DISSERTATION

SPATIAL PROCESSES WITH STOCHASTIC HETEROSCEDASTICITY

Submitted by

Wenying Huang

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2008

COLORADO STATE UNIVERSITY

June 4, 2008

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY WENYING HUANG ENTITLED SPATIAL PROCESSES WITH STOCHASTIC HETEROSCEDASTICITY BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

F. Jay Breidt (Adviser)

Richard A. Davis (Co-Adviser)

Haonan Wang (Committee Member)

Donald Estep (Outside Committee Member)

F. Jay Breidt (Department Head)

ABSTRACT OF DISSERTATION

SPATIAL PROCESSES WITH STOCHASTIC HETEROSCEDASTICITY

Stationary Gaussian processes are widely used in spatial data modeling and analysis. Stationarity is a relatively restrictive assumption regarding spatial association. By introducing stochastic volatility into a Gaussian process, we propose a stochastic heteroscedastic process (SHP) with conditional nonstationarity. That is, conditional on a latent Gaussian process, the SHP is a Gaussian process with non-stationary covariance structure. Unconditionally, the SHP is a stationary non-Gaussian process. The realizations from SHP are versatile and can represent spatial inhomogeneities. The unconditional correlation of SHP offers a rich class of correlation functions which can also allow for a smoothed nugget effect.

For maximum likelihood estimation, we propose to apply importance sampling in the likelihood calculation and latent process estimation. The importance density we constructed is of the same dimensionality as the observations. When the sample size is large, the importance sampling scheme becomes infeasible and/or inaccurate. A low-dimensional approximation model is developed to solve the numerical difficulties. We develop two spatial prediction methods: PBP (plug-in best predictor) and PBLUP (plug-in best linear unbiased predictor). Empirical results with simulated and real data show improved out-of-sample prediction performance of SHP modeling over stationary Gaussian process modeling.

We extend the single-realization model to SHP model with replicates. The spatial replications are modeled as independent realizations from a SHP model conditional on a common latent process. A simulation study shows substantial

improvements in parameter estimation and process prediction when replicates are available. In a example with real atmospheric deposition data, the SHP model with replicates outperforms the Gaussian process model in prediction by capturing the spatial volatilities.

Wenying Huang
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Summer 2008

ACKNOWLEDGEMENTS

First and foremost I would like to gratefully and sincerely thank my advisers Dr. F. Jay Breidt and Dr. Richard A. Davis, for their guidance, understanding and patient support during the past few years. They led me into the research field of spatial data modeling, provided inspirations and many valuable suggestions. Dr. Breidt formulated the idea on which this dissertation is based. He has always been available for discussions and guided me toward fruitful directions. Dr. Davis consistently provided sound advice and managed the research progress. My advisers are rigorous with scientific writing and helped me improve my skills considerably.

Thanks also go to the other members of my committee. Dr. Donald Estep provided motivations for us to start this research topic and continuously supported with advice and perspectives. Thank you to Dr. Wang for being my committee member. In addition, I would like to thank Dr. Jan Hannig for being my previous committee member and such a great teacher. Ke Wang has been my study and research pal for the whole time. I will always remember the hardship, frustration, encouragement and success we shared together and appreciate our friendship. I am thankful to all great professors, staff and fellow graduate students at Department of Statistics, Colorado State University. They made my graduate experience very joyful and unforgettable. Moreover, I acknowledge the support of NSF grant MSPA-CSE-0434354 "Novel A Posteriori Analysis of Ecological Models: The Carbon Cycle".

Special thanks go to my husband, Zhiqiang Cui. He always encourages me, celebrates my successes and comforts me through tough time. He always puts my priority as his first priority. I cannot imagine being here without him.

Finally and most importantly, I want to thank my parents and my brother, especially my Mom. She had always been there for me no matter how she suffered from the disease over decades. She had sacrificed much in her life to give me the opportunities that I have enjoyed in mine. I was deeply grieved when she passed away on my way toward Ph.D. I was, am and will always be, grateful for all her love she gave to me in life.

DEDICATION

To memory of my dear Mama .....

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION AND MOTIVATION

The development and application of spatial models to analyze spatial data have grown considerably during the past 20 years. Spatial data are geographically referenced and can be presented by 2-dimensional or 3-dimensional maps. Driven by new location technologies such as global positioning systems (GPS), there are huge amounts of spatial data collected in research areas such as meteorology, ecology, environmental health and so on. Statisticians and researchers in the corresponding areas seek methods to describe the trends and correlation structures among the data and make predictions of observations at unobserved locations. For example, the Northeast Fisheries Center of the National Marine Fisheries Service in Woods Hole, Massachusetts, samples the continental shelf off the Northeastern United States to estimate the abundance of sea scallops and other shellfish. The scallops data have been studied by many researchers to explore the spatial association and make predictions of abundance at unsampled locations; see Ecker and Heltshe (1994) and Ecker and Gelfand (1997) for more information about the data and modeling issues.

In this chapter, we will provide an overview of modern spatial data modeling techniques. Specifically, we will give more details on research and development of nonstationary spatial process models. In addition, we will introduce the stochastic volatility (SV) model and the spatial stochastic volatility (SSV) model, which motivates us to propose the stochastic heteroscedastic spatial (SHP) model. At the end of the chapter, we will outline this dissertation.

1.1 Spatial Data and Models

1.1.1 Introduction of spatial data modeling

A spatial process is a collection of random variables, $Y(\boldsymbol{x}) : \boldsymbol{x} \in D$, where D is a subset of d-dimensional Euclidean space $\mathcal{R}^d$. For a spatial data set, we assume it is a partial realization of $Y(\cdot)$, i.e., $Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n)$.

Types of spatial data

A fundamental problem of spatial data analysis is how to define the spatial region and locations of the entities being studied, which has crucial effects on the techniques which can be used for the analysis and on the conclusions which can be obtained. In much of the literature, spatial data is classified into three types by the nature of the spatial domain D:

- *Geostatistical data*, where the location $\boldsymbol{x}$ varies continuously over domain D, which means $Y(\boldsymbol{x})$ can be observed anywhere within D. It is also called point-referenced data. Examples of geostatistical data include observations on rainfall, temperature and air quality variables. In this dissertation, we will focus on geostatistical data modeling.

- *Lattice data*, where D is a fixed finite or countable set. The area being studied is partitioned into a regular or irregular lattice. Examples of lattice data include yields on agricultural trials, where the domain is partitioned into fields, and spatial econometrics data, where the domain is partitioned into census tracts or counties. Markov random field (MRF) models are often used to model lattice data.

- *Point pattern data*, where D is a pre-defined plane and the data are the co-ordinates $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ of event locations. That is, the random process $Y(\boldsymbol{x})$ is

simply equal to 1 for all locations $\boldsymbol{x}$ where the event occurs and 0 elsewhere. Examples of point pattern data include lightening strikes in certain areas and locations of a certain species of tree in the forest.

Stationary geostatistical data modeling

Gaussian processes are commonly used to model geostatistical data. Excellent references include Ripley (1981), Cressie (1993), Stein (1999) and Banerjee et al. (2003). A Gaussian process is a stochastic process $Y(\boldsymbol{x})$ for which every finite collection of random variables $(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n))$ has a multivariate normal distribution. Alternatively, a Gaussian process is a stochastic process for which any linear combination of any finite collection of $(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n))$ is normally distributed.

A spatial process is strictly stationary if for any set of locations $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) \in D$ and any location shift $\boldsymbol{h}$ such that $(\boldsymbol{x}_1 + \boldsymbol{h}, \boldsymbol{x}_2 + \boldsymbol{h}, ..., \boldsymbol{x}_n + \boldsymbol{h}) \in D$,

$$(Y(\boldsymbol{x_1}), ..., Y(\boldsymbol{x_n})) \stackrel{d}{\equiv} (Y(\boldsymbol{x_1} + \boldsymbol{h}), ..., Y(\boldsymbol{x_n} + \boldsymbol{h})), \tag{1.1}$$

i.e., the probability distribution at any set of fixed positions is the same for that of the positions after a common translation. As a result, moments such as the mean and variance, if they exist, do not depend on location.

Second-order stationarity is less stringent than strict stationarity. It requires constant mean and that the covariance between any two locations only depends on the difference of the two locations, i.e., $\mathrm{E}(Y(\boldsymbol{x})) \equiv \mu$ and $\mathrm{Cov}(Y(\boldsymbol{x} + \boldsymbol{h}), Y(\boldsymbol{x})) = C(\boldsymbol{h})$ for all $\boldsymbol{h} \in \mathcal{R}^d$ such that $\boldsymbol{x}+\boldsymbol{h}$ and $\boldsymbol{x}$ both lie in D. For a Gaussian process, the second-order stationarity is equivalent to strict stationarity because its distribution is completely characterized by the mean and covariance structure. There is a third type of stationarity called intrinsic stationarity, under which $\mathrm{E}(Y(\boldsymbol{x}+\boldsymbol{h})-Y(\boldsymbol{x})) = 0$ and

$$\mathrm{E}(Y(\boldsymbol{x} + \boldsymbol{h}) - Y(\boldsymbol{x}))^2 = \mathrm{Var}(Y(\boldsymbol{x} + \boldsymbol{h}) - Y(\boldsymbol{x})) = 2\gamma(\boldsymbol{h}). \tag{1.2}$$

That is, the left-hand side of (1.2) only depends on $\boldsymbol{h}$ no matter how $\boldsymbol{x}$ varies. In this case, the process is said to be intrinsically stationary, and $2\gamma(\boldsymbol{h})$ is called the variogram while $\gamma(\boldsymbol{h})$ is called the semivariogram. The intrinsic stationarity is defined through the first and second moments of the difference $Y(\boldsymbol{x}+\boldsymbol{h})-Y(\boldsymbol{x})$. It does not specify the distribution of $Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n)$ and therefore provides no likelihood. It is easy to derive the relationship between the variogram and the covariance function. Given C, a covariance function of a stationary spatial process, we can recover γ by $\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h})$. If the spatial process is ergodic, $C(\boldsymbol{h}) \to 0$ as $||\boldsymbol{h}|| \to \infty$, we have

$$C(\boldsymbol{h}) = C(\boldsymbol{0}) - \gamma(\boldsymbol{h}) = \lim_{||\boldsymbol{u}||\to\infty} \gamma(\boldsymbol{u}) - \gamma(\boldsymbol{h}).$$

A stationary process is isotropic if the covariance between $Y(\boldsymbol{x})$ and $Y(\boldsymbol{x}+\boldsymbol{h})$ is a function solely of $||\boldsymbol{h}||$; otherwise, it is anisotropic. Isotropic processes are popular because they are simple, easily-interpreted and there are a number of nice parametric isotropic covariance functions.

The classic Gaussian process model has the form

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + W(\boldsymbol{x}), \tag{1.3}$$

where $\mu(\boldsymbol{x})$ is a deterministic large-scale trend and usually takes the form of $\boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta}$, where $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), ..., g_p(\boldsymbol{x}))$ are known regression functions and $\boldsymbol{\beta} = (\beta_1, ...\beta_p)^T$ is a vector of unknown regression coefficients. The spatially correlated error process $W(\boldsymbol{x})$ is usually assumed to be a stationary Gaussian process with mean 0, variance σ^2 and (isotropic) correlation function ρ.

Equation (1.3) is the basic model for a spatial surface. It is common practice to model the *observed* geostatistical data by

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + W(\boldsymbol{x}) + \epsilon(\boldsymbol{x}), \tag{1.4}$$

where $\epsilon(\boldsymbol{x})$ is called nugget and is usually assumed to be iid normally distributed. Its variance corresponds to the height of the jump discontinuity of the covariance at the origin. The nugget is considered to come from physical measurement error, i.e., the replication variability and microscale variability which accounts for possible model misspecification at a very fine scale.

Correlation functions

In equation (1.3), the correlation function ρ and the variance σ^2 characterize the distributional properties of the spatially correlated error process $W(\boldsymbol{x})$. From *Bochner's Theorem*, ρ is a valid correlation function if and only if it is the characteristic function of a symmetric random variable. In practice, parametric correlation functions are usually used due to the simplicity, estimability and nice statistical inference properties. For an isotropic spatial process, exponential, Gaussian, Matérn and spherical correlation functions are commonly used.

Most isotropic correlation functions have a range parameter, denoted by ϕ. The correlation function is usually monotonic in ϕ. But the formulations of the correlation function in terms of ϕ are different in different literatures. Throughout this dissertation, the correlation increases with range parameter ϕ decreasing. We define several popular correlation functions as follows:

- Gaussian: $\rho(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\phi\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$,
- Exponential: $\rho(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\phi\|\boldsymbol{x} - \boldsymbol{x}'\|)$,
- Spherical: $\rho(\boldsymbol{x}, \boldsymbol{x}') = (1 - 1.5\phi\|\boldsymbol{x} - \boldsymbol{x}'\| + 0.5(\phi\|\boldsymbol{x} - \boldsymbol{x}'\|)^3)I_{\{\|\boldsymbol{x}-\boldsymbol{x}'\|<1/\phi\}}$.

Different correlation functions can have very close correlation plots by adjusting their range parameters. But even so, the underlying sample paths have different smoothness properties. Referring to Lindgren (2004) Chapter 2, denoting the covariance function as $r(t)$, a Gaussian process satisfying $r(t) = r(0) - C|t|^\alpha + o(|t|^\alpha)$ as

$t \to 0$ for some $0 < \alpha < 2$ has continuous sample path almost surely. Furthermore, a Gaussian process is continuously differentiable if $-r''(t) = -r''(0) - C|t|^{\beta} + o(|t|^{\beta})$ with $0 < \beta < 2$. According to Solak et al. (2003), because differentiation is a linear operation, the derivative of a Gaussian process remains a Gaussian process. The covariance function of the derivative Gaussian process is the second derivative of the covariance function for the Gaussian process. Therefore, the Gaussian correlation function leads to very smooth realizations of the spatial process because it is an analytic function. If using exponential correlation function, in which case $\alpha = 1$, the sample path is continuous but not differentiable. For more theoretical discussions about the differentiability of a spatial process, refer to Lindgren (2004) and Stein (1999). It is practical to prespecify the form of the correlation function based on the preassumed smoothness of the underlying spatial process. Some model selection criterion, such as AIC, can also be employed.

Matérn (1986) introduced a more flexible family of correlation functions given by

$$\rho(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\phi\|\boldsymbol{x} - \boldsymbol{x}'\|)^{\nu}\mathcal{K}_{\nu}(2\sqrt{\nu}\phi\|\boldsymbol{x} - \boldsymbol{x}'\|),$$

where $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of order ν. The smoothness parameter ν controls the differentiability of the sample path. Let $\lceil \nu \rceil$ denote the integer ceiling of ν, i.e., the smallest integer that is greater than or equal to ν. The functions drawn from a Gaussian process with Matérn correlation have almost surely continuously differentiable sample paths of order $(\lceil \nu \rceil - 1)$. When $\nu = 1/2$, the Matérn correlation function becomes $\exp(-\sqrt{2}\phi\|\boldsymbol{x} - \boldsymbol{x}'\|)$, which is an exponential correlation function with range parameter $\sqrt{2}\phi$. As $\nu \to \infty$, $\rho(\boldsymbol{x}, \boldsymbol{x}')$ converges to $\exp(-\phi^2\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, the Gaussian correlation function with range parameter ϕ^2. Stein (1999) provides a detailed discussion about the spectra of Matérn correlation and its properties.

Kriging

The classical approach to spatial prediction is called kriging. Kriging is simply a special case of optimal linear prediction applied to spatial random processes. This method is named after a South African mining engineer, D. G. Krige, who developed the technique in an attempt to more accurately predict ore reserves. Over the past several decades kriging has become a fundamental tool in the field of geostatistics.

Kriging refers to making inference on unobserved values of the spatial process $Y(\boldsymbol{x})$ based on the observed data $(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n))$. The kriging predictor at a point $\boldsymbol{x}_0$, denoted by $\hat{Y}(\boldsymbol{x}_0)$, satisfies the following conditions:

- it is linear in the observations: $\hat{Y}(\boldsymbol{x}_0) = \sum_{i=1}^{n} l_i Y(\boldsymbol{x}_i)$,
- it is unbiased: $\mathrm{E}(\hat{Y}(\boldsymbol{x}_0)) = Y(\boldsymbol{x}_0)$,
- it minimizes the mean-squared prediction error (MSPE), $\mathrm{E}(Y(\boldsymbol{x}_0) - \sum_{i=1}^{n} l_i Y(\boldsymbol{x}_i))^2$ over l_i's.

Based on different assumptions on the spatial mean μ, it is customary to further define simple kriging, ordinary kriging and universal kriging. Simple kriging assumes that μ is zero. In terms of a Gaussian process, where $Y(\boldsymbol{x}_0)$ and $(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n))$ are jointly multivariate normal, the simple kriging predictor is just the conditional expectation of $Y(\boldsymbol{x}_0)$ given $(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2), ..., Y(\boldsymbol{x}_n))$, i.e.,

$$\hat{Y}(\boldsymbol{x}_0) = \mathrm{Cov}(Y(\boldsymbol{x}_0), \boldsymbol{Y}(\boldsymbol{x}))\mathrm{Var}(\boldsymbol{Y}(\boldsymbol{x}))^{-1}\boldsymbol{Y}(\boldsymbol{x}). \tag{1.5}$$

For ordinary kriging, one assumes that μ is an unknown constant. We restrict $\sum_{i=1}^{n} l_i = 1$ to guarantee unbiasedness. It turns out that ordinary kriging results in the best linear unbiased estimator (BLUE). Universal kriging refers to the situation in which μ is assumed to be an (unknown) linear combination of known functions that depend on the locations. The universal kriging predictor has the form

$$\hat{Y}(\boldsymbol{x}_0) = \boldsymbol{g}(\boldsymbol{x}_0)^T\hat{\boldsymbol{\beta}}_{GLS} + \mathrm{Cov}(Y(\boldsymbol{x}_0), \boldsymbol{Y}(\boldsymbol{x}))\mathrm{Var}(\boldsymbol{Y}(\boldsymbol{x}))^{-1}(\boldsymbol{Y}(\boldsymbol{x}) - \boldsymbol{G}(\boldsymbol{x})\hat{\boldsymbol{\beta}}_{GLS}), \tag{1.6}$$

where $\boldsymbol{g}(\boldsymbol{x}_0)$ is a $p \times 1$ regression function and $G(\boldsymbol{x}) = (\boldsymbol{g}(\boldsymbol{x}_1), ..., \boldsymbol{g}(\boldsymbol{x}_n))^T$. Here $\hat{\boldsymbol{\beta}}_{GLS}$ refers to the generalized least square estimates for the regression coefficients . Cressie (1993) has shown the details on obtaining the prediction equations for each case.

1.1.2 Nonstationary spatial process modeling

Isotropy and stationarity are often useful as working assumptions for modeling spatial data. But it is desirable to allow anisotropy and spatial heterogeneity in many applications. Several anisotropic and nonstationary spatial processes have been developed.

Anisotropy refers to the case in which the spatial correlation depends upon the distances and directions between spatial locations rather than merely distances. Anisotropy is generally difficult to handle. But there is a special case, known as geometric anisotropy, that is tractable. In the isotropic case, iso-correlation contours are spherical. When a linear transformation of the coordinate system applies, the spherical contours are changed to elliptical contours. In practice, it is customary to define geometric anisotropy by two parameters, anisotropy angle ψ_A and anisotropy ratio ψ_R. The transformation consists in multiplying the original coordinates $\boldsymbol{x}$ by a rotation matrix R_A and a shrinking matrix T_A, as follows,

$$\boldsymbol{x}_A = \boldsymbol{x} R_A T_A, \tag{1.7}$$

where $T_A = \text{Diag}(1, \psi_R)$ and

$$R_A = \begin{bmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{bmatrix}.$$

Then the geometric anisotropic correlation is given by

$$\rho(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \rho_0(||\boldsymbol{x}_A - \boldsymbol{x}'_A||; \boldsymbol{\theta}), \tag{1.8}$$

provided that ρ_0 is an isotropic correlation function with parameters $\boldsymbol{\theta}$. This linear transformation can be reversed to correct geometric anisotropy into isotropy. A geometric anisotropic spatial process is stationary because $\rho_0(||\boldsymbol{x}_A - \boldsymbol{x}'_A||; \boldsymbol{\theta}) = \rho_0(||(\boldsymbol{x} - \boldsymbol{x}')R_A T_A||; \boldsymbol{\theta})$.

A variety of approaches for nonstationary spatial process modeling have been developed over the last decade. Research regarding their properties, applications and improvements are ongoing. We summarize here in brief several of these approaches.

Multiplicative model

One of the approaches to model nonstationarity is through scaling (Banerjee et al. (2003)). Assuming that $W(\boldsymbol{x})$ is a mean 0, variance 1 stationary process with correlation function ρ, a nonstationary process can be constructed by $\sigma(\boldsymbol{x})W(\boldsymbol{x})$. Here, $\sigma(\boldsymbol{x})$ is a pre-specified deterministic function. A customary choice is $\sigma(\boldsymbol{x}) = g(u(\boldsymbol{x}))\sigma$ where $u(\boldsymbol{x})$ is a suitable positive covariate and g is a strictly increasing positive function. But it is hard to specify functions u and g that are adapted to the data and physically meaningful.

Space deformation

Geometric anisotropy is obtained by linear transformation of spatial coordinate systems of an isotropic process, as shown in equations (1.7) and (1.8). Sampson and Guttorp (1992) generalize this transformation idea by permitting a one-to-one nonlinear mapping over the geographic domain of interest. They refer to the plane of geographical coordinates $\boldsymbol{x}$ of the observed locations as the G plane and the transformed plane as the D (stands for dispersion) plane, where the stationarity and, in fact, isotropy holds. The variance (dispersion) between two locations can be written as

$$\mathrm{Var}(Y(\boldsymbol{x}) - Y(\boldsymbol{x}')) = g(||f(\boldsymbol{x}) - f(\boldsymbol{x}')||), \tag{1.9}$$

where f is a nonlinear transform function and g is a valid isotropic variogram in D space. Both of f and g are unknown and need to be estimated. It is desirable that f be bijective, otherwise there exist two different points in G space for which the correlation of the Y process is 1. Perrin and Meiring (1999) investigated some identifiability issues of this model. They showed that the transformation f is unique up to translation, rotation and reflection about a line or any combination of these transformations, provided that the correlation function is strictly decreasing or differentiable.

Sampson and Guttorp (1992) suppose that a random function is sampled repeatedly at a fixed number of sampling locations so that the point estimates of spatial covariance among the sampling stations can be computed. They smooth the sample covariance in a nonparametric way. The model is implemented through two steps. First, using nonmetric multidimensional scaling (MDS), they compute a two-dimensional representation of the sampling stations for which a monotone function of interpoint distances δ_{ij} approximates the spatial dispersions. MDS transforms the problem into one for which the covariance structure, expressed in terms of spatial dispersions, is stationary and isotropic. Second, they compute thin-plate splines to provide smooth mappings of the geographic representation of the sampling stations into their MDS representation. The above methods are based on computationally intensive algorithms.

This approach has been applied and extended in a variety of ways. Mardia and Goodall (1993) model multivariate spatial fields, assuming a Kronecker structure for the space$\times$time covariance structure and find the estimated D-plane locations and the parameters of the covariance function by solving the likelihood equations. Smith (1996) implements maximum likelihood to estimate the parameters of the Matérn correlation, the coefficients of radial basis functions and the components of the spatial deformation f.

Without stationarity and/or any parametric assumption, it is impossible to estimate the spatial covariance (dispersion) by using a single realization of the underlying spatial process. The deformation approach requires independent replications of the spatial process in order to get an estimated (sample) covariance matrix. In practice, we seldom have independent replications of a spatial process. Typically repeated measurements across time are collected. The approximately independent and identically distributed (iid) observations are obtained by removing the trend and seasonality first. Some researchers, however, prefer spatio-temporal modeling to analyze the spatial correlation and temporal evolution simultaneously.

Convolution methods

Higdon et al. (1998) propose an attractive and powerful way of introducing nonstationarity. It is based on a moving average specification of a Gaussian process. Any stationary Gaussian process $Y(\boldsymbol{x})$ having correlation

$$\rho(\boldsymbol{h}) = \int_{\mathcal{R}^2} k(\boldsymbol{x})k(\boldsymbol{x}-\boldsymbol{h})d\boldsymbol{x},$$

where $k(\cdot)$ is a smoothing kernel, can be expressed as the convolution of a Gaussian white noise process $\omega(\boldsymbol{x})$ (Brownian motion) by

$$Y(\boldsymbol{x}) = \int_{\mathcal{R}^2} k(\boldsymbol{u}-\boldsymbol{x})\omega(\boldsymbol{u})d\boldsymbol{u},$$

where $\omega(\boldsymbol{u})d\boldsymbol{u}$ corresponds to $d\omega(\boldsymbol{u})$ in the definition of an Itô integral. To account for nonstationarity, the smoothing kernel is allowed to depend on the spatial location $\boldsymbol{x}$. Then, the nonstationary process is represented by

$$Y(\boldsymbol{x}) = \int_{\mathcal{R}^2} k_{\boldsymbol{x}}(\boldsymbol{u}-\boldsymbol{x})\omega(\boldsymbol{u})d\boldsymbol{u},$$

where $k_{\boldsymbol{x}}(\cdot)$ is a bivariate normal kernel with center at $\mathbf{0}$ and with covariance $\Sigma_{\boldsymbol{x}}$ that varies spatially. The correlation between two points $\boldsymbol{x}$ and $\boldsymbol{x}'$ becomes

$$\rho(\boldsymbol{x},\boldsymbol{x}') = \int_{\mathcal{R}^2} k_{\boldsymbol{x}}(\boldsymbol{u}-\boldsymbol{x})k_{\boldsymbol{x}'}(\boldsymbol{u}-\boldsymbol{x})d\boldsymbol{u}.$$

Higdon et al. (1998) provided a closed-form expression for $\rho(\boldsymbol{x}, \boldsymbol{x}')$. They developed a Bayesian hierarchical model which can incorporate uncertainties in the resulting inference. The problem of prior formulations was left for further discussion since the nature of the models does not give rise to any inviting conjugate formulations. This model fitting does not require repeated realizations from the spatial process. The non-stationary spatial dependence is explained through a constructive "process-convolution" approach, which ensures that the resulting covariance structure is valid.

The convolution method of Higdon et al. (1998) varies parameters of the kernel function spatially. Fuentes and Smith (2001) propose an alternative convolution approach to model nonstationarity. They vary the stationary processes instead of the kernel, i.e.,

$$Y(\boldsymbol{x}) = \int_D k(\boldsymbol{x} - \boldsymbol{u}) Y_{\boldsymbol{\theta}(\boldsymbol{u})}(\boldsymbol{x}) d\boldsymbol{u},$$

where k is a fixed kernel function and $Y_{\boldsymbol{\theta}(\boldsymbol{u})}$ is a family of independent stationary Gaussian processes indexed by parameter $\boldsymbol{\theta}(\boldsymbol{u})$, which varies substantially over the whole space D. The covariance of $Y(\cdot)$ is a convolution of the local covariances $C_{\boldsymbol{\theta}(\boldsymbol{u})}(\boldsymbol{x}, \boldsymbol{x}')$,

$$C(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \int_D k(\boldsymbol{x} - \boldsymbol{u}) k(\boldsymbol{x}' - \boldsymbol{u}) C_{\boldsymbol{\theta}(\boldsymbol{u})}(\boldsymbol{x}, \boldsymbol{x}') d\boldsymbol{u}.$$

If k is a sharply peaked kernel function and $\boldsymbol{\theta}(\boldsymbol{u})$ varies slowly with $\boldsymbol{u}$, this has the property that for $\boldsymbol{x}$ near $\boldsymbol{u}$, the process "looks like" a stationary process with parameter $\boldsymbol{\theta}(\boldsymbol{u})$. On the other hand, since $\boldsymbol{\theta}(\boldsymbol{u})$ may vary substantially over the whole space, it also allows significant nonstationarity. Fuentes and Smith (2001) discussed model fitting through exact and approximate likelihood maximization, and proposed a hierarchical Bayesian approach to allow predictive inference.

Most of the nonstationary approaches outlined in this section have certain limitations. Since there is no universally accepted approach so far, it is desirable to develop new methods that are capable of modeling a wide variety of spatial processes and are attractively interpretable.

1.2 Stochastic Volatility Model

1.2.1 Stochastic volatility in time series data analysis

In time series analysis, numerous models have been proposed to capture changing variance and covariance structure. Autoregressive conditional heteroscedasticity (ARCH) is one such model, in which the conditional variance is modeled as a deterministic function of the available information, i.e., the past observations. Stochastic volatility (SV) attempts to achieve the same objective as the ARCH. In an SV model, the conditional variance is modeled as a latent stochastic process. The most popular stochastic volatility model from Taylor (1986) assumes

$$y_t = \epsilon_t \exp(h_t/2), \quad h_t = \gamma_0 + \gamma_1 h_{t-1} + \eta_t, \tag{1.10}$$

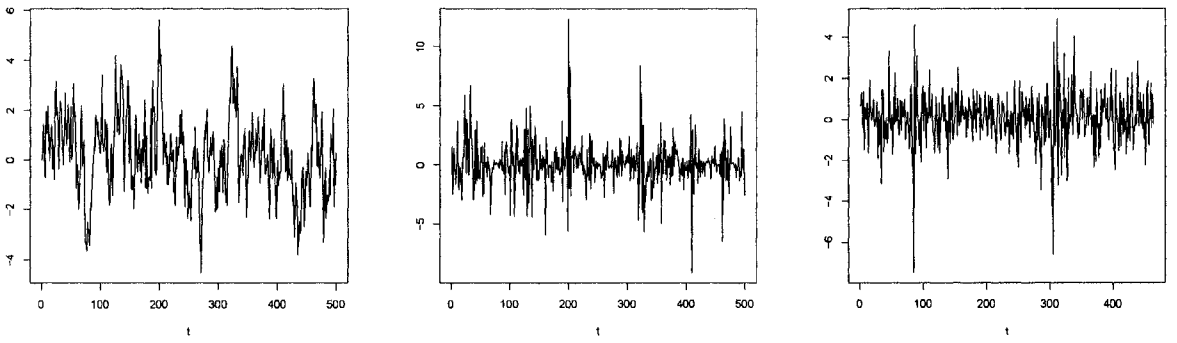
where ϵ_t and η_t are assumed to be independent Gaussian white noise with variances 1 and σ_η^2, respectively. The log-volatility h_t is unobserved but can be estimated using the observations. It is referred to as a latent process. The properties of the SV model are easy to derive. Particularly, it is worthwhile to mention that the kurtosis of the SV model is $3\exp(\sigma_\eta^2)$, which is greater than 3, the normal distribution kurtosis. This means that the SV model has heavier tails than the corresponding normal distribution. The dynamic properties of the SV model are conveniently investigated by taking log of y_t^2, i.e.,

$$\log y_t^2 = h_t + \log \epsilon_t^2, \quad h_t = \gamma_0 + \gamma_1 h_{t-1} + \eta_t. \tag{1.11}$$

This is a linear process, which adds the iid $\log \epsilon_t^2$ error to the AR(1) process h_t. Consequently $\log y_t^2$ follows an ARMA(1, 1) process. The random variable $\log \epsilon_t^2$ is log chi-square distributed and it has mean -1.27 and variance 4.93.

We plot a simulated time series from a SV model generated by equation (1.10) and a real time series in Figure 1.1. The "nonstationary" features in the series

are clearly seen in that values large in absolute value tend to be clustered, corresponding to high volatility, and values small in absolute value tend to be clustered, corresponding to low volatility.



(a) (b) (c)

Figure 1.1: SV time series data plots. Panel (a) shows the h_t process of a simulated SV model. Panel (b) shows the y_t process of the simulated SV model. We take $\gamma_0 = 0, \gamma_1 = 0.8, \sigma_\eta^2 = 0.1, h_0 = 0$ for the SV simulation in panels (a) and (b). Panel (c) shows the percentage daily returns of the Dow Jones Industrial Index for the period July 1st, 1997, through April 9th, 1999.

The estimation of the SV model is difficult because it is not immediately clear how to evaluate the likelihood as the distribution of $y_t|y_{t-1}$ is specified implicitly through a latent process. Researchers have developed many methods to estimate the SV model and the efforts are still going on. Shephard (1996) provides a comprehensive reference about estimation using generalized method of moments (GMM), quasi-likelihood, importance sampling and several Markov Chain Monte Carlo (MCMC) approaches. We summarize some important estimation methods here.

Quasi-likelihood estimation

Since $\log \epsilon_t^2$ is iid, $\log y_t^2$ can be written as a non-Gaussian linear state-space model (see Brockwell and Davis (2002) Chapter 8):

$$\begin{aligned} \text{observation equation: } & \log y_t^2 = h_t + \log \epsilon_t^2, \\ \text{state equation: } & h_t = \gamma_0 + \gamma_1 h_{t-1} + \eta_t. \end{aligned} \tag{1.12}$$

The Kalman filter can be used to provide the best linear unbiased estimator of h_t given $(\log y_1^2, ..., \log y_{t-1}^2)$. The parameter $\boldsymbol{\theta} = (\sigma_\eta^2, \gamma_0, \gamma_1)$ can be estimated using quasi-likelihood (Harvey et al. (1994))

$$lq(\boldsymbol{\theta}, \boldsymbol{y}) = -\frac{1}{2}\sum_{t=1}^{T} \log F_t - \frac{1}{2}\sum_{t=1}^{T} v_t^2/F_t, \tag{1.13}$$

where v_t is the one-step-ahead prediction error and F_t is the corresponding mean squared error from the Kalman filter. If (1.12) had been a Gaussian state-space model then (1.13) would be the exact likelihood. But as $\log \epsilon_t^2$ follows a log chi-square distribution, (1.13) is referred as quasi-likelihood and can be used to provide consistent parameter estimates and asymptotically normal inferences.

Besides parameter estimation, we are interested in estimating and predicting the latent process h_t. Typically the posterior mode $\boldsymbol{h}|\boldsymbol{y}$ is a good representation. Durbin and Koopman (1993), along with others, proposed to recursively solve the linearized approximation to $\partial l(\boldsymbol{h}, \boldsymbol{y})/\partial \boldsymbol{h} = 0$.

Likelihood and importance sampling

The exact likelihood can be computed by integrating out the latent process h_t,

$$f(y_1, ..., y_T) = \int f(y_1, ..., y_T|\boldsymbol{h}) f(\boldsymbol{h}) d\boldsymbol{h}. \tag{1.14}$$

As this integral has no closed form it has to be computed numerically. Monte Carlo integration is a direct approach, i.e., draw a large number of realizations of $\boldsymbol{h}$ from its unconditional distribution, then approximate (1.14) by $(1/N)\sum_{j=1}^{N} f(y_1, ..., y_T|\boldsymbol{h}_j)$. This estimation usually performs poorly especially when the number of observations T is large. An importance sampling strategy is often used to improve the efficiency and accuracy of a Monte Carlo integral approximation. Given an importance density $g(\cdot)$, rewrite (1.14) as

$$f(\boldsymbol{y}) = \int \frac{f(\boldsymbol{y}|\boldsymbol{h}) f(\boldsymbol{h})}{g(\boldsymbol{h}|\boldsymbol{y})} g(\boldsymbol{h}|\boldsymbol{y}) d\boldsymbol{h}. \tag{1.15}$$

Danielsson and Richard (1993) developed an Accelerated Gaussian Importance Sampler (AGIS) which recursively improves its performance, converging towards the optimal g. Their method is highly efficient for very high-dimensional integration of density and likelihood functions, as illustrated by an application to a first-order SV model for daily stock returns with $N = 1000$ Monte Carlo replications for $T = 2022$ observations. The algorithms of Danielsson and Richard (1993) are quite involved even for the simplest model.

