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 dis-

tributions 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):

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

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 $\nu \to \infty$, 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,

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

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

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

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

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

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 fash-

ion, 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 $\boldsymbol{f} \sim \mathrm{N}(0, \eta^2 B D(\delta) B^T)$. 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 imple-

mentation 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 in-

finitely many hidden units. A common form of the neural network model specifies the regression function to be

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

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, $C_{\boldsymbol{f}} = B C_{\boldsymbol{\beta}} B^T$ 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.