A COMPARISON OF FREQUENTIST AND BAYESIAN APPROACHES FOR LINEAR GAUSSIAN PROCESS MODELS

A Thesis in
Statistics
by
Muhammad Atiyat

© 2008 Muhammad Atiyat

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2008
The thesis of Muhammad Atiyat was reviewed and approved* by the following:

Murali Haran  
Assistant Professor of Statistics  
Thesis Advisor

Donald Richards  
Professor of Statistics  
Associate Head of the Department of Statistics

Runze Li  
Associate Professor of Statistics  
Chair of Graduate Program

*Signatures are on file in the Graduate School.
Abstract

Spatial data (data that are geographically referenced) are commonly encountered in varied fields such as ecology, epidemiology, public health, and geoscience. We consider practical issues with using linear Gaussian process models, which are among the most popular models for analyzing spatial data. We summarize some commonly used frequentist and Bayesian approaches for modeling spatial data via Gaussian processes. In the Bayesian context we review some standard approaches for selecting appropriate priors. We also compare estimation and prediction for Gaussian process models via a simulation study and through an application of our methods to a spatial data set used for studying crop epidemics. We conclude with some practical recommendations based on our study.
# Table of Contents

List of Figures ................................................................. vii
List of Tables ................................................................. ix
Acknowledgments ............................................................... x

## Chapter 1

Introduction ................................................................. 1
1.1 Spatial Data .............................................................. 1
1.2 Types of Spatial Data ..................................................... 2
   1.2.1 Geostatistical Data ............................................... 2
   1.2.2 Lattice Data ....................................................... 3
   1.2.3 Point Pattern Data ................................................ 3
1.3 Spatial Autocorrelation ................................................. 3
1.4 Geostatistics Basics ..................................................... 4
   1.4.1 Types of Stationarity .............................................. 4
   1.4.2 Semivariograms and Covariograms .............................. 5
   1.4.3 Isotropy .................................................................. 6
   1.4.4 Semivariogram Model Fitting ..................................... 10
1.5 Gaussian Processes ....................................................... 10

## Chapter 2

Maximum Likelihood Inference for Gaussian Process Models 13
2.1 Overview ................................................................. 13
2.2 Computing ............................................................... 14
2.3 The Likelihood Surface ................................................ 15
2.4 Prediction Based on ML Inference .................................... 16
5.5 Bayesian Inference ........................................ 54
5.6 Conclusion ................................................. 55

Chapter 6
Conclusion and Future Work .................................. 57
6.1 Conclusion .................................................... 57
6.2 Future Work ................................................... 58

Bibliography ......................................................... 59
List of Figures

1.1 Plots of realizations of two Gaussian processes with linear and exponential covariance functions and their corresponding semivariograms 8
1.2 Exponential semivariogram where \( \psi = 1, \kappa = 2, \) and \( \phi = 0.5. \ldots \) 9
2.1 Plot of the log-likelihood surface. ............................................. 16
2.2 Log-likelihood profiles of \( \kappa \) and \( \phi. \) ..................................... 17
4.1 Plots of realizations from parameter sets 1-3 and their corresponding theoretical semivariograms. ...................................................... 33
4.2 Comparison of standard errors for \( \beta \) where \( \beta = 2, \kappa = 3, \psi = 1, \) and \( \phi = 0.1. \) Solid line: posterior density. Dashed line: bootstrap density. .......................................................... 39
4.3 Comparison of standard errors for \( \kappa \) where \( \beta = 2, \kappa = 3, \psi = 1, \) and \( \phi = 0.1. \) Solid line: posterior density. Dashed line: bootstrap density. .......................................................... 40
4.4 Comparison of standard errors for \( \psi \) where \( \beta = 2, \kappa = 3, \psi = 1, \) and \( \phi = 0.1. \) Solid line: posterior density. Dashed line: bootstrap density. .......................................................... 41
4.5 Comparison of standard errors for \( \phi \) where \( \beta = 2, \kappa = 3, \psi = 1, \) and \( \phi = 0.1. \) Solid line: posterior density. Dashed line: bootstrap density. .......................................................... 42
4.6 Prior plots for \( \kappa \) and \( \phi \) where \( \beta = 2, \kappa = 3, \) and \( \phi = 0.1. \) Solid line: data-based prior. Dashed line: independence Jeffreys prior. .......................................................... 43
4.7 Posterior plots for \( \beta \) where \( \beta = 3, \kappa = 2, \) and \( \phi = 0.2. \) Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior. .......................................................... 44
4.8 Posterior plots for \( \kappa \) and \( \phi \) where \( \beta = 3, \kappa = 2, \) and \( \phi = 0.2. \) Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior. .......................................................... 45
4.9 Posterior plots for $\beta$ where $\beta = 2$, $\kappa = 3$, and $\phi = 0.1$. Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior. .............................................. 46

4.10 Posterior plots for $\kappa$ and $\phi$ where $\beta = 2$, $\kappa = 3$, and $\phi = 0.1$. Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior. .............................................. 47

5.1 North Dakota flowering dates, 2006. Squares: June 8-July 5, triangles: July 6-July 12, gray circles: July 13-July 18, white circles: July 26 -July 28. .............................................. 50

5.2 Empirical directional semivariogram. .............................................. 51

5.3 Empirical directional semivariograms for flowering dates. ................. 52

5.4 Plots of the predicted mean flowering dates for the RUC sites based on ML inference. Squares: June 8-July 5, triangles: July 6-July 12, gray circles: July 13-July 18, white circles: July 26 -July 28. .............................................. 53

5.5 Plots of the probability of an FHB epidemic at the RUC sites on July 13 2006 based on ML inference. Squares: $P(s,t) > 0.3$, triangles: $0.1 < P(s,t) \leq 0.3$, gray circles: $0.05 < P(s,t) \leq 0.1$, white circles: $P(s,t) \leq 0.05$. .............................................. 53

5.6 Plots of the predicted mean flowering dates for the RUC sites based on Bayesian inference. Squares: June 8-July 5, triangles: July 6-July 12, gray circles: July 13-July 18, white circles: July 26 -July 28. .............................................. 54

5.7 Plots of the probability of an FHB epidemic at the RUC sites on July 13 2006 based on Bayesian inference. Squares: $P(s,t) > 0.3$, triangles: $0.1 < P(s,t) \leq 0.3$, gray circles: $0.05 < P(s,t) \leq 0.1$, white circles: $P(s,t) \leq 0.05$. .............................................. 55
List of Tables

2.1 ML estimates for $\beta$, $\kappa$, and $\phi$. ........................................ 15
4.1 Parameter sets used for simulation studies. ................................. 32
4.2 MSE results for $\beta = 3$, $\kappa = 2$, and $\phi = 0.2$. ....................... 36
4.3 IEMSE for $\beta = 3$, $\kappa = 2$, and $\phi = 0.2$. ........................... 36
4.4 MSE results for $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. ................. 36
4.5 IEMSE for $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. .................... 37
4.6 Initial parameter values for MCMC algorithm. ............................... 42
4.7 Confidence intervals based on parametric bootstrap and observed
Fisher information for parameter Set 3 in Table 4.1. .......................... 45
5.1 Percentage of mean flowering dates based on the ML approach that
were matched by those based on the Bayesian approach. ................. 56
5.2 Percentage of the probabilities of an FHB epidemic based on the
ML approach that were matched by those based on the Bayesian
approach. ................................................................................. 56
Acknowledgments

I would like to thank my advisor, Dr. Murali Haran, for guiding me and motivating me throughout. I would also like to thank Dr. James Rosenberger for taking the time out to review my thesis and giving me helpful comments. I am also grateful for Dr. Lindsay’s comments about the likelihood surface. Finally, I would like to thank my colleagues: Chris Groendyke, and K. Sham Bhat. Chris took the time out to review the first three chapters of the thesis and provided some very helpful comments. Sham was helpful in answering some of my questions when I got stuck.
Dedication

To my beloved wife, Zareen, who has been a great support throughout my graduate career.
Chapter 1

Introduction

1.1 Spatial Data

Research in areas such as ecology, geology, epidemiology, and climatology often involves data that are geographically referenced and spatially dependent. For such data, we can no longer make the simplifying independent and identically distributed (iid) assumption. Statistical methods for spatial data analysis have evolved into a self contained discipline over time. These methods are used to help capture the dependent nature of this data in order to make inference such as the estimation of model parameters and the prediction of a process at unobserved locations.

This paper seeks to further explore and compare spatial data analysis methods based on maximum likelihood and Bayesian inferences. There has been much debate about which of the two methods yields better estimates of model parameters and predictions at unobserved locations. Herein, we will shed some light on this issue.

This paper focuses on Gaussian processes in the context of geostatistical data which will be introduced in Section 1.5. After introducing Gaussian processes, we will review maximum likelihood inference in the context of the former. We will discuss some of the issues that are raised when using this type of inference to analyze spatial data. In the following sections, we will introduce Bayesian inference and discuss the issue of prior selection. After that we will show the results of a few simulation studies and compare them in the light of maximum likelihood and
Bayesian approaches. The paper concludes with an application of the methods discussed therein to a dataset from crop epidemiology.

The remainder of this chapter describes basic ideas about spatial models and introduces Gaussian processes. Much of this discussion borrows from Cressie (1993) and Schabenberger and Gotway (2005).

1.2 Types of Spatial Data

Due to the rich origin of spatial data, it has been classified into three categories: geostatistical data, lattice data, and point pattern data. In the description of these categories, we follow standard notation. We begin with defining a spatial process. A spatial process in d dimensions is denoted as

\[ \{Z(s) : s \in D \subseteq \mathbb{R}^d\} \] (1.1)

Here, \( Z(s) \) is the attribute \( Z \) observed at location \( s \) where \( s \) is a d-dimensional vector of coordinates in the spatial domain \( D \) and \( d \geq 1 \). Note that \( Z \) could be either a vector or a scalar.

1.2.1 Geostatistical Data

Geostatistical data often appeared in the context of geology. Hence the prefix “geo”. For geostatistical data, the domain \( D \) is continuous and fixed. This means that the process \( Z \) can be observed at any location within \( D \) and the points in \( D \) are not random. For example, let’s say we are interested in the humidity levels throughout the state of Pennsylvania. In this case, the attribute \( Z \) would be the humidity level in the air at a specific location. The location \( s \) in this case could be a vector of length 2 specifying the longitude and latitude. Therefore, \( D \) would be the set of all longitude and latitude coordinates within the state of Pennsylvania (i.e. \( D \subseteq \mathbb{R}^2 \) and is a continuous fixed set).

When dealing with geostatistical data, we are unable to observe \( Z \) at every location in \( D \) because \( D \) contains an infinite number of locations. Therefore, researchers are often interested in reconstructing the surface of attribute \( Z \) over the entire continuous domain \( D \).
1.2.2 Lattice Data

The domain, $D$, for lattice data is discrete and fixed. This means that the locations in $D$ are countable and not random. For example, let’s say we are interested in the rate of heart disease at the county level in the state of Pennsylvania. In this case, $Z(s_i)$ would be the rate of heart disease at county $i$ where $s_i$ could denote the longitude and latitude of the county’s centroid. The set $D$ would denote the centroid locations of all the counties in the state of Pennsylvania. This set is countable and fixed.