Markov chain Monte Carlo (MCMC)

Early work on using MCMC for the SV model focused on "single move" algorithms, drawing h_t individually, ideally from its conditional distribution $h_t|\boldsymbol{h}_{\setminus t}, \boldsymbol{y}$, where $\boldsymbol{h}_{\setminus t}$ refers to all elements of $\boldsymbol{h}$ except h_t. For the first-order SV model, we have

$$f(h_t|\boldsymbol{h}_{\setminus t}, \boldsymbol{y}) = f(h_t|h_{t-1}, h_{t+1}, y_t) \propto f(y_t|h_t)f(h_{t+1}|h_t)f(h_t|h_{t-1}). \tag{1.16}$$

Since the normalization constant is unknown, it is impossible to sample directly from (1.16). Some elaborate rejection sampling methods have been developed to implement Gibbs sampling. For example, it is easy to see that

$$\log f(h_t|h_{t-1}, h_{t+1}, y_t) = \text{const} - \frac{1}{2}h_t - \frac{1}{2\sigma_t^2}(h_t - h_t^*)^2 - \frac{1}{2}y_t^2 \exp(-h_t), \tag{1.17}$$

where σ_t^2 and h_t^* are functions in terms of $h_{t-1}, h_{t+1}, \sigma_\eta^2, \gamma_0$ and γ_1. As $f(h_t|h_{t-1}, h_{t+1}, y_t)$ is a log-concave density function, the adaptive rejection sampling proposed by Wild and Gilks (1993) can be used here.

This "single move" sampler may converge slowly especially when γ_1 is close to 1, i.e., $\boldsymbol{h}|\boldsymbol{y}$ are highly correlated. To overcome this problem, one can work with blocks or a "multi-move" sampler instead of sampling $h_t|\boldsymbol{h}_{\setminus t}, \boldsymbol{y}$ one at a time. We introduce the idea by de Jong and Shephard (1995). They work on the linear state-space model (1.12). They approximate $\log \epsilon_t^2$ by a mixture of normals so that

$$\log \epsilon_t^2|(w_t = j) \sim \mathrm{N}(\mu_j, \sigma_j^2), \quad j = 1, ..., J. \tag{1.18}$$

Here the $\{w_t\}$ are iid with $P(w_t = j) = \pi_j$. Kim et al. (1998) selected μ_j, σ_j^2, π_j for $j = 1, ..., 7$ to match the moments and various other features of this approximation to the truth, i.e., log chi-square distribution with one degree of freedom.

The advantage of this representation of the model is that conditionally on w, the state-space (1.12) is now Gaussian. One can simultaneously sample $(\boldsymbol{h}|\boldsymbol{y}, w)$ by use of the Gaussian simulation smoother proposed by de Jong and Shephard (1995). This approach avoids the correlation in the h_t process and therefore expedites the convergence. But (1.18) is only an approximation. It is a challenging problem to come up with multi-move algorithms without transforming the model.

1.2.2 Regression with input-dependent noise

In time series, the stochastic volatility models describe the time-varying variance by use of a latent stochastic process. In the regression literature, there is some similar treatment. For a classic regression model, the response can be described by a deterministic function of the inputs, together with additive Gaussian noise having constant variance. In many applications a more realistic model would allow the noise variance itself to depend on the input variables. Bishop and Qazaz (1997) propose such a regression model and apply Bayesian methodology for inference. They illustrate their algorithm by a simulated example in which the noise variance depends on x^2. It is natural to extend the deterministic dependence relationship between noise variance and input variable to a latent stochastic process model. Goldberg et al. (1998) suggest modeling the noise variance using a Gaussian process, similar to the SV model structure. In their paper, the input vector is denoted by $\boldsymbol{x}$ and the observed output vector is denoted by $\boldsymbol{t}$. Given n data points $\mathcal{D} = ((\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), ..., (\boldsymbol{x}_n, t_n))$ and assuming that any offset or trend in the data

has been removed, i.e., $\mu(\boldsymbol{x}) \equiv 0$, the Gaussian process model is given by

$$\begin{aligned} t_i &= y_i + \epsilon_i, \\ \boldsymbol{y} &\sim \mathrm{N}(0, C_y(\boldsymbol{x}, \boldsymbol{x})), \\ \epsilon_i &\sim N(0, r_i), \\ r_i &= \exp(z_i), \\ \boldsymbol{z} &\sim \mathrm{N}(0, C_z(\boldsymbol{x}, \boldsymbol{x})), \end{aligned} \tag{1.19}$$

where

$$C_y(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = v_y \exp\left(-\frac{1}{2}\sum_{i=1}^{d} w_{yl}(x_l^{(i)} - x_l^{(j)})^2\right) + J_y\delta(i,j), \tag{1.20}$$

and v_y specifies the overall y-scale and $w_{yl}^{-1/2}$ is the length-scale associated with the lth coordinate. So a geometric anisotropic Gaussian covariance function is adopted here. Note that J_y is a "jitter" term, which is added to prevent ill-conditioning of the covariance matrix of the outputs. Typically J_y is given a small value, e.g., 10^{-6}. C_z has the same form as (1.20) except different parameter values.

In the estimation and prediction procedure, it is vital to sample the noise rates $\boldsymbol{z}$ from the posterior distribution $p(\boldsymbol{z}|\boldsymbol{t})$. As this is quite difficult, they instead sample from $p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{t})$ and ignore $\boldsymbol{y}$ values to get the sample of $\boldsymbol{z}$. They use Gibbs sampling to sample from $p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{t})$ by alternatively sampling from $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{t})$ and $p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{t})$. It is clear that

$$p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{z})p(\boldsymbol{z}). \tag{1.21}$$

It is not easy to sample $\boldsymbol{z}$ as a vector. They propose to sample from $p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{y}, \boldsymbol{t})$ using rejection sampling. They claim that the average rejection rate is approximately two-thirds. For making predictions for t^*, an output at an unobserved point $\boldsymbol{x}^*$, they propose to use Monte Carlo integration. For sampling from the posteriors on parameters, they apply Metropolis algorithms within the Gibbs sampling.

In brief, model (1.19) allows heteroscedastic variance structure and has attractive probabilistic properties. The model is written in a hierarchical way and it is natural to use MCMC method for estimation and prediction. But sampling the latent process $\boldsymbol{z}$ is difficult. The one-at-a-time rejection sampling is not very efficient. The prior specification for parameters in Bayesian approach is also a delicate issue.

1.2.3 Spatial stochastic volatility model for lattice data

The ideas of SV in time series can be applied to the spatial context. Spatial heteroscedasticity for lattice data has been studied by Yan (2007).

Lattice data are collected over a finite set of regular or irregular geographic units, so that we have measurements $Y_1, ..., Y_n$ associated with units $1, ..., n$. Instead of working directly on the joint distribution of $Y_1, ..., Y_n$, the lattice data are usually modeled through the conditional distribution of Y_i given Y_i's in a neighborhood N_i of the unit i. A Markov Random Field (MRF), i.e.,

$$p(y_i|y_j, j \neq i) = p(y_i|y_j, j \in N_i) \tag{1.22}$$

is a special case in which a joint distribution is completely determined from the locally-specified conditionals. The Hammersley-Clifford Theorem (Besag (1974)) proves that if we have a MRF, then the joint distribution uniquely determined by (1.22) is a Gibbs distribution. Geman and Geman (1984) state that if we have a joint Gibbs distribution, then we have a MRF.

Now we introduce the popular conditional autoregressive (CAR) model, proposed by Besag et al. (1991),

$$\begin{aligned} &Y_i = \mu + \phi_i + \epsilon_i, \\ &\phi_i|\phi_{j\neq i} \sim N\left(\frac{\sum_{j\neq i} b_{ij}\phi_j}{\sum_{j\neq i} b_{ij}}, \frac{\sigma_\phi^2}{\sum_{j\neq i} b_{ij}}\right), \\ &\epsilon_i \sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{1.23}$$

where ϕ_i is conditionally specified and follows a Gaussian Markov Random Field. The b_{ij}'s are known weights with $b_{ij} = b_{ji}$. A common choice for b_{ij} is 1 if units i and j are adjacent and 0 otherwise.

In Yan (2007), a spatial stochastic volatility (SSV) component was introduced into the CAR model. The SSV model is given as follows:

$$\begin{aligned} Y_i &= \mu + \phi_i + \epsilon_i, \\ \phi_i|\phi_{j\neq i} &\sim N\left(\frac{\sum_{j\neq i} b_{ij}\phi_j}{\sum_{j\neq i} b_{ij}}, \frac{\sigma_\phi^2}{\sum_{j\neq i} b_{ij}}\right), \\ \epsilon_i &\sim N(0, \exp(\mu_h + h_i)), \\ h_i|h_{j\neq i} &\sim N\left(\frac{\sum_{j\neq i} c_{ij}h_j}{\sum_{j\neq i} c_{ij}}, \frac{\sigma_h^2}{\sum_{j\neq i} c_{ij}}\right). \end{aligned} \tag{1.24}$$

By introducing the latent spatial process $\boldsymbol{h}$ in the log volatility, the error variance ϵ_i has a spatially clustered structure instead of iid $N(0, \sigma_\epsilon^2)$. The latent process $\boldsymbol{h}$ follows another Gaussian Markov random field and σ_h^2 is the variance parameter for $\boldsymbol{h}$. The weights b_{ij}'s and c_{ij}'s are prespecified and they can be different in general. To make the parameters μ and μ_h identifiable, Yan (2007) put the constraints $\sum_{i=1}^n \phi_i = 0$ and $\sum_{i=1}^n h_i = 0$. As the SSV model (1.24) is presented in a hierarchical structure, the author uses Bayesian methods for statistical inference.

The SSV process brings more flexibility in modeling lattice data. As stated in Yan (2007), in the case of volatility clustering, researchers may want to detect "hot spots" of volatilities, and monitor these spots more closely in the future. This can be done naturally by the SSV model. When prediction is of interest, the SSV model may be preferred to constant volatility models by allowing the variance to vary spatially. SSV can also be an approximation of a more complex model and can pick up the effects of omitted variables.

1.2.4 Non-Gaussian Bayesian geostatistical modeling

In Section 1.2.1 through Section 1.2.3, we have introduced stochastic volatility model in time series and its extensions in regression and spatial lattice data modeling. As a matter of fact, there have been also some geostatistical models developed in which the variance is modeled through a random process instead of a constant value. For spatial data with independent replications, Damian et al. (2001) and Schmidt and O'Hagan (2003) proposed to specify the location-dependent variance when formulating the deformation model (Sampson and Guttorp (1992)) in fully Bayesian frameworks. The former applied thin-plate splines for deformation and modeled the temporal variances as a random field by use of a Gaussian process prior. The latter defined the mapping between G-plane and D-plane (see Section 1.1.2 for definitions) by an unknown function $\boldsymbol{d}(\cdot)$ and a Gaussian process prior distribution is assigned to $\boldsymbol{d}(\cdot)$. They assigned an iid inverse Gamma prior distribution to the variances.

For a single spatial process realization, a non-Gaussian geostatistical model was proposed by Palacios and Steel (2006). From their paper, the basic geostatistical model is expressed as

$$Z(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + \sigma\epsilon(\boldsymbol{x}) + \tau\rho(\boldsymbol{x}), \tag{1.25}$$

where $\epsilon(\boldsymbol{x})$ is a second-order stationary error process with mean 0, variance 1 and isotropic correlation function $C_{\boldsymbol{\theta}}$. In order to catch the non-Gaussian feature, such as heavy tails, they propose to extend model (1.25) by adding a mixing variable $\lambda_i \in \mathcal{R}_+$ to each observation $i = 1, ..., n$. The new model at ith location becomes

$$z_i = \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta} + \sigma\frac{\epsilon(\boldsymbol{x}_i)}{\sqrt{\lambda_i}} + \tau\rho_i, \tag{1.26}$$

where $\rho_i \sim \mathrm{N}(0,1)$, iid and independent of $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \sim \mathrm{N}(\boldsymbol{0}, C_{\boldsymbol{\theta}})$. The mixing variables λ_i are independent of ρ_i and $\boldsymbol{\epsilon}$. In order to obtain a valid stochastic process

and ensure nice properties, they propose to model the mixing variable by

$$\ln(\boldsymbol{\lambda}) = (\ln(\lambda_1), \ldots \ln(\lambda_n))^T \sim \mathrm{N}\left(-\frac{\nu}{2}\mathbf{1}, \nu C_{\boldsymbol{\theta}}\right), \tag{1.27}$$

i.e., model $\ln(\boldsymbol{\lambda})$ by use of a Gaussian process with constant mean $-\nu/2$ and covariance function $\nu C_{\boldsymbol{\theta}}$. Note that the correlation function they used for $\ln(\boldsymbol{\lambda})$ is identical to that of the $\boldsymbol{\epsilon}$ process. They argue that the two correlation functions can be different in principle but in that case it would be extremely difficult to estimate parameters with a practically relevant sample size.

They call the new model "Gaussian-log-Gaussian (GLG)". They derive expressions for the moments. Bayesian inference is performed for estimation and prediction. Proper priors are applied and random-walk Metropolis-Hastings algorithms are used to draw samples from posteriors for all parameters. The latent process vectors were partitioned into blocks and the proposal distribution was constructed by use of log-normal distributions to approximate truncated normal distributions for the conditional posterior.

1.3 Outline of the Dissertation

We have introduced the traditional stationary Gaussian process modeling for geostatistical data and some nonstationary approaches in Section 1.1. In Section 1.2, we review the stochastic volatility model in time series and discuss some extended applications in regression and spatial data analysis. The GLG model in Section 1.2.4 is actually an extension from the multiplicative model in Section 1.1.2. The multiplicative model introduces nonstationarity by scaling. The location-dependent scale $\sigma(\boldsymbol{x})$ is specified using a deterministic function, while the GLG model uses a latent stochastic process to model the scale variable. Our motivation comes from the SV model in a similar fashion. We think about modeling the scale $\sigma(\boldsymbol{x})$ by revising the SV model (1.10). The discrete time t is replaced by the continuous (non-lattice)

d-dimensional spatial index $\boldsymbol{x}$ and the iid noise ϵ_t is replaced by a stochastic process $Z(\boldsymbol{x})$. By doing so, we propose a new model called the stochastic heteroscedastic process (SHP). We will model spatial surfaces as realizations from SHP.

Chapter 2 starts by defining the SHP model. Some important properties, especially the features of conditional and unconditional covariance functions, are investigated. By simulation, we show the versatility of SHP realizations, together with the parameter confounding effects that exist in the unconditional correlation functions and sample paths. In Chapter 3, we propose to apply importance sampling in likelihood calculation and latent process estimation. We derive methods of predicting the spatial processes at unobserved locations. Some delicate implementation issues are discussed in detail. A low-dimensional approximation model is introduced to overcome the computational issues when the sample size is large. We also extend the single-realization SHP model to the SHP model with replicates. Chapter 4 provides a variety of simulation studies to evaluate the estimation and prediction procedures for the SHP model and compare different spatial prediction methods. In Chapter 5, we present applications of the SHP model on three data sets: Enhanced Vegetation Index (EVI) data, China precipitation data and NO_3 deposition data. The advantages of SHP over a stationary Gaussian process model are illustrated from a few aspects: SHP is better at capturing spatial heterogeneities, it yields superior prediction performance, and it gives efficient selection probabilities (using SHP prediction variance) for adaptive sampling. Finally, Chapter 6 gives a overview of the thesis and discusses possible future research directions.

Chapter 2

THE STOCHASTIC HETEROSCEDASTIC PROCESS MODEL

As we briefly discussed in Section 1.1.2, one of the approaches to model nonstationarity is through scaling (Banerjee et al. (2003)). Suppose $Z(\boldsymbol{x})$ is a mean 0, variance 1 stationary process with correlation function ρ and $\sigma(\boldsymbol{x})$ is a pre-specified deterministic function, then $W(\boldsymbol{x}) = \sigma(\boldsymbol{x})Z(\boldsymbol{x})$ is a nonstationary process. The $\sigma(\boldsymbol{x})$ is a pre-specified deterministic function. Motivated by the SV model in time series, we think about modeling $\sigma(\boldsymbol{x})$ as a random process such that $W(\boldsymbol{x})$ retains the nonlinear flavor in terms of sample path with nice probabilistic structure.

2.1 Definition of the Stochastic Heteroscedastic Process (SHP)

We now represent the SHP model as follows:

$$\begin{aligned} Y(\boldsymbol{x}) &= \boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta} + W(\boldsymbol{x}), \\ W(\boldsymbol{x}) &= \sigma \exp\left(\frac{\tau\alpha(\boldsymbol{x})}{2}\right) Z(\boldsymbol{x}), \qquad \sigma > 0, \qquad \tau > 0, \end{aligned} \tag{2.1}$$

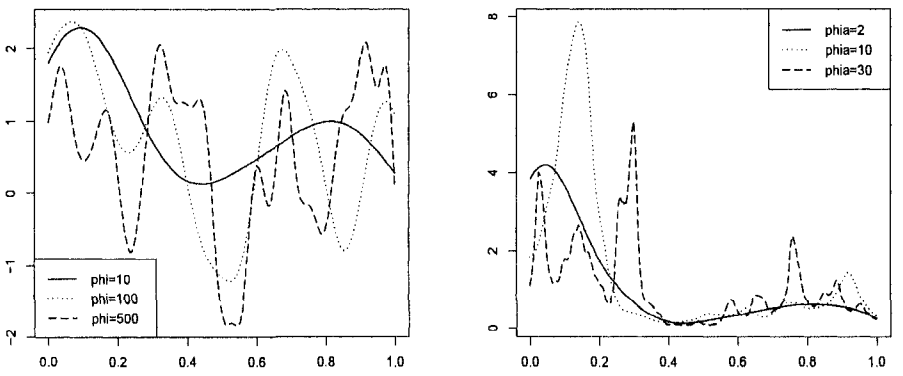
where $\alpha(\boldsymbol{x})$ and $Z(\boldsymbol{x})$ are two independent stationary Gaussian processes with mean 0, variance 1 and correlation functions ρ_α and ρ_z respectively. Throughout this dissertation, we take ρ_α and ρ_z to be isotropic correlation functions with range parameters ϕ_α and ϕ_z respectively. The overall trend is represented by linear regression $\boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta}$, where $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), ..., g_p(\boldsymbol{x}))$ and $\boldsymbol{\beta}$ a $p \times 1$ coefficient vector. The correlated error process $W(\boldsymbol{x})$ captures residual spatial association. The spatial dependence structure is described by the covariance function of $W(\boldsymbol{x})$, which is

traditionally assumed to be stationary. But for the SHP model (2.1), conditional on the α process, the$W(\boldsymbol{x})$ process has nonstationary covariance function.

The latent process $\alpha(\boldsymbol{x})$ is used to model the clustering effect of volatility, which makes the realizations generated from the SHP model more versatile than those from Gaussian processes. The sample paths simulated from a Gaussian process model exhibit homogenous features, as shown in panel (a) of Figure 2.1. Increasing the range parameter ϕ, the whole sample path becomes simultaneously more variable. While for sample paths generated from the SHP model, fixing the Z process, the sample path can be Gaussian-like when ϕ_α is small but becomes more heterogenous by increasing ϕ_α, as shown in panel (b) of Figure 2.1. Also note that by use of different correlation functions for α and Z processes, the sample paths can have rich smoothness (differentiability) properties. By setting $\tau^2 = 0$, the SHP model (2.1) reduces to the Gaussian process model (1.3). In Figure 2.2, we plot a Gaussian process surface by letting $\tau^2 = 0$ in panel (a). As we increase τ^2 to 0.4, the SHP model surface still has a relatively uniform degree of smoothness over the whole input domain, similar to those simulated from a Gaussian process, as seen in Figure 2.2 panel (b). That is, we do not need to restrict $\tau^2 = 0$ to generate Gaussian-like realizations. The SHP model can recover the Gaussian process properties by allowing small values of ϕ_α (as shown by Figure 2.1) and/or τ^2 (as shown by Figure 2.2). On the other hand, the realization from the SHP model has some local volatility, which gets more obvious with increasing ϕ_α and/or τ^2, as seen in panel (b) of Figure 2.1 and Figure 2.2 panel (c).

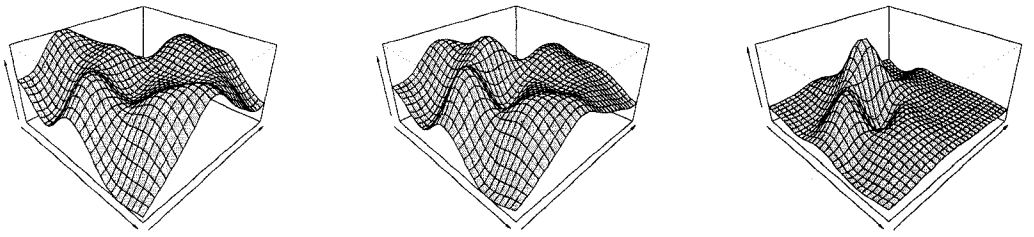
2.2 Properties of the SHP Model

Conditionally on α, the SHP model is a Gaussian process with non-stationary covariance structure. Unconditionally, the SHP is a stationary non-Gaussian process. In this section, we describe some key properties of SHP that are useful for modeling.



(a) (b)

Figure 2.1: Gaussian process and SHP 1-d simulation plots. Panel (a) shows some sample paths simulated from a Gaussian process. We take $\sigma^2 = 1$ and a Gaussian correlation function. Panel (b) shows some sample paths simulated from a SHP model. We take $\sigma^2 = 1$, $\tau^2 = 1$, ρ_z a Gaussian correlation with $\phi_z = 10$ and ρ_α a Matérn correlation with $\nu = 2.5$. Each of the α process realizations on the right panel use a common simulated random noise sequence. The Z process for all of the 6 sample paths use another common random noise sequence.



(a) (b) (c)

Figure 2.2: Gaussian process and SHP 2-d simulation surface plots (I). We simulate 2-d realizations from a SHP model by use of $\sigma^2 = 0.2, \phi_\alpha = 0.15$ and $\phi_z = 0.3$. Both ρ_α and ρ_z are Gaussian correlations. Panel (a) corresponds to $\tau^2 = 0$ which is a Gaussian process realization. Panel (b) corresponds to $\tau^2 = 0.4$. Panel (c) corresponds to $\tau^2 = 4$. The α process realizations in panels (b) and (c) use a common simulated random noise sequence. The Z process realizations in all panels use another common random noise sequence.

The Y process has mean $\boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta}$, variance $\sigma^2\exp(\tau^2/2)$ and kurtosis

$$\frac{\mathrm{E}[(Y-\mathrm{E}(Y))^4]}{(\mathrm{Var}(Y))^2} = 3\exp(\tau^2).$$

The normal distribution has kurtosis equal to 3, so the data from the SHP model has "excess kurtosis" or heavier tails than the normal.

2.2.1 SHP covariance function

Using the independence of α and Z processes, W has unconditional correlation function given by

$$\rho_Y(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{1}{4}\tau^2 + \frac{1}{4}\tau^2\rho_\alpha(\|\boldsymbol{x}-\boldsymbol{x}'\|)\right)\rho_z(\|\boldsymbol{x}-\boldsymbol{x}'\|), \tag{2.2}$$

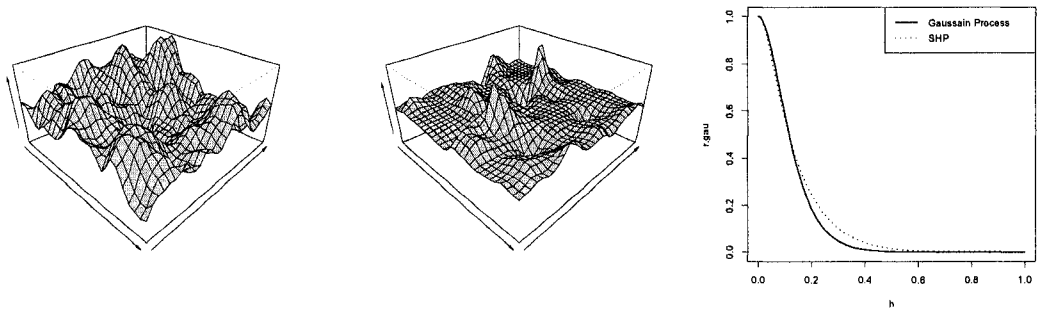
which is isotropic. This unconditional correlation function is flexible because it combines the properties of two isotropic correlation functions ρ_α and ρ_z. It can be used independently of the non-Gaussian SHP model as a rich isotropic correlation class for Gaussian processes. We will discuss its unique smoothed nugget effect property later in this chapter.

Conditioning on the latent process α, the covariance function between two points is

$$\gamma(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\alpha}) = \sigma^2\exp\left(\frac{\tau\alpha(\boldsymbol{x})}{2}\right)\rho_z(\boldsymbol{x},\boldsymbol{x}')\exp\left(\frac{\tau\alpha(\boldsymbol{x}')}{2}\right). \tag{2.3}$$

Equations (2.2) and (2.3) indicate that the W process is conditionally heteroscedastic and unconditionally weakly stationary. For a single realization of SHP, it is the covariance conditional on the latent process α that decides the heteroscedastic features. In Figure 2.3, we plot two 2-d surfaces, one realization simulated from a Gaussian process (panel (a)) and another simulated from a SHP model (panel (b)). While the unconditional correlation functions for the two models are nearly identical as shown in panel (c), the realizations are remarkably different. The realization from the Gaussian process has a relatively uniform degree of smoothness

over the whole input domain. In contrast, the smoothness of the SHP realization varies over different parts of the region, which is due largely to the inhomogeneity of the conditional covariance function.



(a) (b) (c)

Figure 2.3: Gaussian process and SHP 2-d simulation surface plots (2). Panel (a) is a Gaussian process surface with $\phi = 6.4$. Panel (b) is a SHP realization with $\phi_\alpha = 8.5$ and $\phi_z = 4.4$. All correlation functions are Matérn with $\nu = 2.5$. Panel (c) shows the correlation functions of the Gaussian process in (a) and SHP in (b). The Z processes in panels (a) and (b) use a common simulated random noise sequence.

2.2.2 Smoothed nugget effect of the unconditional correlation function

The unconditional correlation function (2.3) motivates us to explore the limiting processes as $\phi_\alpha \to 0$ and $\phi_\alpha \to \infty$. When $\phi_\alpha = 0$, the α process degenerates to a single $N(0,1)$ random variable and the unconditional correlation function for Y becomes ρ_z. While the process Y remains non-Gaussian, a single realization of Y is not distinguishable from a realization of a Gaussian process. It is a realization of the Gaussian process Z scaled by the multiplicative constant $\exp(\tau\alpha/2)$. On the other hand, as $\phi_\alpha \to \infty$, the unconditional correlation function converges to

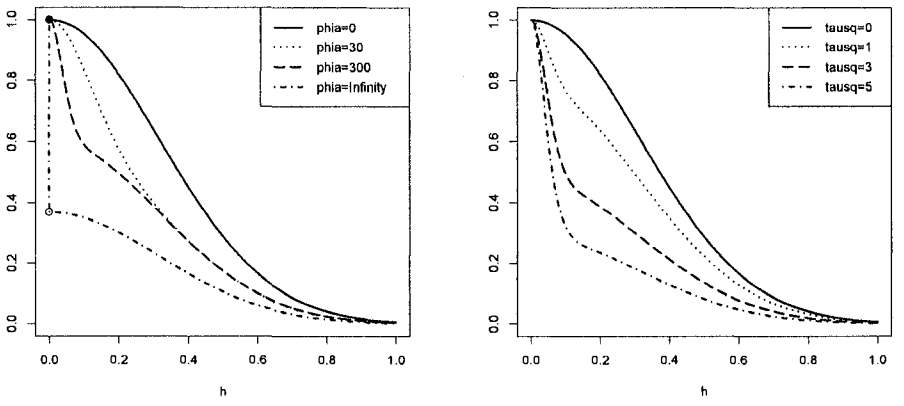
$$\rho(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} 1, & \text{if } \|\boldsymbol{x} - \boldsymbol{x}'\| = 0, \\ \exp(-\frac{\tau^2}{4})\rho_z(\boldsymbol{x}, \boldsymbol{x}'), & \text{if } \|\boldsymbol{x} - \boldsymbol{x}'\| > 0, \end{cases}$$

which is simply ρ_z with the addition of the relative nugget $\delta = 1 - \exp(\tau^2/4)$.

The effect of varying the correlation parameters on the correlation function can be seen from Figure 2.4. The correlation decreases with ϕ_α or τ^2 increases for

each value of h, the distance between sample points. As ϕ_α approaches infinity, one can see the emergence of a smoothed nugget. At $\phi_\alpha = \infty$, we have a full-fledged nugget of size $1 - \exp(-\frac{\tau^2}{4})$. In the other direction, the correlation function decays smoothly for small values of ϕ_α. When $\tau^2 = 0$, the correlation in panel (b) reduces to a Gaussian correlation. With τ^2 increasing, the smoothed nugget effect becomes more pronounced and the correlation decreases more sharply.

The unconditional correlation of SHP offers a rich class of correlation functions that can also allow for a smoothed nugget effect. For the traditional Gaussian process model, a nugget is added to the covariance structure to model the measurement error and microscale variability. These effects are modeled with a single parameter and cannot be separated. The additive nugget makes the covariance discontinuous at the origin, which is undesirable for microscale variation that accounts for possible model misspecification at a very fine scale. The smoothed nugget of the SHP unconditional correlation function explains the microscale variation in a natural way, using a parameterization that is distinct from an additive measurement error.



(a) (b)

Figure 2.4: SHP unconditional correlation function plots. The left panel shows the effect of ϕ_α($\phi_z = 5$ and $\tau^2 = 2$). The right panel shows the effect of τ^2 ($\phi_z = 5$ and $\phi_\alpha = 200$). Both ρ_α and ρ_z are Gaussian correlations.

2.3 Confounding Effects of the Model Parameters

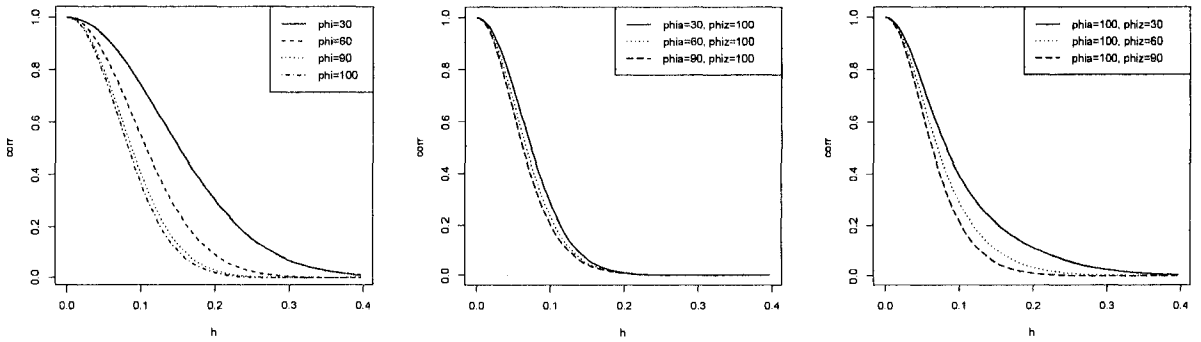
The realizations simulated from the SHP model are versatile and can represent spatial inhomogeneities. But the flexibility comes with a certain price, i.e., a more complicated model with a more complicated likelihood, due to the latent process, than that of a stationary Gaussian process model. In this section, we will explore the confounding effects of model parameters, illustrated via simulation plots.

2.3.1 Confounding in the unconditional correlation function

In Figure 2.5, we plot several Gaussian correlation functions using different range parameter values in panel (a). By applying Gaussian correlation functions on both ρ_α and ρ_z, and using the same range values in (a) interactively for ϕ_α and ϕ_z, we create some SHP unconditional correlation function plots in panels (b) and (c). We see that for Gaussian correlation functions in panel (a), changing the range parameter ϕ makes the correlation function vary substantially. For the SHP unconditional correlation in panel (b), however, the correlation plots almost overlap for a large range of ϕ_α values when fixing ϕ_z at 100. Further the correlations in panel (c) vary much less than the Gaussian correlations when varying ϕ_z with ϕ_α fixed at 100. In Figure 2.6 panel (a), we show several unconditional correlation plots by varying τ^2 with fixed ϕ_z and ϕ_α values. The correlations for four different τ^2 values are very close. In Figure 2.6 panel (b), two correlation plots obtained with very different τ^2 and ϕ_α values are almost identical.

Comparing Figures 2.5 and 2.6 with Figure 2.4, we see that ϕ_z is most dominant in the unconditional correlation shape. When ϕ_z is small, adjusting τ^2 and/or ϕ_α values can substantially change the correlation features, especially the smoothed nugget effect. But when ϕ_z is large enough, adjusting τ^2 and ϕ_α across a wide ranges does almost nothing to the correlation plot. Also the confounding between ϕ_α and ϕ_z, ϕ_α and τ^2 are obvious. So, although the SHP unconditional correlation

(2.2) is flexible and produces the smoothed nugget effect property, it is difficult to estimate the parameters due to confounding if we use it as an isotropic correlation function in a Gaussian process model.

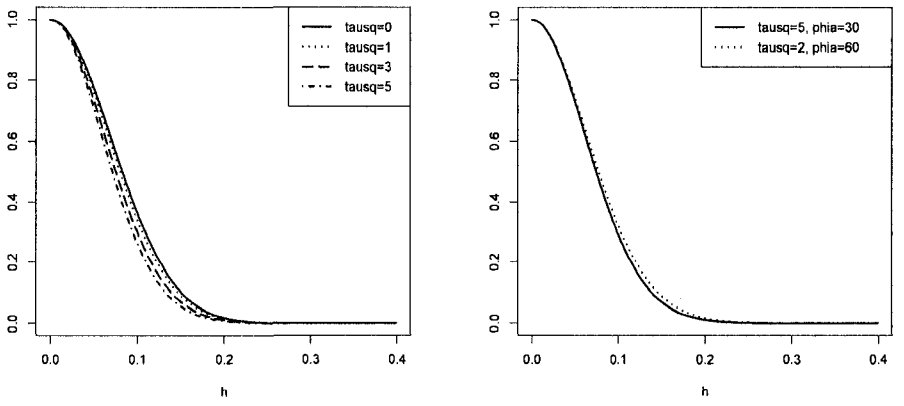


(a) (b) (c)

Figure 2.5: Confounding correlation plot (I). Panel (a) shows Gaussian correlation plots under different range values. Panel (b) shows SHP unconditional correlations with ϕ_z fixed and ϕ_α varying. Panel (c) shows SHP unconditional correlations with ϕ_α fixed and ϕ_z varying. We take $\tau^2 = 4$. Both ρ_α and ρ_z are Gaussian correlation functions.

