(\cdot)$, while the lower limit is set based on not wanting the correlation scale of the eigenprocesses to become smaller than the distances between covariates. I force $\lambda_p(\boldsymbol{x_i}) < 9P$ to avoid having the chain wander off to parts of the parameter space in which the eigenvalues are very large and from which it would take a long time to return because of the flatness of the likelihood in that region. This limit varies with P because as the number of covariates increases, in order for the model to ignore a covariate, it is necessary for the size of the kernels to become increasingly large in that direction of covariate space. However, this limit still prevents the model from completely ignoring a covariate, as discussed in Section $3.5.1.3$.

The nonstationary covariance is non-negative and decays with distance, but can decay at a different rate in different parts of the covariate space, allowing the degree of smoothing to vary. For most applications, the restriction to non-negative correlation is sensible and probably desirable, but perhaps not for some functions, such as oscillating functions.

To assess the effectiveness of the nonstationary correlation model relative to the simpler stationary alternative, I also implement the Gaussian process-based nonparametric regression method using a stationary correlation model, replacing $R_f^{NS}(\cdot,\cdot)$ with $R_f^S(\cdot)$, where the latter is of the same form as $R_\phi^S(\cdot)$ in the nonstationary model. This stationary GP prior for the function takes the same form as the stationary GP priors for the eigenprocesses described above. However, I let ν_f in the stationary Matérn correlation vary in the same way that ν_f varies in the nonstationary model.

4.3 MCMC Sampling Scheme

The sampling scheme for the regression model is built on the proposals discussed in Section $3.6.2.2$. In particular, any time I propose a hyperparameter of $\boldsymbol{f}$ or new process values for any of the eigenprocesses, $\boldsymbol{\phi}$, I also propose $\boldsymbol{f}^*$conditional on the proposed hyperparameter or eigenprocess values. Here I describe the proposal scheme step by step, starting at the bottom of the model hierarchy (Figure 4.1). In all cases, v indicates a user-tuneable proposal standard deviation that differs between parameters and $\boldsymbol{\omega}_\phi = (\sigma_\phi L_{\boldsymbol{\phi}})^{-1}(\boldsymbol{\phi} - \mu_\phi)$, where $\phi(\cdot)$ is an eigenprocess and $L_{\boldsymbol{\phi}}$ is the Cholesky factor of $R_{\boldsymbol{\phi}}^S$.

First let me describe two prototype steps that are used in the sampling scheme. The first prototype step, S1, is the basic posterior mean centering proposal involving $\boldsymbol{f}$.

1. Propose either μ_f, σ_f, or ν_f using a Metropolis-Hastings proposal or propose a single process vector, $\boldsymbol{\phi}$, or propose all of the eigenprocess vectors simultaneously, as will be described below. In the notation that follows, I will indicate that all of the potential parameters have been proposed, but this is merely for notational convenience.

2. Propose $\boldsymbol{f}$ conditionally on $\boldsymbol{\theta}^* = \{\mu_f, \sigma_f, \nu_f, \boldsymbol{\phi}\}$ as

$$\boldsymbol{f}^* \mid \boldsymbol{\theta}^* \sim \mathrm{N}(\widetilde{\boldsymbol{f}(\boldsymbol{\theta}^*)} + \sigma_f^* L_{\boldsymbol{f}}^* \boldsymbol{\chi}, v^2 R_{\boldsymbol{f}}^*),$$

   where $\boldsymbol{\chi} = (\sigma L_{\boldsymbol{f}})^{-1}(\boldsymbol{f} - \widetilde{\boldsymbol{f}(\boldsymbol{\theta})})$, $\widetilde{\boldsymbol{f}(\boldsymbol{\theta})}$ is the posterior mean of $\boldsymbol{f}$ conditional on the current parameter values, and $\widetilde{\boldsymbol{f}(\boldsymbol{\theta}^*)}$ is the posterior mean conditional on the proposed values,

$$\begin{aligned}\widetilde{\boldsymbol{f}(\boldsymbol{\theta})} &= \sigma_f^2 R_{\boldsymbol{f}} \left(\sigma_f^2 R_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1} \boldsymbol{y} + C_{\boldsymbol{Y}} \left(\sigma_f^2 R_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1} \mu_f \\ &= \mu_f + \sigma_f^2 R_{\boldsymbol{f}} \left(\sigma_f^2 R_{\boldsymbol{f}} + C_{\boldsymbol{Y}}\right)^{-1} (\boldsymbol{y} - \mu_f).\end{aligned}$$

   Note that $C_{\boldsymbol{Y}}$ is the diagonal residual covariance matrix. Note again that, as mentioned in Chapter 3, it is allowable to set $v = 0$, so long as one includes the Jacobian of the deterministic mapping, $\boldsymbol{f} \rightarrow \boldsymbol{f}^*$, which is the same as the Hastings ratio one uses if $v > 0$.

3. The Hastings ratio for the proposal is a ratio of determinants, which cancels the determinant ratio from the prior for $\boldsymbol{f}$ as described in Section 3.6.2.2.

The second prototype step, S2, is a joint proposal for an eigenprocess hyperparameter with the values of the process. Since I do not know the posterior mean of $\boldsymbol{\phi}$ conditional on the other parameters, I am forced to make a joint proposal that takes account only of the prior correlation of the process and not the posterior correlation.

1. This proposal applies to $\theta \in \left\{\mu_{\lambda_1}, \ldots, \mu_{\lambda_P}, \log \kappa_1, \ldots, \log \kappa_P, \rho_1, \ldots, \rho_{Q'}\right\}$. Propose θ using a Metropolis-Hastings proposal (for $\log \kappa_p$ I change the proposal variance as a function of the current value, requiring a Hastings correction).

2. If $\theta = \mu_{\lambda_p}$, I next propose only $\boldsymbol{\lambda}_p$. If θ is any of the hyperparameters involved in Σ_ϕ in the stationary correlation matrix $R_{\boldsymbol{\phi}}^S$, propose all of the eigenprocesses. For an individual eigenprocess, take

$$\boldsymbol{\phi}^* \mid \theta^* \sim \mathrm{N}(\mu_\phi^* + \sigma_\phi L_{\boldsymbol{\phi}}^* \boldsymbol{\omega}_\phi, v^2 R_{\boldsymbol{\phi}}^*).$$

   Note again that, as mentioned in Chapter 3, it is allowable to set $v = 0$, so long as one includes the Jacobian of the deterministic mapping, $\boldsymbol{\phi} \rightarrow \boldsymbol{\phi}^*$, which is the same as the Hastings ratio one uses if $v > 0$.

3. The Hastings ratio for this proposal is a ratio of determinants, which cancels the determinant ratio in the prior for $\boldsymbol{\phi}$.

Now let's consider the steps in the proposal scheme for the regression model, making use of prototype steps S1 and S2 as necessary.

1. Sample $\log(\eta)$ using a simple Metropolis step.

2. Separately, for each of the pairs, $(\nu_f, \boldsymbol{f}), (\mu_f, \boldsymbol{f})$, and $(\sigma_f, \boldsymbol{f})$, sample the pair jointly using a proposal of type S1. For ν_f I increase the proposal variance when ν_f is large, so the proposal requires a Hastings correction to the acceptance ratio.

3. Separately, for each of the eigenprocesses, propose $(\boldsymbol{\phi}, \boldsymbol{f})$ using a joint proposal of type S1. The marginal proposal for $\boldsymbol{\phi}$ is
$$\boldsymbol{\phi}^* \sim \mathrm{N}(\boldsymbol{\phi}, v^2 R_{\boldsymbol{\phi}}).$$

4. Separately, for each hyperparameter involved in the eigenprocesses, propose the hyperparameter using a proposal of type S2. As part of the same proposal, propose $\boldsymbol{f}$ conditionally on the new eigenprocess(es) values using a proposal of type S1.

5. Propose $\boldsymbol{f}$ using a simple Metropolis step with correlation amongst the elements of $\boldsymbol{f}$: $\boldsymbol{f}^* \sim \mathrm{N}(\boldsymbol{f}, v^2 R_{\boldsymbol{f}})$. It is also straightforward to do a Langevin update here, however I did not use such an update in the model runs reported in this thesis.

With a Gaussian likelihood, I can of course integrate $\boldsymbol{f}$ out of the model, which simplifies the scheme above and avoids using the PMC proposal scheme. In the development of the model, I was interested in being able to extend the model to non-Gaussian data, so I have sampled from the model without integrating the function out. To sample from a model for non-Gaussian data, the steps are as above, but with the modifications given in Section 3.6.2.3.

The sampling scheme for the stationary GP prior model is a simplified version of this sampling scheme, with proposals of type S1 for $(\log \kappa_p, \boldsymbol{f}), p = 1, \ldots, P$ and $(\gamma_q, \boldsymbol{f}), q = 1, \ldots, Q'$.

Since the function at test locations is conditionally independent of the observations given the function at training locations, I can generate samples of the function at test locations based on the usual conditional normal calculation,

$$\boldsymbol{f_2} \mid \boldsymbol{f_1} \sim \mathrm{N}(\mu_f + C_{\mathbf{21}}C_{\mathbf{11}}^{-1}(\boldsymbol{f_1} - \mu_f), C_{\mathbf{22}} - C_{\mathbf{21}}C_{\mathbf{11}}^{-1}C_{\mathbf{12}}),$$

where the $\mathbf{1}$ subscript indicates the training set and the $\mathbf{2}$ subscript the test set. In practice, one can compute $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ as a single vector, $\boldsymbol{f}' = \mu_f + \sigma_f L'_{\boldsymbol{f}}\boldsymbol{\omega}'_f$, where $L_{\boldsymbol{f}}$ is the Cholesky factor of the full covariance of all the locations, with the training locations in the first block and the test locations in the second, and $\boldsymbol{\omega}'_f$ is the concatenation of the original $\boldsymbol{\omega}_f = (\sigma_f L_{\boldsymbol{f}})^{-1}(\boldsymbol{f} - \mu_f)$ vector, used in sampling the training locations, with m standard normal deviates, where m is the number of test locations. These samples are used in calculating the evaluation criteria (Section 4.4) for test locations.

4.4 Evaluation Procedures

To compare the GP model to other methods, I use three criteria. In this section I use $\mathbf{1}$ to indicate the set of training locations and $\mathbf{2}$ to indicate a set of test locations.

4.4.1 Mean squared error

The first criterion is mean squared error (MSE), which judges the accuracy of the posterior mean of the model. For simulated data, the MSE is calculated with respect to the true mean function, $\check{\boldsymbol{f}}$, used to generate the data,

$$\mathrm{MSE}_{\mathbf{1}} = \frac{\sum_{i=1}^{n}(\check{f}_i - \tilde{f}_i)^2}{n},$$

where $\tilde{f}_i$ is the posterior mean, calculated by averaging over the MCMC simulations. Following DiMatteo et al. (2002) I do not calculate the MSE for test covariates in the one-dimensional simulated datasets because I am using the true mean function in the calculation of MSE rather than using the data values, and because the training data are rather dense in the covariate space and the posterior means are smooth. In the two-dimensional simulated dataset, with the training data less dense in covariate space, and the possibility of extrapolation errors, I calculate MSE for both the training and test sets.

For real datasets, I calculate MSE with respect to both test and training data as

$$\begin{aligned}\mathrm{MSE}_{\mathbf{1}} &= \frac{\sum_{i=1}^{n}(y_i - \widetilde{f}_i)^2}{n},\\ \mathrm{MSE}_{\mathbf{2}} &= \frac{\sum_{j=1}^{m}(y_j - \widetilde{f}_j)^2}{m}.\end{aligned}$$

Finally to facilitate comparison of MSE across datasets, I normalize the MSE by the variance of the function values (for simulated data), $V_{\mathbf{1}} = \sum_{i=1}^{n}(\check{f}_i - \bar{f})^2/n$, or variance of the data (for real data), $V_{\mathbf{2}} = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$, to calculate the Fraction of Variance Unexplained (FVU),

$$\mathrm{FVU}_{\mathbf{1}} = \frac{\mathrm{MSE}_{\mathbf{1}}}{V_{\mathbf{1}}},$$

which is used in Denison et al. (1998a) and Holmes and Mallick (2001), among others. I calculate the analogous quantity for test data, $\mathrm{FVU}_{\mathbf{2}}$, based on $V_{\mathbf{2}}$.

4.4.2 Predictive density

The MSE only assesses the point predictions of the model, not the distribution of the data under the model. As an alternative when analyzing the real datasets, I also calculate the usual predictive density on held-out data. This measure assesses how well the model estimates the residual variance and how well the individual draws from the posterior explain the data. I average the conditional predictive density over the posterior distribution of the parameters using the MCMC draws from the posterior. The log predictive density (LPD) for test data is

$$\begin{aligned}\mathrm{LPD}_{\mathbf{2}} &= \log h(\boldsymbol{y_2}|\boldsymbol{y_1})\\ &= \log \int h(\boldsymbol{y_2} \mid \boldsymbol{f}, \eta, \boldsymbol{y_1}) d\Pi(\boldsymbol{f}, \eta|\boldsymbol{y_1})\\ &\approx \log \frac{1}{K}\left(\sum_{k=1}^{K} \frac{1}{\sqrt{2\pi}\eta_{(k)}^{m}} \exp\left(-\frac{\sum_{j=1}^{m}(y_j - f_{j,(k)})^2}{2\eta_{(k)}^2}\right)\right), \qquad (4.1)\end{aligned}$$

where $h(\cdot)$ is the normal density function and (k) indicates the kth MCMC draw from the posterior. I also calculate (4.1) at the training observations, y_i, $i = 1, \ldots, n$, which should give a sense for how well the model predicts the data on which it was trained, although (4.1) cannot be interpreted as a predictive density, since $h(\boldsymbol{y_1}|\boldsymbol{y_1}) = 1$. I calculate $\mathrm{LPD}_{\mathbf{2}}$ for the two real datasets by calculating the LPD on each data point when held out as a test point and averaging over the data points to report one value for the whole dataset. For the training set, I average over both the training points and the data splits of the cross-validation procedure.

4.4.3 Kullback-Leibler divergence

For simulated data, I know the distribution of the data, so I can calculate the Kullback-Leibler (KL) divergence between the true distribution of the data and the posterior predictive distribution for the data. I do this by generating many test observations from the true model and evaluating the KL divergence between the density of the observations under the true distribution and the predictive density of the observations under the model. I can calculate this divergence for both the set of training covariates and a set of test covariates. In either case, we have

$$\begin{align*}
\text{KL} &= \int \log\left(\frac{h(\boldsymbol{y_2} \mid \check{\boldsymbol{f}}, \check{\eta})}{h(\boldsymbol{y_2} \mid \boldsymbol{y_1})}\right) h(\boldsymbol{y_2} \mid \check{\boldsymbol{f}}, \check{\eta}) d\boldsymbol{y_2} \\
&= \int \log\left(\frac{h(\boldsymbol{y_2} \mid \check{\boldsymbol{f}}, \check{\eta})}{\int h(\boldsymbol{y_2} \mid \boldsymbol{f}, \eta) d\Pi(\boldsymbol{f}, \eta | \boldsymbol{y_1})}\right) h(\boldsymbol{y_2} \mid \check{\boldsymbol{f}}, \check{\eta}) d\boldsymbol{y_2} \\
&= -\frac{\log(2\pi) + \log(\check{\eta}^2) + 1}{2} - \int \log\left(\int h(\boldsymbol{y_2} \mid \boldsymbol{f}, \eta)) d\Pi(\boldsymbol{f}, \eta | \boldsymbol{y_1})\right) \\
&\quad \times h(\boldsymbol{y_2} \mid \check{\boldsymbol{f}}, \check{\eta}) d\boldsymbol{y_2} \tag{4.2} \\
&\approx -\frac{\log(2\pi) + \log(\check{\eta}^2) + 1}{2} - \int \frac{1}{K} \sum_{k=1}^{K} \log(h(\boldsymbol{y_2} \mid \boldsymbol{f}_{(k)}, \eta_{(k)})) h(\boldsymbol{y_2} \mid \check{\boldsymbol{f}}, \check{\eta}) d\boldsymbol{y_2} \\
&\approx -\frac{\log(\check{\eta}^2) + 1}{2} - \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(-\log \eta_{(k)} - \frac{(y_j - f_{j,(k)})^2}{2\eta_{(k)}^2}\right), \tag{4.3}
\end{align*}$$

where the first term in (4.2) is found by integrating the log of the true normal density against itself. In calculating (4.3) I use 500 observations drawn at each covariate of interest, hence for the training covariates, I have $J = 500n$, and for test covariates, I have $J = 500m$. I use $K = 1000$ subsamples from the 20000 MCMC draws (only so as to save on storage space). Note that the KL divergence is equivalent to the LPD for comparing models because the two differ only in the first term in (4.3). However, the KL divergence allows me to compare the quantity to the absolute baseline of zero, which is the KL divergence between the true distribution and itself. In all cases I calculate the KL divergence for new observations generated from the true model and not based on training observations. However, as I discussed in the case of MSE, for the simulated data in one dimension, the KL divergence calculated for test observations generated at test covariates is unlikely to differ substantially from that calculated for test observations at training covariates, so I calculate it only for the training covariates. For the simulated dataset in two dimensions, I calculate the KL divergence separately for test observations at the test and training covariates. Note that in the papers describing the free-knot spline methods, the authors compare their methods to other methods solely on the basis of MSE (or equivalently, FVU) and do not use predictive density estimates.

4.5 Test Datasets and Competing Methods

I compare the GP model to several successful methods from the literature. For one-dimensional functions, I compare the method to the Bayesian free-knot spline model (Bayesian Adaptive Regression Splines or BARS) of DiMatteo et al. (2002). DiMatteo et al. (2002) assess the performance of BARS relative to other popular methods using three test functions, one a smoothly varying function, the second a spatially inhomogeneous function, and the third a function with a sharp jump. I use these same functions to compare the nonstationary GP regression model (NSGP) developed here to BARS. I also compare the results to the stationary GP model (SGP) to evaluate the importance of including nonstationarity in the GP model. For each test function, I generate 50 sets of noisy data from the underlying function, using the error variance given in DiMatteo et al. (2002). I then run BARS, NSGP, and SGP on these data. Since BARS outperformed the other methods to which DiMatteo et al. (2002) compared it, I simply compare the GP methods to BARS and then compare the GP methods to the other methods using DiMatteo et al. (2002, Table 1). I follow DiMatteo et al. (2002) in calculating the assessment criteria at the training locations, since in one dimension, the training locations are dense and the function estimates are smooth enough that evaluation based on training locations should not differ from that based on test locations. I will refer to these three test functions as examples 1, 2, and 3.

For higher-dimensional functions, I compare the GP methods to two free-knot spline methods, Bayesian Multivariate Linear Splines (BMLS) (Holmes and Mallick 2001) and Bayesian Multivariate Automatic Regression Splines (BMARS) (Denison et al. 1998b), a Bayesian version of the original MARS algorithm of Friedman (1991). BMLS uses piecewise, continuous linear splines, while BMARS uses tensor products of univariate splines; both are fit via reversible jump MCMC

(RJMCMC). For both methods, I consider interactions between the spline basis functions up to order two; Holmes and Mallick (2001) suggest that even when $P > 2$, using higher-order interactions in the basis functions may not be necessary. To compare the methods, I use three datasets. The first dataset is a two-dimensional test function first introduced by Hwang, Lay, Maechler, Martin, and Schimert (1994). Both Holmes and Mallick (2001) and Denison et al. (1998b) assess the performance of their methods using 225 training locations and 10000 test locations on a single simulated dataset. I also use 225 training locations, but because of the computational difficulty of working with large covariance matrices (I would need to calculate the Cholesky decomposition of matrices of size 10000 by 10000), for each simulation, I use only 225 test locations. Once again, I simulate 50 sets of noisy data from the test function. By using different test locations for each of the 50 simulations, I assess the method at $50 \cdot 225 = 11250$ test locations. The second dataset is a two-dimensional dataset of real air temperature anomalies (departures from the mean) for December 1993 used by Wood et al. (2002). Fitting the NSGP model to the full 445 observations is very slow, and I was unable to achieve convergence in reasonable time, so I chose a 109 observation subset of the original data, focusing on the Western hemisphere, $222.5°$-$322.5°$E and $62.5°$S-$82.5°$N. This allows me to assess the methods on a spatial dataset and on a dataset with clear inhomogeneous smoothness. I fit the models on 55 test/training splits of the 109 observations, with 54 splits of 107 training examples and two test examples and one split of 108 training examples and one test example, thereby including each data point as a test point once. Finally, I use a real dataset of ozone measurements (Bruntz, Cleveland, Kleiner, and Warner 1974) included in the S-plus statistical software package as the ‘air’ dataset. This dataset has been used frequently in the past (Cleveland and Devlin 1988; Denison et al. 1998b; Holmes and Mallick 2001) for methodological evaluation. The data are 111 daily records of ozone from the New York metropolitan area from the summer of 1973, and the goal is to predict the cube root of ozone based on three covariates: radiation, temperature, and wind speed. Here I do 56 test/training splits of the 111 observations, with 55 splits of 109 training examples and two test examples and one split of 110 training examples and one test example, thereby including each data point as a test point once. I will refer to these three datasets as the Hwang, Wood, and ozone datasets.

For the evaluation runs for all the methods, I use 5000 iterations for burn-in and then collect

20000 iterations for evaluation. For each dataset, I separately scale and translate each covariate so that the values lie in $[0, 1]$.

In running the GP models for these evaluations, I use different random seeds for each run as well as different initial values and different permutations of the covariate vectors. In running the BARS model, I use different random seeds, but use the default initial values in the code of DiMatteo et al. (2002). In running the BMLS and BMARS models, I use different random seeds and different permutations of the covariate vectors, the latter because knots are positioned on the data points and are added randomly during the MCMC. I use the default initial values given in the BMLS and BMARS code, which is available from the website of Dr. Chris Holmes. Note that for some runs, the acceptance rate of the RJMCMC was as high as 90%, although it's not clear if this is a problem.

In Section 3.5.1.2 I discuss the possibility of imposing a prior that constrains the degrees of freedom of the conditional posterior mean of the regression function in the GP model based on the trace of the smoothing matrix. In one dimension, this does not seem to be necessary in practice as the Occam's razor effect is sufficient to avoid undersmoothing. In higher dimensions, smoothness becomes more of a concern, and I do impose a prior on the degrees of freedom. For the simulations that follow I use a $\Gamma(2, 5)$ prior for the Hwang and ozone datasets, which gives a mean of 10 and variance of 50. For the temperature dataset of Wood et al. (2002) I use a $\Gamma(2, 10)$ prior, since I think that the function may not be very smooth.

In adjusting the proposal variances in the GP models, I attempted to achieve the acceptance rates recommended in Roberts and Rosenthal (2001), namely, 0.44 for scalar parameters and 0.23 for vector parameters. Because I used the same proposal variances for all of the simulations for a given dataset, there were parameters for which it was difficult to find proposal variances that worked well for all the simulations, but I do not believe this has much impact on the broad comparisons with other methods, although this may have had some impact on individual simulations. Of course in practice with real data, one could tune these proposal variances in a more optimal way than done in these evaluation runs. For the initial values of the GP models, I use a combination of dispersed values and estimates based on the data. I set $\mu_f = \bar{\boldsymbol{Y}}$ and $\sigma_f = \eta = S_{\boldsymbol{Y}}$. I take $\boldsymbol{\omega} \sim \mathrm{N}(0, I)$ and construct $\boldsymbol{f} = \mu_f + \sigma_f L_{\boldsymbol{f}} \boldsymbol{\omega}_f$. For the parameters determining the covariance structure, on which $L_{\boldsymbol{f}}$ is based, I generate initial values from the following distributions, $\nu_f \sim \mathrm{U}(0.5, 30)$, $\log \kappa_\phi \sim \mathrm{U}(-3.5, 1.5)$, $\gamma_\phi \sim \mathrm{U}(-\pi, \pi)$, while setting $\boldsymbol{\omega}_\phi = \mathbf{0}$. For the eigenvalue processes, $\lambda_p(\cdot)$, I take $\mu_{\lambda_p} \sim \mathrm{U}(-7, 2)$.

For the simulated data in one dimension, I report the 50 values (from the 50 data draws from the underlying distribution) of the MSE and KL for each of the methods for the training set. For the simulated two-dimensional Hwang dataset, I report the 50 values of MSE and KL for both the training and test sets. For the real datasets, I calculate the MSE and LPD for each observation when held out of the fitting and average over the data points to report one value for each of the criteria. For the training set, I average the MSE and LPD over both the data splits and the observations.

4.6 Results

4.6.1 One-dimensional assessment

On the slowly-varying smooth function (example 1), BARS and both the stationary and nonstationary GP models give very similar results for MSE (Figure 4.2). However, the KL divergences for the GP models are much lower than for BARS (Figure 4.2), because BARS systematically overestimates the error variance when $\boldsymbol{y}^T\boldsymbol{y}$ is large. This occurs because the unit information prior on the regression spline coefficients used by DiMatteo et al. (2002) is conditional on the error variance, η^2, and the resulting conditional posterior mean of η^2 is

$$\mathrm{E}(\eta^2|\boldsymbol{f}, \boldsymbol{y}, k, \boldsymbol{\xi}) = \frac{\boldsymbol{y}^T\boldsymbol{y} - \frac{n}{n+1}\boldsymbol{y}^T\boldsymbol{f}}{n} = \frac{\frac{n+2}{n+1}\boldsymbol{f}^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} + \frac{1}{n+1}\boldsymbol{f}^T\boldsymbol{f}}{n},$$

where $\boldsymbol{\epsilon} = \boldsymbol{f} - \boldsymbol{y}$, while the conditional posterior mean without the unit information prior would be

$$\mathrm{E}(\eta^2|\boldsymbol{f}, \boldsymbol{y}, k, \boldsymbol{\xi}) = \frac{\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{f}}{n} = \frac{\boldsymbol{f}^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{n}.$$

When the function mean is far from zero, as in example 1, the term $\frac{1}{n+1}\boldsymbol{f}^T\boldsymbol{f}$ can have a large impact on the estimate of η^2; this is a general effect in models in which the prior for the mean structure is dependent on the variance parameter. A better approach would be to subtract off $\bar{\boldsymbol{y}}$ from the data before running BARS, although I have not done that here. Interestingly, for samples of data in which it was difficult for the algorithms to determine the underlying function (i.e., the

MSE was high for both methods), the nonstationary GP method tended to perform better than BARS, while for samples in which the MSE was low, BARS tended performed better. In Figure 4.3, I show two data samples with BARS and NSGP fit to the data. Next I compare the results with those from Table 1 of DiMatteo et al. (2002), in which they compare BARS to the SARS and DMS methods (Figure 4.4). We see that based on a 95% confidence interval for the mean MSE over simulated datasets, BARS and the GP methods seem to outperform SARS and DMS, but that we cannot be certain of this conclusion relative to DMS because of high variability.

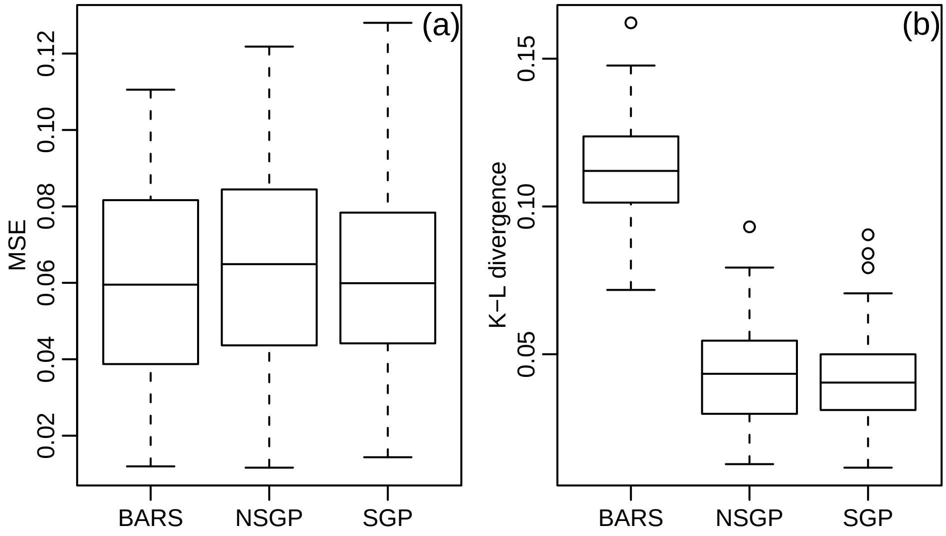

Figure 4.2. Boxplots of (a) MSE and (b) KL divergence for the three methods over 50 simulated datasets of example 1: Bayesian adaptive regression splines (BARS), nonstationary GP (NSGP), and stationary GP (SGP).

On the spatially inhomogeneous function (example 2), the nonstationary GP appears slightly worse than BARS in terms of MSE, while the stationary GP is noticeably worse than either BARS or the nonstationary model (Figure 4.5). In most (42 of 50) data samples, BARS has lower MSE than NSGP. The KL divergence for BARS may be slightly better than for the nonstationary GP (Figure 4.5), although this is difficult to interpret in light of the poor KL divergence for BARS in example 1. In Figure 4.6, I show BARS and NSGP fit to one of the 50 data samples. The nonstationary GP model smooths the data somewhat less than BARS, resulting in undersmoothing, apart

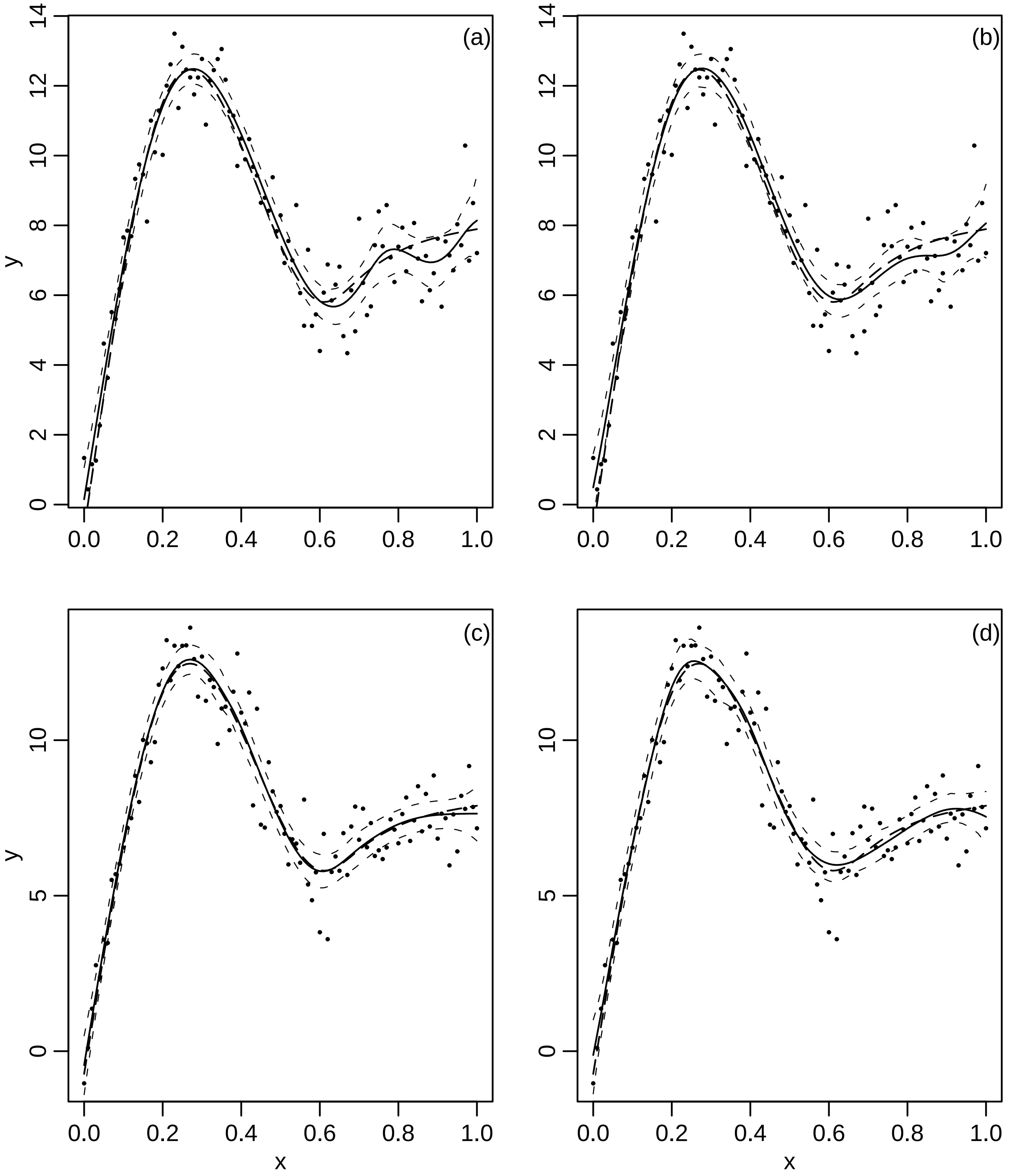

Figure 4.3. (a) BARS and (b) NSGP fit to one data sample in which NSGP has lower MSE than BARS. (c) BARS and (d) NSGP fit to a second data sample in which BARS has lower MSE. The thick dashed line is the true function, the solid line is the posterior mean estimate, and the thin dashed lines are 95% pointwise credible intervals.

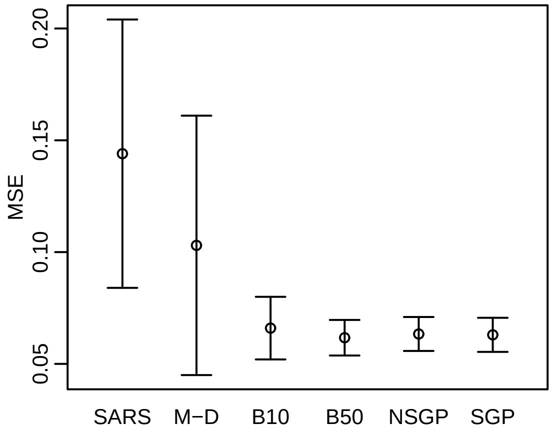

Figure 4.4. 95% confidence intervals for the mean MSE over simulated datasets of example 1. SARS, M-D (Modified-DMS) and B10 (BARS) are based on 10 sample datasets as calculated in DiMatteo et al. (2002), while B50 (BARS), NSGP (nonstationary GP) and SGP (stationary GP) are based on 50 sample datasets as calculated here.

from the jump at zero. Interestingly, for this data sample NSGP better captures the peak than does BARS. Comparing 95% confidence intervals for the mean MSE over simulated datasets (Figure 4.7), we see that the nonstationary GP may be comparable to what DiMatteo et al. (2002) term ”Modified-DMS”, namely BARS without the use of the locality heuristic for locating knots, and better than SARS, although the variability in the modified-DMS estimates makes this conclusion tentative. This result suggests the importance of the locality heuristic for the success of BARS. Note that the stationary GP is clearly worse than the various adaptive methods, illustrating the danger in using a non-adaptive method when the function is inhomogeneous. In the stationary GP model, κ_f and ν_f trade off, with at least one at a very low value during each iteration of the Markov chain. The model overfits in the regions where the true function is smooth, producing a very jagged posterior mean, and somewhat underfits the jump in the function.