A key feature of lattice data is aggregation. Lattice data is often aggregated over areal regions. In the example above, each disease rate is aggregated over a whole county. Another key feature of lattice data is exhaustive observation; meaning the attribute $Z$ is observed at all possible locations in the set $D$. Prediction is not a problem of interest in this case. Instead, smoothing the observed attribute over an entire region such as a county is of interest. Often, lattice data are modeled via Gaussian random fields.

1.2.3 Point Pattern Data

For point pattern data, the domain $D$ is random, meaning the set $D$ changes from one realization of the spatial process to another. For this type of data, the domain where the process or attribute $Z$ is observed is of interest. For example, we may be interested in the locations at which a certain crop disease occurs in the state of Pennsylvania. If in addition to the location, a random attribute is of interest then the point pattern is called a marked pattern. Otherwise, it is simply known as a point pattern or an unmarked pattern. An example of a marked pattern would be if we are interested in the locations at which a certain crop disease occurs in the state of Pennsylvania and the temperature at each location.

1.3 Spatial Autocorrelation

Whenever statistical inference is done using spatial data, it is important to take the dependency of the data into account. In fact, not doing so would be detrimental (see Schabenberger and Gotway (2005) for a discussion on this). Spatial
autocorrelation is the correlation of the same attribute at two different locations. For example, if there is positive spatial autocorrelation then we expect that $Z(s_i)$ and $Z(s_j)$ to be close in value if $s_i$ and $s_j$ are close in terms of distance.

We will discuss the details of spatial autocorrelation concepts and functions in the context of geostatistical data in the next section.

1.4 Geostatistics Basics

In this section we introduce the concepts of stationarity, isotropy, semivariograms, and covariograms. The organization of this section and its content borrow from Banerjee et al. (2004).

1.4.1 Types of Stationarity

As in time series analysis, methods for spatial data analysis make assumptions about the covariance function of a spatial process. These assumptions are known as stationarity assumptions. In this subsection we discuss the three types of stationarity: strict, weak, and intrinsic stationarity.

A process $Z$ as in (1.1) where $s \in D$ is said to be **strictly stationary** if for any set of locations $\{s_1 \ldots s_n\} \in D$ the distribution of $(Z(s_1) \ldots Z(s_n))'$ is the same as the distribution of $(Z(s_1 + h) \ldots Z(s_n + h))'$ for all $h$ such that $\{s_1 + h \ldots s_n + h\} \in D$ where $h$ is a scalar or vector of $d$ dimensions. Here, we are assuming that the variance of process $Z$ exists for all $s_i \in D$. Strict stationarity is a stringent condition. Most methods for spatial analysis are satisfied with less stringent stationarity conditions such as those relying on the moments of the distribution rather than the distribution itself.

Let $\mu(s) = E(Z)$ denote the mean of a spatial process $Z$. Then $Z$ is said to be **weakly stationary** if $\mu(s) = \mu$ (i.e. the mean of the process is constant) and $\text{Cov}(Z(s), Z(s + h)) = C(h)$ (where $C$ is a covariance function sometimes referred to as the **covariogram**) such that $s$ and $s + h$ are in $D$. So the covariance of a weakly stationary process at two locations depends on the distance $h$ between the two locations not on the actual locations. Weak stationarity is also known as **second-order stationarity**. If $Z$ is weakly stationary then
Cov \((Z(s), Z(s + 0)) = \text{Var} \((Z(s)) = C(0)\). This means that a weakly stationary process has constant variance. Some properties of a weakly stationary process adapted from (Schabenberger and Gotway, 2005) are:

1. \(C(0) \geq 0\)
2. \(C(h) = C(-h)\)
3. \(C(0) \geq |C(h)|\)

The third type of stationarity is known as intrinsic stationarity. The process \(Z\) is intrinsically stationary if \(\mu(s) = \mu\) and

\[
E[Z(s + h) - Z(s)]^2 = \text{Var} \((Z(s + h) - Z(s)) = 2\gamma(h). \tag{1.2}\]

Here, \(2\gamma(h)\) is called the variogram and \(\gamma(h)\) is called the semivariogram. Semivariograms will be discussed in Section 1.4.2.

The relationship between these three types of stationarity is that strict stationarity implies weak stationarity and weak stationarity implies intrinsic stationarity.

In geostatistical applications we are often interested in predicting a spatial process at unobserved locations. To do this, we must take the spatial dependence into account. The primary tool used to estimate spatial dependence is the semivariogram.

### 1.4.2 Semivariograms and Covariograms

The semivariogram is a common tool that is used to capture the second moment structure of a spatial process. The semivariogram is defined as

\[
\gamma(h) = \frac{1}{2} \left( \text{Var} \((Z(s + h) - Z(s))\right) \tag{1.3}
\]

\[
= \frac{1}{2} \left( \text{Var} \((Z(s + h)) + \text{Var} \((Z(s)) - 2\text{Cov}(Z(s + h), Z(s))\right) \tag{1.4}
\]

Thus if a process is weakly stationary then

\[
\gamma(h) = C(0) - \text{Cov}(Z(s + h), Z(s)) = C(0) - C(h) \tag{1.5}
\]
Since the true values of the parameters of a spatial process are unknown in most cases, the true semivariogram is unknown. Therefore, we resort to estimating the semivariogram with

\[ \hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (Z(s + h) - Z(s))^2 \]  

(1.6)

where \(|N(h)|\) is the number of distinct location pairs \((s_i, s_j)\) that are distance \(h\) apart. This estimator is due to Matheron (1962) and is often referred to as the Matheron estimator. A plot of \(\hat{\gamma}(h)\) versus \(h\) is called the empirical semivariogram. The Matheron estimator has a number of desirable properties such as unbiasedness and evenness; however, it is sensitive to outliers.

If a spatial process is ergodic,

\[ \lim_{\|h\| \to \infty} C(h) = 0 \]

(1.7)

i.e. the covariance between two points disappears as the distance between them, \(\|h\|\), increases. Therefore, (1.5) implies \(\lim_{\|h\| \to \infty} \gamma(h) = C(0)\). If \(\lim_{\|h\| \to \infty} \gamma(h)\) exists, then the process is weakly stationary and we can derive the covariogram using the semivariogram as follows:

\[ C(h) = \lim_{\|d\| \to \infty} \gamma(d) - \gamma(h). \]

(1.8)

For an alternative definition of ergodicity, see Section 3.3.1. There are a number of conditions that a valid semivariogram and covariogram need to satisfy. For example, a valid semivariogram needs to be conditionally negative definite. On the other hand, a valid covariogram needs to be positive definite. For more details about these conditions, see Cressie (1993).

In Section 1.4.4 we discuss a related topic to semivariogram estimation known as semivariogram model fitting but first we discuss the notion of isotropy.

### 1.4.3 Isotropy

If the semivariogram \(\gamma(h)\) of a spatial process depends only on the length of \(h\), \(\|h\|\), then we say the process is isotropic. If the semivariogram depends also on the
direction of the process we say the process is **anisotropic**. There exists a number of simple models for isotropic processes; this is one reason why isotropic processes are popular. We list three common models given in terms of semivariogram functions below:

1. Linear
   \[
   \gamma(||h||) = \begin{cases} 
   \psi + \kappa ||h|| & ||h|| > 0, \psi > 0, \kappa > 0 \\
   0 & \text{otherwise}.
   \end{cases}
   \]  
   (1.9)

2. Matérn (see Matérn (1986))
   \[
   \gamma(||h||) = \begin{cases} 
   \psi + \kappa \left( 1 - \frac{(2\sqrt{\nu}||h||)\nu}{\phi^{2\nu-1}1(\nu)} K_{\nu} \left( 2\sqrt{\nu} ||h|| \phi \right) \right) & ||h|| > 0, \nu > 0, \phi > 0 \\
   \psi & \text{otherwise}.
   \end{cases}
   \]  
   (1.10)

   where \( K_{\nu} \) is the modified Bessel function of order \( \nu \) (See Abramowitz and Stegun (1964)).

3. Exponential
   \[
   \gamma(||h||) = \begin{cases} 
   \psi + \kappa \left( 1 - \exp \left( -\frac{||h||}{\phi} \right) \right) & ||h|| > 0, \phi > 0 \\
   0 & \text{otherwise}.
   \end{cases}
   \]  
   (1.11)

The exponential semivariogram function is the Matérn function with \( \nu = \frac{1}{2} \).

Figure 1.1 illustrates plots of realizations of two Gaussian processes with linear and exponential covariance functions and plots of their corresponding semivariograms. The plot on the top left shows a realization of a Gaussian process which has a linear covariance function with parameters \( \psi = 1 \) and \( \kappa = 2 \), and a fitted lowess line; the plot on the top right is its corresponding semivariogram. Similarly, the plot on the bottom left shows a realization of a Gaussian process which has an exponential covariance function with parameters \( \psi = 1 \), \( \kappa = 2 \), and \( \phi = .5 \), and a fitted lowess line; the plot on the bottom right shows its corresponding semivariogram.
There are three characteristics that are associated with variograms and semivariograms. They are the **nugget**, **sill**, and **range**. The nugget is defined as

\[
\gamma(0^+) \equiv \lim_{\|h\| \to 0^+} \gamma(\|h\|),
\]

the sill is defined as

\[
\lim_{\|h\| \to \infty} \gamma(\|h\|),
\]

and the range is defined as the value of \(\|h\|\) at which \(\gamma(\|h\|)\) first attains the sill. An additional characteristic that is often referred to is the **partial sill**. The partial sill is the sill minus the nugget. Consider the exponential semivariogram in (1.11). Then its nugget is equal to \(\psi\), its sill is equal to \(\psi + \kappa\), its range is \(\infty\), and its partial sill is \(\kappa\). Figure 1.2 illustrates an exponential semivariogram where \(\psi = 1\), \(\kappa = 2\), and \(\phi = 0.5\).
\(\kappa = 2\), and \(\phi = 0.5\). The nugget is shown by the lower horizontal line and the sill is shown by the upper horizontal line.

![Exponential semivariogram](image)

**Figure 1.2:** Exponential semivariogram where \(\psi = 1\), \(\kappa = 2\), and \(\phi = 0.5\).

Before we end this section, we list the covariograms that correspond to (1.10), and (1.11); there does not exist a covariogram for (1.9).

1. Matérn

\[
C(||h||) = \begin{cases} 
\kappa \frac{\Gamma(\nu)}{2^{\nu-1}} \left( \frac{2\sqrt{\nu}||h||}{\phi} \right)^\nu K_\nu \left( \frac{2\sqrt{\nu}||h||}{\phi} \right) & ||h|| > 0, \nu > 0, \phi > 0 \\
\psi + \kappa & \text{otherwise.}
\end{cases}
\]

(1.14)

2. Exponential

\[
C(||h||) = \begin{cases} 
\kappa \exp \left( \frac{-||h||}{\phi} \right) & ||h|| > 0, \phi > 0 \\
\psi + \kappa & \text{otherwise.}
\end{cases}
\]

(1.15)
1.4.4 Semivariogram Model Fitting

As Cressie points out (Cressie, 1993) estimating the semivariogram is not adequate for spatial prediction. For example, semivariogram estimators such as the Matheron estimator are not guaranteed to be conditionally negative-definite. Historically, researchers used to plot the empirical semivariogram and compare it visually to the various parametric semivariogram models such as (1.9)-(1.11). Once a model was found whose shape approximately fitted that of the empirical semivariogram, researchers used trial and error to determine the values of the nugget, sill, and range parameters. The current, more formal, approach is to use methods such as ordinary least squares, weighted least squares, maximum likelihood, and restricted maximum likelihood estimation to estimate the nugget, sill, and range parameters of the spatial process. In this paper, we use maximum likelihood estimation and Bayesian inference to estimate the parameters of a spatial process.