2.3.2 Confounding in the sample paths

In Figure 2.7, we plot two Gaussian process realizations on each panel using $\phi = 30$, 60 and 90, respectively. It is clear that increasing ϕ leads to considerably more variable sample paths. In Figures 2.8 and 2.9, we show some SHP sample paths using ϕ_α and ϕ_z values as in Figure 2.5 panel (b) and (c), respectively. From Figure 2.8, we see that the sample path features are very similar for a large range of ϕ_α values when fixing $\phi_z = 100$. Also, Figure 2.9 shows that the sample path features are similar for $\phi_z = 60$ or 90 when fixing $\phi_\alpha = 100$. Realizations using $\phi_z = 30$ have slightly smoother features than $\phi_z = 60$ or 90. Note that sample paths in Figure 2.8 panel (c) ($\phi_z = 100$) are more volatile in some ways than those in Figure 2.9 ($\phi_z = 30, 60$ and 90). Based on these observations, we conclude that ϕ_z (or Z process) provides more control on the sample path features than ϕ_α (or



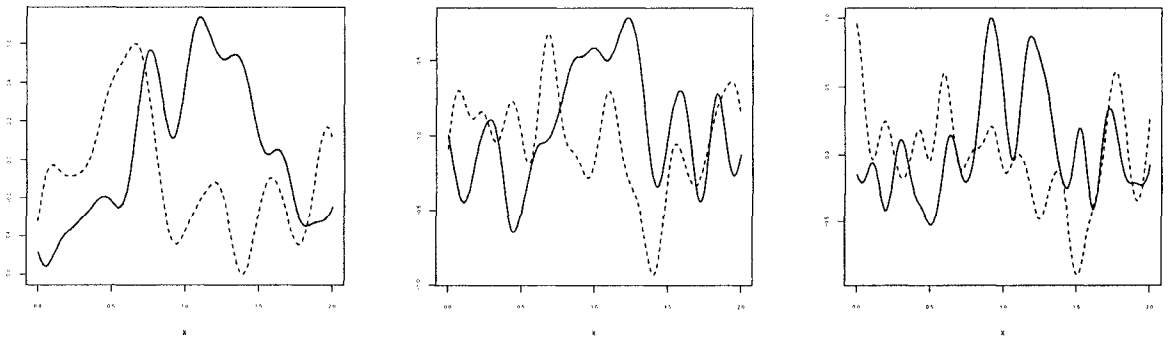
(a) (b)

Figure 2.6: Confounding correlation plot (II). Panel (a) shows SHP unconditional correlation with $\phi_z = 100$ and $\phi_\alpha = 30$ fixed and τ^2 varying. Panel (b) shows the effect of using very different τ^2 and ϕ_α values with $\phi_z = 90$. Both ρ_α and ρ_z are Gaussian correlation functions.

α process). The sample features are not as sensitive with respect to ϕ_α and ϕ_z as those for Gaussian process with respect to ϕ, which will possibly make parameter estimation more problematic than the Gaussian case. Also ϕ_α will probably be even harder to identify than ϕ_z.

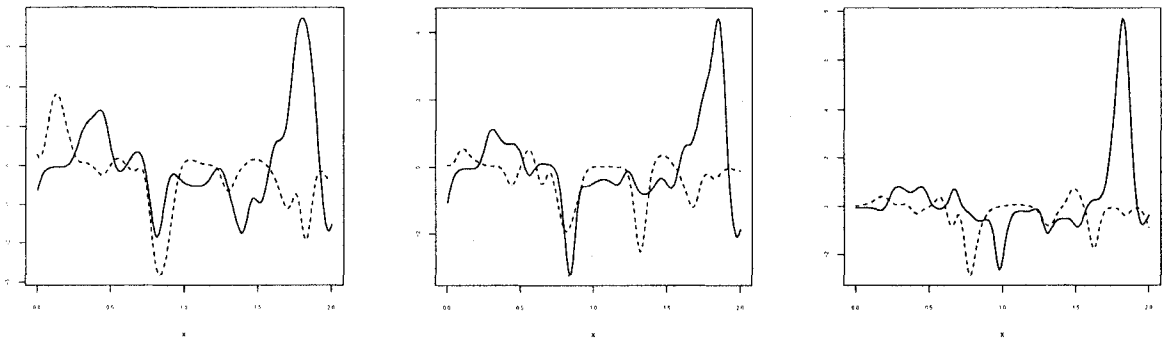
In Figure 2.10, we show three SHP realizations in each panel by interchanging ϕ_α and ϕ_z values. By use of the same common random number sequences for both α and Z across panels, the plots in two panels show very similar shapes. This indicates the confounding effects between ϕ_α and ϕ_z in representing sample paths. Figure 2.11 shows SHP realizations by use of different correlation functions for α and Z processes. Even so, sample paths produced with large ϕ_z and small ϕ_α are similar to those using small ϕ_z and large ϕ_α. For realizations shown in Figures 2.10 and 2.11 and using a finite number of sampled points which are not too dense, it is difficult to recognize from which parameter combinations the realizations are generated.

In Chapter 4, we will explore the influence of these confounding effects on parameter estimation and sample path prediction.



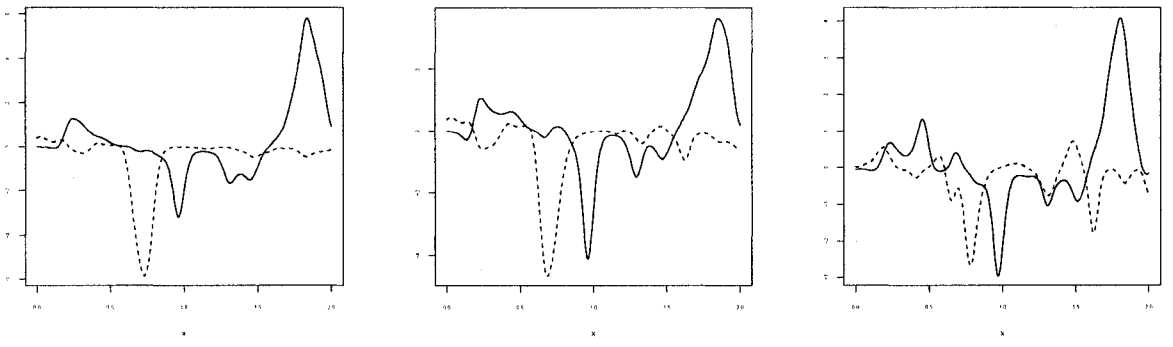
(a) (b) (c)

Figure 2.7: Gaussian process sample paths. Panel (a) – Panel (c) each shows two simulated Gaussian process sample paths using $\phi = 30$, $\phi = 60$ and $\phi = 90$ respectively. We take $\beta = 0$ and $\sigma^2 = 0.2$ and use a Gaussian correlation function. Common random number sequences are applied across panels.



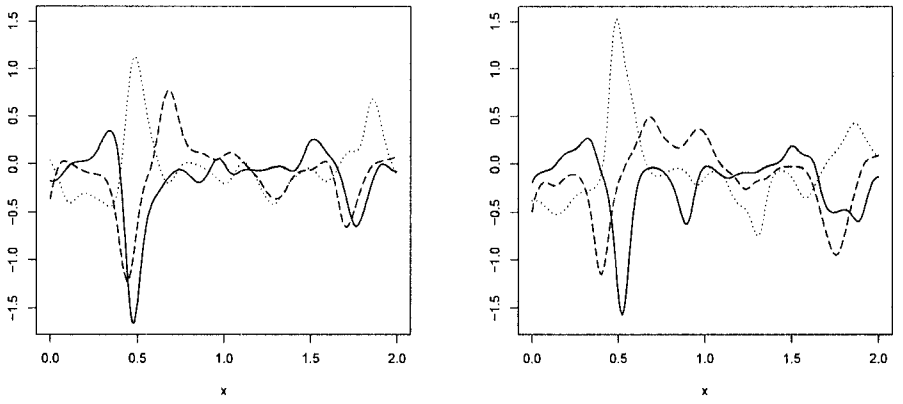
(a) (b) (c)

Figure 2.8: SHP sample paths (I). Panel (a) – Panel (c) each shows two simulated SHP sample paths using $\phi_\alpha = 30$, $\phi_\alpha = 60$ and $\phi_\alpha = 90$ respectively. We take $\sigma^2 = 0.2$, $\tau^2 = 4$, $\beta = 0$ and $\phi_z = 100$. Both ρ_α and ρ_z are Gaussian correlations. Common random number sequences are applied for both α and Z realizations respectively across three panels.



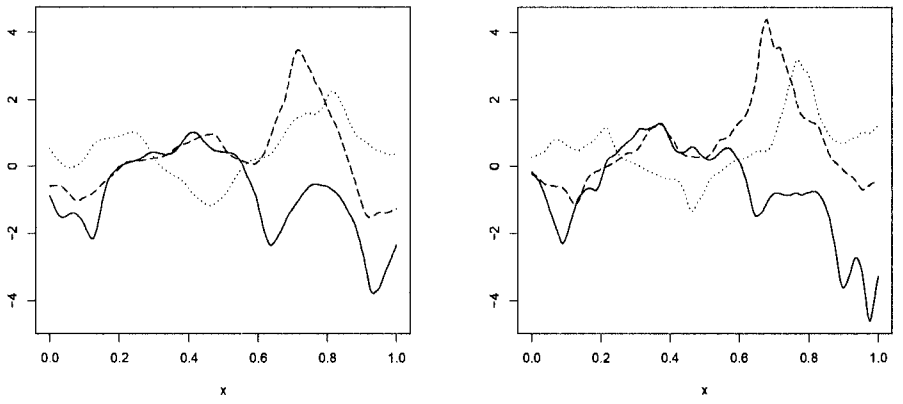
(a) (b) (c)

Figure 2.9: SHP sample paths (II). Panel (a) – Panel (c) each shows two simulated SHP sample paths using $\phi_z = 30$, $\phi_z = 60$ and $\phi_z = 90$ respectively. We take $\sigma^2 = 0.2$, $\tau^2 = 4$, $\beta = 0$ and $\phi_\alpha = 100$. Both ρ_α and ρ_z are Gaussian correlations. Common random number sequences are applied for both α and Z realizations respectively across three panels and are the same as in Figure 2.8.



(a) (b)

Figure 2.10: SHP confounding sample paths (I). Panel (a) shows three SHP sample paths based on $\phi_\alpha = 40$ and $\phi_z = 80$. Panel (b) shows three SHP sample paths based on $\phi_\alpha = 80$ and $\phi_z = 40$. Both ρ_α and ρ_z are Gaussian correlation functions. We take $\beta = 0$, $\sigma^2 = 0.2$ and $\tau^2 = 4$. The α and Z processes realizations in two panels use common random noise sequences, respectively.



(a) (b)

Figure 2.11: SHP confounding sample paths (II). Panel (a) shows three SHP sample paths based on $\phi_\alpha = 10$ and $\phi_z = 60$. Panel (b) shows three SHP sample paths based on $\phi_\alpha = 15$ and $\phi_z = 20$. The correlation function ρ_α is taken to be Matérn with $\nu = 2.5$ and ρ_z is Gaussian. We take $\beta = 0$, $\sigma^2 = 1$ and $\tau^2 = 1$. The α and Z processes realizations in two panels use common random noise sequences, respectively.

Chapter 3

LIKELIHOOD INFERENCE

As we discussed in Section 1.2.1, the SV model in time series has easily-derived probabilistic properties, but the estimation is difficult since the likelihood is not in a closed form. The distribution of $y_t|y_{t-1}$ is specified implicitly through the latent process h_t. Researchers have developed many methods to estimate the SV model, among which maximum likelihood by use of an importance sampling device and Markov Chain Monte Carlo (MCMC) have had the most impact. For the MCMC approach, early work focused on "single move" algorithms, drawing h_t one at a time. The drawback of "single move" is slow convergence especially when the latent process has high correlation. A "multi-move" sampler is obtained by approximating $\log \epsilon_t^2$ by a mixture of normals so that $\log y_t^2$ can be written in the form of a Gaussian linear state-space model, and a Gaussian simulation smoother can be applied to draw $\boldsymbol{h}|\boldsymbol{y}$ simultaneously. One can argue that this approach is only based on an approximation. For the SHP model, the correlation of the latent process will lead to slow convergence in MCMC algorithm if using a "single move" sampler. Since the α process is continuous, the correlation has even more impact that in a discrete-time AR(1) process in a SV model. Also, $\log(Z^2)$ is correlated instead of iid log chi-square distributed, which prevents us from readily applying a "multi-move" sampler analogous to the SV model. We will explore methodologies to improve MCMC convergence behavior in future work. In this dissertation, we use a maximum likelihood method and importance sampling strategy. We describe the scheme of likelihood calculation in Section 3.1. We derive an importance density

used for likelihood approximation and latent process estimation. In Section 3.2, we present strategies for predicting the latent processes α and Y at unobserved location $\boldsymbol{x}_0$. We go through some implementation details in Section 3.3. In Section 3.4, we introduce a low-dimensional importance density to overcome the high dimensionality problems when the data set is large. Section 3.5 discusses how to model the SHP with replicates.

3.1 Importance Density and Likelihood Approximation

Let $\boldsymbol{\alpha} := (\alpha(\boldsymbol{x}_1), ..., \alpha(\boldsymbol{x}_n))^T$ be the vector of the latent process values at the observed locations and $\boldsymbol{\psi} := (\boldsymbol{\theta}, \phi_\alpha)$ the model parameters. Here $\boldsymbol{\theta} := (\sigma^2, \tau^2, \phi_z, \boldsymbol{\beta})$. The joint density of $(\boldsymbol{Y}, \boldsymbol{\alpha})$ $(\boldsymbol{Y} := (Y(\boldsymbol{x}_1), ..., Y(\boldsymbol{x}_n))^T)$ of the SHP model is given by

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\alpha}|\boldsymbol{\psi}) &= p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\phi_\alpha) \\ &= p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) |R_\alpha|^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\alpha}^T R_\alpha^{-1}\boldsymbol{\alpha})(2\pi)^{-\frac{n}{2}}, \end{aligned} \tag{3.1}$$

where R_α is the correlation matrix for $\boldsymbol{\alpha}$, which only depends on ϕ_α and

$$p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) \sim \mathrm{N}\left(G(\boldsymbol{x})\boldsymbol{\beta}, \sigma^2 \mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\} R_z \mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}\right),$$

where R_z is the $n \times n$ correlation matrix for $\boldsymbol{z}$ and $G(\boldsymbol{x})$ is the $n \times p$ design matrix for the regression term. It follows that the likelihood of the observed data is given by the n-fold integral

$$L(\boldsymbol{\psi}; \boldsymbol{Y}) = \int p(\boldsymbol{Y}, \boldsymbol{\alpha}|\boldsymbol{\psi}) d\boldsymbol{\alpha} = \int p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\phi_\alpha) d\boldsymbol{\alpha}. \tag{3.2}$$

The likelihood (3.2) cannot be computed explicitly. There are some simulation-based procedures in the literature to approximate such integration (see Robert and Casella (1999)). Importance sampling may be used to increase computational efficiency and improve the accuracy of the approximation. Some methodologies have been

developed to find efficient important densities, e.g., Danielsson and Richard (1993) and Durbin and Koopman (1997), etc. Suppose we have an importance density $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$, by implementation of Monte Carlo integration, the integral in (3.2) can be rewritten as

$$\begin{aligned} L(\boldsymbol{\psi};\boldsymbol{Y}) &= \int \frac{p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)}{p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})} p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})d\boldsymbol{\alpha} \\ &= \mathrm{E}_a\left[\frac{p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)}{p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})}\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\left[\frac{p(\boldsymbol{Y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\theta})p(\boldsymbol{\alpha}^{(i)}|\phi_\alpha)}{p_a(\boldsymbol{\alpha}^{(i)}|\boldsymbol{Y},\boldsymbol{\psi})}\right], \end{aligned} \tag{3.3}$$

where $\boldsymbol{\alpha}^{(1)},...,\boldsymbol{\alpha}^{(N)}$ are drawn from $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$.

As mentioned in Durbin and Koopman (1997), to achieve efficiency the importance density $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ should be chosen to be as close as can be managed to $p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ within the class of conditional densities that are feasible and efficient for drawing simulation samples. The reason for this choice is that, if $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ is exactly equal to $p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$, then a sample of only $N = 1$ is required for accurate likelihood calculation, as is easily shown. In this dissertation, we refer to the likelihood approximation method in Davis and Rodriguez-Yam (2005) and obtain a density $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ that is an approximation of the posterior density $p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$. Their work was applied to state-space models so that recursive prediction algorithms, such as the Kalman recursions or innovations algorithm, were available to accelerate the calculation in finding the importance density. In Section 3.1.1, we will derive $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ by modifying the method in Davis and Rodriguez-Yam (2005) to fit the SHP model framework.

3.1.1 Derivation of the importance density

To find a good approximation of $p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$, Davis and Rodriguez-Yam (2005) start with a Taylor series expansion of $\log p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ in a neighborhood of the posterior mode of $p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$. The log-density of $(\boldsymbol{Y},\boldsymbol{\alpha})$, denoted by $l(\boldsymbol{\psi};\boldsymbol{Y},\boldsymbol{\alpha})$, is given

by

$$l(\boldsymbol{\psi};\boldsymbol{Y},\boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|R_\alpha|^{-1} + l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^T R_\alpha^{-1}\boldsymbol{\alpha}, \tag{3.4}$$

where $l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha}) := \log p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})$.

Now, let

$$\boldsymbol{k}^* := \frac{\partial}{\partial \boldsymbol{\alpha}} l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}, \tag{3.5}$$

where $\boldsymbol{\alpha}^*$ is the mode of $p(\boldsymbol{Y},\boldsymbol{\alpha}|\boldsymbol{\psi})$, which solves $\frac{\partial}{\partial \boldsymbol{\alpha}} l(\boldsymbol{\psi};\boldsymbol{Y},\boldsymbol{\alpha}) = \mathbf{0}$. From (3.4), it follows that

$$\boldsymbol{k}^* := R_\alpha^{-1}\boldsymbol{\alpha}^*. \tag{3.6}$$

Hence, the second order Taylor expansion of $l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha})$ with respect to the latent process $\boldsymbol{\alpha}$ around the posterior mode $\boldsymbol{\alpha}^*$ is

$$l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha}) = h^* + \boldsymbol{\alpha}^{*T} R_\alpha^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) - \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T K^*(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + e(\boldsymbol{\alpha},\boldsymbol{\alpha}^*), \tag{3.7}$$

where $h^* := l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}$, $K^* := -\frac{\partial^2}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} l(\boldsymbol{\theta};\boldsymbol{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}$ and $e(\boldsymbol{\alpha},\boldsymbol{\alpha}^*)$ is the corresponding remainder. Thus,

$$\begin{aligned} l(\boldsymbol{\psi};\boldsymbol{Y},\boldsymbol{\alpha}) &= -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|R_\alpha|^{-1} + h^* - \frac{1}{2}\boldsymbol{\alpha}^{*T} R_\alpha^{-1}\boldsymbol{\alpha}^* \\ &\quad -\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T(K^* + R_\alpha^{-1})(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) + e(\boldsymbol{\alpha},\boldsymbol{\alpha}^*). \end{aligned} \tag{3.8}$$

Let $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})$ be the approximation of the posterior (or joint distribution (3.8)) when the remainder term is omitted. It follows that

$$p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi}) \sim \mathrm{N}(\boldsymbol{\alpha}^*, (K^* + R_\alpha^{-1})^{-1}), \tag{3.9}$$

which is the equation (6) of Davis and Rodriguez-Yam (2005), except replacing V by R_α^{-1}.

For the SHP model, by careful derivation, we get K^* calculated as follows:

$$
\begin{aligned}
&K^* = \frac{\tau^2}{4\sigma^2}(B + \mathrm{diag}\{\boldsymbol{c}\}),\\
&B = \mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right\}\mathrm{diag}\{\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}\}R_z^{-1}\mathrm{diag}\{\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}\}\\
&\times\mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right\},\\
&\boldsymbol{c} = \left(\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right)^T\mathrm{diag}\{\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}\}R_z^{-1}\mathrm{diag}\{\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}\}\\
&\times\mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right\}.
\end{aligned}
\tag{3.10}
$$

Thus we have created an importance density and can approximate the likelihood value by (3.3). Maximizing (3.3) with respect to $\boldsymbol{\psi}$, we can get the maximum likelihood estimate $\hat{\boldsymbol{\psi}}$. Some implementation details will be discussed in Section 3.3.

3.1.2 Estimation of functions of the volatility

If $\boldsymbol{\psi}$ were known, a function of the latent process at observed locations, $h(\boldsymbol{\alpha})$, can be estimated as the conditional expectation $\mathrm{E}[h(\boldsymbol{\alpha})|\boldsymbol{Y}]$, written as

$$
\begin{aligned}
\mathrm{E}[h(\boldsymbol{\alpha})|\boldsymbol{Y}] &= \int h(\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})d\boldsymbol{\alpha}\\
&= \int h(\boldsymbol{\alpha})\frac{p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)}{p(\boldsymbol{Y}|\boldsymbol{\psi})}d\boldsymbol{\alpha}\\
&= \frac{\int h(\boldsymbol{\alpha})p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)d\boldsymbol{\alpha}}{\int p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)d\boldsymbol{\alpha}}\\
&= \frac{\mathrm{E}_a[h(\boldsymbol{\alpha})p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)/p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})]}{\mathrm{E}_a[p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha)/p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\psi})]}.
\end{aligned}
\tag{3.11}
$$

Once the parameter estimate $\hat{\boldsymbol{\psi}}$ is obtained, the conditional expectation in (3.11) is approximated using Monte Carlo integration by sampling from $p_a(\boldsymbol{\alpha}|\boldsymbol{Y},\hat{\boldsymbol{\psi}})$. Then equation (3.11) can be approximated by Monte Carlo integration. In this way, we are able to estimate many functions $h(\boldsymbol{\alpha})$ of interest, e.g., $\boldsymbol{\alpha}$ and $\exp(\tau\boldsymbol{\alpha}/2)$.

3.2 Prediction

A critical issue in spatial data analysis is prediction, i.e., to predict the variable Y at an unobserved location $\boldsymbol{x}_0$ given observations of a random field $\boldsymbol{Y} = (Y(\boldsymbol{x}_1), ..., Y(\boldsymbol{x}_n))$. For the SHP model, we are not only interested in the process Y but also the latent process α which describes the spatial volatility. We have presented the scheme of estimating the α process at observed locations in Section 3.1.2. In this Section, we will show the methods of predicting α and Y at an unobserved location $\boldsymbol{x}_0$.

3.2.1 Prediction of the latent process

Let α_0 be the value of the latent process at some unobserved location $\boldsymbol{x_0}$. We seek for the best predictor $\mathrm{E}(\alpha_0|\boldsymbol{Y})$. Noting that $p(\alpha_0|\boldsymbol{\alpha}, \boldsymbol{Y})$ is equivalent to $p(\alpha_0|\boldsymbol{\alpha})$, since

$$\begin{aligned} p(\alpha_0|\boldsymbol{\alpha}, \boldsymbol{Y}) &= \frac{p(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{Y})}{p(\boldsymbol{Y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})} \\ &= \frac{p(\boldsymbol{Y}|\alpha_0, \boldsymbol{\alpha})}{p(\boldsymbol{Y}|\boldsymbol{\alpha})} \times \frac{p(\alpha_0, \boldsymbol{\alpha})}{p(\boldsymbol{\alpha})} \\ &= \frac{p(\alpha_0, \boldsymbol{\alpha})}{p(\boldsymbol{\alpha})} \\ &= p(\alpha_0|\boldsymbol{\alpha}). \end{aligned} \tag{3.12}$$

As the joint distribution of $(\alpha_0, \boldsymbol{\alpha})$ is multivariate normal, the conditional distribution of α_0 is also Gaussian with mean and variance given by

$$\mathrm{E}(\alpha_0|\boldsymbol{\alpha}, \boldsymbol{Y}) = \mathrm{E}(\alpha_0|\boldsymbol{\alpha}) = \boldsymbol{r}_\alpha R_\alpha^{-1} \boldsymbol{\alpha}, \tag{3.13}$$

$$\mathrm{Var}(\alpha_0|\boldsymbol{\alpha}, \boldsymbol{Y}) = \mathrm{Var}(\alpha_0|\boldsymbol{\alpha}) = 1 - \boldsymbol{r}_\alpha R_\alpha^{-1} \boldsymbol{r}_\alpha^T, \tag{3.14}$$

where $\boldsymbol{r}_\alpha$ is the correlation vector of the α process between $\boldsymbol{x}_0$ and $\boldsymbol{x}$. Furthermore, it can be seen that

$$\begin{aligned} \mathrm{E}(\alpha_0|\boldsymbol{Y}) &= \mathrm{E}\{\mathrm{E}(\alpha_0|\boldsymbol{\alpha},\boldsymbol{Y})|\boldsymbol{Y})\} \\ &= \mathrm{E}\{\mathrm{E}(\alpha_0|\boldsymbol{\alpha})|\boldsymbol{Y}\} \\ &= \mathrm{E}\{\boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{\alpha}|\boldsymbol{Y}\} \\ &= \boldsymbol{r}_\alpha R_\alpha^{-1}\mathrm{E}(\boldsymbol{\alpha}|\boldsymbol{Y}). \end{aligned} \tag{3.15}$$

Consequently, we plug the estimates of ϕ_α and $\mathrm{E}(\boldsymbol{\alpha}|\boldsymbol{Y})$ into (3.15) to get a plug-in best predictor of α_0.

It will also be of interest to predict a function of α_0, e.g., $\sigma^2\exp(\tau\alpha_0)$ and $\sigma\exp(\tau\alpha_0/2)$, the conditional variance and standard deviation of Y at unobserved location $\boldsymbol{x}_0$. Following the same scheme, we predict $h(\alpha_0)$ by $\mathrm{E}(h(\alpha_0)|\boldsymbol{Y})$ and the corresponding prediction variance by $\mathrm{Var}(h(\alpha_0)|\boldsymbol{Y})$. By integrating out $\boldsymbol{\alpha}$ in $\mathrm{E}(h(\alpha_0)|\boldsymbol{\alpha},\boldsymbol{Y})$ and $\mathrm{Var}(h(\alpha_0)|\boldsymbol{\alpha},\boldsymbol{Y})$, the predictor and prediction variance will be a function of $\mathrm{E}(\boldsymbol{\alpha}|\boldsymbol{Y})$, which is straightforward to evaluate.

3.2.2 Prediction of the Y process

Plug-in Best Predictor (PBP)

Given the latent process α and parameter $\boldsymbol{\psi}$, the joint distribution of Y_0 at unobserved location $\boldsymbol{x}_0$ and the vector $\boldsymbol{Y}$ at sampled sites $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ is heteroscedastic Gaussian. The covariance of $(Y_0, \boldsymbol{Y})$ conditional on $(\alpha_0, \boldsymbol{\alpha})$ is given by

$$\begin{aligned} \mathrm{Cov}(Y_0, \boldsymbol{Y}|\alpha_0, \boldsymbol{\alpha}) &= \sigma^2\exp\left(\frac{\tau\alpha_0}{2}\right)\boldsymbol{r}_z(\boldsymbol{x}_0,\boldsymbol{x})\mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}, \\ \mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{Y}|\boldsymbol{\alpha}) &= \sigma^2\mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}R_z\mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}, \\ \mathrm{Cov}(Y_0, Y_0|\alpha_0) &= \sigma^2\exp(\tau\alpha_0). \end{aligned} \tag{3.16}$$

where R_z is the $n \times n$ correlation matrix for $\boldsymbol{z}$ and $\boldsymbol{r}_z(\boldsymbol{x}_0, \boldsymbol{x})$ is the $1 \times n$ correlation vector between z_0 and $\boldsymbol{z}$. For ease of notation, we will write $\boldsymbol{r}_z$ for $\boldsymbol{r}_z(\boldsymbol{x}_0, \boldsymbol{x})$ and $\boldsymbol{r}_z^T$ for $\boldsymbol{r}_z(\boldsymbol{x}, \boldsymbol{x}_0)$. We will adopt similar notation for correlations of the α process. The diag$\{\exp(\tau\boldsymbol{\alpha}/2)\}$ refers to the $n \times n$ diagonal matrix with $\exp(\tau\boldsymbol{\alpha}/2)$ being the diagonal elements. The conditional mean and variance of the predictive distribution for Y_0 can be written as:

$$\mathrm{E}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0) = \boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta} + \exp(\tau\alpha_0/2)\boldsymbol{r}_z R_z^{-1}\mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}(\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}), \tag{3.17}$$

$$\mathrm{Var}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0) = \sigma^2\exp(\tau\alpha_0)(1 - \boldsymbol{r}_z R_z^{-1}\boldsymbol{r}_z^T), \tag{3.18}$$

where $G(\boldsymbol{x}) = (\boldsymbol{g}(\boldsymbol{x}_1), ..., \boldsymbol{g}(\boldsymbol{x}_n))^T$. As such, the best predictor $\mathrm{E}(Y_0|\boldsymbol{Y})$ can be obtained by integrating out $(\alpha_0, \boldsymbol{\alpha})$ with respect to $p(\alpha_0, \boldsymbol{\alpha}|\boldsymbol{Y})$ in (3.17). By use of equations (3.13) and (3.14) (the mean and variance of conditional distribution $\alpha_0|\boldsymbol{Y}, \boldsymbol{\alpha}$), integrating out $\alpha_0|\boldsymbol{Y}, \boldsymbol{\alpha}$, we obtain using properties of the lognormal distribution,

$$\begin{aligned}\mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{Y}, \boldsymbol{\alpha}) &= \mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{\alpha}) \\ &= \exp\left(\frac{\tau}{2}\boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{\alpha} + \frac{\tau^2}{8}(1 - \boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{r}_\alpha^T)\right),\end{aligned} \tag{3.19}$$

$$\begin{aligned}\mathrm{E}(\exp(\tau\alpha_0)|\boldsymbol{Y}, \boldsymbol{\alpha}) &= \mathrm{E}(\exp(\tau\alpha_0)|\boldsymbol{\alpha}) \\ &= \exp\left(\tau\boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{\alpha} + \frac{\tau^2}{2}(1 - \boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{r}_\alpha^T)\right).\end{aligned} \tag{3.20}$$

Combining equations (3.17) and (3.19), the best predictor $\mathrm{E}(Y_0|\boldsymbol{Y})$ can be obtained by

$$\begin{aligned}\mathrm{E}(Y_0|\boldsymbol{Y}) &= \mathrm{E}[\mathrm{E}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0)|\boldsymbol{Y}] \\ &= \mathrm{E}\{\mathrm{E}[\mathrm{E}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0)|\boldsymbol{Y}, \boldsymbol{\alpha}]|\boldsymbol{Y}\} \\ &= \mathrm{E}\big([\boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta} + \mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{Y}, \boldsymbol{\alpha})\boldsymbol{r}_z R_z^{-1} \\ &\quad \times\mathrm{diag}\{\exp(-\tau\boldsymbol{\alpha}/2)\}(\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta})]|\boldsymbol{Y}\big),\end{aligned} \tag{3.21}$$

where the last expectation is taken with respect to $p(\boldsymbol{\alpha}|\boldsymbol{Y})$. Since we do not have closed form expression for this posterior distribution, we cannot calculate (3.21) explicitly. Fortunately, we can evaluate such a function of $\boldsymbol{\alpha}$ by applying importance sampling and Monte Carlo integration as in Section 3.1.2. Similarly, by integrating out $\alpha_0|\boldsymbol{Y}, \boldsymbol{\alpha}$, the prediction variance $\mathrm{Var}(Y_0|\boldsymbol{Y})$ can be written as

$$\begin{aligned}
\mathrm{Var}(Y_0|\boldsymbol{Y}) &= \mathrm{E}(\mathrm{Var}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0)|\boldsymbol{Y}) + \mathrm{Var}(\mathrm{E}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0)|\boldsymbol{Y}), \\
\mathrm{E}(\mathrm{Var}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0)|\boldsymbol{Y}) &= \mathrm{E}\left([\sigma^2 \mathrm{E}(\exp(\tau\alpha_0)|\boldsymbol{Y}, \boldsymbol{\alpha})(1 - \boldsymbol{r}_z R_z^{-1} \boldsymbol{r}_z^T)]|\boldsymbol{Y}\right), \\
\mathrm{Var}(\mathrm{E}(Y_0|\boldsymbol{Y}, \boldsymbol{\alpha}, \alpha_0)|\boldsymbol{Y}) &= \mathrm{E}\big([\mathrm{E}(\exp(\tau\alpha_0)|\boldsymbol{Y}, \boldsymbol{\alpha})(\boldsymbol{r}_z R_z^{-1} \mathrm{diag}\{\exp(-\tau\boldsymbol{\alpha}/2)\} \\
&\quad \times (\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}))^2]|\boldsymbol{Y}\big) \\
&\quad - (\mathrm{E}(Y_0|\boldsymbol{Y}) - \boldsymbol{g}(\boldsymbol{x}_0)^T \boldsymbol{\beta})^2, \qquad (3.22)
\end{aligned}$$

where $\mathrm{E}(\exp(\tau\alpha_0)|\boldsymbol{Y}, \boldsymbol{\alpha})$ is calculated by use of (3.20) and $\mathrm{E}(Y_0|\boldsymbol{Y})$ is obtained through (3.21). The final expectations are taken with respect to $p(\boldsymbol{\alpha}|\boldsymbol{Y})$, which will be approximated by importance sampling and Monte Carlo integration. Once we have the parameter estimates $\hat{\boldsymbol{\psi}}$, we get the plug-in best predictor (PBP) and the plug-in prediction variance by plugging $\hat{\boldsymbol{\psi}}$ into (3.21) and (3.22).

Note that the plug-in prediction variance calculated above depends on the observations $\boldsymbol{Y}$. Recall the prediction variance for Gaussian process model,

$$\mathrm{Var}(Y_0|\boldsymbol{Y}) = \sigma_g^2(1 - \boldsymbol{r}_g R_g^{-1} \boldsymbol{r}_g^T), \qquad (3.23)$$

where the subscript "g" indicates parameter/correlation matrices for Gaussian process model. Equation (3.23) does not depend on the observation vector $\boldsymbol{Y}$. We would expect that the observation-dependent SHP prediction variance accounts for spatial heterogeneity.

