ON LARGE SAMPLE SIZE ISSUES IN SPATIAL STATISTICS: INFORMATION, COMPUTATION, AND DESIGN

by

Gangqiang Xia

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Alan E. Gelfand, Supervisor

_____
Dr. Robert L. Wolpert

_____
Dr. Merlise A. Clyde

_____
Dr. Nils L. Hjort

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2006

ABSTRACT

(Spatial Statistics)

ON LARGE SAMPLE SIZE ISSUES IN SPATIAL STATISTICS: INFORMATION, COMPUTATION, AND DESIGN

by

Gangqiang Xia

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Alan E. Gelfand, Supervisor

_____
Dr. Robert L. Wolpert

_____
Dr. Merlise A. Clyde

_____
Dr. Nils L. Hjort

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2006

Abstract

The focus of this thesis is on large sample size issues in spatial data analysis. Specifically, it consists of three major parts: spatial asymptotics; model fitting for large spatial datasets; and spatial design for one-time sampling. Our major contributions involve (i) providing various new results for spatial asymptotics; (ii) development of three new spatial process approximation methods useful for handling large spatial datasets; and (iii) development of approximately optimal sampling approaches for extensive spatial sampling.

Performing large sample analysis for spatially dependent data is challenging. Based on different spatial sampling schemes, we consider three types of asymptotics: *infill* asymptotics, *expansion* asymptotics, and so called *"middle-ground"* asymptotics. The first two asymptotics are well known but not fully studied. Middle-ground asymptotics is a new territory. We study the limiting behavior of the Fisher information matrix, the asymptotic properties of various estimators, and the weak identifiability of the parameters in spatial models under these three asymptotics scenarios.

Historically, it has been difficult to apply spatial modeling techniques to analyze large spatial datasets. The problem is that we have to handle the inversion and determinant computation of a covariance matrix with the size same as the sample size. Consider fitting a Gaussian spatial model for a spatial dataset with a large sample size n. Likelihood based or Bayesian modeling suffer from severe computational difficulties since each evaluation of the exact likelihood requires an $O(n^3)$ operation. We refer to this computational challenge as the "large n problem." We develop a new finite sum process approximation model which is both theoretically attractive and computationally efficient. The model is implemented in a Bayesian framework and

applied to analyze several large spatial datasets.

Finally, we consider the problem of approximately optimal design in the special case of one-time sampling at a large number of spatial locations. Our goal is to develop a good design strategy to help practitioners select sampling locations.

Acknowledgements

This dissertation would not have been possible without the support of many people.

First of all, I would like to express my life-long gratitude to my advisor Prof. Alan E. Gelfand, for his encouragement, insightful guidance and endless support throughout my graduate school experience. His wide knowledge and logical way of thinking have been of great value to me. It has been a privilege, both from an academic and personal point of view, to work with him.

I would also like to thank Prof. Robert L. Wolpert, who provided much help to me on stochastic process theory and oriented me in the right research direction. I also thank Prof. Merlise Clyde, for her detailed and helpful comments on my research and thesis, and Prof. Michael Lavine, for his inspiring discussions on my research work.

I owe my most sincere gratitude to Prof. Nils L. Hjort, who introduced me to the field of asymptotic analysis and contributed much in developing Chapter 3 of my dissertation.

I warmly thank Prof. Marie Lynn Miranda for her funding support. She has given me the opportunities to involve in several research projects which led to part of this dissertation.

Thanks go as well to Prof. Bibhuti Bhattacharyya and Prof. Subhashis Ghosal for their valuable advice and suggestions on my research, and to Dr. Sourabh Bhattacharya, for his helpful discussions and friendly help on my work. And special thanks to Krista, Pat, Susan, Anne, Eric and Lance for all their administrative support during my study at ISDS.

Last but not least, I wish to dedicate this work to my mother and my wife Qingyi for their love, understanding and support.

Contents

Abstract **iv**

Acknowledgements **vi**

List of Tables **xi**

List of Figures **xii**

1 Introduction **1**

1.1 Overview 1

1.2 Outline 4

2 Spatial Models **6**

2.1 Spatial processes 6

2.1.1 Stationarity 7

2.1.2 Covariance/correlation functions and variograms 10

2.1.3 Isotropy 12

2.1.4 The spectral method 14

2.2 Modeling spatial data 15

2.2.1 Differences between spatial data and time-series data 15

2.2.2 Models for point-referenced data 16

2.2.3 Generalized linear spatial models 18

2.2.4 Spatio-temporal models 18

2.2.5 Prediction 19

2.3 Bayesian methods for spatial data analysis 21

2.3.1 Bayesian paradigm and computation 21

2.3.2 Bayesian spatial models 24

3 Spatial Asymptotics **26**

3.1 Introduction 26

3.2 The information matrix 30

3.3 Equivalent measures 33

3.4 Only the location parameter is unknown 36

3.4.1 Estimating μ: the sample average and BLUE 36

3.4.2 Bayesian beware 43

3.4.3 Equivalent measures and estimation of μ 43

3.4.4 Properties of $A_n(\boldsymbol{\phi})$ 44

3.4.5 On computing $\lim_{n\to\infty} \mathbf{1}^T R_n^{-1} \mathbf{1}$. 45

3.4.6 Extension to $\mu(s) = X^T(s)\boldsymbol{\beta}$ 52

3.4.7 Under expansion and middle-ground asymptotics 53

3.5 Unknown center and scale 53

3.5.1 Information and MLEs 53

3.5.2 Bayesian caveat 54

3.6 Known μ, σ: estimating $\boldsymbol{\phi}$ only 55

3.6.1 Information calculations 56

3.6.2 Asymptotic normality? 57

3.6.3 Decay parameters versus smoothness parameters 58

3.7 All parameters unknown 59

3.7.1 Basic information calculations 59

3.7.2 Information for the OU model 61

3.7.3 Beyond OU 64

3.8 Middle-ground asymptotics 68
3.8.1 Information formula 69
3.8.2 Infill, expansion, and middle ground 70
3.9 Including a nugget 77
3.10 Prediction 79
3.11 Discussion 80

4 Analysis for Large Spatial Datasets **81**

4.1 Introduction 81
4.2 Theoretical preliminaries 87
4.2.1 Representations for stationary process 88
4.2.2 Kernel mixing in detail 90
4.3 Finite sum approximation 94
4.3.1 Kernel mixing process approximation 94
4.3.2 Choice of D_r and m 97
4.3.3 Projection process approximation 101
4.3.4 Karhunen-Loève approximation 105
4.4 Formalizing the analysis in a Bayesian framework 106
4.5 Examples 107
4.5.1 Simulation example 108
4.5.2 Baton Rouge housing data example 110
4.6 Extensions 113

5 Spatial Design **114**

5.1 Introduction 114
5.2 Background and literature review 116

5.3 An overview of the issues 119
5.4 The information criterion 121
5.5 Approaches for approximately optimal design 123
5.6 Information gain and comparison to the entropy criterion 125
5.6.1 Calculation of the information gain 125
5.6.2 Comparison of the information criterion and the entropy criterion 128
5.7 Some remarks 130
5.8 Modified utility spatial designs 134
5.9 Computational issues and a simulation illustration 135
5.10 A prospective real application 139
5.11 Extensions 144

A Details in Chapter 3 **146**

B Details in Chapter 4 **153**

References **156**

Biography **162**

List of Tables

3.1 Simulation results for the middle-ground exponential model. 73

4.1 Posterior summaries for the simulated data example. 109

4.2 Posterior summaries for Baton Rouge housing data. 112

List of Figures

3.1 BLUE weights for a Gaussian process on $[0, 1]$ with the correlation function $\exp(-10|s-t|^2)$, $n = 30$. 40

3.2 BLUE weights $\mathrm{d}A(t)$ for a Gaussian process on $[0, 1]$ with Matérn covariance function $(n = 50, \sigma^2 = 1, \phi = 1, \nu = 0.1, 0.5, 1.5, 3)$. 42

3.3 First five eigenfunctions for the exponential correlation function on $[-1, 1]$, $\varrho(s, t) = \exp(-2|s-t|)$. 49

3.4 Solutions for (3.32) and (3.33) 50

3.5 Karhunen-Loève expansion approximation for the exponential correlation function on $[-1, 1]$: $\varrho(s, t) = \exp(-2|s-t|)$. 51

3.6 Information for power exponential covariance model: $\sigma^2 \exp(-|\phi h|^{0.5})$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 10)$. 65

3.7 Information for Matérn covariance model when $\nu = 1.5$: $\sigma^2(1+\phi|h|)\exp(-\phi|h|)$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 8)$. 65

3.8 Information for Gaussian covariance model: $\sigma^2 \exp\{-(\phi h)^2\}$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 100)$. 66

3.9 Information for Cauchy covariance function: $\sigma^2/\{1 + (\phi h)^2\}$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 50)$. 67

3.10 MLEs for the exponential covariance model ($C(h) = \sigma^2 \exp(-\phi|h|)$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 5)$), with n=100. We fit a linear model of $\log(\hat{\sigma}^2)$ on $\log(\hat{\phi})$ and obtain the curve $\sigma^2\phi^{0.96} = 6.95$. . . . 68

3.11 MLEs for the Matérn covariance model with ν=1.5 ($C(h) = \sigma^2(1 + \phi|h|)\exp(-\phi|h|)$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 8)$), n=100. We fit a linear model of $\log(\hat{\sigma}^2)$ on $\log(\hat{\phi})$ and plot the curve $\sigma^2\phi^{2.76} = 460.14$. 69

3.12 MLEs of (σ, ϕ) from simulation runs. The true exponential model parameters $(\sigma, \phi) = (1, 1)$ with $L_n = \sqrt{n}$ and $\Delta = 1/\sqrt{n}$ and $n = 100, 500, 1000, 4000$. 73

4.1 The effect of choice of D_r. 99

4.2 Performance of approximating Gaussian covariance function $C(h) = \sigma^2 \exp(-\|h\|^2/\phi)$ (σ^2=1, ϕ=15) based on kernel mixing. 100

4.3 Approximate Matérn covariance function $C(h) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)}(\phi\|h\|)^\nu \kappa_\nu(\phi\|h\|)$ (σ^2=1, ϕ=2/3, ν=2) using projection process approximation. 104

4.4 10050 data points in a $[0, 10] \times [0, 10]$ square with 100 grid points(∘). 108

4.5 Prediction at 50 locations. 111

4.6 Baton Rouge house locations (∘ denotes the grid points). 112

5.1 Density of $\mathbf{1}^T R^{-1}\mathbf{1}$ given $s_1, ..., s_4 \sim \text{unif}(0, 1)$ 131

5.2 Sequential selection vs. block selection 132

5.3 Information ($I(\beta_0)$) growth in sample size for the four sampling schemes 133

5.4 Information ($I(\phi)$) growth in sample size for the four sampling schemes 136

5.5 Study region 137

5.6 Mean surface 138

5.7 Sequential design based on $I(\beta_0)$ 139

5.8 Sequential design based on $I(\beta_0, \beta_1, \beta_2)$ 140

5.9 Sequential design based on $I(\phi)$ 141

5.10 Sequential design based on $Y * I(\beta_0)$ 142

5.11 Sequential design based on $Y * I(\beta_0, \beta_1, \beta_2)$ 143

Chapter 1

Introduction

1.1 Overview

Data in areas such as environmental health, ecology, meteorology, and real estate markets often have a geographical and temporal label associated with them. Usually, data that are close together in space (and time) are more alike than those that are far apart.

There are three basic types of spatial data: *point-referenced data*, *areal data*, and *point-process data*. Point-referenced data are also known as *geocoded* or *geostatistical* data where data are observed at a collection of locations in a set D in $\mathbb{R}^d$ and d is the number of dimensions. Typically, the locations are represented in two or three spatial coordinates, e.g. longitude, latitude, and altitude. The Baton Rouge house price dataset which we will analyze in Chapter 4 provides an example of this case. The study region D is the city of Baton Rouge, LA. We observe house selling prices and house characteristic variables (e.g. age of the house, square feet of living area, and number of bathrooms) at a set of locations. In practice, we can treat a house's location as a point since the area of that house usually is small relative to the study region D. Conceptually, observations can be taken at every location in D so we

envision uncountably many data points. We can imagine a random surface over D and view the observations as a realization from a *spatial process*. We denote this spatially varying quantity (spatial process) by $Y(s)$, where s indexes location. Often, the goals are to make statistical inference about $Y(s)$ and predict Y at new locations based on the current data.

For areal data, the study region D is again a fixed subset in $\mathbb{R}^d$, but now partitioned into a finite number of areal units with well-defined boundaries. For example, in an environmental health investigation, for purpose of confidentiality, counts of some adverse health outcome (e.g. lung cancer) are aggregated by county in a particular state. And environmental risk factors are supplied for these areal units to explain the counts.

Point-process data describe the locations of "interesting" events. Examples of such data include locations of trees, bird nests, or cancer cases. A spatial point process is a collection of random points, where each point indicates the location of an event (e.g. the occurrences of the earthquakes and the incidence of a disease). A point process $N(\cdot)$ is defined as a random measure on $D \subset \mathbb{R}^d$, taking non-negative integer values. So $N(A)$ means the number of points falling in the set $A \subset D$.