Example 3 has a sharp jump at $x = 0.4$, where the function is not differentiable. On this example, BARS clearly outperforms NSGP in terms of both MSE and KL divergence, while the stationary GP again performs poorly relative to the nonstationary GP (Figure 4.8). It’s clear what is happening based on fits from BARS and NSGP (Figure 4.9). The NSGP model captures the

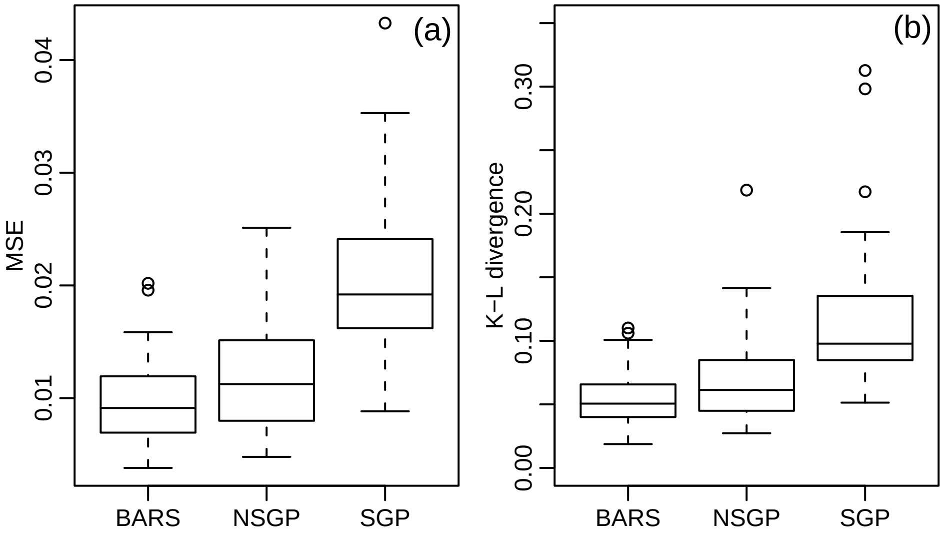

Figure 4.5. Boxplots of (a) MSE and (b) KL divergence for the three methods over 50 simulated datasets of example 2: Bayesian adaptive regression splines (BARS), nonstationary GP (NSGP), and stationary GP (SGP). For SGP, one outlier with KL divergence of 0.58 is not plotted.

sharp jump by using small kernels at the jump point, but because the kernels are forced to vary smoothly, the model undersmooths the data in the vicinity of the jump. Comparing 95% confidence intervals for the mean MSE over simulated datasets (Figure 4.10), we see that the nonstationary GP is roughly comparable to and probably somewhat better (the two outlying MSE values for NSGP appear to be caused by poor MCMC mixing) than modified-DMS, namely BARS without the use of the locality heuristic for location knots, and worse than SARS. As in example 2, this result suggests the importance of the locality heuristic for the success of BARS. The relatively poor performance of the nonstationary GP method is expected because the kernels are forced to vary smoothly, limiting the ability of the method to model sharp function changes. In Figure 4.11 I show how the GP model defines an implicit kernel smoother (locally-weighted averaging of the observations) based on the smoothing matrix,

$$S = C_{\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1},$$

as I discussed in Section 1.4.3. We see that the kernel at $x = 0.4$, the location of the sharp jump

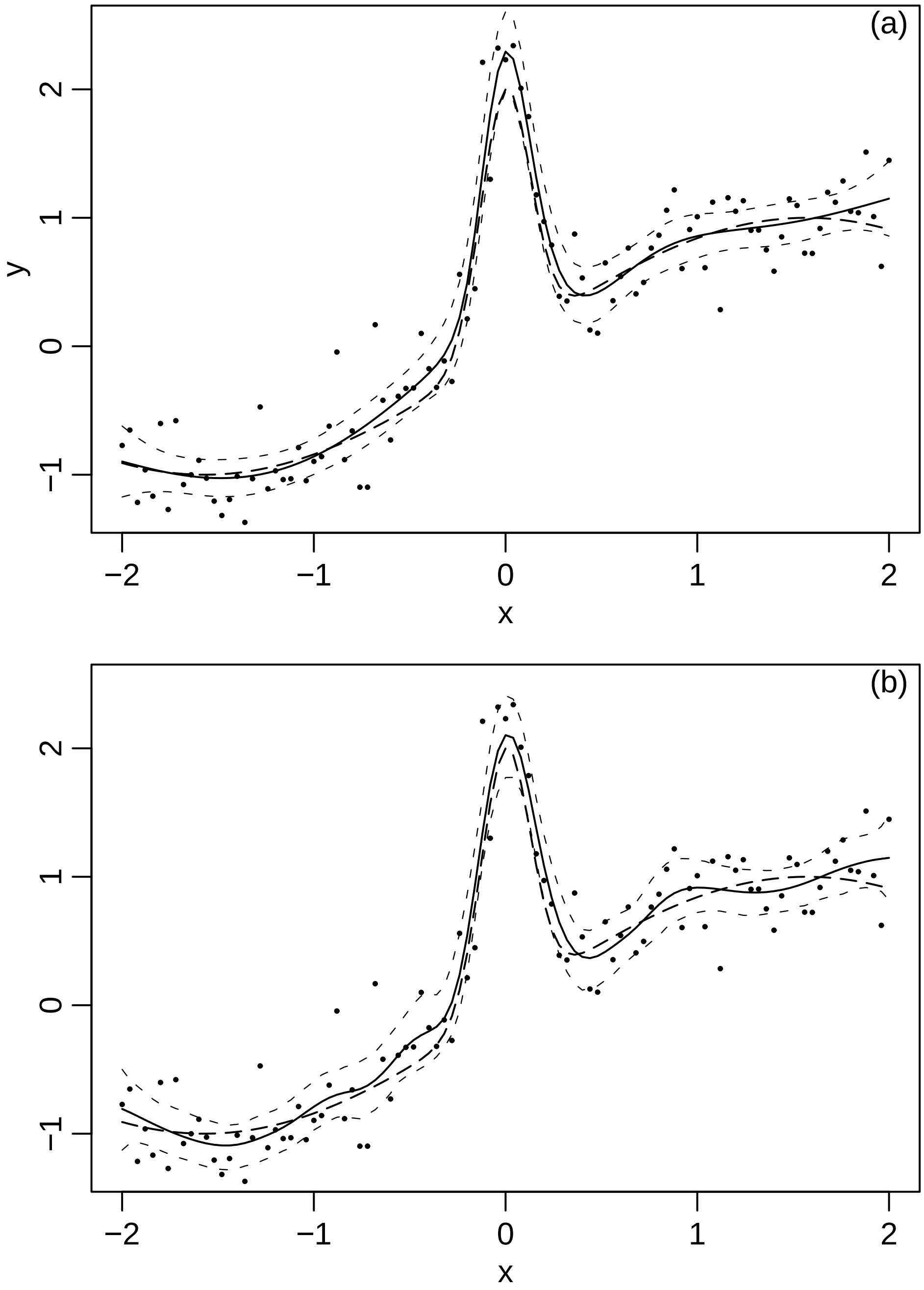

Figure 4.6. (a) BARS and (b) nonstationary GP fit to one data sample of example 2. The thick dashed line is the true function, the solid line is the posterior mean estimate, and the thin dashed lines are 95% pointwise credible intervals.

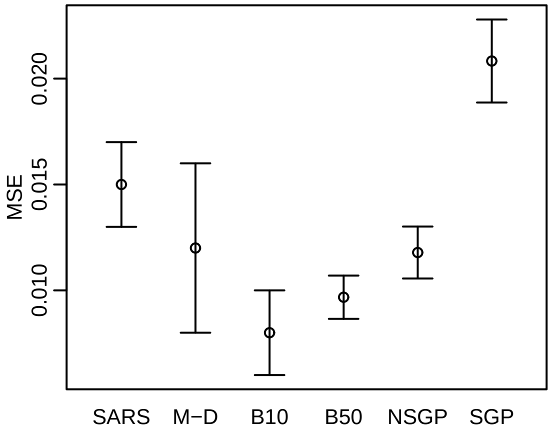

Figure 4.7. 95% confidence intervals for the mean MSE over simulated datasets of example 2. SARS, M-D (Modified-DMS) and B10 (BARS) are based on 10 sample datasets as calculated in DiMatteo et al. (2002), while B50 (BARS), NSGP (nonstationary GP) and SGP (stationary GP) are based on 50 sample datasets as calculated here.

in the function, is quite narrow, resulting in very local averaging, while the kernels elsewhere are much broader, resulting in more smoothing. As in example 2, the stationary GP model overfits in the regions where the true function is smooth, producing a very jagged posterior mean, and somewhat underfits the jump in the function. In this example, either κ_f or ν_f, or both, are at very low values during each iteration of the Markov chain.

The one-dimensional examples allow me to assess the GP model relative to BARS qualitatively. In general BARS sample paths are much smoother than GP sample paths. Once the MCMC in BARS settles on a small number of knots, it does not visit the parts of the function space with many knots. This may be partly because BARS is known to oversmooth to some extent, although to its credit, in these simulations, the method performs better than the GP method in areas where the function is less smooth, through its ability to place multiple knots in small intervals. The GP method shows less smoothness for several reasons. The main reason seems to be that the nonstationary GP model inherently samples less smooth functions. However, other reasons contribute as well. First, on the occasions when it samples from small values of ν_f, such as $\nu_f \leq 2$, which give non-differentiable sample paths, the sample paths are locally unsmooth, as expected. We can see

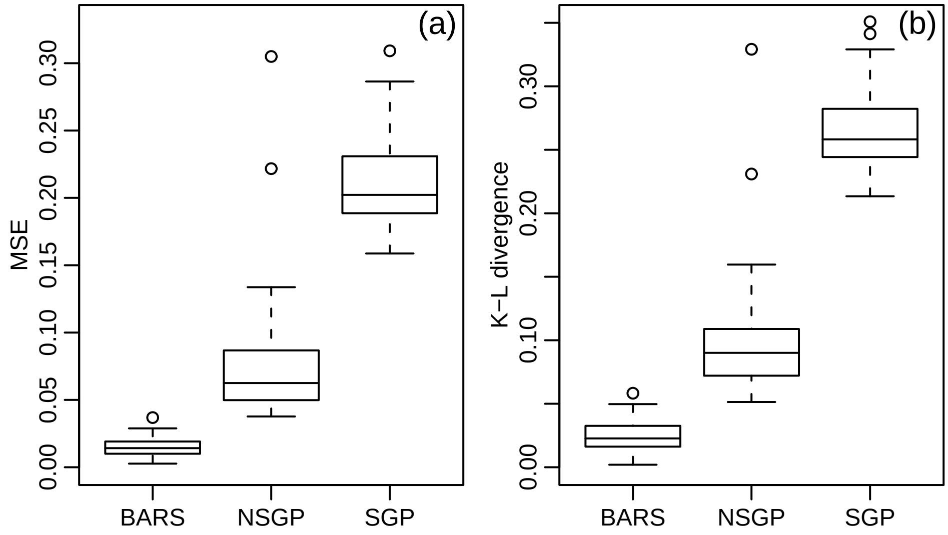

Figure 4.8. Boxplots of (a) MSE and (b) KL divergence for the three methods over 50 simulated datasets of example 3: Bayesian adaptive regression splines (BARS), nonstationary GP (NSGP), and stationary GP (SGP). One outlying value for SGP is omitted in both plots (MSE=0.87, KL=0.54). The two outliers for NSGP appear to be datasets for which the MCMC did not mix well because the proposal variances were the same for the MCMCs for all 50 data samples.

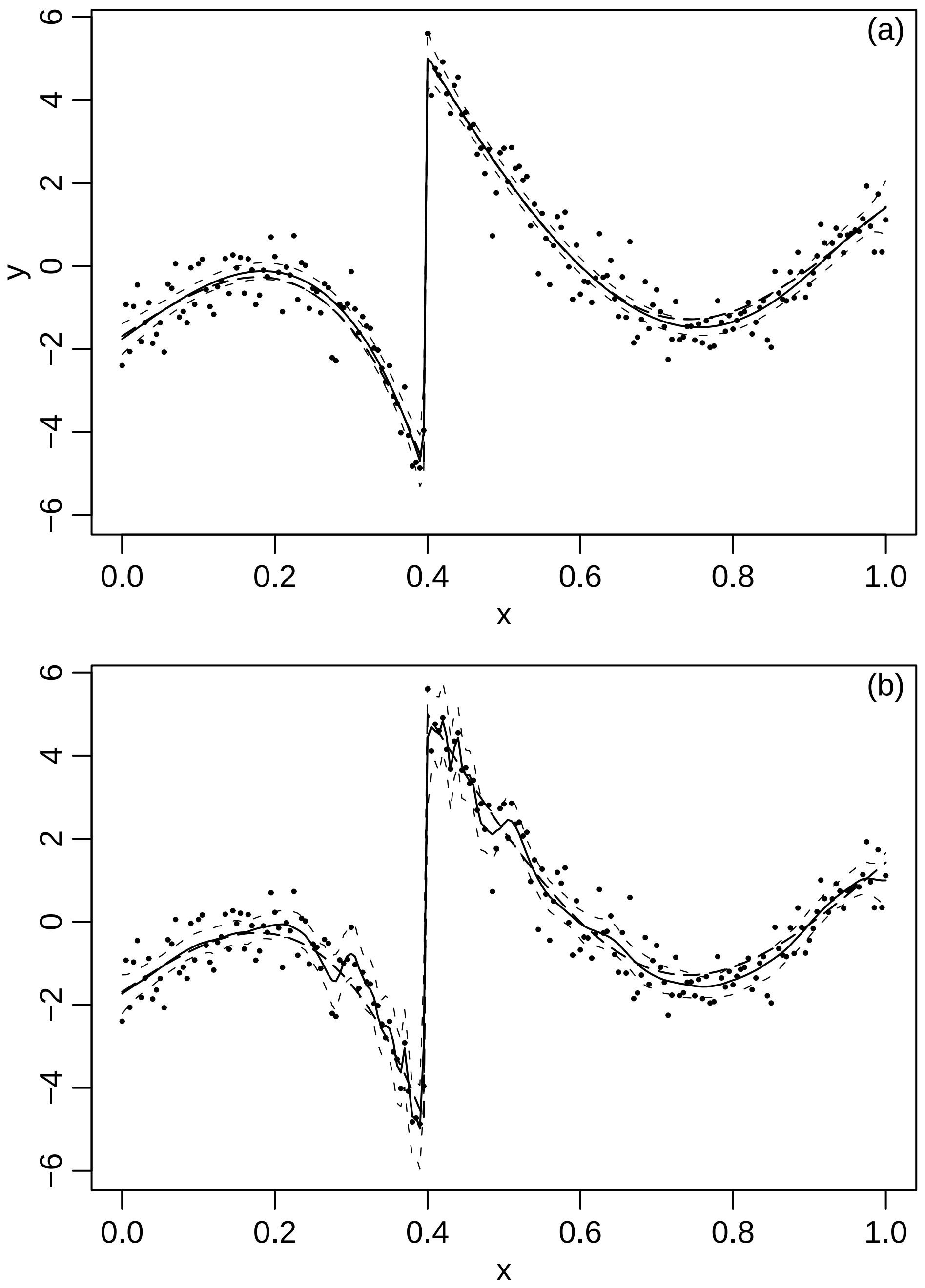

Figure 4.9. (a) BARS and (b) nonstationary GP fit to one data sample of example 3. The thick dashed line is the true function, the solid line is the posterior mean estimate, and the thin dashed lines are 95% pointwise credible intervals.

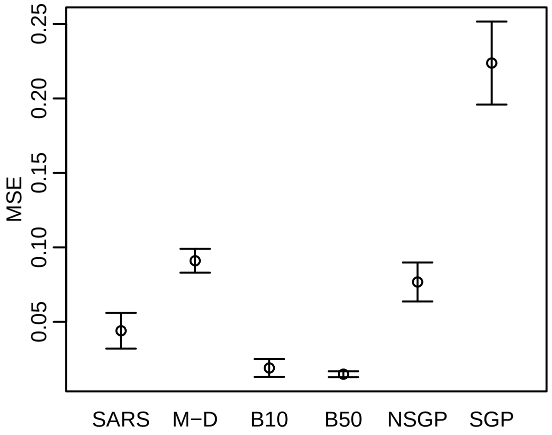

Figure 4.10. 95% confidence intervals for the mean MSE over simulated datasets of example 3. SARS, M-D (Modified-DMS) and B10 (BARS) are based on 10 sample datasets as calculated in DiMatteo et al. (2002), while B50 (BARS), NSGP (nonstationary GP) and SGP (stationary GP) are based on 50 sample datasets as calculated here.

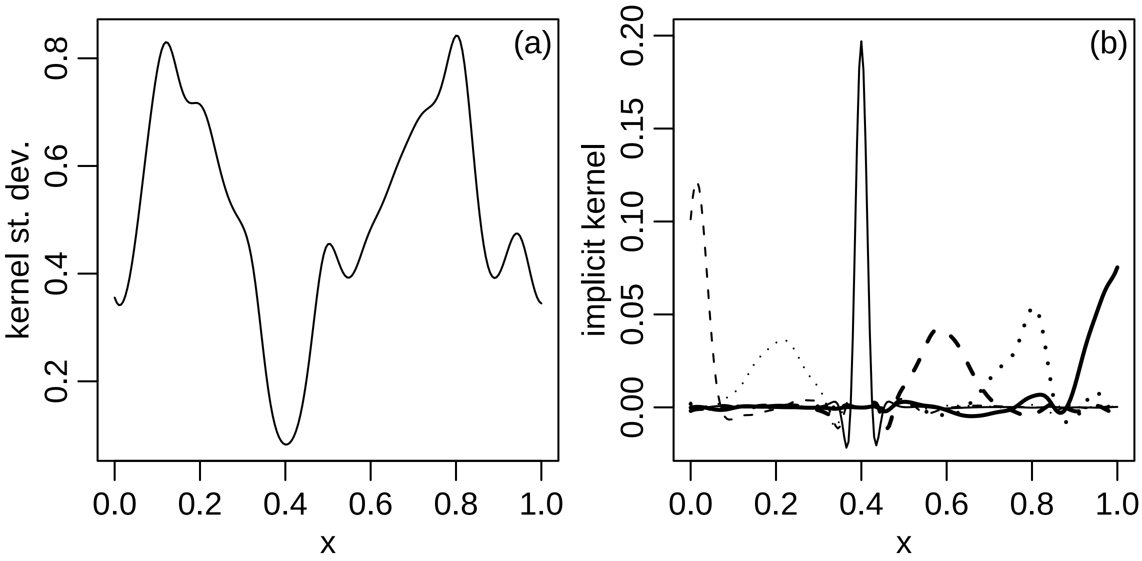

Figure 4.11. (a) Kernel size (standard deviation of kernels) as a function of the covariate for one posterior sample for example 3. (b) Implicit smoothing kernels at six covariate values, with different line types and/or widths for each implicit kernel.

an example of this in Figure 4.12 where the regression function is locally rough when $\nu_f = 1.3$ and locally smooth when $\nu_f = 4.5$, even though the other hyperparameter values are similar and the kernel sizes are larger in the case of $\nu_f = 1.27$. In other cases with small ν_f, the sample function can be even more jagged. A second reason relates to the observation of Gibbs (1997), namely that when the kernel sizes change rapidly, the correlation drops off relatively quickly in an intermediate neighborhood before levelling off and declining slowly at larger distances (Section 2.2). When the chain samples kernel sizes that are relatively large in magnitude, but which change rapidly, the sample functions tend to be rather unsmooth at intermediate scales (though still smooth at small scales). In Figure 4.13, we see an example of a sample function in which sharp changes in the kernel size induce lack of smoothness in the function, even though the kernel size is quite large. Also in Figure 4.13, we see a different example of a sample function in which the sharp change in the kernel size induces lack of smoothness in the function, although on this occasion, the model is using this feature of the correlation model to capture the jump in the function at $x = 0$ by having an unintuitively large kernel size right at the point of the jump. A third reason relates to a numerical problem that also occurred in some of my development runs. On occasion, the generalized Cholesky algorithm creates a Cholesky factor for which $LL^T \approx R$ did not hold. The result is localized jaggedness in the sample paths accounted for by not using a reasonable approximation to the Cholesky of the covariance matrix. It is not entirely clear why this happens, but it seems to occur because of very small errors, $O(10^{-14})$, in the calculation of elements of the covariance matrix.

For the stationary GP method on the inhomogeneous functions, the posterior for the Matérn smoothness parameter, ν_f, tends to concentrate on very small values in the range $(0.5, 2)$. These values are much lower than the bulk of the posterior for ν_f in the nonstationary model or in the stationary model for the homogeneous function. This suggests that the model is attempting to capture sharp changes in the function using the smoothness parameter, because it is unable to adequately model the data with the constant scale parameter, κ_f. In Figure 4.14 we see a demonstration of this for example 2 in which $\nu_f = 0.69$, verging on the exponential correlation function, but $\kappa_f = 4.0$, which is extremely large. (But note that κ_f is also quite small in many of the data samples and Markov chain iterations.) Instead of using a small value of κ_f to be able to capture the jump at

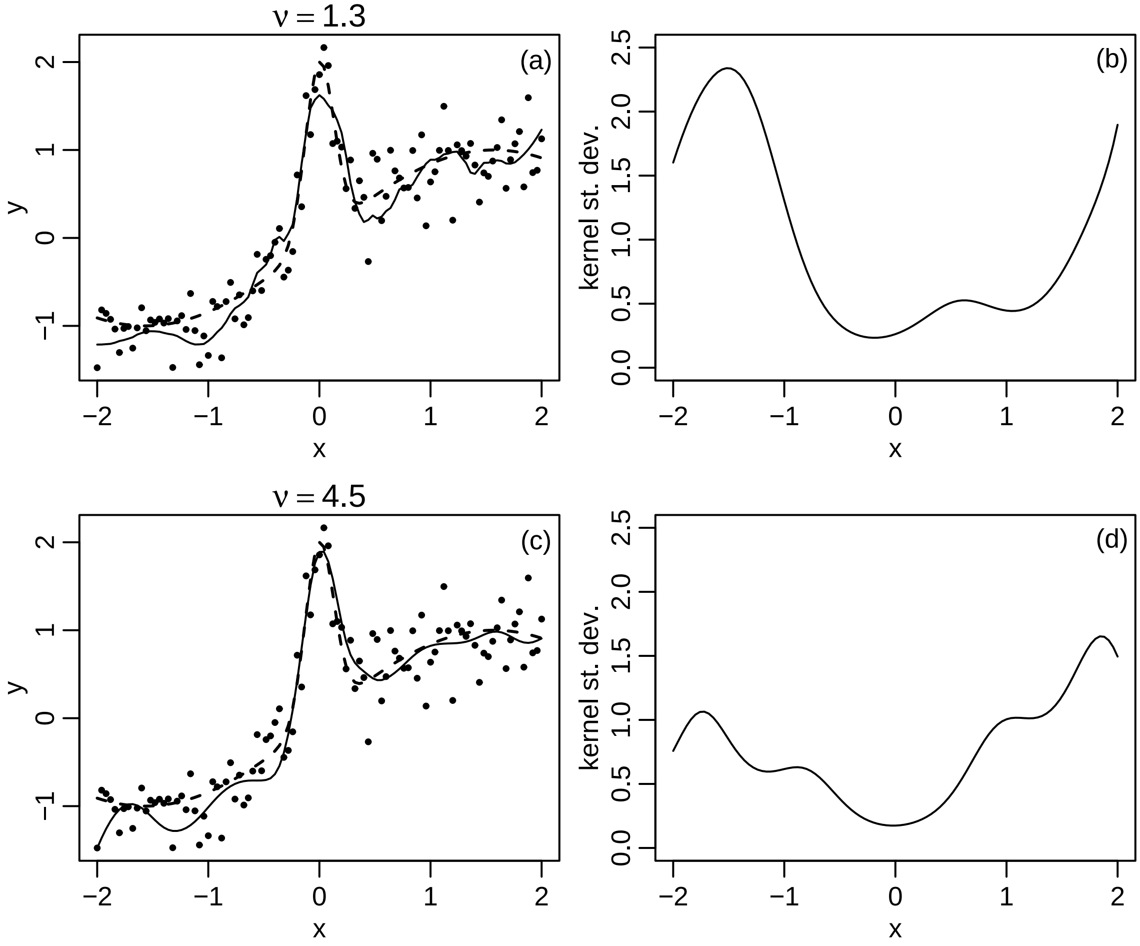

Figure 4.12. (a) Sample posterior regression function with $\nu_f = 1.3$ (solid line) and true function (dashed line) from example 2. (b) Kernel size (standard deviation of kernel) as a function of the covariate for the sample shown in (a). (c) Sample posterior regression function with $\nu_f = 4.5$ (solid line) and true function (dashed line). (d) Kernel size (standard deviation of kernel) as a function of the covariate for the sample shown in (c).

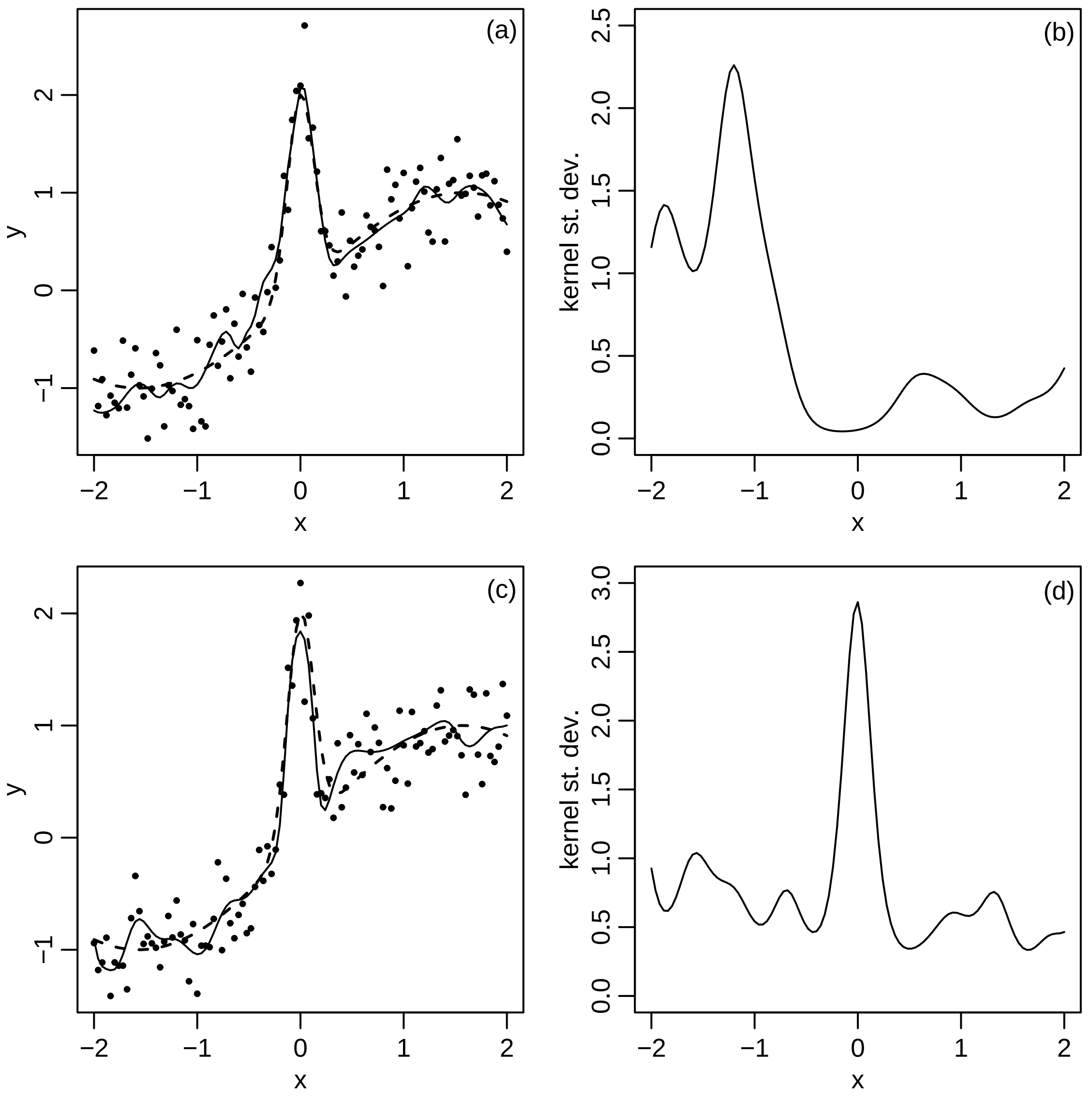

Figure 4.13. (a) Sample posterior regression function from one data sample (solid line) and true function (dashed line) for example 2. (b) Kernel size (standard deviation of kernel) as a function of the covariate for the sample shown in (a). Notice the lack of smoothness in the function for $-2 < x < -0.5$*, where the kernel sizes are large but variable. (c) Sample posterior regression function for a different data sample of example 2 (solid line) and true function (dashed line). (d) Kernel size (standard deviation of kernel) as a function of the covariate for the sample shown in (c). Notice that the nonintuitive sharp increase in the kernel size is what allows the model to capture the function jump at* $x = 0$*.*

$x = 0$, the model uses a very small value of ν_f. This suggests that one should be wary of interpreting the Matérn smoothness parameter as the degree of differentiability of the underlying function, since the parameter may just be compensating for misspecification of the correlation structure at coarser scales.

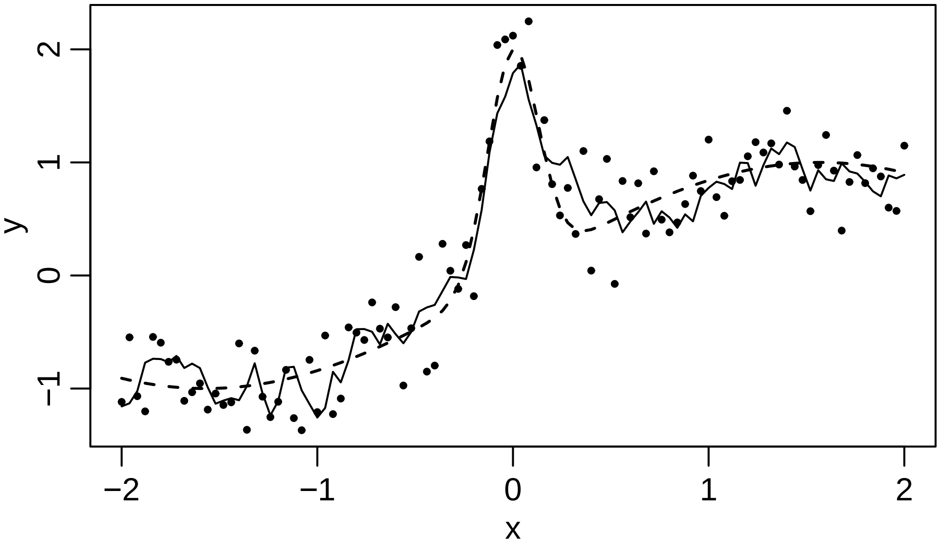

Figure 4.14. Sample posterior regression function from example 2 using a stationary GP model; here $\nu_f = 0.69$ and $\kappa_f = 4.0$.

Note that in earlier runs, I set $\nu_\lambda = 4$ rather than the value of 30 used here, and the performance of the nonstationary GP was very similar, suggesting that differentiability properties of the kernel elements may be of limited practical import, although I suspect that specifying ν approaching $\frac{1}{2}$ (the exponential correlation) might have an effect.

4.6.2 Higher-dimensional assessment

The simulated two-dimensional covariate dataset of Hwang et al. (1994) is based on a complex interaction surface. In Figure 4.15, I show perspective plots of the true function and the posterior mean estimate from the nonstationary GP model based on 225 training observations. The model does a good job of estimating the surface. Note that the residual standard deviation is only 0.25, so it is a fairly easy smoothing problem. Figure 4.16 shows contour plots of the true function and posterior mean estimate, as well as the difference between the two, which highlights that the model performs worst at the extremes of the covariate values, as one would expect because of extrapolation error. Both the stationary and nonstationary Gaussian process models perform well in comparison with the spline methods, based on both the training (Figure 4.17) and test (Figure 4.18) locations. For FVU, both the stationary and nonstationary GP models perform better than either spline model, while the stationary and nonstationary models are very similar, reflecting the relative homogeneity of the true function. All methods show higher FVU and KL divergence on the test covariates than the training covariates, probably because of poor extrapolation for outlying test covariates. The average test FVU estimate for BMLS is essentially the same (0.033) as reported in Holmes and Mallick (2001) based on a single simulated dataset. They also report test FVU estimates of 0.07 for a neural network model and 0.043 for a projection pursuit model, so it appears that the GP models do well relative to these methods as well, since the mean FVU for the NSGP is 0.023 and that for the SGP is 0.024. Denison et al. (1998b) report a training FVU of 0.0575 and a test FVU of 0.0411 for BMARS on this function, both somewhat different than the averages over the 50 datasets that I found (training FVU of 0.0288 and test FVU of 0.0761), presumably because I used different simulated data samples than they did.

The temperature dataset used by Wood et al. (2002) allows me to compare the methods on a spatial dataset, for which inhomogeneity in the underlying surface seems likely. NSGP outperforms the other methods in terms of both MSE (Table 4.1) and LPD (Table 4.2). Comparing the data and the model estimates (not shown), it appears that the success of the NSGP relative to the other methods occurs because the other methods oversmooth the data and don't capture real peaks and troughs in the surface. Note that in my initial fits of the models to the full 445 observations in the dataset, the MSE of the NSGP on test data was poor, but given its success on the reduced dataset and the very slow mixing of the MCMC for the full dataset, it appears that this occurs because the NSGP model has not burned in, rather than because the model is unable to fit the data. Using the prior on the degrees of freedom of the conditional posterior mean regression function appeared to help in the NSGP model. The MSE on test data without the df prior was 1.24, higher than the 1.10 with the prior and the LPD was slightly lower,-0.43 compared to -0.40.

For the real ozone dataset of Bruntz et al. (1974), the NSGP does a better job of prediction

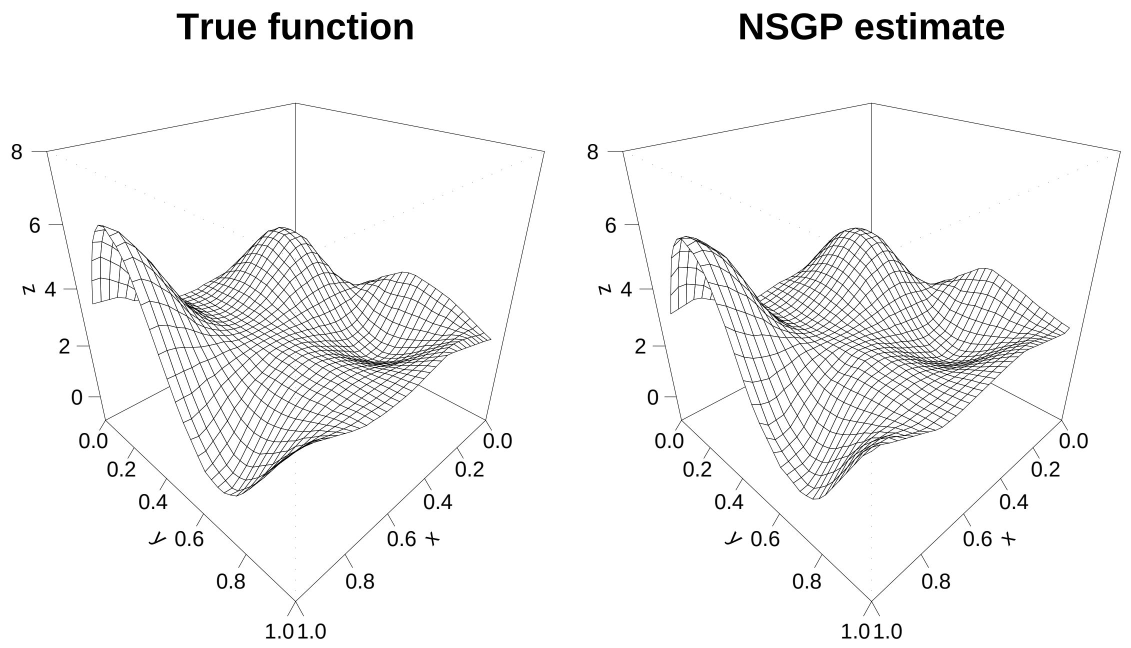

Figure 4.15. Perspective plots of (left) true Hwang function and (right) posterior mean function using the nonstationary GP model. Note that these plots involve interpolation by the interp and persp functions in the R statistical package; the interp function is found in the akima library.

Table 4.1. MSE for training and test sets for the four methods on a portion of the Wood dataset..

model	train	test
BMARS	0.55	1.74
BMLS	0.97	2.40
SGP	0.62	1.40
NSGP	0.25	1.10

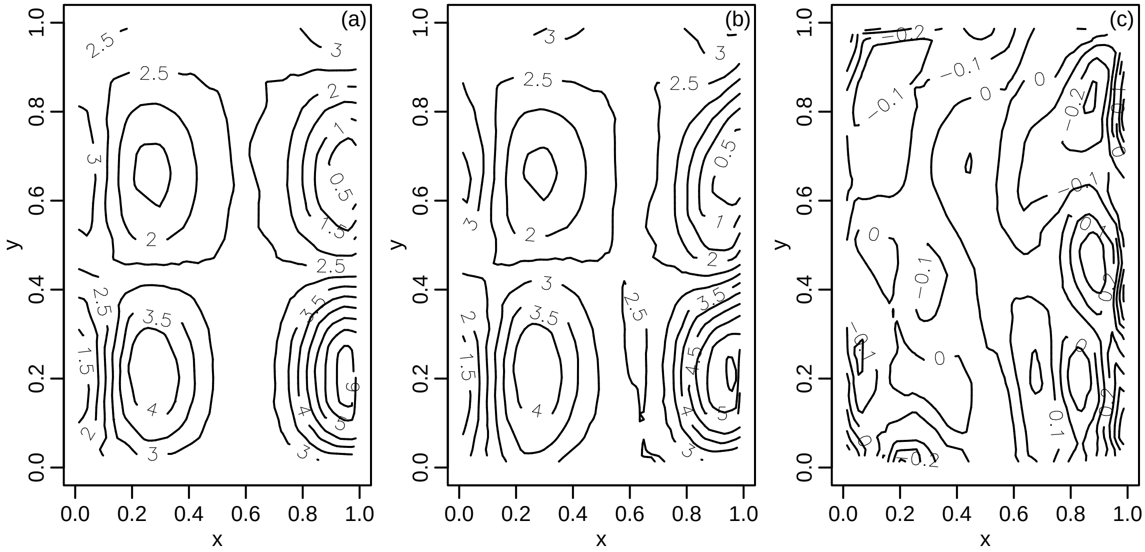

Figure 4.16. Contour plots of (a) true Hwang function, (b) posterior mean function using the nonstationary GP model, and (c) difference between the true function and the estimated function. Note that these plots involve interpolation by the interp and contour functions in the R statistical package; the interp function is found in the akima library.