Plug-in Best Linear Unbiased Predictor (PBLUP)

If the parameter $\boldsymbol{\psi}$ were known, the best linear unbiased predictor (BLUP) could be readily computed. By integrating out the latent process α, the mean

vector and covariance matrix of the unconditional joint distribution for $(Y_0, \boldsymbol{Y})$ is given by

$$\begin{bmatrix} Y_0 \\ \boldsymbol{Y} \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{g}(\boldsymbol{x}_0)^T \\ G(\boldsymbol{x}) \end{bmatrix} \boldsymbol{\beta}, \sigma^2 \exp(\tau^2/2) \begin{bmatrix} 1 & \boldsymbol{r}_y \\ \boldsymbol{r}_y^T & R_y \end{bmatrix} \right), \tag{3.24}$$

where $\boldsymbol{r}_y$ stands for $\boldsymbol{r}_y(\boldsymbol{x}_0, \boldsymbol{x})$, and R_y and $\boldsymbol{r}_y$ are computed by (2.2).

Since the parameters are assumed known, the BLUP (in fact, the best linear predictor, BLP) of Y_0 is given by

$$\hat{Y_0} = \boldsymbol{g}(\boldsymbol{x}_0)^T \boldsymbol{\beta} + \boldsymbol{r}_y R_y^{-1} (\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}), \tag{3.25}$$

and the corresponding prediction variance is given by

$$\mathrm{Var}(\hat{Y_0}) = \sigma^2 \exp(\tau^2/2)(1 - \boldsymbol{r}_y R_y^{-1} \boldsymbol{r}_y^T). \tag{3.26}$$

For a Gaussian process model, the empirical BLUP is obtained by plugging in the estimated correlation parameters together with the profiled mean and variance estimates, which are also GLS (Generalized Least Squares) estimates. For the SHP model, there is no explicit expression for the likelihood. It is impossible to profile out $\boldsymbol{\beta}$ and σ^2. Therefore, we will plug in the whole parameter estimates $\hat{\boldsymbol{\psi}}$ into (3.25) and (3.26) to get the Plug-in Best Linear Unbiased Predictor and the corresponding prediction variance.

In Chapter 4, we will compare the prediction performance of PBP and PBLUP through simulations. The advantage of PBP is that it accounts for the spatial inhomogeneities by incorporating the α process estimates. The PBLUP is computationally faster and its correlation function with smoothed nugget property accounts for small scale variations. When we plug in the estimated parameter $\hat{\boldsymbol{\psi}}$ in equations (3.22) and (3.26) to get the prediction variances, we underestimate the prediction variances by ignoring the variances arising from estimating $\boldsymbol{\psi}$. This is a common issue in statistical literature. It is difficult to incorporate the variance from parameter estimation into prediction inference when using maximum likelihood methods. A Bayesian approach has the advantage of incorporating uncertainties in the prediction inference. We leave this as a topic of future research.

3.3 Implementation

When using (3.3) to calculate the likelihood, it is a common practice to use "common random numbers" (CRNs) to generate $\boldsymbol{\alpha}^{(1)}, ..., \boldsymbol{\alpha}^{(N)}$ for different parameter values $\boldsymbol{\psi}$. CRNs will help improve the smoothness of the likelihood and facilitate the convergence of estimates. Due to the numerical optimization in high dimensional space, we have more computational issues to discuss for maximum likelihood estimation procedure.

3.3.1 Estimating the posterior mode

The performance of the importance density (3.9) heavily depends on the posterior mode $\boldsymbol{\alpha}^*$, which is obtained by maximizing (3.4). Since the dimensionality of $\boldsymbol{\alpha}$ is the same as the number of observations, it is difficult to find $\boldsymbol{\alpha}^*$, especially for large data sets. To make this optimization more feasible and accurate, we approximate the α process by a low-dimensional process and do optimization over the low-dimensional space.

In Section 1.1, we have discussed that Higdon (2002) proposes a process convolution method to construct a continuous spatial model. A Gaussian process $\alpha(\boldsymbol{x})$ over a spatial region D can be constructed through convolving a continuous white noise $\omega(\boldsymbol{x})$ with a smoothing kernel $k(\boldsymbol{x})$, i.e.,

$$\alpha(\boldsymbol{x}) = \int_D k(\boldsymbol{u} - \boldsymbol{x})\omega(\boldsymbol{u})d\boldsymbol{u}, \quad \text{for } \boldsymbol{x} \in D. \tag{3.27}$$

The resulting covariance function for $\alpha(\boldsymbol{x})$ is given by

$$\gamma(\boldsymbol{x}, \boldsymbol{x}') = \text{Cov}(\alpha(\boldsymbol{x}), \alpha(\boldsymbol{x}')) = \int_D k(\boldsymbol{u} - \boldsymbol{x})k(\boldsymbol{u} - \boldsymbol{x}')du = \int_D k(\boldsymbol{u} - \boldsymbol{d})k(\boldsymbol{u})d\boldsymbol{u}, \tag{3.28}$$

where $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}'$. In case of isotropy, $\gamma(\boldsymbol{x}, \boldsymbol{x}')$ only depends on $||\boldsymbol{d}|| = ||\boldsymbol{x} - \boldsymbol{x}'||$. There is a one-to-one relationship between the smoothing kernel $k(\boldsymbol{d})$ and the covariance $\gamma(||\boldsymbol{d}||)$ under mild conditions. The relationship is based on the convolution theorem for Fourier transforms. For example, the Gaussian kernel $k(\boldsymbol{d}) \propto$

$\exp(-2\phi||\boldsymbol{d}||^2)$ corresponds to Gaussian covariance function $\gamma(\boldsymbol{d}) \propto \exp(-\phi||\boldsymbol{d}||^2)$. More details about this relationship and various kernels and their induced covariance functions can be found in Higdon (2002).

We restrict the latent process $\omega(\boldsymbol{u})$ to be nonzero at a coarse lattice of spatial sites $\boldsymbol{\kappa}_1, ..., \boldsymbol{\kappa}_J$. We take J far less than the dimensionality of the observations. In this case, a small number of parameters $\omega(\boldsymbol{\kappa}_1), ..., \omega(\boldsymbol{\kappa}_J)$ effectively controls the entire spatial process $\alpha(\boldsymbol{x})$. The resulting continuous α process is then approximated by

$$\alpha(\boldsymbol{x}) = \sum_{j=1}^{J} \omega_j k_{\phi_\alpha}(\boldsymbol{x} - \boldsymbol{\kappa}_j). \tag{3.29}$$

where k_{ϕ_α} is the kernel with parameter ϕ_α. In matrix multiplication notation, we can write $\boldsymbol{\alpha} = K\boldsymbol{\omega}$, where K is the $n \times J$ matrix given by

$$K = \begin{bmatrix} k_{\phi_\alpha}(\boldsymbol{x}_1 - \boldsymbol{\kappa}_1) & k_{\phi_\alpha}(\boldsymbol{x}_1 - \boldsymbol{\kappa}_2) & \cdots & k_{\phi_\alpha}(\boldsymbol{x}_1 - \boldsymbol{\kappa}_J) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ k_{\phi_\alpha}(\boldsymbol{x}_n - \boldsymbol{\kappa}_1) & k_{\phi_\alpha}(\boldsymbol{x}_n - \boldsymbol{\kappa}_2) & \cdots & k_{\phi_\alpha}(\boldsymbol{x}_n - \boldsymbol{\kappa}_J) \end{bmatrix}.$$

We substitute (3.29) into (3.4) and maximize the likelihood with respect to $\boldsymbol{\omega}$ in dimension J. The $\alpha^*(\boldsymbol{x})$ is then approximated by

$$\alpha^*(\boldsymbol{x}) \approx \sum_{j=1}^{J} \omega_j^* k_{\phi_\alpha}(\boldsymbol{x} - \boldsymbol{\kappa}_j), \tag{3.30}$$

where $\boldsymbol{\omega}^* = (\omega_1^*, \omega_2^*, ..., \omega_J^*)$ is the mode of the low-dimensional approximate joint likelihood.

We show the efficacy of applying the low-dimensional approximation on estimating $\boldsymbol{\alpha}^*$ through a 1-d example. Figure 3.1 shows the true sample paths of α and Y that we simulated. We set $J = 10$ in this example. We choose five fixed knots that are equally-spaced locations within the domain $[0, 2]$. The other five knots are randomly sampled with probability proportional to $|\boldsymbol{Y}|$. As such, the kernel centers can cover the whole domain well and put more emphasis on bump locations. We will

use this strategy to determine kernel centers throughout this dissertation, i.e., fixed knots which evenly spread within the domain and the remaining knots randomly selected with probabilities proportional to $|\boldsymbol{Y}|$.

Since the joint likelihood (3.4) is not convex in terms of $\boldsymbol{\alpha}$, there could be multiple maxima. Even for $J = 10$, the optimization is in a mildly high-dimensional space and it is better to try different initial values in order to find a good estimate. We try three different initial values $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$.

We try to find meaningful initial values instead of randomly drawing from the prior $N(0, \rho_\alpha)$. The first initial value $\boldsymbol{\alpha}_1$ comes from "solving" for $\boldsymbol{\alpha}$ from its generating equation after plugging in a rough estimate of Z. Recall the Gaussian process model (with nugget) and the SHP model:

$$\text{Gaussian process model: } Y(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta} + \sigma Z(\boldsymbol{x}) + \epsilon(\boldsymbol{x}), \tag{3.31}$$

$$\text{SHP model: } Y(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta} + \sigma \exp\left(\frac{\tau\alpha(\boldsymbol{x})}{2}\right) Z(\boldsymbol{x}). \tag{3.32}$$

We first fit the data using the Gaussian process model (3.31). Denoting the maximum likelihood estimates from the Gaussian process model (with nugget, $\epsilon(\boldsymbol{x}) \sim$ iid $\mathrm{N}(0, \sigma_\epsilon^2)$) by $(\tilde{\boldsymbol{\beta}}, \tilde{\phi}, \tilde{\sigma}^2, \tilde{\sigma_\epsilon}^2)$, the "smoothed" response $\tilde{Y}$ at an observed location $\boldsymbol{x}_o$ is computed by

$$\tilde{Y}(\boldsymbol{x}_o) = \boldsymbol{g}(\boldsymbol{x}_o)^T\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{r}}\tilde{R}^{-1}(\boldsymbol{Y} - G\tilde{\boldsymbol{\beta}}),$$

where $\tilde{\boldsymbol{r}}$ is the estimated correlation vector between $Y(\boldsymbol{x}_o)$ and $(Y(\boldsymbol{x}_1), ..., Y(\boldsymbol{x}_n))$ by ignoring the nugget term, $\tilde{R}$ is the estimated covariance matrix for $(Y(\boldsymbol{x}_1), ..., Y(\boldsymbol{x}_n))$ and $G = (\boldsymbol{g}(\boldsymbol{x}_1), ..., \boldsymbol{g}(\boldsymbol{x}_n))^T$. An estimate of $Z(\boldsymbol{x}_o)$ is obtained by by $\tilde{Z}(\boldsymbol{x}_o) = (\tilde{Y}(\boldsymbol{x}_o) - \boldsymbol{g}(\boldsymbol{x}_o)^T\tilde{\boldsymbol{\beta}})/\tilde{\sigma}$. In order to guarantee that $\tilde{Y}(\boldsymbol{x}_o) \neq \boldsymbol{g}(\boldsymbol{x}_o)^T\tilde{\boldsymbol{\beta}}$, we set a lower bound for estimating σ_ϵ^2 to avoid a zero estimate. We treat this $\tilde{Z}(\boldsymbol{x}_o)$ as an estimate of the Z process realization at the observed locations $\boldsymbol{x}_o$ in equation (3.32). We get the first initial value $\boldsymbol{\alpha}_1(\boldsymbol{\beta}, \sigma^2, \tau)$ by solving (3.32),

$$\alpha_1(\boldsymbol{x}_o; \boldsymbol{\beta}, \sigma^2, \tau) = [\log(Y(\boldsymbol{x}_o) - \boldsymbol{g}(\boldsymbol{x}_o)^T\boldsymbol{\beta})^2 - \log\sigma^2 - \log\tilde{Z}(\boldsymbol{x}_o)^2]/\tau. \tag{3.33}$$

The second initial value $\boldsymbol{\alpha}_2(\boldsymbol{\beta}, \sigma^2, \tau)$ is also obtained by solving (3.32). Instead of using the rough estimate $\tilde{Z}$, we set all $\log Z^2$ values equal to -1.27, which is the mean of the log chi-square distribution with one degree of freedom, i.e.,

$$\alpha_2(\boldsymbol{x}_o; \boldsymbol{\beta}, \sigma^2, \tau) = [\log(Y(\boldsymbol{x}_o) - \boldsymbol{g}(\boldsymbol{x}_o)^T\boldsymbol{\beta})^2 - \log\sigma^2 + 1.27]/\tau. \tag{3.34}$$

The third initial value $\boldsymbol{\alpha}_3$ is simply a vector of zeros, which is the mean of the prior. When applying the low-dimensional approximation scheme on estimating $\boldsymbol{\alpha}^*$, we first transfer the initial vectors $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ to $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_3$, by taking $\boldsymbol{\omega} = (K^TK)^{-1}\boldsymbol{\alpha}$. By plugging $\boldsymbol{\alpha} = K\boldsymbol{\omega}$ into equation (3.4), we apply three initial values $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_3$ and obtain three estimated modes $\boldsymbol{\omega}_1^*$, $\boldsymbol{\omega}_2^*$ and $\boldsymbol{\omega}_3^*$, from which we chose the one corresponding to the largest joint likelihood value as the final mode estimate $\boldsymbol{\omega}^*$. The $\boldsymbol{\alpha}^*$ is then approximated by equation (3.30).

For illustration, we use true parameters to estimate $\boldsymbol{\alpha}^*$ for the SHP realization shown in Figure 3.1. Figures 3.2, 3.3 and 3.4 show the results of estimating $\boldsymbol{\alpha}^*$ by use of three initial values and two methods: optimizing with respect to $\boldsymbol{\alpha}$ directly and optimization in low-dimensional space (with respect to $\boldsymbol{w}$). We can see that the $\boldsymbol{\alpha}^*$ obtained by the low-dimensional optimization approach are close to the true $\boldsymbol{\alpha}$ values, while the $\boldsymbol{\alpha}^*$ obtained by direct optimization with respect to $\boldsymbol{\alpha}$ do not move from the initial values too much and are quite different from the true $\boldsymbol{\alpha}$ values. Although the posterior mode does not exactly equal the true $\boldsymbol{\alpha}$, it is expected that they are close. We also notice that the estimate is robust to initial values for low-dimensional optimization in the sense that different initial values lead to similar $\boldsymbol{\alpha}^*$ estimates. But the results of optimization with respect to $\boldsymbol{\alpha}$ depend quite strongly on the initial values. In Table 3.1, we list the mean square error of the estimated mode with respect to the true $\boldsymbol{\alpha}$ values, $n^{-1}\sum_{i=1}^{n}(\alpha^* - \alpha)^2$, which illustrates the above conclusions quantitatively. From Table 3.1, we see that the joint likelihood values of the low-dimensional approach are very close to each other

for different initial values and much greater than those from optimizing with respect to $\boldsymbol{\alpha}$ directly. Overall, the low-dimensional approach improves the posterior mode estimate dramatically over optimizing in the original high-dimensional space.

In summary, to evaluate a likelihood value at a set of parameter values $\boldsymbol{\psi} = (\sigma^2, \tau^2, \phi_\alpha, \phi_z, \boldsymbol{\beta})$, we take the following steps:

Step 1 Fit a Gaussian process model with nugget (equation (3.31)) and get a rough estimate of Z, i.e., $\tilde{Z}$.

Step 2 Solve equations (3.33) and (3.34) to get initial values $\boldsymbol{\alpha}_1(\boldsymbol{\beta}, \sigma^2, \tau)$ and $\boldsymbol{\alpha}_2(\boldsymbol{\beta}, \sigma^2, \tau)$.

Step 3 Transform the three initial values $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ (vector of zeros) to initial values in $\boldsymbol{\omega}$ space, i.e., $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_3$.

Step 4 Optimize (3.4) with respect to $\boldsymbol{\omega}$ by plugging in $\boldsymbol{\alpha} = K\boldsymbol{\omega}$, using three initial values $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_3$.

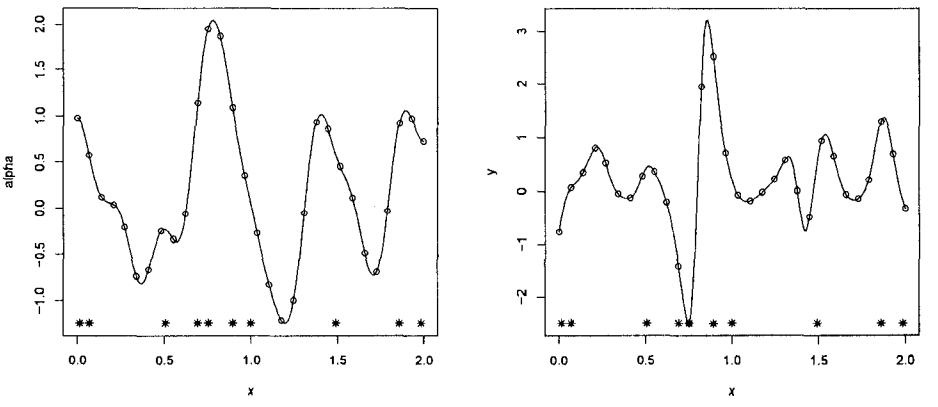
Step 5 Choose $\boldsymbol{\omega}^*$ from $\boldsymbol{\omega}_1^*$, $\boldsymbol{\omega}_2^*$ and $\boldsymbol{\omega}_3^*$, the one corresponding to the largest joint likelihood value.

Step 6 Take $\boldsymbol{\alpha}^* = K\boldsymbol{\omega}^*$ and plug $\boldsymbol{\alpha}^*$ and $\boldsymbol{\psi}$ into formulas (3.10) to get the importance density.

Step 7 Sample a large number of $\boldsymbol{\alpha}$'s from the importance density and calculate the likelihood by equation (3.3).

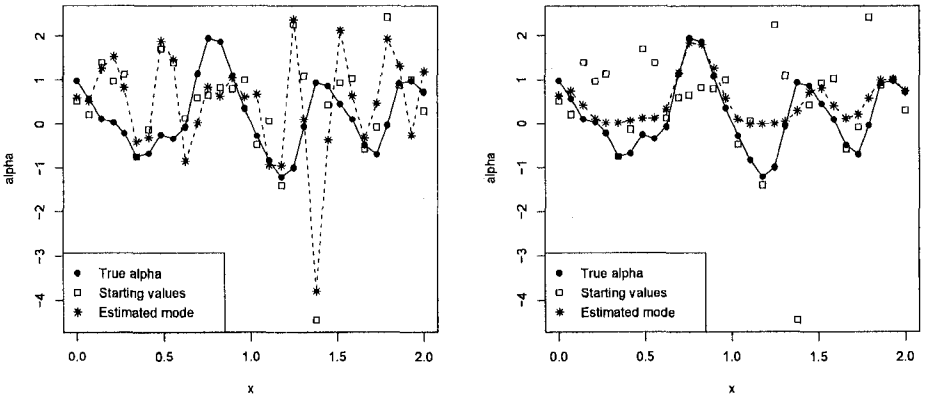
3.3.2 Estimating σ^2

By simulation, we found that the likelihood is flat for a wide range of σ^2 values, which brings difficulties in maximum likelihood estimation for σ^2. It is impossible to profile out σ^2 in equation (3.2). We tried to profile σ^2 out of the joint density



(a) (b)

Figure 3.1: Estimating the posterior mode (I). Panel (a) shows the true α sample path. Panel (b) shows the true Y process realization. Circles are sampled points. Stars on the x-axis are locations of kernel centers. We take $\beta = 0$, $\sigma^2 = 0.2$, $\tau^2 = 4$, $\phi_\alpha = 40$ and $\phi_z = 80$. Both ρ_α and ρ_z are Gaussian correlation functions. The true sample path is based on 200 equal-spaced points on $[0, 2]$. The sampled locations are 30 equal-spaced points.



(a) (b)

Figure 3.2: Estimating the posterior mode (II). Panel (a) shows the optimization with respect to $\boldsymbol{\alpha}$ directly. Panel (b) shows the optimization by applying low-dimensional approximation. We take the starting values $\boldsymbol{\alpha}_1$ in each panel and plot the true $\boldsymbol{\alpha}$, the starting values and the estimated posterior mode for comparison.

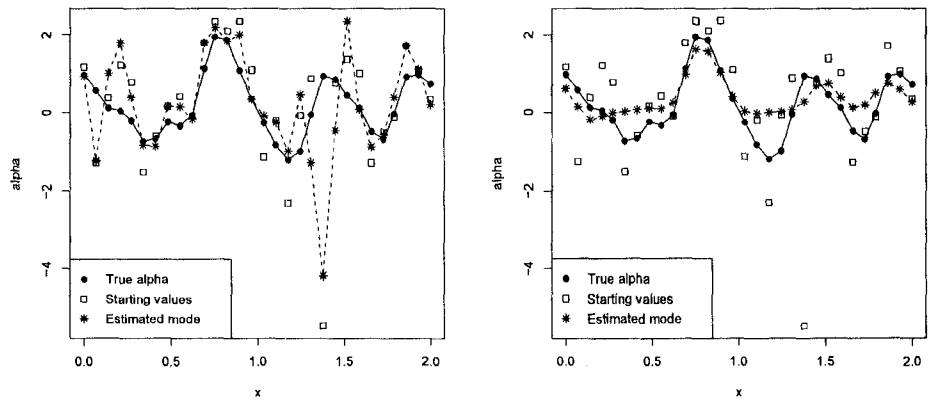


Figure 3.3: Estimating the posterior mode (III). Panel (a) shows the optimization with respect to $\boldsymbol{\alpha}$ directly. Panel (b) shows the optimization by applying low-dimensional approximation. We take the starting values $\boldsymbol{\alpha}_2$ in each panel and plot the true $\boldsymbol{\alpha}$, the starting values and the estimated posterior mode for comparison.

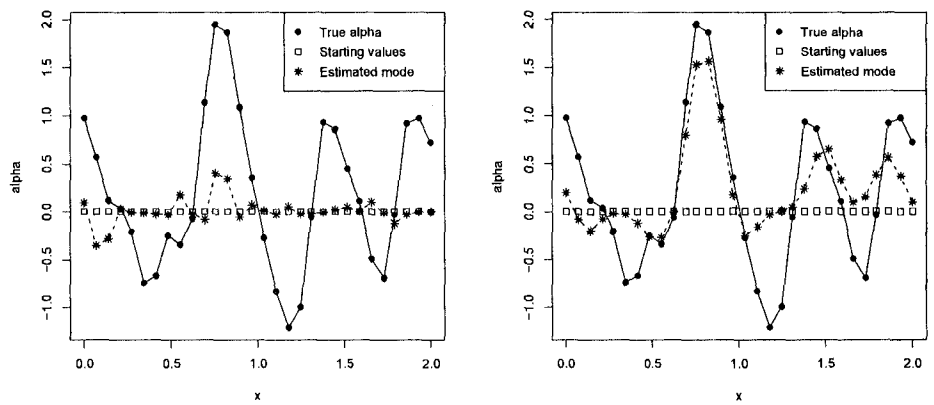


Figure 3.4: Estimating the posterior mode (IV). Panel (a) shows the optimization with respect to $\boldsymbol{\alpha}$ directly. Panel (b) shows the optimization by applying low-dimensional approximation. We take the starting values $\boldsymbol{\alpha}_3$ (a vector of zeros) in each panel and plot the true $\boldsymbol{\alpha}$, the starting values and the estimated posterior mode for comparison.

Table 3.1: Comparison of MSE and log-likelihood values for estimating posterior mode by direct optimization and low-dimensional approximation approach.

	MSE.α	MSE.ω	loglik.α	loglik.ω
$\boldsymbol{\alpha}_1$	2.12	0.25	-1396.4	-70.5
$\boldsymbol{\alpha}_2$	1.54	0.26	-1581.8	-66.8
$\boldsymbol{\alpha}_3$	0.63	0.27	-185.2	-68.9

[a] loglik.α and loglik.ω refer to the log likelihood values (up to a constant); MSE.α and MSE.ω refer to the mean square errors of the estimated mode with respect to the true α values at 30 sampled locations. The suffixes α and ω refer to optimization with respect to α directly and by applying low-dimensional methodology, respectively.

of $(\boldsymbol{Y}, \boldsymbol{\alpha})$, i.e., equation (3.1) when finding $\boldsymbol{\alpha}^*$. Accordingly, we maximize (3.1) with respect to $\boldsymbol{\alpha}$ and σ^2 (having closed form) simultaneously. It turns out that in simulations the σ^2 estimate obtained in this way is highly positively biased. Therefore, we explore an alternative way to estimate σ^2. From SHP model (2.1), by incorporating the correlations in the observations, the expected value of the sample variance is

$$\mathrm{E}(s^2) = \frac{\sigma^2 \exp(\frac{\tau^2}{2})(n^2 - \sum_i \sum_j \rho_Y(i,j))}{n(n-1)}.$$

Therefore, an unbiased estimator for σ^2 would be

$$\hat{\sigma}^2 = \frac{n(n-1)\exp(-\frac{\tau^2}{2})s^2}{n^2 - \sum_i \sum_j \rho_Y(i,j)} \tag{3.35}$$

if τ^2 and ρ_Y were known. We propose to fix σ^2 at the sample variance and obtain maximum likelihood estimates for τ^2, ϕ_α, ϕ_z and $\boldsymbol{\beta}$. Then by plugging maximum likelihood estimates of the other parameters in equation (3.35), we get the final estimate for σ^2.

Because the data have a heavy-tailed distribution, we use a robust alternative to s^2 in estimating σ^2. We use the scaled Median Absolute Deviation (MAD) estimate, a robust estimate for σ developed by Johnstone and Silverman (1997),

$$\mathrm{MAD} = \frac{\mathrm{median}(|\boldsymbol{Y} - \mathrm{median}(\boldsymbol{Y})|)}{0.6745}, \tag{3.36}$$

where 0.6745 is the standard normal MAD. We replace s^2 in equation (3.35) by the square of MAD in equation (3.36).

3.4 A Low-Dimensional Approximation Model

In Section 3.3.1, we introduced a low-dimensional approximation for the latent process α. We replace $\boldsymbol{\alpha}$ by its low-dimensional approximation (3.29) in the joint likelihood equation (3.4). The mode $\boldsymbol{\alpha}^*$ is obtained by (3.30) after optimizing (3.4) to get $\boldsymbol{\omega}^*$. This low-dimensional approximation scheme reduces computational load and increases accuracy in the numerical optimization procedure. But the importance density calculated from (3.9) and (3.10) is still of dimension n. When the data set is large, this n-dimensional importance density is cumbersome and not feasible for reliable computation. We would like to develop a low-dimensional importance density by approximating the α process completely.

We approximate the SHP model by

$$\begin{aligned} Y(\boldsymbol{x}) &= \boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta} + W(\boldsymbol{x}), \\ W(\boldsymbol{x}) &= \sigma \exp\left(\frac{\tau K\boldsymbol{\omega}}{2}\right) Z(\boldsymbol{x}), \qquad \sigma > 0, \qquad \tau > 0, \end{aligned} \tag{3.37}$$

where $\boldsymbol{\omega} \sim \mathrm{N}(0, I_J)$. We have introduced the notation K in Section 3.3.1. The log joint density of $(\boldsymbol{Y}, \boldsymbol{\omega})$ is

$$l(\boldsymbol{\psi}; \boldsymbol{Y}, \boldsymbol{\omega}) = -\frac{n}{2}\log(2\pi) + l(\boldsymbol{\psi}; \boldsymbol{y}|\boldsymbol{\omega}) - \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}, \tag{3.38}$$

where $l(\boldsymbol{\psi}; \boldsymbol{y}|\boldsymbol{\omega}) := \log p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta})$ and

$$p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}) \sim \mathrm{N}\left(G(\boldsymbol{x})\boldsymbol{\beta}, \sigma^2 \mathrm{diag}\left\{\exp\left(\frac{\tau K\boldsymbol{\omega}}{2}\right)\right\} R_z \mathrm{diag}\left\{\exp\left(\frac{\tau K\boldsymbol{\omega}}{2}\right)\right\}\right).$$

For this approximation model, following the same derivation procedure in Section 3.1, the J-dimensional importance density is given by

$$p_a(\boldsymbol{\omega}|\boldsymbol{Y}, \boldsymbol{\psi}) \sim \mathrm{N}(\boldsymbol{\omega}^*, (K^* + I_J)^{-1}), \tag{3.39}$$

where

$$
\begin{aligned}
K^* &= \frac{\tau^2}{4\sigma^2} K^T (B + \mathrm{diag}\{\boldsymbol{c}\}) K \\
B &= \mathrm{diag}\left\{\exp\left(-\frac{\tau K \boldsymbol{\omega}^*}{2}\right)\right\} \mathrm{diag}\{\boldsymbol{Y} - \boldsymbol{g}^T\boldsymbol{\beta}\} R_z^{-1} \\
&\times \mathrm{diag}\{\boldsymbol{Y} - \boldsymbol{g}^T\boldsymbol{\beta}\} \mathrm{diag}\left\{\exp\left(-\frac{\tau K \boldsymbol{\omega}^*}{2}\right)\right\}, \\
\boldsymbol{c} &= \left(\exp\left(-\frac{\tau K \boldsymbol{\omega}^*}{2}\right)\right)^T \mathrm{diag}\{\boldsymbol{Y} - \boldsymbol{g}^T\boldsymbol{\beta}\} R_z^{-1} \mathrm{diag}\{\boldsymbol{Y} - \boldsymbol{g}^T\boldsymbol{\beta}\} \\
&\times \mathrm{diag}\left\{\exp\left(-\frac{\tau K \boldsymbol{\omega}^*}{2}\right)\right\}.
\end{aligned} \tag{3.40}
$$

Using the approximation model (3.37) and the low-dimensional importance density formulated by (3.39) and (3.40), we can get maximum likelihood estimates of the parameters through a faster optimization procedure. A function of the latent vector $\boldsymbol{\omega}$ can be estimated through a strategy similar to that for estimating $\boldsymbol{\alpha}$ as shown in Section 3.1.2, i.e., the best predictor $\mathrm{E}(h(\boldsymbol{\omega})|\boldsymbol{Y})$ is

$$
\begin{aligned}
\mathrm{E}[h(\boldsymbol{\omega})|\boldsymbol{Y}] &= \int h(\boldsymbol{\omega}) p(\boldsymbol{\omega}|\boldsymbol{Y}, \boldsymbol{\psi}) d\boldsymbol{\omega} \\
&= \int h(\boldsymbol{\omega}) \frac{p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\phi_\alpha)}{p(\boldsymbol{Y}|\boldsymbol{\psi})} d\boldsymbol{\omega} \\
&= \frac{\int h(\boldsymbol{\omega}) p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\phi_\alpha) d\boldsymbol{\omega}}{\int p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\phi_\alpha) d\boldsymbol{\omega}} \\
&= \frac{\mathrm{E}_a[h(\boldsymbol{\omega}) p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\phi_\alpha)/p_a(\boldsymbol{\omega}|\boldsymbol{Y}, \boldsymbol{\psi})]}{\mathrm{E}_a[p(\boldsymbol{Y}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\phi_\alpha)/p_a(\boldsymbol{\omega}|\boldsymbol{Y}, \boldsymbol{\psi})]}.
\end{aligned} \tag{3.41}
$$

The last step will be approximated by Monte Carlo integration using samples $\boldsymbol{\omega}^{(1)}, ..., \boldsymbol{\omega}^{(N)}$ drawn from $p_a(\boldsymbol{\omega}|\boldsymbol{Y}, \hat{\boldsymbol{\psi}})$.

Based on the low-dimensional approximation model (3.37), the prediction of Y_0 at unobserved location $\boldsymbol{x}_0$ is straightforward. Conditioning on the latent vector $\boldsymbol{\omega}$, $(Y_0, \boldsymbol{Y})$ are heteroscedastic Gaussian. The conditional mean and variance for Y_0 are

$$
\mathrm{E}(Y_0|\boldsymbol{Y}, \boldsymbol{\omega}) = \boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta} + \exp(\tau \boldsymbol{k}_0 \boldsymbol{\omega}) \boldsymbol{r}_z R_z^{-1} \mathrm{diag}\{\exp(-\tau K \boldsymbol{\omega}/2)\}(\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}), \tag{3.42}
$$

$$\mathrm{Var}(Y_0|\boldsymbol{Y},\boldsymbol{\omega}) = \sigma^2 \exp(\tau \boldsymbol{k}_0\boldsymbol{\omega})(1 - \boldsymbol{r}_z R_z^{-1}\boldsymbol{r}_z^T). \tag{3.43}$$

Furthermore, the prediction variance $\mathrm{Var}(Y_0|\boldsymbol{Y})$ can be written as

$$\begin{aligned}
\mathrm{Var}(Y_0|\boldsymbol{Y}) &= \mathrm{E}(\mathrm{Var}(Y_0|\boldsymbol{Y},\boldsymbol{\omega})|\boldsymbol{Y}) + \mathrm{Var}(\mathrm{E}(Y_0|\boldsymbol{Y},\boldsymbol{\omega})|\boldsymbol{Y}), \\
\mathrm{Var}(\mathrm{E}(Y_0|\boldsymbol{Y},\boldsymbol{\omega}|\boldsymbol{Y}) &= \mathrm{E}\big([\exp(\tau \boldsymbol{k}_0\boldsymbol{\omega})(\boldsymbol{r}_z R_z^{-1}\mathrm{diag}\{\exp(-\tau K\boldsymbol{\omega}/2)\} \\
&\quad \times(\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}))^2]|\boldsymbol{Y}\big) \\
&\quad -(\mathrm{E}(Y_0|\boldsymbol{Y}) - \boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta})^2,
\end{aligned} \tag{3.44}$$

where $\boldsymbol{k}_0 = (k_{\phi_\alpha}(\boldsymbol{x}_0 - \boldsymbol{\kappa}_1), k_{\phi_\alpha}(\boldsymbol{x}_0 - \boldsymbol{\kappa}_2), ..., k_{\phi_\alpha}(\boldsymbol{x}_0 - \boldsymbol{\kappa}_J))$. Given the parameter $\boldsymbol{\psi}$, equations (3.42), (3.43) and (3.44) are all functions of the latent vector $\boldsymbol{\omega}$. Therefore, we can plug in the parameter estimates $\hat{\boldsymbol{\psi}}$ and utilize (3.41) to compute the plug-in best predictor and the prediction variance by Monte Carlo integration.

An important issue of estimation and prediction by use of the approximation model (3.37) is the determination of the kernel centers $(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, ...\boldsymbol{\kappa}_J)$. We use the same strategies as in Section 3.3.1, i.e., fixed knot locations that evenly spread within the domain and the remaining knots randomly selected by use of probabilities proportional to $|\boldsymbol{Y}|$. We actually try a large number (ranging from 100 to 1000 depending on the application) of randomly selected knots and choose the one that yields the largest joint likelihood value.