The main focus of this thesis is on the analysis of point-referenced spatial data. In general, it is very challenging to study spatially dependent data. It is considerably harder to derive large sample properties of estimators associated with spatial models. Chapter 3 studies three types of spatial asymptotics: 1) *expansion* asymptotics; 2) *infill* asymptotics; and 3) so called *"middle-ground"* asymptotics. The first two asymptotics are well known but not fully studied. Under the expansion asymptotic setting, the study region grows as the number of the observations increases such that the distance between the neighboring observations remains roughly the same. In some sense, expansion asymptotics is the *higher dimensional version* of traditional

time-series asymptotics. Infill asymptotics is based on sampling increasingly dense observations in a fixed bounded region. Infill asymptotics is most relevant to spatial data analysis because the study region usually is determined in advance. Since the distance between the neighboring observations decreases to zero as sample size increases, the large sample behavior of estimators can become very unusual as we will see in Chapter 3. We propose the term "middle-ground" to refer to a slowly expanding region with a certain rate of infill. Middle-ground asymptotics can provide insights for some practical applications, like spatial design. Again, the asymptotic behavior of estimates can be rather unusual under a middle-ground sampling scheme. We study the limiting behavior of the Fisher information matrix, the asymptotic properties of various estimators, and the weak identifiability of the parameters in spatial models in these three different asymptotics scenarios. Noting that the information contained in the likelihood under infill is limited even when the sample size tends to infinity, this suggests that the effects of prior will not be washed away by the data.

Historically, it has been difficult to apply spatial modeling techniques to analyze large spatial datasets. The problem is that we have to handle the inversion and determinant computation of a variance/covariance matrix with the size same as the sample size. Consider fitting a Gaussian spatial model for a spatial dataset with a large sample size n. Likelihood based or Bayesian modeling suffer from severe computational difficulties since each evaluation of the exact likelihood requires an $O(n^3)$ operation. We refer to this computational challenge as the "large n problem." In Chapter 4, we first review a number of existing methods to handle this "large n problem", then develop a new process approximation model which is both theoretically attractive and computationally efficient. The model is implemented in a Bayesian framework and applied to analyze several large datasets.

In Chapter 5, we consider the problem of approximately optimal design in the

special case of one-time sampling at a large number of spatial locations. For example, how shall we sample individuals within a region to measure contaminant levels in the blood? Or, how shall we sample locations to learn about ambient levels of air toxics or perhaps arsenic levels in the water table? Our goal is to develop a good design strategy to help practitioners select sampling locations. If we plan to use spatial processes in building models to analyze the data, it seems equally appropriate to use such models in developing the sampling design. Our design enables us to learn about a response model as well as the spatial distribution of the response across the study region. The criteria we consider are developed from the Fisher information matrix with the goal of learning about not only the regression structure in the model but also the dependence structure. Under a criteria that attempts to maximize information gain, we consider three strategies to develop an approximately optimal design: sequential sampling, block sampling, and stochastic search sampling.

1.2 Outline

The outline of the thesis is as follows: Chapter 2 reviews the basic elements in spatial and spatio-temporal modeling. Important concepts such as *stationary*, *isotropic*, *variograms*, *covariance/correlation function*, and *kriging* are introduced. Spatial and spatio-temoral models are briefly reviewed. We also discuss why and how we use Bayesian methods to fit spatial models.

Chapter 3 addresses important questions in spatial asymptotics and presents a variety of new results. Section 3.2 derives the Fisher information matrix for spatial model parameters. Section 3.3 introduces equivalent measure ideas and discusses why they are useful for spatial asymptotics. Section 3.4–3.10 provide extensive discussion and various results under different model assumptions and asymptotics scenarios.

Chapter 4 offers a general process approximation approach to handle the "large n

problem" in spatial data analysis. Section 4.2 introduces the theoretical preliminaries for developing the approximation approach. Section 4.3 considers three methods of approximating the spatial process with an emphasis on a finite sum approximation based on kernel mixing. Section 4.4 implements the approximation model in a Bayesian framework. A simulated data example and a Baton Rouge house price data example are given to illustrate our method in Section 4.5. Section 4.6 summarizes the process approximation ideas and discusses the extension to generalized spatial models and spatio-temproal models.

In Chapter 5, we consider the optimal spatial design problem and suggest several design approaches for practitioners. Section 5.2 provides background and literature review of spatial design. Section 5.4 develops design criterion based on the information matrix. Section 5.5 offers three design methods: block design, sequential design, and stochastic search design. Section 5.6 takes up information gain and connection to the entropy criterion. Section 5.7 considers comparison among the proposed sampling approaches. Section 5.9 illustrates our methods by a simulation study.

Chapter 2

Spatial Models

2.1 Spatial processes

One of the fundamental elements in specifying spatial models is the spatial process. We now review a few important basics of spatial processes.

The quantity we are studying varies over space, so we denote it as $W(s)$, where s indexes location and $s \in D \subset \mathbb{R}^d$. For each s, $W(s)$ is a *random variable*. The collection, or family of $W(s)$, when s varies over all its possible values, is called a *spatial process* or *random field*. In essence, $W(s)$ is just a *random function* indexed by the symbol s which belongs to some index set D. Some authors prefer to use the term *stochastic process* or *random process* when $d = 1$, and *random field* when $d \geq 2$. In this thesis, these terms are used interchangeably.

For each s, $W(s)$ is simply a random variable and its properties (e.g. mean and variance) can be described by its *distribution function*. However, we are interested in studying the whole collection of random variables $\{W(s)\}$. So more generally, we wish to specify the *joint distribution* for, say $\{W(s_1), W(s_2), W(s_3)\}$ at three distinct locations s_1, s_2 and s_3. This specification seems still not satisfying, since we want to know the probability structure for the complete process $W(s)$ which con-

sists of *uncountable* many random variables. It seems that we have to consider an *uncountably* infinite-dimensional joint distribution of $W(s)$. Fortunately, thanks to the Kolmogorov consistency theorem (see e.g. Billingsley 1995, Section 36), under fairly general conditions, the probability structure of $W(s)$ is fully specified if the joint distribution of $\{W(s_1), W(s_2), ..., W(s_n)\}$ is given for arbitrary choice of n and $s_1, ..., s_n$. So, we only need to focus our attention on the *finite*-dimensional distribution of $\{W(s_1), W(s_2), ..., W(s_n)\}$ at a finite number of locations $s_1, ..., s_n$.

2.1.1 Stationarity

We are interested in making inference about the probability structure of the spatial process based on what we observe (often just a single realization of the process). The class of all random fields is too large to obtain useful information. A common simplifying assumption is called *stationarity*. Loosely speaking, the statistical structure of a stationary process *looks similar at different parts of study region* D (or *does not change over the space*). More precisely, suppose $D = \mathbb{R}^d$,

Definition 1. *A process* $\{W(s)\}$ *is said strictly stationary if for all* $s_1, ..., s_n$ *and any* $h \in \mathbb{R}^d$*, the joint distribution of* $\{W(s_1), ..., W(s_n)\}$ *is identical with the joint distribution of* $\{W(s_1 + h), ..., W(s_n + h)\}$*, i.e.,*

$$Pr(W(s_1) \leq w_1, ..., W(s_n) \leq w_n) = Pr(W(s_1 + h) \leq w_1, ..., W(s_n + h) \leq w_n), \quad (2.1)$$

where $w_1, ..., w_n \in \mathbb{R}$.

Note that, a process $W(s)$ on $D \in \mathbb{R}^d$ can be viewed as the restriction of $W(s)$ (defined on $\mathbb{R}^d$) on D. By definition, we can say that the probability law of a strictly stationary process is *invariant* under a *shift* in space.

A weaker type of stationarity is defined in terms of the moments of W.

Definition 2. *A process $W(s)$ is stationary up to order m if*

$$E\{W(s_1)^{m_1}W(s_2)^{m_2}...W(s_n)^{m_n}\} = E\{W(s_1+h)^{m_1}W(s_2+h)^{m_2}...W(s_n+h)^{m_n}\} \quad (2.2)$$

for all $h \in \mathbb{R}^d$ and all possible positive integer $m_1, ..., m_n$ such that $m_1+m_2+...+m_n \leq m$.

The assumption of stationarity up to order 2 is most common in practice. In this case,

$$\mathrm{E}\{W(s)\} = \mathrm{E}\{W(0)\} = \mu, \text{a constant independent of } s \quad (2.3)$$

and

$$\begin{aligned} \mathrm{Cov}\{W(s_1), W(s_2)\} &= \mathrm{E}\{W(s_1)W(s_2)\} - \mu^2 \\ &= \mathrm{E}\{W(0)W(s_2 - s_1)\} \\ &= C(s_2 - s_1), \text{ a function of } (s_2 - s_1) \text{ only.} \end{aligned} \quad (2.4)$$

$C(\cdot)$ is called *covariance* or *autocovariance* function of W.

For a weakly stationary process W (from now on, we always mean a weakly stationary process up to oder 2 when we say a weakly stationary process), the correlation between $W(s_1)$ and $W(s_2)$ is also a function of $s_1 - s_2$, defined as

$$\varrho(s_1 - s_2) = C(s_1 - s_2)/C(0). \quad (2.5)$$

$\varrho(\cdot)$ is called *correlation* or *autocorrelation* function.

Note that although the conditions of strict stationarity seem much stronger, they may not always imply weak stationarity. For example, a strictly stationary process W whose finite joint distribution is Cauchy does not have any moments. However, if we assume second moments exist, strict stationarity *does* imply weak stationarity.

Definition 3. *$W(s)$ is called Gaussian process if for all n and admissible $s_1, ..., s_n$, the joint distribution of $\{W(s_1), ..., W(s_n)\}$ is multivariate normal.*

A multivariate normal distribution is characterized by its mean and covariance matrix, so the first two moments of a Gaussian process completely specify its probability structure. Thus for Gaussian processes, weak stationarity implies strict stationarity.

In terms of what kind of assumptions about the spatial process are appropriate for spatial data analysis, it is tempting to make less restrictive assumption, say, weak stationarity. But this generality makes it impossible for making likelihood-based inference since we don't have the joint distribution specification. Furthermore, Stein (1999, p.6–7) gives an example showing that best linear prediction can be very poor by only considering the first two moments of the process.

As the name indicates, Gaussian processes play a central role in modeling spatial data. The advantages of the Gaussian process assumption are obvious: it allows convenient distribution theory (for instance, conditional distributions are easily obtained from the joint distributions); the Gaussian process is well studied and many classical theory and results for Gaussian process are available (several equivalent-Gaussian measure results will be used in Chapter 3). Furthermore, in most applications, we just observe a single replication (realization) of the process at a finite set of locations. It is not easy to criticize a Gaussian assumption since we only have a sample size of *one* from a finite dimensional distribution. Nevertheless, there are situations in which it is more appropriate to use other processes to model spatial data. Wolpert and his colleagues use Lévy processes to model spatial data in a series of their papers (see e.g. Wolpert and Ickstadt, 1998; Ickstadt and Wolpert, 1999).

2.1.2 Covariance/correlation functions and variograms

For a spatial process $W(s)$ with finite second moments, its associated covariance matrix $K(s_i, s_j)$ must be *positive semi-definite*, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(s_i, s_j) \geq 0, \tag{2.6}$$

for any set of $s_1, ..., s_n$ and all real $a_1, ..., a_n$. This simply follows by noting

$$\mathrm{Var}\{\sum_{i=1}^{n} a_i W(s_i)\} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(s_i, s_j).$$

The positive semi-definite condition is necessary for the existence of a random field with finite second moments. On the other hand, if K is positive semi-definite, there *exists* a Gaussian random field with covariance matrix K and mean $\mathrm{E}\{W(s)\}$ (since the finite-dimentional distribution satisfies the Kolmogorov's consistency conditions).

If W is a stationary process, $K(s_i, s_j)$ depends only on $s_i - s_j$ (see (2.4)). So we can use covariance function $C(s_i - s_j)$ to describe the covariance structure of W. As indicated by condition (2.6), not every function can serve as a valid covariance function. Bochner's theorem provides necessary and sufficient conditions for $C(\cdot)$ to be positive semi-definite.

Theorem 1 (Bochner's Theorem)**.** *For a real-valued process on $\mathbb{R}^d$, $C(h)$ is positive semi-definite if and only if it can be represented as*

$$C(h) = \int_{\mathbb{R}^d} e^{i\omega^T h} F(d\omega), \tag{2.7}$$

where F is a positive, symmetric, and finite measure and is called the spectral measure of $C(h)$. If $F(d\omega)$ is absolutely continuous with respect to Lebesgue measure, i.e., $F(d\omega) = f(\omega)d\omega$, $f(\omega)$ is called the spectral density.

In the *geostatistical* literature (e.g. Cressie 1993), researchers commonly use variograms to describe the spatial correlation of the process $Y(s)$. The variogram assumes that the variance of $Y(s_i) - Y(s_j)$ only depends on $s_i - s_j$ and is defined as

$$2\gamma(h) = \mathrm{Var}\{Y(s+h) - Y(s)\}. \tag{2.8}$$

In the above, $\gamma(h)$ is called *semivariogram*.

Provided the covariance function exists, the relationship between the variogram and the covariance function is obvious:

$$\begin{aligned} 2\gamma(h) &= \mathrm{Var}\{Y(s+h)\} + \mathrm{Var}\{Y(s)\} - 2\mathrm{Cov}\{Y(s+h), Y(s)\} \\ &= 2\{C(0) - C(h)\}. \end{aligned} \tag{2.9}$$

A valid variogram needs to satisfy a negative definiteness condition. Specifically, for any set of $\{s_1, ..., s_n\}$ and any set of real $\{a_1, ..., a_n\}$ such that $\sum_{i=1}^n a_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(s_i - s_j) \leq 0. \tag{2.10}$$

This follows by noting

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j \gamma(s_i - s_j) = -\mathrm{E}\{\sum_{i=1}^n a_i Y(s_i)\}^2 \leq 0.$$

Variograms only describe the first two moments (not the probability law) of the spatial process so it is not possible to make likelihood-based inference on its basis. One of the reasons that geostatisticians prefer the variogram is that it can be conveniently estimated by

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{Y(s_i + h) - Y(s_i)\}^2, \tag{2.11}$$

where $\hat{\gamma}(h)$ is called the *empirical variogram* at distance h and $N(h)$ is the number of pairs of observations whose distance is equal to h. Variogram binning is ignored in this formula.