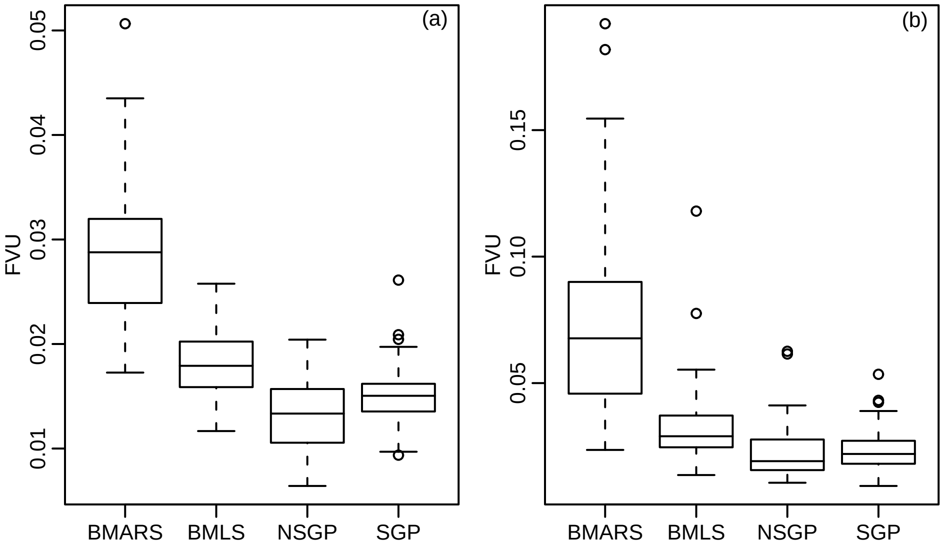

Figure 4.17. Boxplots of FVU for (a) training covariates and (b) test covariates over 50 simulated datasets of the Hwang function for the four methods

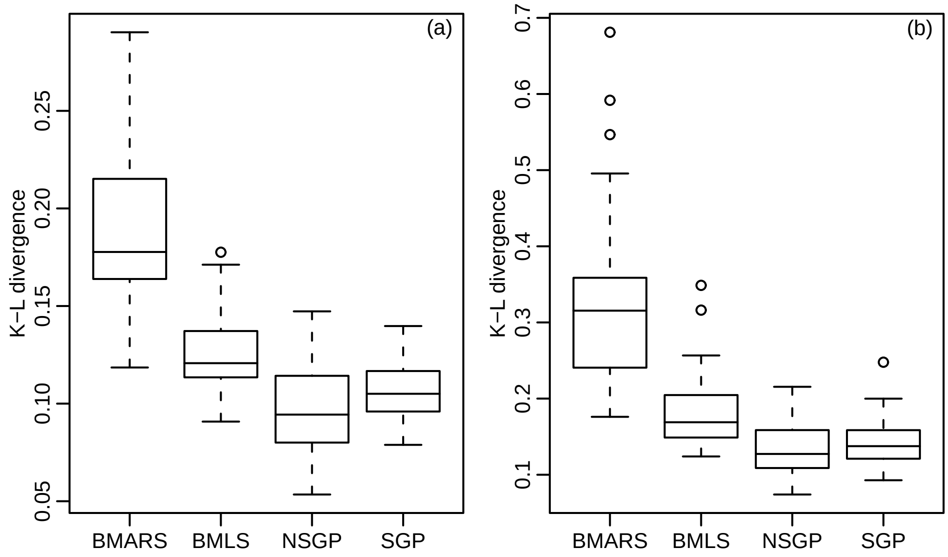

Figure 4.18. Boxplots of KL divergence for (a) training covariates and (b) test covariates over 50 simulated datasets of the Hwang function for the four methods.

Table 4.2. LPD for training and test sets for the four methods on a portion of the Wood dataset, averaged across observations.

model	train	test
BMARS	-0.22	-0.66
BMLS	-0.40	-0.86
SGP	-0.25	-0.65
NSGP	0.11	-0.40

on held-out data than the stationary model or the spline models (Table 4.3). Note that all of the GP and spline models do much better than linear regression or even a generalized additive model (GAM) with a four degree of freedom smooth for each covariate, suggesting the presence of non-additive structure in the data. The NSGP also does better in terms of predictive density than the other methods (Table 4.4), but the relative differences are small (a change of 0.08 in LPD means that an observation is 1.08 times as likely under the NSGP as under BMARS, for example). The similarity between the stationary and nonstationary GP results for MSE suggests that the underlying regression function is relatively homogeneous, but there does appear to be some heterogeneity. For this dataset, the use of the prior on the degrees of freedom seems to have little impact on predictive performance (MSE of 0.0055 and LPD of 2.12 without the prior).

Table 4.3. MSE for training and test sets for the four methods, as well as linear regression and a generalized additive model (GAM) on the ozone dataset.

model	train	test
Lin Regr	not computed	0.021
GAM	not computed	0.020
BMARS	0.0039	0.0062
BMLS	0.0048	0.0062
SGP	0.0037	0.0062
NSGP	0.0027	0.0054

In the code for BMARS and BMLS, available from the website of Dr. Chris Holmes, the spline basis functions in the model are standardized so that they have mean zero and standard deviation one over the basis function values evaluated at the training covariates. However, when a basis function has support only on a small portion of the covariate space, this tends to result in the basis function having very high values in extreme parts of the space. This is discouraged by the likelihood when there is a training covariate in this part of the space, but when there is only a test covariate, it tends to result in very poor prediction on the test covariates. For this reason, I altered the code to avoid the standardization, and achieved better predictive results. I do not know why this

Table 4.4. LPD for training and test sets for the four methods on the ozone data, averaged across observations.

model	train	test
BMARS	2.31	2.05
BMLS	2.17	2.04
SGP	2.13	2.05
NSGP	2.45	2.13

standardization step is included in the code. In particular, this problem arose for the BMLS model in the Wood et al. (2002) dataset, with the MSE on test data being 17.2 compared to 2.4 without the standardization.

4.6.3 Convergence

Based on diagnostic plots for the MCMC runs for the GP models, as well as the free-knot spline models in the higher-dimensional datasets, it appears that while the models have burned in, they have not fully mixed and that longer runs would be necessary to ensure convergence, in particular for the hyperparameters of the kernel eigenvalue processes. However, the evaluation criteria appear to be stable, which is sufficient for my purpose here. My general approach of running many chains with some of the initial parameter values differing between runs is in the spirit of Gelman and Rubin (1992), and the comparable results for the various runs offer evidence that the MCMC is giving reasonable results. Since I know the underlying function in 4 of the 6 cases, it would be clear if the MCMC were performing terribly, and my comparison with other methods gives further evidence that the chain is producing reasonable results, even though I do not directly assess the convergence for each simulation.

4.6.4 Example using non-Gaussian data

As a final evaluation of the method, I consider a dataset with non-Gaussian response, the Tokyo rainfall dataset used by Biller (2000) and others. I do not carry out a full evaluation of the GP method in comparison with other methods nor do I assess whether a nonstationary model is useful for this dataset, but rather I use the dataset to illustrate the ability of the nonstationary GP method to model non-Gaussian data. In particular, I compare two MCMC algorithms with this dataset, one the centered parameterization with joint sampling and the other the PMC sampling scheme. I also compare sampling with and without the Langevin-style update for $\boldsymbol{f}$, which uses the gradient of the log posterior. The model and sampling are described in more detail in Section $3.6.2.3$, with details of the Langevin sampling in Section $3.6.2.1$.

Biller (2000) uses the Tokyo rainfall dataset, originally from Kitagawa (1987), to illustrate his Bayesian free-knot spline method for non-Gaussian data. The data are the presence or absence of rainfall greater than 1 mm for every calendar day in the years 1983 and 1984. Assuming independence between years, the likelihood for a given calendar day, x_i, is binomial with two trials and unknown probability of rainfall, $p(x_i)$. Following Biller (2000) I take the response at different calendar days to be independent, conditional on $f(\cdot) = \text{logit}(p(\cdot))$. In a careful analysis, one might want to constrain the function and the eigenvalue function to be continuous between calendar day 365 and calendar day 1 by using distance on the circle rather than Euclidean distance, but I have not done so here.

I sampled 20000 iterations during the MCMC, following a burn-in period of 5000 iterations. Based on time series plots (not shown) and the effective sample size (ESS) approach (described in Section 3.6.2.3), the PMC and PMC with Langevin sampling schemes mix better than using the centered-joint scheme, either with or without Langevin sampling (Table 4.5). For reasons that are unclear, the PMC-Langevin scheme mixed much better when I attempted to have an acceptance rate of 0.23 for $\boldsymbol{f}$ than when I attempted to achieve a rate of 0.57, the theoretically optimal value given in Roberts and Rosenthal (2001).

Since the PMC-Langevin scheme with acceptance rate of approximately 23% mixed best, I focus on the results from that scheme. In Figure 4.19, I show time series plots for the hyperparameters and in Figure 4.20, time series plots for the function values and kernel eigenvalues at four locations. These plots indicate that the MCMC needs more iterations to mix fully but that it does seem to be exploring the parameter space, suggesting that the model is being fit successfully. In Figure 4.21 I show the posterior mean of the modelled probability of rainfall, $p(\cdot)$, with pointwise

Table 4.5. Effective sample size (ESS) by sampling scheme for key model parameters for the Tokyo rainfall dataset. $\bar{f}$ is the mean ESS for the function values, averaged over 10 randomly sampled calendar days, and $\bar{\lambda}$ is the mean ESS for the log of the kernel eigenvalues, averaged over 10 randomly sampled calendar days.

sampling method	μ_f	σ_f	ν_f	μ_λ	κ_λ	$\bar{f}$	$\bar{\lambda}$
centered-joint	37	47	21	39	26	47	17
centered-joint, Langevin, acc.≈23%	32	68	23	30	37	46	12
PMC	1114	81	241	48	59	85	62
PMC, Langevin, acc.≈57%	902	34	646	27	47	55	29
PMC, Langevin, acc.≈23%	1269	91	639	72	87	101	71

95% credible intervals. The response function seems to reasonably follow the data. Note that the data in some areas seem quite clustered, which explains why the response function is not estimated to be smoother. This response function closely matches the response function in Biller (2000, Fig. 1) using natural splines with a Poisson(60) prior on the number of knots, but is much less smooth than the spline models with Poisson(30) priors on the number of knots. Biller (2000) prefers the Poisson(30) curves, saying that the Poisson(60) curve is too rough and shows too many details. If one's goal is to see long-term patterns, then a smoother function is desirable, but my assessment is that the less smooth curves (namely the Poisson(60) curve and the GP curve here) are more true to the apparent clustering in the underlying data.

In Figure 4.22 I show the geometric mean kernel size as a function of calendar day, indicating that the model is detecting inhomogeneity in the underlying function, with more smoothness in the first few months and less smoothness later in the year. I use the geometric mean since the kernel eigenvalues are fit on the log scale.

The GP method is clearly much slower than the spline approach. Biller (2000) reports a running time of about 5 minutes for 15000 iterations on a Pentium II 333 MHz machine, while the runs here took about 30 hours for 15000 iterations on a Pentium 4 2.2 GHz machine.

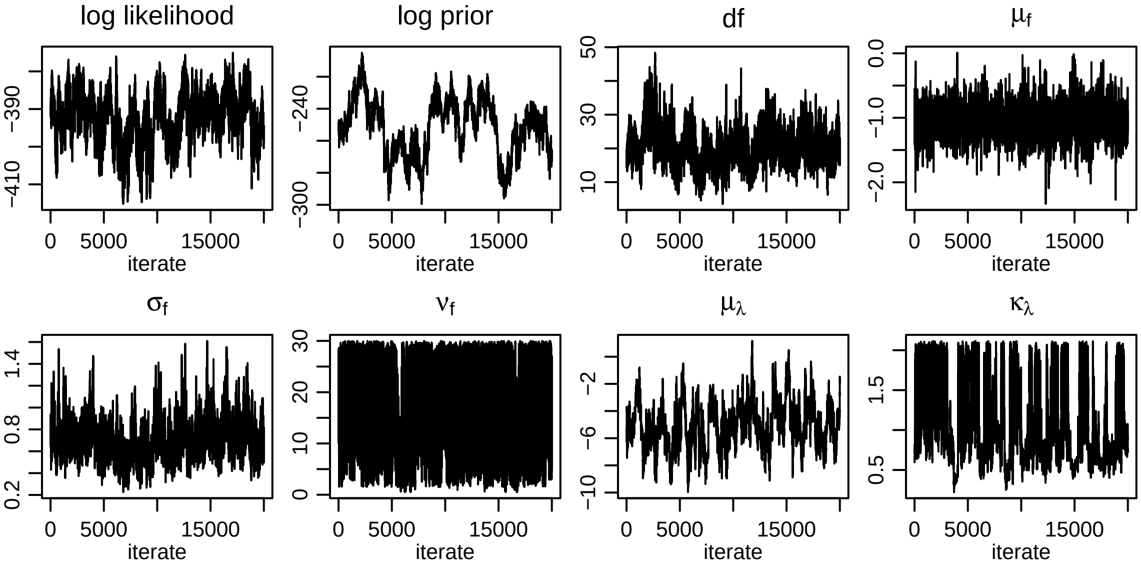

Figure 4.19. Time series plots for the Tokyo rainfall dataset for model log likelihood, log prior density, degrees of freedom of the conditional posterior mean function, and hyperparameters.

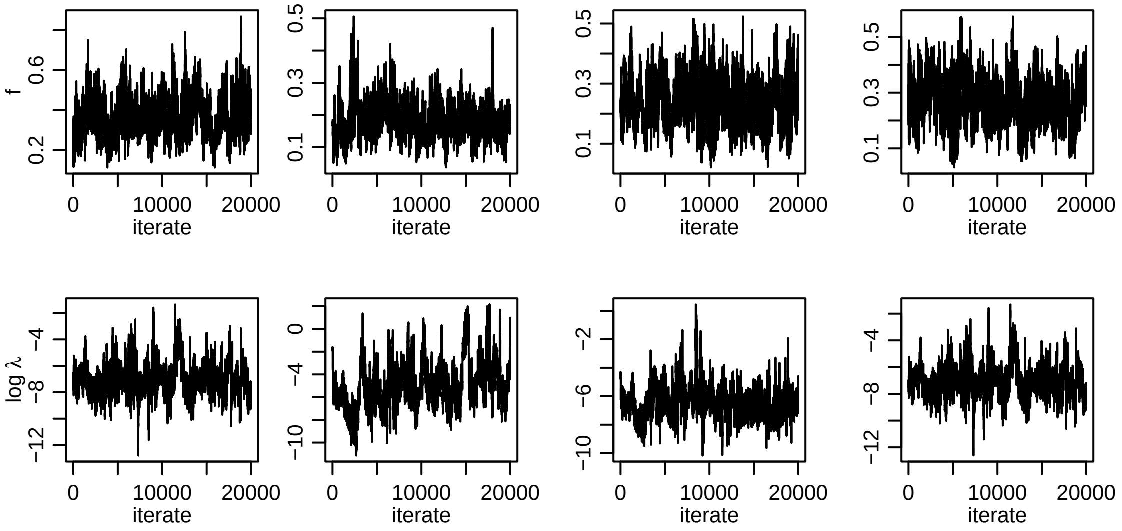

Figure 4.20. Time series plots for the Tokyo rainfall dataset for function values, $f(\cdot)$ (first row), and log of kernel eigenvalues, $\log\lambda(\cdot)$ (second row), at four covariate values.

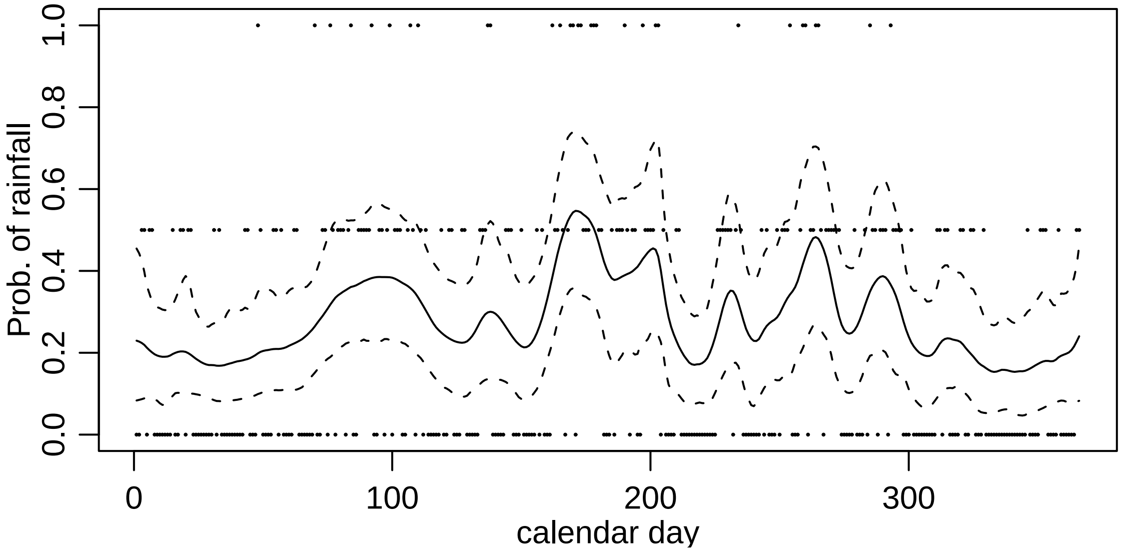

Figure 4.21. Posterior mean estimate of $p(\cdot)$, the probability of rainfall as a function of calendar day, with 95% pointwise credible intervals. Dots are empirical probabilities of rainfall based on the two binomial trials.

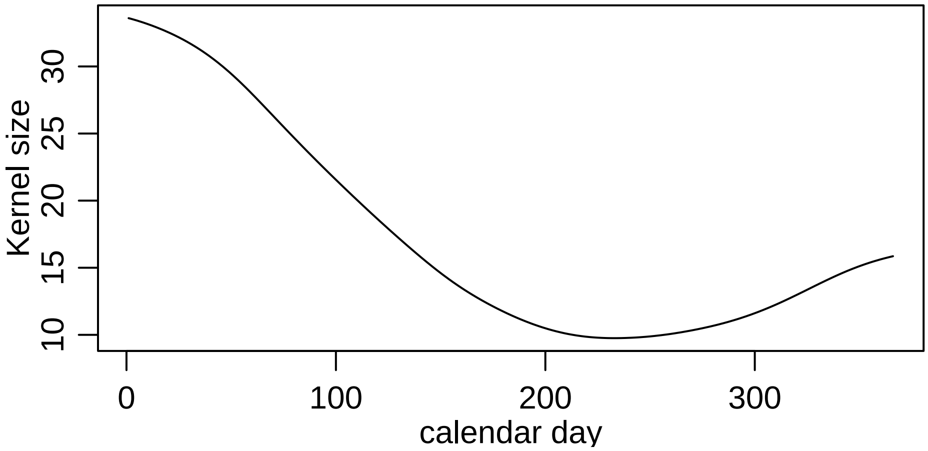

Figure 4.22. Posterior geometric mean kernel size as a function of calendar day. The kernel sizes are plotted as the square roots of the geometric means of the kernel eigenvalues, and hence can be thought of as correlation scale parameters, with units of days.

4.7 Discussion

The results for the nonstationary GP model in fitting nonparametric regression models are mixed. In one dimension, the nonstationary GP does indeed improve the fit compared to a stationary GP model when the true function is inhomogeneous, as for examples 2 and 3 of DiMatteo et al. (2002). However, for very sharp jumps, such as in example 3, the smoothness constraints on the kernel structure cause the nonstationary model to undersmooth in the vicinity of the jump. While the nonstationary GP method generally outperforms some earlier spline methods, in all three one-dimensional examples examined here, the free-knot spline method, BARS, which employs a locality heuristic for locating knots, outperforms the nonstationary GP method. For one-dimensional problems, I would suggest the use of BARS rather than a GP model, although for functions that are not too inhomogeneous, the nonstationary GP method performs quite well. An added advantage of BARS is that it is much faster than the GP method and allows one to easily compute the function estimate in closed form based on the linear model representation conditional on the location of the knots.

The story changes somewhat in higher dimensions. Here I have compared the GP methods to free-knot spline methods that generalize the usual one-dimensional spline representation. If one is willing to use an additive model, then an additive model version of BARS is probably one's best choice. In particular, based on some incomplete experimentation, I believe the nonstationary GP method has trouble ignoring unimportant covariates, and so should probably be employed only in situations in which one has reason to believe that all or most of the covariates are important. Employing a non-additive model such as the multivariate spline models or the GP models is likely to be most useful if important interactions are present. The evidence presented in this chapter suggests that GP models are effective competitors to the spline-based methods, provided the sample size and number of covariates are not too large, allowing the GP model to be fit in reasonable time. Additional comparisons with standard nonparametric regression models such as kernel smoothers, wavelets, radial basis function networks, and neural networks, if successful, would strengthen the case for the nonstationary GP model. It would also be useful to compare model performance excluding possible boundary effects by assessing in the interior portion of the covariate space, although as the covariate dimension increases, the importance of good estimation near the boundaries increases as well.

In more standard regression problems, to contrast with what one might think of as surface-fitting problems, nonstationarity in higher dimensions may not be an important concern; in particular the curse of dimensionality limits our ability to detect such features of the data even if they are present. This suggests that a stationary GP model may be a good choice. For two of the three examples here, which seem to be relatively homogeneous, the stationary GP model outperforms the spline-based models, and in the ozone example, the performance is comparable to the spline-based model. The stationary GP model is also simpler conceptually and in its parameterization than the nonstationary GP model or the spline-based models. The MCMC methods outlined in Chapter 3, possibly in conjunction with the work of Christensen et al. (2003) will help in fitting the stationary model when the likelihood is not normal. However, full MCMC is still slow and subject to poor mixing. Thoughtful choices about hyperparameters that can be fixed in some way, such as via an empirical Bayes approach, may be useful. The relative success of the stationary GP model should not be too surprising, given that work in the machine learning literature has reported success with the stationary model on a variety of tasks. For example, Rasmussen (1996) reports that, along with a Bayesian neural network model, GP models were the most successful of the models he used for a variety of regression problems.

Chapter 5

Spatial Model Results

5.1 Introduction

In the previous chapter, I demonstrated that nonstationary covariance functions could be used in a Gaussian process (GP) prior for regression functions. In the regression setting, a single set of observations is observed, and the nonstationary covariance is inferred based on the similarity between responses at nearby locations. This neighborhood-based inference is the same type of inference done in time series analysis in the standard case in which there is only one time profile. Recall that in Chapter 4 the nonstationary covariance was the prior covariance for the regression function.

In spatial analysis, multiple replicates of a field of data (usually taken over time) are often available, and we may be interested in using the replication to model the covariance structure of the data. Ideally, we would have enough data to estimate the covariance by maximum likelihood as a simple averaged cross-product of residuals. However, for the maximum likelihood (ML) estimate of the covariance to be non-singular, we need at least as many replicates as we have observations (locations) within each replicate. Even when this many replicates are available, a smoothed estimator of the covariance is likely to perform better than the unsmoothed ML estimator.

In such cases, a covariance model is needed, and in many settings a stationary covariance may not be sufficient. This appears to be the case for the storm activity data analyzed in Paciorek, Risbey, Ventura, and Rosen (2002). As we see in Figure 5.1, the residual spatial correlation structure at two different locations appears to have very different correlation scale, which indicates nonstationarity in the correlation structure. In this chapter, I use the nonstationary covariance model described in Chapter 2 and developed further in Chapter 3 as a model for the residual spatial correlation in a hierarchical space-time model.

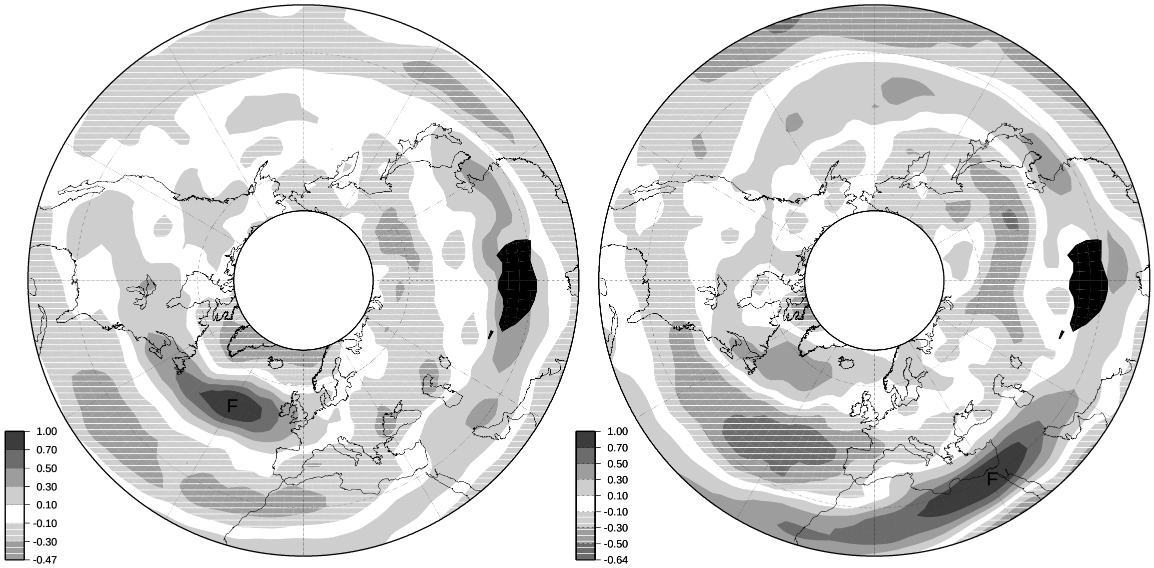

Figure 5.1. Plots of residual correlation of storm activity between all locations and each of two focal locations, each marked with an 'F': (left) $50°$ *N,* $330°$ *E (in the North Atlantic) and (right)* $30°$ *N,* $30°$ *E (in Egypt). The residual correlation is calculated after removing location-specific linear trends. Gray shading indicates the value of the correlation coefficient, with darker colors indicating values large in magnitude. Negative values are indicated by horizontal hatching. The high topography of the Himalayas is blacked out.*

5.2 Scientific Problem

In recent decades, atmospheric science and climatology have become much more prominent, in part because of advances in weather prediction and demand for meteorological information as an important economic product, and in part because of increasing scientific and public concern about climate change induced by anthropogenic changes in atmospheric greenhouse gas concentrations. One aspect of climate change receiving recent attention has been the possibility of changes in extreme events such as hurricanes, tornados, and winter storms. In Paciorek et al. (2002), we analyzed several indices of Northern Hemisphere extratropical winter cyclonic storm activity. Of particular interest was the possibility of long-term changes in storm activity over the 51 years for which we had data (1949-1999). Based on time series analysis, we concluded there was little evidence for temporal autocorrelation, probably because we calculated winter averages, and correlation does not seem to occur from year to year. Since standard linear regression seemed appropriate, we calculated the maximum likelihood linear time trend estimates based on the 51 years and mapped these estimates. To assess the reliability of the estimates in the face of simultaneously testing 3024 locations, we used the False Discovery Rate (FDR) approach introduced by Benjamini and Hochberg (1995). We found that for most of the indices of storm activity, there was significant evidence of trend at some locations but no significant evidence at many locations. Some questions arose as to which version of the FDR methodology to use in the face of correlation between test statistics induced by the spatial correlation of the data. In a separate project, we have been assessing this question in a simulation study of several versions of the methodology (Ventura, Paciorek, and Risbey 2003). The goal of this chapter is to further assess the trends in storm activity by building and fitting a Bayesian model of the time trends that properly accounts for the uncertainty induced by variability in both time and space.

The effect of variability in time on the trend estimates is simple to understand. We fit linear regressions for storm activity as a function of year at each location. Temporal variability causes uncertainty in our estimation of the underlying trends. Analyzing a single location, this uncertainty can be assessed easily in the usual fashion, either classically or from a Bayesian perspective. The difficulty comes in incorporating spatial correlation into a joint estimate of the time trends. The goal is to borrow strength across locations to better estimate the time trend spatial field and its uncertainty in a cohesive fashion. One reason that using spatial information is particularly important is that the data themselves are quite smooth spatially, as are the ML trends (as well as the intercepts and residual variances) at each location. There are two main reasons for this. One is that storm activity is spatially correlated; storms are spatially coherent phenomena that move through the atmosphere, generally along certain preferred paths, called storm tracks. The result is that locations that are close in space tend to have similar amounts of activity because they are affected by the same storms. The second reason for correlation in the data is that the data we use are the out-put of a meteorological data assimilation model (the NCEP-NCAR reanalysis model (Kalnay and Coauthors 1996)) that takes raw weather observations (temperature, wind, pressure, etc.) combines them with a deterministic meteorological model, and estimates the values of weather variables on a regular latitude-longitude grid ($2.5^\circ \times 2.5^\circ$). This assimilation process induces spatial correlation in the variables and in the storm activity indices calculated from these variables.

A worst case scenario for the trend analysis is that there are no true trends in time, but that temporal variability in the data in time cause apparent trends, and the spatial correlation makes these false trends seem more reliable by causing them to be spatially coherent. The key issue lies in determining the extent to which the spatial correlation causes spatially smooth real trends as opposed to spatially smooth noise. The goal of the spatial model that I construct in the next section is to embed the individual linear time trend models in a model for the spatial correlation structure of the data, so that the estimation of the trends is done in a manner that accounts for the spatial structure of the data. There are three main aims; the first is to get a better point estimate of the trend activity by borrowing strength across locations. The second is to come up with uncertainty estimates for the trends that account for the spatial correlation structure. Hopefully the results of the model will allow us to obtain a more complete picture of the trends and their reliability to compare to the results of the classical testing analysis. Separating trend from spatially smoothed noise is essentially the same problem faced in time series analysis in separating signal from correlated noise. Luckily, in the case of these spatial data, we have repeated observations, the multiple years, with which to make headway. The third aim is to evaluate several covariance models, stationary and nonstationary, to see which is the best model for the data and therefore which models might be preferred for similar climatological data.

Performing spatial smoothing and estimating trends at unobserved locations are not primary goals of this analysis for several reasons. First, as mentioned above, the data are already quite smooth, so further smoothing seems of limited importance, and since the latitude-longitude grid is already fairly dense, smooth maps can be produced with little effort directly from the unsmoothed data with little need for spatial interpolation of either data or estimated trends. Second, by working only with observed locations, I have a simple baseline ‘model’ for the data, namely the predictions from the ML trends and intercepts at the locations. I can compare the results from the various covariance models to this simple baseline model, which is not possible if I estimate trends at held-out locations. Finally, most of the variability in the data is accounted for by residual variation about the trends, not by the trends themselves. Therefore, the main driver behind the accuracy in predicting unobserved locations will be the the model's performance in spatially smoothing the residuals from the temporal model, not the model's performance in smoothing the trends. Hence, assessment of model performance with respect to the main scientific problem of trend estimation is better addressed by cross-validating in time (holding out time points) rather than by cross-validating in space (holding out locations). For other datasets, spatial prediction might be an important goal, and cross-validating in space would be an important means of evaluation. Prediction on unobserved years does not fully address the question of model evaluation, but I believe it does the best possible job under the constraints imposed by these data.

In addition to the empirical evidence for nonstationarity discussed briefly above, basic climatology also suggests that nonstationarity would be present because of the morphology of storm activity. Winter storm activity tends to follow certain preferred paths, called storm tracks. The correlation structure in these storm track areas is likely to be different from the correlation structure in areas far from the storm tracks with their differing climatology. Other important aspects of storm activity are that the correlation scale in the west-east direction is likely to be longer than the correlation scale in the north-south direction because the storm tracks tend to be oriented west-east. Also, storm tracks can shift in time, so there may be strong negative correlations between locations. For example negative correlations may occur between locations at the same longitude but separated by latitude, since if a storm track shifts in the north-south direction, storms that would have hit one location instead hit the location at the other latitude.

While many analyses have modelled spatiotemporal structure, few have focused on including spatial context in assessing long-term time trends at multiple locations. Holland et al. (2000) extensively analyzed atmospheric concentrations of sulfur dioxide, estimating 35 location-specific trends after accounting for covariates. Using kriging methods, they smoothed the raw trend estimates. They used the jackknife to estimate both diagonal and unstructured covariances of the unsmoothed trend estimates. They found that accounting for correlation between the location-specific estimates using the unstructured covariance better fit the data than the diagonal covariance and resulted in small changes in regional trend estimates and increases in the standard errors of these regional trends. Oehlert (1993) discusses but does not fit (because of insufficient data) a model in which linear time trends of wet sulfate deposition are embedded in a spatiotemporal model and the estimated residual spatial correlation is used in making inference about the trends. Wikle, Berliner, and Cressie (1998) propose a spatiotemporal model in which location-specific temporal models are parameterized by spatial processes with Markov random field priors. This chapter represents a Bayesian effort to fully account for the residual spatial structure in assessing time trends.

Methods that accomplish these goals are needed. Spatio-temporal data for which the scientific questions of interest involve analysis of changes over time are common. They are of obvious importance in assessing scientific and public policy questions about the evolving state of natural systems of various kinds, from climate stability to changes in biological populations and environmental phenomena of various sorts.

5.3 Basic Spatial Model

The basic spatial model builds on the location-specific models used in Paciorek et al. (2002). In that work, we fit simple linear regressions of storm activity against time,

$$Y_t \sim \mathrm{N}(\alpha + \beta t, \eta^2).$$

In this analysis, I use the same local model, but tie the residuals, now indexed by location i, $Y_{it} - \alpha(\boldsymbol{x_i}) - \beta(\boldsymbol{x_i})t$, together in space using various covariance models. I model $\alpha(\cdot)$ and $\beta(\cdot)$ a priori as Gaussian processes. The joint likelihood for a single time, $t \in \{-25, \cdots, 25\}$, is taken to be

$$\boldsymbol{Y}_t \sim \mathrm{N}(\boldsymbol{\alpha} + \boldsymbol{\beta} t, C_{\boldsymbol{Y}}), \tag{5.1}$$

and $\boldsymbol{Y}_t|\boldsymbol{\alpha}, \boldsymbol{\beta}$ is taken to be independent of $\boldsymbol{Y}_s|\boldsymbol{\alpha}, \boldsymbol{\beta}$ for $t \neq s$, so that $C_{\boldsymbol{Y}}$ is not a function of time. This is justified based on an exploratory analysis of the correlation structure of the data, which indicated that for winter-long averages of the storm indices, there is little correlation from year to year. The covariance matrix of the observations is taken to be

$$C_{\boldsymbol{Y}} = D(\boldsymbol{\eta}) R_{\boldsymbol{Y}} D(\boldsymbol{\eta}) + \delta I,$$

where $D(\boldsymbol{\eta})$ is a diagonal matrix of standard deviations, $R_{\boldsymbol{Y}}$ is a correlation matrix, and δ is the variance of location-independent noise. In the analysis I compare several models for $R_{\boldsymbol{Y}}$, including a stationary model, kernel-based nonstationary model, and smoothed versions of the empirical correlation structure; description of these models is deferred to Section 5.4. A directed acyclic graph (DAG) of the model, with the nonstationary parameterization for $R_{\boldsymbol{Y}}$, is shown in Figure 5.2.