3.5 SHP with Replicates

In practice, some spatial processes are measured at a finite set of locations over a region of interest D at regular times. For example, in studies of air pollution, some measurements of interest are recorded at monitoring stations at regular time intervals (daily, monthly, yearly,.., etc.). In this situation, the stochastic process can be described by $Y(\boldsymbol{x}, t)$ where $\boldsymbol{x} \in D$ and t represents time. Frequently, the aims are to estimate the process $Y(\boldsymbol{x}, t)$ and to predict the process at locations and times

which are not being measured. The introduction of time into spatial modeling increases the model complexity and computation burden. Spatial-temporal modeling (Banerjee et al. (2003)) considers the spatial correlation, temporal correlation and possibly the interaction of spatial and time trend simultaneously. Research in the area of heterogeneous spatial covariance modeling simplifies the temporal aspect, in the sense that the observations are assumed independent in time, probably obtained after detrending, see Sampson and Guttorp (1992), Schmidt and O'Hagan (2003), Damian et al. (2001) and Nychka et al. (2002). We denote by $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T$ the T independent n-dimensional vectors of observations. The sample covariance matrix S can be calculated (S is non-singular if $T > n$). Most approaches of modeling nonstationary spatial covariances work with the likelihood of S or smoothing S by some modern statistical techniques.

We can extend the SHP model (2.1) to fit the framework of such a problem. We assume that the spatial processes $Y_1, ..., Y_T$ come from a SHP model conditional on a common α process, i.e.,

$$\begin{aligned} Y_t(\boldsymbol{x}) &= \boldsymbol{g}_t(\boldsymbol{x})^T\boldsymbol{\beta} + W_t(\boldsymbol{x}), \\ W_t(\boldsymbol{x}) &= \sigma \exp\left(\frac{\tau\alpha(\boldsymbol{x})}{2}\right) Z_t(\boldsymbol{x}), \qquad \sigma > 0, \qquad \tau > 0, \end{aligned} \tag{3.45}$$

where $Z_t(\boldsymbol{x}), t = 1, ..., T$ are independent stationary Gaussian processes with mean 0, variance 1 and correlation functions ρ_z. The latent process $\alpha(\boldsymbol{x})$, independent of t, is used to model the spatially correlated variance process. Therefore, $Y_1, ..., Y_T$ are conditionally independent given α.

The joint likelihood of model (3.45) is given by

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\alpha}|\boldsymbol{\psi}) &= p(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha) \\ &= \prod_{t=1}^{T} p(\boldsymbol{Y}_t|\boldsymbol{\alpha}, \boldsymbol{\theta})p(\boldsymbol{\alpha}|\phi_\alpha). \end{aligned} \tag{3.46}$$

Correspondingly, the log density of $(\boldsymbol{Y}, \boldsymbol{\alpha})$ is given by

$$l(\boldsymbol{\psi}; \boldsymbol{Y}, \boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|R_\alpha|^{-1} + \sum_{t=1}^{T} l(\boldsymbol{\theta}; \boldsymbol{Y}_t|\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^T R_\alpha^{-1}\boldsymbol{\alpha}. \quad (3.47)$$

Comparing this log density with equation (3.4), the only difference is to replace $l(\boldsymbol{\theta}; \boldsymbol{Y}|\boldsymbol{\alpha})$ by $\sum_{t=1}^{T} l(\boldsymbol{\theta}; \boldsymbol{Y}_t|\boldsymbol{\alpha})$. Therefore, we follow a strategy similar to that in Section 3.1.1 to derive the importance density, which is still of the form (3.9), i.e.,

$$p_a(\boldsymbol{\alpha}|\boldsymbol{Y}, \boldsymbol{\psi}) \sim \mathrm{N}(\boldsymbol{\alpha}^*, (K^* + R_\alpha^{-1})^{-1}), \quad (3.48)$$

with K^* calculated by

$$\begin{aligned}
K^* &= \sum_{t=1}^{T} K_t^*, \\
K_t^* &= \frac{\tau^2}{4\sigma^2}(B_t + \mathrm{diag}\{\boldsymbol{c_t}\}), \\
B_t &= \mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right\} \\
&\times \mathrm{diag}\{\boldsymbol{Y}_t - \boldsymbol{g}^T\boldsymbol{\beta}\}R_z^{-1}\mathrm{diag}\{\boldsymbol{Y}_t - \boldsymbol{g}^T\boldsymbol{\beta}\}\mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right\}, \\
\boldsymbol{c_t} &= \left(\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right)^T \mathrm{diag}\{\boldsymbol{Y}_t - \boldsymbol{g}^T\boldsymbol{\beta}\}R_z^{-1} \\
&\times \mathrm{diag}\{\boldsymbol{Y}_t - \boldsymbol{g}^T\boldsymbol{\beta}\}\mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}^*}{2}\right)\right\}
\end{aligned} \quad (3.49)$$

By use of the importance density formulated in equations (3.48) and (3.49), we can do maximum likelihood parameter estimation using very similar procedures as those for the single-realization SHP model. Since we have replicates for the Z process, it is expected to improve the estimation and prediction performances by taking advantage of more information. We will also revise some implementation details accordingly.

Conditional on α, $Y_t(\boldsymbol{x})$'s are independently and identically distributed as $\mathrm{N}(\boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta}, \sigma^2\exp(\tau\alpha)\rho_z)$. We can calculate the sample standard deviation $S_y(\boldsymbol{x})$ at each location $\boldsymbol{x}$. Since $S_y(\boldsymbol{x})$ converges to $\sigma\exp(\tau\alpha(\boldsymbol{x})/2)$ as the number of replicates

goes to infinity, it is immediately seen that $Y_t(\boldsymbol{x})/S_y(\boldsymbol{x})$ is approximately distributed as $\mathrm{N}(\boldsymbol{g}(\boldsymbol{x})^T\boldsymbol{\beta}/S_y(\boldsymbol{x}), \rho_z)$. As such, $\{Y_t(\boldsymbol{x}_1)/S_y(\boldsymbol{x}_1), ..., Y_t(\boldsymbol{x}_n)/S_y(\boldsymbol{x}_n)\}_{t=1}^T := \boldsymbol{\mathcal{Z}}$ is approximately T realizations from a Gaussian process with correlation function ρ_z. We can estimate ϕ_z using the approximate likelihood of $\boldsymbol{\mathcal{Z}}$, which is a good initial value for estimating ϕ_z using maximum likelihood.

For estimating the posterior mode $\boldsymbol{\alpha}^*$, we use different initial values from Section 3.3.1. Since we have replicates, there will be T solutions of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ in solving equations (3.33) and (3.34). It is not computationally efficient to try so many initial values. With replicates, we can obtain the sample variance at each observed location, which will provide a direct estimate of $\boldsymbol{\alpha}$. Therefore, we will apply two initial values here. The first initial value is simply a vector of zeros. Conditional on α, the variance of $Y(\boldsymbol{x}, t)$ is $\sigma^2 \exp(\tau\alpha)$. By equating this variance to the sample variance, we calculate the second initial values by

$$\alpha_0(\boldsymbol{x}) = (\log(S_y^2(\boldsymbol{x})) - \log(\sigma^2))/\tau. \tag{3.50}$$

The strategy of choosing kernel centers is also different from the single-realization SHP model. We still fix a certain number of knots that evenly spread within the domain but the remaining knots will be randomly selected by use of probability proportional to $S_y^2(\boldsymbol{x})$ instead of $|Y(\boldsymbol{x})|$.

For prediction, since $Y_1, ..., Y_T$ are independent with each other (conditional on α), the conditional mean for a new observation at $\boldsymbol{x}_0$ and time t, denoted by $Y_{0,t}$, can be written as:

$$\begin{aligned} \mathrm{E}(Y_{0,t}|(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T), \boldsymbol{\alpha}, \alpha_0) &= \mathrm{E}(Y_{0,t}|\boldsymbol{Y}_t, \boldsymbol{\alpha}, \alpha_0) \\ &= \boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta} + \exp(\tau\alpha_0/2)\boldsymbol{r}_z R_z^{-1}\mathrm{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}(\boldsymbol{Y}_t - G(\boldsymbol{x})\boldsymbol{\beta}). \end{aligned} \tag{3.51}$$

That is, conditional on $\boldsymbol{\alpha}$ and α_0, $Y_{0,t}$ only depends on $\boldsymbol{Y}_t$. By revising equation (3.21), the PBP of $Y_{0,t}$ can be obtained by

$$\begin{aligned}\mathrm{E}(Y_{0,t}|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T)) &= \mathrm{E}(\boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta}|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T)) \\ &+\mathrm{E}\big(\mathrm{E}(\exp(\tau\alpha_0/2)|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T),\boldsymbol{\alpha})\boldsymbol{r}_z R_z^{-1}\mathrm{diag}\{\exp(-\tau\boldsymbol{\alpha}/2)\} \\ &\times(\boldsymbol{Y}_t - G(\boldsymbol{x})\boldsymbol{\beta})|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T)\big), \end{aligned} \tag{3.52}$$

where $\mathrm{E}(\exp(\tau\alpha_0/2)|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T),\boldsymbol{\alpha})$ is the same as equation (3.19) except replacing $\boldsymbol{Y}$ by $(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T)$. The last expectation is taken with respect to $p(\boldsymbol{\alpha}|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T))$, which is computed by applying importance sampling and Monte Carlo integration as in Section 3.1.2, by replacing $\mathrm{E}(h(\boldsymbol{\alpha})|\boldsymbol{Y})$ with $\mathrm{E}(h(\boldsymbol{\alpha})|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T))$. For prediction variance $\mathrm{Var}(Y_{0,t}|(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T))$, the formulas are similar as those in equation (3.22) by replacing the $\boldsymbol{Y}$ vector in condition with $(\boldsymbol{Y}_1,...,\boldsymbol{Y}_T)$ and $(\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta})$ by $(\boldsymbol{Y}_t - G(\boldsymbol{x})\boldsymbol{\beta})$.

Chapter 4

SIMULATION STUDIES OF SPATIAL PREDICTION METHODS

We explored some properties of the SHP model in Chapter 2. Specifically, we showed the parameter confounding effects that exist in the unconditional correlation functions and the sample paths, which would possibly bring difficulties in parameter estimation. Chapter 3 proposed maximum likelihood parameter estimation strategy and two process predictors. The purpose of this chapter is to evaluate the estimation and prediction methods proposed in Chapter 3 through simulations. To assess the performances of different spatial prediction methods, we simulate realizations from the SHP model, stationary Gaussian process model and some nonstationary models, and compare the out-of-sample mean square prediction errors by use of different model fits. We also explore the estimation and prediction performances of the low-dimensional approximation model and the SHP model with replicates.

4.1 Simulation Study for SHP Model

In this section, we will generate 1-dimensional and 2-dimensional realizations from the SHP model. By applying different parameters and different sample sizes, we will explore the properties of the maximum likelihood estimates. We will evaluate the prediction performances of the SHP and Gaussian process model fits on the SHP realizations by comparing their mean square prediction error (MSPE).

4.1.1 Prediction and estimation comparisons for 1-dim SHP simulation

Simulation setup

We simulate 1-dimensional realizations from the SHP model with constant mean $\beta = 0$. We take $\sigma^2 = 0.2$ and change the values of τ^2, ϕ_α and ϕ_z to form four parameter combinations. The parameter values we used are summarized in Table 4.1. Both ρ_α and ρ_z are taken to be Gaussian correlation functions. The true sample path is based on 200 input points equally spaced on $[0, 2]$. To achieve a fair comparison, we use the same random number sequences to generate both α and Z processes respectively for realizations using different parameter values. For each parameter combination, we take sample sizes of 30 and 60. The observed 30 or 60 locations are regularly spaced in the domain $[0, 2]$. For each realization and each sample size, we fit SHP and Gaussian process models using 30 or 60 observations, estimate model parameters, predict Y at the remaining 170 or 140 unobserved locations and calculate the out-of-sample MSPE. For the SHP model, we implement maximum likelihood estimation by use of the techniques presented in Sections 3.1 and 3.3. We apply PBLUP and PBP for predictions. For the Gaussian process model, we employ maximum likelihood estimation and take empirical BLUP using Gaussian, exponential, spherical and Matérn correlation functions. Since both ρ_α and ρ_z are Gaussian correlations, it turns out that Gaussian correlation function leads to the best prediction result, i.e., smallest MSPE. Therefore, we only present the results for the Gaussian process model using a Gaussian correlation function.

In order to evaluate our parameter estimation and process prediction schemes, for some parameter combinations, we try three more SHP predictors in addition to PBLUP and PBP. First, we take BLUP predictor using true parameter values in equation (3.25). Second, given true parameter values, equation (3.21) can be evaluated by plugging in the true parameters, applying importance sampling and

Monte Carlo integration. We call such a predictor as MBP (Monte Carlo best predictor). Third, given true parameters and the true $\boldsymbol{\alpha}$ vector at observed locations, the best predictor (BP) can be easily obtained. Referring to equations (3.17) and (3.19), we immediately see that

$$\begin{aligned} \mathrm{E}(Y_0|\boldsymbol{Y}) &= \mathrm{E}[\mathrm{E}(Y_0|\boldsymbol{Y},\alpha_0)] \\ &= \boldsymbol{g}(\boldsymbol{x}_0)^T\boldsymbol{\beta} + \mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{Y})\boldsymbol{r}_z R_z^{-1}\mathrm{diag}\{\exp(-\tau\boldsymbol{\alpha}/2)\} \\ &\quad \times(\boldsymbol{Y} - G(\boldsymbol{x})\boldsymbol{\beta}), \end{aligned} \tag{4.1}$$

where

$$\mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{Y}) = \exp\left(\frac{\tau}{2}\boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{\alpha} + \frac{\tau^2}{8}(1 - \boldsymbol{r}_\alpha R_\alpha^{-1}\boldsymbol{r}_\alpha^T)\right). \tag{4.2}$$

Everything is known in equations (4.1) and (4.2). This best predictor should performs the "best" among all predictors we try.

Results

We first analyze the results from estimation and prediction by use of 30 or 60 observations. Table 4.1 lists the mean and standard deviation of parameter estimates for the SHP model based on 100 realizations. We see that the estimates for ϕ_α and ϕ_z, the important range parameters, have large variances for sample size 30. The standard deviations decrease considerably when increasing sample size to 60. But the estimates for the second parameter combination, i.e., $\phi_\alpha = 80$ and $\phi_z = 40$, are badly biased. They are in reverse relationship to the corresponding true values, i.e., estimated low ϕ_α (with mean 56.11 at $n = 30$ and 45.22 at $n = 60$) and high estimated ϕ_z (with mean 90.00 at $n = 30$ and 161.74 at $n = 60$) versus high true ϕ_α (80) and low true ϕ_z (40). Another unusual phenomenon is that with increased sample size, the estimates are even more negatively biased for ϕ_α and positively biased for ϕ_z.

We can explain these situations from the discussion of confounding effects in Section 2.3. We have shown that both the unconditional correlation function and sample path features are not as sensitive with respect to ϕ_α and ϕ_z as those for Gaussian process with respect to its range parameter ϕ, which leads to large variances in estimating ϕ_α and ϕ_z. We actually show some realizations from the SHP model with the first two parameter combinations in Figure 2.10. Comparing panels (a) and (b) in Figure 2.10, the sample paths show strong inhomogeneous features and look very similar. For any single realization, it is hard to identify from which parameter combination it was generated. The simulation result tells us that the likelihood "prefers" to let the Z process account for the primary features of the Y process (by leading to greater ϕ_z than ϕ_α). It makes sense that the variance process α is smoother than the overall trend process Z.

We can also explain the reason using stochastic process interpolation theory. Any 1-dimensional curve or 2-dimensional surface can be deemed as a realization from a Gaussian process with properly chosen correlation function to satisfy the required smoothness/differentiability property. With sample points dense enough (particularly easy to achieve for a 1-dimensional sample path), any 1-dimensional curve or 2-dimensional surface can be perfectly interpolated by a Gaussian process. Therefore, with sample size increasing, the SHP model fitting will be more and more reduced to a Gaussian process model fit. As such, the ϕ_z estimate will increase while ϕ_α estimate will decrease.

The estimates for the location parameter β and scale parameter σ^2 look like almost the same for sample sizes 30 and 60 if referring to the mean and standard deviation summaries. The additional sample size does not help improving the estimation precision because the observations are highly correlated instead of iid. As such, estimates for the β and σ^2 are reasonable. The estimate of τ^2 is also satisfying.

Figures 4.1 and 4.2 are the boxplots of MSPE for sample sizes 30 and 60, respectively. Throughout this thesis, boxplot refers to the box and whisker plot

without outliers. We do not show outliers in the boxplot to avoid graphs in which the main information is screened due to some extreme outliers. Because the realizations come from SHP model, we expect that SHP using PBP yields the best prediction results. Therefore we table the ratios of MSPE for Gaussian process model and SHP PBLUP over MSPE for SHP PBP. The larger the ratio is, the better the performance of SHP PBP over the other two prediction methods. Table 4.2 summarizes these MSPE ratios. From the boxplots and summary of MSPE ratios, we see that for the first two parameter combinations, SHP PBP outperforms Gaussian process and SHP PBLUP considerably, reflected by large MSPE ratios and low boxes. We know that realizations from these two parameter combinations are very "SHP" like, i.e. with obvious heterogenous features that stationary Gaussian processes are unable to capture. It is encouraging to see that, for the second parameter combination, although the estimates for ϕ_α and ϕ_z are totally reversed, the prediction performance is very pleasing. This confirms that the realizations from the SHP model with high ϕ_α and low ϕ_z values can be represented by low ϕ_α and high ϕ_z values. SHP PBLUP yields similar prediction results as the fitted Gaussian process. It is not surprising since PBLUP is simply a linear predictor using the SHP unconditional correlation function. Therefore, in order to catch the spatial inhomogeneities, we should apply SHP PBP for prediction. From now on, without further comment, SHP model prediction will refer to prediction by use of PBP. For the third and fourth parameter combinations, the three predictors are very close. From Chapter 2, we know that realizations from the SHP model can be Gaussian-like by allowing small τ^2 and/or ϕ_α values. By setting $\phi_\alpha = 10$ and $\tau^2 = 1$ in the third and fourth parameter combinations, we do generate Gaussian-like SHP realizations. Therefore, it is not surprising (and even encouraging) that SHP PBP yields similar prediction performance as Gaussian process model, as well as SHP PBLUP.

In general, we get satisfying maximum likelihood estimation results. The estimated parameters reflect the model features from which we generate the realizations. SHP PBP outperforms Gaussian process model and SHP PBLUP for predicting those SHP-like realizations. In order to further investigate the prediction performance of SHP PBP, we would like to compare it with three more predictors introduced above: BLUP, MBP and BP. These three predictors are all based on known true parameters. Figure 4.3 gives the boxplots of MSPE by use of all six prediction methods for the first two parameter combinations and two sample sizes. Not surprisingly, BP gives excellent prediction results, far beyond the other predictors. It is interesting to see that for the first parameter combination ($\phi_\alpha = 40, \phi_z = 80$), the MSPEs for PBP and MBP are very close, indicating that without knowing $\boldsymbol{\alpha}$, the true parameters do not improve the prediction comparing to estimated parameters. For sample size 60, PBP and MBP have comparable performance as BP. For the second parameter combination ($\phi_\alpha = 40, \phi_z = 80$), MBP gets worse results and crashes for sample size 60. We discuss this breakdown later in this subsection. BLUP works close to PBLUP (as well as GP). This makes sense if we recall the severe confounding effects of parameters in the unconditional correlation functions in Section 2.3.1. It is confirmed again that the linear predictor for SHP is incapable of capturing the heterogeneous features and making good predictions. Among these six predictors, we are most interested in comparing PBP and MBP.

Both MBP and BP use true parameters. Without knowing $\boldsymbol{\alpha}$, MBP performs much worse than BP. Therefore, we doubt the prediction performance is highly associated with the $\boldsymbol{\alpha}$ estimates. Table 4.3 compares the MSPE for predicting Y process and MSE for estimating $\boldsymbol{\alpha}$ by summarizing the ratios of PBP result over MBP result. For the first parameter combination, PBP estimates $\boldsymbol{\alpha}$ a little worse than MBP for sample size 30 and comparable with MBP for sample size 60. The predictions for Y process are similar. For the second parameter combination, PBP

estimates $\boldsymbol{\alpha}$ and predicts Y much better than MBP for sample size 30. For sample size 60, the estimate of $\boldsymbol{\alpha}$ for MBP has totally failed, which leads to the extremely poor prediction results for Y. Therefore, we see that the $\boldsymbol{\alpha}$ estimate is essential for the prediction performance.

Table 4.1: Performance of maximum likelihood estimates of parameters in the 1-dimensional SHP model. The means and standard deviations are based on 100 simulated realizations from the model.

n		σ^2	τ^2	ϕ_α	ϕ_z	β
	True	0.2	4	40	80	0
30	Mean	0.22	3.85	41.28	90.25	0.01
	Stdev	0.23	2.41	20.80	22.09	0.34
60	Mean	0.22	4.28	27.84	119.33	0.00
	Stdev	0.23	2.61	11.94	19.23	0.35
	True	0.2	4	80	40	0
30	Mean	0.24	3.72	56.11	90.00	0.01
	Stdev	0.25	2.46	24.94	25.32	0.37
60	Mean	0.24	4.00	45.22	161.74	0.01
	Stdev	0.24	2.62	21.55	21.67	0.35
	True	0.2	4	10	80	0
30	Mean	0.28	3.77	17.67	77.93	0.00
	Stdev	0.33	2.56	10.15	19.35	0.34
60	Mean	0.28	4.64	9.95	80.77	0.02
	Stdev	0.35	2.77	5.34	15.70	0.31
	True	0.2	1	40	80	0
30	Mean	0.21	1.09	65.26	84..66	0.01
	Stdev	0.15	0.72	34.55	20.30	0.17
60	Mean	0.21	1.11	32.34	100.45	0.01
	Stdev	0.15	0.81	14.02	12.51	0.17

4.1.2 Prediction and estimation comparisons for 2-dim SHP simulation

In Section 2.1 Figure 2.2 panels (b) and (c), we plot two SHP realizations, one of which has similar features as a realization from a Gaussian process (when $\tau^2 = 0.4$) and the other of which has a SHP-like heterogeneous pattern (when

Table 4.2: Summary of MSPE ratios for the 1-dimensional SHP simulation based on 100 realizations. Ratios greater than one favor the PBP method.

Parameters	n	Min	1st Quartile	Median	3rd Quartile	Max
$\phi_\alpha = 40, \phi_z = 80, \tau^2 = 4$	30^{gp}	0.28	1.06	1.64	3.05	22.65
	60^{gp}	0.59	0.97	1.18	1.70	17.87
	30^{pblup}	0.30	1.04	1.70	2.96	21.02
	60^{pblup}	0.55	1.02	1.21	1.66	15.26
$\phi_\alpha = 80, \phi_z = 40, \tau^2 = 4$	30^{gp}	0.21	1.25	1.77	3.15	23.30
	60^{gp}	0.34	1.29	1.97	3.33	13.42
	30^{pblup}	0.21	1.21	1.77	3.05	26.30
	60^{pblup}	0.35	1.20	1.97	3.27	58.73
$\phi_\alpha = 10, \phi_z = 80, \tau^2 = 4$	30^{gp}	0.10	0.72	1.14	2.04	7.98
	60^{gp}	0.46	0.96	1.01	1.07	2.75
	30^{pblup}	0.08	0.75	1.11	1.97	6.72
	60^{pblup}	0.75	0.98	1.03	1.09	1.91
$\phi_\alpha = 40, \phi_z = 80, \tau^2 = 1$	30^{gp}	0.28	0.84	0.99	1.44	6.68
	60^{gp}	0.67	0.98	1.03	1.09	6.38
	30^{pblup}	0.35	0.89	1.04	1.41	5.68
	60^{pblup}	0.69	0.98	1.04	1.11	3.66

[a] 30^{gp} or 60^{gp} refer to MSPE ratios of Gaussian process model over SHP model using PBP with sample sizes 30 or 60. 30^{pblup} or 60^{pblup} refer to MSPE ratios of SHP model using PBLUP over SHP model using PBP with sample sizes 30 or 60.

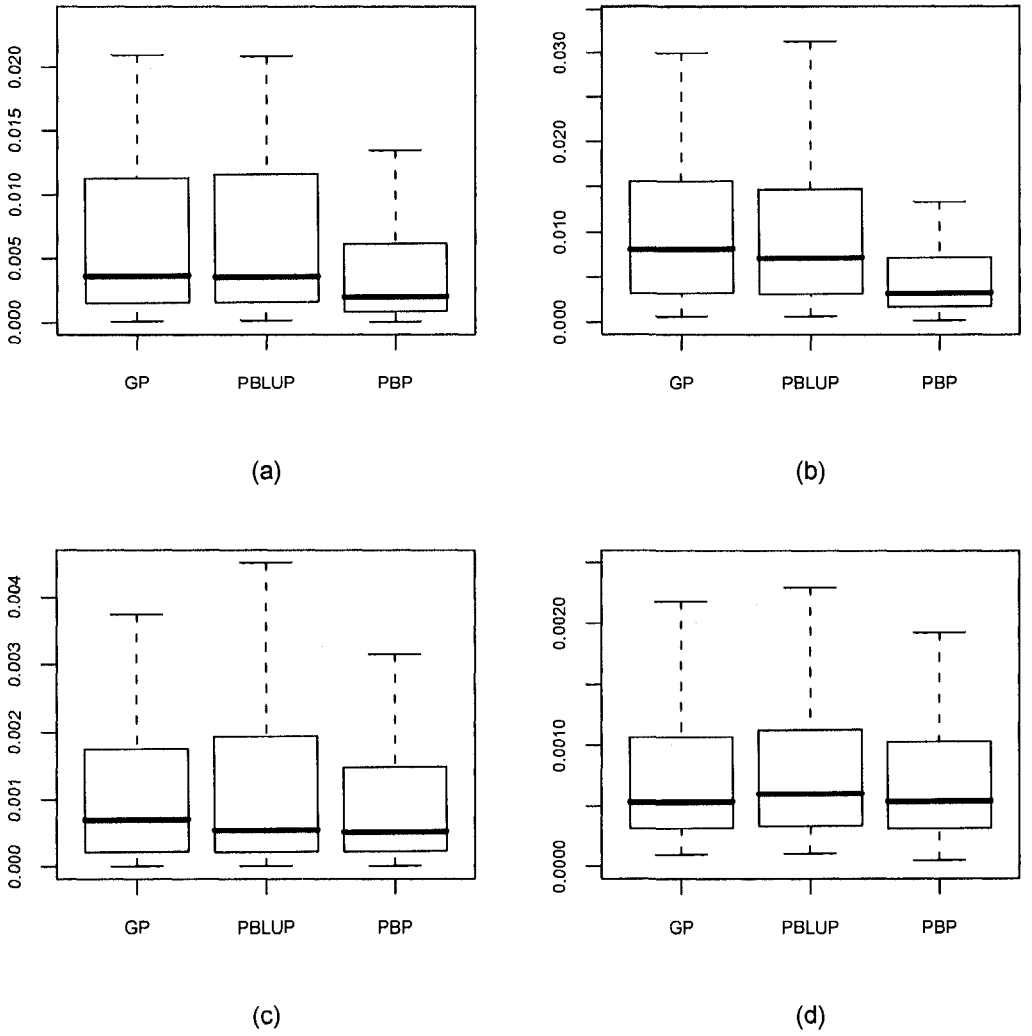


Figure 4.1: The 1-d SHP simulation MSPE boxplots for sample size 30. Panel (a) corresponds to $\phi_\alpha = 40, \phi_z = 80, \tau^2 = 4$. Panel (b) corresponds to $\phi_\alpha = 80, \phi_z = 40, \tau^2 = 4$. Panel (c) corresponds to $\phi_\alpha = 10, \phi_z = 80, \tau^2 = 4$. Panel (d) corresponds to $\phi_\alpha = 40, \phi_z = 80, \tau^2 = 1$.

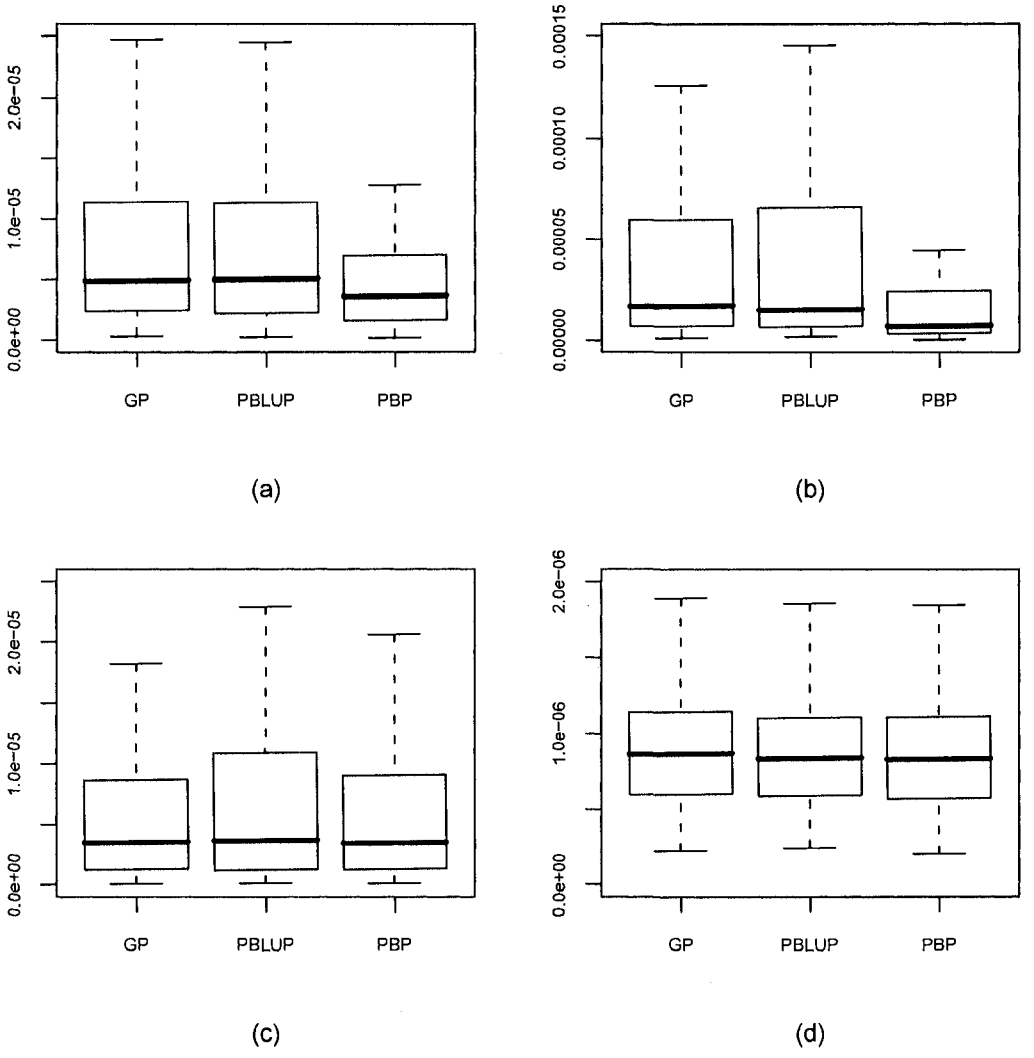


Figure 4.2: The 1-d SHP simulation MSPE boxplots for sample size 60. Panel (a) corresponds to $\phi_\alpha = 40, \phi_z = 80, \tau^2 = 4$. Panel (b) corresponds to $\phi_\alpha = 80, \phi_z = 40, \tau^2 = 4$. Panel (c) corresponds to $\phi_\alpha = 10, \phi_z = 80, \tau^2 = 4$. Panel (d) corresponds to $\phi_\alpha = 40, \phi_z = 80, \tau^2 = 1$.

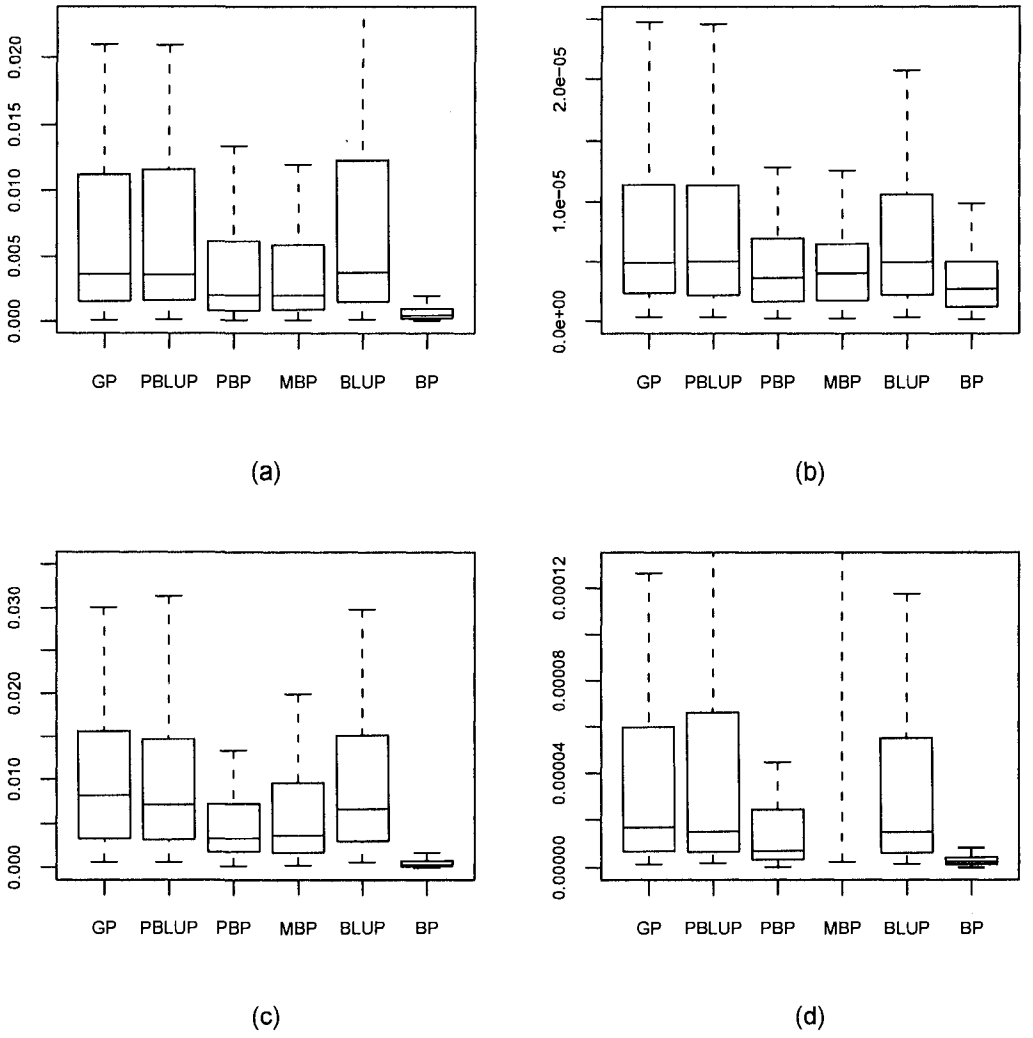


Figure 4.3: The 1-d SHP simulation MSPE boxplots for six predictors. Panel (a) corresponds to $\phi_\alpha = 40, \phi_z = 80, n = 30$. Panel (b) corresponds to $\phi_\alpha = 40, \phi_z = 80, n = 60$. Panel (c) corresponds to $\phi_\alpha = 80, \phi_z = 40, n = 30$. Panel (d) corresponds to $\phi_\alpha = 80, \phi_z = 40, n = 60$.