2.1.3 Isotropy

A stationary random field is said *isotropic* if its covariance function $C(h)$ only depends on $\|h\|$, where $\|\cdot\|$ indicates the Euclidean distance. The isotropy property can be thought as an invariance property under *rotations*. A stronger type of isotropy, called *strict isotropy*, can be defined in terms of the invariance of the probability structure under rigid motions analogous to the definition of strict stationarity (see Stein, 1999, p.17).

The class of all valid continuous covariance functions on $\mathbb{R}^d$ can be characterized by the Fourier transforms of all finite positive measures on $\mathbb{R}^d$ (see Theorem 1). There is an analogous characterization for isotropic covariance functions (see Yaglom 1987, Section 22). Specifically,

Theorem 2. *For $d \geq 2$, a function C is a continuous isotropic covariance function for a random field on $\mathbb{R}^d$ if and only if it can be represented as*

$$C(h) = 2^{(d-2)/2}\Gamma(d/2)\int_0^\infty (\omega\|h\|)^{-(d-2)/2} J_{(d-2)/2}(\omega\|h\|)G(d\omega), \tag{2.12}$$

where J_ν is the Bessel function of the first kind of order ν (Aabramowitz and Stegun 1965, Section 9); G is nondecreasing, bounded on $\mathbb{R}^+$ and $G(0) = 0$.

An isotropic covariance function C must be a real function since $C(\|h\|) = C(\|-h\|) = \overline{C(\|h\|)}$. For $d = 1$, the class of valid isotropic covariance functions is the same as the class of all real covariance functions for stationary processes.

Let C_d be the class of all valid continuous isotropic covariance functions in $\mathbb{R}^d$ and $C_\infty = \bigcap_{d=1}^\infty C_d$, the class of valid continuous isotropic covariance functions in all

dimensions, then $C_1 \supset C_2 \supset ... \supset C_\infty$. The intuition behind $C_{d_1} \supset C_{d_2}$ when $d_1 < d_2$ is that starting from higher dimension d_2, we can restrict the consideration to d_1 coordinates and set the values for the remaining coordinates to zero. A function is in C_∞ if and only if it can be represented as

$$C(h) = \int_0^\infty e^{-\|h\|^2\omega^2} G(\mathrm{d}\omega), \tag{2.13}$$

where G is as in (2.12). The Gaussian covariance function ($C(h) = \sigma^2 \exp(-\phi\|h\|^2)$) is an example of a covariance function valid in all dimensions.

Following is a list of popular parametric isotropic covariance functions:

1. Exponential covariance function:

$$C(h; \sigma^2, \phi) = \sigma^2 e^{-\phi\|h\|}, \tag{2.14}$$

   where σ^2 is the variance parameter and ϕ is called *decay* parameter which tells us how quickly the correlation decays as the distance $\|h\|$ increases. The decay parameter is related with a notion of *effective range* which is often used geostatistics. Effective range is the distance at which there is essentially no lingering spatial correlation. In practice, it is commonly defined as the distance at which the correlation drops to only 0.05. In the exponential correlation function case, the effective range h_0 equals $-\log(0.05)/\phi$.

2. Powered exponential covariance function:

$$C(h; \sigma^2, \phi, \alpha) = \sigma^2 e^{-\phi\|h\|^\alpha}, 0 < \alpha \leq 2. \tag{2.15}$$

3. Gaussian covariance function:

$$C(h; \sigma^2, \phi) = \sigma^2 e^{-\phi\|h\|^2}. \tag{2.16}$$

4. Wave covariance function:

$$C(h;\sigma^2,\phi) = \sigma^2 \sin(\phi\|h\|)/(\phi\|h\|). \tag{2.17}$$

5. Matérn covariance function:

$$C(h;\phi,\alpha,\nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\phi\|h\|)^\nu \kappa_\nu(\phi\|h\|), \tag{2.18}$$

where ϕ is the correlation decay parameter; ν is the smoothness parameter (the larger ν is, the smoother the corresponding process is); and κ_ν is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun 1965, Section 9.6). Note the form (2.18) does not depend on the dimensionality d. A slightly different parameterization of the Matérn class recommended by Handcock and Wallis (1994) is

$$C(h;\phi^*,\alpha,\nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\nu^{1/2}\phi^*\|h\|)^\nu \kappa_\nu(2\nu^{1/2}\phi^*\|h\|), \tag{2.19}$$

The advantage of using $2\nu^{1/2}\phi^*$ rather than ϕ is that the value of ν has minimal effects on the interpretation of the decay parameter ϕ^* (see Stein 1999, p.50).

There is another alternative parameterization of Matérn class in $\mathbb{R}^d$:

$$C(h;\zeta,\phi,\nu) = \frac{\pi^{d/2}\zeta}{2^{\nu-1}\Gamma(\nu+d/2)\phi^{2\nu}}(\phi\|h\|)^\nu \kappa_\nu(\phi\|h\|). \tag{2.20}$$

The process variance is given by $\sigma^2 = \frac{\pi^{d/2}\Gamma(\nu)\zeta}{\Gamma(\nu+d/2)\phi^{2\nu}}$. Note that the exponential covariance function and the Gaussian covariance function are two special cases of the Matérn class with $\nu = 1/2$ and $\nu = \infty$, respectively.

2.1.4 The spectral method

The spectral method is a powerful tool for studying random processes. Consider a random process $W(s)$ on $\mathbb{R}^d$. Traditional Fourier analysis says that each realization

of $W(s)$ can be expressed as a Fourier transform of the form

$$W(s) = \int_{\mathbb{R}^d} e^{i\omega^T s} Z(\mathrm{d}\omega), \tag{2.21}$$

where $Z(d\omega)$ is a set function called the *Fourier transform* of this particular realization of $W(s)$. Of course, different realizations of $W(s)$ will result in different in $Z(d\omega)$. Thus $Z(d\omega)$ is itself random, called the *random measure* corresponding to $W(s)$. $Z(\cdot)$ has the following properties:

- $\mathrm{E}\{Z(\Delta)\} = 0$ for all measurable set Δ.
- $\mathrm{E}\{Z(\Delta_1)\overline{Z(\Delta_2)}\} = 0$ for disjoint measurable sets Δ_1 and Δ_2.
- $Z(\Delta_1 \cup \Delta_2) = Z(\Delta_1) + Z(\Delta_2)$ if measurable sets Δ_1 and Δ_2 are disjoint.

Suppose $\mathrm{E}|Z(\Delta)|^2 = F(\Delta)$ for some positive finite measure F, then the covariance function associated with $W(s)$ can be expressed as

$$C(h) = \int_{\mathbb{R}^d} e^{i\omega^T h} F(\mathrm{d}\omega). \tag{2.22}$$

The function F is called the spectral measure of W.

2.2 Modeling spatial data

Having reviewed the fundamental elements of spatial processes, we now consider basic statistical models, building from spatial processes for spatial data. We start from a brief discussion comparing spatial data with time-series data.

2.2.1 Differences between spatial data and time-series data

Time-series data and spatial data share many similarities. Both settings assume dependent data and much of the theory for time series can be generalized to spatial

data analysis. Nevertheless, there are several important differences between these two types of data.

- Time series naturally have the distinction of past and future and the present observation depends only on the past observations. Usually it is not possible to *go back* from now to obtain more samples because of this time "direction". In contrast, the dependence extends to all directions for spatial data and in principle it is possible to obtain more samples at any locations and in any direction.

- The domain for time series usually is unbounded as we collect more and more observations. In the case of spatial data, the study region is generally fixed and bounded *a priori* in practice. The asymptotic analysis based on the increasing dense samples in a fixed region is called infill asymptotics which is one of the subjects studied in Chapter 3.

2.2.2 Models for point-referenced data

Suppose that data are observed at a collection of sites $\{s_1, s_2, ..., s_n\}$, where $s_i \in D$. A common spatial process model is constructed as follows:

$$Y(s_i) = \mu(s_i) + W(s_i) + \epsilon(s_i), \tag{2.23}$$

where $\mu(s_i)$ is the mean of the response $Y(s_i)$, typically of the form $X^T(s_i)\boldsymbol{\beta}$ ($X(s_i)$ is a p-dimensional vector of explanatory variables at location s_i and $\boldsymbol{\beta}$ is a p-dimensional vector of parameters); $W(s)$ is a mean zero spatial process ($W(s)$ is often assumed to be a stationary Gaussian process with a parametric covariance function $\sigma^2\varrho(s_i - s_j; \boldsymbol{\phi})$); and $\epsilon(s)$ is a pure error process with mean 0 and variance τ^2. In spatial literature, the value of $\sigma^2 + \tau^2$ is called *sill*; the value of σ^2 is called *partial sill*; and the value of τ^2 is called *nugget*.

For each s, $W(s)$ is a random variable and it can be viewed as the *spatial random effect* at location s. $W(s)$ essentially introduces the spatial association into the model and captures spatial variation at site s. To understand the spatial random effect, consider the Baton Rouge house price example. Suppose $Y(s)$ is the (log) selling price of a house at location s and $X(s)$ is a vector of characteristics of the house (e.g. age of the house, square feet of living area, number of bedrooms, etc.) It is well known that the price of a particular house depends not only on its inherent characteristics, but also its location. For instance, if a house is close to a business center or a school, or in a good community, the price of that house is expected to be high. Often, underlying variables with spatial explanation are unobserved, so the spatial random effect term $W(s)$ is introduced to account for these latent effects. Intuitively, $W(s)$ will be high if the location s is "good". Of course, if we observe location related information (e.g. distance from a business center), we can treat them as covariates and put them into the mean structure of the model. However, there will always be some latent or unmeasurable spatial effects in the data. So, with a parsimonious mean, we can model the residual structure more accurately by including a spatial random effect in the model.

$\epsilon(s)$ can be viewed as "noise" associated with replication of measurements at location s. In the housing data example, $\epsilon(s)$ may be caused by a particular buyer and a particular seller. If Y comes from certain measuring devices, $\epsilon(s)$ can be regarded as the measurement error.

Note that we can view model (2.23) as a hierarchical model with a conditionally independent first stage given $W(s)$ and $\mu(s)$. In the second stage, usually we assume $W(s)$ to be a Gaussian random field with mean zero and certain parametric covariance structure.

2.2.3 Generalized linear spatial models

In some situations, the response variable $Y(s)$ (even after transformation) is not appropriate to be treated as a normal random variable. For instance, $Y(s)$ might be a binary variable or a count variable. It is natural to consider an extension of the model (2.23) analogous to the generalized linear model. Following Diggle, Tawn, and Moyeed (1998) , assuming $Y(s_i)$ are conditionally independent given $\boldsymbol{\beta}$ and $W(s_i)$, we formulate a hierarchical model as follows:

$$\textbf{First stage:} \quad f(y(s_i)|\boldsymbol{\beta}, W(s_i),) = h(y(s_i), \gamma) \exp\{\gamma[y(s_i)\eta(s_i) - \psi(\eta(s_i))]\}, \quad (2.24)$$

$$\textbf{Second stage:} \quad W(s_i) \sim \text{Gaussian process}(0, C), \quad (2.25)$$

where $g(\eta(s_i)) = X^T(s_i)\boldsymbol{\beta} + W(s_i)$ for some *link function* g; γ is a *dispersion* parameter; $\psi(\cdot)$ is a known function; and $W(s)$ is a Gaussian process with mean zero and certain covariance structure C as in (2.23). The second stage specifies the spatial structure.

2.2.4 Spatio-temporal models

Recently, there is considerable literature in spatio-temporal modeling. Spatio-temporal data can be represented by $Y(s,t)$, where s indexes location and t indexes time. Examples of such data include pollutant data (monitoring stations measure the pollutant level hourly or daily) and house transaction data (sales can happen at any time during study period). We can model $Y(s,t)$ analogously to the model (2.23):

$$Y(s,t) = \mu(s,t) + W(s,t) + \epsilon(s,t), \quad (2.26)$$

where $\mu(s,t)$ is the mean which may take the form $X^T(s,t)\boldsymbol{\beta}$, and $W(s,t)$ is a mean zero spatial-temporal process.

For the spatio-temporal data, there is both temporal and spatial association. The challenge is to specify an appropriate space-time interaction for the process

$W(s,t)$. Viewing t as having continuous support, there is a substantial literature which discusses the specification of spatio-temporal covariance structure of $W(s,t)$. One of the popular covariance functions is called *stationary separable* and has the following form:

$$\text{Cov}(Y(s,t), Y(s',t')) = \sigma^2 \varrho_1(s-s')\varrho_2(t-t'), \tag{2.27}$$

where ϱ_1 is a valid two-dimensional correlation function and ϱ_2 is a valid one-dimensional correlation function.

2.2.5 Prediction

The basic problem of spatial prediction is as follows: suppose we observe a random field Y on $\mathbb{R}^d$ at a set of locations $\{s_1, ..., s_n\}$, how shall we predict Y at a new location s_0? More generally, how shall we predict Y at more than one new locations?

If the probability law of $Y(s)$ is known, then it is natural to do prediction based on the conditional distribution of $Y(s_0)$ given $Y(s_1), ..., Y(s_n)$. If we only know the mean and covariance structure of $Y(s)$ up to some parameters, then linear prediction is commonly used. The best linear prediction approach is known as *kriging* in geostatistics, named by Matheron (1963) in honor of the South African mining engineer D.G. Krige (Krige, 1951).