1.5 Gaussian Processes

A spatial process defined as (1.1) is said to be Gaussian if every finite dimensional vector $Z = (Z(s_1) \ldots Z(s_n))$ for $n \geq 1$ of the process has a multivariate normal distribution, i.e.

$$Z \sim N(\mu, \Sigma) \quad (1.16)$$

where $\mu$ is the mean vector and $\Sigma$ is the variance-covariance matrix.

Gaussian process models are among the most popular models that are used to model spatial data. Gaussian processes have a number of desirable properties, for instance weakly stationary Gaussian processes are strictly stationary. Also, since Normal distributions are uniquely determined by their mean and covariance structure, Gaussian processes are uniquely determined by their mean and covariance structure. In addition, by knowing the mean and covariance structure of a Gaussian process one could determine the joint, marginal, and conditional distributions of process $Z$.

In this paper, we focus on linear Gaussian process models. This means that
process $Z$ is defined as (1.1) and

$$Z \sim N(X\beta, \Sigma(\Theta))$$

(1.17)

where $X$ is a vector or matrix of covariates, $\beta$ is a scalar or vector of regression parameters, $\Sigma$ is the variance-covariance matrix, and $\Theta$ is a vector of covariance function parameters.

In all of our simulations we focus on the linear Gaussian process where

$$\Sigma(\Theta) = \psi I + \kappa H(\phi),$$

(1.18)

and

$$\{H(\phi)\}_{ij} = \exp\left(-\frac{\|s_i - s_j\|}{\phi}\right)$$

(1.19)

with $s_i \in [0, 1]$ where $i = 1, \ldots, n$, $\Theta = (\kappa, \psi, \phi)'$ where $'$ denotes the transpose of a vector or matrix, $I$ is the identity matrix, $X$ is a $n \times 1$ vector, and $\beta$ is a scalar. In other words, $Z$ is a linear Gaussian process in one dimension with an exponential covariance function.

Researchers are often interested in estimating the covariance function parameters of a spatial process and predicting attributes of a process at unobserved locations. The goal of this paper is to discuss and compare both frequentist and Bayesian approaches in addressing these interests. In Chapter 2, we discuss frequentist inference which is also known as maximum likelihood inference. Furthermore, in Chapter 2, we discuss some of the challenges that arise when using maximum likelihood (ML) inference, how to predict a spatial process at unobserved locations using ML inference, and how to estimate standard errors. Chapter 3 discusses Bayesian inference. Specifically, it provides an overview of Bayesian inference, discusses methods of prior selection, gives a brief overview of Markov Chain Monte Carlo, and finally discusses how to predict a spatial process at unobserved locations using Bayesian inference. In Chapter 4, we summarize the results of some simulation studies that compare the two approaches in regards to estimating the covariance function parameters, predicting a spatial process at unobserved locations, and the variance of the estimates of the covariance function parameters. Chapter 4 also summarizes the results of a simulation study that com-
pares the posterior distributions based on two different priors, a data-based prior and the independence Jeffreys prior discussed in Chapter 3. Chapter 4 ends with an example comparing standard error estimation in the context of ML inference based on the parametric bootstrap and the observed Fisher information. We apply frequentist and Bayesian approaches to analyze a dataset from crop epidemiology and summarize those results in Chapter 5. Finally, the paper ends with Chapter 6 giving a summary of our findings and mentioning possible future work.
Maximum Likelihood Inference for Gaussian Process Models

2.1 Overview

Let $X$ be a random variable from a distribution $F(\theta)$ that has one parameter $\theta$. Assume that we would like to learn about $\theta$. In likelihood inference, we consider $\theta$ to be a fixed unknown quantity. Once we have a sample that is drawn from $F$ then we can learn about $\theta$. This is the basis of likelihood inference.

Let $f(\mathbf{x}|\theta)$ denote the joint probability density function (pdf) or probability mass function (pmf) of the sample $\mathbf{x} = (x_1, \ldots, x_n)$. Then, given $\mathbf{x}$, the function

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the likelihood function.

Maximum likelihood estimation is the most popular method for deriving estimators (Casella and Berger, 1990). A maximum likelihood estimator $\hat{\theta}(\mathbf{x})$ of $\theta$ based on a sample $\mathbf{x}$ is a value of the parameter $\theta$ at which the likelihood, $L(\theta|\mathbf{x})$, attains its maximum. The idea of a maximum likelihood estimator is intuitive. A maximum likelihood estimator (MLE) is the value of a parameter for which a particular observed sample is most likely.

Some issues that arise when attempting to find an MLE are whether the MLE that is found is a global maximum or not, and how sensitive the MLE is to the
observed data, i.e. would the MLE change drastically if we observed a different realization of our process?

Consider ML inference in the context of a Gaussian process as in (1.16). The likelihood of $Z$ is

$$L(\beta, \Theta|z) = \frac{1}{(2\pi)^{n/2} |\Sigma(\Theta)|^{1/2}} \exp \left( -\frac{1}{2} (z - X\beta)' \Sigma(\Theta)^{-1} (z - X\beta) \right)$$  (2.1)

Since the natural logarithm function is a one to one function, it is often easier to work with the log likelihood function, denoted by $\ell$, which is the natural logarithm of the likelihood function. Therefore, the log likelihood function corresponding to (2.1) is

$$\ell(\beta, \Theta|z) \propto -\frac{1}{2} \log |\Sigma(\Theta)| - \frac{1}{2} (z - X\beta)' \Sigma(\Theta)^{-1} (z - X\beta)$$  (2.2)

In the next two sections, we discuss some challenges that come up when using maximum likelihood inference in the context of Gaussian processes.

## 2.2 Computing

Consider the linear Gaussian process $Z(s)$ as in (1.17) with (1.18) and (1.19). As an example, let $\beta = 1$, $\kappa = 1$, $\psi = 0$ (i.e. there is no nugget), and $\phi = .05$. To simplify this example, we consider the nugget parameter $\psi$ to be fixed at 0. We simulate three realizations of this process each having a different number of points (n) and look at the ML estimates of $\beta$, $\kappa$, and $\phi$. The realizations of process $Z$ are simulated using the function $\text{grf}$ in the $\text{geoR}$ package (Ribeiro Jr. and Diggle, 2001) in the statistical software R (Ihaka and Gentleman, 1996). By default, the function $\text{grf}$ simulates a Gaussian process by the Cholesky decomposition. To simulate $Z \sim N(\mu, \Sigma)$ using the Cholesky decomposition, we do the following:

1. Simulate $X \sim N(\mu, I)$.
2. Find the Cholesky factor $U$ of $\Sigma$ such that $UU' = \Sigma$.
3. Then $Z = UX + \mu \sim N(\mu, \Sigma)$

The ML estimates of the parameters are obtained using the function $\text{likfit}$ (Ribeiro Jr. and Diggle, 2001) in the $\text{geoR}$ package. By default, the $\text{likfit}$ func-
Table 2.1: ML estimates for $\beta$, $\kappa$, and $\phi$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>MLE (n=100)</th>
<th>MLE (n=500)</th>
<th>MLE (n=1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1</td>
<td>0.1375</td>
<td>1.1202</td>
<td>0.5186</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>1</td>
<td>1.2459</td>
<td>0.8768</td>
<td>0.7300</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.05</td>
<td>0.0587</td>
<td>0.0453</td>
<td>0.0459</td>
</tr>
</tbody>
</table>

2.3 The Likelihood Surface

Recall that the likelihood is considered to be a function of the parameters while fixing the process at a particular realization. For the linear Gaussian process in Section 1.5, the likelihood function is a function of $\beta$, $\kappa$, $\phi$, and $\psi$. Because plotting the likelihood function as a function of these four parameters would result in a five dimensional plot, we consider $\beta$ and $\psi$ to be fixed at 1 for a moment and look at a three dimensional plot of the log-likelihood function versus $\kappa$ and $\phi$ where the true values of $\kappa$ and $\phi$ are 3 and 0.05 respectively (see Figure 2.1). We notice that the log-likelihood surface in Figure 2.1 is fairly flat. To illustrate the point further, let us look at two dimensional plots derived from Figure 2.1. We look at log-likelihood profile plots of $\kappa$ and $\phi$. To get a likelihood profile of $\kappa$ we maximize the log-likelihood function with respect to $\phi$ for each value of $\kappa$. The log-likelihood profile of $\phi$ is obtained in a similar way. By looking at Figure 2.2, we see that the log-likelihood profiles are fairly flat. By looking at Figure 2.2, we see that the log-likelihood profiles are fairly flat. The problem is that a fairly flat likelihood surface leads to MLEs which have a large variance. Li and Sudjianto (2005) also noticed this problem. They propose using penalized likelihood estimation with the smoothly clipped absolute deviation (SCAD) penalty as the penalty function. Using the proposed penalized likelihood reduces the variance of the MLEs but
introduces a small bias (Li and Sudjianto, 2005).

Having illustrated the challenges in ML estimation for Gaussian process models, we conclude our discussion of ML estimation with an overview of how to predict a spatial process at unobserved locations.

2.4 Prediction Based on ML Inference

Predicting a Gaussian spatial process at unobserved locations is fairly straightforward. Let $Z_1$ be an observed process at $n$ locations $(s_1 \ldots s_n)$. We are interested in predicting the value of the process at $m$ unobserved locations. Denote the process at the unobserved locations by $Z_2$. Then it is clear that

$$
\begin{pmatrix}
Z_1 \\
Z_2
\end{pmatrix}
= N_{n+m} \left( \begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}, \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix} \right). \tag{2.3}
$$
Therefore, the conditional distribution of $Z_2$ is

$$Z_2 | Z_1 \sim N_m(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1}(Z_1 - \mu_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})$$  \hspace{1cm} (2.4)$$

We use (2.4) to predict $Z_2$. However, instead of using the true values of the parameters in (2.4) we use the maximum likelihood estimates of the parameters. More specifically, for the linear Gaussian process described in Section 1.5, we use $\hat{\beta}$ and $\hat{\Theta}$.

### 2.5 Standard Error Estimation

It is important to report some measure of the variability of parameter estimates in order to give an idea of how accurate the estimates are. This is usually done by calculating standard errors. A standard error of an estimation method is the estimated standard deviation of the error of the method. This section discusses standard error estimation for ML inference based on two approaches: bootstrapping, and the observed Fisher information.
2.5.1 Bootstrapping

Bootstrapping (Efron, 1979) is an approach used to estimate properties of a statistic, such as its variance. The basic idea behind bootstrapping, is to draw samples from an approximate distribution to the true distribution, and use these samples to measure the properties of interest. Here, we briefly describe both the parametric and nonparametric bootstrap. For a more detailed discussion on bootstrapping see Efron (1979), Efron (1982), and Efron and Tibshirani (1993).