Table 4.3: Summary of MSPE ratios (PBP/MBP) for the 1-dimensional SHP simulation based on 100 realizations. Ratios greater than one favor the MBP method.

Parameters		Min	1st Quartile	Median	3rd Quartile	Max
$\phi_\alpha = 40, \phi_z = 80$	$n = 30, Y$	0.09	0.78	1.00	1.61	16.58
	$n = 30, \boldsymbol{\alpha}$	0.45	1.27	1.89	3.02	12.64
	$n = 60, Y$	0.24	0.82	0.95	1.09	3.31
	$n = 60, \boldsymbol{\alpha}$	0.11	0.63	1.09	1.90	6.92
$\phi_\alpha = 80, \phi_z = 40$	$n = 30, Y$	0.07	0.54	0.87	1.46	7.96
	$n = 30, \boldsymbol{\alpha}$	0.18	0.54	0.80	1.17	3.64
	$n = 60, Y$	1.28e-10	1.54e-4	3.06e-3	5.05e-2	4.16
	$n = 60, \boldsymbol{\alpha}$	0.02	0.05	0.09	0.14	0.55

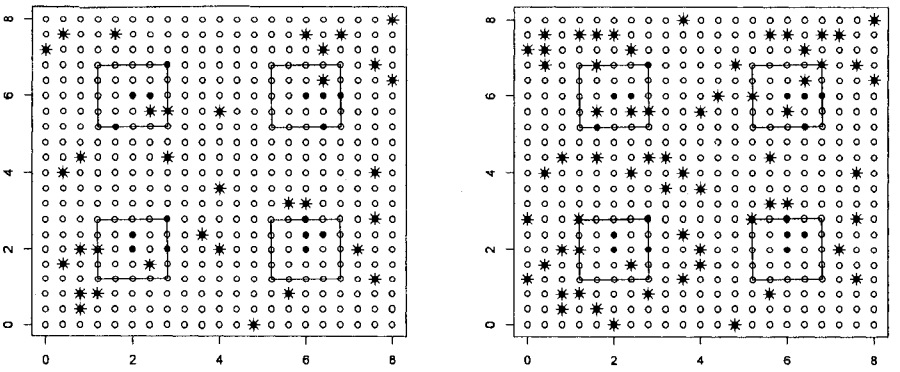
[a] "$n = 30, Y$" and "$n = 60, Y$" refer to the MSPE ratios for prediction of Y process. "$n = 30, \boldsymbol{\alpha}$" and "$n = 60, \boldsymbol{\alpha}$" refer to the MSPE ratios for estimating α process at observed locations.

$\tau^2 = 4$). From the 1-dimensional SHP simulation study, we learn that the MLE has "reverse estimation" behavior for ϕ_α and ϕ_z when the true value of ϕ_α is high and ϕ_z is low. We would like to re-examine this issue by a 2-dimensional SHP simulation study. We have three parameter combinations to try. The true parameter values are listed in Table 4.4. We simulation 100 realizations from SHP model using each of the three parameter combinations. For each realization, the true surface is based on 21×21 grid points on $[0, 8] \times [0, 8]$. To achieve a fair comparison, we use the same random number sequences for both α and Z processes respectively for realizations based on different parameter combinations. For each parameter setting, we try two sample sizes: 50 and 80. The observed locations are obtained by cluster sampling and simple random sampling. As shown in Figure 4.4, we first choose four cluster centers at $(2, 2)$, $(2, 6)$, $(6, 2)$ and $(6, 6)$, then we sample three more locations from the second-order neighborhood around each center. We randomly sample 34 points, together with 16 clustered points, to compose the 50 observed locations, see Figure

4.4 panel (a). For sample size 80, we add 30 more randomly selected points on the base of 50 sampled locations, see Figure 4.4 panel (b). For each realization and each sample size, we fit Gaussian process models by use of Gaussian, exponential, spherical and Matérn correlation functions. Since both ρ_α and ρ_z are Gaussian correlations, it turns out that Gaussian correlation leads to the best prediction result, i.e., smallest MSPE. Therefore, we only present the result of Gaussian process model using Gaussian correlation. For SHP modeling, we implement maximum likelihood estimation, then apply PBLUP and PBP for predictions.

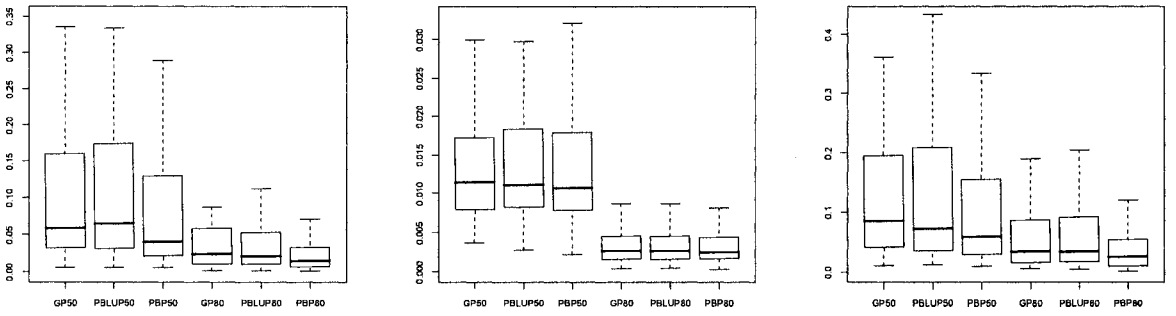
The sample means and standard deviations of model parameter estimates are summarized in Table 4.4. First notice that estimation of σ^2 for the $\tau^2 = 0.4$ model is more accurate than for the $\tau^2 = 4$ models because the realizations from the latter models are more volatile. Estimation of τ^2 for the $\tau^2 = 4$ models are better than for the $\tau^2 = 0.4$ model because the former models provide richer information about the latent process. The estimation in the $\phi_\alpha = 0.3, \phi_z = 0.15$ model is again biased to the point of reversal, as in the 1-dimensional case: ϕ_α estimates have means 0.21/0.18 and ϕ_z estimates have means 0.40/0.46. Increasing sample size effectively decreases the standard deviation.

Table 4.5 summarizes MSPE ratios of Gaussian process fitting and SHP PBLUP over SHP PBP. Figure 4.5 gives the boxplots of all MSPEs. When $\tau^2 = 4$, SHP PBP gives the best prediction results. SHP PBLUP has similar prediction performance as the Gaussian process model. This is another example in which we clearly should apply SHP PBP to catch the spatial heterogeneities and get the best prediction results. When $\tau^2 = 0.4$, SHP PBP has similar performance as SHP PBLUP as well as Gaussian process fitting. So for 2-dimensional Gaussian-like SHP realizations, the out-of-sample prediction performance of Gaussian process modeling is comparable to the SHP model.



(a) (b)

Figure 4.4: Sampling locations of the 2-d SHP simulation. Panel (a) shows 50 sampled locations and panel (b) shows 80 sampled locations. The circles are 21×21 grid points. The 4 squares are clustered neighborhoods. The 16 solid circles are clustered locations. The star points are randomly sampled 34/64 locations.



(a) (b) (c)

Figure 4.5: MSPE boxplots for the 2-dimensional SHP simulation. Panel (a) corresponds to $\tau^2 = 4, \phi_\alpha = 0.15, \phi_z = 0.3$. Panel (b) corresponds to $\tau^2 = 0.4, \phi_\alpha = 0.15, \phi_z = 0.3$ and panel (c) corresponds to $\tau^2 = 4, \phi_\alpha = 0.3, \phi_z = 0.15$.

Table 4.4: Performance of maximum likelihood estimates of parameters in the 2-dimensional SHP model. The means and standard deviations are based on 100 simulated realizations from the model.

n		σ^2	τ^2	ϕ_α	ϕ_z	β
	True	0.2	4	0.15	0.3	0
50	Mean	0.26	3.91	0.16	0.34	-0.03
	Stdev	0.33	2.30	0.08	0.07	0.35
80	Mean	0.26	3.97	0.12	0.37	0.00
	Stdev	0.39	2.15	0.05	0.06	0.31
	True	0.2	0.4	0.15	0.3	0
50	Mean	0.22	0.49	0.25	0.30	-0.02
	Stdev	0.17	0.33	0.12	0.07	0.16
80	Mean	0.21	0.47	0.17	0.31	-0.02
	Stdev	0.13	0.34	0.09	0.04	0.15
	True	0.2	4	0.3	0.15	0
50	Mean	0.21	4.36	0.21	0.40	-0.01
	Stdev	0.25	2.51	0.13	0.10	0.38
80	Mean	0.20	4.37	0.18	0.46	-0.02
	Stdev	0.23	2.56	0.08	0.08	0.35

Table 4.5: Summary of MSPE ratios for the 2-dimensional SHP simulation based on 100 realizations. Ratios greater than one favor SHP model with PBP.

Parameters	n	Min	1st Quartile	Median	3rd Quartile	Max
$\phi_\alpha = 0.15, \phi_z = 0.3, \tau^2 = 4$	50^{gp}	0.18	1.07	1.24	1.80	6.12
	80^{gp}	0.11	1.05	1.36	2.20	8.81
	50^{pblup}	0.12	1.08	1.34	1.95	4.64
	80^{pblup}	0.13	1.07	1.40	2.33	20.78
$\phi_\alpha = 0.15, \phi_z = 0.3, \tau^2 = 0.4$	50^{gp}	0.47	0.87	0.98	1.17	2.34
	80^{gp}	0.52	0.85	1.05	1.28	3.76
	50^{pblup}	0.60	0.93	1.02	1.15	2.16
	80^{pblup}	0.58	0.90	1.02	1.18	3.89
$\phi_\alpha = 0.3, \phi_z = 0.15, \tau^2 = 4$	50^{gp}	0.22	0.92	1.22	1.81	11.47
	80^{gp}	0.39	1.03	1.34	2.32	5.19
	50^{pblup}	0.17	0.95	1.24	1.80	21.09
	80^{pblup}	0.44	1.07	1.54	2.48	7.95

[a] 50^{gp} or 80^{gp} refer to MSPE ratio of Gaussian process model over SHP model with PBP. 50^{pblup} or 80^{pblup} refer to MSPE ratio of SHP model with PBLUP over SHP model with PBP.

4.2 Simulation Study for Stationary Gaussian Process Model

From Chapter 2, we know that the SHP model will reduce to a stationary Gaussian process model by setting $\tau^2 = 0$. Moreover, small τ^2 and/or ϕ_α values will lead to Gaussian-like SHP realizations. In Section 4.1, we show that for the Gaussian-like SHP realizations, SHP and Gaussian process models have similar out-of-sample prediction performance. In this section, we will simulate realizations from stationary Gaussian process model and evaluate the out-of-sample prediction performance of SHP model fitting on the stationary Gaussian process realizations. From now on, we will often abbreviate Gaussian process as GP.

4.2.1 Prediction comparisons for 1-dim Gaussian process simulation

We try four stationary Gaussian process models. Fixing mean 0 and variance 4, we apply Gaussian correlation functions with range parameter ϕ equal to 100 (high) and 10 (low) and exponential correlation functions with ϕ equal to 10 (high) and 1 (low). For each model, we simulate 100 realizations. The true sample paths are based on 200 input points equally spaced on [0,2] and 30 observations are sampled regularly. The random number sequences for generating true sample paths are the same for all four GP models. We fit Gaussian process model and SHP model, predict the process at unobserved locations and compute the out-of-sample MSPE. The summaries of MSPE ratios are given by Table 4.6. Figure 4.6 gives a comparison of MSPE using boxplots. We see that SHP fitting has comparable prediction performance as stationary Gaussian process (the true model) fit for all four models.

4.2.2 Prediction comparisons for 2-dim Gaussian process simulation

For 2-dimensional stationary Gaussian process, we again try four models. Fixing mean 0 and variance 4, we apply Gaussian correlation functions with range parameter ϕ equal to 1 (high) and 0.2 (low) and exponential correlation functions

Table 4.6: Summary of MSPE ratios (SHP/GP) for the 1-dimensional stationary Gaussian process simulation based on 100 realizations.

Model	Min	1st Quartile	Median	3rd Quartile	Max
Gaussian correlation $\phi = 100$	0.59	0.97	1.00	1.03	1.50
Gaussian correlation $\phi = 10$	0.94	1.00	1.00	1.01	1.44
Exponential correlation $\phi = 10$	0.52	0.99	1.00	1.01	1.05
Exponential correlation $\phi = 1$	0.93	1.00	1.00	1.00	1.03

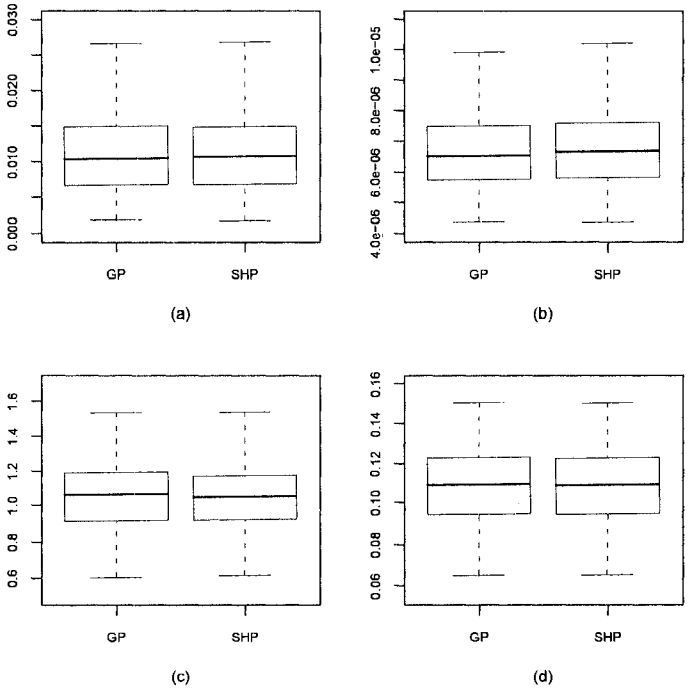


Figure 4.6: MSPE boxplots for the 1-dimensional stationary Gaussian process model simulation. Panel (a) corresponds to Gaussian correlation with $\phi = 100$. Panel (b) corresponds to Gaussian correlation with $\phi = 10$. Panel (c) corresponds to exponential correlation with $\phi = 10$. Panel (d) corresponds to exponential correlation with $\phi = 1$.

with ϕ equal to 0.05 (high) and 0.01 (low). For each model, we simulate 100 realizations. The true sample surfaces are based on 21×21 grid points on $[0, 8] \times [0, 8]$. The 80 observations are sampled according to the sample map shown on panel (b) of Figure 4.4. We fit Gaussian process model and SHP model, predict the process at unobserved locations and compute the out-of-sample MSPE. The summaries of MSPE ratios are given by Table 4.7. Figure 4.7 provides a comparison of MSPE using boxplots. For GP models using exponential correlation functions and Gaussian correlation with $\phi = 1$, SHP fitting has very close prediction performance as stationary Gaussian process (true model) fit. For Gaussian correlation with $\phi = 0.2$, MSPEs for SHP and Gaussian process models are close from the boxplot (Figure 4.7 panel (b)). But from Table 4.7, the ratios have a little wider range. The minimum value 0.00 is due to failure of convergence of GP parameter estimates.

Table 4.7: Summary of MSPE ratios (SHP/GP) for the 2-dimensional stationary Gaussian process simulation of 100 realizations.

	Min	1st Quartile	Median	3rd Quartile	Max
Gaussian correlation $\phi = 1$	0.90	0.99	1.01	1.03	1.21
Gaussian correlation $\phi = 0.2$	0.00	0.85	1.01	1.14	3.68
Exponential correlation $\phi = 0.05$	0.93	0.99	1.00	1.01	1.05
Exponential correlation $\phi = 0.01$	0.39	1.00	1.00	1.01	1.17

4.3 Simulation Study for Nonstationary Spatial Process Models

4.3.1 Deformation model

We have briefly discussed using space deformation to construct heterogeneous spatial covariances in Section 1.1.2. In this section, we will extend the 1-dimensional example from Nychka et al. (2002) to a 2-dimensional nonstationary Gaussian process by use of deformed nonstationary covariance.

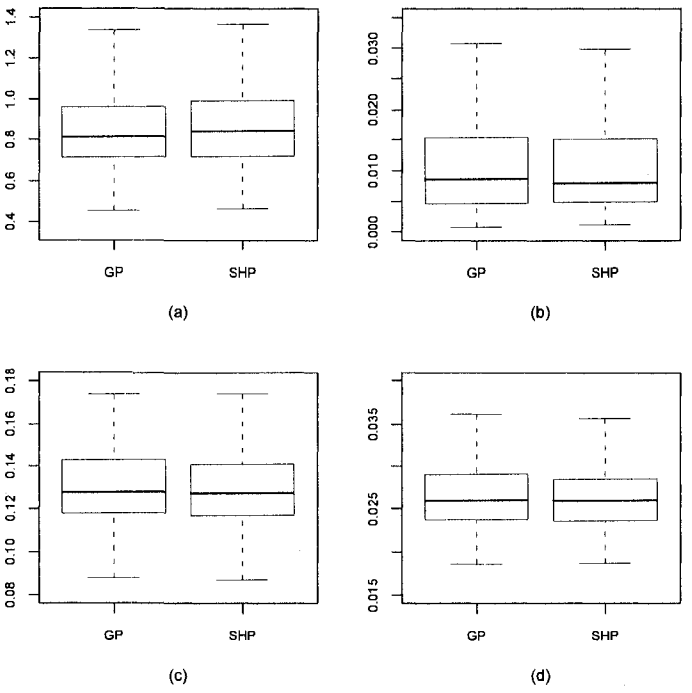


Figure 4.7: MSPE boxplots for the 2-dimensional stationary Gaussian process model simulation. Panel (a) corresponds to Gaussian correlation with $\phi = 1.0$. Panel (b) corresponds to Gaussian correlation with $\phi = 0.2$. Panel (c) corresponds to exponential correlation with $\phi = 0.05$. Panel (d) corresponds to exponential correlation with $\phi = 0.01$.

In Nychka et al. (2002), a one-dimensional nonstationary model is obtained by the following deformation

$$T(x) = px + (1-p)\Phi\left(\frac{x-\mu}{\sigma}\right), \tag{4.3}$$

where x is the 1-dimensional location value and Φ is the standard normal distribution function. This deformation is composed of linear (non-deformed) part and nonlinear (deformed) part with p indicating the fraction of linear part. The parameters μ and σ together with p control the degree of deformation. We restrict x to be between 0 and 1. But the deformation can be applied to arbitrary x values simply by scaling. The nonstationary deformation correlation is obtained by $\rho(|T(x) - T(x')|)$ with ρ a prespecified isotropic correlation function.

In Figure 4.8, we plot several correlation function images to show how this deformation works. Panel (a) shows the stationary correlation by taking the identity transformation since $p = 1$. The correlations are the same for any two locations with the same distances. In panel (b), due to the deformation by setting $p = 0.5$, we see the shorter range correlations in the middle of the interval and longer ranges near the end points. In panel (c), there is no linear part ($p = 0$). We see the correlations are severely distorted and far from stationary. There are very short ranges in the middle and high correlations near the end. Panels (d) through (f) show the positioning effects of μ. Panels (g)-(i) show that smaller σ leads to more deformation.

This deformation strategy can be extended to higher dimension by transforming all dimensions jointly or independently. In our study, we transform x_1 and x_2 in $\boldsymbol{x} = (x_1, x_2)^T$ separately and obtain the nonstationary correlation by

$$\rho(\boldsymbol{x}, \boldsymbol{x}') = \rho\left(\sqrt{(T_1(x_1) - T_1(x_1'))^2 + (T_2(x_2) - T_2(x_2'))^2}\right), \tag{4.4}$$

where T_1 and T_2 are deformations applied to x_1 and x_2 respectively. They can be the same or different.

Nychka et al. (2002) simulate realizations from the 1-dimensional deformation model and estimate the nonstationary covariance by use of multiresolution (wavelet) basis functions. For our simulation study purpose, we will simulate realizations from the 2-dimensional deformation model and then, assuming a correct model specification, estimate parameters p_1, p_2, μ_1, μ_2 and σ_1, σ_2 together with mean β and range parameter ϕ.

4.3.2 Weighted nonstationary model

As we mentioned in Section 1.1.2, a nonstationary model can be obtained by weighting (multiplicative model). Referring to a simulation example in Chang et al. (2007), we propose a nonstationary model by

$$Y(\boldsymbol{x}) = \sigma_1 w_1(\boldsymbol{x}) Y_1(\boldsymbol{x}) + \sigma_2 w_2(\boldsymbol{x}) Y_2(\boldsymbol{x}), \tag{4.5}$$

where $w_1(\boldsymbol{x}) = \sqrt{||\boldsymbol{x} - \boldsymbol{x}_0|| / \max_{\boldsymbol{x} \in D} ||\boldsymbol{x} - \boldsymbol{x}_0||}$ and $w_2(\boldsymbol{x}) = \sqrt{1 - w_1(\boldsymbol{x})}$. The $\boldsymbol{x}_0$ is a prespecified reference point within the domain, and $Y_1(\boldsymbol{x})$ and $Y_2(\boldsymbol{x})$ are stationary Gaussian processes with (isotropic) stationary correlation functions ρ_1 and ρ_2 respectively. The two positive scalers σ_1 and σ_2 can be equal or different. The correlation functions ρ_1 (with ϕ_1) and ρ_2 (with ϕ_2) can be the same or different. In this example, we take ρ_1 and ρ_2 to be the same correlation functions but using different range parameters. It can be seen that the covariance of Y is nonstationary, given by

$$\gamma(\boldsymbol{x}, \boldsymbol{x}') = \sigma_1^2 w_1(\boldsymbol{x}) w_1(\boldsymbol{x}') \rho_1(\boldsymbol{x}, \boldsymbol{x}') + \sigma_2^2 w_2(\boldsymbol{x}) w_2(\boldsymbol{x}') \rho_2(\boldsymbol{x}, \boldsymbol{x}'). \tag{4.6}$$

For our simulation study purpose, we will simulate realizations from the weighted nonstationary model, assuming that the model and the reference point $\boldsymbol{x}_0$ are known, then estimate parameters $\sigma_1^2, \sigma_2^2, \phi_1, \phi_2$ together with the mean β.

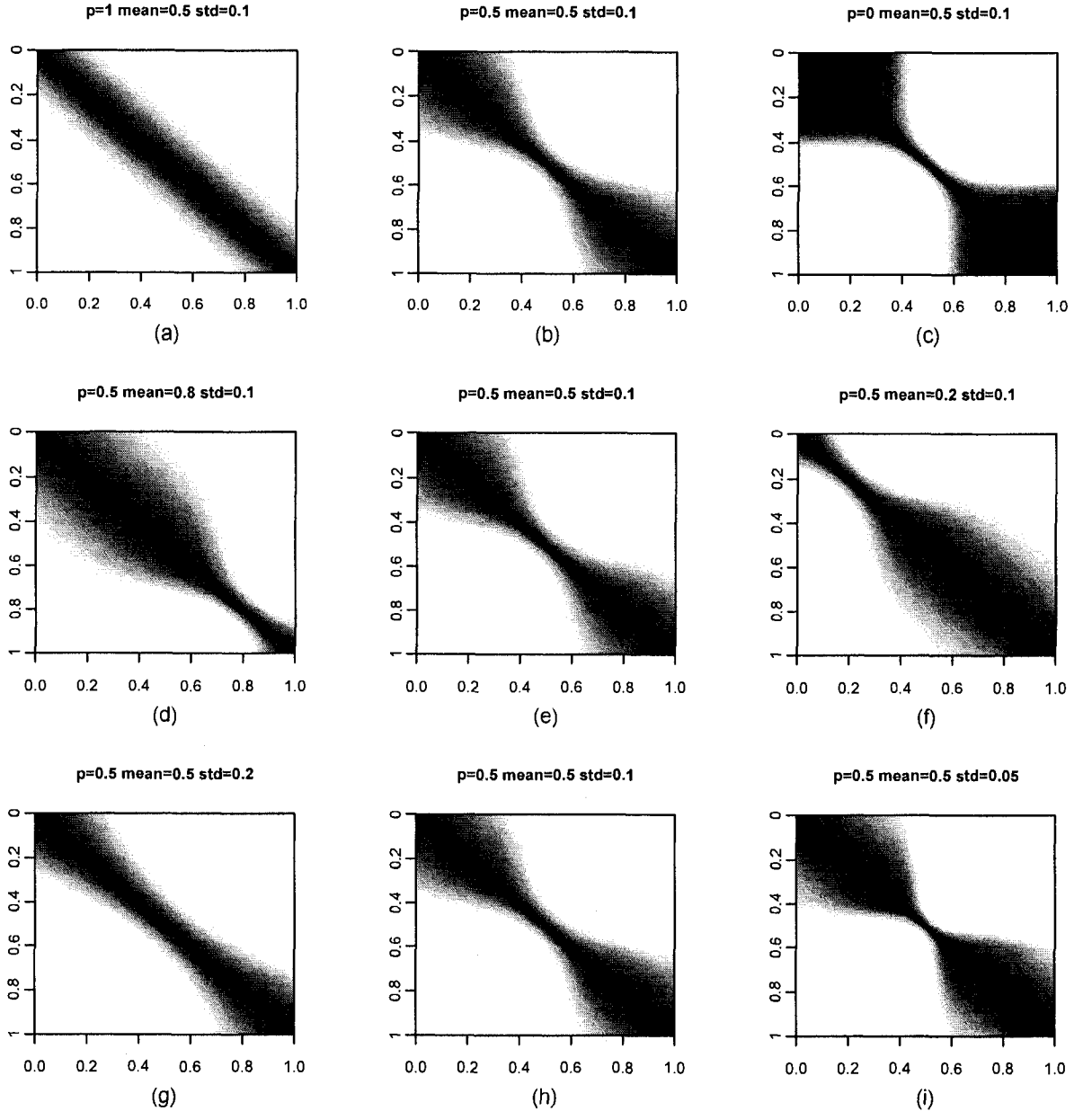


Figure 4.8: The 1-dimensional nonstationary correlations through deformation. The correlation function in the deformation space is Gaussian with range parameter $\phi = 50$.

4.3.3 Prediction comparisons for SHP and nonstationary model simulations

We want to simulate from the deformation model, weighted nonstationary model and SHP. The idea here is to compare model fittings of stationary Gaussian process model, deformation model, weighted nonstationary model and SHP model on the simulated realizations.

We simulate 100 realizations from each model. The true surface is based on 21×21 grid points on $[0, 8] \times [0, 8]$. We sampled 80 points by use of cluster sampling and simple random sampling, the same strategy and sample locations as in Section 4.1.2 Figure 4.4. For all simulations, we adopt zero mean.

For the deformation model, we deform x_1 and x_2 by use of different parameters. We take $p_1 = 0.6, \mu_1 = 0.2, \sigma_1 = 0.2$ for transforming x_1 and $p_2 = 0.8, \mu_2 = 0.7, \sigma_2 = 0.2$ for transforming x_2. We apply a Gaussian covariance function with variance 6 and range parameter 0.2 in the deformed space. For each realization, we fit a stationary Gaussian process model, deformation model, weighted nonstationary model by use of origin as reference location ($\boldsymbol{x}_0 = (0, 0)^T$), weighted nonstationary model by use of center as reference location ($\boldsymbol{x}_0 = (4, 4)^T$) and SHP model. Figure 4.9 panel (a) summarizes the MSPE for different model fits. The two nonstationary models, together with SHP, outperform the stationary Gaussian process model. It is not surprising that the true deformation model performs the best. The SHP model gives smaller MSPE than the weighted nonstationary models.

For the weighted nonstationary model, we take both ρ_1 and ρ_2 as Gaussian correlation functions. We use the origin as reference location ($\boldsymbol{x}_0 = (0, 0)^T$) and $\phi_1 = 0.1, \phi_2 = 0.3, \sigma_1^2 = 4, \sigma_2^2 = 1$. Figure 4.9 panel (b) provides the boxplot of the MSPE for different model fits. Again, the two nonstationary models, together with SHP, outperform the stationary Gaussian process model. The true model, weighted nonstationary model by use of origin as reference location, performs the best. We

see that the SHP model outperforms the deformation model and the weighted nonstationary model by use of center ($\boldsymbol{x}_0 = (4,4)^T$) as reference location. From this example, we see that SHP outperforms the deterministic weighted nonstationary model when the weight is a little misspecified, which illustrates that SHP has the advantage of model flexibility by using latent (Gaussian) stochastic process to model the scale/weight.

For the SHP model, we take both ρ_α and ρ_z as Gaussian correlation functions and $\sigma^2 = 0.2, \tau^2 = 6, \phi_\alpha = 0.1, \phi_z = 0.2, \beta = 0$. Figure 4.9 panel (c) summarizes the MSPE for different model fits. Not surprisingly, the SHP model performs the best. The weighted nonstationary model by use of origin as reference location gives slightly poorer result than the stationary Gaussian process model.

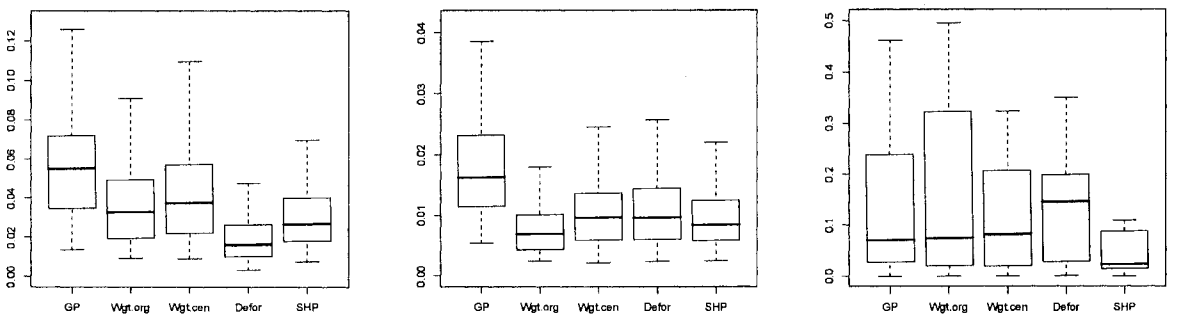
Overall, the two nonstationary models we introduced, together with SHP, are capable of capturing the nonstationary properties even for realizations generated from other nonstationary models, reflected by their better fits than stationary Gaussian process model. SHP can outperform the nonstationary models and comparable to the true model.

4.4 Simulation Study for the Low-Dimensional SHP Approximation Model

In this section, we will revisit some simulation examples presented above and refit models by use of the low-dimensional SHP approximation model proposed in Section 3.4. The purpose is to compare the prediction performances of the regular SHP model and the low-dimensional approximation model.

4.4.1 Prediction and estimation comparisons for 1-dim SHP simulation

In Section 4.1.1, we ran 1-dimensional simulations using four parameter combinations and two different sample sizes. We concluded that the first two parameter combinations, i.e., with $\sigma^2 = 0.2$, $\tau^2 = 4$, $\beta = 0$, setting $\phi_\alpha = 40$, $\phi_z = 80$ or



(a) (b) (c)

Figure 4.9: MSPE boxplots for simulations from nonstationary process models. Panel (a) corresponds to simulations from deformation model by use of $p_1 = 0.6, \mu_1 = 0.2, \sigma_1 = 0.2$ and $p_2 = 0.8, \mu_2 = 0.7, \sigma_2 = 0.2$; Panel (b) corresponds to simulations from weighted covariance model by use of the origin as reference location and $\phi_1 = 0.1, \phi_2 = 0.3, \sigma_1^2 = 4, \sigma_2^2 = 1$; Panel (c) corresponds to simulations from SHP model by use of $\sigma^2 = 0.2, \tau^2 = 6, \phi_\alpha = 0.1, \phi_z = 0.2, \beta = 0$. Wgt.org refers to weighted nonstationary model by use of origin as reference location. Wgt.cen refers to weighted nonstationary model by use of center as reference location. Defor refers to deformation model.

$\phi_\alpha = 80$, $\phi_z = 40$ are representatives of SHP model features. In this section, we use the low-dimensional approximation model to refit realizations generated from SHP model by use of these two parameter combinations.

Table 4.8 summarizes the parameter estimation results. The parameter estimates do not match the true values well, especially for τ^2. The $K\boldsymbol{\omega}$ in equation (3.37) is an approximation of the α process in equation (2.1). They are equivalent in the limiting situation, i.e., when the number of knots goes to infinity and the summation becomes integration. But in the low-dimensional case (we use $J = 10$ knots), they do not have a one-to-one correspondence. We should not expect the parameters estimated by use of the low-dimensional approximation model fit to reflect the true model parameter values. The mean and standard deviation for σ^2 and β almost do not change for sample sizes 30 and 60. This is because the initial values are very close for the two sample sizes (initial value for σ^2 is calculated by

(3.36) and initial value for β is sample mean), and the final optimization results do not differ too much. If we observe the 5-number summaries instead of mean and standard deviation, we do see the differences. The standard deviation for ϕ_α and ϕ_z decrease considerably when sample size increases from 30 to 60. For the second parameter combination, the estimate for ϕ_z (with mean 94.00 at $n = 30$ and 160.80 at $n = 60$) deviates from the true value (40) severely.

To evaluate the low-dimensional approximation model, we are more concerned about the out-of-sample prediction performance than parameter estimation accuracy. We examine prediction performance through the comparison of MSPE as shown in Figure 4.10 and Table 4.9. Figure 4.10 compares the true SHP model fit and low-dimensional approximation model fit by MSPE boxplots. We see that for small sample size ($n = 30$), the true model outperforms the low-dimensional approximation model. For large sample size ($n = 60$), the two model fits are comparable. This indicates that with sample size increasing, the difficulties for true model fit by use of high dimensional importance density increases, while the low-dimensional approximation model is faster and has relatively better performance. For 1-d, 60 evenly located points on [0,2] are dense enough, and so we did not increase sample size further. But we expect that with sample size increasing further, the low-dimensional approximation model will beat the true model fit, which we will see in the next section about 2-dimensional simulations. From Table 4.9, we see that although the boxplots for are similar, the MSPE ratios are widely spread.

4.4.2 Prediction comparisons for 2-dim SHP simulation

From the above 1-dimensional simulation, we see that the parameter estimation based on true SHP model and the low-dimensional approximation model do not match. In this section, we ignore the parameter estimation while concentrating on the out-of-sample prediction performance for 2-dimensional simulations. In

Table 4.8: Summary of parameter estimates for low-dimensional approximation model fits on the 1-dimensional SHP simulations. The means and standard deviations are based on 100 simulated SHP realizations.