Suppose $Y(s)$ has a mean function $\mu(s)$ and a covariance function $C(s_i, s_j)$. The linear prediction takes the form $\hat{Y}(s_0) = a_0 + \sum_{i=1}^n a_i Y(s_i)$ and we wish to minimize the mean squared error of $\hat{Y}(s_0)$, i.e. $\text{E}\{Y(s_0) - \hat{Y}(s_0)\}^2$. It is not hard to find the form of best a_0 and $\mathbf{a} = (a_1, ..., a_n)$:

$$a_0 = K^{-1}\mathbf{h} \quad \text{and} \quad \mathbf{a} = \mu(s_0) - \mathbf{h}^T K^{-1}\boldsymbol{\mu}, \tag{2.28}$$

where K is the covariance matrix of $\mathbf{Y} = (Y(s_1), ..., Y(s_n))$, $\mathbf{h} = \text{Cov}(Y(s_0), \mathbf{Y})$, and $\boldsymbol{\mu} = (\mu(s_1), ..., \mu(s_n))$. The resulting mean squared error is $C(s_0, s_0) - \mathbf{h}^T K^{-1}\mathbf{h}$.

If we assume $Y(s)$ to be a Gaussian process, then the conditional distribution of $Y(s_0)$ given $\mathbf{Y}$ is

$$N\big(\mu(s_0) + \mathbf{h}^T K^{-1}(\mathbf{Y} - \boldsymbol{\mu}), C(s_0, s_0) - \mathbf{h}^T K^{-1}\mathbf{h}\big).$$

So $\mathrm{E}(Y(s_0)|\mathbf{Y})$ is equal to the best linear estimator in this case. If we wish to predict Y at a set of new locations, say, $\{s_{0,1}, ..., s_{0,m}\}$, we simply find the conditional distribution of $\mathbf{Y}_0 = (Y(s_{0,1}), ..., Y(s_{0,m}))$ given $\mathbf{Y}$. From standard multivariate normal theory, we can easily find the conditional distribution from the joint distribution. Suppose

$$\begin{pmatrix} \mathbf{Y}_0 \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} K_{00} & K_{01} \\ K_{10} & K_{11} \end{pmatrix} \right), \tag{2.29}$$

where $\boldsymbol{\mu}_0 = (\mu(s_{0,1}), ..., \mu(s_{0,m}))$, $K_{00} = \mathrm{Cov}(\mathbf{Y}_0, \mathbf{Y}_0)$, $K_{10}^T = K_{01} = \mathrm{Cov}(\mathbf{Y}_0, \mathbf{Y})$, and $K_{1,1} = \mathrm{Cov}(\mathbf{Y}, \mathbf{Y})$. Then the conditional distribution $f(\mathbf{Y}_0|\mathbf{Y})$ is normal with mean and variance:

$$\mathrm{E}(\mathbf{Y}_0|\mathbf{Y}) = \boldsymbol{\mu}_0 + K_{01}K_{11}^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \tag{2.30}$$

$$\mathrm{Var}(\mathbf{Y}_0|\mathbf{Y}) = K_{00} - K_{01}K_{11}^{-1}K_{10}. \tag{2.31}$$

In practice, the parameters in the mean and covariance structure are unknown and must be estimated from the data. In a classical framework, usually the restricted maximum likelihood estimates (REML) are computed and plugged into the above prediction formulas.

In a Bayesian paradigm, we can obtain posterior predictive distribution of $Y(s_0)$ given $\mathbf{Y}$ as follows:

$$\begin{aligned} \pi(y_0|\mathbf{Y}) &= \int f(y_0, \boldsymbol{\theta}|\mathbf{Y})\mathrm{d}\boldsymbol{\theta} \\ &= \int f(y_0|\mathbf{Y}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y})\mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{2.32}$$

where $\boldsymbol{\theta}$ are parameters in the model. So any desired point or interval estimate may be computed based on this posterior distribution. This is a *"Bayesian kriging"* approach. In practice, Markov Chain Monte Carlo (MCMC) methods are often used to obtain estimates of (2.32). In particular, we typically use *composition sampling*, i.e., we first draw posterior samples $\boldsymbol{\theta}^{(t)}$ $(t = 1, ..., T)$ from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{Y})$, then for each $\boldsymbol{\theta}^{(t)}$, we sample a $y_0^{(t)} \sim f(y_0|\mathbf{Y}, \boldsymbol{\theta}^{(t)})$. Thus the collection $\{y_0^{(t)}\}$ is a collection of samples from the predictive distribution (2.32). See Section 2.3 for more details about Bayesian methods.

2.3 Bayesian methods for spatial data analysis

In this section we discuss why we prefer to fit spatial models in a Bayesian framework. We also review important MCMC algorithms for Bayesian computation and then formalize the Bayesian spatial model.

2.3.1 Bayesian paradigm and computation

Bayesian methods enjoy several advantages over frequentist methods for fitting models involving complicated dependence structures. Here we list a few of them: i) with the recent advances in computing technology and MCMC algorithm, fitting spatial models in a Bayesian framework is fairly standard. MCMC output enables us to make exact analysis for any sample size. The subsequent inference (e.g. estimation, hypothesis testing, and prediction) is relatively straightforward. In contrast, frequentist methods for spatial data rely heavily on the asymptotic analysis. It is not easy to study asymptotics for spatial models due to the dependence from all directions of the spatial data (see Chapter 3 for detailed discussion). ii) the Bayesian paradigm also enables us to incorporate prior information into the model in a natural way. For example, in an air pollutant study, historical pollutant information can be translated

into a prior specification and incorporated into the current analysis. iii) in many situations, classical inference can be obtained as a special case of Bayesian inference with a particular choice of prior (often *non-informative* prior). iv) Bayesian inference has advantages in the case of missing value problems (either missing response or covariances) since missing values will be modeled and treated and estimated as additional parameters.

Bayesians treat parameters $\boldsymbol{\theta}$ in the model as random variables and make inference based on the *posterior* distribution of $\boldsymbol{\theta}$ through Bayes formula

$$\pi(\boldsymbol{\theta}|\text{data}) = \frac{f(\text{data}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\text{data}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\text{d}\boldsymbol{\theta}}, \tag{2.33}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ and $f(\text{data}|\boldsymbol{\theta})$ is the likelihood function. In general, $\pi(\boldsymbol{\theta}|\text{data})$ has no closed form. In those situations, MCMC methods play important role in obtaining samples from the posterior distribution (2.33). MCMC posterior samples are correlated since they are recursive draws from a particular Markov chain. The Gibbs sampler (Gelfand and Smith, 1990) and the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) are the two most popular MCMC algorithms.

Suppose $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ is a p-dimensional vector of parameters. The basic algorithm of the Gibbs sampler proceeds as follows:

Step 1 Set $t = 0$ and choose an arbitrary starting point $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$.

Step 2 Generate $\boldsymbol{\theta}^{(t+1)}$ sequentially as follows:

- Generate $\theta_1^{(t+1)}$ from $f(\theta_1|\theta_2^{(t)}, ..., \theta_p^{(t)})$;
- Generate $\theta_2^{(t+1)}$ from $f(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_p^{(t)})$;
- ... ...

– Generate $\theta_p^{(t+1)}$ from $f(\theta_p|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{p-1}^{(t+1)})$.

Step 3 Set $t = t + 1$ and go to Step 1.

Under mild regularity conditions, the sequence $\{\boldsymbol{\theta}^{(t)}\}$ has a stationary distribution $\pi(\boldsymbol{\theta}|\text{data})$ (Gelfand and Smith, 1990). This means that, when t is sufficiently large (say bigger than t_0), $\{\boldsymbol{\theta}^{(t)}, t > t_0\}$ are samples from the true posterior distribution $\pi(\boldsymbol{\theta}|\text{data})$. So we can make inference about any function of $\boldsymbol{\theta}$ based on these posterior samples.

The Gibbs sampler is easy to implement, but requires sampling from each of the *full conditional density* which may not be available in a standard distributional form. The Metropolis-Hastings algorithm is designed to handle this problem.

Suppose we want to sample from the posterior distribution $\pi(\boldsymbol{\theta}|\text{data})$. Let $q(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ be a *candidate density* such that $\int q(\boldsymbol{\theta}, \boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta} = 1$. The basic scheme of the Metropolis-Hastings algorithm is as follows:

Step 1 Set $t = 0$ and choose an arbitrary starting point $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_p^{(0)})$.

Step 2 Generate a candidate point $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}^{(t)}, \cdot)$ and generate a number u from a uniform distribution over (0, 1).

Step 3 Compute the acceptance ratio

$$r = \min\left\{ \frac{\pi(\boldsymbol{\theta}^*|\text{data})q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(t)})}{\pi(\boldsymbol{\theta}^{(t)}|\text{data})q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^*)}, 1\right\} \tag{2.34}$$

and set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ if $u \leq r$ and $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$ otherwise.

Step 3 Set $t = t + 1$ and go to Step 1.

2.3.2 Bayesian spatial models

For the spatial model (2.23), assume $W(s)$ is a stationary Gaussian process with a parametric covariance function $C(h;\sigma^2,\boldsymbol{\phi}) = \sigma^2\varrho(h;\boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a q-dimensional vector of parameters in the correlation function. Let $\boldsymbol{\theta} = (\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi}^T,\tau^2)^T$ and $\pi(\boldsymbol{\theta})$ be a prior distribution for parameters $\boldsymbol{\theta}$. The posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Y} = (Y(s_1),...,Y(s_n))^T$ is

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{2.35}$$

$f(\mathbf{Y}|\boldsymbol{\theta})$ follows a n-dimensional multivariate normal distribution $\mathrm{N}(X\boldsymbol{\beta},\sigma^2R(\boldsymbol{\phi})+\tau^2\mathrm{I}_n)$, where $R(\boldsymbol{\phi})$ is the associated correlation matrix and I_n is a $n\times n$ identity matrix.

The prior specification of $\pi(\boldsymbol{\theta})$ is largely arbitrary but typically independent priors are chosen for the different parameters, i.e.,

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\boldsymbol{\phi})\pi(\tau^2). \tag{2.36}$$

If we want to obtain a particular marginal posterior distribution, say $\pi(\boldsymbol{\beta}|Y)$, we need to integrate $\pi(\boldsymbol{\theta}|\mathbf{Y})$ with respect to all other parameters, i.e.,

$$\pi(\boldsymbol{\beta}|\mathbf{Y}) = \int \pi(\boldsymbol{\theta}|\mathbf{Y})\mathrm{d}\sigma^2\mathrm{d}\boldsymbol{\phi}\mathrm{d}\tau^2. \tag{2.37}$$

Often, we are interested in examining the spatial surface $W(s)|\mathbf{Y}$. The posterior realizations of W can be obtained by noting

$$\pi(W|\mathbf{Y}) = \int \pi(W|\sigma^2,\boldsymbol{\phi})\pi(\sigma^2,\boldsymbol{\phi}|\mathbf{Y})\mathrm{d}\sigma^2\mathrm{d}\boldsymbol{\phi}. \tag{2.38}$$

In general, the posterior distribution $\pi(\boldsymbol{\beta}|Y)$ is not available in a closed-form so we use some suitable MCMC algorithms to draw posterior samples. In expression (2.35), W is marginalized so it is not involved in the MCMC updating. It is also

possible to obtain $\pi(\boldsymbol{\theta}|Y)$ by considering the following hierarchical model:

$$\textbf{Stage 1: } \mathbf{Y}|\boldsymbol{\theta}, \mathbf{W} \sim \mathrm{N}(X\boldsymbol{\theta} + W, \tau^2 \mathrm{I}_n),$$

$$\textbf{Stage 2: } \mathbf{W}|\sigma^2, \boldsymbol{\phi} \sim \mathrm{N}(0, \sigma^2 R(\boldsymbol{\phi})).$$

In this case, the n-dimentional vector $\mathbf{W}$ is treated as parameters and updated in each MCMC step. Generally, we would prefer to work in a low dimensional parameter space in Bayesian computation. So we shall do as much marginalization as possible. But for generalized spatial models, marginalizing over W is usually not possible.

Chapter 3

Spatial Asymptotics

3.1 Introduction

In this chapter, we shall study spatial asymptotics. Consider a stochastic process model of the following form:

$$Y(t) = \mu(t) + W(t), \tag{3.1}$$

where $t \in D$ and D is some set in $\mathbb{R}^d$; $\mu(t)$ is a mean function; and $W(t)$ is a mean zero stationary Gaussian process. Suppose $W(t)$ has a parametric covariance function, say,

$$C(s,t) = \mathrm{Cov}\{W(s), W(t)\} = \sigma^2 \varrho(s-t;\boldsymbol{\phi}), \tag{3.2}$$

where ϱ is a valid correlation function on $\mathbb{R}^d$; σ^2 is the variance of the process W; and $\boldsymbol{\phi}$ is a q-dimensional vector of parameters that determines the correlation function ϱ. In this chapter we focus primarily on the case of $d = 1$ (continuous time series) and $d = 2$ (spatial processes).

The primary issues we address here are as follows. If one observes Y at locations $t_1, \ldots, t_n$, how well can one estimate the mean function and the covariance

parameters $\sigma^2, \boldsymbol{\phi}$? In particular, we ask the usual questions, such as, how much information is there in the data about these parameters? If the number of locations (sample size) tends to infinity, does the amount of information tend to infinity? Can we consistently estimate these parameters? If some of them are assumed known, can we consistently estimate the others? What functions of these parameters can we consistently estimate? Can we develop any asymptotic distribution theory, e.g., asymptotic normality? We summarize and extend the literature but are only able to provide partial answers in special cases. Perhaps surprisingly, general answers are remarkably difficult to obtain. So, in this regard, this chapter serves as both a review as well as offering several new results.

We have to be more precise with regard to the statement that "sample size tends to infinity". Potentially, there are three possible sampling schemes:

1. The (Euclidean) distance Δ between two neighboring locations remains constant when the sample size increases. In this case, D is necessarily unbounded as the sample size tends to infinity. The asymptotic analysis under this situation is called *expansion* asymptotics.