2.5.1.1 Parametric Bootstrap

The idea behind the parametric bootstrap is to fit a parametric model, often using maximum likelihood, to the data. We then draw samples from the parametric model and use these samples to estimate properties of interest. In this paper, we are interested in estimating the standard errors for the parameter estimates of the linear Gaussian process model discussed in Section 1.5. Therefore, in our simulation studies, we will implement the parametric bootstrap method as follows:

1. Fit a linear Gaussian process model to the observed realization of a spatial process using maximum likelihood estimation.
2. Simulate a realization from the fitted model.
3. Use likfit to obtain MLEs of $\beta$ and $\Theta$.
4. Repeat steps 2 and 3 until the desired number of samples is obtained.

To estimate the bootstrap standard error of an estimator, $\widehat{\theta}$ of a parameter $\theta$, we use the following formula:

$$SE(\widehat{\theta}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\widehat{\theta}_i - \overline{\widehat{\theta}})^2} \quad (2.5)$$

where $n$ is the total number of bootstrap samples, $\widehat{\theta}_i$ is the estimator of $\theta$ based on the $i^{th}$ sample, and $\overline{\widehat{\theta}}$ is the mean of all $\widehat{\theta}_i$. 
2.5.1.2 Nonparametric Bootstrap

The nonparametric bootstrap method, uses observed data of a process and re-samples from this data with replacement. This can be difficult to do in general, particularly because one needs to resample the data in blocks where the block size depends on the degree of dependence. For a detailed discussion see Lahiri (2002), Politis et al. (1999), and Zhu and Lahiri (2007).

2.5.2 Fisher Information

In the univariate case, under certain regularity conditions, such as the parameter $\theta$ is identifiable, the density $f(z|\theta)$ is three times differentiable with respect to $\theta$, the third derivative of $f(z|\theta)$ is continuous in $\theta$, and $\int f(z|\theta)dz$ can be differentiated three times under the integral sign,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{I(\theta)}\right)$$

where $\hat{\theta}$ is the MLE of $\theta$, $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution or law, and $N\left(0, I^{-1}(\theta)\right)$ is a Normal distribution with mean 0 and variance $I^{-1}(\theta)$, see for instance Casella and Berger (1990). $I(\theta)$, is called the (expected) Fisher information, and it is defined as

$$I(\theta) = E\left(\frac{\partial}{\partial \theta} \ell(\theta|Z)\right)^2. \quad (2.6)$$

An alternative form of the Fisher information, $I(\theta) = -E\left(\frac{\partial^2}{\partial \theta^2} \ell(\theta|Z)\right)$, may be used when we have an exponential family distribution such as the Normal distribution. This form is used to define the observed Fisher information which is

$$\mathcal{I}(\theta) = -\frac{\partial^2}{\partial \theta^2} \ell(\theta|Z). \quad (2.7)$$

Efron and Hinkley (1978) show that using the observed Fisher information is better than the expected Fisher information for several reasons; the reciprocal of the observed Fisher information is shown to approximate the conditional variance of $\hat{\theta}$, the MLE of $\theta$, given an appropriate ancillary statistic, the observed Fisher
information is easier to compute, and it is closer to the data than the expected Fisher information.

For a Gaussian process we can estimate the standard error of an estimate based on maximum likelihood estimation by calculating the observed Fisher information. For related work, see Ying (1991, 1993), and Abt and Welch (1998). For a guide on how to compute the observed Fisher information matrix for a Gaussian process see Wolfinger et al. (1994).
Bayesian Inference for Gaussian Process Models

3.1 Introduction

Bayesian inference, named after the British mathematician Thomas Bayes, takes advantage of any prior knowledge about a parameter of interest. This information is then combined with information from observed data to make inference. Let $X$ be a random variable of a distribution which has one parameter, $\theta$. Assume we would like to estimate $\theta$. Bayesian inference is different from maximum likelihood inference in that the parameter $\theta$ is no longer considered an unknown fixed quantity. Instead, $\theta$ is considered to be a random variable. Any knowledge or belief that we may have about $\theta$ before looking at the data is described by a distribution known as the prior distribution. Once we look at the data, we describe our updated knowledge about $\theta$ by the posterior distribution using Bayes’ Rule (Bayes, 1763).

If $X$ is a random variable of a one parameter ($\theta$) distribution then Bayes’ Rule states

$$
\pi(\theta|x) = \frac{p(\theta)f(x|\theta)}{f(x)} \tag{3.1}
$$

where $p(\theta)$ is the prior density function, $f(x|\theta)$ is the conditional density function of $X$ given $\theta$, $f(x)$ is the marginal density function of $X$, and $\pi(\theta|x)$ is the posterior density function. Note, we are using the term “density function” to refer to either a probability density function (pdf) or a probability mass function (pmf). The
The marginal density of $X$ is defined as $f(x) = \int p(\theta) f(x|\theta) \, d\theta$. Historically, frequentists have criticized Bayesian methods for relying on computationally convenient priors (Carlin and Louis, 2000). On the other hand, Bayesians have criticized frequentists for not being able to incorporate prior information and for not being able to process information systematically (Carlin and Louis, 2000).

Bayesian methods are particularly useful for Gaussian process models. They enable us to easily incorporate variability in parameter estimates when making predictions. Also, with Bayesian methods, one can utilize prior information when such information is available and it is easy to build multilevel and/or hierarchical models. Some of the general advantages of Bayesian methodology listed in Berger (1985) are:

1. They provide us with a way to incorporate prior information.

2. Inferences based on Bayesian approaches are conditional on actual data.

3. Bayesian answers are easier to interpret.

As an example of point 3 above, consider a credible set or the Bayesian equivalent to a frequentist confidence interval. The interpretation of a 95% (frequentist) confidence interval for a parameter $\theta$ is that if we repeat the experiment many times and calculate a 95% confidence interval for $\theta$ each time, we expect that 95% of the intervals will include the true value of $\theta$. On the other hand, a 95% Bayesian confidence interval (or credible set) is an interval which will contain the true value of $\theta$ with probability 0.95. A $(1-\alpha)\% \times 100$ credible set for a parameter, $\theta$, is defined as a subset $C$ of the parameter space of $\theta$ which satisfies $P(\theta \in C | Z = z) \geq 1 - \alpha$. Credible sets are of different types. For example, there are equal-tail credible sets and highest posterior density (HPD) credible sets. In our simulation studies we use HPD credible sets. A HPD set is defined as the set $C$ such that

$$C = \{ \theta : \pi(\theta|z) \geq k_\alpha \}$$

(3.2)

where $k_\alpha$ is the largest number such that $P(C|z) \geq 1 - \alpha$. HPD credible sets are shortest in length (Carlin and Louis, 2000).
3.2 Prior Selection

The question of how to select a prior for a parameter of interest is an open area of research. Researchers have recommended many different types of priors. These include subjective priors, default priors, and data-based priors. One important point that needs to be kept in mind is that priors need to yield proper posteriors i.e. posterior distributions whose densities integrate to 1. We will briefly describe the three types of priors just mentioned and we will discuss two of them in the context of linear Gaussian process models.

3.2.1 Subjective Priors

Subjective priors are priors that are based on similar past studies or historical information. To understand subjective priors one needs to be familiar with the notion of subjective Bayesian analysis which we discussed in Section 3.1. The following discussion is based on Goldstein (2006).

The basic idea behind the subjective Bayesian approach is that we quantify our uncertainties about the parameter(s) of interest and the observations we might make given those parameters as (subjective prior) probabilities and conditional probabilities respectively. Once the data is collected, we update our prior probabilities to new conditional probabilities. This approach can be seen as a methodology for representing and connecting prior information from previous studies. The newly collected data is a useful resource that may confirm or change our prior beliefs about the uncertainties in question. Therefore, the subjective Bayesian approach uses the collected data to apply it to a wider context as opposed to considering the study at hand to be as separate and not relevant to other prior information.

Two main objections have been brought up against the subjective Bayesian approach. The first is based on the idea that in scientific analyses, there is no room for subjective judgment. The second, is that its implementation is too complicated. On the other hand, Goldstein (2006) argues that two problems are faced when one turns away from the subjective Bayesian approach. The first, is that we may calculate results in which we have no confidence. The second problem is that we may be unclear about what our analysis claims to represent. See Goldstein (2006)
for more details.

\subsection{Default Priors}

Default priors are different from subjective priors in that they are not based on past studies nor historical information. Default priors are less restrictive and more general because not much information is known about the parameter of interest to begin with. Many default priors are based on the parameter space.

Default priors are sometimes referred to as objective priors. For a detailed discussion about objective priors see Berger (2006). Sometimes, default priors are referred to as “noninformative” priors. However, we prefer not to use this term as we feel it is misleading. For example, a prior for a parameter \( \theta \) that is uniformly distributed on the positive real line (known as a flat prior) is an example of a default prior. It would be inaccurate to call this prior a noninformative prior as it tells us that \( \theta \) is positive and every value of \( \theta \) is equally likely.

The class of default priors includes flat priors, vague priors, and the Jeffreys prior or modified versions of the Jeffreys prior. We will concentrate our discussion on the independence Jeffreys prior. But first we will introduce the Jeffreys prior.

\subsubsection{Jeffreys Prior}

The Jeffreys prior (Jeffreys, 1961) for a parameter \( \theta \) of a process \( Z \) is defined as

\[ \pi(\theta) \propto \sqrt{I(\theta)} \]

where \( I(\theta) \), is the Fisher information as defined in (2.6). A main desirable property of the Jeffreys prior is that it is invariant under one to one transformations. In other words, if \( \beta = h(\theta) \), and \( h \) is an invertible function with inverse function \( \theta = g(\beta) \) then the Jeffreys prior for \( \beta \) is defined as:

\[ \pi(\beta) = \pi(g(\beta)) \left| \frac{dg(\beta)}{d\beta} \right| = \pi(\theta) \left| \frac{d\theta}{d\beta} \right| . \]

We illustrate this with a simple example.
Example

Let $X_1, \ldots, X_n$ be i.i.d $N(0, \theta)$ variables. Assume that we would like to derive the Jeffreys prior for $\theta$. Since the $X_i$’s are normally distributed we can use the alternative form of the Fisher information discussed in Section 2.5.2. We derive the Jeffreys prior as follows:

\[
\ell(\theta | X) \propto -\frac{n}{2} \log \theta - \frac{\sum X_i^2}{2\theta}
\]

\[
\frac{\partial}{\partial \theta} \ell(\theta | X) = -\frac{n}{2\theta} + \frac{\sum X_i^2}{2\theta^2}
\]

\[
\frac{\partial^2}{\partial \theta^2} \ell(\theta | X) = \frac{n}{2\theta^2} - \frac{\sum X_i^2}{\theta^3}
\]

So the Fisher information is

\[
I(\theta) = -E \left( \frac{n}{2\theta^2} - \frac{\sum X_i^2}{\theta^3} \right) = -\frac{n}{2\theta^2} + \frac{n}{\theta^2} = \frac{n}{2\theta^2}
\]

Therefore, the Jeffreys prior is

\[
\pi(\theta) \propto \frac{1}{\theta} \sqrt{\frac{n}{2}}.
\]

Consider transforming $\theta$ to $\beta = \frac{1}{\theta}$. Following (3.4), we get

\[
\pi(\beta) \propto \pi(\theta) \left| \frac{d\theta}{d\beta} \right| = \frac{1}{\theta} \sqrt{\frac{n}{2}} \left| \frac{d\frac{1}{\beta}}{d\beta} \right| = \frac{1}{\beta^2} \sqrt{\frac{n}{2}} = \frac{1}{\beta} \sqrt{\frac{n}{2}}.
\]

As can be seen, the Jeffreys prior for $\beta$ is the same as $\pi(\theta)$. All we have to do is replace $\theta$ by $\beta$. This is important, as it allows researchers to perform one to one transforms on parameter(s) of interest without having to select a different prior.