		σ^2	τ^2	ϕ_α	ϕ_z	β
	True	0.2	4	40	80	0
30	Mean	0.22	0.13	75.16	100.52	0.00
	Stdev	0.23	0.18	56.57	27.04	0.35
60	Mean	0.22	0.21	50.81	125.51	0.00
	Stdev	0.23	0.18	27.85	20.86	0.35
	True	0.2	4	80	40	0
30	Mean	0.24	0.15	130.32	94.00	0.00
	Stdev	0.25	0.25	137.19	23.11	0.35
60	Mean	0.24	0.15	89.20	160.80	0.01
	Stdev	0.25	0.12	56.50	23.77	0.37

Table 4.9: Summary of MSPE ratios for low-dimensional approximation model study on 1-dimensional SHP simulations. We take the ratio of MSPE using low-dimensional approximation model over SHP model on 100 1-dimensional SHP realizations.

		Min	1st Quartile	Median	3rd Quartile	Max
$\phi_\alpha = 40, \phi_z = 80$	$n = 30$	0.20	0.82	1.13	1.62	22.60
	$n = 60$	0.49	0.91	0.99	1.09	4.56
$\phi_\alpha = 80, \phi_z = 40$	$n = 30$	0.10	0.92	1.23	1.95	16.65
	$n = 60$	0.01	0.82	1.04	1.52	13.67

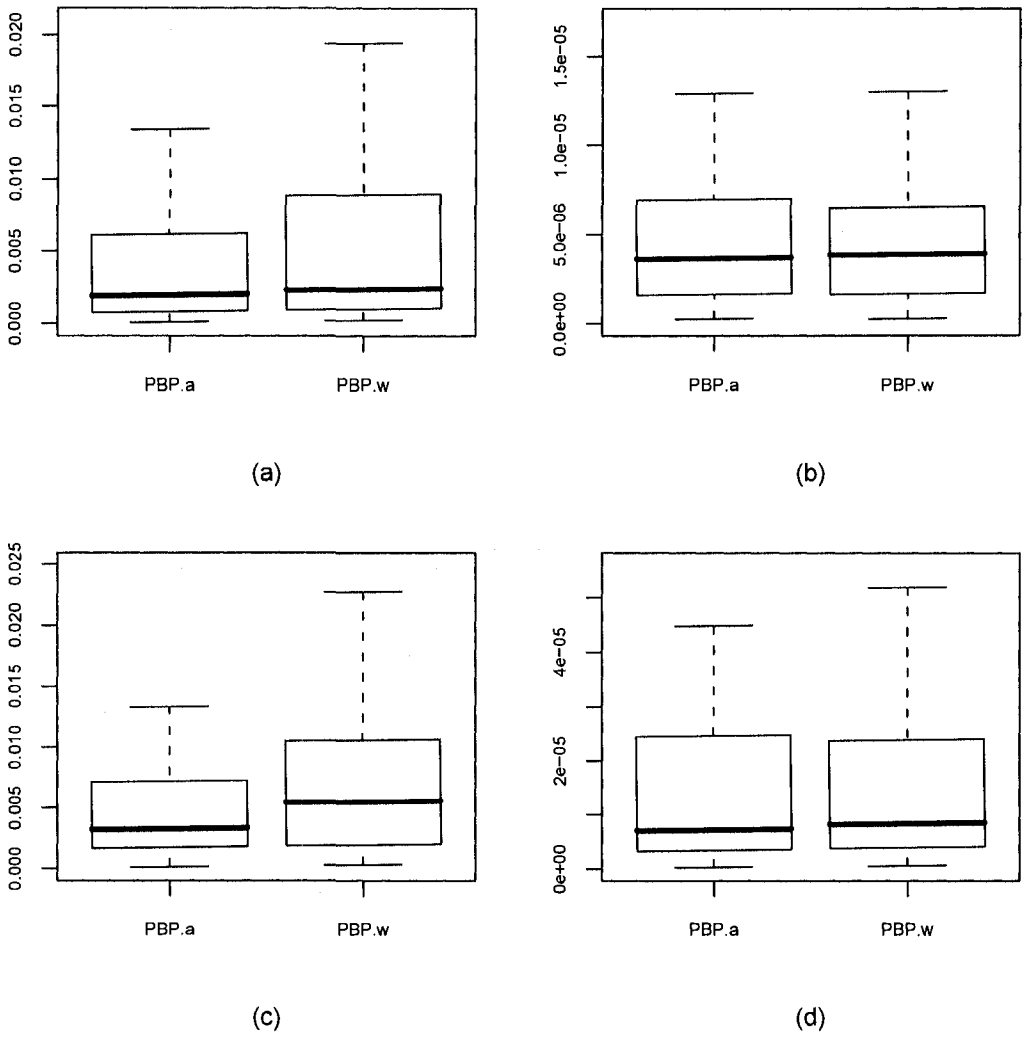


Figure 4.10: MSPE boxplots for comparing low-dimensional approximation model and SHP model on 1-dimensional SHP realizations. Panel (a) corresponds to $\phi_\alpha = 40, \phi_z = 80, n = 30$. Panel (b) corresponds to $\phi_\alpha = 40, \phi_z = 80, n = 60$. Panel (c) corresponds to $\phi_\alpha = 80, \phi_z = 40, n = 30$. Panel (d) corresponds to $\phi_\alpha = 80, \phi_z = 40, n = 60$. PBP.a refers to regular SHP model fit and PBP.w refers to low-dimensional approximation model fit.

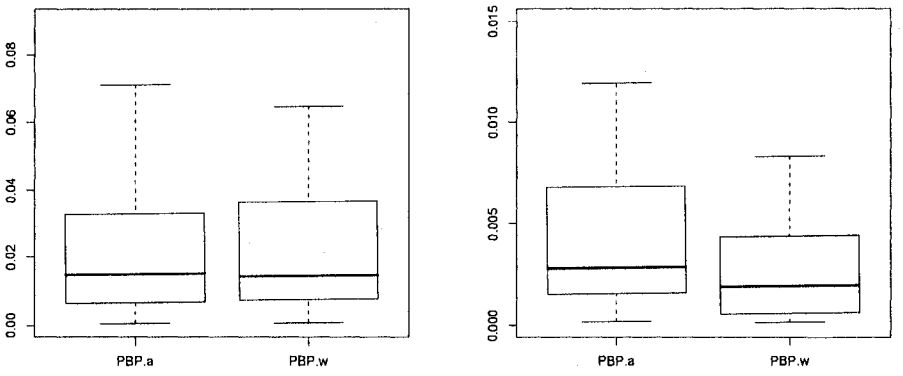
Section 4.1.2, we try three parameter sets to evaluate the estimation and prediction procedures for regular SHP model. The first parameter combination, i.e., $\sigma^2 = 0.2, \tau^2 = 4, \phi_\alpha = 0.15, \phi_z = 0.3, \beta = 0$ is most favorable. The realizations generated from this parameter set are SHP-like and we get very good estimation and prediction results. In this section, we refit realizations from this parameter set by use of low-dimensional approximation model. We take two sample sizes: 80 and 160. We summarize the MSPE ratios in Table 4.10 and compare the boxplots in Figure 4.11. For sample size 80, the low-dimensional approximation model has comparable prediction performance to the regular SHP model. For sample size 160, the approximation model yields considerably smaller MSPE than regular SHP model. This example provides some evidence that the low-dimensional approximation model can solve the numerical difficulties and improve predictions when the sample size is large.

Table 4.10: Summary of MSPE ratios for low-dimensional approximation model study on 2-dimensional SHP simulations. We take the ratio of MSPE using low-dimensional approximation model over SHP model on 100 2-dimensional SHP realizations.

	Min	1st Quartile	Median	3rd Quartile	Max
n=80	0.30	0.87	1.06	1.36	4.10
n=160	0.06	0.46	0.78	1.09	5.29

4.4.3 Prediction comparisons for nonstationary simulations

We have shown that for SHP realizations, the true SHP model fitting outperforms the low-dimensional approximation model when sample size is small, while the low-dimensional approximation model works better as sample size increases. Overall, the low-dimensional approximation model runs faster than regular SHP model. With sample size increasing, the computational advantage of the low-dimensional approximation model becomes more obvious. It is of interest to compare these



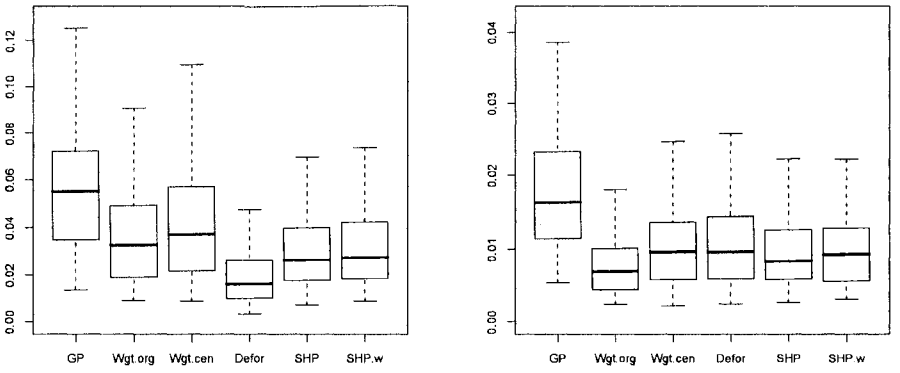
(a) (b)

Figure 4.11: MSPE boxplots for comparing low-dimensional approximation model and SHP model on 2-dimensional SHP realizations. Panel (a) and (b) correspond to sample sizes 80 and 160, respectively. PBP.a refers to regular SHP model fit and PBP.w refers to low-dimensional approximation model fit.

two model fittings on some other nonstationary models. We have introduced deformation model and weighted nonstationary model in Section 4.3. We simulated realizations from these two nonstationary models and compare different model fittings in Section 4.3.3. In this section, we refit those nonstationary realizations by use of the low-dimensional approximation model.

Figure 4.12 shows the boxplots of MSPE. We see that for both nonstationary models, the low-dimensional approximation model fit is slightly worse than SHP model while better than stationary Gaussian process model and other nonstationary models except the true model. Note that for these two examples, the sample size is 80.

In summary, the low-dimensional approximation model has similar performance as the SHP model. When sample size is small, it is recommended to apply SHP model. When sample size is large, SHP model becomes infeasible and it is better to use the low-dimensional approximation model, which is faster and more accurate.



(a) (b)

Figure 4.12: MSPE boxplots for comparing low-dimensional approximation model and other prediction methods on nonstationary simulations. Panel (a) corresponds to simulations from deformation model by use of $p_1 = 0.6, \mu_1 = 0.2, \sigma_1 = 0.2$ and $p_2 = 0.8, \mu_2 = 0.7, \sigma_2 = 0.2$; Panel (b) corresponds to simulations from weighted nonstationary model by use of the origin as reference location and $\phi_1 = 0.1, \phi_2 = 0.3, \sigma_1^2 = 4, \sigma_2^2 = 1$. Wgt.org refers to weighted nonstationary model by use of origin as reference location. Wgt.cen refers to weighted nonstationary model by use of center as reference location. Defor refers to deformation model. SHP.w refers to the low-dimensional approximation model.

4.5 A Simulation Study for SHP Model with Replicates

We introduced the SHP model with replicates and outlined the estimation and prediction procedures in Section 3.5. In this section, we will illustrate the estimation and prediction implementations via a simulation.

We adopt the 1-dimensional simulation framework from Section 4.1.1. We will use the first parameter set in the 1-d simulation example, i.e., $\sigma^2 = 0.2, \tau^2 = 4, \phi_\alpha = 40, \phi_z = 80$ and $\beta = 0$. The realizations from a SHP model with replicates are generated as follows:

- On [0,2], let $x_1, ...x_{200}$ be equally spaced points.
- We generate 100 realizations of $\boldsymbol{\alpha}_i = (\alpha_i(x_1), ..., \alpha_i(x_{200}))^T$ and $\boldsymbol{Z}_i = (Z_i(x_1), ..., Z_i(x_{200}))^T$, from which we form $\boldsymbol{Y}_i = (Y_i(x_1), ..., Y_i(x_{200}))^T$, $i = 1, ..., 100$. This is how we generate 100 $\boldsymbol{Y}$'s from the single-realization SHP model.
- For each $i = 1, ..., 100$, we generate $\boldsymbol{Z}_{i2} = (Z_{i2}(x_1), ..., Z_{i2}(x_{200}))^T, ..., \boldsymbol{Z}_{iT} = (Z_{iT}(x_1), ..., Z_{iT}(x_{200}))^T$. We then use the same $\boldsymbol{\alpha}_i$ to generate $(T-1)$ more $\boldsymbol{Y}$'s for each i, given by: $\boldsymbol{Y}_{it} = \beta + \sigma \exp(\tau \boldsymbol{\alpha}_i/2)\boldsymbol{Z}_{it}$ for $t = 2, ..., T$.

In this example, we try $T = 20$. We select 30 equally-spaced points as observed locations, the same as in the previous 1-d single-realization SHP simulation study. We fit SHP and GP models using a total of $30 \times 20 = 600$ points and predict the remaining $170 \times 20 = 3400$ points. For the SHP model with replicates, we apply the estimation and prediction methods introduced in Section 3.5. For the GP model, the 20 replicates are regarded as iid replicates. The likelihood is simply the product of 20 likelihoods for 20 realizations from the same Gaussian process.

We compare the parameter estimation results from the single-realization SHP model and SHP with replicates in Table 4.11. First we see that the standard deviations for the τ^2 and β estimates decrease considerably and the standard deviation for

the σ^2 estimate decreases mildly for SHP model with 20 replicates. The estimation for ϕ_z improves significantly, while the estimate for ϕ_α still has large variance. Since we have 20 realizations for the Z process, it is reasonable that we can estimate ϕ_z much better than the single-realization SHP model. It may not be surprising that we do not estimate ϕ_α better because we do not have replicates in $\boldsymbol{\alpha}$.

Table 4.12 provides summaries of MSPE ratios. First we observe the ratios of MSPE for GP model over SHP model for realizations each with 20 replicates (GP.20/SHP.20). The ratios are remarkably greater than 1, recognizing that the SHP model (with replicates) can capture the heterogenous features in the sample paths and lead to better prediction performance than the GP model fit.

The first realization of the 20 replicates is from our previous single-realization SHP simulation. We would like to compare the prediction performance on the first realization for the single-realization SHP model fit and replicate SHP model fit. Ratios SHP.1/SHP.20 provide such information. We see that the ratio is considerably greater than 1, indicating the improvement of prediction performance of SHP model with replicates. In Section 4.1.1, we conclude that the $\boldsymbol{\alpha}$ estimate is essential for the prediction performance, by comparing the MSPE of $\boldsymbol{\alpha}$ for PBP and MBP. Here we also summarize the ratios of MSPE for $\boldsymbol{\alpha}$ from two model fits ($\boldsymbol{\alpha}.1/\boldsymbol{\alpha}.20$). It is clearly seen that the replicate model has much better $\boldsymbol{\alpha}$ estimates than the single-realization model.

We are also interested in the ratios of MSPE for GP model fits using single realization over replicates (GP.1/GP.20). We see that these two GP model fits are comparable. Therefore replicates from the SHP model do not help improve the GP (wrong) model fit.

Figure 4.13 compares the MSPE of every first realization of 20 replicates for predictions using SHP PBP, SHP BP and SHP PBP with 20 replicates model. We see the substantially smaller MSPE for SHP PBP with replicates comparing to the single realization SHP PBP.

Overall, by taking advantage of the replicates, we improve the parameter estimation and process prediction performances considerably.

The performance of our algorithms does not hold uniformly in the parameter space. We have also considered the second parameter set from the 1-d single-realization SHP simulation study, i.e., the $\phi_\alpha = 80, \phi_z = 40$ model. Numerical difficulties arise from computing likelihood values using equation (3.3), due to the decreased correlation in the α process and therefore inefficiency of the importance sampling. Improvements to the algorithms or an alternative estimation procedures are a subject for future research.

Table 4.11: Comparison of Parameter estimation for the 1-dimensional SHP simulation. The means and standard deviations are based on 100 simulated realizations from the single-realization model ($T = 1$) or SHP model with 20 replicates ($T = 20$).

T		σ^2	τ^2	ϕ_α	ϕ_z	β
	True	0.2	4	40	80	0
1	Mean	0.22	3.85	41.28	90.25	0.01
	Stdev	0.23	2.41	20.80	22.09	0.34
20	Mean	0.22	4.03	40.08	90.49	-0.01
	Stdev	0.20	0.77	23.71	4.09	0.05

Table 4.12: Summary of MSPE ratios for the 1-dimensional SHP with replicates based on 100 simulated realizations.

	Min	1st Quartile	Median	3rd Quartile	Max
GP.20/SHP.20	1.27	2.57	3.44	5.71	29.06
SHP.1/SHP.20	0.22	1.22	1.79	3.63	35.86
$\boldsymbol{\alpha}.1/\boldsymbol{\alpha}.20$	0.61	1.97	3.43	5.57	29.16
GP.1/GP.20	0.58	0.95	1.00	1.04	16.92

[a]GP.20/SHP.20 refers to ratios of MSPE for GP model over SHP model for realizations each with 20 replicates. MSPE for Y is based on 170 (unobserved) out of 200 (true) observations and 20 replicates. GP.1/GP.20 and SHP.1/SHP.20 refer to the ratios of MSPE for GP/SHP model using single-realization over GP/SHP model using 20 replicates. MSPE for Y is based on 170 (unobserved) out of 200 (true) observations for the first realization out of 20 replicates. $\boldsymbol{\alpha}.1/\boldsymbol{\alpha}.20$ refers to the ratios of MSPE for SHP model using single-realization over SHP model using 20 replicates. MSPE for α is based on 30 observed locations for the first realization out of 20 replicates.

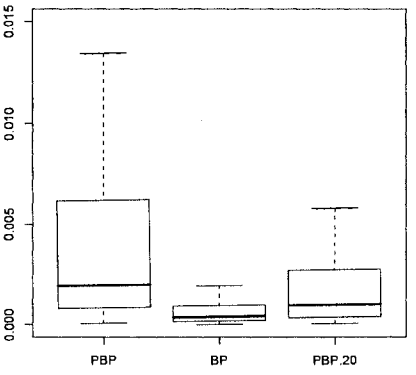


Figure 4.13:
Boxplot for 1-dimensional SHP with replicates. The sample size is 30. MSPE is based on 170 (unobserved) out of 200 (true) observations for the first realization out of 20 replicates. PBP refers to single-realization SHP model fit and PBP.20 refers to SHP model fit with 20 replicates.

Chapter 5

APPLICATIONS

Chapter 4 assessed the parameter estimation and process prediction performances of the SHP model through simulations. In this chapter, we will apply the SHP model to three real data applications. The enhanced vegetation index (EVI) data analysis is a 1-dimensional application. The 2-dimensional example, China precipitation data, shows typical spatial heterogeneities and illustrates an adaptive sampling scheme. The NO_3 deposition data have replicates over 20 years. We would like to see the advantages of SHP model over GP model for fitting these three data sets. The out-of-sample MSPE is the criterion of comparing prediction performances for the SHP model and the stationary Gaussian process model in these examples.

5.1 Enhanced Vegetation Index (EVI) Data Analysis

EVI (the enhanced vegetation index) is the most common index used to assess Earth's vegetation from space. It was developed by Huete et al. (2002) and uses remote sensing data collected by NASA thanks to its satellite *Terra*. *Terra*'s goal is to assess the health of the planet by providing comprehensive information about Earth's land, oceans and atmosphere. EVI describes the relative "greenness" of Earth's vegetation, which is in mathematical terms "a comparison of amounts of visible and near-infrared sunlight that are absorbed and reflected by plants". In other words, it allows recognition of the type of vegetation and crop and the density and size of leaves for every pixel (size of pixels can be 250, 500 or 1000 meters depending on the spatial resolution of a sensor).

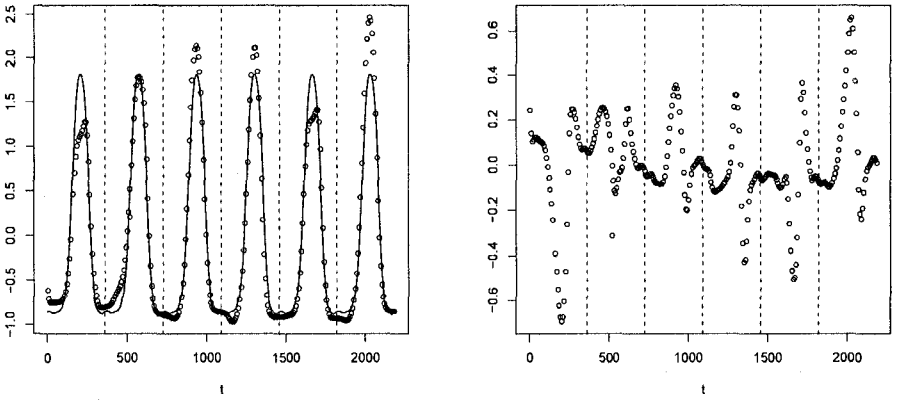
The data we use for analysis come from the Natural Resource Ecology Laboratory (NREL) at Colorado State University. The data set contains EVI values from January 2000 to December 2005 in Iowa (longitude 91.91°N and latitude 42.84°W). The data is recorded every 8 days for 6 full years of observations and EVI values range from 259 to 8034. We standardized the EVI values (subtracted the sample mean and divided by the sample standard deviation) before model fitting. From panel (a) of Figure 5.1, we see the sinusoidal shape of the EVI curve and the yearly trend. Therefore, we decide to regress on Fourier bases with period 365, i.e.,

$$f(t) = \beta_0 + \sum_{r=1}^{k} (\beta_{cr} \cos(2r\pi t/365) + \beta_{sr} \sin(2r\pi t/365)), \tag{5.1}$$

where k is the number of sinusoidal functions used in regression. We fit the regression curves using up to $k = 6$ sinusoidal functions. Because the sine and cosine functions are orthogonal to each other, the regression coefficients are uncorrelated. The coefficient estimates do not change by adding or dropping regression terms. We summarize the coefficient estimates in Table 5.1. We see that the coefficients for sine functions are significant at level 0.05 up to $k = 5$. In Figure 5.1 panel (a), we see that the regression curve fits the data well. In Figure 5.2, we plot the residuals from regression on sinusoidal functions for $k = 1$ up to $k = 6$. It can be seen that the residuals are smooth and highly correlated. The residuals did not change too much after $k = 4$. We finally fit the data using $k = 6$ sinusoidal functions. Figure 5.1 panel (b) shows the enlarged plot of the residuals. Since the residuals are obviously correlated and show inhomogeneous features, we should not treat them as iid normally distributed. We think that the SHP model will be a good fit for the residuals.

We want to compare the performance of SHP modeling and Gaussian process fitting by examining the out-of-sample prediction performance. We sample 36 time points from the residuals where six points are randomly selected from each year. We

use the 36 points to establish the model and predict the residuals at the remaining 216 time points. We repeat this procedure 100 times by independently sampling the 36 time points. For Gaussian process modeling, we tried using Gaussian, exponential and spherical correlation functions. The spherical correlation function leads to the best prediction results, i.e., the smallest MSPE. Therefore, we only list the result of Gaussian process using spherical correlation function. For SHP modeling, we apply Gaussian correlation functions for both α and Z processes. Figure 5.3 provides the MSPE boxplot and Table 5.2 summarizes the MSPE ratios. SHP model by use of PBP prediction outperforms SHP PBLUP and Gaussian process modeling considerably. SHP PBLUP has smaller MSPE than the Gaussian process model. For this example, SHP unconditional correlation function outperforms the other isotropic correlation functions.



(a) (b)

Figure 5.1:
EVI data analysis (I). Panel (a) plots the original data and regression curve using 6 sinusoidal functions. Panel (b) shows the residuals after regression. The dashed lines are separation of different years.

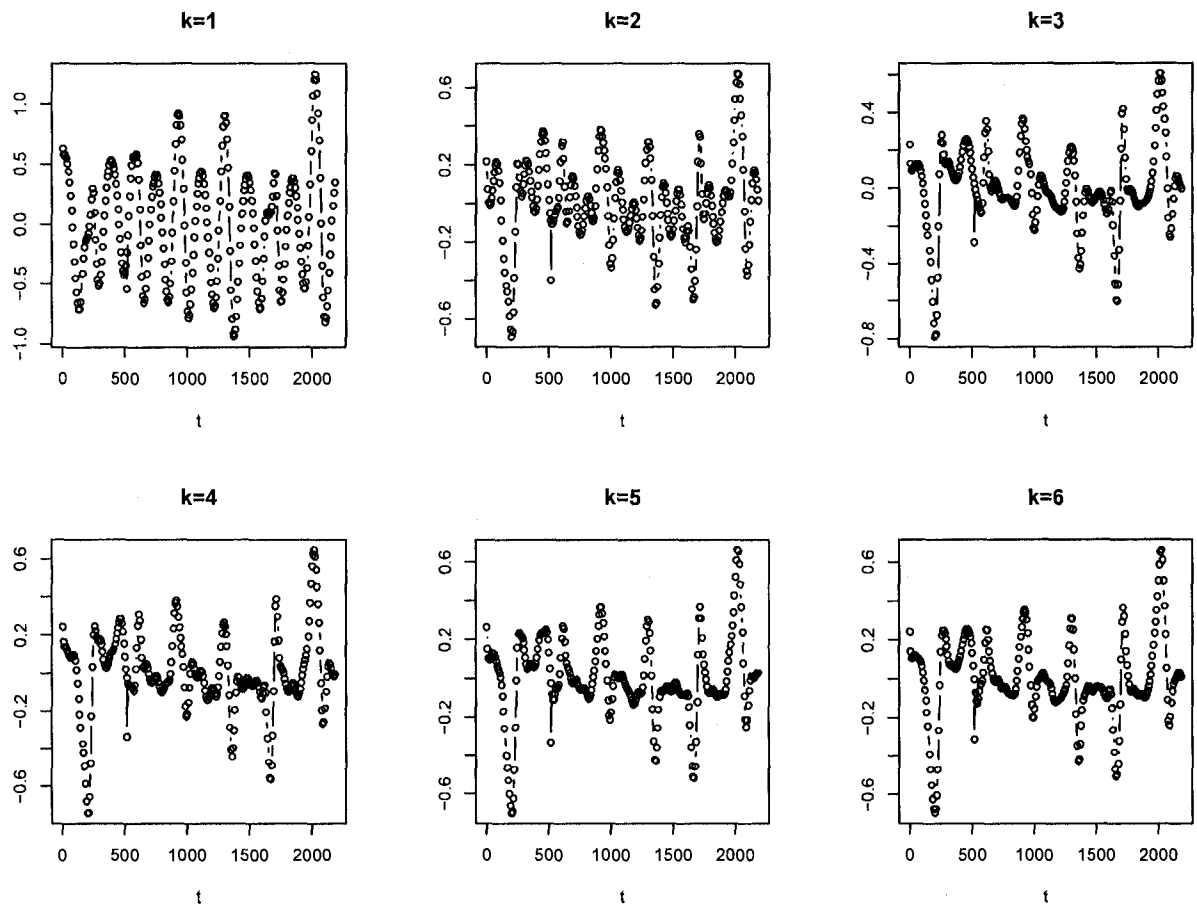


Figure 5.2: EVI data analysis (II): Regression residuals for fits with increasing numbers of sinusoidal basis functions.

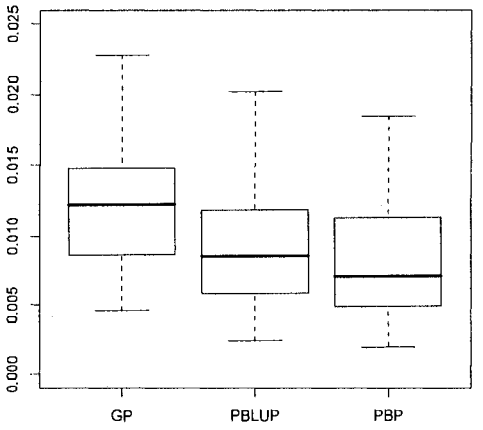


Figure 5.3: EVI data analysis (III): MSPE boxplots for 100 replicates of out-of-sample prediction of 216 days given 36 days selected by stratified simple random sampling, six days per year.

Table 5.1: Summary of regression coefficients for EVI data analysis.

	Estimate	Standard error	t value	$\Pr(>\|t\|)$
β_0	-0.052	0.013	-4.01	$6.53e-05$
β_{c1}	-1.19	0.018	-66.01	$<2e-16$
β_{s1}	-0.47	0.018	-25.80	$<2e-16$
β_{c2}	0.41	0.018	22.52	$<2e-16$
β_{s2}	0.43	0.018	23.57	$<2e-16$
β_{c3}	-0.0058	0.018	-0.33	0.75
β_{s3}	-0.11	0.018	-6.23	$1.82e-09$
β_{c4}	-0.012	0.018	-0.69	0.49
β_{s4}	-0.044	0.018	-2.41	0.017
β_{c5}	-0.024	0.018	-1.33	0.18
β_{s5}	0.039	0.018	2.13	0.034
β_{c6}	0.022	0.018	1.22	0.22
β_{s6}	0.00015	0.018	0.008	1.00

Table 5.2: Summary of MSPE ratios for EVI data residuals fitting. Results are based on 100 replicates of out-of-sample prediction of 216 days given 36 days selected by stratified simple random sampling, six days per year.

	Min	1st Quartile	Median	3rd Quartile	Max
GP/(SHP PBP)	0.03	1.18	1.59	2.06	17.55
(SHP PBLUP)/(SHP PBP)	0.29	0.98	1.05	1.19	2.03
GP/(SHP PBLUP)	0.10	1.14	1.50	1.83	13.19

5.2 China Precipitation Data Analysis

In this section, we will demonstrate the advantages of SHP model over stationary Gaussian process model by an example used in `Spherekit`, the spatial interpolation toolkit. `Spherekit` was developed at the National Center for Geographic Information and Analysis (NCGIA) at the University of California, Santa Barbara. The data set they use for tutoring consists of precipitation in millimeters for 160 weather stations in China. The data was based on the Global Historical Climate Network (GHCN). Both the software and data set can be downloaded from `http://www.ncgia.ucsb.edu/pubs/spherekit/`. The values are long term average (year 1961 – 1990) for January precipitation. Figure 5.4 panel (a) plots locations of the 160 weather stations. Bubbles are used to show different scales of data values. We see that the data is somehow deficient in that high altitude locations in the Himalayan mountains are under-represented. It shows obvious spatial inhomogeneous feature. There are more clustered stations and the precipitation is much more abundant in the south area. For north and west areas, the data are relatively sparse and the precipitation is lower.

In order to compare the performance of SHP modeling and Gaussian process modeling, we sample 30 stations as observed locations and predict the precipitation at the remaining of 130 stations. Due to the unevenly distributed locations of

the data, we want to do stratified sampling to avoid unusual sampled locations. We divide the whole country into three subareas, as shown in Figure 5.4 panel (b). There are 20, 96 and 44 stations in subregions I, II and III respectively. We randomly sampled 5, 15 and 10 stations from subregions I, II and III respectively. We repeat this stratified sampling procedure 300 times by using different random seeds. For each sample, we fit Gaussian process model and SHP model. For each model fitting, we calculate the out-of-sample MSPE. For Gaussian process modeling, we apply Gaussian, exponential, spherical and Matérn correlation functions. It turns out that exponential and spherical correlations lead to very close and the best (i.e., the smallest) MSPE results. For simplicity, we choose to report the results based on the exponential correlation fit only. For SHP modeling, we use exponential correlation functions for both α and Z processes. It is hard to get an explicit form of the kernel that induces the exponential correlation function, for use in the procedure of estimating the posterior mode $\boldsymbol{\alpha}^*$ (discussed in Section 3.3.1). Therefore we simply apply $k_{\phi_\alpha}(\boldsymbol{d}) = \exp(\phi_\alpha||\boldsymbol{d}||)$. In this example and the following NO_3 deposition data analysis, $\boldsymbol{\alpha}^*$ obtained using this kernel for the low-dimensional approximation is effective in our importance sampling procedure.

We only report the PBP prediction results here since it works better than PBLUP. Figure 5.4 panel (b) gives the MSPE boxplots. We see that for sample size 30, SHP outperforms Gaussian process model substantially. Table 5.3 gives the summary of MSPE ratios. The first numerical row corresponds to the MSPE ratio summary for sample size 30 based on 300 samples. It is clear that for most samples, SHP model yields smaller out-of-sample prediction errors than Gaussian process model.

Another advantage of the SHP model is that the prediction variance, calculated by (3.18), provides a means to compute selection probabilities for efficient adaptive sampling. For each sample of 30 observed locations, we randomly sample 20 more

stations with probability proportional to either the SHP prediction variance or the Gaussian process prediction variance (calculated by equation (3.23)). We will refer to the former as SHP adaptive sampling and the latter as GP adaptive sampling. For each sample of size 50 obtained by either SHP adaptive sampling or GP adaptive sampling, we refit the Gaussian process model and SHP model and predict values at the remaining 110 unobserved locations. We actually fit models for 600 samples, 300 of which come from SHP adaptive sampling and another 300 of which come from GP adaptive sampling. From Table 5.3, by observing summaries for GP-GP50/GP-SHP50 and SHP-GP50/SHP-SHP50, we see that for samples from both adaptive sampling schemes, SHP model yields smaller out-of-sample prediction errors than Gaussian process model for most samples, although the ratios are less remarkable than the 30-station case given the large sample size of 50. Note that for GP-GP50/SHP-GP50 and GP-SHP50/SHP-SHP50, we actually take the ratios of MSPE for different samples each extended from the same 30 base sampled locations by use of the same model fits. This ratio indicates that SHP adaptive sampling is more efficient on improving the out-of-sample prediction performance than GP adaptive sampling. Also from Figure 5.4 panel (b), it is clear that by use of probability proportional to the prediction variance for adaptive sampling, the 50-station MSPE is reduced substantially compared with 30-station MSPE. While the GP adaptive sampling does not reduce the MSPE as significantly as SHP adaptive sampling does.

It is not surprising that the SHP model fits better than the Gaussian process because the data show an obvious inhomogeneous pattern from Figure 5.4 panel (a). Since the data value scales differ considerably among subregions I, II and III, we would like to compare the MSPE for each subregion. For the previous analysis, we fitted Gaussian process and SHP interpolation models, i.e., there is no nugget term. The nugget usually refers to measurement error and is assumed to be iid normally distributed with mean 0. Because the data are long-term averages, it is

reasonable to assume that measurement errors have been averaged away. The nugget may also account for microscale variability, i.e., possible model misspecification at very fine scale. So we would like to investigate whether adding a nugget term for the Gaussian process model would help capturing heteroscedasticity and improve prediction performance.