2. The study region D is fixed and bounded so that the distance between neighboring locations decreases to zero as the sample size increases to infinity. Eventually, we obtain everywhere dense samples inside D. Cressie (1993) refers to asymptotics based on this sampling scheme as *infill* asymptotics. Stein (1999) calls it *fixed-domain* asymptotics.

3. We introduce a third sampling scheme that is a mixed version of the above two. The sample region D is slowly growing with a certain rate of infill sampling as the sample size increases. We call asymptotics based on this situation *middle-ground* asymptotics.

We are interested in studying the limiting behavior of the (Fisher) information matrix for model (3.1) under the above three asymptotics scenarios. We are also interested in large sample properties of the maximum likelihood estimators (MLEs) and Bayesian estimators under the different asymptotics.

Expansion asymptotics have been studied extensively and there are many classical results, especially for time series (d=1, $t \in \mathbb{Z}$, the integer set) analysis. Roughly speaking, under mild regularity conditions, the MLEs are consistent and asymptotical normal under expansion asymptotics. Mardia and Marshall (1984) give general results for expansion asymptotics. Stein (1999, p.62) argues that in a spatial setting, it is more natural to consider infill asymptotics than expansion asymptotics since usually the study region of interest is fixed *a priori*.

The basic difficulty in studying asymptotics for data with structured dependence is that for a parametric covariance function, except for a very few cases, the closed form expressions of the determinant and inverse of the covariance matrix are not available. So in general it is hard to study the properties of the likelihood. In the literature, usually the mean function $\mu(t)$ is assumed to be zero with attention focused on the parameters of the covariance function. Ying (1991) studies infill asymptotics in $\mathbb{R}^1$ with the exponential covariance function ($C(s,t) = \sigma^2 \exp(-\phi|s-t|)$). He shows that the variance parameter σ^2 and the decay parameter ϕ cannot be estimated consistently. Nevertheless, the MLE for $\sigma^2\phi$ is consistent and asymptotic normal with the usual $\sqrt{n}$ convergence rate. In Ying (1993) , these results are extended to $\mathbb{R}^2$ for a *separable* (*product*) exponential covariance function. Ying's results rely critically on the Markov property of the process with an exponential covariance function. More specifically, for the correlation matrix R_n (arising from n locations), $|R_n|$ and R_n^{-1} have closed form expressions under these two cases. So the likelihood functions can be written explicitly. Stein (1999) studies the consistency of optimal prediction under

infill asymptotics using the equivalent-measure idea. Zhang (2004) gives some infill asymptotic consistency results for the Matérn covariance function. Most recently, Loh (2005) is able to find the explicit forms of $|R_n|$ and R_n^{-1} for a separable Matérn subclass covariance function in $\mathbb{R}^2$ with the smoothness parameter $\nu = 1.5$.

Middle-ground asymptotics have received little attention in the literature (though see Stein 1995). Such asymptotics have potential usefulness for spatial design in which one wants to determine the size of study region as well as the number and locations of samples. In this rather untravelled territory of "middle-ground asymptotics" we can obtain interesting results. For example, in $\mathbb{R}^1$ with the exponential correlation function and sampling distance $\Delta = O(1/\sqrt{n})$, the rate of convergence of estimators is $n^{1/4}$ rather than $n^{1/2}$ (see Section 3.8).

In this chapter, we examine information growth under three above mentioned different types of asymptotics. We also study the asymptotic properties of various estimators for the mean parameters and covariance parameters. For instance, when parameters σ^2 and $\boldsymbol{\phi}$ in the covariance structure are known, i.e. with only mean parameters unknown, then in general the information for estimating mean is bounded under infill sampling, regardless of the denseness of the sampling sites. In other words, since one may estimate mean parameters only with limited precision, there are no consistent estimators. When $\boldsymbol{\phi}$ is known, but both center and scale are unknown, then information about the mean is again limited while information about σ^2 is unlimited. Thus, somewhat mysteriously, based on a densely observed path one may estimate the scale arbitrarily well (there are consistent estimators), whereas there are no consistent estimators for the center parameters. This also implies that a Bayesian needs utmost care when specifying a prior for the mean parameters and σ^2. Aspects of the prior related to the mean parameters will be remembered for ever; the data will never entirely *overwhelm* the prior. When all parameters are unknown, certain functions of

these parameters can be learned with arbitrary precision. Reparametrization using these functions will lead to better behaved likelihoods and MCMC algorithms. See, for instance, the parametrization of the Matérn covariance function in Handcock and Stein (1993). More specifically, for the case of an exponential covariance function on $\mathbb{R}^1$, say $C(s-t) = \exp(-\phi|s-t|)$, it appears that information is (i) limited in the mean; (ii) limited in σ; (iii) limited in ϕ; (iv) but nevertheless unlimited in the parameter $\sigma\sqrt{\phi}$.

3.2 The information matrix

In asymptotic analysis, usually the first question to ask is whether the parameters in the model can be consistently estimated or not. If the answer is yes, then we can turn to the asymptotic distributions and rates of convergence of various estimators. Usually the Fisher information matrix enables assessment of consistency and the rate of convergence. Under regularity conditions, the Cramér-Rao inequality provides lower bounds for the variance of estimators. In the limit, the asymptotic variance associated with an asymptotically unbiased estimator is no less than the value of the corresponding element in the inverse information matrix. If the variance of an estimator asymptotically reaches the Cramér-Rao lower bound, this estimator is asymptotically efficient.

For the classical *independent identically distributed* (i.i.d.) sampling set up, usually, the MLEs are $\sqrt{n}$ consistent and asymptotically normally distributed with the covariance matrix equal to the inverse information matrix. For data from independent but not identical distributions, the information for estimating some model parameters might be limited. Consider a simple example: let $X_i \overset{ind.}{\sim} \mathrm{N}(\mu, i^2)$, $i = 1, ..., \infty$, where μ is an unknown parameter. It is easy to see that μ cannot be consistently estimated and the limiting information for is $\sum_{i=1}^{\infty} 1/i^2 = \pi^2/6$. In this example,

we have limited information about μ even though we have an infinite sample size. The general behavior of the information matrix for dependent data will be studied in subsequent sections.

Note that when the number of the parameters is greater than one, the interpretation of the entries of the information matrix requires care. If the information matrix is denoted by $I(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ is a vector of parameters, then the i^{th} diagonal element $I(\boldsymbol{\theta})_{i,i}$ is the *conditional* information for the i^{th} parameter, i.e., assuming all other parameters are known. The *marginal* information for a parameter (assuming all other parameters are unknown) is $1/I(\boldsymbol{\theta})^{i,i}$, the reciprocal of the i^{th} diagonal element of the inverse information matrix. Here $I(\boldsymbol{\theta})^{i,i} = (I(\boldsymbol{\theta})^{-1})_{i,i}$. Also it is worth noting that $I(\boldsymbol{\theta})_{i,i} \to \infty$ (as the sample size tends to infinity) does not always imply $I(\boldsymbol{\theta})^{i,i} \to 0$. However, $I(\boldsymbol{\theta})^{i,i} \to 0$ implies $I(\boldsymbol{\theta})_{i,i} \to \infty$.

Here we calculate the Fisher information matrix associated with model (3.1). Suppose that we observe Y at $t_1, \ldots, t_n \in D \subset \mathbb{R}^d$ and denote $\mathbf{Y} = (Y(t_1), ..., Y(t_n))$. We also assume that $\mu(t_i) = X^T(t_i)\boldsymbol{\beta}$. Let $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi})^T$, with covariance matrix $K_{\boldsymbol{\theta}} = \sigma^2 R(\boldsymbol{\phi})$. Here, $R(\boldsymbol{\phi})$ is the correlation matrix associated with the n locations whose $(i,j)^{th}$ entry is $\varrho(t_i - t_j; \boldsymbol{\phi})$ and $\boldsymbol{\phi}$ indexes the parameters of the correlation function, for example, in the Matérn case, a smoothness parameter and a decay parameter (see Section 2.1.3).

The log likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K_{\boldsymbol{\theta}}| - \frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta})^T K_{\boldsymbol{\theta}}^{-1}(\mathbf{Y} - X\boldsymbol{\beta}). \tag{3.3}$$

The score function $S(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ is

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = (\mathbf{Y} - X\boldsymbol{\beta})^T K_{\boldsymbol{\theta}}^{-1} X$$

and the associated Hessian is

$$\mathrm{H}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{\partial^2\ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -X^T K_{\boldsymbol{\theta}}^{-1} X.$$

So, the expected information matrix for $\boldsymbol{\beta}$ is $I(\boldsymbol{\beta}) = -\mathrm{E}\{\mathrm{H}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\boldsymbol{\theta})\} = X^T K_{\boldsymbol{\theta}}^{-1} X$.

The score function for the i^{th} component of $\boldsymbol{\theta}$ is

$$\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i} = -\frac{1}{2}\mathrm{Tr}(K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i}) + \frac{1}{2}(\mathbf{Y}-X\boldsymbol{\beta})^T K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i}K_{\boldsymbol{\theta}}^{-1}(\mathbf{Y}-X\boldsymbol{\beta})$$

and the $(i,j)^{th}$ entry of the associated Hessian matrix $\mathrm{H}_{\boldsymbol{\theta}}\ell(\boldsymbol{\beta},\boldsymbol{\theta})$ is

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j} &= \frac{1}{2}\mathrm{Tr}[K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_j}K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i} - K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}] \\ &+ \frac{1}{2}(\mathbf{Y}-X\boldsymbol{\beta})^T[-2K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_j}K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i}K_{\boldsymbol{\theta}}^{-1} + K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}K_{\boldsymbol{\theta}}^{-1}](\mathbf{Y}-X\boldsymbol{\beta}).\end{aligned}$$

Hence, the $(i,j)^{th}$ entry of the expected information matrix of θ is

$$I(\boldsymbol{\theta})_{i,j} = -\mathrm{E}[\frac{\partial\ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}] = \frac{1}{2}\mathrm{Tr}[K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i}K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_j}]. \tag{3.4}$$

Finally, the expected information matrix for (β,θ) has the block diagonal form

$$I(\boldsymbol{\beta},\boldsymbol{\theta}) = \begin{pmatrix} X^T K_{\boldsymbol{\theta}}^{-1} X & 0 \\ 0 & \left(\frac{1}{2}\mathrm{Tr}[K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_i}K_{\boldsymbol{\theta}}^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}_j}]\right) \end{pmatrix}. \tag{3.5}$$

The block diagonal form in (3.5) reveals that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are *orthogonal* parameters (Cox and Reid, 1987). Informally, this means that information for estimating $\boldsymbol{\beta}$ is "independent" of the information for estimating $\boldsymbol{\theta}$.

3.3 Equivalent measures

For a general parametric covariance function, closed form expressions for the determinant and inverse of the covariance matrix will not be available. As a result, it is hard to derive the properties of the likelihood function. The use of equivalent measures enables study of consistency through examination of the spectral density of the covariance function, avoiding matrix calculation. The basic idea in using equivalent measures is as follows: *if we know that the data are generated by one of two equivalent measures P_1 and P_2, we are not able to distinguish with probability one which measure is correct regardless of what we observe.* In other words, we will not be able to provide consistent estimation.

First let us recall the definition of equivalent/orthogonal measures. Let two probability measures to be defined on the same measurable space $(X, \mathcal{F})$. We say P_1 is absolutely continuous with respect to P_2, denoted by $P_1 \ll P_2$, if for any $A \in \mathcal{F}$, $P_2(A) = 0$ implies $P_1(A) = 0$. If $P_1 \ll P_2$ and $P_2 \ll P_1$, we call P_1 and P_2 equivalent, denoted by $P_1 \equiv P_2$. If there exists two disjoint measurable sets A_1 and A_2 such that $P_1(A_1) = 1$ and $P_2(A_2) = 1$, we call P_1 and P_2 orthogonal, denoted by $P_1 \perp P_2$. In our problem, $\mathcal{F}$ is the σ-algebra generated by the spatial process $Y(s)$, where $s \in D$ and $D \subset \mathbb{R}^d$ is a compact set. The equivalence or orthogonality of two measures is informative. For example, suppose we have two uniform measures P_1 on $[0, 1]$ and P_2 on $[2, 3]$. Assume the data are generated by either P_1 or P_2. It is trivial to see P_1 and P_2 are orthogonal and we can identify the true measure perfectly based on any number of observations. On the other hand, if $P_1 \equiv P_2$, then we are not able to distinguish perfectly which measure generate the data based on what we observe. The following are a few clarifying remarks.

Remark 1. *If $P_1 \equiv P_2$, we cannot distinguish P_1 and P_2 with P_1-probability 1. In other words, we cannot make a perfect decision to choose P_1 or P_2 no matter what*

we observe.

Suppose we observe an event A (i.e. $P_1(A) > 0$ and $P_2(A) > 0$), if we can perfectly distinguish P_1 and P_2, we must be able to construct an event B based on A such that B occurs with P_1-probability 1 and P_2-probability 0. But if $P_1(B|A) = 1$, we have $P_1(B^c|A) = 0$ which implies $P_2(B^c|A) = 0$ (B^c means the complement of B). This further implies $P_2(B|A) = 1$ which leads to contradiction. Thus we cannot distinguish P_1 and P_2 perfectly.