### 3.2.2.2 Independence Jeffreys Prior

This section follows the description in Paulo (2005). Consider the linear Gaussian process $Z$ as described in Section 1.5 with mean $\mathbb{E}Z(s) = X(s)\beta$. Assume that the nugget parameter $\psi$ is fixed. Recall, $X(s)$ is an $n \times 1$ vector of locations. Define $\eta = (\kappa, \beta, \phi)'$. The resulting random vector $Z$ satisfies $Z|\eta \sim N(X(s)\beta, \kappa\Sigma)$ where $\Sigma = H(\phi)$ and $\{H(\phi)\}_{ij} = \exp \left( -\frac{||s_i - s_j||}{\phi} \right)$. The independence Jeffreys prior on $\eta$ is defined to be of the form

\[
\pi(\eta) \propto \frac{\pi(\phi)}{\kappa^a}
\]

with $a = 1$ and $\pi(\phi) \propto |I_J(\phi)|^{1/2}$ where
\[ I_J(\phi) = \begin{pmatrix} n & trU \\ trU & trU^2 \end{pmatrix}. \]

\( U = \hat{\Sigma}\Sigma^{-1}, \) where \( \hat{\Sigma} \) is the matrix that results by differentiating each of the elements of \( \Sigma \) with respect to \( \phi \), and \( trU \) is the trace of matrix \( U \). This prior is flat on the parameters that specify the mean i.e. \( \beta \). With the independence Jeffreys prior we assume \( \beta \) and \((\kappa, \phi)\) a priori independent. Paulo (2005) states that under certain conditions, the independence Jeffreys prior (as well as two others) yields a proper posterior.

Later on, through a simulation study, we will look at the posterior distributions of \( \beta, \kappa, \) and \( \phi \) when using the independence Jeffreys prior and compare them to those based on a data-based prior discussed in the next section.

### 3.2.3 Data-Based Priors

Data-based priors are priors for which observed data are used to specify their parameters. In this section, we explore a very informal approach to deriving priors as described in Finley et al. (2007). This method borrows information from the empirical semivariograms to help define the prior parameters. Strictly speaking, this is not full Bayesian inference since the prior is obtained from the data, so the data are used twice. According to this approach, an inverse gamma prior is specified for the partial sill parameter \( \kappa \) and for the nugget parameter \( \psi \), and a uniform prior is specified for \( \phi \). The inverse gamma density with shape parameter \( a \) and scale parameter \( b \) has the form

\[
\frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp \left( \frac{-b}{x} \right)
\]

where \( \Gamma(\cdot) \) is the gamma function. Following Finley et al. (2007), the shape parameter, \( a \), is set to 2 and the scale parameter, \( b \), is set to the estimate of the partial sill or the nugget from the empirical semivariogram. A shape of 2 in the inverse gamma distribution implies a distribution with infinite variance centered on the scale parameter \( b \) (Finley et al., 2007).
Note that in Finley et al. (2007), $\phi$ in the exponential covariance function is reparameterized. The function $H(\phi)$ is specified as

$$\{H(\phi)\}_{ij} = \exp (-\phi \| s_i - s_j \|).$$

(3.7)

Before we discuss the uniform prior for $\phi$ we need to introduce the term effective range. The effective range is defined as the distance, $d = \| s_i - s_j \|$, at which the correlation is negligible. By common convention, 0.05 is considered as negligible correlation. If $H(\phi)$ has form (3.7) the effective range, $d$, is

$$\exp (-\phi d) = 0.05 \implies d \approx \frac{3}{\phi}.$$ 

Alternatively, if $H(\phi)$ has form (1.19) the effective range is $d = 3\phi$. Therefore, if the empirical semivariogram suggests that $d = 1$ for example, then $\phi$ is either 3 or $\frac{1}{3}$ depending on the form of $H(\phi)$ that we are using. Taking information from the empirical semivariogram and the data one can specify values for the parameters $a$ and $b$ of the Uniform($a, b$) prior for $\phi$. Our strategy will be to choose a support for $\phi$ that will allow for a series of Gaussian process models that range from low dependency to high dependency. We will give an example of what values to select for the parameters of the uniform prior in Chapter 4.

When using Bayesian inference, researchers often use Markov Chain Monte Carlo (MCMC) to draw samples from the posterior distribution. The next section gives an overview of MCMC.

### 3.3 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is a collection of algorithms that are used to make approximate draws from complicated probability distributions. There are a number of MCMC algorithms. We will concentrate our discussion on two of them: the Metropolis-Hastings algorithm and the Gibbs sampler. However, first we will give a basic overview of Markov chains.
3.3.1 Markov Chains

A Markov chain, named after Andrey Andreyevich Markov, is a sequence of random variables $X_0, X_1 \ldots$ which has the Markov property. The Markov property states that given the current state, the future is independent of the past states. Stated in terms of probabilities, the Markov property satisfies

$$P(X_{i+1} = x_{i+1}|X_0 = x_0, X_1 = x_1 \ldots X_i = x_i) = P(X_{i+1} = x_{i+1}|X_i = x_i).$$

The state space, $S$, is the set of possible values for the random variables $X_0, X_1 \ldots$. A move from one state to another is known as a step. If the Markov chain is currently at state $i$ then the next state of the chain will be $j$ with transition probability $p_{ij}$.

We briefly discuss some of the properties of Markov chains. A state $j$ is accessible from state $i$ if there exists an $n \geq 0$ such that $P(X_n = j|X_0 = i) > 0$. States $i$ and $j$ are said to communicate if $i$ is accessible from $j$ and $j$ is accessible from $i$. If starting in state $i$, any return to state $i$ must occur in a multiple of $k$ steps, then we say state $i$ has period $k$. If $k = 1$, then we say state $i$ is aperiodic. State $i$ is said to be recurrent, if starting in state $i$ the probability of never coming back to it is 0. We say that state $i$ is positive recurrent if it is recurrent and starting in state $i$, the expected time to return to it is finite. If a state is aperiodic and positive recurrent then we say it is ergodic. A Markov chain is said to be ergodic if all of its states are ergodic. For a more detailed discussion on Markov chains see Ross (2007).

3.3.2 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is named after Nicholas Metropolis (Metropolis et al., 1953) and Hastings (1970) who helped generalize the algorithm. Suppose we are interested in sampling from $f(x)$, the density of $x$, where $x$ is a vector or scalar. We first choose a conditional density $q(x^*|x^{(i)})$ known as a proposal density where $x^*$ is a possible value of the chain and $x^{(i)}$ is the current state of the chain. Drawing samples from the proposal density should be a straightforward process. The proposal density may or may not depend on the current state of the
Markov chain. The outline of the Metropolis-Hastings algorithm is as follows:

1. Start with $i = 0$. Choose a value for $x^{(0)}$, the starting value of the chain.

2. Draw $x^*$ from $q(x^*|x^{(0)})$.

3. Accept $x^*$, i.e. set $x^{(i+1)} = x^*$ with probability

   $$
   \alpha(x^{(i)}, x^*) = \min \left( \frac{f(x^*)q(x^{(i)}|x^*)}{f(x^{(i)})q(x^*|x^{(i)})}, 1 \right)
   $$

   otherwise set $x^{(i+1)} = x^{(i)}$.

4. Set $i = i + 1$.

5. Repeat steps 2-4 until the desired Markov chain length is attained.

Now we have a Markov chain $\{x^{(0)}, x^{(1)}, \ldots, x^{(n)}\}$ with desired length $n$.

The Metropolis-Hastings algorithm has a number of useful properties. Some of which are: ease of implementation, ability to use any choice of proposal density that has the same support as $f$, and the ability to use it to draw samples from $f$ or any other function that is proportional to $f$ without requiring the knowledge of what the normalizing constant of $f$ is. See Jones and Hobert (2002) for examples and more details.

### 3.3.3 The Gibbs Sampler

The Gibbs sampler (Gelfand and Smith, 1990) is a special case of the Metropolis-Hastings algorithm. Following the discussion in Jones and Hobert (2002), suppose we are interested in sampling from $f(x)$, the density of $x$ where $x = (x_1, \ldots, x_n)'$.

Furthermore, suppose that it easy to sample from the full conditional densities, $f_i(x_i|x_{-i})$ where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)'$. Note that

$$
 f_i(x_i|x_{-i}) = \frac{f(x)}{\int f(x)dx_i} \propto f(x)
$$

The algorithm for the Gibbs sampler is as follows:

1. Start with $i = 0$. Choose a value for $x^{(0)}$, the starting value of the chain.
2. Draw \( \mathbf{x}^{(i+1)} \) as follows:

\[
\begin{align*}
    x_1^{(i+1)} &\sim f_1(x_1|x_2^{(i)}, \ldots, x_n^{(i)}) \\
    x_2^{(i+1)} &\sim f_2(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)}) \\
    &\vdots \\
    x_n^{(i+1)} &\sim f_n(x_n|x_1^{(i+1)}, \ldots, x_{n-1}^{(i+1)})
\end{align*}
\]

3. Set \( i = i + 1 \).

4. Repeat steps 2 and 3 until the desired Markov chain length is attained.

If the conditional densities are proportional to a known density, then it would be easy to sample our \( x_i \)'s. However, if this is not the case, then one can use the Metropolis-Hastings algorithm discussed in the previous section.

### 3.3.4 Standard Error Estimation

One important question that needs to be addressed is how to assess the accuracy of our MCMC methods. To do this, we use the consistent batch means method in Jones et al. (2006). There is a vast literature on sophisticated methods for computing Monte Carlo standard errors for MCMC output. However, batch means has the advantage of being easy to implement and it appears to work reasonably well in practice. Suppose we are interested in estimating an expectation \( \mu = E(g(X)) \) where \( X \) is a draw from a distribution \( f \). The basic idea of the batch means method is to divide the Markov chain \( \{X_n\} \) into batches and compute, for each batch, the Monte Carlo estimate of \( \mu \); call it \( Y_k \) where \( k \) refers to the batch. The batch means estimate of the Monte Carlo standard error is \( \frac{\hat{\sigma}}{N} \) where

\[
\hat{\sigma}^2 = \frac{b}{a-1} \sum_{k=1}^{a} (Y_k - \hat{\mu})^2,
\]

and \( N = ab \) the length of the Markov chain, \( a \) is the number of batches, \( b \) is the size of each batch, and \( \hat{\mu} \) is the estimate of \( \mu \) based on the whole Markov chain. For ideas on how to select the batch sizes see Jones et al. (2006).
3.4 Prediction Based on Bayesian Inference

Prediction based on Bayesian inference is done in a similar fashion as in ML inference. The only difference is that for each posterior sample obtained we calculate (2.4). However, instead of using the true values of the parameters in (2.4) we use the posterior sample mean for each parameter. So if we have $n$ posterior samples then we calculate (2.4) $n$ times obtaining posterior predictive samples. In other words, as in Section 2.4, let $Z_1$ be an observed process at $n$ locations ($s_1 \ldots s_n$). We are interested in predicting the value of the process at $m$ unobserved locations. Denote the process at the unobserved locations by $Z_2$. Then for each posterior sample $i = 1 \ldots n$, simulate $Z_2^{(i)} \sim N(X\beta^{(i)}, \Sigma(\Theta^{(i)}))$ where $\beta^{(i)}$ and $\Theta^{(i)}$ are the posterior means of $\beta$ and $\Theta$ respectively obtained from posterior sample $i$. Obtaining posterior predictive samples is computationally expensive as the variance-covariance matrix is recalculated for each posterior sample.
Chapter 4

Simulation Studies

This chapter discusses the details and results of simulation studies that were conducted in order to compare ML and Bayesian approaches in terms of estimation, prediction, and variability of estimation, and also to compare the posterior distributions based on the data-based prior discussed in Chapter 3 and the independence Jeffreys prior. All simulation studies were done for the linear Gaussian process model discussed in Section 1.5. Recall that each location, \( s_i \), for this process is one-dimensional, continuous, and \( s_i \in [0, 1] \).