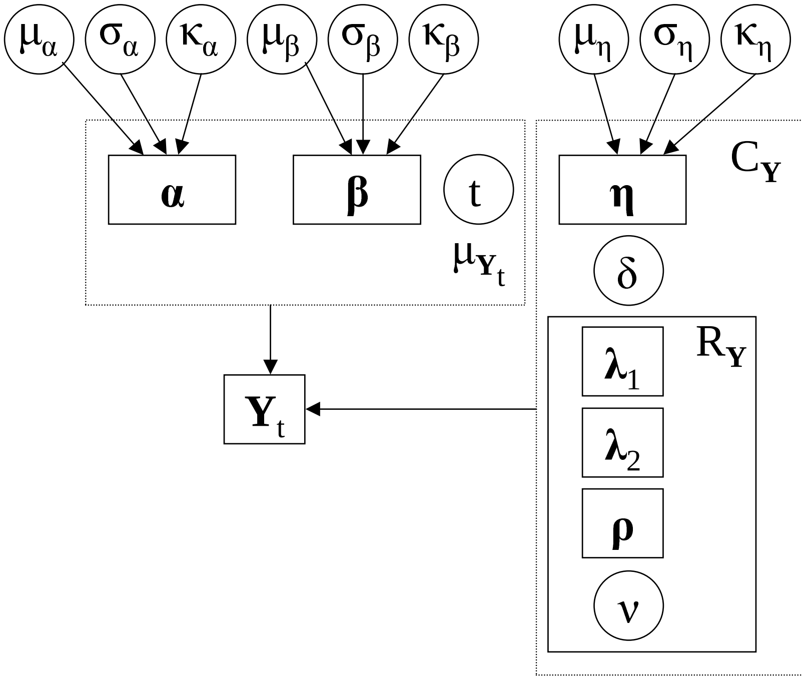

Figure 5.2. Directed acyclic graph of nonstationary spatial model. Bold letters indicate vectors..

The mean function parameters and the residual variance parameters are taken to be spatial fields. For $\phi(\cdot) \in \{\alpha(\cdot), \beta(\cdot), \log \eta(\cdot)^2\}$, we have

$$\phi(\cdot) \sim \text{GP}\left(\mu_\phi, \sigma_\phi^2 R_\phi^S(\cdot; \kappa_\phi, \nu_\phi)\right),$$

where $R_\phi^S(\cdot)$is the stationary Matérn correlation function with scale parameter κ_ϕ and fixed smoothness parameter, $\nu_\phi = 4.0$. This value of ν_ϕ reflects my belief that the parameter processes vary smoothly in space. The atmospheric phenomena that generate storms are based on masses of air and energy that move through the atmosphere. Furthermore, the data, Y_{it}, are winter-long averages; this averaging should increase the smoothness. For $\nu_\phi = 4$ we have $\lceil \nu_\phi - 1 \rceil = 3$ sample path derivatives of the $\phi(\cdot)$ processes. I fix ν_ϕ because I don't believe the data can reasonably inform this parameter based on the model structure and coarseness of the data grid. Use of this value does limit the sample path differentiability (Section 2.5.5) of the underlying spatial residual processes.

To use the Matérn function on the sphere, I transform angular distance to Euclidean distance in $\Re^3$, calculating the chordal distance, $\tau = \sin\left(\frac{\rho}{2}\right)$, where ρ is angular distance. As a function of chordal distance, the usual Matérn function defined by

$$R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{2\sqrt{\nu}\tau}{\kappa}\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}\tau}{\kappa}\right)$$

is a legitimate correlation function. While the chordal distance underestimates the true distance, I believe it is a reasonable distance measure for pairs of locations close enough in proximity to have non-negligible correlation.

One might argue on principle with my use of stationary priors for $\alpha(\cdot), \beta(\cdot)$, and $\log \eta(\cdot)^2$ based on the climatological evidence I discuss for why nonstationary models are more reasonable for the actual observations. However, using nonstationary priors for these fields is not practical because of computation limitations. Adding the necessary parameters high in the model hierarchy will make it even harder for the model to mix adequately. In practice, stationary priors may be sufficient, because the posterior covariance of these fields incorporates the data covariance model and will therefore be nonstationary if the data covariance model is nonstationary. As more data are collected, the posteriors for the fields will be less and less influenced by their stationary priors. For $\boldsymbol{\phi} \in \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ we can see this in closed form. Conditional on the other parameters, represented as $\boldsymbol{\theta}$, including the variance vector $\boldsymbol{\eta}$, these vectors are a posteriori normal with moments,

$$\begin{aligned} \mathrm{E}\boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{\theta} &= C_{\boldsymbol{\phi}}\left(C_{\hat{\boldsymbol{\phi}}} + C_{\boldsymbol{\phi}}\right)^{-1}\hat{\boldsymbol{\phi}} + C_{\hat{\boldsymbol{\phi}}}\left(C_{\hat{\boldsymbol{\phi}}} + C_{\boldsymbol{\phi}}\right)^{-1}\mu_\phi \\ &= \mu_\phi + C_{\boldsymbol{\phi}}\left(C_{\hat{\boldsymbol{\phi}}} + C_{\boldsymbol{\phi}}\right)^{-1}(\hat{\boldsymbol{\phi}} - \mu_\phi) \qquad (5.2) \\ \mathrm{Cov}(\boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{\theta}) &= \left(C_{\hat{\boldsymbol{\phi}}}^{-1} + C_{\boldsymbol{\phi}}^{-1}\right)^{-1}, \qquad (5.3) \end{aligned}$$

where $\hat{\boldsymbol{\phi}}$ is the vector of MLEs for the field, and $C_{\hat{\boldsymbol{\phi}}}$ is the variance matrix of $\hat{\boldsymbol{\phi}}$. For $\boldsymbol{\alpha}$ we have $\hat{\alpha}_i = \frac{\sum_t Y_{it}}{T}$and $C_{\hat{\boldsymbol{\alpha}}} = \frac{C_{\boldsymbol{Y}}}{51}$, while for $\boldsymbol{\beta}$ we have $\hat{\beta}_i = \frac{\sum_t tY_{it}}{\sum_t t^2}$ and $C_{\hat{\boldsymbol{\beta}}} = \frac{C_{\boldsymbol{Y}}}{\sum_t t^2}$ with $C_{\boldsymbol{Y}} = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta}) + \delta I$ as defined previously. So with a nonstationary correlation model, $R_{\boldsymbol{Y}}$, the posterior specified by (5.2-5.3) is nonstationary for $\boldsymbol{\phi}$. With little data, the prior will dominate and the data will be smoothed in a generally stationary fashion, but with more and more data, the denominators of $C_{\hat{\boldsymbol{\phi}}}$ will increase, driving the posterior variance down and causing the posterior mean to be more and more influenced by the MLEs. The smoothing will then be primarily nonstationary in structure, based on the model-inferred data covariance.

It is possible to integrate both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ out of the model. Doing so produces the marginal likelihood

$$\boldsymbol{Y}_t|\boldsymbol{\theta} \sim \mathrm{N}(\mu_\alpha + \mu_\beta t, C_{\boldsymbol{Y}} + C_{\boldsymbol{\alpha}} + t^2 C_{\boldsymbol{\beta}}),$$

which for computational efficiency can be expressed in an alternate form (not shown) that requires inverting only $C_{\boldsymbol{Y}}, (C_{\hat{\boldsymbol{\alpha}}} + C_{\boldsymbol{\alpha}})$, and $(C_{\hat{\boldsymbol{\beta}}} + C_{\boldsymbol{\beta}})$. To generate samples of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the marginal distributions are

$$\begin{aligned}
\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \mathrm{N}\left(\mu_\alpha + C_{\boldsymbol{\alpha}}\left(C_{\boldsymbol{\alpha}} + C_{\hat{\boldsymbol{\alpha}}}\right)^{-1}\left(\hat{\boldsymbol{\alpha}} - \mu_\alpha\right), C_{\boldsymbol{\alpha}}\left(C_{\boldsymbol{\alpha}} + C_{\hat{\boldsymbol{\alpha}}}\right)^{-1} C_{\hat{\boldsymbol{\alpha}}}\right) \\
\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \mathrm{N}\left(\mu_\beta + C_{\boldsymbol{\beta}}\left(C_{\boldsymbol{\beta}} + C_{\hat{\boldsymbol{\beta}}}\right)^{-1}\left(\hat{\boldsymbol{\beta}} - \mu_\beta\right), C_{\boldsymbol{\beta}}\left(C_{\boldsymbol{\beta}} + C_{\hat{\boldsymbol{\beta}}}\right)^{-1} C_{\hat{\boldsymbol{\beta}}}\right).
\end{aligned}$$

For the results reported in this thesis, I sampled from the full model, choosing not to integrate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ out of the model. This is primarily because while I was developing the model, I was interested in devising a sampling scheme that would not require marginalization because I was interested in non-Gaussian likelihoods. Also, sampling $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ conditionally on the chain for the marginalized model still involves a great deal of calculation and in the end, these are the parameters I am most interested in, so I do need to sample them, although integrating them out would allow me to avoid sampling them at every iteration of the chain. While it is conceivable that integrating out the intercepts and trends would make it easier to sample the remaining parameters, the fact that most of the variability in the data is related to the residual variance and not to the mean structure suggests this will not be the case.

The spatial model defined in this section has obvious similarities with the regression models in Chapter 4, except that here I have built a hierarchical model based on the components used in the regression modelling, and the nonstationary covariance is used to model the residual structure. In fact, we can reexpress the model in a way that makes clear that there is an underlying nonstationary GP prior for functions, just as in the regression modelling. The model is

$$\boldsymbol{Y}_t \sim \mathrm{N}(\boldsymbol{f}_t, \delta I)$$

$$\begin{aligned} \boldsymbol{f}_t &\sim \mathrm{N}\left(\boldsymbol{\mu}_t, D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta})\right) \\ \boldsymbol{\mu}_t &= \boldsymbol{\alpha} + \boldsymbol{\beta}t, \end{aligned} \tag{5.4}$$

with each function, $f_t(\cdot)$, sharing common correlation, $R_{\boldsymbol{Y}}(\cdot,\cdot)$ and variance $\eta(\cdot)^2$ functions, but differing in their mean function, $\mu_t(\cdot)$. In contrast to the regression modelling, in which the regression function has constant mean, μ, and variance, σ^2, hyperparameters, here we have the hyperparameters $\mu_t(x) = \alpha(x) + \beta(x)t$ and $\eta(x)^2$ that vary over the covariate space and themselves have GP priors. I avoid having to sample $\boldsymbol{f}_t$ by integrating them out of the model, which gives the model as originally stated (5.1).

5.4 Covariance Modelling

The previous section did not specify a correlation model for the data. There is a limited literature on fitting spatial covariance models to replicated data of the sort that I use here. In this section, I describe some possible correlation models that could be used, including a stationary model, and discuss potential advantages and disadvantages of the models. I choose three models to compare in this thesis and describe in detail how the three are parameterized. Recall that the covariance of the data is modelled as

$$C_{\boldsymbol{Y}} = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta}) + \delta I.$$

This construction leads me to focus on models for the correlation structure, since I treat the variance structure separately.

5.4.1 Correlation models used in this work

5.4.1.1 Independence model

The simplest model takes $R_{\boldsymbol{Y}} = I$ and estimates trends for each location independently, thereby disregarding the spatial correlation of the data. A second baseline model takes $R_{\boldsymbol{Y}} = I$ but assumes $\boldsymbol{\beta} = \boldsymbol{0}$, the joint null hypothesis at all locations.

Unfortunately, I can't use the MLE of the residual covariance $\hat{C}_{\boldsymbol{Y}} = Y^{*}Y^{*T}$ where Y^{*} is a matrix of the standardized residuals, because Y^{*} has T columns, one for each year of data, and therefore, $\hat{C}_{\boldsymbol{Y}}$, has rank $T-2$, based on estimating $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. This situation is not unusual with climatological and other geophysical data, in which data are available at more locations than time points. The singularity of $\hat{C}_{\boldsymbol{Y}}$ prevents me from using it in the predictive distribution calculations, which involve the inverse of the covariance matrix, used to compare models, as discussed in Section 5.6.

5.4.1.2 Stationary Matérn correlation

The simplest way to include the covariance structure of the data in a model is to use a stationary covariance model. While examination of the data suggests that stationarity is unlikely, it may be that a stationary model fits the data adequately and that the additional complexity of modelling the nonstationarity is not warranted. One drawback to the stationary model, which seems to be realized in the storm data, is that in regions in which the stationary correlation does not fit well, the residual variance portion of the covariance model, $\boldsymbol{\eta}^2$, is driven up relative to the MLE variances for the locations. It seems undesirable that the modelled variance be much larger than is reflected in the empirical variability about the fitted time trend, merely because of lack of fit in the correlation model. A more sophisticated approach would be to embed the stationary model in the deformation model of Sampson and Guttorp (1992), possibly including the deformation in a fully Bayesian fashion (Schmidt and O'Hagan 2000; Damian et al. 2001), but I have not used the deformation approach here.

I use a stationary Matérn model, with the transformation of angular distance to chordal distance in $\Re^3$ and a uniform prior on $\nu_{\boldsymbol{Y}} \in [0.5, 15]$. Note that for simplicity I had initially fixed $\nu_{\boldsymbol{Y}} = 2$ but this resulted in fitted correlations and variances not consistent with the data and different results than using fixed $\nu_{\boldsymbol{Y}} = 4$. This suggests that it is better to allow the model to choose the value of ν, as I have done here.

5.4.1.3 Nonstationary smoothed empirical correlation

Nychka et al. (2001) introduced a wavelet-based method for smoothing the empirical covariance of data that lie on a regular grid. As an alternative to the kernel-based nonstationary covariance described next, I use this smoothing method on the residuals, $\boldsymbol{U}_t = \boldsymbol{Y}_t - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}t$, of the storm activity data obtained by subtracting the ML trends from the data. To explain the wavelet method, first consider the eigendecomposition of a positive definite matrix. The decomposition expresses the empirical covariance as $\hat{C} = \Gamma\Lambda\Gamma^T$, based on the matrix of eigenvectors, Γ, and a diagonal matrix of eigenvalues, Λ. If we consider the dual space in which each year of data is produced as a linear combination of basis functions, we have $\boldsymbol{U}_t = \Gamma\boldsymbol{\omega}_t$, with $\boldsymbol{\omega}_t \sim \mathrm{N}(0, \Lambda)$. Instead of the eigendecomposition, the wavelet method uses the W wavelets, which are a nonorthogonal wavelet basis whose functions are piecewise quadratic splines. Denoting the basis matrix as Ψ, each year of residuals can be expressed as $\boldsymbol{U}_t = \Psi\boldsymbol{\omega}_t$, so collecting the vectors $\boldsymbol{\omega}_t$ into a matrix Ω we have $\hat{C} = \frac{1}{T}UU^T = \frac{1}{T}\Psi\Omega\Omega^T\Psi^T = \Psi\hat{\Lambda}\Psi^T$ where the empirical covariance matrix of the coefficients, $\hat{\Lambda}$, is no longer diagonal because we are not using the eigendecomposition. To smooth the empirical covariance instead of reproducing it exactly, the method thresholds the off-diagonal elements of a square root matrix, $\hat{H}$, of $\hat{\Lambda}$ and reconstructs a smoothed version of the original empirical covariance, $\tilde{C} = \Psi\tilde{H}\tilde{H}^T\Psi^T = \Psi\tilde{\Lambda}\Psi^T$. The calculations are very fast because $\hat{\Lambda}$ can be calculated using the discrete wavelet transform, as can elements of $\tilde{C}$ once $\tilde{\Lambda}$ is computed. Calculating $\hat{H}$ is slow $O(n^3)$ if one naively takes the square root of the n by n matrix, $\hat{\Lambda}$, where n is the number of locations. Instead, one can calculate $\hat{H}$ from the SVD of Ω, which is $O(n^2T)$, which greatly speeds computation if the number of replicated observational fields, T, is much less than the number of locations. Even greater efficiency can be obtained by only calculating the square root matrix for the leading submatrix of $\hat{\Lambda}$, which is possible if one decides to threshold only the leading submatrix and zero out all of the remaining elements of $\hat{\Lambda}$, save the diagonals.

Nychka et al. (2001) have shown that such a smoothing approach can closely approximate a stationary Matérn covariance in the sense that the elements of the resulting smoothed matrix are similar to those of the original Matérn covariance matrix. The method can give a nonstationary covariance because the original empirical covariance is nonstationary, and nothing in the thresholding enforces stationarity. The intuition behind the method is that real structure is modelled through the retained coefficients, while noise is zeroed out during the thresholding. Unfortunately, there is no principled way to choose the degree of or exact procedure for the thresholding. In their empirical example, Nychka et al. (2001) set the smallest (in magnitude) 90% of the off-diagonal elements in the leading block (which corresponds to the coefficients of the coarsest father wavelets, which look like smooth bumps on a grid) of $\hat{H}$ to zero, retain all the diagonal elements of $\hat{H}$, and zero out the remaining elements.

Since it is not clear how one should do the thresholding, I create two smoothed correlation matrices based on the wavelet decomposition. The first aims to mimic the empirical residual correlation closely, while being positive definite. To create this correlation matrix, I use the empirical residual covariance, $\hat{C}_{\boldsymbol{Y}}$, and implement the thresholding in the following way. In the leading block of $\hat{H}$ I retain all the elements, while in the remainder of $\hat{H}$ I retain the diagonals and the largest (in magnitude) 50% of the off-diagonals. Because I want to model the variances as spatial processes with Gaussian process priors in the same way for all of the correlation models, I then create a smoothed correlation matrix, $\tilde{R}_{\boldsymbol{Y}}$ from $\tilde{C}_{\boldsymbol{Y}}$. In the MCMC, I treat $R_{\boldsymbol{Y}} = \tilde{R}_{\boldsymbol{Y}}$ as fixed, and model $\alpha(\cdot)$, $\beta(\cdot)$, $\eta(\cdot)$, and δ as before. Comparing the matrix $\tilde{R}_{\boldsymbol{Y}}$ to the empirical residual correlation matrix, the differences between elements of the matrices are at most 0.02. The second matrix aims to smooth the empirical correlation matrix, while retaining obvious structure. To do this I follow the procedure above, but include only the largest 0.5% of the off-diagonals that are not in the leading block.

The wavelet decomposition could be done within the context of a full Bayesian model, with priors on the elements of either Λ or H, but this would seem to require sampling matrix elements one by one. This would be very slow in a MCMC sampling scheme because, as each element changes, $C_{\boldsymbol{Y}}$ and its Cholesky factor would need to be recomputed. Furthermore, developing a prior framework for Λ is beyond the scope of this work.

5.4.1.4 Kernel-based Matérn nonstationary correlation

To fit a nonstationary covariance model within a fully Bayesian framework and account for uncertainty in the nonstationary covariance, I use the Matérn form of the kernel-based nonstationary covariance defined in Chapters 2 and 3. I use the basis kernel approach described in Section 3.2.4 to limit the computations. I parameterize the basis kernels, using the Givens angle approach given in Section 3.2.3.1. The pqth element of the kernel matrix at location i is given by

$$(\Sigma_i)_{pq} = \frac{\sum_{k=1}^{K} w_{ik} (\Sigma'_k)_{pq}}{\sum_{k=1}^{K} w_{ik}},$$

where Σ'_k is the kth basis kernel matrix, and w_{ik} are the weights defined below. Since sums of positive definite matrices are positive definite, the resulting kernel matrix, Σ_i, obtained by averaging the basis kernel matrices element by element, is positive definite. I weight the basis kernels based on a squared exponential weighting function

$$w_{ik} = \exp\left(-\left(\frac{\rho_{ik}}{\kappa_{\boldsymbol{Y}}}\right)^2\right), \tag{5.5}$$

where ρ_{ik} is angular distance between location i and the center of basis kernel k and $\kappa_{\boldsymbol{Y}}$ determines how quickly the weight decays with distance. The squared exponential weighting function (5.5) is infinitely differentiable as a function of ρ_{ik}, and we can express $\rho_{ik} = 2\sin^{-1}(\tau_{ik}) = \frac{1}{2}\cos^{-1}(-2\tau_{ik}^2 + 1)$, where τ_{ik} is distance in $\Re^3$. If we consider any element of Σ_i as a function in $\Re^3$ we can see that the function is infinitely differentiable as a function of location, which satisfies the conditions needed in Section 2.5.5 to show sample path differentiability of GPs using this nonstationary covariance model. This means that the differentiability of the residual functions, $f_t(\cdot)$ (5.4), which are integrated out of the model to give (5.1), are based on $\nu_{\boldsymbol{Y}}$ and ν_ϕ. I take $\nu_{\boldsymbol{Y}} \sim U(0.5, 15)$.

In Section 2.4, I discussed how to define kernel matrices on the sphere in a way that generates a nonstationary positive definite covariance on the sphere. To avoid having to do numerical integration, I use a shortcut here that does not seem to cause numerically non-positive definite matrices. To calculate the covariance between two locations, I locate the two points in the Lambert azimuthal equidistant Euclidean projection of the sphere that is centered at the midpoint of the great circle connecting the two locations. This projection accurately preserves the distances and directions from the centerpoint of the projection such that any circle in the Euclidean projection whose origin is the centerpoint of the projection gives a locus of points that are truly equidistant from the centerpoint. The effect of the projection is as if one took the globe with the centerpoint pointed straight up and squashed it directly down onto a plane. I calculate the closed form kernel covariance given in (2.5) based on the kernel matrices defined in this projection. Since I change the midpoint of the projection each time I change the locations under consideration, it is not strictly true that the kernels are solely a function of location, as is required for positive definiteness (2.2), but for the portion of the hemisphere used here, this approximation does not seem to cause problems.

A disadvantage of this nonstationary model is that even though it is more flexible than a stationary model, I still rely on the formulation of convolving kernels, which constrains the types of correlation structure that can be modelled. In particular, the kernel convolution model will not model correlations that do not drop off monotonically from the focal location well, and negative correlations are not possible in the model. This latter inability is a drawback for these data because the empirical data suggest that negative correlations are present, particularly between latitude bands (Figure 5.1). Presumably these negative correlations occur as the storm tracks shift in latitude over time.

5.4.2 Alternate approaches not used in this work

5.4.2.1 Smoothing the MLE fields

An approach that does not require modelling the full covariance of the data is to take the MLE fields for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ and smooth these spatially. If one wanted to do this in a nonstationary way, it would essentially involve doing the regression modelling of Chapter 4, taking the MLEs as the observations. This approach is similar to that of Holland et al. (2000). The main drawback to this approach is that it does not provide uncertainty estimates of the trends that take account of the data correlation unless this is included, as done by Holland et al. (2000) in an ad hoc way. Because the ML estimates are smooth spatially, the estimated intercept and trend functions will be smooth, and the modelled noise about these functions will have very small variance. Represent model (5.1) as

$$Y_{it} = \alpha(\boldsymbol{x_i}) + \beta(\boldsymbol{x_i})t + r_t(\boldsymbol{x_i}) + \epsilon_{it},$$

where $r_t(\cdot)$ is a spatial residual process with $\text{Cov}(\boldsymbol{r_t}) = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta})$. Naive smoothing of the ML estimates, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, ignores the $r_t(\cdot)$ component of the model even though exploratory analysis suggests that $\alpha(\cdot)$ and $r_t(\cdot)$ account for most of the variability in the data. This is precisely what I want to avoid.

5.4.2.2 Smoothing the empirical covariance toward a stationary covariance

Like Nychka et al. (2001), Loader and Switzer (1992) focus on smoothing an empirical estimate of the covariance structure based on replicated data. They smooth toward a stationary model of the covariance structure, $C^S(\hat{\theta})$, giving

$$\tilde{C}_{\boldsymbol{Y}} = w\hat{C}_{\boldsymbol{Y}} + (1-w)C^S(\hat{\theta}),$$

where $w \in (0,1)$ is a smoothing parameter. The estimate, $\tilde{C}_{\boldsymbol{Y}}$, can be derived as the posterior mean of Bayesian model in which the prior for $C_{\boldsymbol{Y}}$ is inverse Wishart, and Loader and Switzer (1992) use this derivation to optimize w. It would be interesting to assess the success of this simple smoothing approach relative to those used in this work, particularly because the optimization of the degree of smoothing in this approach is straightforward, whereas choosing the degree of thresholding in the Nychka et al. (2001) model is difficult.

5.4.2.3 Fitting an unconstrained data covariance

Another possibility is to define a simple prior for the data correlation matrix, such as a uniform prior over the compact space of correlation matrices (Lockwood 2001), and then sample the matrices. It is possible to sample from unconstrained correlation matrices element by element in a way that ensures positive definiteness (Barnard, McCulloch, and Meng 2000; Lockwood 2001), but element by element sampling would be extremely computationally intensive, since the data covariance and its Cholesky factor would change each time an element changed. With 288 locations, this does not seem feasible; furthermore with only 51 years of data we do not have much data for fitting an unconstrained correlation matrix, and it is not clear how such a prior weights various types of correlation structures.

Using Bayesian methods, Daniels and Kass (1999) consider freely estimating small covariance matrices based on a variety of priors, while Smith and Kohn (2002) estimate sparse Cholesky decompositions of covariance matrices in a computationally efficient way. Some combination of the ideas in these papers may be feasible for the storm activity data, but would entail further methodological development, so I have not pursued these avenues further.

5.4.2.4 Covariances modelled indirectly through basis function regression

An approach that shares some features with the wavelet decomposition is recommended by Minka and Picard (1997). They suggest fitting each replicate of the residuals, $\boldsymbol{U}_t$, using a multilayer perceptron (feed-forward neural network), as $\boldsymbol{U}_t = B\boldsymbol{\omega}_t + \boldsymbol{\epsilon}_t$, and then reconstructing the implicitly modelled covariance as $\widetilde{C}_{\boldsymbol{Y}} = B\Omega\Omega^T B^T$ where Ω is a matrix whose columns are the vectors $\boldsymbol{\omega}_t$. The basis B could in principle be any basis; the critical choices that must be made involve the basis used, the number of basis functions, and the fitting method, since the degree to which $\widetilde{C_{\boldsymbol{Y}}}$ smooths the empirical covariance, $\hat{C}_{\boldsymbol{Y}}$, will be determined by these. The attraction in this approach is that one is doing regression modelling, which does not involve the constraints of positive definiteness, rather than covariance modelling. One can in principle model covariances involving locations not in the training set as well as incorporate uncertainty in the estimated covariance, but this requires one to model Cov$(\boldsymbol{\omega}_t)$ in some fashion, which leads back to the difficulties involved in covariance modelling.

5.5 Fitting the Model

5.5.1 Data and data-dependent model specifications

I fit the model for two of the storm indices in Paciorek et al. (2002). First, I fit the model to the logarithm of the temperature variance index in the Pacific region of the Northern Hemisphere, defined as $20° - 75°$ N, $130° - 245°$ E (Figure 5.3). Second, I fit the model to the Eady growth rate in the Atlantic region of the Northern Hemisphere, defined as $20° - 75°$ N, $250° - 5°$ E (Figure 5.3). Using a $5° \times 5°$ subgrid of the original $2.5° \times 2.5°$ NCEP-NCAR reanalysis grid, this gives 288 locations. While we would ideally like to fit the data from the whole hemisphere, $20° - 70°$ N, that was analyzed in Paciorek et al. (2002) and to the finer grid, the computational limitations of this hierarchical model with GP priors and nonstationary covariance limit the number of locations that can be fit. Even with only 288 locations, a full MCMC takes several weeks on a moderately-fast computer. The wavelet code of Nychka et al. (2001) requires a regular grid based on powers of two (with at most one factor of three allowed), so I include $75°$ N (which was not used in the analysis of Paciorek et al. (2002)) so that with one-third of a hemisphere, I have a grid with 24 longitude values and 12 latitude values. I use this same grid for the non-wavelet-based correlation models to facilitate comparison between models.

For the wavelet representation, at the coarsest level, I use a grid with 12 longitude and 6 latitude

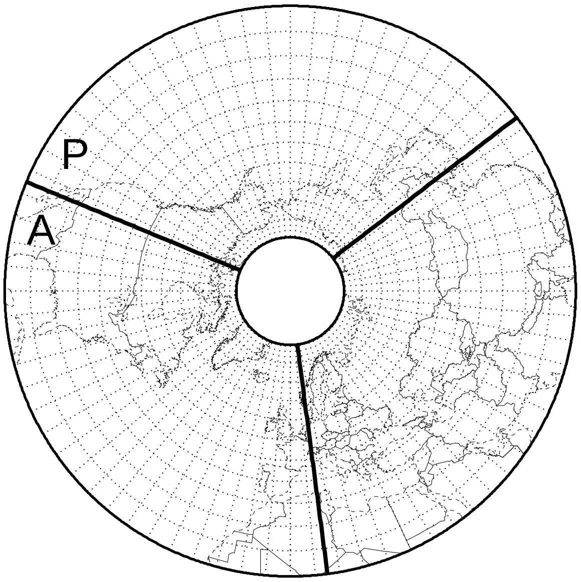

Figure 5.3. Map of the Northern hemisphere, $20^\circ - 75^\circ$ *N, with* $5^\circ \times 5^\circ$ *grid overlaid as dotted lines and Pacific (P) and Atlantic (A) region boundaries indicated by the thick dark lines of longitude.*

values, which means there are 72 'smooth' basis functions (using the terminology of Nychka et al. (2001)). These basis functions are essentially bumps centered on the subgrid; they are able to pick up patterns in the covariance structure of resolution approximately 10°. The leading submatrix of $\hat{H}$, to which I refer in Section 5.4.1.3, is therefore a 72 by 72 matrix.

While the latitude-longitude grid of the data distorts distances and areas, I rely on the ability of the wavelet-based decomposition to model nonstationarity to account for the changes in distances with latitude. Even if the data truly were stationary, this would require that the decomposition give a longer correlation scale at high latitude than at low latitude, to account for the distances between grid points being shorter at high latitude.

In constructing the kernel convolution nonstationary covariance model, I use nine basis kernels, which I believe are sufficient to represent the basic nonstationarity in the data. I position the nine kernels on a three by three latitude-longitude grid, using three kernels at each of the latitudes 30° N, 45° N, and 60° N, with the kernels 40 degrees apart in longitude. For the Atlantic region, the longitudes are 270° E, 310° E and 350° E, and for the Pacific, 150° E, 190° E, and 230° E.

5.5.2 MCMC sampling procedure

5.5.2.1 Posterior mean centering

I fit the model via MCMC. For the spatial parameters, $\boldsymbol{\phi} \in \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \log(\boldsymbol{\eta}^2)\}$, I use the posterior mean centering (PMC) scheme outlined in Section 3.6.2.2. Because the years are equally-spaced, I can center time about the mean of the years and sample $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ independently. This simplifies the sampling and justifies using independent priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In the PMC sampling for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ I can use the exact conditional posterior mean, since the prior and likelihood are both of Gaussian form for these parameters. Note that I could integrate parameters out of the model, which might speed the calculations and improve mixing. For $\boldsymbol{\phi} = \log(\boldsymbol{\eta}^2)$, I use an approximation to the posterior mean, based on (5.2) using the usual MLE $\hat{\boldsymbol{\phi}}$:

$$\widehat{\phi(\boldsymbol{x_i})} = \log \widehat{\eta(\boldsymbol{x_i})}^2 = \log \left(\frac{\sum_t (Y_{it} - \widehat{\alpha(\boldsymbol{x_i})} - \widehat{\beta(\boldsymbol{x_i})} t)^2}{T} \right),$$

where the hats indicate the usual MLEs calculated from the data independently by location. To approximate $C_{\hat{\boldsymbol{\phi}}}$ in (5.2), I first calculate the covariance of $\hat{\boldsymbol{\eta}}^2$ empirically and then use the delta method to calculate the approximate covariance of $\log(\hat{\boldsymbol{\eta}}^2)$. I derive the empirical covariance of $\hat{\boldsymbol{\eta}}^2$ as follows. Let the residual at location i and time t be denoted $U_{it} = Y_{it} - \alpha(\boldsymbol{x_i}) - \beta(\boldsymbol{x_i})t$ and let $C = C_{\boldsymbol{Y}}$ be the covariance matrix of the residuals with ijth element, C_{ij}. I need the following moments

$$\begin{aligned} \mathrm{E}U_{it}^2 &= C_{ii} \\ \mathrm{E}\left(U_{it}^2 U_{js}^2\right) &= C_{ii}C_{jj} + 2C_{ij}^2, \end{aligned}$$

where the second expression is derived straightforwardly, but tediously, based on (U_{it}, U_{jt}) being bivariate normal with covariance, C_{ij}. Now, consider

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\eta}}^2)_{ij} &= \mathrm{E}\left(\frac{\sum_{t=1}^T \sum_{s=1}^T U_{it}^2 U_{js}^2}{T^2}\right) - \mathrm{E}\left(\frac{\sum_{t=1}^T U_{it}^2}{T}\right) \mathrm{E}\left(\frac{\sum_{t=1}^T U_{jt}^2}{T}\right) \\ &= \frac{1}{T^2} \sum_t \sum_s \mathrm{E}\left(U_{it}^2 U_{js}^2\right) - C_{ii}C_{jj} \\ &= \frac{T(T-1)}{T^2} C_{ii}C_{jj} + \frac{T}{T^2}(C_{ii}C_{jj} + 2C_{ij}^2) - C_{ii}C_{jj} \\ &= \frac{2C_{ij}^2}{T}. \end{aligned}$$

Next, I use a second-order Taylor expansion applied to $\log \widehat{\eta(\boldsymbol{x_i})}^2$. Suppressing the dependence on $\boldsymbol{x_i}$, this gives me

$$\hat{\phi} = \log \hat{\eta}^2 \approx \log \eta^2 + (\eta^2 - \hat{\eta}^2)\frac{1}{\eta^2}.$$

Calculating the covariance of the right-hand side and plugging in $\widehat{\eta(\boldsymbol{x_i})}$ for $\eta(\boldsymbol{x_i})$, I can now approximate $C_{\hat{\boldsymbol{\phi}}}$ as

$$\text{Cov}(\hat{\boldsymbol{\phi}}) \approx D(\hat{\boldsymbol{\eta}}^{-2})\text{Cov}(\hat{\boldsymbol{\eta}}^2)D(\hat{\boldsymbol{\eta}}^{-2}),$$

where $D(\hat{\boldsymbol{\eta}}^{-2})$ is a diagonal matrix with the reciprocals of the MLE variances on the diagonal. Using $\hat{\boldsymbol{\eta}}$ as a plug-in estimator (which simplifies the acceptance ratio calculations) again in the expression $C_{\boldsymbol{Y}} = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta}) + \delta I$, the final result is

$$C_{\hat{\boldsymbol{\phi}}} \approx \frac{2}{T}D(\hat{\boldsymbol{\eta}}^{-2})(D(\hat{\boldsymbol{\eta}})R_{\boldsymbol{Y}}D(\hat{\boldsymbol{\eta}}) + \delta I)^{*2}D(\hat{\boldsymbol{\eta}}^{-2}), \tag{5.6}$$

where the $*2$ notation indicates squaring element by element. Note that for $\delta \approx 0$, which is the case for the spatial model, we have

$$C_{\hat{\boldsymbol{\phi}}} \approx \frac{2}{T}R_{\boldsymbol{Y}}^{*2},$$

which does not involve $\hat{\boldsymbol{\eta}}^2$, which makes sense because the logarithm is variance-stabilizing.

5.5.2.2 Sampling steps

As in Section 4.3, I define the basic posterior mean centering proposal S1. Proposal S1 in the context of the spatial model is as follows.

1. This proposal applies to either μ or to the pair (σ, κ). Propose the hyperparameter(s), using a Metropolis or Metropolis-Hastings proposal. In the notation that follows, I will indicate that all three of the hyperparameters μ, σ, and κ have been proposed, but this is merely for notational convenience.

2. Propose $\boldsymbol{\phi} \in \boldsymbol{\alpha}, \boldsymbol{\beta}, \log \boldsymbol{\eta}^2$ conditionally on $\boldsymbol{\theta}^* = \{\mu^*, \sigma^*, \kappa^*\}$ as

$$\boldsymbol{\phi}^* \sim \text{N}(\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta}^*)} + \sigma^* L(\kappa^*)\boldsymbol{\chi}, v^2 R(\kappa^*)),$$

where $\boldsymbol{\chi} = (\sigma L(\kappa))^{-1}(\boldsymbol{\phi} - \widetilde{\phi(\boldsymbol{\theta})})$. $\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta})}$is the posterior mean of $\boldsymbol{\phi}$ conditional on the current hyperparameters and $\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta}^{*}}$)is the posterior mean of $\boldsymbol{\phi}$ conditional on the proposed hyperparameter(s),

$$\begin{aligned}\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta})} &= \sigma^2 R_{\boldsymbol{\phi}}(\kappa)(C_{\hat{\boldsymbol{\phi}}} + \sigma^2 R_{\boldsymbol{\phi}}(\kappa))^{-1}\hat{\boldsymbol{\phi}} + C_{\hat{\boldsymbol{\phi}}}(C_{\hat{\boldsymbol{\phi}}} + \sigma^2 R_{\boldsymbol{\phi}}(\kappa))^{-1}\mu \\ &= \mu + \sigma^2 R_{\boldsymbol{\phi}}(\kappa)(C_{\hat{\boldsymbol{\phi}}} + \sigma^2 R_{\boldsymbol{\phi}}(\kappa))^{-1}(\hat{\boldsymbol{\phi}} - \mu).\end{aligned}$$

$\hat{\boldsymbol{\phi}}$ is the MLE for $\boldsymbol{\phi}$, and $C_{\hat{\boldsymbol{\phi}}}$ is the covariance of the MLE, either the approximation (5.6) for $\log \boldsymbol{\eta}^2$ or the exact covariance in Section 5.3 for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Note again that, as mentioned in Chapter 3, it is allowable to set $v = 0$, so long as one includes the Jacobian of the deterministic mapping $\boldsymbol{\phi} \rightarrow \boldsymbol{\phi}^{*}$, which is the same as the Hastings ratio one uses if $v > 0$.