Remark 2. *Let $P_{\boldsymbol{\theta}}$ be a measure indexed by a vector of parameters $\boldsymbol{\theta} = (\theta^{(1)}, ..., \theta^{(p)})$. If $P_{\boldsymbol{\theta}_1} \equiv P_{\boldsymbol{\theta}_2}$, where $\boldsymbol{\theta}_i = (\theta_i^{(1)}, ..., \theta_i^{(p)})$ for $i = 1, 2$; and suppose $\theta_1^{(j)} \neq \theta_2^{(j)}$ for $j = 1, ..., q$ with $q \leq p$, then there is no weakly consistent estimator for each component in the subset $(\theta^{(1)}, ..., \theta^{(q)})$. Here weakly consistent means convergence in probability.*

Consider $\theta^{(1)}$ in the subset $(\theta^{(1)}, ..., \theta^{(q)})$ and suppose $\hat{\theta}_n$ is a weakly consistent estimator for $\theta^{(1)}$. Then there exists a subsequence $\hat{\theta}_{n_k}$ which is strongly consistent (i.e. $P_{\boldsymbol{\theta}_1}(\hat{\theta}_{n_k} \to \theta_1^{(1)}) = 1$). $P_{\boldsymbol{\theta}_1} \equiv P_{\boldsymbol{\theta}_2}$ implies $P_{\boldsymbol{\theta}_2}(\hat{\theta}_{n_k} \to \theta_1^{(1)}) = 1$. For the sequence $\hat{\theta}_{n_k}$ which converges weakly to θ_2 under probability $P_{\boldsymbol{\theta}_2}$, there is a sub-subsequence $\hat{\theta}_{n_{k_j}}$ such that $P_{\boldsymbol{\theta}_2}(\hat{\theta}_{n_{k_j}} \to \theta_2^{(1)})$=1. This leads to contradiction since $\theta_1^{(1)} \neq \theta_2^{(1)}$.

Remark 3. *If the limiting value of the inverse information of a parameter (by inverse information of a parameter we mean the corresponding element in the inverse information matrix) is bounded below by a positive number as sample size tends to infinity, then there is no L^2-consistent estimator for that parameter. If there is no consistent estimator for a parameter and the variance of MLE attains the Cramér-Rao lower bound asymptotically, then the associated inverse information is bounded below by a positive number and hence the information is bounded above by a finite number no matter how large the sample size is.*

Finally, a theorem of Blackwell and Dubins (1962) tells us that the parametriza-

tion under which measures are equivalent is important for prediction. Two measures with different parametric covariance functions could be equivalent. Then, the predictions based on the wrong measure are asymptotically optimal as long as this wrong measure is equivalent to the correct measure.

From the above remarks, we can use the idea of equivalent measures to study consistency and limiting behavior of the Fisher information. In general, two measures could be equivalent, orthogonal, or neither equivalent nor orthogonal. However, it is well known that two Gaussian measures are either equivalent or orthogonal (see e.g. Gikhman and Skorokhod, 1974, Chapter 7). Furthermore, there are many results that provide conditions for checking the equivalence of Gaussian measures.

A Gaussian measure is characterized by its mean and covariance function. Suppose $Y_j(t)$ $(j = 1, 2)$ are stationary Gaussian random fields on $D \in \mathbb{R}^d$ with mean $\mu_j(t)$ and covariance function $C(s-t) = \sigma_j^2 \varrho_j(s-t)$ and $\varrho_j(s-t)$ has a spectral representation

$$\varrho_j(s-t) = \int_{\mathbb{R}^d} e^{i(\omega, s-t)} F_j(\mathrm{d}\omega), \tag{3.6}$$

where $F_j(\cdot)$ is the spectral measure of $Y_i(t)$ and $F_j(\mathrm{d}\omega) = f_j(\omega)\,\mathrm{d}\omega$ if $F_j(\cdot)$ is absolutely continuous with respect to Lebesgue measure. The random fields $Y_j(t)$ have the representation

$$Y_j(t) = \sigma_j \int_{\mathbb{R}^d} e^{i(\omega, t)} Z_j(\mathrm{d}\omega), \tag{3.7}$$

where $Z_j(\cdot)$ is a random measure such that $\mathrm{E}\{Z_j(A_1)Z_j(A_2)\} = F_j(A_1 \cap A_2)$ for measurable sets A_1 and A_2. Gaussian random fields $Y_j(t), t \in D$ induces Gaussian measure P_j on the measurable space $(\mathbb{R}^D, \mathcal{B}^D)$.

Let $L_2(D)$ be the space of square-integrable functions on D. $L_2(D)$ is a Hilbert

space with inner product

$$(x(t), y(t)) = \int_D x(t)\overline{y(t)}\,\mathrm{d}t \tag{3.8}$$

for $x(t), y(t) \in L_2(D)$. The covariance operator $\mathcal{K}$ is defined as

$$\mathcal{K}x(t) = \int_D \sigma^2 \varrho(t,s)x(s)\,\mathrm{d}s \tag{3.9}$$

for $x(t) \in L_2(D)$. We use the notation $P(\mu, C)$, $P(\mu, \sigma^2\varrho)$, or $P(\mu, \mathcal{K})$ for a Gaussian measure with mean μ and covariance function $C = \sigma^2\varrho$, or covariance operator $\mathcal{K}$.

The fundamental theorems about Gaussian equivalent measures state the conditions in terms of operators which are not easy to verify. Fortunately, there are many more specific conditions expressed in terms of either the covariance functions or the spectral measures. See Gikhman and Skorokhod (1974), Ibragimov and Rozanov (1978), and Yadrenko (1983) for details.

3.4 Only the location parameter is unknown

We begin with the simplest case in model (3.1) assuming the parameters in covariance function, σ^2 and $\boldsymbol{\phi}$ are known, and we want to estimate the unknown mean.

3.4.1 Estimating μ: the sample average and BLUE

Suppose a one dimensional process $Y(t)$ with a constant mean μ is observed at $t_1, ..., t_n \in [0, L]$ where L is a bounded positive real number. For concreteness, let $t_i = iL/n$, $i = 1, ..., n$. Suppose we use the sample average $\bar{\mu}_n = \sum_{i=1}^n Y(t_i)/n$ to estimate μ. $\bar{\mu}_n$ has variance

$$\zeta_n^2 = \frac{1}{n^2}\sum_{i,j} C(t_i, t_j) \to \zeta^2 = \int_0^L \int_0^L C(s,t)\,\mathrm{d}s\,\mathrm{d}t/L^2. \tag{3.10}$$

ζ^2 is the variance associated with the stochastic integral $\bar{\mu}_L = \int_0^L Y(t)\,\mathrm{d}t/L$, clarifying that, even with an arbitrarily large sample, precision is limited. One cannot estimate the mean parameter arbitrarily well with a single realisation, however dense, on a bounded region. For instance, consider $C(s,t) = \exp(-\phi|s-t|)$, $L = 1$, and $\phi = 1$, corresponding to a time-transformed *Ornstein-Uhlenbeck* (OU) process. Then, $\zeta^2 = 2/e = 0.7358$. But even with a quite small n, ζ_n^2 is close to ζ^2, e.g., for $n = 5, 20, 100$ the values are 0.7466, 0.7364, 0.7358, respectively.

Now suppose we allow L to grow large. It can be shown that (Yaglom, 1987, Section 16)

$$\lim_{L\to\infty} \bar{\mu}_L = \lim_{L\to\infty} \frac{1}{L}\int_0^L Y(t)\,\mathrm{d}t = \mu \text{ if and only if } \lim_{L\to\infty} \frac{1}{L}\int_0^L \varrho(h)\,\mathrm{d}h = 0. \quad (3.11)$$

Note the condition that $\lim_{L\to\infty}\int_0^L \varrho(h)\,\mathrm{d}h/L = 0$ is weaker than the condition that $\varrho(h) \to 0$ as $h \to \infty$. It is also worth noting that

$$\lim_{L\to\infty}\int_0^L \varrho(h)\,\mathrm{d}h/L = F(+0) - F(-0), \quad (3.12)$$

where $F(\cdot)$ is the spectral measure of $\varrho(\cdot)$. So $\bar{\mu}_L$ is consistent as $L \to \infty$ if and only if the spectrum of the correlation function is continuous at 0. So for middle-ground and expansion asymptotics, under fairly general conditions, μ can be consistently estimated.

Suppose we consider the *best linear unbiased estimator* (BLUE) instead of the sample average, i.e., consider

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} a(iL/n)Y(iL/n), \quad (3.13)$$

the limit of which is $\widehat{\mu}_L = \int_0^L Y(s)\,\mathrm{d}A(s)$. Unbiasedness requires $(1/n)\sum_{i=1}^n a(iL/n) = 1$, which in the limit means $\int_0^L \mathrm{d}A(s) = 1$. We may think of $A(s)$ as a *signed*

measure with *unit mass* or as a *"generalized"* distribution function. The variance of $\hat{\mu}_n$ is

$$\begin{aligned} \sigma_n^2 &= \frac{1}{n^2}\sum_{i,j} a(iL/n)a(jL/n)C(iL/n, jL/n) \\ &\to \int_0^L \int_0^L C(s,t)\,\mathrm{d}A(s)\,\mathrm{d}A(t)/L^2 = \sigma^2. \end{aligned} \tag{3.14}$$

There should be a well-defined best *weighting function* $\mathrm{d}A(s)$ that minimizes the limit variance among those with $\int_0^L \mathrm{d}A(s) = 1$. It is easy to consider finding the best weighting function for finite n, then passing to the limit. A Lagrange multiplier calculation shows that the best linear combination sets $\mathbf{a}_0 = R_n^{-1}\mathbf{1}/\mathbf{1}^T R_n^{-1}\mathbf{1}$ with resulting

$$\mathrm{Var}\,(\mathbf{a}_0^T Y) = \sigma^2/\mathbf{1}^T R_n^{-1}\mathbf{1}. \tag{3.15}$$

This corresponds to the optimal linear estimator

$$\widehat{\mu}_n = \mathbf{a}_0^T Y = \frac{\mathbf{1}^T R_n^{-1} Y}{\mathbf{1}^T R_n^{-1}\mathbf{1}}. \tag{3.16}$$

For the OU process, R_n^{-1} can be calculated explicitly (see Appendix A.1) leading to the information expression for μ

$$A_n(\phi) \equiv \mathbf{1}^T R_n^{-1}\mathbf{1} = 1 + (n-1)(1-\rho_n)/(1+\rho_n), \tag{3.17}$$

where $\rho_n = \exp(-\phi L/n)$. As $n \to \infty$, the limit is

$$\lim_{n\to\infty} A_n(\phi) = 1 + \phi L/2, \tag{3.18}$$

explicitly providing the bound on information.

Furthermore, the weighting function $\mathbf{a}_0 = (a_{0,1}, ..., a_{0,n})^T$ with $a_{0,1} = a_{0,n} = (n - (n-2)\rho_n)^{-1}$ and $a_{0,i} = (1-\rho_n)(n-(n-2)\rho_n)^{-1}$ for $i = 2, ..., n-1$. So, in a limit,

for the OU process on $[0, L]$, the BLUE weights $\mathrm{d}A(t)$ is

$$\mathrm{d}A(t) = \frac{1}{2+\phi L}\{\delta(t) + \delta(t-L) + \phi 1_{(0,L)}(t)\}\,\mathrm{d}t, \tag{3.19}$$

where $\delta(t)$ is a Dirac delta function with $\int_{-\infty}^{\infty} f(x)\delta(t-x)\,\mathrm{d}t = f(x)$.

How much smaller is the variance of $\hat{\mu}_n$ compared with that of $\bar{\mu}_n$? The relative efficiency

$$\mathrm{Var}\,\hat{\mu}_n / \mathrm{Var}\,\bar{\mu}_n$$

can be studied. For the OU case with $L = 1$, the ratio is rather close to 1 for small ϕ and large ϕ, and, for ϕ around 2.5, the ratio reaches at its maximum value of 1.14. This may suggest that for ordinary correlation functions, the relative efficiency never gets large. In other words, the simple sample average estimation may not lose much to the optimal strategy.

For a general covariance function, R_n^{-1} has no closed form expression so we are not able to find $\hat{\mu} = \lim_{n\to\infty}(\mathbf{1}^T R_n^{-1} Y)/(\mathbf{1}^T R_n^{-1}\mathbf{1})$ by direct calculation. As an alternative approach, for a random process $Y(t)$ $(t \in D)$ with mean $x(t)\mu$ ($x(t)$ is a known covariate function. For example, $x(t)$=1 in the above discussion) and covariance function $\sigma^2\varrho$, Grenander (1981, Chapter 4) shows the existence and uniqueness of the BLUE for μ. He also shows that BLUE $\hat{\mu}$ can be characterized by a so-called *normal equation*

$$\mathrm{E}\{\hat{\mu}Y(s)\} = (\mathrm{Var}(\hat{\mu}) + \mu^2)x(s) \equiv cx(s) \quad \text{for any } s \in D, \tag{3.20}$$

where c is a constant. This is easy to see by noting that the variance of any other linear unbiased estimator $\hat{\mu}^* = \hat{\mu} + \epsilon\{Y(t)x(s) - Y(s)x(t)\}$ is greater than the variance of BLUE $\hat{\mu}$ for arbitrary ϵ and any $s, t \in D$. As a special case, if $x(s) = 1$ for $s \in D$ and $\hat{\mu} = \int_D Y(t)\,\mathrm{d}A(t)$, the normal equation becomes

$$\mathrm{E}\{\hat{\mu}Y(s)\} = \int_D \varrho(s,t)\,\mathrm{d}A(t) + \mu^2 = c. \tag{3.21}$$

So if we can find $\mathrm{d}A(t)$ (up to a normalizing constant) such that $\int_D \varrho(s,t)\,\mathrm{d}A(t)$ is a constant for each s, we are done.

Example 1: For the OU case on $[0, L]$, it is easy to verify that when $\mathrm{d}A(t) = \{\delta(t) + \delta(t-L) + \phi 1_{(0,L)}(t)\}\,\mathrm{d}t$, $\int_0^L e^{-\phi|s-t|}\,\mathrm{d}A(t) = 2$ for all $s \in [0, L]$. We normalize $\mathrm{d}A(t)$ to get back (3.19) and find the variance of the BLUE is $2/(2+\phi L)$.