We selected three parameter sets to be used in the simulation studies. They are shown in Table 4.1. Note the “-” means that the nugget parameter \( \psi \) was fixed at 0 and was not estimated. Figure 4.1 shows a plot of a realization from each parameter set in Table 4.1, along with a fitted lowess line, and a plot of the corresponding theoretical semivariogram. Looking at the theoretical semivariograms in Figure 4.1, we notice that the theoretical semivariogram takes longer to reach the sill for parameter Set 1; therefore, Set 1 represents a more dependent process than sets 2 and 3. To illustrate this further, the effective ranges for sets 1 and 2 are \( 3\phi = 0.6 \), and 0.3 respectively. This means that for sets 2 and 3, two locations have to be at

<table>
<thead>
<tr>
<th>Set</th>
<th>( \beta )</th>
<th>( \kappa )</th>
<th>( \psi )</th>
<th>( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>-</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 4.1: Parameter sets used for simulation studies.
least 0.3 distance units apart to be considered to have a negligible or no effect on each other; whereas for Set 1, two locations have to be at least 0.6 distance units apart to be considered to have a negligible or no effect on each other. Based on the effective range, we conclude that Set 1 represents a more dependent process than sets 2 and 3. In the next sections we discuss each simulation study and its results.

Figure 4.1: Plots of realizations from parameter sets 1-3 and their corresponding theoretical semivariograms.

All simulation studies were coded in R (Ihaka and Gentleman, 1996), version 2.6.1.

4.1 Comparison of Estimation and Prediction

The aim of this simulation study is to compare parameter estimation and prediction at unobserved locations based on ML and Bayesian inference. We look at two different spatial processes. We simulate realizations from each process using the
function grf (Ribeiro Jr. and Diggle, 2001) in R. To estimate the parameters using ML inference, we obtain maximum likelihood estimates using the function likfit (Ribeiro Jr. and Diggle, 2001). In likfit we set the initial values of the parameters $\kappa$ and $\phi$ as 0.5. To predict the process at unobserved locations we follow Section 2.4.

For the Bayesian approach, we use MCMC to obtain posterior means to estimate the parameters. The Markov chain is run using the function ggt.sp in the spBayes (Finley et al., 2007) package. We fit a model to the empirical variogram using the function variofit in the geoR package. In variofit, we set the initial values for $\kappa$ and $\phi$ as 1 and 0.1 respectively. We also set the weights for the loss function to equal. The minimization function that was used is nls, which stands for “nonlinear least squares.” This is the default in R when weights are set to equal. Other options are the general purpose optimization function (optim) and nonlinear minimization (nlm). To predict the process at unobserved locations we use posterior predictive sampling as outlined in Section 3.4. To summarize, the design of this simulation study is as follows:

1. Simulate a realization at $n + m = 150$ locations from linear Gaussian process $Z$ with chosen values for $\beta$ and $\Theta$. Hide $m = 50$ observations from the simulated realization. We will predict the process at the locations of the hidden observations.

2. Obtain maximum likelihood estimates for $\beta$ and $\Theta$ based on the $n = 100$ observations.

3. Predict the process at the $m = 50$ locations using the ML approach.

4. Fit a linear Gaussian process model with an exponential covariance function to the empirical variogram. Set the priors of the parameters using the fitted model.

5. Run a Markov chain of length 10,000. Obtain sample-based estimates of the posterior distribution of $\beta$ and $\Theta$.

6. Predict the process at the $m = 50$ locations using the posterior predictive distribution of the process.
7. Repeat steps 1-6 3,000 times.

The priors are specified as in Section 3.2.3. Recall that $\phi$ is 0.2 for Set 1 and 0.1 for sets 2 and 3. According to the reparameterization in Finley et al. (2007) shown in (3.7), $\phi$ is 5 and 10 respectively. Our strategy for selecting the support for the uniform prior is to choose a support that will allow the process to vary in its dependency. For example, since we know that the maximum distance between locations for the process that we are considering is 1, an effective range of 15 implies a highly dependent process; this is because a value of 15 means that the correlation is negligible once two locations are at least 15 distance units apart (i.e. more than 15 times the maximum distance apart!). On the other hand, an effective range of 0.15 implies fairly low dependent process. Therefore, assuming the parameterization in (3.7) which implies that $\phi \approx \frac{3}{d}$ where $d$ is the effective range, we choose (0.2, 20) as the support on the uniform prior. The Uniform(0.2, 20) prior was used as a prior for $\phi$ for all three parameter sets.

To compare the estimation of the parameters we use the mean squared error (MSE). Recall that the MSE of an estimator $\hat{\theta}$ for a parameter $\theta$ is $E_\theta(\hat{\theta} - \theta)^2$.

To evaluate how well each approach predicted the process at the 50 locations, we use a modification of the integrated empirical mean squared error (IEMSE) (Abt, 1999). The modified version which we still refer to as IEMSE is:

$$
\frac{1}{m} \sum_{i=1}^{m} \frac{1}{N} \sum_{j=1}^{N} \left( \hat{Z}^j(s_i) - Z^j(s_i) \right)^2
$$

where $i$ indicates the $i$th hidden location, $m$ is the total number of locations we would like to predict at (i.e. the total number of hidden locations), $j$ refers to the $j$th simulated realization, and $N = 3,000$. $\hat{Z}^j(s_i)$ is the predicted value at location $s_i$ based on the $j$th simulated sample, and $Z^j(s_i)$ is the true value at location $s_i$ for the $j$th simulated sample. In Abt (1999), $\hat{Z}^j(s_i)$ is the predicted value of the process based on maximum likelihood estimation of the parameters. Here, $\hat{Z}^j(s_i)$ could also be based on posterior predictive distributions. Also, in Abt (1999), $Z^j(s_i)$ is replaced by the $j$th simulated sample from the marginal distribution of the process $Z$.

To compare the prediction of the two approaches we also compare their coverage
<table>
<thead>
<tr>
<th>Parameter</th>
<th>ML MSE(SE)</th>
<th>Bayes MSE(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.552 (0.041)</td>
<td>1.507 (0.04)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>1.406 (0.087)</td>
<td>2.109 (0.053)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.016 (0.001)</td>
<td>0.025 (0.001)</td>
</tr>
</tbody>
</table>

Table 4.2: MSE results for $\beta = 3$, $\kappa = 2$, and $\phi = 0.2$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Coverage(SE)</th>
<th>IEMSE(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>0.9456 (7e-04)</td>
<td>0.0925 (0.00031)</td>
</tr>
<tr>
<td>Bayes</td>
<td>0.949 (7e-04)</td>
<td>0.0938 (0.00032)</td>
</tr>
</tbody>
</table>

Table 4.3: IEMSE for $\beta = 3$, $\kappa = 2$, and $\phi = 0.2$.

probabilities. For each of the 50 locations, we look at the 95% confidence interval (CI) for the ML approach and the 95% credible set (described in Section 3.1) for the Bayesian approach and calculate the proportion of times each covers the true value of the process at a particular location. For example, if the coverage probability for the ML approach is 0.95, this means that, on average, the ML approach was able to include the true value of the process for each of the $m = 50$ locations 95% of the time.

Note: While running some preliminary simulations, we noticed that at times, variofit (Ribeiro Jr. and Diggle, 2001) returned a huge estimate of $\kappa$. In this case, we ignored the estimate and simulated another 150 observations.

### 4.1.1 Results and Conclusions

The results of the simulation study for parameter sets 1 and 3 are shown in Tables 4.2-4.5.

Note that (SE) refers to the standard error of our estimates. For example, consider the column labeled ML MSE (SE); the number in parenthesis is the MSE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ML MSE(SE)</th>
<th>Bayes MSE(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>1.54 (0.041)</td>
<td>1.535 (0.041)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>1.769 (0.146)</td>
<td>1.936 (0.058)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>0.052 (0.001)</td>
<td>0.045 (0.001)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.015 (0.006)</td>
<td>0.008 (0.001)</td>
</tr>
</tbody>
</table>

Table 4.4: MSE results for $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. 
<table>
<thead>
<tr>
<th>Method</th>
<th>Coverage(SE)</th>
<th>IEMSE(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>0.9409 (7e-04)</td>
<td>2.7833 (0.00652)</td>
</tr>
<tr>
<td>Bayes</td>
<td>0.6911 (0.0015)</td>
<td>1.8026 (0.00542)</td>
</tr>
</tbody>
</table>

Table 4.5: IEMSE for $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$.

standard error. Based on other similar simulation studies that we have conducted and looking at tables 4.2-4.5, we note that the MSEs based on the two approaches are fairly comparable. This suggests that the ML and Bayesian approaches yield fairly comparable results when estimating the covariance function parameters. In terms of coverage probability, the ML and Bayesian approaches both have a coverage probability of approximately 95% as shown in Table 4.3. However, this does not seem to be the case when we have a nugget. Table 4.5 shows that the coverage probability for the Bayesian approach is approximately 70%. This is much worse than the coverage probability of the ML approach which is approximately 94%; it is possible that this is an error. Other simulations need to be done for parameter sets that are different from parameter Set 3 but include a nugget parameter before we can conclude that including a nugget parameter lowers the coverage probability for the Bayesian approach. Comparing the IEMSEs in tables 4.3 and 4.5 and considering past simulations, we conclude that the IEMSEs are fairly comparable for the two approaches. An interesting note that the results in Table 4.5 imply, is that when introducing a nugget parameter, the Bayesian approach performs worse than the ML approach in terms of coverage probability. However, the Bayesian approach is better than the ML approach in prediction.

### 4.2 Comparison of Sampling Distribution and Posterior Distribution

Of interest is to get an idea of how the variabilities of the parameter estimates based on ML and Bayesian inference compare. We use the parametric bootstrap to get an approximate distribution based on ML estimation for each parameter. We also obtain a posterior distribution for each parameter and compare the standard errors of each approach by overlaying both distributions on the same plot. For
the Bayesian approach we ran a Markov chain of length 100,000. The priors were specified as in Section 3.2.3 and the prior for $\phi$ was the same prior as in Section 4.1.

The design of the simulation study is outlined below.

1. Simulate a realization at 100 locations from process $Z$ with specified values for $\beta$ and $\Theta$.

2. Obtain maximum likelihood estimates for $\beta$ and $\Theta$.

3. Repeat 10,000 times:
   
   (a) Simulate a linear Gaussian process with parameter values equal to the ML estimates in step 2.
   
   (b) Obtain ML estimates of the parameters of the simulated process. Save the estimates.