3. The Hastings ratio (the proposal ratio portion of the acceptance ratio) for the proposal is a ratio of determinants, which cancels the determinant ratio from the prior for $\boldsymbol{\phi}$ as described in Section 3.6.2.2.

Next I describe the sampling steps in a single iteration of the Markov chain for the nonstationary correlation model using the basis kernels. The steps for the stationary correlation and for the fixed wavelet-based correlation models are straightforward simplifications of this scheme.

1. Sample δ using a simple Metropolis step.

2. Sample $\log \kappa_{\boldsymbol{Y}}$, the weighting parameter for the basis kernel averaging, using a simple Metropolis step.

3. For $\boldsymbol{\theta} = (\log \lambda_{1,1}, \ldots, \log \lambda_{1,K})$, the first eigenvalues of the $K = 9$ basis kernels, use a simple multivariate Metropolis step with the same proposal standard deviation for each element. The basis kernels are sampled jointly in this fashion to avoid having to recalculate the Cholesky of the data covariance matrix too frequently. In practice the basis kernel parameters seem to mix much more quickly than the hyperparameters of $\boldsymbol{\phi}$, so this does not seem to be a problem.

4. Repeat step 3 for $\boldsymbol{\theta} = (\log \lambda_{2,1}, \ldots, \log \lambda_{2,K})$, the second eigenvalues of the basis kernels.

5. Repeat step 3 for $\boldsymbol{\theta} = (\rho_1, \ldots, \rho_K)$, the Givens angles of the basis kernels. If I propose ρ_k outside of $(0, \pi)$ I add or subtract π as necessary to bring the proposal back into the parameter space.

6. For $\boldsymbol{\phi} = \boldsymbol{\alpha}$, do the following proposals:

   (a) Sample $(\mu_\phi, \boldsymbol{\phi})$ jointly using a proposal of type S1.

   (b) To account for the high posterior correlation between κ_ϕ and σ_ϕ, sample $(\kappa_\phi, \sigma_\phi, \boldsymbol{\phi})$ jointly using a modified proposal of type S1. The modification of substep (1) of S1 is as follows. First, sample $\log \kappa_\phi^* \sim \mathrm{N}(\log \kappa_\phi, v_1^2)$ using a simple Metropolis step. Next, sample

   $$\log \sigma_\phi^* \mid \kappa_\phi^* \sim \mathrm{N}\left(\log \sigma_\phi \frac{v_1}{r}(\log \kappa_\phi^* - \log \kappa_\phi), v_2^2\right).$$

   I take v_2 to be very small, so that κ_ϕ and σ_ϕ move together closely with the ratio r being a constant scale factor chosen to speed mixing. Finally use substep 2 of S1 to sample $\boldsymbol{\phi}^*$ conditional on $(\kappa_\phi^*, \sigma_\phi^*)$.

   (c) Next, allow $\log \sigma_\phi$ to move independently of $\log \kappa_\phi$. Sample $\log \sigma_\phi$ using a simple Metropolis step based on the centered parameterization without changing $\boldsymbol{\phi}$. In other words, the acceptance ratio will only involve the ratio of the priors for $\log \sigma_\phi^*$ and $\log \sigma_\phi$ and the ratio of the prior for $\boldsymbol{\phi}$ as a function of $\log \sigma_\phi^*$ to that of the prior for $\boldsymbol{\phi}$ as a function of $\log \sigma_\phi$. Note that this does not involve having to invert $R_{\boldsymbol{\phi}}(\kappa_\phi)$. I sample $\log \sigma$ in this way rather than with a PMC step of type S1 because joint sampling mixes very slowly, with the chain spending a long time in parts of the space with large values for the elements of the implicit parameter, $\boldsymbol{\omega} = (\sigma_\phi L(\kappa_\phi))^{-1}(\boldsymbol{\phi} - \mu_\phi)$, and correspondingly low values for the log of the prior density, which is a function of $\boldsymbol{\omega}^T \boldsymbol{\omega}$. The reason for this is still unclear to me, but the phenomenon seems to occur because it is difficult for σ_ϕ and $\boldsymbol{\phi}$ to 'trade-off' when employing joint proposals for $(\sigma_\phi, \boldsymbol{\phi})$. In the joint proposals, when large values of $\sigma_{\boldsymbol{\phi}}$ are proposed using PMC, the result is that $\boldsymbol{\phi}$ is proposed conditionally on σ_ϕ^* such that it has higher variability, and it is difficult to increase the prior density of $\boldsymbol{\phi}$ by increasing σ_ϕ. Using a straightforward centered parameterization style proposal for σ_ϕ allows σ_ϕ and $\boldsymbol{\omega}$ to stabilize with $\boldsymbol{\omega}$ approximately $N(0, I)$ and $\boldsymbol{\omega}^T\boldsymbol{\omega} \approx n = 288$.

(d) Propose $\boldsymbol{\phi}$ using a simple Metropolis step with correlation amongst the elements of $\boldsymbol{\phi}$: $\boldsymbol{\phi}^* \sim \mathrm{N}(\boldsymbol{\phi}, v^2 R(\kappa_{\boldsymbol{\phi}}))$. It is also straightforward to do a Langevin update here, however, I did not use such an update in the model runs reported in this thesis.

7. Repeat step 7 for $\boldsymbol{\phi} = \boldsymbol{\beta}$. Again a Langevin update in substep (d) would be straightforward.

8. Repeat step 7 for $\boldsymbol{\phi} = \log \boldsymbol{\eta}^2$. Note that a Langevin update here in substep (d) is not straightforward because there is no simple form for the gradient of the log posterior with respect to $\log \boldsymbol{\eta}^2$.

5.5.2.3 Running the chain

The priors and initial values are listed in detail in the Appendix. Priors are taken to be relatively noninformative. For parameters involved in the correlation structures I use reasonable lower and upper limits based on the induced correlations between the nearest and most distant pairs of locations analyzed. For the initial values, I would like to use the MLEs for the spatial fields and reasonable values for the hyperparameters based on ad hoc analysis of the MLEs. Unfortunately while I am able to do the latter, it is difficult to know what initial prior correlation structure to use so that the MLEs are consistent with the initial prior correlation, i.e., so that we have

$$\hat{\boldsymbol{\phi}} = \mu_\phi + \sigma_\phi L(\kappa_\phi)\boldsymbol{\omega}$$

with the magnitude of $\boldsymbol{\omega}$ remaining reasonable. Instead, I chose to initialize $\boldsymbol{\phi}$ by taking $\boldsymbol{\omega} \sim \mathrm{N}(0, I)$, namely simulating $\boldsymbol{\omega}$ from its prior.

To fit the model, I run a long MCMC chain. Because of the slow mixing and long computation time, I can only run a limited number of runs and cannot do a full convergence assessment using multiple chains. Instead I assess time series and acf plots and compare the beginning of the chain to the end to get some feel for the behavior of the chain. I cannot be sure that the chain has fully converged, but based on many trial runs during which I was adjusting the parameterizations and sampling schemes, I believe that the general results from the runs, in particular the comparison between models, are legitimate, even if some of the parameters cannot be shown to have completely mixed. However, the slow mixing and potential lack of convergence is a cause for real concern with this model and fitting procedure, although perhaps not any more so than for any large Bayesian model of temporal or spatial structure. Furthermore, since the models are compared via cross-validation, the model comparison results still give a valid indication of which of the correlation structures best model these data.

I used at least 50000 iterations for burn-in (longer in some cases) and then sampled 500,000 iterations for inference and model comparison. To economize on storage space, I retained only every 10th iteration, giving a final sample of size 50,000. In adjusting the proposal variances during burn-in, I attempted to achieve the acceptance rates recommended in Roberts and Rosenthal (2001), namely, 0.44 for scalar parameters and 0.23 for vector parameters, and generally come quite close to these rates. The parameters μ_ϕ and δ mix much more quickly than the other parameters, so to reduce computation, I chose to sample those parameters only once for every five updates of the remaining parameters.

5.6 Model Comparison Criteria

As I discussed in Section 5.2, the main scientific goal of the spatial model is to better estimate the slopes and their uncertainty. Therefore I focus on prediction in time rather than prediction in space. To do this I split the data into $T = 44$ training years and $T^* = 7$ test years, and hold out all the data for the 288 locations in the 7 years, which are distributed throughout the 51 year period (1949,1954,1964,1974,1989,1994, and 1999). Based on these 7 years of data, I use two main criteria to compare the success of the models in prediction. The first is the posterior predictive distribution of the test data,

$$\begin{aligned}\log \prod_{t^*} h(\boldsymbol{Y}_{t^*}|\boldsymbol{Y}_t) &= \log \int \prod_{t^*} h(\boldsymbol{Y}_{t^*}|\boldsymbol{\alpha},\boldsymbol{\beta},C_{\boldsymbol{Y}})\pi(\boldsymbol{\alpha},\boldsymbol{\beta},C_{\boldsymbol{Y}}|\boldsymbol{Y}_t)d\boldsymbol{\theta} \\ &\approx \log \sum_{k=1}^{K} |C_{\boldsymbol{Y}}^{(k)}|^{-\frac{T^*}{2}} \\ &\quad \times \exp\left(-\frac{1}{2}\sum_{t^*}(\boldsymbol{Y}_{t^*}-\boldsymbol{\alpha}^{(k)}-\boldsymbol{\beta}^{(k)}t^*)^T C_{\boldsymbol{Y}}^{(k)-1}(\boldsymbol{Y}_{t^*}-\boldsymbol{\alpha}^{(k)}-\boldsymbol{\beta}^{(k)}t^*)\right),\end{aligned}$$

where the integral is approximated by the average over the posterior simulations. Models that perform well should have high predictive density on test data. This criterion is heavily influenced by how well the models predict the covariance structure of the data. As an alternative, I use mean squared error as my second criterion,

$$\text{MSE} = \frac{\sum_{t^*}(\boldsymbol{Y}_{t^*} - \boldsymbol{\alpha} - \boldsymbol{\beta}t)^T(\boldsymbol{Y}_{t^*} - \boldsymbol{\alpha} - \boldsymbol{\beta}t)}{nT^*}.$$

This criterion focuses on the point predictions and ignores the spatial covariance structure of the test data.

5.7 Results

5.7.1 Convergence

All of the models show evidence of slow mixing. In Figure 5.4, I show time series plots of the log posterior densities for the four models for temperature variance, as suggested in Gelman and Rubin (1992) as a way to monitor convergence to the full joint distribution. The wavelet-empirical model has not burned-in, with the log-likelihood still increasing, reaching 53065, which is far greater than for the model with the next highest log-likelihood, the kernel nonstationary model which spends 95% of iterations in the range (28186,28289). Clearly the wavelet-empirical model is closely fitting the covariance structure of the training data, although as we will see in Section 5.7.3, it does a terrible job in predicting held-out data. While the three remaining models appear to be sampling from their posterior distributions, the posterior density plots (Figure 5.4) suggest that mixing is slow. The log-likelihood for the stationary model is approximately (27481,27583) while the wavelet-smooth model is much lower at (23565,23668). The time series plots of the log posterior density are qualitatively similar for the Eady growth rate dataset (not shown). In Figure 5.5 I show time series plots for the hyperparameters $\mu, \sigma,$ and κ of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\log \boldsymbol{\eta}^2$ for the kernel nonstationary model for temperature variance. These indicate that while the μ parameters have mixed, σ and κ have long-term trends (less pronounced for $\boldsymbol{\beta}$) and have not adequately mixed. The mixing properties of the process hyperparameters in the stationary model and the wavelet-smooth model appear to be similar (not shown). For the kernel nonstationary model for Eady growth rate, the hyperparameter mixing is somewhat better, but still not satisfactory, while mixing of σ_η and κ_η are much worse for the stationary model for Eady growth rate (not shown). The lack of hyperparameter mixing appears to impact the process values to some degree, as seen for three locations in the kernel nonstationary model for temperature variance in Figure 5.6. Mixing in the stationary model is similar while mixing in the wavelet-smooth model is even worse (not shown). Interestingly, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}^2$ appear to mix better than those of $\boldsymbol{\alpha}$; in particular, the values of $\boldsymbol{\beta}$, in which we are most interested, do seem to be mixing reasonably well. Mixing of the process values for the Eady growth rate models is qualitatively similar (not shown). Of course these plots only assess the marginal distributions of the process values and not the joint distributions, so the problems with hyperparameter and log posterior density mixing remain a concern. In general, it does appear that the parameters are staying within a range of values, so the sample from the chain may give us some idea of reasonable parameter values even though I do not have confidence that the sample truly reflects the posterior. Note that the PMC scheme greatly improves the mixing of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and their hyperparameters. It also helps in the mixing of the $\boldsymbol{\eta}$ and its hyperparameters. In particular, μ_η mixes well now, and σ_η and κ_η much better than without PMC, albeit still glacially slowly. The impact of the sampling scheme on the mixing of $\boldsymbol{\eta}$ is of most interest, since in a real application, I would integrate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ out of the model.

I have not carried out a full MCMC convergence diagnosis based on running multiple chains from dispersed starting values and ensuring that the chains all converged to the same posterior. However, for both the stationary and kernel nonstationary models for both datasets, I have run three chains each from dispersed starting values. Although I did not run the chains long enough for the chains to completely converge to the distribution found in the primary runs, the parameters did appear to be converging to the same values. Note that to achieve this, I needed to include a new sampling step in the MCMC, in which I used a step of type S1 (Section 5.5.2.2) for $(\kappa_\phi, \boldsymbol{\phi})$. I believe this is necessary to allow the chain to move quickly from the dispersed starting values in which the pair $(\sigma_\phi, \kappa_\phi)$ are not in the right ratio relative to each other. Once the chain burns-in, this step does not seem to be necessary, and the correlated proposal for σ_ϕ and κ_ϕ seem to be sufficient. For some of the chains, a few of the hyperparameters and some of the process values were somewhat different than the values seen in the primary runs, but I suspect this is an effect of

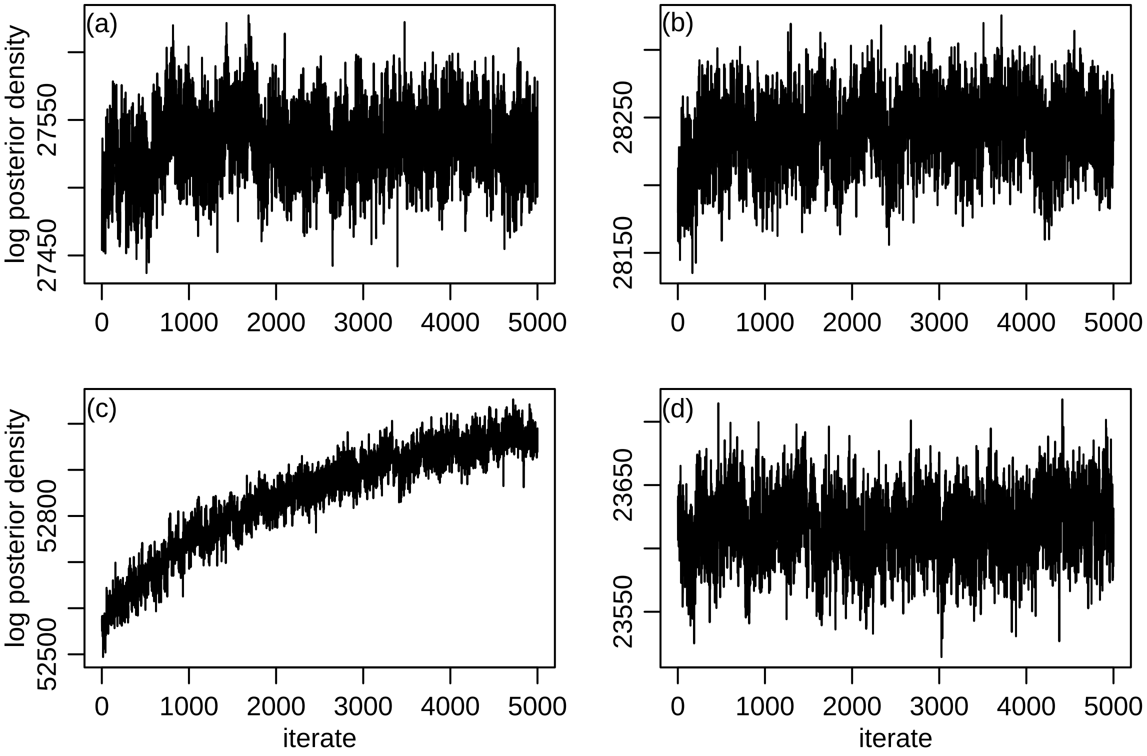

Figure 5.4. Time series plots of the log posterior density for temperature variance for the four Bayesian models: (a) stationary, (b) kernel nonstationary, (c) wavelet-empirical, and (d) wavelet-smooth.

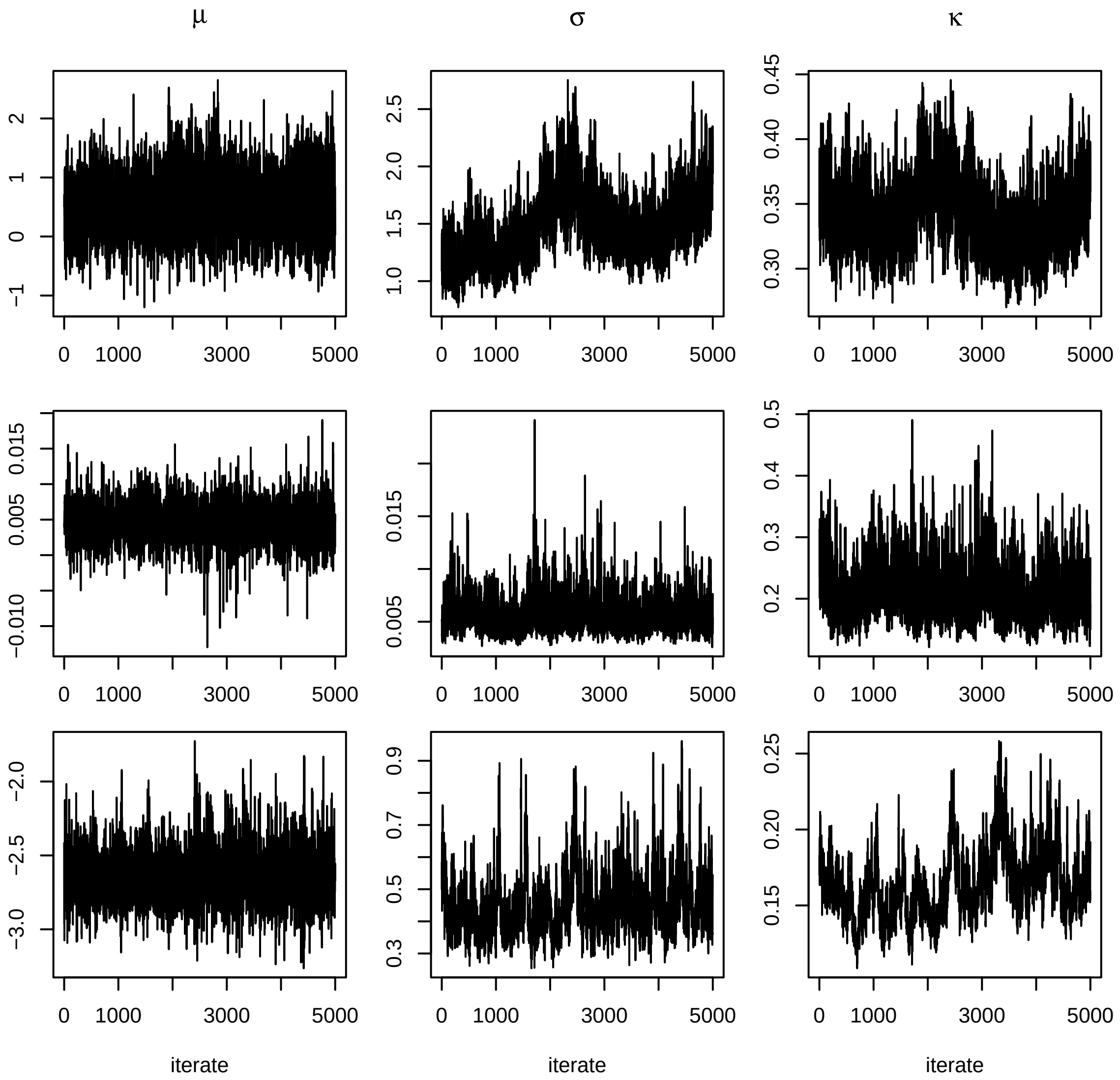

Figure 5.5. Time series plots of the hyperparameters, μ (first column), σ (second column), and κ (third column) for $\boldsymbol{\alpha}$ (first row), $\boldsymbol{\beta}$ (second row), and $\boldsymbol{\eta}^2$ (third row) from the kernel nonstationary model fit to the temperature variance data by MCMC.

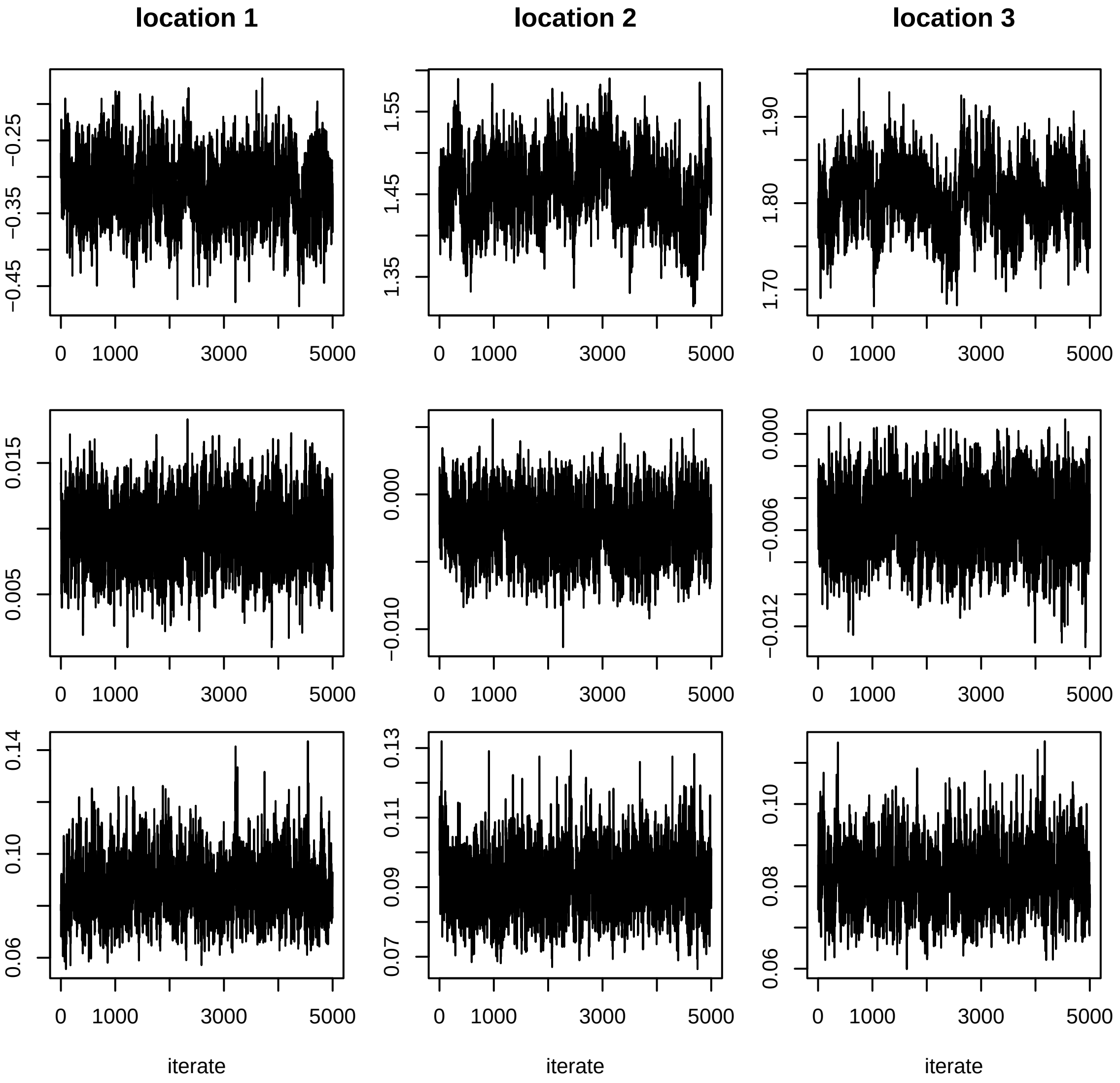

Figure 5.6. Time series plots of three process values of $\boldsymbol{\alpha}$ (first row), $\boldsymbol{\beta}$ (second row) and $\boldsymbol{\eta}^2$ (third row) for the kernel nonstationary model fit to the temperature variance by MCMC.

not having burned-in. In general, this occurs for the σ and κ hyperparameters, and coincides with the chain having log-likelihood values lower than achieved in the primary runs. Also, I have run many runs with these data during the development of the model and therefore have some sense for the range of plausible values of the parameters. While some parameters in the primary runs have not fully mixed, the predictive quantities all seem very stable, suggesting that the results are robust with respect to the comparison of models. In partial defense of the model, mixing is likely to be an issue for most other Bayesian models of similar complexity, and classical fitting methods for large models are prone to finding local minima and not fully exploring the parameter space.

5.7.2 Model estimates

To evaluate the effect of the correlation model used, I first compare the posterior mean estimates of the slopes and residual variances from the four models for temperature variance to the MLEs for the slopes and variances, all based on the training data. In mapping the slopes (5.7), we see that the stationary and kernel nonstationary models smooth the MLE field but retain much of its structure, while the wavelet methods drastically smooth out the peaks and troughs in the MLE field. In the residual variance maps (Figure 5.8), the stationary estimates bear no particular resemblance to the MLEs, while the kernel nonstationary estimates appear to be a smoothed version of the MLEs. The wavelet-smooth estimates are an order of magnitude larger than the MLEs, but the spatial pattern largely mimics that of the MLEs. The wavelet-empirical estimates closely match the pattern of the MLEs but are smaller in value. Note that in mapping these and subsequent quantities, I have used the contour function in the R statistical package, which performs some additional interpolation on top of the interpolation done by the Bayesian model.

In Figure 5.9 I focus further on the smoothing being done by the models of temperature variance using scatterplots of the posterior mean values for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\eta}^2$ for the four full models against the MLE values. Consider first the stationary and kernel nonstationary models. For the intercepts, the posterior means coincide closely with the MLEs. The two models appear to smooth the MLE slopes in similar fashion. For the residual variance, the nonstationary model seems to be doing some smoothing of the MLEs, while the estimates from the stationary model are not clearly related to the MLEs and some of the residual variances are rather high. These high variance estimates

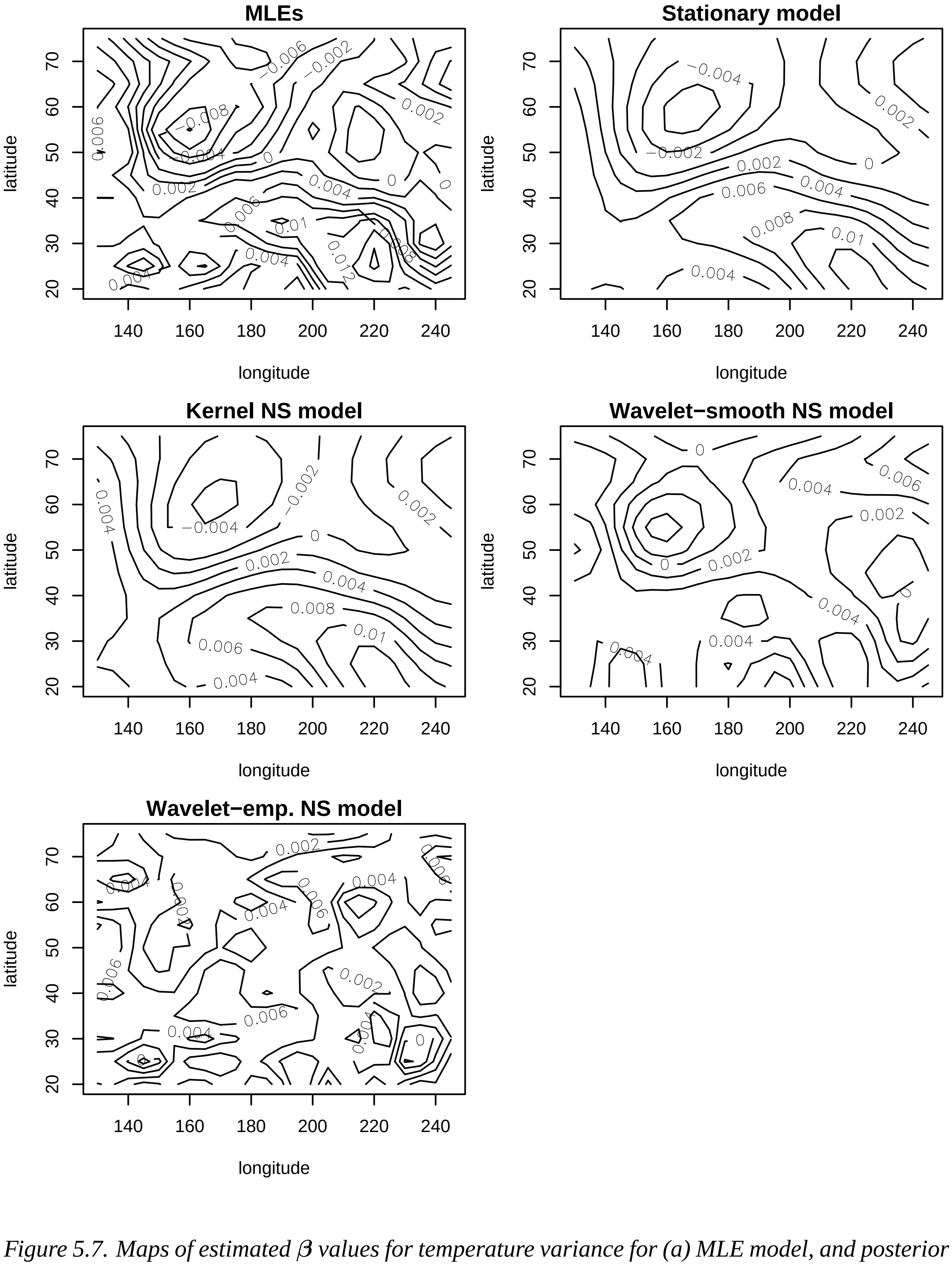

Figure 5.7. Maps of estimated **β** *values for temperature variance for (a) MLE model, and posterior means from (b) stationary model, (c) kernel nonstationary model, (d) wavelet-smoothed covariance model, and (e) wavelet-empirical covariance model.*

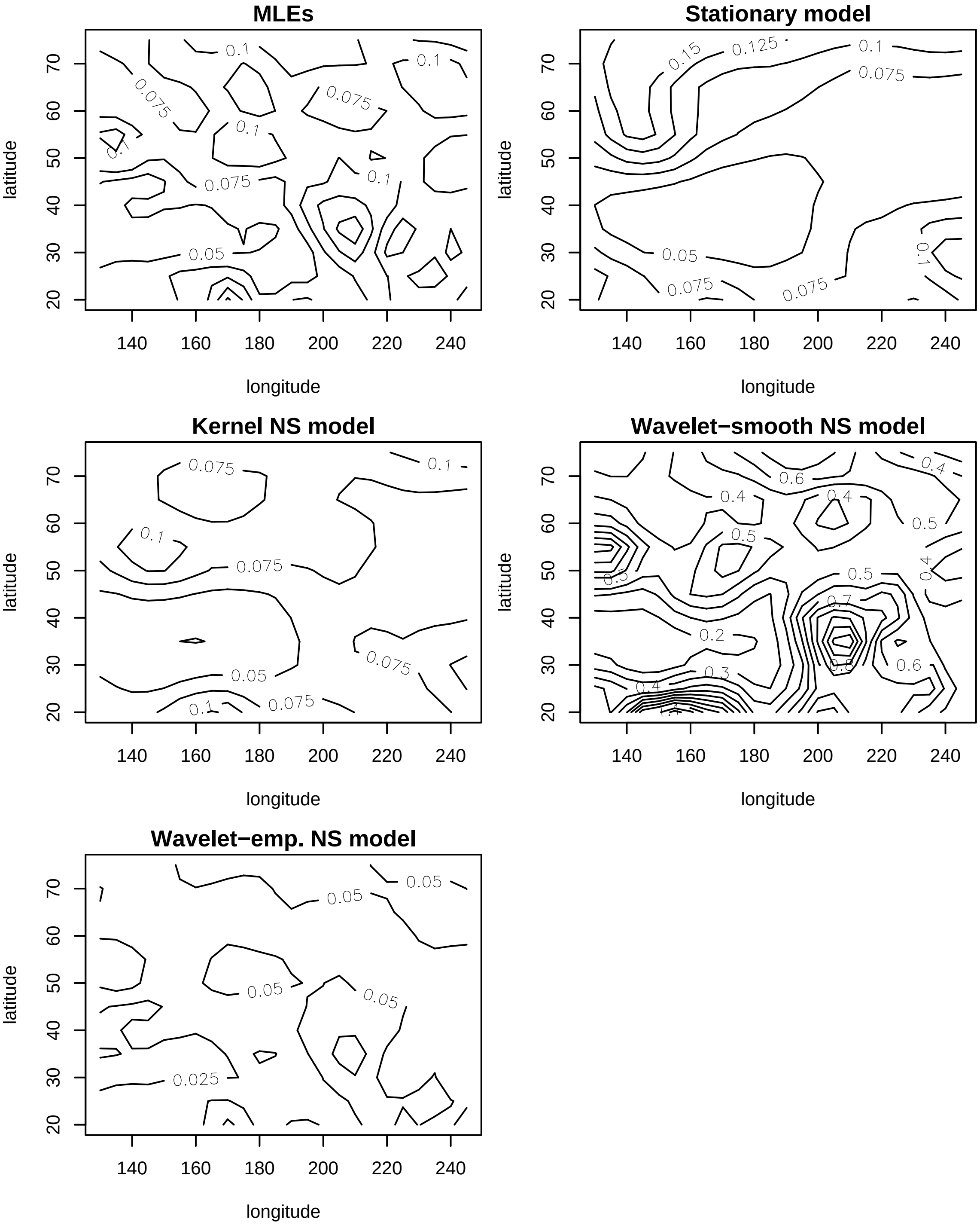

Figure 5.8. Maps of estimated η^2 values for temperature variance for (a) MLE model, and posterior means from (b) stationary model, (c) kernel-based nonstationary model, (d) wavelet-smoothed covariance model, and (e) wavelet-empirical covariance model.

are occurring at locations where the correlation structure is not well modelled by the stationary model; the variance increases at these locations to compensate. We can see this by calculating the standardized residuals, $\boldsymbol{Y}_t^* = L_{\boldsymbol{Y}}(\hat{\boldsymbol{\eta}}^2)^{-1}(\boldsymbol{Y}_t - \tilde{\boldsymbol{\alpha}} - \tilde{\boldsymbol{\beta}}t)$ where $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are the posterior mean estimates and where the Cholesky factor $L_{\boldsymbol{Y}}(\hat{\boldsymbol{\eta}}^2)$ of the residual covariance matrix is calculated based on the posterior means for $\kappa_{\boldsymbol{Y}}, \nu$, and δ, but using the MLEs, $\hat{\boldsymbol{\eta}}^2$ rather than the posterior mean estimates. If the covariance structure fits the data, then $\boldsymbol{Y}_t$ should be uncorrelated white noise with standard deviation of one. I calculate the average squared values of these residuals, averaging over the training years, and plot this by location (Figure 5.10). The areas of lack of fit, where there are large, correlated squared residual values, coincide with the areas in which the ratios of the posterior mean variances to the MLE variances,

$$\frac{\widetilde{\eta_i}^2}{\widehat{\eta_i}^2},$$

are larger than one, generally falling in the corners of the plot, particularly the upper left corner.