Example 2: For a Gaussian correlation function $\varrho(s,t) = \exp(-\phi(s-t)^2)$ $(t, s \in [0, L])$, we want to find $\mathrm{d}A(t)$ such that $\int_0^L e^{-\phi(s-t)^2}\,\mathrm{d}A(t)$ is constant for all s. We differentiate both sides of the equation with respect to s to get $\int_0^L e^{-\phi(s-t)^2}(t-s)\,\mathrm{d}A(t) = 0$ for any s. If we let $s = 0$, we have $\int_0^L e^{-\phi t^2} t\,\mathrm{d}A(t) = 0$ so there must be negative weights for t in the interval $[0, L]$. Though an exact solution for $\mathrm{d}A(t)$ is elusive, Figure 3.1 shows how the BLUE weights behave for a finite n ($n = 30$, $L = 1$ and $\phi = 10$).

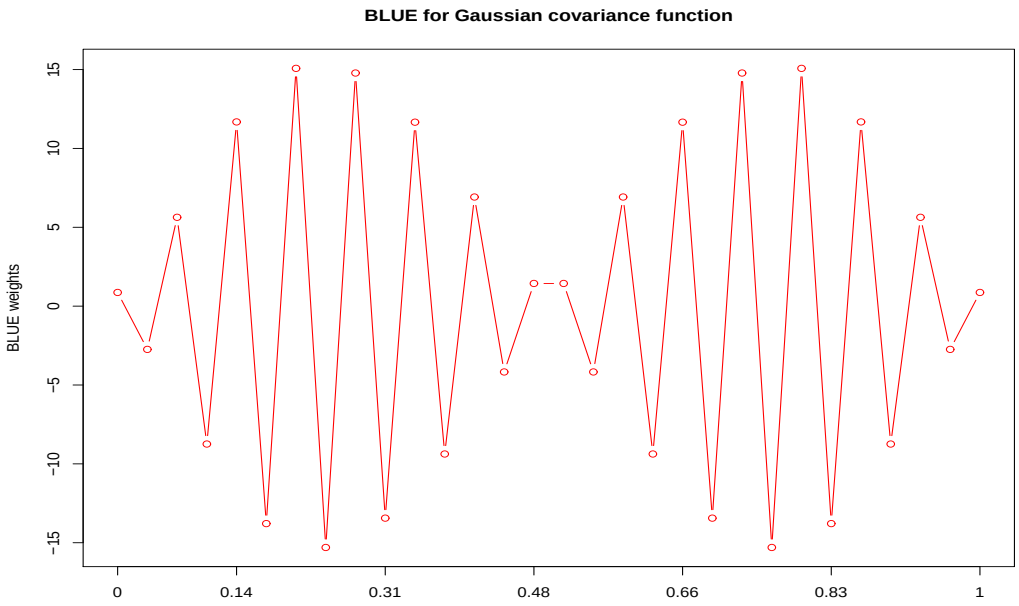

Figure 3.1: BLUE weights for a Gaussian process on $[0, 1]$ with the correlation function $\exp(-10|s-t|^2)$, $n = 30$.

Example 3: Turning to the Matérn covariance function (see Section 2.1.3),

$$C(s,t;\sigma^2,\phi,\nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\phi\|s-t\|)^{\nu}\kappa_{\nu}(\phi\|s-t\|), \tag{3.22}$$

where ϕ is correlation decay parameter; ν is the smoothness parameter; σ^2 is the variance parameter; and κ_ν is the modified Bessel function of the second kind of order ν. The modified Bessel function is not easy to work with. Even for the special case where the smooth parameter $\nu = 3/2$ which implies the correlation function has the closed form $e^{-\phi|s-t|}(1+\phi|s-t|)$, it is not easy to find the exact form of $\mathrm{d}A(t)$. Figure 3.2 shows some numerical results for $\mathrm{d}A(t)$ with different values of ν ($\sigma^2 = 1, \phi = 1, n = 50$). It can be seen that $\nu = 0.5$ is a *critical* point. When $\nu \leq 0.5$, $\mathrm{d}A(t)$ is positive and less than 1 for all t and the shape of $\mathrm{d}A(t)$ converges to the shape of (3.19) as ν goes to 0.5. When $\nu > 0.5$, $\mathrm{d}A(t)$ becomes very irregular with enormous negative and positive weights. The shape of $\mathrm{d}A(t)$ converges to that of $\mathrm{d}A(t)$ for the Gaussian covariance function as ν goes to infinity. Perhaps most important is that it seems very *unattractive* to use such extreme weights to estimate the mean with positively correlated data. For example,

$$\widehat{\mu} = 3519.189y_1 - 10560.157y_2 + 12122.415y_3 - 7506.865y_4 + 3674.883y_5 - \cdots$$

does look strange, for data with standard deviation 1. These weights are for $n = 50$ and $\nu = 3$. The weights escalate further with increasing ν and increasing samples intensity n. This aspect of traditional geostatistics theory *does not* seem to be noted in the literature.

Grenander (1981, Chapter 4) suggests a way to understand why the weights at boundary (or near boundary) points are so different. If we want to estimate the mean

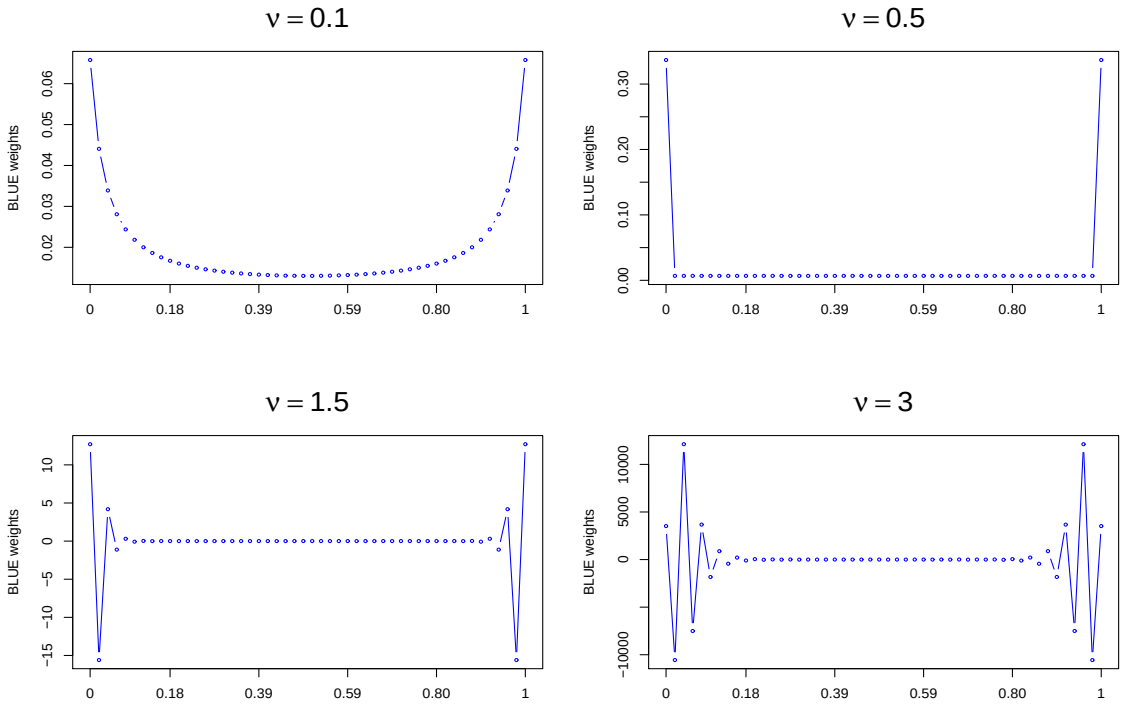

Figure 3.2: BLUE weights $\mathrm{d}A(t)$ for a Gaussian process on $[0, 1]$ with Matérn covariance function ($n = 50, \sigma^2 = 1, \phi = 1, \nu = 0.1, 0.5, 1.5, 3$).

of a random process based on observations $\{Y(t), t \in D\}$, we could conceptually obtain the unbiased prediction $\hat{Y}(t)$ for $t \notin D$. Then we can estimate the mean based on the combined set of $\{Y(t)\}$ and $\{\hat{Y}(t)\}$. This procedure can not change the information contained in the original data in terms of estimating μ. Again, consider the OU process on $[0, L]$. The OU process is Markovian (see Appendix A.1) so the best unbiased prediction of $Y(t)$ for $t < 0$ only depends on $Y(0)$ and $\hat{Y}(t)$ only depends on $Y(L)$ for $t > L$. Now we have three pieces of information, the first one is from $\{\hat{Y}(t), t < 0\}$, the second one comes from $\{\hat{Y}(t), t > L\}$ and the third one is from $\{Y(t), t \in [0, L]\}$. The contributions from these three information sources are weighted according to the decay parameter ϕ as in (3.19). Note that if ϕ increases, the correlation will decrease so that the contribution from $\{Y(t), t \in [0, L]\}$ will increase. For a process with a covariance function other than exponential, $\hat{Y}(t)$ $(t > L)$ usually

depends more on the samples close to the right boundary (similarly, $\hat{Y}(t)$ depends more on the samples close to the left boundary for $t < 0$). As a result, samples close to boundaries will have bigger impact on $\hat{\mu}$.

3.4.2 Bayesian beware

In terms of Bayesian analysis, suppose we place a prior $\pi(\mu)$ on the mean parameter, with consequent posterior density

$$\pi_n(\mu \,|\, \text{data}) \propto \pi(\mu) \exp\{-\tfrac{1}{2} A_n (\mu - \widehat{\mu}_n)^2\}, \tag{3.23}$$

in terms of the information level A_n and the ML estimator $\widehat{\mu}_n$. Since A_n is limited, the posterior will *for ever* depend on aspects of the prior. The limiting posterior, for infinite fill-in, is

$$\pi(\mu \,|\, \text{data}) \propto \pi(\mu) \exp\{-\tfrac{1}{2} A (\mu - \widehat{\mu})^2\},$$

with A the finite limit of A_n. If the prior is a normal $(\mu_0, 1/A_0)$, for example, then the Bayes estimator, and its limit for large n, are

$$\widetilde{\mu}_n = \frac{A_0 \mu_0 + A_n \widehat{\mu}_n}{A_0 + A_n} \quad \text{and} \quad \widetilde{\mu} = \frac{A_0 \mu_0 + A \widehat{\mu}}{A_0 + A}. \tag{3.24}$$

Thus one needs to be particularly careful with one's prior in such contexts.

3.4.3 Equivalent measures and estimation of μ

As discussed in Section 3.3, the approach of equivalent measures can help us understand the limiting behavior of the information about μ. Let us start with two theorems that give some specific conditions for verifying equivalence of Gaussian measures.

Theorem 3 (Ibragimov and Rozanov, 1978, p.77)**.** *Suppose two Gaussian measures* $P(\mu_1(t), \sigma_1^2 \varrho_1)$ *and* $P(\mu_2(t), \sigma_2^2 \varrho_2)$ *are corresponding to two Gaussian processes* $Y_1(t)$

and $Y_2(t)$ defined on some set D. $P(\mu_1(t), \sigma_1^2\varrho_1) \equiv P(\mu_2(t), \sigma_2^2\varrho_2)$ if and only if $\lim_{n\to\infty}(E_1 \log \frac{f_{1,n}}{f_{2,n}} + E_2 \log \frac{f_{2,n}}{f_{1,n}}) < \infty$, where $f_{j,n}$ $(j = 1, 2)$ is the joint density of $\{Y_j(t_1), ..., Y_j(t_n), t_i \in D\}$, E_1 and E_2 denotes expectations under $P(\mu_1(t), \sigma_1^2\varrho_1)$ and $P(\mu_2(t), \sigma_2^2\varrho_2)$, and $\{t_i\}$ is dense in D.

We add the following corollary:

Corollary 3.1. *If $\mu_1(t) = u_1$, $\mu_2(t) = u_2$ for $t \in D$, then $P(\mu_1, \sigma^2\varrho) \equiv P(\mu_2, \sigma^2\varrho)$ if and only if $(u_1 - u_2)^2 \lim_{n\to\infty} \mathbf{1}^T R_n^{-1} \mathbf{1} < \infty$.*

Proof. $\log \frac{f_{1,n}}{f_{2,n}} = (u_1 - u_2) Y^T R_n^{-1} 1 - \frac{1}{2} u_1^2 \mathbf{1}^T R_n^{-1} \mathbf{1} + \frac{1}{2} u_2^2 \mathbf{1}^T R_n^{-1} \mathbf{1}$ and $\mathrm{E}_1 \log \frac{f_{1,n}}{f_{2,n}} = \frac{1}{2} u_1^2 \mathbf{1}^T R_n^{-1} \mathbf{1} + \frac{1}{2} u_2^2 \mathbf{1}^T R_n^{-1} \mathbf{1} - u_1 u_2 \mathbf{1}^T R_n^{-1} \mathbf{1}$. So $\lim_{n\to\infty}(\mathrm{E}_1 \log \frac{f_{1,n}}{f_{2,n}} + \mathrm{E}_2 \log \frac{f_{2,n}}{f_{1,n}}) = (u_1 - u_2)^2 \mathbf{1}^T R_n^{-1} \mathbf{1}$. □

Theorem 4 (Ibragimov and Rozanov, 1978, p.86; Yadrenko, 1983, p.137)**.** *The Gaussian measures $P(0, \sigma^2\varrho)$ and $P(\mu(t), \sigma^2\varrho)$ defined on the σ-algebra $\mathcal{B}^D$ are equivalent if and only if the mean $\mu(t)$ permits a spectral representation*

$$\mu(t) = \int e^{-i\lambda t} \psi(\lambda) F(d\lambda) \tag{3.25}$$

for $t \in D$, where the function $\psi(\lambda)$ satisfies the condition $\int |\psi(\lambda)|^2 F(d\lambda) < \infty$. Here $F(d\lambda)$ is the spectral measure of the covariance function.