4. Run a Markov Chain of length 100,000.

### 4.2.1 Results and Conclusions

The plots of the distributions, of each parameter, based on the bootstrap and posterior estimates are shown in Figures 4.2-4.5 for parameter Set 3. Other simulations, with different parameter sets yielded similar results.

In Figures 4.2-4.5, the solid density is based on the posterior samples, the dashed density is based on the parametric bootstrap, and the dotted vertical line denotes the true value of each parameter. We notice that the density from the parametric bootstrap and the posterior density are surprisingly similar for each parameter. We also note that the shapes of the densities are similar. We calculated 95% CIs (based on parametric bootstrap) and 95% credible sets (based on HPD) and found that they both included the true value of each parameter. Looking at figures 4.2 and 4.3, we notice that, for $\beta$ and $\kappa$, the density based on the parametric bootstrap is narrower than the density based on the posterior samples. The opposite is true for $\psi$ and $\phi$ as shown in figures 4.4 and 4.5. These observations were the same for other simulations that we have conducted using different parameter sets. In conclusion, it seems that both approaches, ML and Bayesian,
Figure 4.2: Comparison of standard errors for $\beta$ where $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. Solid line: posterior density. Dashed line: bootstrap density.

are fairly comparable in that the sampling distribution of the parameters is similar to the posterior distribution of the parameters.

4.3 Comparison of Different Prior Specifications

Another point of interest is to compare the posterior distributions of each parameter based on the two methods of prior specification discussed in Chapter 3. To do this we ran an MCMC algorithm to sample from the posterior distributions using the two different priors: independence Jeffreys and data-based priors. We used the function `ggt.sp` for sampling from the posterior using the data-based prior. We obtained 100,000 samples for each approach.

Note that when using the independence Jeffreys prior, the likelihood of $Z$ is specified in a slightly different notation than (2.1). The difference is that $\kappa$ is taken
Figure 4.3: Comparison of standard errors for $\kappa$ where $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. Solid line: posterior density. Dashed line: bootstrap density.

out of the variance-covariance matrix $\Sigma$. Therefore, the likelihood is:

$$L(\eta|z) = \frac{1}{(2\pi)^{n/2}\kappa^{n/2}}|\Sigma|^{1/2} \exp\left(-\frac{1}{2\kappa} (z - X\beta)'\Sigma^{-1}(z - X\beta)\right),$$

and the log-likelihood is

$$\ell(\eta|z) \propto -\frac{1}{2} \log(\kappa) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2\kappa} (z - X\beta)'\Sigma^{-1}(z - X\beta)$$

The unnormalized joint density is $L(\eta|z) \times \pi(\eta)$ where $\pi(\eta)$ is the independence Jeffreys prior (3.5). The full conditional distribution for $\beta$ is derived as follows. First, note that,

$$z - X\beta)'\Sigma^{-1}(z - X\beta) = z'\Sigma^{-1}z - 2\beta z'\Sigma^{-1}X + \beta^2 X'\Sigma^{-1}X$$

$$= X'\Sigma^{-1}X(\beta^2 - 2\beta(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}z + (X'\Sigma^{-1}X)^{-1}z'\Sigma^{-1}z)$$
Figure 4.4: Comparison of standard errors for $\psi$ where $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. Solid line: posterior density. Dashed line: bootstrap density.

Therefore, the full conditional distribution of $\beta$ is, $\pi(\beta|z, \kappa, \phi) \propto \text{Normal}((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}z, \kappa(X'\Sigma^{-1}X)^{-1})$

Also, according to Paulo (2005), the full conditional distribution of $\kappa$ is

$$\pi(\kappa|z, \beta, \phi) \propto \text{IG} \left( \frac{n}{2}, \frac{1}{2}z'\Sigma^{-1}z \right)$$

Finally, the full conditional distribution of $\phi$ is $\pi(\phi|\beta, z) \propto L(z) \times f(\eta)$. Since the full conditional distributions of $\beta$ and $\kappa$ are known distributions, we will update both of these parameters using a Gibbs update. On the other hand, we will update $\phi$ using a Metropolis-Hastings step with a Normal proposal.

An outline of the MCMC algorithm written for the independence Jeffreys prior is as follows:

1. Initialize all parameters.
Figure 4.5: Comparison of standard errors for $\phi$ where $\beta = 2$, $\kappa = 3$, $\psi = 1$, and $\phi = 0.1$. Solid line: posterior density. Dashed line: bootstrap density.

Table 4.6: Initial parameter values for MCMC algorithm.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>0</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.01</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.01</td>
</tr>
</tbody>
</table>

2. Update $\beta$ using a Gibbs update.

3. Update $\kappa$ using a Gibbs update.

4. Update $\phi$ using a Metropolis-Hastings with a Normal proposal; meaning the proposed value of $\phi$, $\phi^*$ is drawn from a $\text{Normal}(\phi^{(i)}, \sigma^2)$ where $\phi^{(i)}$ is the current value of $\phi$. $\sigma^2$ is called the tuning parameter and was set to $0.15^2$.

5. Repeat steps 2-4 $m$ times where $m$ is the length of the Markov Chain.

We initialized the values of the parameters as shown in Table 4.6. To get an idea of the shape of priors for $\kappa$ and $\phi$, Figure 4.6 shows the priors, for parameter...
Set 2, based on the data-based and independence Jeffreys priors. The dashed priors are those based on independence Jeffreys.

Figure 4.6: Prior plots for $\kappa$ and $\phi$ where $\beta = 2$, $\kappa = 3$, and $\phi = 0.1$. Solid line: data-based prior. Dashed line: independence Jeffreys prior.

### 4.3.1 Results and Conclusions

The results of this simulation study, for parameter sets 1 and 2, are displayed in figures 4.7-4.10. The dashed posteriors are based on the independence Jeffreys prior, and the solid posteriors are based on the data-based prior. The dotted vertical lines indicate the true value of each parameter. Figures 4.7-4.10 indicate that the independence Jeffreys prior yields posterior distributions that are less variable and hence narrower distributions than the posterior distributions yielded by the data-based prior considered. The posterior distributions of $\beta$ for parameter sets 1 and 2, shown in figures 4.7 and 4.9, both include the true value of $\beta$. The posterior distribution for $\beta$ based on the data-based prior is approximately centered at the true value of $\beta$. The posterior distributions of $\kappa$ for parameter sets 1 and 2, shown in figures 4.8 and 4.10, both include the true value of $\kappa$, although the
posterior based on the independence Jeffreys prior barely includes the true value. The posterior distributions of $\phi$ for parameter Set 1, shown in Figure 4.8 both include the true value of $\phi$; this is not the case for the posterior distributions of $\phi$ for parameter Set 2. In Figure 4.10, the posterior distribution based on the independence Jeffreys prior does not include the true value of $\phi$ and the posterior distribution based on the data-based prior barely includes the true value of $\phi$.

Figure 4.7: Posterior plots for $\beta$ where $\beta = 3$, $\kappa = 2$, and $\phi = 0.2$. Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior.

### 4.4 Comparison of ML Standard Error Estimation

In this section we discuss the results of a few trials that compare standard error estimation based on the parametric bootstrap and the observed Fisher information. In these trials, a confidence interval for each parameter was calculated based on the two methods of standard error estimation. Table 4.7 displays the results of one
Figure 4.8: Posterior plots for $\kappa$ and $\phi$ where $\beta = 3$, $\kappa = 2$, and $\phi = 0.2$. Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bootstrap CI [width]</th>
<th>Observed Fisher CI [width]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>(-0.2134, 2.622) [2.835]</td>
<td>(-0.0789, 2.6290) [2.7079]</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>(0.5533, 2.8578) [2.3042]</td>
<td>(0.5369, 2.8708) [2.3339]</td>
</tr>
<tr>
<td>$\psi$</td>
<td>(0.6158, 1.4762) [0.8604]</td>
<td>(0.6128, 1.4387) [0.8259]</td>
</tr>
<tr>
<td>$\phi$</td>
<td>(0.0068, 0.1088) [0.102]</td>
<td>(0.0013, 0.0890) [0.0877]</td>
</tr>
</tbody>
</table>

Table 4.7: Confidence intervals based on parametric bootstrap and observed Fisher information for parameter Set 3 in Table 4.1.

example; for each parameter a confidence interval was calculated along with the width of the interval. Based on this example, it seems that both methods yield similar confidence intervals and hence standard error estimates. However, more simulation studies need to be done to verify this. An interesting point to note, is that for this example, the true value of $\kappa$, 3, was not included in the confidence intervals based on both approaches.
Figure 4.9: Posterior plots for $\beta$ where $\beta = 2$, $\kappa = 3$, and $\phi = 0.1$. Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior.
Figure 4.10: Posterior plots for $\kappa$ and $\phi$ where $\beta = 2$, $\kappa = 3$, and $\phi = 0.1$. Solid line: posterior based on data-based prior. Dashed line: posterior based on independence Jeffreys prior.
Chapter 5

Application to Crop Epidemics

In this chapter, we apply some of the methods discussed earlier to a data set used for studying crop epidemics.

5.1 Motivation

Fusarium Head Blight (FHB), also known as “Scab,” is a destructive disease that affects wheat crops. FHB causes a reduction in the crop yield and reduces grain quality. FHB is primarily spread to adjacent regions when spores of the fungi are windblown or splashed onto the heads of the crops. Researchers are interested in developing a system that would help determine the risk of getting FHB in an area at a certain time of the current year. This will aid farmers in growing wheat plants in low risk areas and avoid high risk areas.

Forecasting FHB is entirely based on weather conditions, such as temperature, humidity, and precipitation. Based on the weather of the current day, researchers would like to forecast the risk of FHB three weeks from the current day. In order to do this, weather data was obtained for each location using radar. Our interest is in developing a model to forecast FHB.

In this study, data was collected at experimental plots set up by plant pathologists across 12 states. Weather information and FHB disease rates were observed at each plot. Using knowledge of disease biology and the data from the experimental plots, we can obtain a model that predicts an FHB epidemic based on weather. Therefore, the probability of an FHB epidemic in site \( s \) at time \( t \), \( P(s,t) \), is a
function of the weather at site $s$ and time $t$.

Two factors help determine the risk of FHB in a certain area: weather conditions and flowering dates. High humidity, rainfall, and optimal temperatures create ideal weather conditions for an FHB epidemic. In addition, an outbreak can only occur within a small window of time around the flowering date of the crop. An FHB epidemic is most likely when optimal weather conditions occur prior to or during flowering. If the flowering and weather conditions for FHB do not coincide, there will be no epidemic.

A model needs to be developed that will take into account the spatial and temporal dependencies of an outbreak of FHB and the chance that the wheat crop is flowering to predict the probability $P(s,t)$. In this paper, we examine the data collected for the state of North Dakota in the year 2006.

5.2 Overview of Data

Weather data was obtained from 545 Rapid Update Cycle (RUC) sites for the state of North Dakota. Plant pathologists’ FHB risk predictions are available for these sites over a total of 44 days in the summer of 2006, specifically from June 8 through July 28, 2006 (though data for July 19 - July 25 are missing). Information about the flowering dates was collected from 365 surveyed sites which are different from the 545 RUC sites. The flowering dates of wheat crops were estimated in one of two ways: either the surveyors estimated the flowering date after flowering had happened or they predicted when it was going to happen for a particular site. This means that the information about the flowering dates is estimated. A map of the flowering dates for the 365 sites is shown in Figure 5.1. In this dataset, each site’s location is a two dimensional vector specifying the latitude and longitude of the site.