The wavelet models do not match the MLE intercepts as closely as the stationary and kernel nonstationary models and smooth the slopes to a much higher degree (Figure 5.9). For the wavelet model with the smoothed correlation matrix, the residual variances are very large relative to the MLEs, suggesting that the correlation model is fitting the correlation structure of the data very poorly, and the variances must increase to compensate. I am not surprised that these variance estimates are larger than the MLEs since the correlation matrix is fit without reference to the likelihood, so the variance estimates and/or the value of δ must compensate for any lack of fit of the correlation matrix to the data. However, I am surprised that the variance estimates are so high, given that for the stationary model, the posterior mean residual variances are not more than about twice as large as the MLEs. The estimate of δ for the wavelet smooth model is three orders of magnitude larger than for the other models, again suggesting compensation for lack of fit of some sort in the model. For the wavelet model that mimics the empirical correlation matrix, the variances are a linear function of the MLEs and are lower than the MLEs. It appears that this correlation model fits the training data so well (as shown by the large log-likelihood) that the residual variance estimates are lower than the MLEs even though the intercept estimates appear quite different than their MLEs, which one would expect to detract from the model fit and drive the variances up. In any event, it is clear that the wavelet models are fitting the data in rather strange ways. Given that there

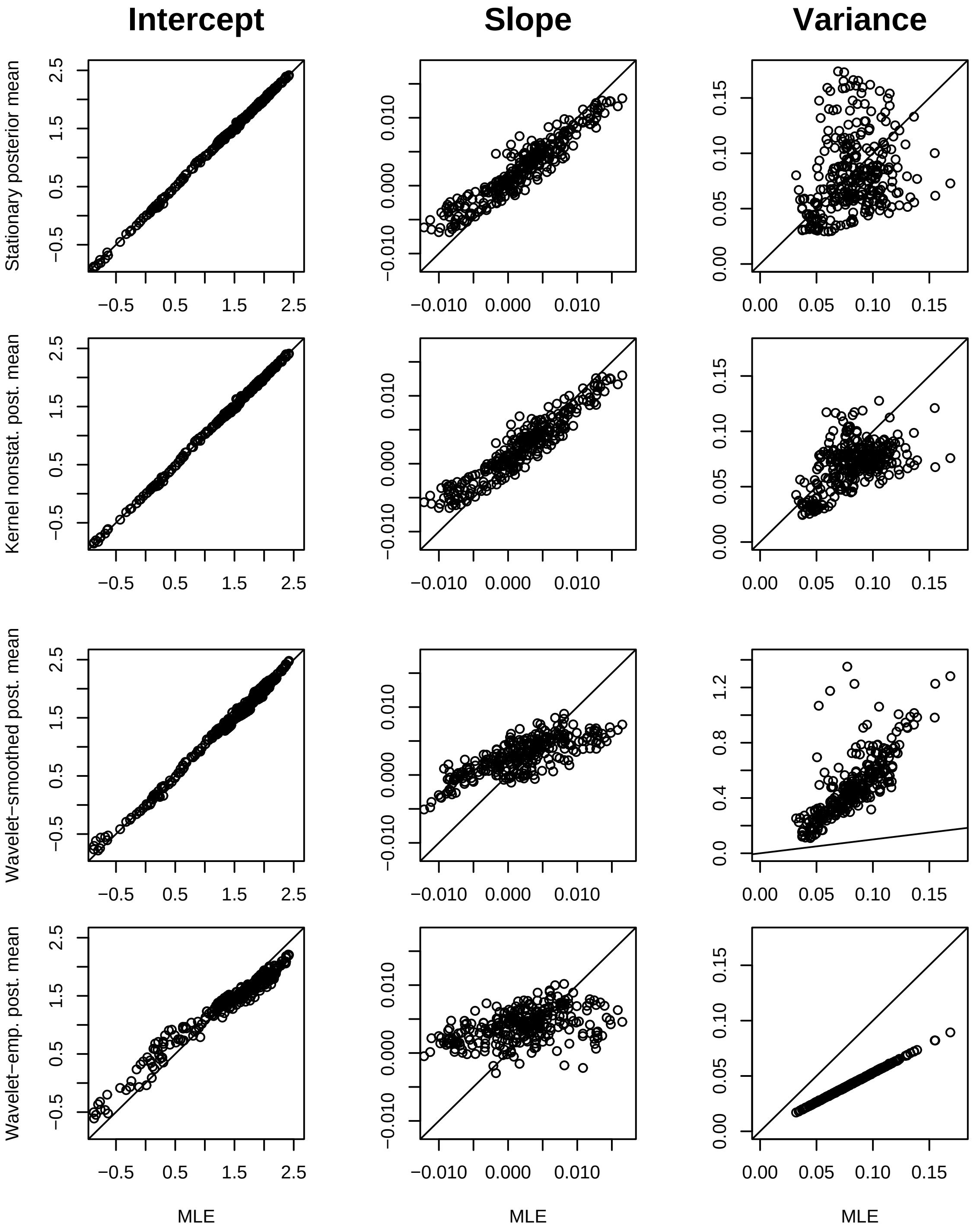

Figure 5.9. Scatterplots of model estimates (posterior means) of intercept (column 1), slope (column 2), and residual variance (column 3) fields compared to the MLE values for the four models: stationary (row 1), kernel nonstationary (row 2), wavelet-smoothed (row 3) and wavelet-empirical (row 4) for temperature variance.

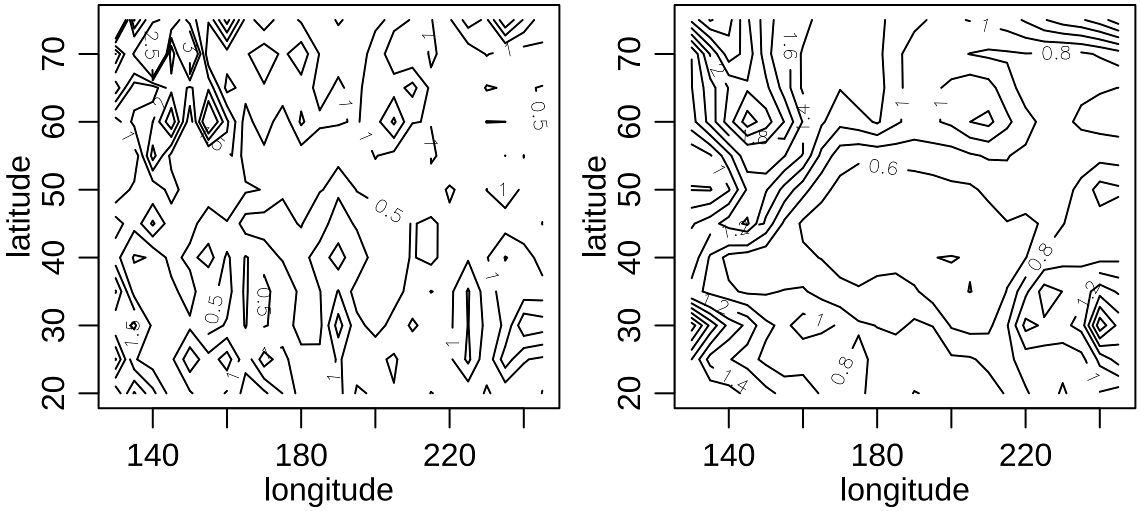

Figure 5.10. (a) Plot of standardized residuals (defined in text) as a function of location for temperature variance; these residuals are calculated based on the posterior mean parameters, but using the MLEs for the residual variances. (b) Plot of the ratio of the posterior mean residual variance estimates to the MLE variance estimates.

is no clear theory behind the thresholding, that the predictive performance of the wavelet models is poor (Section 5.7.3), and that the smoothing being done is not intuitive, I would be reluctant to use the wavelet models in the fashion employed in this work without further development.

In Figure 5.11, we see that for Eady growth rate, the slopes are smoothed much more toward zero in both the stationary and nonstationary models than was the case for temperature variance (Figure 5.9). Given that the predictive calculations suggest that the trends are not strong in the Eady growth rate dataset (Section 5.7.3), this level of smoothing is not surprising, and suggests that the model is correctly adjusting to the information in the data about the certainty in the trend estimates. For the variance estimates, the picture is similar to that for temperature variance; the nonstationary model estimates are somewhat similar to the MLEs, while the stationary model has some estimates that are much larger than the MLEs, suggesting that once again lack of fit in the correlation model is forcing the residual variance to rise to compensate.

The nonstationary model does appear to estimate nonstationary correlation structure, based on the posterior mean basis kernels shown in Figure 5.12 for temperature variance in the Pacific and in Eady growth rate in the Atlantic. We see that the basis kernels vary in size and orientation. However, the degree of nonstationarity based on the correlation structure induced by the basis kernel model is much less striking. In the following figures, I estimate the correlations between each of nine focal locations (the nine locations at which the basis kernels are centered) and all 288 other locations. For the two models in which the correlations are fit during the MCMC, I do this by calculating the induced correlations between locations at each MCMC iterate and averaging over the values. I compare these posterior mean correlations to the empirical correlations in the data by comparing correlation maps. In Figure 5.13, which shows the correlation structure for the kernel nonstationary model for temperature variance, we see that the correlation structure is somewhat nonstationary, primarily at 30° N, but that much of the variability in the basis kernels (Figure 5.12) has been smoothed out by taking the kernels to be spatial averages of the basis kernels. The correlation structure for the kernel nonstationary model for Eady growth rate is somewhat more nonstationary than for temperature variance (Figure 5.14), although the variability in the basis kernels is still greatly smoothed. Comparing the nonstationary correlation structure and the stationary correlation structure (Figures 5.15 and 5.16, for temperature variance and Eady growth,

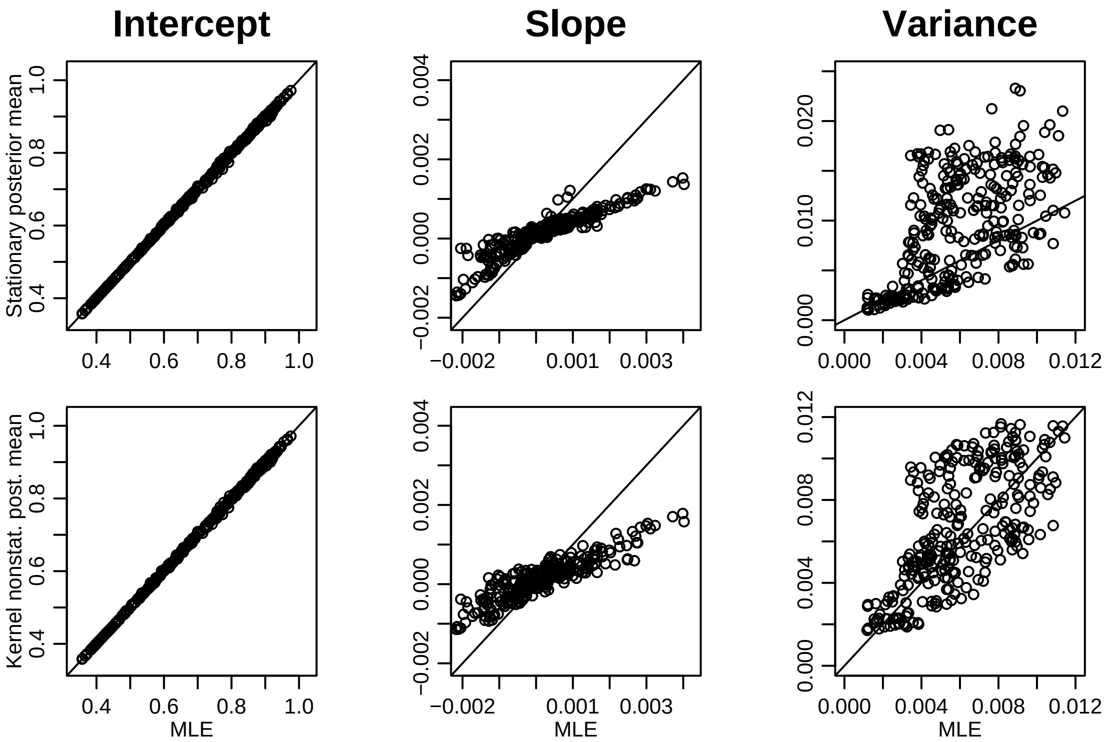

Figure 5.11. Scatterplots of model estimates (posterior means) of intercept (column 1), slope (column 2), and residual variance (column 3) fields compared to the MLE values for the models for Eady growth rate: stationary (row 1) and kernel nonstationary (row 2).

respectively) to the empirical correlations (Figures 5.17 and 5.18) we see that the nonstationary model is reflecting some of the gross patterns of the empirical correlation structure, but not all of the gross structure that one might expect the model to capture. This may be because the empirical correlations are noisy estimates that the model is correctly smoothing out, or because the basis kernel approach is not flexible enough to capture the true structure. Use of more kernels or a different approach to the spatial averaging that does less smoothing of the basis kernels might result in more nonstationarity being modelled. However, note that the model is free to choose the parameter that controls the degree of spatial smoothness. In contrast to the stationary and kernel nonstationary models, the wavelet correlations are fixed by the thresholding chosen. In Figure 5.19 we see that the wavelet-empirical model most closely matches the empirical correlation structure. Yet, this model drastically overfits (Section 5.7.3), suggesting that the structure is not matching the correlation structure of the test data. In Figure 5.20 we see that the wavelet-smooth method is is in fact smoothing out much of the empirical structure, while retaining much of the gross structure that is not retained in the nonstationary model. The poor predictive performance of the wavelet-smooth model could be occurring because it overfits the gross correlation structure, retaining structure that should be smoothed out. Alternately, the model may not correctly capture the intricate high-dimensional structure of the test data, perhaps because the test data do not satisfy certain constraints specified by the wavelet-smooth correlation matrix. One indication that this may be the case is that the correlation scale of the wavelet-smooth structure tends be longer than that of the stationary and kernel nonstationary models. Longer correlation scales tend to produce linear combination constraints amongst the locations. By using shorter scales the stationary and nonstationary correlation structures may avoid imposing such constraints on the test data (assuming that in the fitting process similar constraints are discouraged by the training data). This suggests that it is difficult to closely model the correlation structure and expect it to match the structure of test data, and also that modelling the structure may tend to give lower correlations than truly exist, if by doing so the model can avoid imposing constraints that are not satisfied by the data. The general lesson is that modelling the correlation structure of high dimensional data in a joint fashion is very difficult.

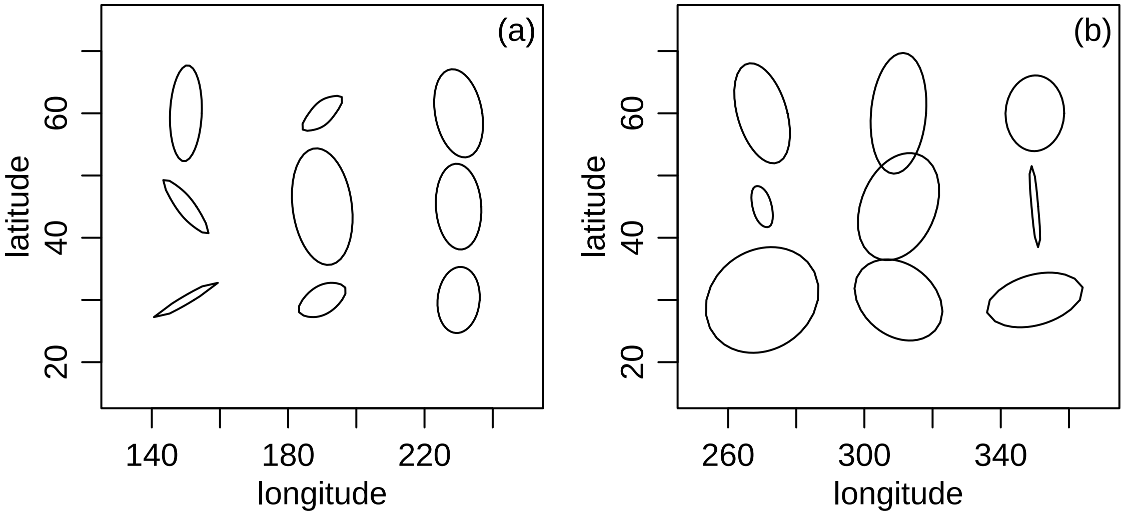

Figure 5.12. Mean basis kernels for (a) temperature variance in the Pacific region and (b) Eady growth rate in the Atlantic. For each posterior sample, I represent the kernel as a constant density ellipse from a normal density with covariance matrix equal to the basis kernel matrix. Mean ellipses are then plotted using the average distances from the ellipse origin to the ellipse itself at 44 different angles, averaged over the posterior samples. Note that since the kernels are plotted on a latitude-longitude grid, distances toward the top of the plot are exaggerated and the true basis kernels there are smaller in size than represented here.

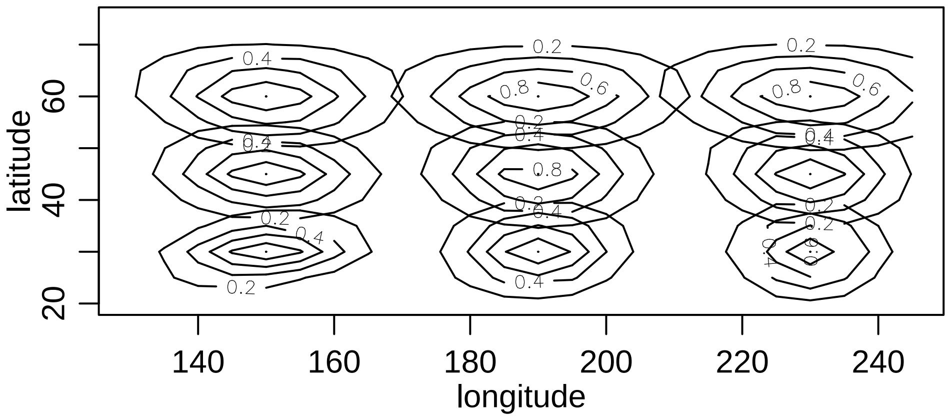

Figure 5.13. Plot of posterior mean correlation structure from the kernel nonstationary model for temperature variance between each of nine focal locations and all 288 locations. Correlation structures at the nine focal locations are overlaid on the same plot because correlations are less than 0.20 except in the bullseye areas. The nine focal locations are at the centers of the bullseyes and are the same locations as the centers of the basis kernels, as listed in the text.

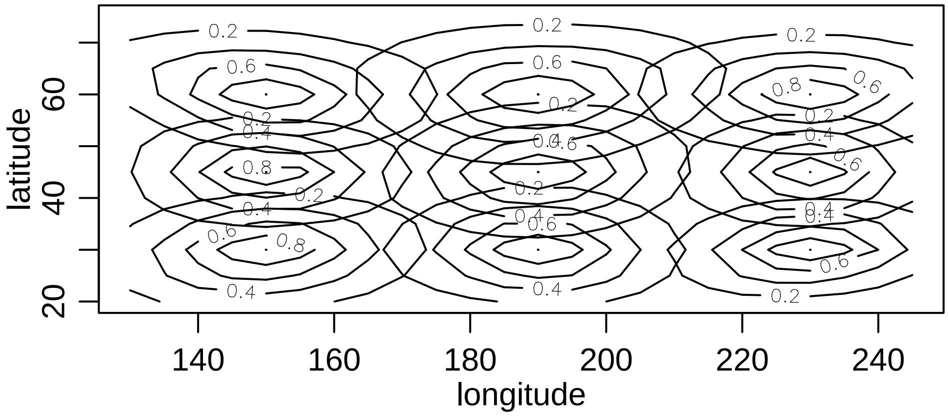

Figure 5.14. Plot of posterior mean correlation structure from the kernel nonstationary model for Eady growth between each of nine focal locations and all 288 locations. Plots are overlaid because correlations are less than 0.20 except in the bullseye areas. The nine focal locations are at the centers of the bullseyes and are the same locations as the centers of the basis kernels, as listed in the text.

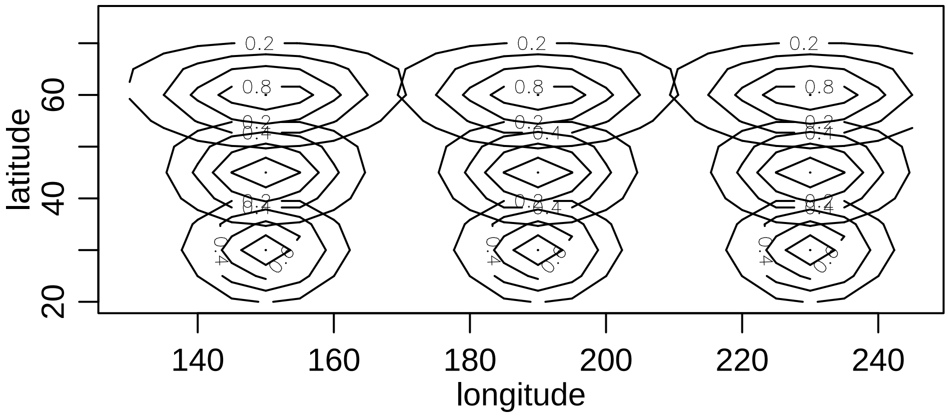

Figure 5.15. Plot of posterior mean correlation structure from the stationary model for temperature variance between each of nine focal locations and all 288 locations. Correlation structure appears different at different latitudes because of the distortion induced by the latitude-longitude grid. Other details are as in Figure 5.13.

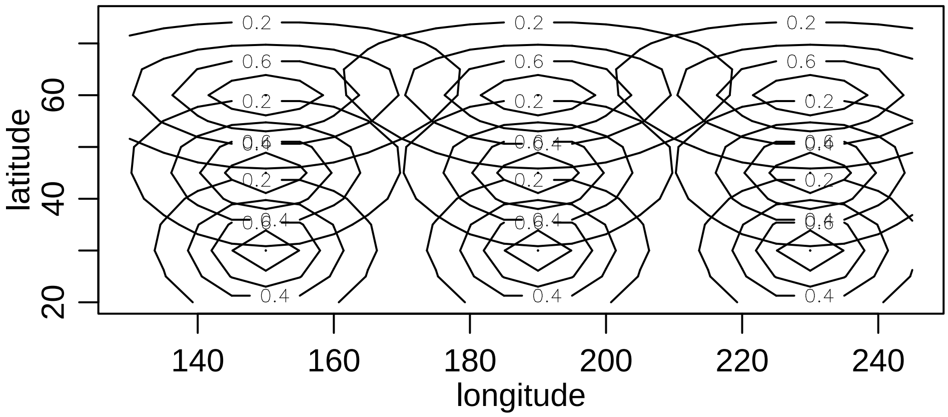

Figure 5.16. Plot of posterior mean correlation structure from the stationary model for Eady growth between each of nine focal locations and all 288 locations. Correlation structure appears different at different latitudes because of the distortion induced by the latitude-longitude grid. Other details are as in Figure 5.14.

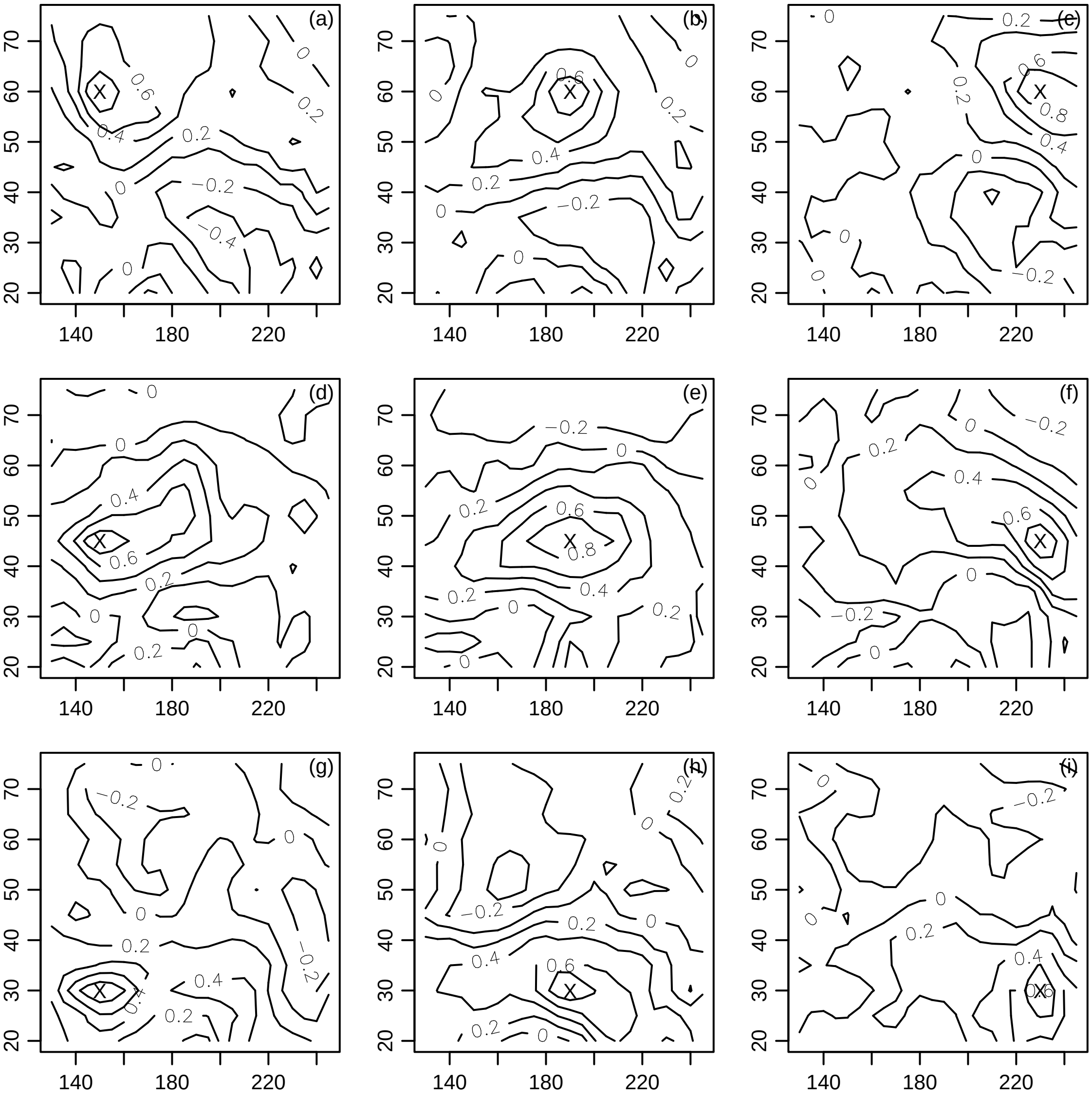

Figure 5.17. Plots of empirical correlations for temperature variance between each of the nine focal locations and all 288 locations. Each subplot displays the correlation structure for one focal location (marked by 'X') with latitude and longitude increasing from bottom to top and left to right respectively: (a) 150° *E,* 60° *N, (b)* 190° *E,* 60° *N, (c)* 230° *E,* 60° *N, (d)* 150° *E,* 45° *N, (e)* 190° *E,* 45° *N, (f)* 230° *E,* 45° *N, (g)* 150° *E,* 30° *N, (h)* 190° *E,* 30° *N, (i)* 230° *E,* 30° *N.*

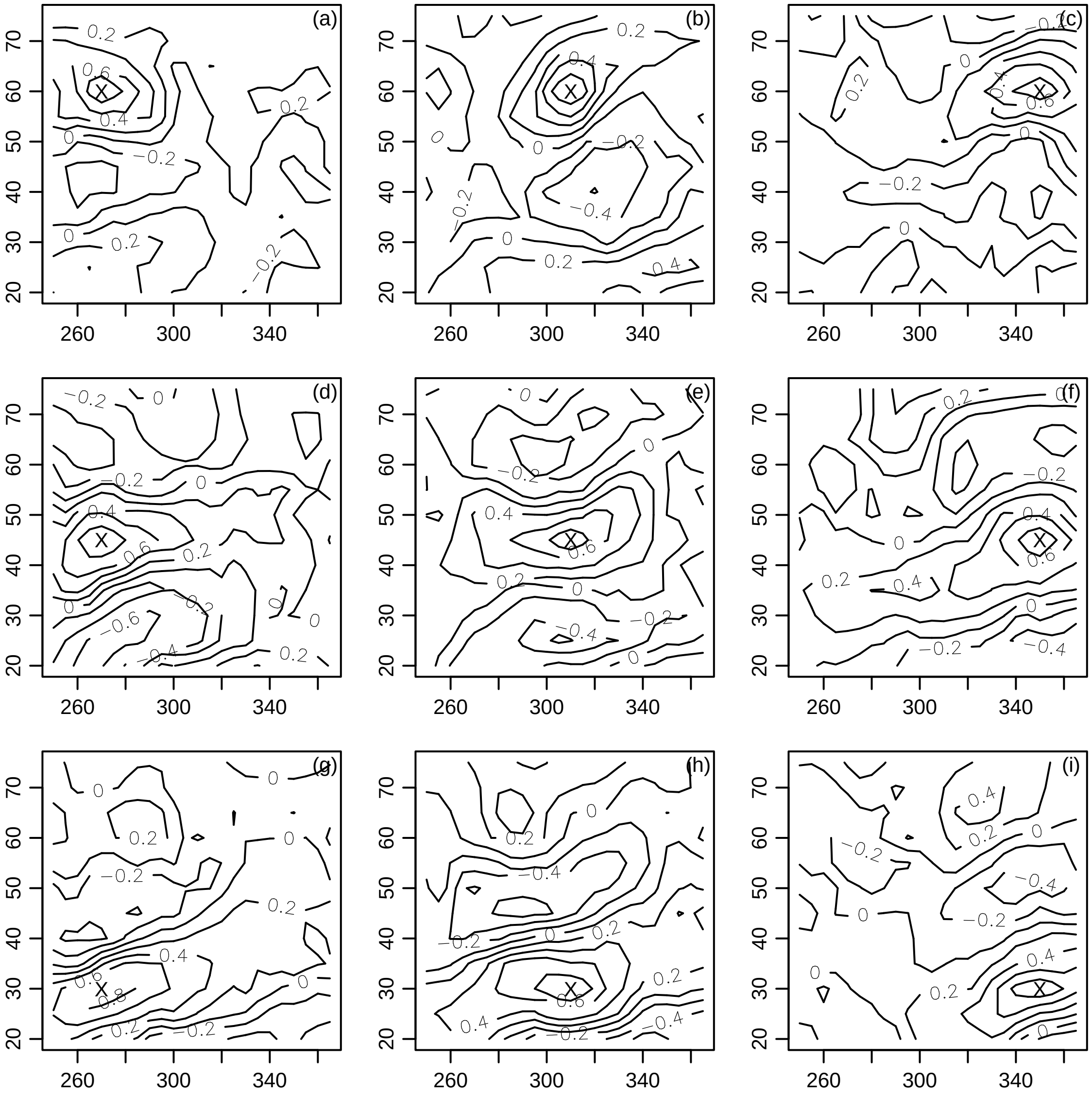

Figure 5.18. Plots of empirical correlations for Eady growth between each of the nine focal locations and all 288 locations. Each subplot displays the correlation structure for one focal location (marked by 'X') with latitude and longitude increasing from bottom to top and left to right respectively: (a) 150° *E,* 60° *N, (b)* 190° *E,* 60° *N, (c)* 230° *E,* 60° *N, (d)* 150° *E,* 45° *N, (e)* 190° *E,* 45° *N, (f)* 230° *E,* 45° *N, (g)* 150° *E,* 30° *N, (h)* 190° *E,* 30° *N, (i)* 230° *E,* 30° *N.*

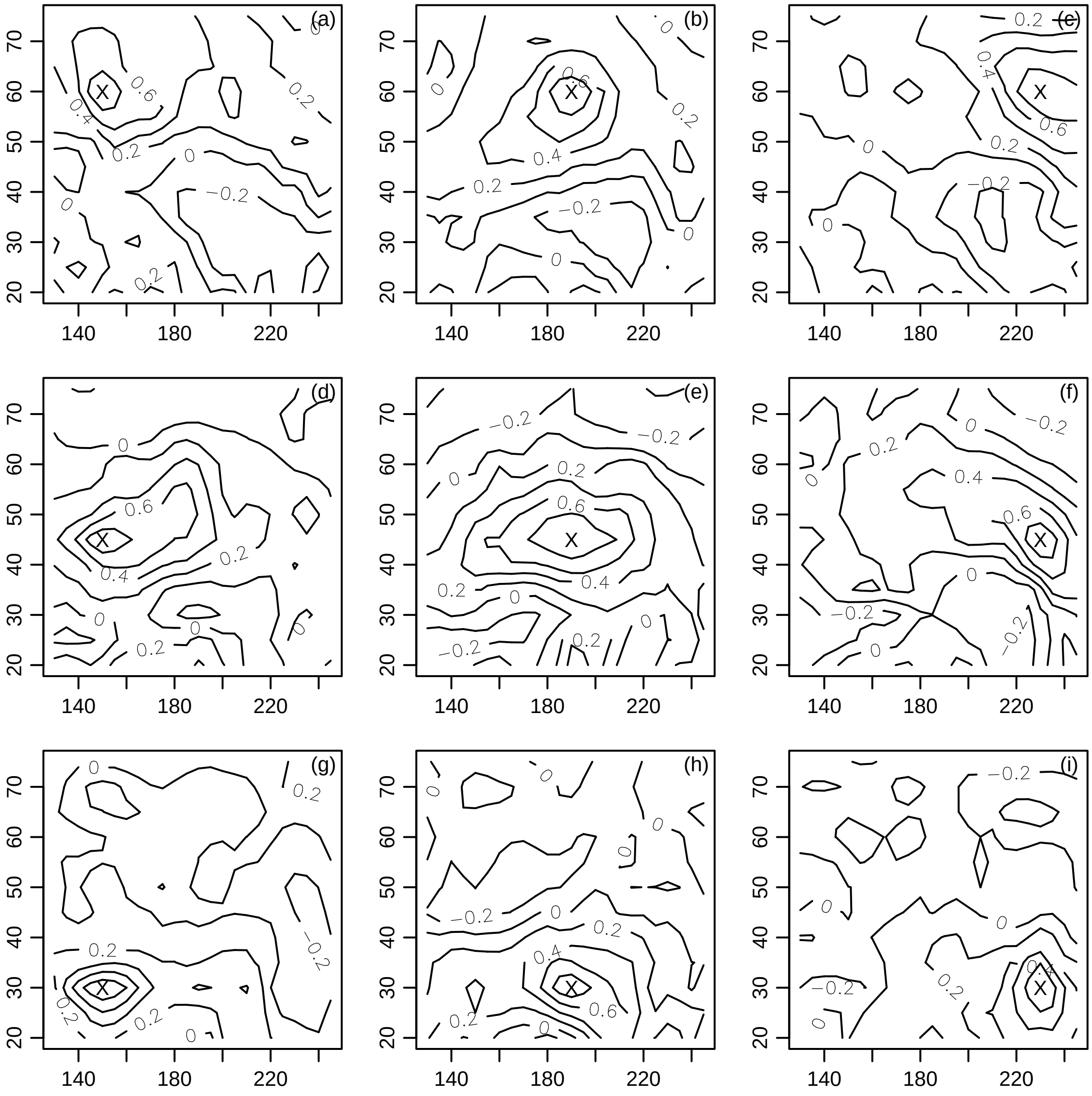

Figure 5.19. Plots of wavelet-empirical model correlations between each of nine focal locations and all 288 locations for temperature variance. Details are as in Figure 5.17.