Figure 5.5 gives MSPE boxplots for subregions. The first column of panels refers to 30-location MSPE boxplots. We observe that subregion II dominates the whole MSPE with the largest scale. The Gaussian process with nugget model improves the MSPE over Gaussian process interpolation model and is comparable to the SHP model for subregions I and II. But for subregion III where the low volatilities occur, both Gaussian process models perform much more poorly than the SHP model. By comparing the second and third columns across rows, we see that for subregions I and III, 50 observations obtained from GP adaptive sampling perform slightly better than 50 observations from SHP adaptive sampling. But for subregion II, SHP adaptive sampling outperforms GP adaptive sampling considerably. Since subregion II dominates the overall MSPE, this explains the results on Figure 5.4 panel (b). We know that Gaussian process prediction variances are independent of the observations. The observed locations have 0 prediction variances and locations further away from the existing sampled points have larger prediction variances. Therefore the GP adaptive sampling selects new locations fairly uniformly across the whole region. The SHP prediction variances depend on the observations and can reflect the heteroscedasticity. Due to the presence of high volatilities in subregion II, the SHP adaptive sampling selects new locations intensively from this hot spot. For the second column, the two Gaussian process models perform comparably. Because the 20 additional locations are selected by non-nugget GP adaptive sampling, the 50 observations favor the interpolation Gaussian process model. For the third column, the Gaussian process with nugget model works better than the interpolation Gaussian process model for subregions I and III. In subregion II, they have similar model

performances and comparable to SHP due to large sample size. Because most of the 20 additional locations come intensively from the subregion II by use of SHP adaptive sampling, the relative prediction performance in subregions I and III does not change too much from the 30-observation setup.

Overall, adding a nugget to the Gaussian process model helps improve the prediction performance especially for small sample size but SHP still performs the best. SHP model successfully captures the low volatilities and outperforms Gaussian process models significantly in subregion III. This example illustrates the fact that the SHP prediction variance can represent spatial volatilities and can be efficient in adaptive sampling.

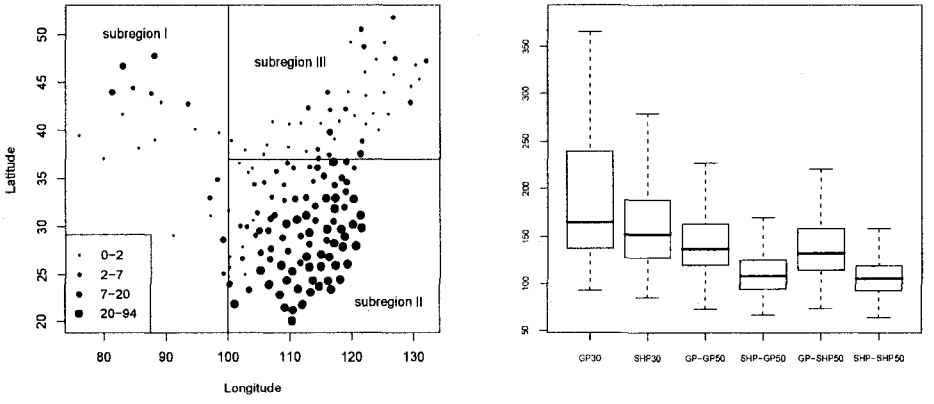
Table 5.3: Summary of MSPE ratios for China precipitation data analysis. Results are based on 300 samples.

	min	1st Quartile	Median	3rd Quartile	Max	Percent
GP30/SHP30	0.62	0.98	1.06	1.17	5.07	79
GP-GP50/GP-SHP50	0.74	0.98	1.02	1.07	5.17	65
SHP-GP50/SHP-SHP50	0.81	0.99	1.02	1.06	6.82	68
GP-GP50/SHP-GP50	0.23	1.05	1.25	1.51	6.90	79
GP-SHP50/SHP-SHP50	0.58	1.08	1.28	1.49	3.29	83

[a] GP30 refers to Gaussian process modeling and SHP30 refers to SHP modeling with 30 observations; GP-GP50 and GP-SHP50 refer to GP and SHP model fits using 50 observations where the extra 20 locations are selected by GP adaptive sampling. SHP-GP50 and SHP-SHP50 refer to GP and SHP model fits based on 50 observations where the extra 20 locations are selected by SHP adaptive sampling. The last column (Percent) indicates the percentage of MSPE ratios being greater than 1 out of 300 samples.

5.3 NO_3 Deposition Data Analysis

The National Atmospheric Deposition Program (NADP) monitors wet atmospheric deposition (chemical constituents deposited from the atmosphere via rain,



(a) (b)

Figure 5.4: China precipitation data analysis. Panel (a) is the precipitation data map. The whole region is manually separated into 3 subregions for stratified sampling. A bubble plot is used to show different value scales. Panel (b) is the MSPE boxplots. GP30 refers to Gaussian process modeling and SHP30 refers to SHP modeling by use of 30 observations; GP-GP50 and GP-SHP50 refer to GP and SHP model fits using 50 observations where the extra 20 locations are selected by GP adaptive sampling. SHP-GP50 and SHP-SHP50 refer to GP and SHP model fits based on 50 observations where the extra 20 locations are selected by SHP adaptive sampling.

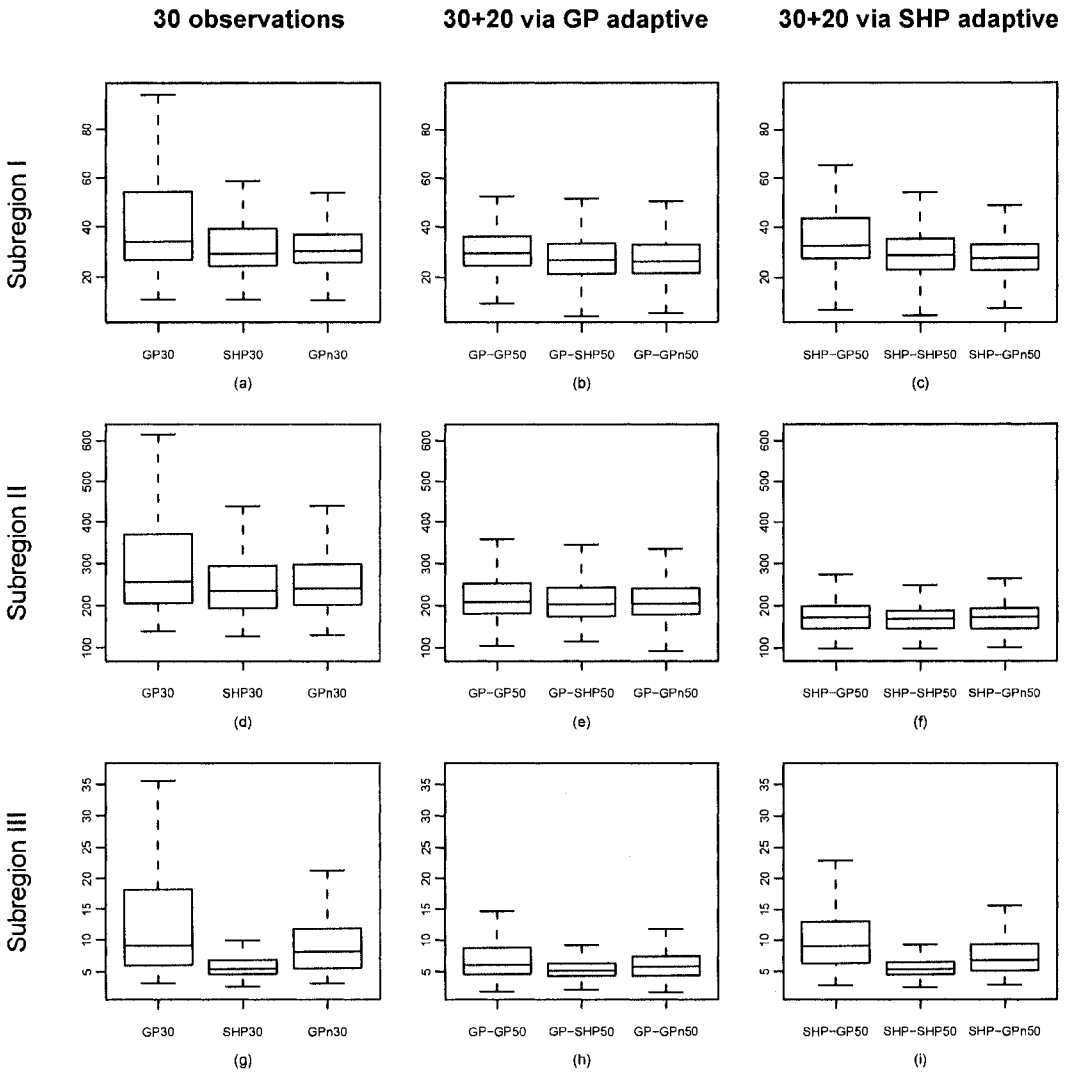


Figure 5.5: China precipitation data analysis - Subregion MSPE boxplots. The rows correspond to subregions I, II and III from top to bottom. From left to right, the columns correspond to 30 observations, 50 observations where the 20 more stations are sampled through GP adaptive sampling, 50 stations where the 20 more stations are sampled through SHP adaptive sampling. For each boxplot, the boxes from left to right correspond to Gaussian process interpolation model, SHP interpolation model and Gaussian process model by allowing a nugget term.

sleet, and snow) in the United States. From `http://nadp.sws.uiuc.edu/`, the NADP data set provides fundamental research support in the areas of air quality, water quality, agricultural effects, forest productivity, materials effects, ecosystem studies, watershed studies, and human health. In this section, we will analyze the annual average of NO_3 concentration (mg/L) from 1986 to 2005. The domain of interest covers 17 states in the west United States with longitude between 100° and 120° west and latitude between 30° and 50° north. There are a total of 79 monitoring sites, 52 of which have complete data for 20 years. The remaining 27 sites have an average of 43% complete data (232 records/(20 $\times$ 27)), with records varying from two years to 18 years. Figure 5.6 plots locations of the monitoring sites. The complete data set can be downloaded from `http://nadp.sws.uiuc.edu/sites/ntnmap.asp`?.

In practice, it will be desired to predict those un-monitored/un-recorded data values. In order to assess the prediction performance, we fit the stationary Gaussian process model and SHP model using only the data from the 52 sites with complete records, then predict the data from the 27 sites with incomplete records. We calculate the out-of-sample MSPE using the 232 observations recorded irregularly across 20 years at the 27 sites. For the SHP model, the observations across years are taken as independent replicates conditional on a common α process. Therefore we will fit a SHP model with replicates, i.e., model (3.45). For the Gaussian process model, the observations over years are regarded as iid replicates. The likelihood for the Gaussian process model will be the product of 20 likelihoods for 20 realizations from the same random process. For prediction using the Gaussian process model, since the 20 replicates are independent, we predict the 27 sites for each year independently. For prediction using the SHP model, we apply equation (3.52).

For Gaussian process models, we try Gaussian correlation functions with/without nugget and exponential correlation functions with/without nugget. It turns out that the exponential correlation function without nugget model yields

the smallest overall MSPE (taken over 232 observations across 27 locations and 20 years), which equals 0.057. We fit the SHP model with replicates by use of exponential correlation functions for both α and Z processes. The overall MSPE for SHP model is 0.050. So the SHP model reduces the overall MSPE by about 12% compared to the best Gaussian process model.

In addition to computing the overall MSPE, we summarize the out-of-sample prediction errors for the 27 locations (computing MSPE across all recorded years within each location) and for the 20 years (computing MSPE across all locations within each year). Table 5.4 provides the summary of MSPE ratios over 27 locations and across 20 years. Figure 5.8 compares relative MSPE ratios over 27 locations and across 20 years. For MSPE across 27 locations, the SHP model outperforms the Gaussian process model at 16 out of 27 sites (60%). From Figure 5.8 panel (a), we see that the Gaussian process model yields slightly smaller MSPE at 11 out of 27 sites. But SHP outperforms the Gaussian process model considerably at a number of sites, e.g., locations 1, 6, 7, 8, 9, 10, 12, 13, 15, 16 and 25. The Gaussian process model only yields obviously smaller MSPE at sites 21 and 23. From Table 5.4 and Figure 5.8 panel (b), we see that SHP outperforms Gaussian process model remarkably for 16 years out of 20 years (80%) and has comparable MSPE as Gaussian process model for the remaining four years.

We want to further investigate the advantage of the SHP model over the Gaussian process model. From Section 5.2, we see that the SHP model outperforms the Gaussian process model by capturing the volatilities successfully. Since we have replicates over 20 years, we can calculate the sample standard deviation at each of the 52 locations with complete data. In Figure 5.7 panel (a), we plot the image and contours of the sample standard deviations. We observe low variances in the northwest corner and high variances around the Colorado area. The image also shows large variances in the southwest corner but they come from the cubic interpolation because we are short of observations there.

We are curious how the prediction variance images look like for the SHP model and Gaussian process model. Because the prediction variance for SHP model calculated by (3.18) is dependent of the observations, the prediction variances for the same location are different across years. But the 20 replicates share the common α process and the prediction variances conditional on α are identical for the same location across different years. So we try to estimation the conditional standard deviation $\sigma \exp(\tau\alpha/2)$. Equation (3.19) gives the formula of $\mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{Y}, \boldsymbol{\alpha})$, the conditional standard deviation at an unobserved location $\boldsymbol{x}_0$. By applying importance sampling and Monte Carlo integration as discussed in Section 3.2, we can calculate $\mathrm{E}(\exp(\tau\alpha_0/2)|\boldsymbol{Y})$. The final estimate of $\sigma \exp(\tau\alpha_0/2)$ is obtained by plugging in the estimated parameters in the Monte Carlo approximation of $\mathrm{E}(\sigma \exp(\tau\alpha_0/2)|\boldsymbol{Y})$. We plot the estimated conditional standard deviation image in Figure 5.7 panel (c). The prediction variances for the Gaussian process model calculated by equation (3.23) are independent of observations and therefore identical across years. Figure 5.7 panel (b) gives the image plot of the GP prediction variance.

For the Gaussian process model, the prediction variances are independent of the observations. The observed locations have 0 prediction variances and locations further away from the observed sites have larger prediction variances. Therefore the standard deviation image shows small contours around each of the 52 sites. The conditional prediction standard deviation image for the SHP model, however, is smooth. We see that the SHP standard deviation image matches the sample standard deviation image in most areas, e.g., greatest variances around the Colorado areas and small variances in the northwest corner. We show the SHP standard deviation contours and 27 prediction locations together in Figure 5.7 panel (d). It is obvious that the locations for which SHP has remarkably smaller MSPE are clustered in the highly volatile areas around Colorado and the northwest corner with low volatility. Comparing Figure 5.6 panel (b) and Figure 5.7 panel (d), we see that

the three locations yielding largest relative MSPE ratios (locations 12, 25 and 13) are clustered in the low volatility area, the northwest corner. Locations 6, 7, 8, 9, 10, which give mild large MSPE ratios, are clustered in the highly volatile Colorado area. From the above analysis, we conclude that the SHP model outperforms the Gaussian process model by capturing the spatial volatilities.

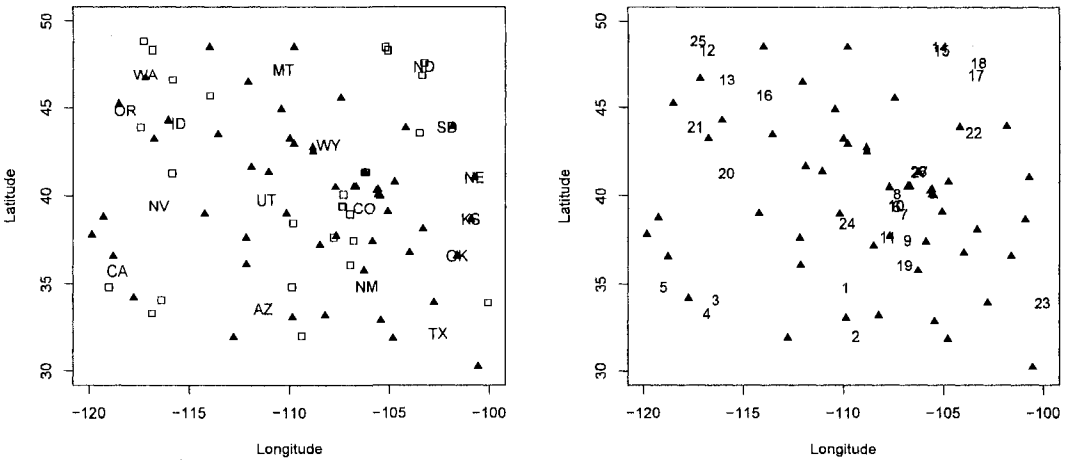


Figure 5.6: Deposition (NO_3) data map. In panel (a), the solid triangles correspond to 52 sites with complete 20 years data and the empty squares correspond to 27 sites with incomplete data. In panel (b), we index the 27 sites by numbers.

Table 5.4: Summary of MSPE ratios (GP/SHP) for the deposition (NO_3) data analysis.

	min	1st Quartile	Median	3rd Quartile	Max	Percent
27 locations	0.44	0.95	1.10	1.44	19.99	60
20 years	0.92	1.06	1.19	1.28	1.70	80

[a]The last column (Percent) indicates the percentage of MSPE ratios being greater than 1.

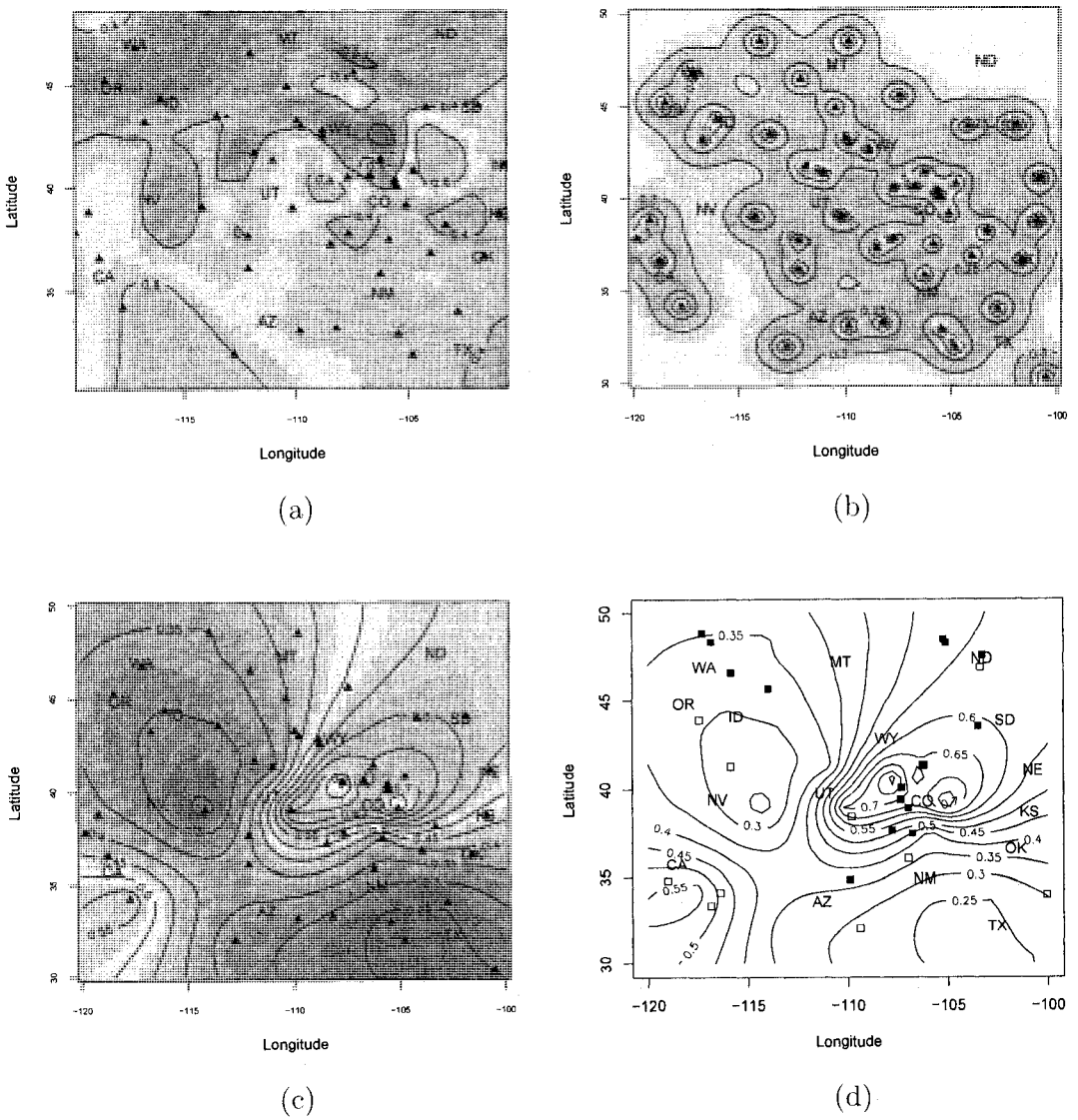
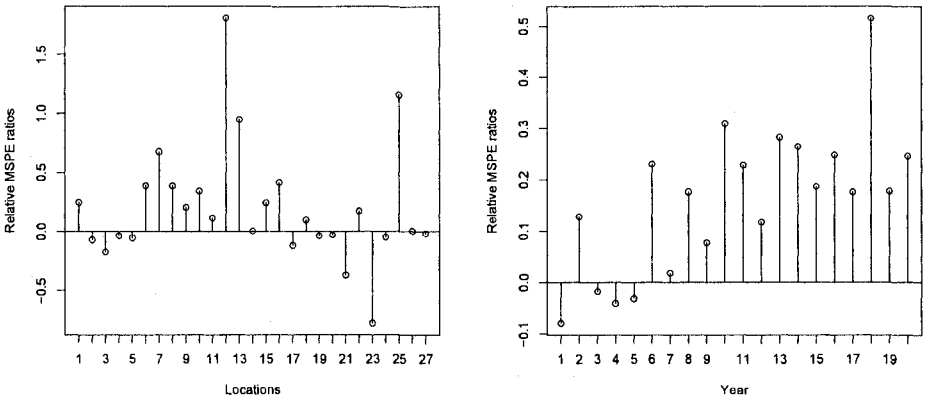


Figure 5.7: NO_3 deposition data standard deviation images. Panel (a) is the image and contour of sample standard deviation using cubic spline interpolation based on 52 sites. Panel (b) is Gaussian process prediction standard deviation. Panel (c) is SHP conditional standard deviation, i.e., the estimates of $\sigma \exp(\tau\alpha/2)$. The triangles are 52 complete data locations. All images are based on 50×50 grid points. Panel (d) plot the 27 predicting locations along with SHP conditional standard deviation contours. Solid squares refer to locations that SHP model yields smaller MSPE than Gaussian process model and empty squares refer to locations that Gaussian process model yields smaller MSPE.



(a) (b)

Figure 5.8: Relative MSPE ratios for NO_3 deposition data analysis. Panel (a) plots the relative MSPE ratios at each of 27 locations. Panel (b) shows the relative MSPE ratios across 20 years. We calculate the relative MSPE ratio by [MSPE(GP)−MSPE(SHP)]/mean[MSPE(GP),MSPE(SHP)]. Ratios greater than 1 favor the SHP model.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1 General Conclusions

Spatial data modeling and analysis aim at the description, explanation and prediction of a spatial process based on a sample of observations. The Gaussian process with a stationary and isotropic correlation function is customarily used to model spatial data. In recent years, a lot of research focuses on modeling the spatial covariance structure more realistically by relaxing the stationary assumptions. This dissertation has developed a new method that is capable of modeling a wide variety of spatial processes and are attractively interpretable.

By analogy to temporal and spatial lattice stochastic volatility models, we propose a stochastic heteroscedastic process (SHP). Conditional on a latent Gaussian process, the SHP is a Gaussian process with non-stationary covariance structure. Unconditionally, the SHP is a stationary non-Gaussian process. The realizations from SHP are versatile and can show obvious heterogeneous features. By adjusting parameter values, SHP can also produce Gaussian-like realizations. The unconditional correlation of SHP offers a rich class of correlation functions which can also allow for a smoothed nugget effect. This smoothed nugget explains the microscale variation in a natural way. The unconditional correlation function can be used independently as a flexible isotropic correlation class. The SHP model has more parameters involved than the stationary Gaussian process model, which brings flexibility but also leads to confounding effects that complicate parameter estimation.

We investigate the difficulties of identifying parameter values through the unconditional correlation and sample path simulation plots.

For statistical inference, we have proposed to apply importance sampling and Monte Carlo integration to evaluate the likelihood values. The importance density we proposed is a multivariate normal distribution with mean equal to the posterior mode. The performance of this importance density heavily depends on the posterior mode estimate. By use of a low-dimensional approximation of the latent process in the optimization procedure, we improve the posterior mode estimation dramatically. This low-dimensional approximation scheme is extended to construct a fully low-dimensional approximation model for SHP, which reduces computational load and increases accuracy considerably when dealing with large sample size. For prediction, we derive the formulas for BLUP (best linear unbiased predictor) and BP (best predictor). By plugging in the maximum likelihood estimators for parameters and applying importance sampling in Monte Carlo integration, we call the final predictors PBLUP (plug-in BLUP) and PBP (plug-in BP).

To evaluate the estimation procedures and compare the prediction performances for SHP and other spatial prediction methods, we conduct various simulation studies. From 1-dimensional and 2-dimensional SHP simulation studies, we see that our importance sampling strategy for parameter estimation and process prediction works well. We also investigate the interesting "reverse estimation" phenomenon for the two range parameters, which can be explained by confounding effects and stochastic process interpolation theory. We conclude that SHP PBP yields the best prediction performance by incorporating the latent process to capture the spatial heteroscedasticity. We show that SHP can fit data generated from Gaussian processes as well as from the true model. Two parameterized nonstationary models — deformation model and weighted nonstationary model are introduced. From simulation, these two nonstationary models, together with SHP, are capable of catching

the inhomogeneous features for realizations simulated from the nonstationary models and SHP model, reflected by their better prediction performances than stationary Gaussian process model. It has been demonstrated that the low-dimensional approximation model we proposed has slightly worse or similar prediction performance as regular SHP model for mild sample size. But with sample size increasing, the low-dimensional approximation model becomes more feasible and accurate.

In Sampson and Guttorp (1992) and many other research work of modeling nonstatonary spatial covariance structure, the spatial data collected over time are deemed as iid replicates, probably obtained after detrending. They work with likelihood of the sample covariance matrix S or smooth S by modern statistical techniques. We have extended the single-realization SHP model to SHP model with replicates to fit the framework of such problem. We assume the replicates come from a SHP model conditioning on a common latent process. From simulation, we see the remarkable improvements in parameter estimation and process prediction by having replicates over single-realization SHP model.

We present applications of SHP model on three real data sets. The advantages of SHP over stationary Gaussian process model are illustrated in several ways. The enhanced vegetation index (EVI) data analysis shows the advantage of SHP model over stationary Gaussian process model for 1-dimensional case. The China precipitation data application revealed an important feature of SHP model, the efficiency of using probability proportional to the SHP prediction variance in adaptive sampling for additional site selections. From the NO_3 deposition data analysis, we see that the SHP model with replicates outperforms the Gaussian process model in prediction by capturing the spatial volatilities.

6.2 Future Topics for Research

Besides improvements of the algorithms for estimation and prediction, there are several aspects in which this work can be further developed to make SHP more

generally applicable to a wider class of spatial processes. We summarize below what we consider the main future directions.

6.2.1 Bayesian approach

The Bayesian approach has been intensively applied in spatial data modeling and analysis. It is easy to rewrite the SHP model (2.1) in a hierarchical structure,

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\psi},\boldsymbol{\alpha} &\sim \mathrm{N}\left(G\boldsymbol{\beta}, \sigma^2 \mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\} R_z \mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}\right), \\ \boldsymbol{\alpha}|\phi_\alpha &\sim \mathrm{N}(\mathbf{0}, \rho_\alpha). \end{aligned} \tag{6.1}$$

By specifying the priors on parameters $\boldsymbol{\psi} = (\sigma^2, \tau^2, \phi_\alpha, \phi_z, \boldsymbol{\beta})$, we complete a hierarchical Bayesian model. The posterior is given by

$$p(\boldsymbol{\psi}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\psi},\boldsymbol{\alpha}) p(\boldsymbol{\alpha}|\phi_\alpha) p(\boldsymbol{\psi}). \tag{6.2}$$

The parameters are usually drawn by use of Metropolis-Hastings algorithms. The difficulties consist in drawing $\boldsymbol{\alpha}$ from its conditional (posterior) distribution. For MCMC approach in time series SV model, the most popular approach to draw the latent vector h_t is a "multi-move" sampler, which is obtained by approximating $\log \epsilon_t^2$ by a mixture of normals so that the $\log y_t^2$ can be written in the form of a Gaussian linear state-space model, and a Gaussian simulation smoother can be applied to draw $\boldsymbol{h}|\boldsymbol{Y}$ simultaneously. For the GLG (Gaussian-log-Gaussian) model proposed by Palacios and Steel (2006), the latent process vector is partitioned into blocks by use of certain clustering algorithm or regular partitioning of the space. For each cluster, they apply a Metropolis-Hastings step and the proposal distribution is constructed by use of log-normal distributions to approximate truncated normal distributions for the conditional posterior. For SHP model, the α process is continuous and the correlation is more severe than the AR(1) process in SV model. Also, $\log(Z^2)$

is correlated instead of iid log chi-square distributed which impedes us to apply "multi-move" sampler analogous to the SV model directly. But we have successfully developed a low-dimensional approximation model in Section 3.4, which reduces the dimensionality of the latent process dramatically. The latent vector $\boldsymbol{\omega} = (\omega_1, ..., \omega_J)$ are iid standard normally distributed. We should be able to draw the latent vector $\boldsymbol{\omega} = (\omega_1, ..., \omega_J)$ simultaneously without being bothered by the high correlations. But from the simulation studies in Section 4.4, the parameter estimates based on the low-dimensional approximation model do not match the true SHP model parameters well. Therefore we will need to further explore how to "tune up" the parameter estimates.

6.2.2 Measurement error

The extension of SHP model by adding a measurement error term is immediate. By allowing the measurement error (nugget) term, SHP becomes more flexible and more adapted to model a wider variety of spatial data. Our importance sampling estimation strategy cannot be extended directly because the importance density we proposed is not readily revised to adapt the nugget model. But it is easy to write the extended SHP model in a hierarchical structure,

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\psi}, \boldsymbol{W} &\sim \mathrm{N}(G\boldsymbol{\beta} + \boldsymbol{W}, \sigma_\epsilon^2 I), \\ \boldsymbol{W}|\boldsymbol{\psi}, \boldsymbol{\alpha} &\sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\} R_z \mathrm{diag}\left\{\exp\left(\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}\right), \\ \boldsymbol{\alpha}|\phi_\alpha &\sim \mathrm{N}(\mathbf{0}, R_\alpha), \end{aligned} \tag{6.3}$$

where we assume the nugget is iid normally distributed with mean 0 and variance σ_ϵ^2. If we are able to implement the full Bayesian approach proposed in Section 6.2.1 well, it will be straightforward to extend the schemes to model (6.3).

Bibliography

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC.

Besag, J. (1974). Spatial interaction and the statistical analysis of the lattice systems (with discussion). *Journal of the Royal Statistical Society: Series B*, 36:192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (disc: p21-59). *Annals of the Institute of Statistical Mathematics*, 43:1–20.

Bishop, C. M. and Qazaz, C. S. (1997). Regression with input-dependent noise: a Bayesian treatment. *Advances in Neural Information Processing Systems*, 9:347–353. MIT Press.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting.* Springer.

Chang, Y.-M., Hsu, N.-J., and Huang, H.-C. (2007). Semiparametric estimation of nonstationary spatial covariance function. Submitted to *Journal of Computational and Graphical Statistics.*

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* New York: Wiley, 2nd edition.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semiparametric non-stationary spatial covariance structures. *Environmetrics*, 12:161–178.

Danielsson, J. and Richard, J.-F. (1993). Accelerated Gaussian importance sampler with applications to dynamic latent variable models. *Journal of Applied Econometrics*, 8:153–173.

Davis, R. A. and Rodriguez-Yam, G. (2005). Estimation for state-space models: an approximate likelihood approach. *Statistica Sinica*, 15:381–406.

de Jong, P. and Shephard, N. (1995). The simulation smoother for time series models. *Biometrika*, 82(2):339–350.

Durbin, J. and Koopman, S. J. (1993). Filtering, smoothing and estimation for time series models when the observations come from exponential family distributions. In *Bulletin of the International Statistical Institute, Book 1.*

Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84:669–684.

Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological and Environmental Statistics*, 2:347–369.

Ecker, M. D. and Heltshe, J. F. (1994). Geostatistical estimates of scallop abundance. In *Case studies in Biometry*, pages 107–124. Wiley, New York. In: Lange, N., Ryan, L., Billard, L., Brillinger, D., Conquest, L. and Greenhouse, J. (Eds.).

Fuentes, M. and Smith, R. L. (2001). A new class of nonstationary spatial models. Technical report, North Carolina State University, Department of Statistics.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Goldberg, P. W., Williams, C. K. I., and Bishop, C. M. (1998). Regression with input-dependent noise: a Gaussian process treatment. In *Advances in Neural Information Processing Systems 10*, Cambridge, MA. MIT Press.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61:247–264.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Enviromental Issues*, pages 37–56. London: Springer-Verlag.

Higdon, D., Swall, J., and Kern, J. (1998). Non-stationary spatial modeling. In *Bayesian Statistics 6*, pages 761–768. Oxford: Oxford University Press. Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (Eds.).

Huete, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., and Ferreira, L. (2002). Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote Sensing of Environment*, 83:195–213.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society: Series B*, 59:319–351.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Lindgren, G. (2004). Lectures on stationary stochastic processes. Lund Uinversity.

Mardia, K. V. and Goodall, C. R. (1993). Spatial-temporal statistical analysis of multivariate environmental monitoring data. In *Multivariate Enviromental Statistics*, pages 347–386. In Patil, G.P. and Rao, C.R. (Eds.).

Matérn, B. (1986). *Spatial Variation.* Spinger-Verlag, New York.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–331.

Palacios, M. B. and Steel, M. F. J. (2006). Non-Gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101:604–618.

Perrin, O. and Meiring, W. (1999). Identifiability for non-stationary spatial structure. *Applied Probability*, 36:1244–1250.

Ripley, B. (1981). *Spatial Statistics.* John Wiley & Sons.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Spinger-Verlag, New York, 2nd edition.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B*, 65:743–758.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields*, pages 1–67. Chapman and Hall, London. In: Cox, D. R., Hinkley, D. V. and Barndork-Nielsen, O. E. (Eds.).

Smith, R. L. (1996). Estimating nonstationary spatial correlations. Technical report, Cambridge University, UK.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D., and Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems 15*, pages 1033–1040. (Eds.) Becker, S., S. Thrun and K. Obermayer, MIT Press (2003).

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Spinger-Verlag.

Taylor, S. (1986). *Modelling Financial Time Series.* Chichester: John Wiley.

Wild, P. and Gilks, W. R. (1993). Algorithm as 287: Adaptive rejection sampling from log-concave density functions. *Applied Statistics*, 42:701–709.

Yan, J. (2007). Spatial stochastic volatility for lattice data. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(1):25–40.