Based on the information collected from the 545 RUC sites, we know the probability of an FHB epidemic in site $s$ at time $t$, given that the crop at site $s$ is flowering at time $t$; we express this probability as $P(s,t|\text{flowering})$. We also know the approximate flowering dates for the 365 surveyed sites. Our strategy is to fit a linear Gaussian process model to the flowering data information from the 365 sites. Then we predict the flowering dates for the 545 sites and calculate the probability
that a crop at site $s$ is flowering at time $t$ denoted as $P(\text{flowering})$. Finally, to get $P(s, t)$, we multiply $P(s, t|\text{flowering}) \times P(\text{flowering})$.

In the next sections we will obtain a model that will predict the risk of FHB ($P(s, t)$) in the state of North Dakota using ML and Bayesian inference.

5.3 Exploratory Data Analysis (EDA)

Before carrying out a statistical analysis, it is preferable to do an EDA. An EDA helps us learn about the data at hand. We first plot the empirical directional semivariogram. A directional semivariogram helps us assess whether a spatial process is anisotropic. Recall that if the semivariogram depends also on the direction of a spatial process, we say the process is anisotropic. Figure 5.2 shows the empirical directional semivariogram. It shows lines for four directions (angles): $0^\circ$, $45^\circ$, $90^\circ$, and $135^\circ$. Based on the empirical directional semivariogram, it seems that the process is anisotropic because the four lines are not lined up with each other. We next remove a first order trend on both latitude and longitude; the plot of the new empirical directional semivariogram is shown on the left in Figure 5.3. We notice an improvement because the lines are now more lined up with each other. The plot on the right in Figure 5.3, shows the empirical directional semivariogram with
a first order trend removed from latitude only. It appears that this new directional semivariogram is very similar to the one on the left in the same figure. For this reason, we removed a first order trend on latitude only.

Now we are ready to discuss our analysis.

Figure 5.2: Empirical directional semivariogram.

5.4 ML Inference

In carrying out ML inference we fit a linear Gaussian process model with an exponential covariance function to the approximate flowering dates of the 365 sites using maximum likelihood estimation. We fit two models; one with the nugget parameter $\psi$ and one without it.

Based on the fitted models, we predict the mean flowering date for each of the 545 RUC sites. The results are shown in Figure 5.4. Looking at Figure 5.4 we notice that the two maps of the mean flowering dates are different. The map on the right shows three horizontal parallel bands of dates which indicate that as one travels northwards in North Dakota the flowering dates will be later in the
Figure 5.3: Empirical directional semivariograms for flowering dates.

<table>
<thead>
<tr>
<th>(a) Empirical directional semivariogram with 1st order trend removed.</th>
<th>(b) Empirical directional semivariogram with 1st order trend removed on latitude.</th>
</tr>
</thead>
</table>

Calendar year; this is not the case for the map on the left. The map on the left of Figure 5.4, which includes the parameter $\psi$, shows a region, in the shape of a semi-circle, in the southern part of North Dakota which has the earliest flowering dates denoted by black squares. Furthermore, the map on the left does not show three horizontal parallel bands of dates. The nugget parameter, $\psi$, accounts for non-spatial effect variation and is often attributed to measurement error. We know that the observed flowering dates for the 365 sites were only estimates. Therefore, not putting $\psi$ in the model ignores the fact that there may be measurement error; this is not valid as we know that there is measurement error. For this data, the map of the predicted mean flowering dates for the model which includes $\psi$ is more realistic. Using the predicted mean flowering dates, we can predict the risk of having an FHB epidemic at a certain location and time. We use a seven day window around the mean flowering date. This means that for a certain day of the year at a particular location, an outbreak of FHB will occur if the weather conditions are suitable for an outbreak and the current date is between seven days before and after the flowering date for the specific location.

Figure 5.5 shows the probability of an FHB epidemic at the 545 sites on July 13 2006. The map on the left illustrates the calculated probabilities when including $\psi$. 
in the model; whereas the map on the right illustrates the calculated probabilities when excluding $\psi$ in the model. The two maps in Figure 5.5 are quite similar although we see more triangles towards the bottom of the map excluding $\psi$. We

Figure 5.4: Plots of the predicted mean flowering dates for the RUC sites based on ML inference. Squares: June 8-July 5, triangles: July 6-July 12, gray circles: July 13-July 18, white circles: July 26-July 28.

Figure 5.5: Plots of the probability of an FHB epidemic at the RUC sites on July 13 2006 based on ML inference. Squares: $P(s, t) > 0.3$, triangles: $0.1 < P(s, t) \leq 0.3$, gray circles: $0.05 < P(s, t) \leq 0.1$, white circles: $P(s, t) \leq 0.05$.

now consider using Bayesian inference in the analysis of the data.
5.5 Bayesian Inference

It is of interest to use the data-based and independence Jeffreys priors mentioned in Chapter 3 to draw inference about the crop epidemics data set considered in this chapter. However, our analysis with the independence Jeffreys prior was unsuccessful. A possible reason is the tuning parameter for the Normal proposal for \( \psi \). We hope to look into this in the future. This section summarizes our analysis based on the data-based prior.

Figure 5.6 shows the mean flowering dates for the RUC sites based on the two models considered; the first model including \( \psi \) and the second excluding it. Based on this figure, we almost observe the same relationship between the maps of the two models in Figure 5.4. Therefore, based on the same reasoning for ML inference in the previous section, the map of the predicted mean flowering dates for the model which includes \( \psi \) is more realistic. Using the predicted mean flowering dates, and a seven day window around the mean flowering data, we predict the risk of having an FHB epidemic at a certain location and time. Figure 5.7 shows the probability of an FHB epidemic at the 545 sites on July 13 2006. The map on the left illustrates the calculated probabilities when including \( \psi \) in the model; whereas the map on the right illustrates the calculated probabilities when excluding \( \psi \) in the model. The two maps in Figure 5.7 are different. For example, the map on
the left in Figure 5.7 has black squares, and more triangles and gray circles than the map on the right. Since $\psi$ accounts for measurement error, we opt for the probability map which includes $\psi$.

![Figure 5.7: Plots of the probability of an FHB epidemic at the RUC sites on July 13 2006 based on Bayesian inference. Squares: $P(s,t) > 0.3$, triangles: $0.1 < P(s,t) \leq 0.3$, gray circles: $0.05 < P(s,t) \leq 0.1$, white circles: $P(s,t) \leq 0.05$.](image)

(a) With nugget parameter $\psi$. (b) Without nugget parameter $\psi$.

5.6 Conclusion

To compare Figure 5.4 to Figure 5.6, we calculated the percentage of time the Bayesian approach matched the ML approach. Overall, when there was no nugget parameter and compared to the flowering dates map based on ML, the Bayesian approach matched the ML map for 84.22% of the points. This percentage rose to 98.34% for the comparison of the maps including the nugget parameter. Table 5.1 shows the percentage of matched points specified by the different categories of dates. Based on the percentages, the two approaches are more similar when the nugget parameter is included. Comparing Figure 5.5 to Figure 5.7, the probability maps including $\psi$ are quiet similar. The probability maps which exclude $\psi$ are also similar; although the map based on the data-based prior has more white circles than the map based on ML inference, note that the white circles are close in value to the gray circles because white circles denote $P(s,t) \leq 0.05$ while gray circles denote $0.05 < P(s,t) \leq 0.1$. Table 5.2 shows the percentage of points in Figure 5.5 that were matched by the points in Figure 5.7 specified by the different categories of the probability $P(s,t)$. Note we combine the categories for $P(s,t) \leq 0.05$ and
Table 5.1: Percentage of mean flowering dates based on the ML approach that were matched by those based on the Bayesian approach.

<table>
<thead>
<tr>
<th>Date Category</th>
<th>With nugget ($\psi$)</th>
<th>Without nugget ($\psi$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 8-July 5</td>
<td>96.4%</td>
<td>NA</td>
</tr>
<tr>
<td>July 6-July 12</td>
<td>99%</td>
<td>92.9%</td>
</tr>
<tr>
<td>July 13-July 18</td>
<td>97.8%</td>
<td>81.22%</td>
</tr>
<tr>
<td>July 26 -July 28</td>
<td>98.6%</td>
<td>69%</td>
</tr>
<tr>
<td>Overall</td>
<td>98.34%</td>
<td>84.22%</td>
</tr>
</tbody>
</table>

Table 5.2: Percentage of the probabilities of an FHB epidemic based on the ML approach that were matched by those based on the Bayesian approach.

<table>
<thead>
<tr>
<th>Probability Category</th>
<th>With nugget ($\psi$)</th>
<th>Without nugget ($\psi$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(s, t) &gt; 0.3$</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>$0.1 &lt; P(s, t) \leq 0.3$</td>
<td>81%</td>
<td>49.6%</td>
</tr>
<tr>
<td>$P(s, t) \leq 0.1$</td>
<td>83.2%</td>
<td>87.4%</td>
</tr>
<tr>
<td>Overall</td>
<td>82.9%</td>
<td>77.6%</td>
</tr>
</tbody>
</table>

0.05 < $P(s, t) \leq 0.1$ since they are close in value. Based on the overall matching percentage, the ML and Bayesian approach yield more similar results when the nugget parameter is included. We end the paper with a summary of our conclusions and a list of possible problems to be considered in the future.
Conclusion and Future Work

6.1 Conclusion

In terms of estimating the covariance function parameters, both approaches, ML and Bayes, yielded fairly comparable results, as their respective MSEs were close in value. According to our findings, either approach would be fine to use for estimating parameters and prediction. We hypothesize, based on our experience here, that this may be due to the relative flatness of the likelihood and posterior surfaces. A more rigorous basis for such conclusions may be drawn from the theoretical developments on the equivalence and orthogonality of Gaussian measures (see Stein (1999) Section 6.2 and Zhang (2004)). If time permits, we recommend carrying out both approaches as a way to validate the results.

We have only examined two types of priors. The results based on the independence Jeffreys prior were not promising; sometimes the posterior distribution based on the independence Jeffreys prior did not include the true value of a parameter. The posterior distributions based on the independence Jeffreys prior were narrower than those based on the data-based prior. Our results based on the independence Jeffreys prior depended on the MCMC algorithm that we coded. Certainly, there is room for improving the MCMC algorithm. We should look into this further; for example, improperly tuning the Normal proposal for $\phi$ might have resulted in our unpromising findings. Also, no special effort was made to ensure that a certain acceptance rate for $\phi$ was attained.
6.2 Future Work

It is of interest to calculate the observed Fisher information and the simple bootstrap discussed in Section 2.5 to assess the accuracy of the parameter estimates based on ML inference and compare them to the standard errors based on the parametric bootstrap and Bayesian approaches. It would also be interesting to look at other ways of selecting priors. For example, looking at another default prior similar to the independence Jeffreys prior but one which includes a nugget parameter such as in De Oliveira (2007) would be a good start. It would be helpful to further examine why the independence Jeffreys prior performed poorly.

This paper focused on Gaussian process models. An interesting problem would be to compare ML and Bayesian inference for other models such as Gaussian Markov random field (GMRF) models and spatial generalized linear (GLM) models. Another interesting problem would be to try to find priors that achieve what Li and Sudjianto (2005) accomplish: sharper posterior surfaces and hence tighter credible regions for the parameters.
Bibliography