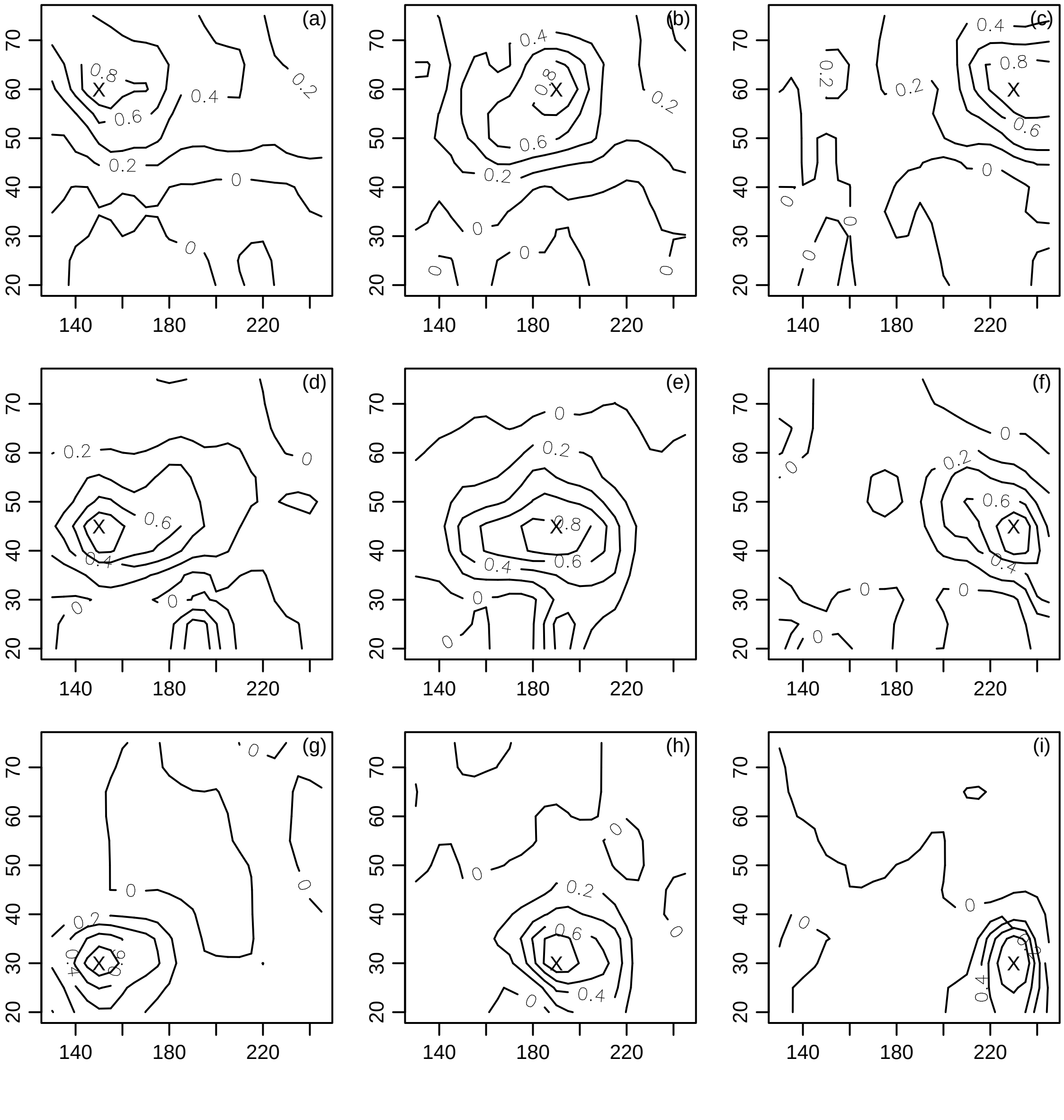

Figure 5.20. Plots of wavelet-smooth model correlations between each of nine focal locations and all 288 locations for temperature variance. Details are as in Figure 5.17.

5.7.3 Model comparison

I first assess the log posterior predictive density. In Table 5.1 I show the posterior predictive density values for test data for both datasets for the the six models. The kernel nonstationary model has the highest density. For the Bayesian models, I assess the uncertainty in these estimates by calculating the log predictive density from 10 subsets of 5000 contiguous MCMC samples and assessing the variability in the blocked estimates. This approach makes it clear that the stationary and kernel nonstationary methods give by far the highest log predictive densities and that the kernel nonstationary model is clearly better than the stationary model. Comparing the predictive density between the stationary and kernel nonstationary models by year, the kernel nonstationary is better in each of the seven test years (not shown), although the relative difference varies somewhat by year. Clearly when we assess performance in a way that takes account of the covariance structure of the data, the kernel nonstationary model is the best model. However, is this improvement substantial? If the observations were independent, we could consider the improvement in the predictive density on a per observation basis, by considering the difference in log predictive densities from the two models divided by the number of observations. This works for the time variable, since I consider the seven years to be independent, and the improvement in the log predictive density is 17.6 and 20.6 per year for temperature variance and Eady growth rate, respectively. If one is interested in predicting a whole year of observations, then one may be interested in the likelihood ratio for all locations which would be $\exp(17.6)$ or $\exp(20.6)$, suggesting great improvement based on the nonstationary model relative to the stationary model. However, if one is thinking about individual locations or wants to consider the improvement in predictive density per observation, the picture is clouded. It is not clear how many effective observations we have each year because of the correlation between locations. If all the locations were independent, the improvement would be only 0.061 or 0.071 per observation, corresponding to a likelihood ratio of 1.06 or 1.07. However, if there is substantial correlation between observations, the likelihood ratio is more compelling. Given the empirical correlations in the data, having the number of effective observations be one-fifth or even one-tenth of the number of nominal observations may be reasonable. Further investigation of this issue would involve estimating the effective number of locations in an appropriate manner, but I have not pursued this issue.

Table 5.1. Log predictive density comparisons of the six models on test data for the two datasets. Values in parentheses are ranges based on blocked values from the posterior simulations.

Model	Eady growth rate - Atlantic	Temperature variance - Pacific
MLE	3832	1267
MLE,$\boldsymbol{\beta} \equiv 0$	3886	1393
stationary	6393(6375-6395)	4439 (4427-4441)
wavelet-empirical	not modelled	~ -100000
wavelet-smoothed	not modelled	2527 (2473-2529)
kernel-nonstationary	6537 (6524-6539)	4562 (4550-4565)

The MLE models perform poorly because I assume independence between locations, which is a poor model for the joint structure of the data. The wavelet-empirical model has a terrible predictive density, presumably because of overfitting. The wavelet-smooth model is not as unreasonable, but is much worse than the stationary or kernel nonstationary models. Note that based on the strange behavior of the wavelet models seen in the previous section, and the poor predictive performance on the temperature variance index seen here, I did not fit the wavelet models to the Eady growth rate data.

The results change somewhat when we look only at point predictions, as shown by the MSE values in Table 5.2. In the temperature variance case, the wavelet models are clearly not performing as well, with the wavelet-empirical model appearing to overfit drastically and the smoothing imposed by the wavelet-smoothed model decreasing predictive performance as well. The MLE models, stationary model, and kernel nonstationary model all give similar MSE values, with relative differences of only a few percent. For temperature variance, the differences do appear to be somewhat robust with respect to simulation error, as seen from the range of MSE values in blocks of the MCMC iterations, as well as when assessing MSE by year. Comparing the MSE values by year for the temperature variance case, the stationary model is worse than the kernel nonstationary model in all years while the kernel-based nonstationary model is better than the MLE model in one year, 1954. For Eady growth, the differences do not appear to be robust with respect to simulation error, as seen from the range of MSE values in blocks of the MCMC iterations, although the nonstationary model does slightly outperform the stationary model in all seven years. The nonstationary model outperforms the MLE model with slopes fixed at zero in one of seven years, while it outperforms the unrestricted model in only two of seven years, but in those two years (1949 and 1999, the extremal years), it sufficiently outperforms the MLE model so that averaged across all seven years, it has slightly lower MSE.

Given the relatively small differences in MSE, it is unwise to overinterpret the results. However, it does seem that the nonstationary model is performing slightly better than the stationary model. Also, in the case of temperature variance, the MLEs seem to perform better than the nonstationary model, suggesting that given the smoothness of the original data, there is little to be gained by borrowing strength across locations. In the Eady growth rate case, the slopes appear not to help with prediction, suggesting that for most locations, there is little real trend over time, although it is still possible that for a subset of locations, there are real trends that are masked by assessing only the effect of forcing all the slopes to be zero simultaneously. Unfortunately the full models do not appear to improve prediction by compromising between the MLEs and the joint null hypothesis. Note that the model comparison results might change dramatically with less smooth data. Also, fitting the full model does allow us to assess joint uncertainty in our estimates, as discussed next.

Table 5.2. MSE comparisons of the six models on test data for the two datasets. Values in parentheses are ranges based on blocked values from the posterior simulation.

Model	Eady growth rate - Atlantic	Temperature variance - Pacific
MLE	0.00796	0.0879
MLE,$\boldsymbol{\beta} \equiv \mathbf{0}$	0.00765	0.0904
stationary	0.00798 (0.00793-0.00805)	0.0901 (0.0891-0.0910)
wavelet-empirical	not modelled	0.155 (0.140-0.177)
wavelet-smoothed	not modelled	0.101 (0.0964-0.105)
kernel-nonstationary	0.00793 (0.00788-0.00798)	0.0887 (0.0882-0.0894)

5.7.4 Trend significance

Here I compare the trend estimates from the Bayesian nonstationary model with the MLEs. First, let's consider the temperature variance model. For the Bayesian model I use the posterior means and take the posterior standard deviation as a standard error, while for the MLEs, I estimate the standard error as $\frac{\widehat{\eta_i}^2}{\sum t^2}$. In Figure 5.21 I plot the standard errors as function of the point estimates for both models. The Bayes point estimates have been shrunk somewhat toward zero, but the most obvious difference is that the standard error estimates are on average about two-thirds as much for the Bayesian model as for the MLEs. It appears that accounting for the spatial correlation has made us more certain about the trend estimates. Using a conventional significance threshold of plus or minus two standard errors, the MLEs suggest that 103 locations have significant trends while the Bayesian model suggests 146 locations. Such an analysis does not account for simultaneously conducting 288 tests. In Paciorek et al. (2002) we used the False Discovery Rate approach to assess joint significance. Using the standard FDR method for independent data, for which Ventura et al. (2003) report successful results with spatially correlated storm activity data, 49 of the 288 locations are significant. To see the effect of the Bayesian shrinkage, I calculated the p-values and FDR result that would apply if one used the point estimates and standard errors from the Bayesian model, and found that I rejected 117 locations. Of course there is no theoretical justification for this, but it gives a sense for the impact of the changes in the point estimates and their standard errors in a multiple testing context. The Bayesian shrinkage has substantially increased our certainty in the estimates of the slopes. However, given that the Bayesian model produced MSE values that are slightly worse than those based on the MLEs (Table 5.2), it is not clear that such certainty is warranted.

Next let's consider the Eady growth rate data; the shrinkage results are very different here. In Figure 5.22 I plot the standard errors as function of the point estimates for both models. Once again, the estimates have been shrunk toward zero and the standard error estimates are much smaller, suggesting that accounting for the spatial correlation has made us more certain about the trend estimates. The linear trends in the Eady growth rate data are much less pronounced than in the temperature variance case. Using a conventional significance threshold of plus or minus two standard errors, the MLEs suggest that 49 locations have significant trends while the Bayesian model

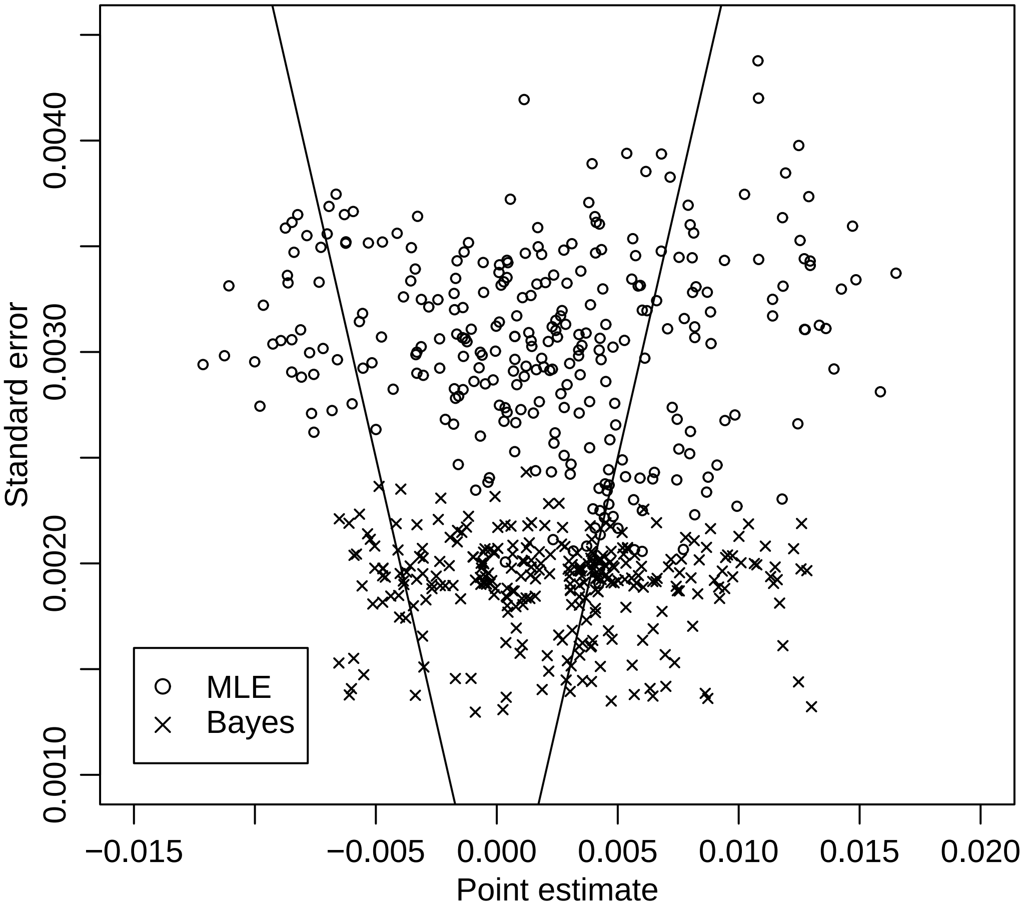

Figure 5.21. Scatterplot of standard error estimates as a function of the point estimates for the linear trends in both the MLE and Bayesian nonstationary models for the 288 locations of temperature variance in the Pacific. Points in the areas toward the outer sides of the plot relative to the nearly vertical lines are individually significant based on the point estimates being at least two standard errors away from zero.

suggests 12 locations. In contrast to the temperature variance data, the Bayesian model suggests many fewer locations are significant than based on the MLEs. Using FDR, 19 of the 288 locations are simultaneously significant. Once again, to see the effect of the Bayesian shrinkage, I calculated the p-values and FDR result that would apply if one used the point estimates and standard errors from the Bayesian model, and found that I rejected no locations.

One might also take a Bayesian approach to multiple testing based on the posterior sample. The first difficulty lies in defining H_0. One might define it in an ad hoc way as $\mathrm{H}_{0,i} : \beta_i\tilde{\beta}_i < 0$, namely that the true slope is of the opposite sign from the posterior mean. Of course this hypothesis depends on the data, which is anathema to a frequentist, but poses no real problems to the Bayesian. Another concern is that this is essentially a one-sided test in which the side is determined based on the data, so this approach will result in more significant locations than a classical approach because the level of the test is essentially twice that in the classical test. However, if one proceeds in this fashion, one might choose to reject the null for all locations at which $P(\mathrm{H}_0|\boldsymbol{Y}) < 0.05$. Such an approach satisfies a Bayesian version of the FDR criterion because

$$\mathrm{E}(\text{FDP}|\boldsymbol{Y}) \leq 0.05, \tag{5.7}$$

where FDP is the proportion of false rejections (rejections for which the null is actually true). This is just the classical FDR, except for the conditioning on $\boldsymbol{Y}$. In fact, so long as the average value of $P(\mathrm{H}_{0,i}|\boldsymbol{Y})$ taken over the rejected locations is less than 0.05, the expectation property (5.7) will hold:

$$\begin{aligned}
\mathrm{E}(\text{FDP}|\boldsymbol{Y}) &= \mathrm{E}\frac{n_{fr}}{n_r} \\
&= \mathrm{E}\frac{\sum_i I(\mathrm{H}_{0,i}|\boldsymbol{Y})}{n_r} \\
&= \frac{\sum_i \mathrm{E}I(\mathrm{H}_{0,i}|\boldsymbol{Y})}{n_r} \\
&= \frac{\sum_i P(\mathrm{H}_{0,i}|\boldsymbol{Y})}{n_r},
\end{aligned}$$

where n_r is the number of hypotheses rejected and n_{fr} is the number rejected falsely, which is a random variable. We estimate $P(\mathrm{H}_{0,i}|\boldsymbol{Y})$ using the posterior samples, performing the calculation location by location; the effect of the other locations is already incorporated through the fitting

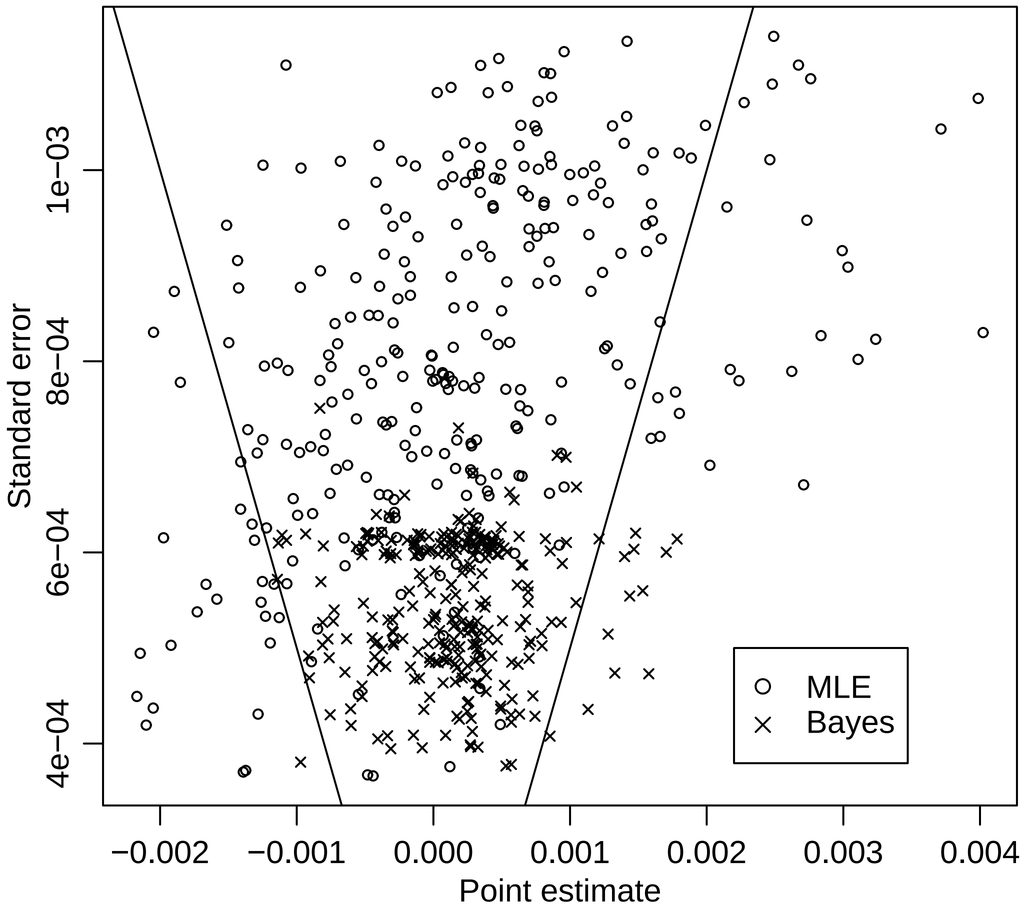

Figure 5.22. Scatterplot of standard error estimates as a function of the point estimates for the linear trends in both the MLE and Bayesian nonstationary models for the 288 locations of Eady growth rate in the Atlantic. Points in the areas toward the outer sides of the plot relative to the nearly vertical lines are individually significant based on the point estimates being at least two standard errors away from zero.

process. The result using this Bayesian version of FDR is that many more locations are rejected than in the frequentist FDR approach. For temperature variance, if we reject only those locations with posterior probability of the null less than 0.05, 170 are rejected (146 if we use 0.025 to account for the use of a ’one-sided’ test). If one is willing to reject locations with posterior probability of the null exceeding 0.05, so that the Bayesian FDR is very close to 0.05, one rejects 241 of the 288 locations (based on rejecting the locations with the smallest posterior probabilities). Note that these are more than are rejected using the Bayesian estimates to get a ‘p-value’ and then using the classical FDR, which suggests that not only are the smaller standard errors having an effect, but also that the use of the Bayesian approach to FDR plays a role. For Eady growth rate, the Bayesian approach accounts for the apparent lack of robust trends; 17 locations are rejected with posterior probability of the null less than 0.05, while 12 are rejected if we use 0.025, and 33 if we reject such that the average probability of the null is 0.05. Thus the Bayesian approach suggests that for temperature variance most of the locations have slopes that can be distinguished from zero, but that the point estimates of those slopes are closer to zero than the MLEs. For Eady growth rate, the Bayesian approach suggests that few locations have real trends. The optimism on the part of the Bayesian approach for the temperature variance data is rather troubling given that the cross-validation results for MSE in Section 5.7.3 give little indication that the trends are so certain. However, it may also be the case that the FDR approach is overly conservative. It is illustrative to consider an extreme example of positive dependence amongst the test statistics. Suppose the test statistics are perfectly correlated. In that case, we are really only doing one test and would want to reject if $p < 0.05$. However, the standard FDR approach will only reject if $p < \frac{.05}{n}$, a much harder threshold to reach. Also, the Bayesian approach does not appear to be overly optimistic in the case of Eady growth rate.

We might also take the approach of trying to determine locations at which the slopes can be determined with high certainty to not be near zero. This is motivated by the fact that there might be small trends that with enough data we could be certain were not zero, but are not substantively of interest. For example, in conjunction with the subject matter experts, we might decide that trends less than some value ϵ are not substantively of interest. One would then determine $P(|\beta_i| > \epsilon)$ and apply the Bayesian FDR approach outlined above to determine a set of locations for which the FDR for this new type of null hypothesis is less than, say, 0.05. Of course, this requires choosing a value for ϵ. A variation on the theme is to plot the number of rejected hypotheses as a function of different values of ϵ. I have not pursued these possibilities here as in-depth data analysis is somewhat outside the scope of this work. Also the fact that the temperature variance index modelling is done on the log scale makes it somewhat difficult to interpret the substantive importance of different values of ϵ.

5.8 Discussion

5.8.1 Covariance modelling assessment

Accounting for spatial covariance in a rigorous manner when fitting trends at multiple locations is a difficult task. Even after reducing the number of locations from 3024 to 288, fitting the full spatial model was very computationally intensive, and the MCMC exhibited very slow mixing. This was the case for the stationary covariance model and the models in which the wavelet-based correlations were fixed, as well as the full kernel-based nonstationary model.

Based on cross-validation, the kernel-based nonstationary model seems to offer some improvement over the stationary model. The nonstationary model reflects some of the local correlation structure, although it is unable to model irregular, long-distance, and negative correlations. In both datasets, the MLEs outperform the stationary model estimates in terms of point estimates for test data, which suggests that unless one is willing to use a nonstationary model, one may not want to model the correlation structure. Of course, if one wants to predict off-grid, some sort of model is necessary, although for point predictions, one might just consider smoothing the MLE fields. For one dataset the MLEs outperform the nonstationary model in terms of MSE, while in the other the nonstationary model performs slightly better. For less smooth data, there may be much greater benefit to fitting the full model rather than relying on the MLEs. With respect to the posterior predictive density, which assesses how well the models jointly predict test data assuming the normal likelihood is reasonable, the nonstationary model is clearly better than the other models. However, on a per-location basis, the gain relative to the stationary model is small.

A less smooth dataset may allow a better evaluation of the various covariance models and better let us determine how critical is the choice of covariance structure in smoothing noisy data. One potential dataset is the storm activity index based on extreme wind speeds (Paciorek et al. 2002), which is much less smooth than the storm activity indices used here. Another possibility is the storm count index in Paciorek et al. (2002). However, these latter index values are non-normal count data and modelling them would require additional methodological development (discussed in Section 6.3). Climatological fields can range from the very smooth, such as pressure fields, to the very noisy, such as precipitation and wind fields. It seems plausible that the advantages of modelling the fields and thereby smoothing the observations will be largest for the noisy fields.

There are several important limitations built into the nonstationary model constructed here. The model cannot account for long-distance and negative correlations, and the use of simple Gaussian kernels limits the correlation structures that can be modelled. These limitations may be important in many datasets, including possibly the storm activity data, as suggested by the long-distance and negative correlations in Figure 5.1. I do not have any particular suggestions for modelling long-distance or highly irregular correlation structures, apart from the wavelet decomposition approach explored in this work, but some relatively simple approaches may model some of the negative correlation structure. One idea is to use kernels that are allowed to take negative values. However, even if such kernels could be parameterized sufficiently simply, the advantage of the closed form expression based on Gaussian kernels (2.5) would be lost. Another potential drawback is that the resulting correlation function is somewhat non-intuitive, because the product of two negative values is positive. An alternative is to use a product of the Matérn nonstationary correlation function used in this work and a stationary correlation function that can take negative values, such as those defined on the sphere (Das 2000) and in $\Re^3$ (Abrahamsen 1997), although this approach has the same drawback as described at the end of the next paragraph.

I have used stationary models that allow only non-negative correlations, but I could easily use the stationary correlation functions mentioned above that take negative values. It would be interesting to see if such correlation functions better fit the data than the Matérn stationary correlation function fit here. However, I suspect that the improvements would be minimal, because the negative correlations seen in Figure 5.1 occur in latitude bands. It does not seem likely that stationary correlation functions that require negative correlations to occur in equidistant rings around the focal location would fit such data well.

In the form that I have used here, the wavelet approach does not perform well. This work points out several areas in which the approach needs further development. First, it is unclear how much thresholding to do or exactly how to carry out the thresholding. I tried two levels of thresholding and neither performed well when incorporated into the full spatial model. One might argue that an intermediate level of thresholding would work much better, but how does one find that level? In particular, this highlights the need for a criterion by which to choose the level of smoothing. Nychka et al. (2001) report that with the thresholding they chose, they could mimic a stationary Matérn correlation with little error in terms of element by element differences between the true correlation and the approximation. However, this was done by smoothing the true correlation matrix, not by smoothing an empirical correlation matrix. In my experience applying the approach to empirical covariance matrices, it is very difficult to decide on the thresholding. Furthermore, even if one is able to choose a level of thresholding that appears satisfactory on an ad hoc basis, there are several reasons for concern. First, covariance matrices are very tricky to work with because of the constraints involved in being positive definite, and they involve very high-level structure when modelling many locations. Indeed, once the level of correlations is high enough, the values at some locations are for all practical purposes linear combinations of the values at other locations. It is entirely feasible that one could choose a smoothed covariance that by eye seems reasonable but in reality does not fit the data well at all when used in a likelihood-based context. This would occur if the data violate a linear combination constraint imposed in the covariance matrix. This lack of fit seems to be the case for the wavelet-thresholded covariance that smoothed the empirical covariance. This covariance gives a very low log-likelihood to the training data when embedded in the Bayesian model, presumably because the thresholding was done without reference to the likelihood, unlike the MCMC fitting of the stationary and kernel nonstationary models. Second, based on my experience here with the wavelet covariance that closely matched the empirical covariance, overfitting is a real concern. The log-likelihood of training data with the wavelet-empirical covariance was extremely high, but the spatial model with this covariance did a very poor job of predicting test data.

If one is interested merely in smoothing the empirical covariance for the purpose of display, ad hoc thresholding may be sufficient, but for model building and extrapolation to other locations, the exact covariance structure and the inherent high-level structure is critical. One potential avenue for development of the wavelet approach is to consider more rigorous criteria for optimizing the thresholding. Choosing the thresholding with reference to a likelihood or similar criterion is one possibility, possibly in conjunction with cross-validation, although this might limit the computational advantages of the approach. Other loss functions for estimating covariance matrices (Yang and Berger 1994) might be useful.

5.8.2 Other approaches

Given the difficulties encountered in the covariance modelling done in this work, other approaches may be more practical, albeit possibly more ad hoc. Holland et al. (2000) smooth estimated trends based on a kriging style approach, but with the 'data' covariance based on jackknifing the trend estimates. Such resampling procedures may be the most practical and computationally feasible approaches to data such as these. An alternative that makes use of the model development in this work would be to employ more empirical Bayes methodology to fix as many parameters as possible without unduly influencing the final inference. Hopefully this would get around some of the mixing issue encountered here as well as speeding the computations. Wikle et al. (1998) build hierarchical space-time models in which the analogues to my $\alpha(\cdot)$ and $\beta(\cdot)$ processes are given Markov random field priors, and spatial structure within the residual fields is modelled using nearest neighbor autoregressive models with spatial dependence on the field at the previous time point. Their approach is to build more complicated temporal models than are used here, while attempting to model the spatial structure in a relatively simple nearest neighbor fashion that avoids specifying a joint spatial covariance structure explicitly. This may be more computationally feasible than my approach, but it's not clear if the spatial structure is sufficiently flexible to capture the important structure in the data.

In problems with many locations, such as with the full storm activity dataset of 3024 locations, computationally-efficient methods are needed. While computationally feasible, smoothing the empirical covariance using the wavelet approach suffers from lacking a defined objective function for optimizing the degree and structure of the smoothing. Further work on this approach is needed.

A final area that I have not addressed in this work involves the linearity of the temporal model. Relaxing the assumption of linearity in the temporal model is an obvious choice for improving the model. In Paciorek et al. (2002) we present evidence of nonlinearity in all the indices for at least some locations. Two main issues arise in moving to nonlinear models. The first is how to parameterize and fit such models. The goal is to extract useful high-level information about long-term trends; deciding how to and the extent to which to partition between signal and noise is a difficult task. One possibility would be to represent the time series at each location by regressing on an orthogonal basis, such as low-order orthogonal polynomials, and then independently smooth the estimated coefficients. Because of the orthogonality, one could then estimate the smoothed trend using the smoothed coefficients at each location. Of course one still needs to account for the residual spatial correlation in estimating the coefficients, which becomes an even larger challenge if one moves away from simple linear models. The second issue is that even if one is able to estimate nonlinear trends, displaying the results is an important challenge. Linear trends can be portrayed as a unidimensional quantity on a map. Nonlinear trends cannot be so easily summarized. One possibility is to display the nonlinear trends as time series plots for a moderate number of representative locations.

Chapter 6

Conclusions and Future Work

6.1 Summary and Contributions

This thesis makes several original contributions. The first contribution is my extension of the Higdon et al. (1999) (HSK hereafter) kernel convolution nonstationary covariance function. The HSK covariance is a straightforward way to construct a flexible nonstationary covariance function, albeit with some restrictions, such as non-negativity. The HSK covariance function is a generalization of the well-known squared exponential stationary correlation function. When used in Gaussian process (GP) distributions, both produce sample functions that are infinitely differentiable in both a mean square and sample path sense. This high degree of smoothness is generally undesirable. In Chapter 2, I generalize the HSK covariance function to produce a class of nonstationary correlation functions, one of which is a nonstationary version of the Matérn correlation, which has a parameter that controls the degree of differentiability. I prove that the smoothness properties of the new nonstationary correlation functions follow from those of the stationary correlation functions on which they are based, under certain smoothness conditions on the kernels used to construct the nonstationarity. In Chapters 4 and 5, I show that the Matérn nonstationary correlation can be used within a Gaussian process prior in nonparametric regression and spatial smoothing models.

Given the class of nonstationary covariance functions, we need a way to parameterize kernels that vary smoothly in space. In Chapter 3, I describe one parameterization based on the eigendecomposition of the kernel matrices and an overparameterized model for the eigenvectors. This parameterization seems to be feasible for two- and three-dimensional covariate spaces, but the number of parameters, and presumably the difficulty in fitting the model, increases with the dimension of the covariate space. In higher dimensions, simpler alternatives may be required and even abandoning nonstationarity may be a reasonable approach. When GPs are embedded in a hierarchical model, or used in any nonconjugate fashion, the process cannot be integrated out of the model. This nullifies the general approach taken to fitting GP models via Markov chain Monte Carlo (MCMC) and causes major problems in MCMC convergence. I introduce a sampling scheme, which I call posterior mean centering, that can be used in some cases in which the process cannot be integrated out of the model. The scheme greatly improves mixing of the GP hyperparameters, although mixing is still quite slow. Hyperparameter mixing is important because the covariance hyperparameters control the flexibility of the function; they perform the same role that the number of knots does in a spline model. The sampling scheme can be applied successfully to generalized nonparametric regression models with non-Gaussian response using an analogue to the familiar iteratively-reweighted least squares algorithm, as I describe in Chapter 3 and demonstrate in Chapter 4. Unfortunately, when Gaussian processes are used to construct the kernel matrices of the nonstationary kernel covariance, there is no way to make use of the PMC scheme, and we are left with a naive, slowly-mixing scheme for sampling the processes controlling the kernel structure.

The nonparametric GP regression model that I define based on the generalized kernel convolution covariance successfully models a number of experimental datasets, when the dimension of the covariate is between one and three. In one dimension, the method is successful provided the function does not change too sharply, but the BARS method of DiMatteo et al. (2002) is more successful under all conditions examined. In two and three dimensions, the nonstationary GP model outperforms free-knot spline methods generalized to higher-dimensional covariate spaces, as well as a stationary GP model, although the improvement is relatively limited in some cases. While these results are encouraging, successful comparison with standard nonparametric regression methods such as kernel smoothing, wavelet regression, and neural network models, would strengthen my conclusions, and I hope to perform additional comparisons. Based on the current results, for datasets in which a high degree of inhomogeneity is expected, use of a nonstationary GP model may be advantageous, such as in two- and three-dimensional geophysical datasets. However, in many regression settings, the inhomogeneity may not be drastic and the spline methods (which are nonstationary in nature) and stationary GP model may perform well, with the advantage of being faster computationally and having less complicated parameterizations. The GP model gives smooth function realizations and allows one to model the degree of differentiability of the function, as well as allowing one to place a prior on the degree of flexibility of the function using the trace of the smoothing matrix. In the former aspects, these features of the model are advantageous relative to some of the spline-based models, which must deal with the difficulty of generalizing one-dimensional spline basis functions to higher dimensions. The GP model moves smoothly between functions with varying degrees of flexibility, based on the sizes of the kernels in the nonstationary case and the correlation scale parameters in the stationary case. These parameters control the implicit model dimension and allow one to embed a range of models in one structure without the need for reversible jump MCMC, but this comes at the cost of slow mixing and computation. The model structure implicitly favors smoother functions if these are consistent with the data, because of the Occam's razor effect. This aspect of the Bayesian modelling approach is very attractive, since it encompasses our desire for simpler models, all else being equal. This type of prior information is an aspect of Bayesian modelling that even diehard frequentists may find appealing, since a subjective preference for simpler models is widespread and is usually taken into account implicitly in the choice of likelihood.

I also use the nonstationary correlation function to model correlation in the residual structure of a spatial model designed to jointly assess trends in time of climatological data at multiple spatial locations. The nonstationary model better accounts for the residual covariance structure than do a stationary model and a wavelet-based model. However, probably because the data are pre-smoothed by a deterministic climatological model, the nonstationary model does not predict held-out time points any better than simply using the maximum likelihood estimates. The full Bayesian model does allow one to jointly assess trend significance and compare the results to the frequentist False Discovery Rate (FDR) approach to multiple testing. The Bayesian model shrinks both the point estimates and the estimated uncertainty in those estimates by borrowing strength across space. For one index of storm activity, the Bayesian model indicates the existence of more real trends than does the FDR approach, while for a second index, the Bayesian model indicates that few locations have strong trends, somewhat fewer than the FDR results. The nonstationary GP method also allows one to perform inference at unobserved locations, unlike some methods for smoothing the empirical correlation structure. Perhaps the most important result from the spatial model analysis is that the comparison of covariance models indicates the difficulty in flexibly fitting residual covariance without overfitting, as I discuss further in Section 6.3. Accounting for residual spatial structure is important in assessing trends in spatial variables and is an area in need of more research.

6.2 Shortcomings and Potential Improvements

Here I describe a number of drawbacks to the nonstationary Gaussian process approach to nonparametric regression and spatial modelling.

The first obvious shortcoming of the model is computational speed. Fitting the model via MCMC involves computing the Cholesky factor of n by n covariance matrices within every iteration of the Markov chain. This is the case even in the regression model in which the function can be integrated out of the model because of the use of GPs to model the eigenstructures of the kernels. Ongoing work in the machine learning community on reduced rank approximations to the covariance may introduce techniques for improving the speed of the models. Another possibility is to use simpler functional forms for processes high in the model hierarchy where the full flexibility of a GP prior may not be needed. The difficulty with GP models is that even though the implicit kernel smoothing that is being performed is essentially local, the computations involve the full covariance matrices and do not make use of sparsity in any sense.

The difficulties involving computational speed are compounded by the slow mixing of GP models. The PMC method introduced here improves the mixing of the hyperparameters in generalized nonparametric regression models, while Langevin updates improve the mixing of the process values. If it can be extended to work with numerically singular covariance matrices, the adaptive reparameterization of Christensen et al. (2003), possibly in conjunction with PMC, may greatly improve mixing, although I suspect that datasets with thousands of observations will still be difficult to fit. However, neither the adaptive reparameterization nor the PMC scheme will work with the GPs used for the kernel covariance structure. This suggests that a parameterization of the kernel structure that is easier to fit may be worth pursuing.

In Chapter 4, I demonstrate that the nonstationary GP model is unable to capture a sharp jump in the regression function without undersmoothing in the neighborhood of the jump. This is inherent in the smooth parameterization of the kernels used to construct the nonstationary covariance function. One possibility that may help somewhat in capturing jumps is the use of asymmetric kernels. This would allow a point to have high correlation with points to one side where the function is smooth and low correlation with points on the other side where the jump occurs. In higher dimensions, this might involve asymmetry across a hyperplane; such non-Gaussian kernels would seem to be difficult to model. Ultimately, the use of a standard kernel shape restricts the types of functions that can be approximated by the GP model. Models that can adaptively add and delete basis functions, such as the spline models and neural network models, may perform better in some cases, although fitting such models and assessing convergence can be difficult, just as in the GP case.

The eigendecomposition model for the kernel structure seems feasible in two and three dimensions, but the number of parameters increases quickly with covariate dimension, even with the simplified scheme of sharing a single correlation structure across the eigenprocesses. However, simpler parameterizations carry the danger of not being as flexible in modelling features of the data. It may be that there are other parameterizations that do a better job of capturing features of the data without having too many parameters or causing mixing problems.

In addition to being faithful to the real features of the data, a regression model should also ignore covariates that appear to be independent of the response. In some initial experimentation, it appeared that my parameterization of the nonstationary GP model did a poor job of ignoring unimportant covariates, in part because this requires that the relevant correlation scale parameter (in the nonstationary model, this is the size of the kernel in the direction of the covariate) become very large, so that the response is highly correlated in the direction of the unimportant covariate. This is difficult to model without causing mixing problems, since as the parameter increases, the likelihood becomes quite flat. One possibility for addressing this shortcoming would be to reparameterize the model so that once that parameter becomes sufficiently large, one enters a discrete part of the parameter space in which the covariate is completely ignored. Such an approach would have the benefits of ignoring unimportant covariates, allowing one to assess which covariates are relevant, and improving mixing by removing a troublesome part of the covariate space, but would require moving between models of different dimension during the MCMC.

Examination of the empirical correlations in Chapter 5 suggests that the covariance structure of the storm data is complicated; such complicated structure is likely to be present in many spatial datasets. The basis kernel construction of the nonstationary covariance seems to capture only a portion of the local non-negative nonstationary structure apparent in the empirical correlations. This may be because the apparent structure is noise that the nonstationary model is correctly ignoring. However, it seems plausible that there is structure present that the model is unable to capture for some reason. There are structural reasons that the nonstationary model may not be able to capture the true underlying correlation patterns, including its inability, as parameterized here, to account for the long-distance and negative correlations, and the constraints on the correlation structure induced by using the particular kernel form of simple Gaussian kernels. Application of the model in this thesis requires the assumption that these limitations have little effect on the resulting posterior distribution. While this may be reasonable in many applications, methods for dealing with complicated nonstationarity would be desirable.

However, flexibly modelling covariance structure appears to be a difficult task, much more difficult than modelling mean structure. Wikle et al. (1998) build hierarchical models in which the modelling focuses on the mean structure at each stage in the hierarchy in an attempt to avoid joint covariance modelling for a large number of locations. The difficulties in joint covariance modelling are apparent in the model comparison results here. The wavelet model that closely follows the empirical correlation drastically overfits, with very poor generalizability. The smoothed wavelet model does a poor job of predicting test data, both in terms of point predictions and covariance structure, presumably because it is fit without reference to the likelihood or rigorous fitting criteria. This poor predictive ability occurs even though the speed at which the correlations in the wavelet model decay appears to more closely mirror the empirical correlations than does the speed at which the correlations decay in the nonstationary model. While the kernel-based nonstationary model has its problems, as discussed above, the stationary model suffers in comparison. In the stationary model, many of the residual variance estimates are much larger than the ML estimates.

The variances seem to be accommodating the inability of the stationary model to fit the correlation structure. One difficulty that may play an important role in these results is that when modelling many locations simultaneously, the response at some locations can be nearly a linear combination of the response at other locations. When this constraint is part of a covariance structure and is violated, the fit can be very poor, even if the covariances look reasonable to the eye and on an element by element basis. It is possible that the quicker decay with distance in the correlations in the GP models relative to the wavelet models serves to minimize the inclusion of such linear combinations in the covariance matrix, at the cost of modelling other features of the covariance structure less closely. A possible alternative to the methods used here is to employ the wavelet approach with a specific criteria for fitting, such as making use of the likelihood and performing the thresholding in a Bayesian context. This might produce a flexible covariance model that does a better job of prediction and gets around the difficulty in choosing the thresholding in the current approach, although it would seemingly obviate many of the computational advantages of the method.

6.3 Future Work

In addition to the ideas mentioned in the previous section, there are some general areas of potential future work involving GP models that may be fruitful. My discussion here focuses first on generalizing the Gaussian process approach in two ways and then comments on including nonlinear time structure in the spatial model.

Previous approaches to modeling non-normal data using Gaussian processes have generally assumed independence of the observations conditional on a Gaussian process prior for a function determining location (Diggle et al. 1998). In many spatial datasets with replicated data, including the discrete-valued storm count index of Paciorek et al. (2002), residual correlation suggests this assumption of independence is violated. Here I suggest a generalized Gaussian process distribution for non-normal data with a single observation per location. To describe the distribution of the data, I show how to generate a single sample at each of n covariate values. Let $Z(\cdot)$ be a Gaussian process with covariance function $C(\cdot,\cdot)$. For each covariate, $\boldsymbol{x_i}$, let $p(\boldsymbol{x_i}) = \Phi(Z(\boldsymbol{x_i}))$, the standard normal CDF transformation of the Gaussian process at each covariate; $p(\cdot)$ might be called the quantile process. Next let $H_{\mu(\cdot)}$ be an appropriate parametric distribution function indexed by a mean/location process, $\mu(\cdot)$, which could be a function of additional covariates or have a Gaussian process prior itself. For count data, one would likely choose H to be Poisson, while for binary data, it would be Bernoulli. Then the data are generated such that $Y_i = H^{-1}_{\mu(\boldsymbol{x_i})}(p(\boldsymbol{x_i}))$, i.e. the $p(\boldsymbol{x_i})$ quantile of $H_{\mu(\boldsymbol{x_i})}$. If the variance terms in the covariance function of $Z(\cdot)$ are equal to one, then the marginal distribution of $Z(\cdot)$ at each covariate is standard normal and therefore $p(\cdot)$ is $U(0,1)$, which means that $Y_i \sim H_{\mu(\boldsymbol{x_i})}$. If the variance at a covariate is less (greater) than one, then the distribution of the observation is under(over)-dispersed with median equal to that of $H_{\mu(\boldsymbol{x_i})}$. This approach mimics the inherent separation of mean and covariance in a Gaussian distribution by modelling the mean/location separately from the correlation structure.

The storm indices of Paciorek et al. (2002) are closely related to each other, measuring different aspects of the same phenomenon, yet I have modelled them independently, in part because I do not have a reasonable model for the joint distribution of the indices as a function of spatial location. To my knowledge little work has been done to define models for regression problems involving multiple responses. As one approach to defining a covariance structure for two responses, consider the kernel convolution of Higdon et al. (1999), but introduce a cross-covariance matrix process, $C'_{a,b}(\cdot)$, that relates the responses, a and b, at any location. Now if we have two separate kernel matrix processes, one for each response, we can do the convolution,

$$C_{a,b}(x,y) = \int_u K^a_x(u) K^b_y(u) C'_{a,b}(u) du,$$

to produce the covariance between the two responses at any pair of locations (x, y). The parameterizations for smoothly spatially-varying covariance matrices in Chapter 3 allow one to model the kernel matrix processes as before, as well as the newly-introduced cross-covariance matrix processes. Note that the covariance between the responses at the same location, $C_{a,b}(x,x)$ is not determined solely by the cross-covariance at x; this makes the underlying cross-covariance matrices, $C'_{a,b}(\cdot)$, somewhat difficult to interpret. In principle, this approach extends simply to more than two response variables. Other parameterizations may also be useful.

A final area for future research, which I discuss in more detail in Section 5.8.2, involves modelling nonlinear trends at multiple locations while accounting for spatial structure. The goal is to be able to understand the broad-scale time trends while allowing for nonlinearities that are likely to be present. Even if a reasonable temporal model can be constructed and fit, presenting the results in an easily accessible manner will be a challenge because one is working with high-dimensional quantities in addition to the two-dimensional covariate space. Of course as the temporal modelling becomes more complicated and additional parameters are introduced, accounting for spatial structure also becomes more of a challenge, and I have found that doing this even in the linear trend case is difficult.

Appendices

A Spatial Model Prior Distributions and Starting Values

A.1 Prior distributions

I chose prior distributions to be diffuse but proper, based on the mean values of temperature variance and Eady growth over 1949-1999 for the Northern hemisphere, $20^\circ - 70^\circ$ N, using Paciorek et al. (2002, Fig. 1). These climatological mean fields gave me a sense for the reasonable range of values for the parameters. In particular, for temperature variance, I take

$$\begin{aligned}
\mu_\alpha &\sim \mathrm{N}(1.2, 0.7^2) \\
\sigma^2_\alpha &\sim \mathrm{IG}(0.1, 0.1) \\
\mu_\beta &\sim \mathrm{N}(0, 0.01^2) \\
\sigma^2_\beta &\sim \mathrm{IG}(0.1, 1 \times 10^{-7}) \\
\mu_\eta &\sim \mathrm{N}(-4.0, 3.0^2) \\
\sigma^2_\eta &\sim \mathrm{IG}(0.1, 0.1) \\
\log \kappa_\phi &\sim \mathrm{N}(-5.4, 1.2^2) \\
\log \delta &\sim \mathrm{U}(-23.0, 2.3) \\
\nu_{\boldsymbol{Y}} &\sim \mathrm{U}(0.5, 15.0),
\end{aligned}$$

where κ_ϕ indicates that I use the same prior for $\phi \in \{\alpha, \beta, \eta\}$. For the inverse gamma distributions, these are parameterized such that the mean is $\frac{\beta}{\alpha - 1}$. For the stationary model, I take $\log \kappa_{\boldsymbol{Y}} \sim \mathrm{N}(-5.4, 1.2^2)$. For the nonstationary model, $k = 1, \ldots, 9$, I take $\log \lambda_k \sim \mathrm{U}(-5.6, 7.8)$ for each of the two eigenvalues in the kth basis kernel matrix. I take the kth basis kernel matrix Givens angle, $\gamma_k \sim \text{U}(0,\pi)$. Finally, I take the weight decay parameter, $\log \kappa_{\boldsymbol{Y}} \sim \text{U}(-2.3, 1.6)$. The priors for parameters that affect various correlation scales are informed by the fact that I do not want the scale less than the smallest distance between grid points or much larger than the largest distance between grid points.

For Eady growth rate, I use the same parameters, with the following exceptions:

$$\begin{aligned}
\mu_\alpha &\sim \text{N}(0.5, 0.5^2) \\
\sigma_\alpha^2 &\sim \text{IG}(0.1, 0.001) \\
\mu_\beta &\sim \text{N}(0.0, 0.005^2) \\
\sigma_\beta^2 &\sim \text{IG}(0.1, 1 \times 10^{-8}) \\
\mu_\eta &\sim \text{N}(-7.0, 6.0^2).
\end{aligned}$$

For the wavelet models, $R_{\boldsymbol{Y}}$ is fixed, so I only have priors for the remaining parameters, which I take to be the same as for the kernel nonstationary model.

A.2 Starting values

For starting values for the hyperparameters of the $\alpha(\cdot), \beta(\cdot)$ and $\log \eta(\cdot)^2$ processes, I calculated approximate maximum likelihood estimates (MLEs) based on $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\eta}}^2$, namely the MLEs for the process values based on assuming independent locations, and used these approximate MLEs for the hyperparameters to come up with reasonable starting values. For temperature variance I use $\mu_\alpha = 1.2$, $\log \sigma_\alpha = -1.2$, $\log \kappa_\alpha = -1.3$, $\mu_\beta = 0.0$, $\log \sigma_\beta = -6.9$, $\log \kappa_\beta = -1.3$, $\mu_\eta = -4.0$, $\log \sigma_\eta = 0.0$, $\log \kappa_\eta = -1.3$, $\log \delta = -9.2$, and $\nu_{\boldsymbol{Y}} = 4.0$. For the stationary model, I use $\log \kappa_{\boldsymbol{Y}} = -2.2$ and for the nonstationary model, I take $\log \kappa_{\boldsymbol{Y}} = -1.2$ and started the basis kernel matrix eigenvalues at 1.0 and the Givens angles at $\frac{\pi}{2}$. I construct the process values, $\boldsymbol{\phi} = \mu_\phi + \sigma_\phi L_{\boldsymbol{\phi}} \boldsymbol{\omega}_\phi$, using $\boldsymbol{\omega}_\phi \sim \text{N}(0, I)$.

For the Eady growth models, I used the same initial values, with the following exceptions. I took $\mu_\alpha = 0.72$, $\log \sigma_\alpha = -1.8$, $\log \kappa_\alpha = -1.1$, $\mu_\beta = 0.00012$, $\log \sigma_\beta = -6.6$, $\log \kappa_\beta = -1.3$, $\mu_\eta = -5.2$, $\log \sigma_\eta = -0.7$, and $\log \kappa_\eta = -1.5$. For the stationary model, I used $\log \kappa_{\boldsymbol{Y}} = -2.38$.

For the wavelet models, I again used the same values, except for the parameters involved in $R_{\boldsymbol{Y}}$, which are not used.

B Notation

I have attempted to be consistent in my notation, both in my use of alphabets and cases, as well as my use of individual letters and symbols. However, in a work as large as this, I have needed in some situations to use the same symbol in different contexts, and there are also undoubtedly places where I have not been entirely consistent.

In general, I have indicated functions with both lower and upper case Arabic letters, matrices with upper case Arabic and Greek letters, vectors with bold Arabic and Greek letters, and parameters with lower-case Greek letters. For indices, I have used lower case Arabic letters.

For random variables that are parameters, I have been lax and used lower case Greek letters to indicate the random variable itself and realizations of the random variable.

In various places in the thesis, I need a vector-valued mean for a vector-valued random variable; as necessary I take $\mu = \mu\mathbf{1}$.

Next I list the notation and meanings, broadly grouped.

B.1 Parameters, stochastic processes, matrices, data, and covariates

$f(\cdot), \boldsymbol{f}, f_i, f(\boldsymbol{x_i})$: a regression function/process, a vector of values of the function evaluated at a finite set of covariates, the value of the function at the ith covariate value, the value of the function at $\boldsymbol{x_i}$

$\phi(\cdot), \boldsymbol{\phi}, \phi_i, \phi(\boldsymbol{x_i})$: a stochastic process, a vector of values of the process evaluated at a finite set of covariates, the value of the process at the ith covariate value, the value of the process at $\boldsymbol{x_i}$

$Z(\cdot), Z(\boldsymbol{x_i})$: a stochastic process, the value of the process at $\boldsymbol{x_i}$

$\alpha(\cdot), \beta(\cdot), \eta(\cdot)^2$: intercept, slope, and residual variance processes in the spatial model

η^2: error (noise) variance in the regression model

$\boldsymbol{Y}, \boldsymbol{y}$: vector of data values as a random variable, as a realization

$\boldsymbol{x}_i, \boldsymbol{x}_j$: two different covariate values, $\boldsymbol{x}_i \in \Re^P$

Σ_i : positive definite kernel matrix

R : correlation matrix

C : covariance matrix

Γ: eigenvector matrix

Λ: eigenvalue matrix

Q: quadratic form in nonstationary correlation function

τ : Euclidean distance

κ : correlation scale parameter, in units of distance

ν : smoothness parameter in Matérn correlation function

μ : mean of a stochastic process

σ^2 : variance of a stochastic process

ρ : angle or angular distance

$\theta, \boldsymbol{\theta}$: a parameter or vector of parameters

$\boldsymbol{\omega}$: a vector of values with a standard normal prior or drawn from a standard normal, or white noise values in general

$\psi, \boldsymbol{\psi}$: value(s) used in generating an MCMC proposal

v : proposal standard deviation in an MCMC

ϵ : tolerance in numerical calculations

$\boldsymbol{u}$: spatial location

S, s : scale parameter

W, w : spectral random variable

λ : eigenvalue

γ : parameter used in constructing eigenvectors

c: a constant

B.2 Indices

i : indexes training set values $(1, \ldots, n)$

j : indexes test set values $(1, \ldots, m)$

k : indexes MCMC draws or number of components in a model (1,. . . ,K)

m : indexes derivatives $(1, \ldots, M)$

p : indexes dimension of the covariate space $(1, \ldots, P)$

t : indexes time in the spatial model $(1, \ldots, T)$

$\boldsymbol{x}_i, \boldsymbol{x}_j$: two different covariate values, $\boldsymbol{x}_i \in \Re^P$

$\boldsymbol{f_1}, \boldsymbol{f_2}$: training set values of f, test set values of f

B.3 Symbols, superscripts, and subscripts

$f^{(m)}$: the mth derivative of the function, f

$f_{(k)}$: kth MCMC draw

$\hat{\boldsymbol{\phi}}$: maximum likelihood estimate

$\check{f}$: the true value of a parameter

$\tilde{f}$: posterior mean

$\widetilde{f|\mu}$: conditional posterior mean

B.4 Functions

$R(\cdot), R(\cdot, \cdot)$: stationary correlation function, nonstationary correlation function

$C(\cdot), C(\cdot, \cdot)$: stationary covariance function, nonstationary covariance function

$g(\cdot)$: used to indicate functions in various contexts

$h(\cdot), H(\cdot)$: density function, distribution function

References

Abrahamsen, P. (1997), "A review of Gaussian random fields and correlation functions," Technical Report 917, Norwegian Computing Center.

Abramowitz, M. and Stegun, I. (1965), *Handbook of Mathematical Functions*, New York: Dover.

Adler, R. (1981), *The Geometry of Random Fields*, New York: John Wiley & Sons.

Anderson, T. W., Olkin, I., and Underhill, L. G. (1987), "Generation of random orthogonal matrices," *SIAM Journal on Scientific and Statistical Computing*, 8, 625–629.

Bakin, S., Hegland, M., and Osborne, M. (2000), "Parallel MARS algorithm based on B-splines," *Computational Statistics*, 15, 463–484.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000), "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, 10, 1281–1311.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, 57, 289–300.

Biller, C. (2000), "Adaptive Bayesian regression splines in semiparametric generalized linear models," *Journal of Computational and Graphical Statistics*, 9, 122–140.

Billingsley, P. (1995), *Probability and Measure* (third ed.), New York: John Wiley & Sons.

Bochner, S. (1959), *Lectures on Fourier Integrals*, Princeton, N.J.: Princeton University Press.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *J. Am. Stat. Assoc.*, 88, 9–25.

Brockmann, M., Gasser, T., and Herrmann, E. (1993), "Locally adaptive bandwidth choice for kernel regression estimators," *J. Am. Stat. Assoc.*, 88, 1302–1309.

Brockwell, A. and Kadane, J. (2002), "Identification of regeneration times in MCMC simulation, with application to adaptive schemes," Technical Report 770, Department of Statistics, Carnegie Mellon University.

Bruntz, S., Cleveland, W., Kleiner, B., and Warner, J. (1974), "The dependence of ambient ozone on solar radiation, temperature, and mixing height," in *Symposium on Atmospheric Diffusion and Air Pollution*, ed. A. M. Society, pp. 125–128.

Buck, R. C. (1965), *Advanced Calculus* (second ed.), New York: McGraw-Hill.

Cambanis, S. (1973), "On some continuity and differentiability properties of paths of Gaussian processes," *Journal of Multivariate Analysis*, 3, 420–434.

Casella, G. and Berger, R. (1990), *Statistical Inference*, Belmont, California: Duxbury Press.

Chipman, H., George, E., and McCulloch, R. (1998), "Bayesian CART model search," *J. Am. Stat. Assoc.*, 93, 935–960.

Christensen, O., Møller, J., and 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.

——— (2001), "Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalized linear mixed models," *Methodology and Computing in Applied Probability*, 3, 309–327.

Christensen, O., Roberts, G., and Sköld, M. (2003), "Robust MCMC methods for spatial GLMMs," *in preparation*.

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

Churchill, R. (1960), *Complex variables and applications* (second ed.), New York: McGraw-Hill.

Cleveland, W. and Devlin, S. (1988), "Locally-weighted regression: an approach to regression analysis by local fitting," *J. Am. Stat. Assoc.*, 83, 597–610.

Cohen, A. and Jones, R. H. (1969), "Regression on a random field," *J. Am. Stat. Assoc.*, 64, 1172–1182.

Cramér, H. and Leadbetter, M. (1967), *Stationary and Related Stochastic Processes*, New York: John Wiley & Sons.

Cressie, N. (1993), *Statistics for Spatial Data* (Revised ed.): Wiley-Interscience.

Damian, D., Sampson, P., and Guttorp, P. (2001), "Bayesian estimation of semi-parametric non-stationary spatial covariance structure," *Environmetrics*, 12, 161–178.

Daniels, M. and Kass, R. (1999), "Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models," *J. Am. Stat. Assoc.*, 94, 1254–1263.

Das, B. (2000), *Global Covariance Modeling: a Deformation Approach to Anisotropy*, unpublished Ph.D. dissertation, University of Washington, Department of Statistics.

Denison, D., Holmes, C., Mallick, B., and Smith, A. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Chichester, England: Wiley.

Denison, D., Mallick, B., and Smith, A. (1998a), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society, Series B*, 60, 333–350.

——— (1998b), "Bayesian MARS," *Statistics and Computing*, 8, 337–346.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics," *Applied Statistics*, 47, 299–326.

DiMatteo, I. (2001), *Bayesian Curve Fitting Using Free-Knot Spline*, unpublished Ph.D. dissertation, Carnegie Mellon University, Department of Statistics.

DiMatteo, I., Genovese, C., and Kass, R. (2002), "Bayesian curve-fitting with free-knot splines," *Biometrika*, 88, 1055–1071.

Donoho, D. L. and Johnstone, I. M. (1995), "Adapting to unknown smoothness via wavelet shrinkage," *J. Am. Stat. Assoc.*, 90, 1200–1224.

Doob, J. (1953), *Stochastic Processes*, New York: John Wiley & Sons.

Duane, S., Kennedy, A., Pendleton, B., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters B*, 195, 216–222.

Friedman, J. (1991), "Multivariate adaptive regression splines," *Annals of Statistics*, 19, 1–141.

Fuentes, M. (2001), "A high frequency kriging approach for non-stationary environmental processes," *EnvironMetrics*, 12, 469–483.

Fuentes, M. and Smith, R. (2001), "A New Class of Nonstationary Spatial Models," Technical report, North Carolina State University, Department of Statistics.

Gaspari, G. and Cohn, S. (1999), "Construction of correlation functions in two and three dimensions," *Quart. J. Roy. Meteor. Soc.*, 125, 723–757.

Gelfand, A., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003), "Spatial modelling with spatially varying coefficient processes," Technical report, Department of Statistics, University of Connecticut.

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

Gelman, A. and Rubin, D. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–511.

Gibbs, M. (1997), *Bayesian Gaussian Processes for Classification and Regression*, unpublished Ph.D. dissertation, University of Cambridge.

Gibbs, M. and MacKay, D. (1997), "Efficient implementation of Gaussian processes," Technical report, University of Cambridge.

Gihman, I. and Skorohod, A. (1974), *The Theory of Stochastic Processes, I*, New York: Springer-Verlag.

Gneiting, T. (1999), "Correlation functions for atmospheric data analysis," *Quart. J. Roy. Meteor. Soc.*, 125, 2449–2464.

——— (2001), "Compactly supported correlation functions," *Journal of Multivariate Analysis*, 83, 493–508.

Golub, G. and van Loan, C. (1996), *Matrix Computations* (third ed.), Baltimore, Maryland: Johns Hopkins University Press.

Gradshteyn, I. and Ryzhik, I. (1980), *Tables of Integrals, Series and Products: Corrected and Enlarged Edition*, New York: Academic Press, Inc.

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

Green, P. and Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models*, Boca Raton: Chapman & Hall/CRC.

Handcock, M. and Stein, M. (1993), "A Bayesian analysis of kriging," *Technometrics*, 35, 403–410.

Hansen, M. and Kooperberg, C. (2002), "Spline adaptation in extended linear models," *Statistical Science*, 17, 2–51.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag Inc.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall Ltd.

Higdon, D. (1998), "A process-convolution approach to modeling temperatures in the North Atlantic Ocean," *Journal of Environmental and Ecological Statistics*, 5, 173–190.

——— (2002), "Space and space-time modeling using process convolutions," in *Quantitative Methods for Current Environmental Issues*, ed. C. Anderson, London: Springer, pp. 37–54.

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

Hilbert, D. and Cohn-Vassen, S. (1952), *Geometry and the Imagination*, New York: Chelsea Pub. Co., translated by P. Nemenyi. Translation of Anschauliche Geometrie.

Holland, D., Oliveira, V. D., Cox, L., and Smith, R. (2000), "Estimation of regional trends in sulfur dioxide over the eastern United States," *Environmetrics*, 11, 373–393.

Holmes, C. and Mallick, B. (2001), "Bayesian regression with multivariate linear splines," *Journal of the Royal Statistical Society, Series B*, 63, 3–17.

Holmes, C., Mallick, B., and Kim, H. (2002), "Analyzing non-stationary spatial data using piecewise Gaussian processes,", Poster presented at Bayesian Statistics 7.

Huerta, G., Sansó, B., and Guenni, L. (2001), "A spatio-temporal model for Mexico City ozone levels," Technical Report 2001-04, Centro de Estadística y Software Matemático,Universidad Simón Bolívar.

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

Kalnay, E. and Coauthors (1996), "The NCEP/NCAR 40-year reanalysis project," *Bull. Amer. Meteor. Soc.*, 77, 437–471.

Kennedy, M. C. and O'Hagan, A. (2001), "Bayesian calibration of computer models," *Journal of the Royal Statistical Society, Series B*, 63, 425–464.

Kitagawa, G. (1987), “Non-Gaussian state-space modeling of nonstationary time series,” *J. Am. Stat. Assoc.*, 82, 1032–1063.

Lee, H. K. (1998), *Model Selection and Model Averaging for Neural Networks*, unpublished Ph.D. dissertation, Carnegie Mellon University, Department of Statistics.

Leithold, L. (1968), *The Calculus with Analytic Geometry*, New York: Harper & Row.

Loader, C. and Switzer, P. (1992), “Spatial covariance estimation for monitoring data,” in *Statistics in the Environmental and Earth Sciences*, eds. A. Walden and P. Guttorp, London: Edward Arnold.

Lockwood, J. (2001), *Estimating Joint Distributions of Contaminants in U.S. Community Water System Sources*, unpublished Ph.D. dissertation, Carnegie Mellon University, Department of Statistics.

Lockwood, J., Schervish, M., Gurian, P., and Small, M. (2001), “Characterization of arsenic occurrence in source waters of U.S. community water systems,” *J. Am. Stat. Assoc.*, 96, 1184–1193.

Loève, M. (1978), *Probability Theory II* (fourth ed.), New York: Springer-Verlag Inc.

MacKay, D. (1997), “Introduction to Gaussian Processes,” Technical report, University of Cambridge.

Matérn, B. (1986), *Spatial Variation* (second ed.), Berlin: Springer-Verlag Inc.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models* (second ed.), Boca Raton: Chapman & Hall.

McLeish, D. (1982), “A robust alternative to the normal distribution,” *The Canadian Journal of Statistics*, 10, 89–102.

Meiring, W., Monestiez, P., Sampson, P., and Guttorp, P. (1997), “Developments in the modelling of nonstationary spatial covariance structure from space-time monitoring data,” in *Geostatistics Wallongong 1996*, eds. E. Baafi and N. Schofield, Dordrecht: Kluwer, pp. 162–173.

Minka, T. P. and Picard, R. W. (1997), “Learning how to learn is learning with point sets,”, http://citeseer.nj.nec.com/minka97learning.html.

Møller, J., Syversveen, A., and Waagepetersen, R. (1998), “Log Gaussian Cox processes,” *Scandinavian Journal of Statistics*, 25, 451–482.

Neal, R. (1993), “Probabilistic Inference Using Markov Chain Monte Carlo Methods,” Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

——— (1996), *Bayesian Learning for Neural Networks*, New York: Springer.

——— (1997), “Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification,” Technical Report 9702, Department of Statistics, University of Toronto.

Nott, D. and Dunsmuir, W. (2002), “Estimation of nonstationary spatial covariance structure,” *Biometrika*, 89, 819–829.

Nychka, D., Wikle, C., and Royle, J. (2001), “Multiresolution Models for Nonstationary Spatial Covariance Functions,” Technical report, Geophysical Statistics Project, National Center for Atmospheric Research.

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

Odell, P. and Feiveson, A. (1966), “A numerical procedure to generate a sample covariance matrix,” *J. Am. Stat. Assoc.*, 61, 199–203.

Oehlert, G. (1993), “Regional trends in sulfate wet deposition,” *J. Am. Stat. Assoc.*, 88, 390–399.

O’Hagan, A. (1978), “Curve fitting and optimal design for prediction,” *Journal of the Royal Statistical Society, Series B*, 40, 1–42.

Olmsted, J. M. (1961), *Advanced Calculus*, New York: Appleton-Century-Crofts.

Paciorek, C., Risbey, J., Ventura, V., and Rosen, R. (2002), “Multiple indices of Northern Hemisphere Cyclone Activity, Winters 1949-1999,” *J. Climate*, 15, 1573–1590.

Paige, R. and Butler, R. (2001), “Bayesian inference in neural networks,” *Biometrika*, 88, 623–641.

Papaspiliopoulos, O., Roberts, G., and Sköld, M. (2003), “Non-centered parameterisations for hierarchical models and data augmentation,” in *Bayesian Statistics 7*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford, U.K.: Oxford University Press, pp. 319–333.

Paulo, R. (2002), *Problems on the Bayesian-Frequentist Interface*, unpublished Ph.D. dissertation, Duke University, Institute of Statistics and Decision Sciences.

Rasmussen, C. (1996), *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*, unpublished Ph.D. dissertation, University of Toronto, Graduate Department of Computer Science.

Rasmussen, C. and Ghahramani, Z. (2001), “Occam’s razor,” in *Advances in Neural Information Processing Systems 13*, eds. T. Leen, T. Dietterich, and V. Tresp, Cambridge, Massachusetts: MIT Press.

——— (2002), “Infinite mixtures of Gaussian process experts,” in *Advances in Neural Information Processing Systems 14*, eds. T. G. Dietterich, S. Becker, and Z. Ghahramani, Cambridge, Massachusetts: MIT Press.

Robert, C. and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Roberts, G. O. and Rosenthal, J. S. (1998), “Optimal scaling of discrete approximations to Langevin diffusions,” *Journal of the Royal Statistical Society, Series B, Methodological*, 60, 255–268.

——— (2001), “Early sample measures of variability,” *Statistical Science*, 16(4), 351–367.

Ruppert, D. and Carroll, R. J. (2000), “Spatially-adaptive penalties for spline fitting,” *Australian and New Zealand Journal of Statistics*, 42(2), 205–223.

Sampson, P., Damian, D., and Guttorp, P. (2001), “Advances in modeling and inference for environmental processes with nonstationary spatial covariance,” in *GeoENV 2000: Geostatistics* *for Environmental Applications*, eds. P. Monestiez, D. Allard, and R. Froidevaux, Dordrecht: Kluwer, pp. 17–32.

Sampson, P. and Guttorp, P. (1992), “Nonparametric estimation of nonstationary spatial covariance structure,” *J. Am. Stat. Assoc.*, 87, 108–119.

Sansó, B. and Guenni, L. (2002), “Combining Observed Rainfall and Deterministic Prediction Using a Bayesian Approch,” Technical Report 2002-02, Centro de Estadística y Software Matemático,Universidad Simón Bolívar.

Schervish, M. (1995), *Theory of Statistics*, New York: Springer-Verlag.

Schmidt, A. and O’Hagan, A. (2000), “Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations,” Technical Report 498/00, University of Sheffield, Department of Probability and Statistics.

Schoenberg, I. (1938), “Metric spaces and completely monotone functions,” *Ann. of Math.*, 39, 811–841.

Schucany, W. R. (1995), “Adaptive bandwidth choice for kernel regression,” *J. Am. Stat. Assoc.*, 90, 535–540.

Seeger, M. and Williams, C. (2003), “Fast forward selection to speed up sparse Gaussian process regression,” in *Workshop on AI and Statistics 9*.

Skilling, J. (1989), “The eigenvalues of mega-dimensional matrices,” in *Maximum entropy and Bayesian methods, Cambridge, England, 1988*, ed. J. Skilling, Dordrecht: Kluwer Academic Publishers, pp. 455–466.

——— (1993), “Bayesian numerical analysis,” in *Physics and Probability*, eds. J. Grandy, W.T. and P. Milonni, Cambridge: Cambridge University Press, pp. 207–221.

Smith, M. and Kohn, R. (2002), “Parsimonious covariance matrix estimation for longitudinal data,” *J. Am. Stat. Assoc.*, 97, 1141–1153.

Smith, R. (2001), "Environmental Statistics," Technical report, Department of Statistics, University of North Carolina.

Smola, A. and Bartlett, P. (2001), "Sparse greedy Gaussian process approximation," in *Advances in Neural Information Processing Systems 13*, eds. T. Leen, T. Dietterich, and V. Tresp, Cambridge, Massachusetts: MIT Press.

Stein, M. (1999), *Interpolation of Spatial Data : Some Theory for Kriging*, New York: Springer.

Swall, J. (1999), *Non-Stationary Spatial Modeling Using a Process Convolution Approach*, unpublished Ph.D. dissertation, Duke University, Institute of Statistics and Decision Sciences.

Vanmarcke, E. (1983), *Random Fields: Analysis and Synthesis*, Cambridge, Massachusetts: The MIT Press.

Ventura, V., Paciorek, C., and Risbey, J. (2003), "Controlling the proportion of falsely-rejected hypotheses when conducting multiple tests with geophysical data," *in preparation*.

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York: John Wiley & Sons.

Vivarelli, F. and Williams, C. (1999), "Discovering hidden features with Gaussian processes regression," in *Advances in Neural Information Processing Systems 11*, eds. M. Kearns, S. Solla, and D. Cohn.

Wahba, G. (1990), *Splines for observational data*, Philadelphia: Society for Industrial and Applied Mathematics.

Wand, M. (2003), "Smoothing and mixed models," *Under Review*.

Wikle, C., Berliner, L., and Cressie, N. (1998), "Hierarchical Bayesian space-time models," *Environmental and Ecological Statistics*, 5, 117–154.

Williams, C. (1997), "Prediction with Gaussian processes: from linear regression to linear prediction and beyond," Technical Report NCRG/97/012, Neural Computing Research Group, Aston University.

Williams, C., Rasmussen, C., Schwaighofer, A., and Tresp, V. (2002), "Observations on the Nyström Method for Gaussian Process Prediction," Technical report, Gatsby Computational Neuroscience Unit, University College London.

Williams, C. and Seeger, M. (2001), "Using the Nyström Method to Speed Up Kernel Machines," in *Advances in Neural Information Processing Systems 13*, eds. T. Leen, T. Dietterich, and V. Tresp, Cambridge, Massachusetts: MIT Press.

Williams, C. K. I. and Barber, D. (1998), "Bayesian classification With Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1342–1351.

Williams, C. K. I. and Rasmussen, C. E. (1967), "Gaussian processes for regression," in *Advances in Neural Information Processing Systems, 8*, eds. D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, MIT Press.

Wood, S., Jiang, W., and Tanner, M. (2002), "Bayesian mixture of splines for spatially adaptive nonparametric regression," *Biometrika*, 89, 513–528.

Yaglom, A. (1987), *Correlation Theory of Stationary and Related Random Functions I: Basic Results*, New York: Springer-Verlag, Inc.

Yang, R. and Berger, J. O. (1994), "Estimation of a covariance matrix using the reference prior," *The Annals of Statistics*, 22, 1195–1211.

Zhou, S. and Shen, X. (2001), "Spatially adaptive regression splines and accurate knot selection schemes," *J. Am. Stat. Assoc.*, 96, 247–259.