If we can verify two measures (with different means but a same covariance structure) are equivalent using Theorem 4, then by Remark 2, we cannot estimate mean consistently. And by Corollary 3.1, $\mathbf{1}^T R_n^{-1} \mathbf{1}$ is bounded.

3.4.4 Properties of $A_n(\boldsymbol{\phi})$

We now summarize a few basic properties of $A_n(\boldsymbol{\phi})$. Perhaps the most useful result is the explicit form of the information gain upon sampling an additional location, say

t_0. We have

$$A_{n+1}(\boldsymbol{\phi}) - A_n(\boldsymbol{\phi}) = \frac{(1 - \mathbf{1}^T R_n^{-1}(\boldsymbol{\phi})\mathbf{r}_{n0}(\boldsymbol{\phi}))^2}{1 - \mathbf{r}_{n0}(\boldsymbol{\phi})^T R_n^{-1}(\boldsymbol{\phi})\mathbf{r}_{n0}(\boldsymbol{\phi})} \tag{3.26}$$

where $\mathbf{r}_{n0}$ is an $n \times 1$ vector with i^{th} entry $\varrho(t_i - t_0; \boldsymbol{\phi})$. This result is a special case of the recursion of Brimkulov, Krug, and Savanov (1986).

Next, since $A_1 = 1$, we have $A_n > 1$ and A_n increasing in n. As the scalar decay parameter $\phi \to \infty$, $A_n \to n$. Is $A_n \leq n$? A simple calculation shows that $A_2 - A_1 = (1-\rho)/(1+\rho)$, where $\rho = \varrho(t_1 - t_2; \phi)$. So, if ρ is allowed to be less than 0, $A_2 > 2$ and, in fact, $A_2 - A_1 \to \infty$ as $\rho \to -1$. If ϱ only takes on positive values then $A_{n+1}(\phi) - A_n(\phi) \leq 1$ and $A_n \leq n$.

For the OU process, we have the following results (see Appendix A.1 for proofs)

Result 1: For an OU process on $[0, L]$, $\mathbf{1}^T R_n^{-1}(\phi)\mathbf{1}$ increases as ϕ increases, i.e. $\mathbf{1}^T R_n^{-1}(\phi_1)\mathbf{1} > \mathbf{1}^T R_n^{-1}(\phi_2)\mathbf{1}$ if $\phi_1 > \phi_2 > 0$.

Result 2: For an OU process on $[0, L]$, $\lim_{\phi \to 0} \mathbf{1}^T R_n^{-1}\mathbf{1} = 1$ and $\lim_{\phi \to \infty} \mathbf{1}^T R_n^{-1}\mathbf{1} = n$.

Finally, as we have noted, the only covariance functions in $\mathbb{R}^1$ that have proven amenable to explicit computation of R_n^{-1} and $|R_n|$ are two cases of the Matérn class, namely smoothness parameter $\nu = 0.5$ (the exponential case) and $\nu = 1.5$ (as in Loh, 2005). To obtain explicit forms for $\mathbb{R}^d$ the only solution to date appears to be a specification that is separable in the coordinates, i.e., $\varrho(s - s'; \phi) = \Pi_{l=1}^d \exp(-|s_l - s'_l|)$ (as in Ying, 1993 and Loh, 2005). The resulting R_n^{-1} has a Kronecker product form.

3.4.5 On computing $\lim_{n \to \infty} \mathbf{1}^T R_n^{-1}\mathbf{1}$.

We wish to calculate the limiting value of $\mathbf{1}^T R_n^{-1}\mathbf{1}$, the information for μ. Motivated by Parseval's equality and Theorem 4, we have the following result:

Theorem 5. *Let the model be* $Y(t) = \mu x(t) + W(t)$ $(t \in D,$ *a compact set in* $\mathbb{R}^d)$*, where* $x(t)$ *is a known covariate function and* $W(t)$ *is a mean 0 stationary process with a correlation function* ϱ*. Then the information for estimating* μ *can be computed as follows:*

$$\lim_{n\to\infty} x(t_1, ..., t_n)^T R_n^{-1} x(t_1, ..., t_n) = \sum_{i=1}^{\infty} \frac{|(\psi_i(\cdot), x(\cdot))|^2}{\lambda_i}, \tag{3.27}$$

where $\{\lambda_i\}$ *is a countable set of eigenvalues and* $\{\psi_i(t)\}$ *are corresponding orthonormal eigenfunctions associated with* ϱ*; and* $(\psi_i(\cdot), x(\cdot)) = \int_D \psi_i(t)x(t)\,dt$ *is the inner product.*

See Appendix A.2 for a proof.

Below we demonstrate how to calculate $\lim_{n\to\infty} \mathbf{1}^T R_n^{-1} \mathbf{1}$ for the OU model using Theorem 5. The Karhunen-Loève expansion (see, for example, Adler 1981) is one way of finding eigenvalues and eigenfunctions for a correlation function ϱ on D. It decomposes $\varrho(s,t)$ $(s,t \in D \subset \mathbb{R}^d)$ into a sum of orthogonal series (see Section 4.2.1 for details):

$$\varrho(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(s)\psi_i(t), \tag{3.28}$$

where $\{\lambda_i\}$ is a countable set of eigenvalues and $\{\psi_i(t)\}$ are corresponding orthonormal eigenfunctions of ϱ. We solve

$$\int_D \varrho(s,t)\psi(t)dt = \lambda\psi(s) \quad \text{and} \quad \int_D \psi_i(t)\psi_j(t)\,\mathrm{d}t = \delta_{ij}, \tag{3.29}$$

where δ_{ij} equals 1 when $i = j$ and 0 otherwise, to obtain λ_i and $\psi_i(t)$.

Example 4: Consider a Gaussian process $Y(t)$ on $D = [-L, L]$ (the symmetric interval is chosen for computation simplicity) with mean 0 and exponential correlation

function $\varrho(t,s) = e^{-\phi|t-s|}$ where $t, s \in [-L, L]$. We solve the integral equation $\lambda\psi(t) = \int_{-L}^{L} e^{-\phi|t-s|}\psi(s)\,\mathrm{d}s$ with $\int_{-L}^{L} \psi_i(s)\psi_j(s)\,\mathrm{d}s = \delta_{ij}$ to find λ_i and ψ_i as follows (see Appendix A.3 for details):

$$\lambda_{1,i} = \frac{2\phi}{w_{1,i}^2 + \phi^2}, \quad \psi_{1,i}(t) = \frac{\cos(w_{1,i}t)}{\sqrt{L + \sin(2w_{1,i}L)/(2w_{1,i})}} \tag{3.30}$$

and

$$\lambda_{2,i} = \frac{2\phi}{w_{2,i}^2 + \phi^2}, \quad \psi_{2,i}(t) = \frac{\sin(w_{2,i}t)}{\sqrt{L - \sin(2w_{2,i}L)/(2w_{2,i})}}, \tag{3.31}$$

where $w_{1,i}$ and $w_{2,i}$ are the solutions of the following two equations (3.32) and (3.33) respectively:

$$\tan(wL) = \phi/w, \quad \text{and} \tag{3.32}$$

$$\tan(wL) = -w/\phi. \tag{3.33}$$

Thus we have

$$\varrho(s,t) = \sum_{i=1}^{\infty} \Big(\lambda_{1,i}\psi_{1,i}(s)\psi_{1,i}(t) + \lambda_{2,i}\psi_{2,i}(s)\psi_{2,i}(t)\Big). \tag{3.34}$$

If we arrange the eigenvalues in decreasing order, we can approximate the correlation function by truncating the infinite sum (3.34). Let us look at a concrete example. Specifically, letting $L = 1$ and $\phi = 2$, we find the 10 largest eigenvalues to be $\lambda_{1\ldots10}$ =(0.7752, 0.4329, 0.2315, 0.1339, 0.0846, 0.0575, 0.0413, 0.0310, 0.0241, 0.0192) and the corresponding eigenfunctions (see Figure 3.3). (The solutions of (3.32) and (3.33) are $w_{1,1\ldots5}$=(1.08, 3.64, 6.58, 9.63,12.72), $w_{2,1\ldots5}$ =(2.29, 5.09, 8.10, 11.72, 14.28). See Figure 3.4.) We approximate the correlation function using these 10 terms. The approximation is fairly good as we can see from Figure 3.5. In general,

the weaker the correlation is, the slower the decrease in the eigenvalues and more terms are needed in order to achieve accurate approximation.

According to Theorem 5, in order to calculate $\lim_{n\to\infty} \mathbf{1}^T R_n^{-1}\mathbf{1}$, we only need to compute $\sum_{i=1}^{\infty} |(\psi_i(\cdot), 1)|^2/\lambda_i$. Note

$$(\psi_{i,1}(\cdot), 1) = \int_{-L}^{L} \psi_{1,i}(t)\,\mathrm{d}t = \frac{2\sin(w_{1,i}L)}{w_{1,i}\sqrt{L + \sin(2w_{1,i}L)/(2w_{1,i})}} \quad \text{and} \quad \int_{-L}^{L} \psi_{2,i}(t)\,\mathrm{d}t = 0.$$

So, by plugging in the first 10 eigenvalues and eigenfunctions we obtain

$$\sum_{i=1}^{10} |(\psi_i(\cdot), 1)|^2/\lambda_i = 2.91.$$

The true value of $\lim_{n\to\infty} \mathbf{1}^T R_n^{-1}\mathbf{1} = 1 + \phi L = 3$. For the $\phi = 1$ and $L = 0.5$ case, using only six eigenvalues, the approximated value of $\lim_{n\to\infty} \mathbf{1}^T R_n^{-1}\mathbf{1}$ is 1.46 while the true value is 1.5.

Remark 4. *The use of the Karhunen-Loève expansion to calculate the information depends on the ability to solve the integral equation (3.29), which is typically a hard task. There exists numerical methods, for example, the "Galerkin method" (see e.g. Ghanem and Spanos, 1991), to find approximate solutions enabling approximate limiting information.*

Another way to handle the difficulty of finding exact eigenvalues is motivated by the following solution for (3.29),

$$\int_{\mathbb{R}^d} \varrho(s-t)e^{i\omega^T t}\,\mathrm{d}t = e^{i\omega^T s}\int_{-\mathbb{R}^d} \varrho(u)e^{-i\omega^T u}\,\mathrm{d}u = \lambda(\omega)e^{i\omega^T s}, \tag{3.35}$$

where $\lambda(\omega) = \int_{\mathbb{R}^d} \varrho(u)e^{-i\omega^T u}\,\mathrm{d}u$.

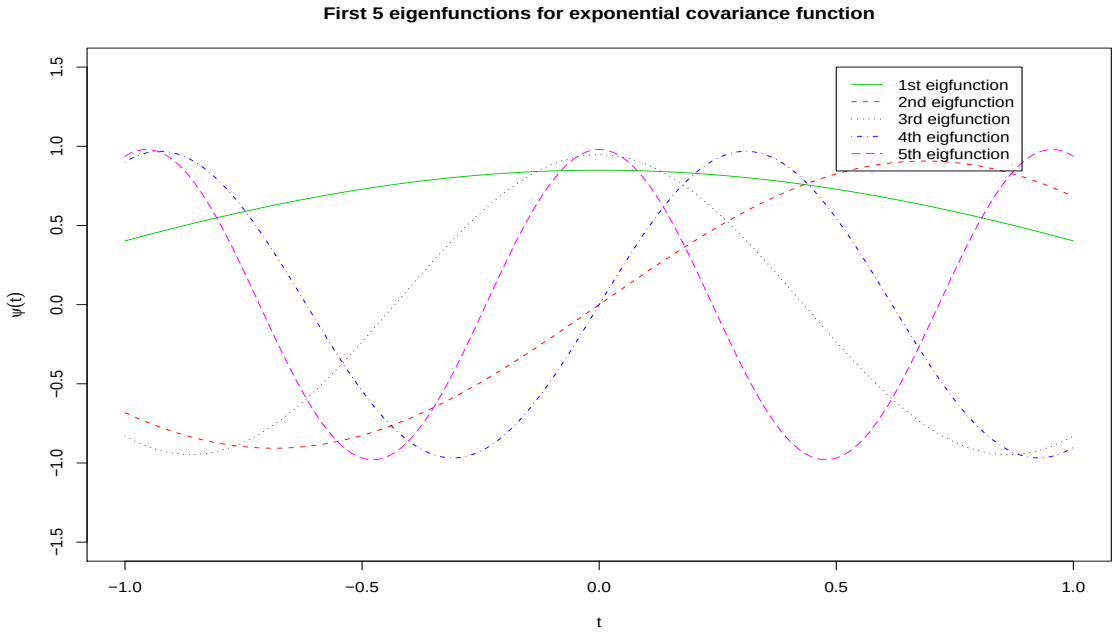

Figure 3.3: First five eigenfunctions for the exponential correlation function on $[-1, 1]$, $\varrho(s,t) = \exp(-2|s-t|)$.

Note that $\lambda(\omega)$ is the spectral density associated with ϱ. For a process $Y(t), t \in [-L/2, L/2]$, we expect the eigenvalues and eigenfunctions to be close to $\lambda(\omega)$ and $e^{i\omega t}$ when L is large. Since we know the spectral density for many covariance functions of interest, we can use the approximated eigenvalues and eigenfunctions to compute (3.27).

We follow the approach of Van Trees (1968, Chapter 3). For the integral equation

$$\int_{-L/2}^{L/2} \varrho(s-t)\psi(t)\,\mathrm{d}t = \lambda\psi(s),$$

we try solutions of the form $\psi_n(t) = e^{i\pi nt/L}, n = 0, \pm 1, \pm 2....$ and substitute in the

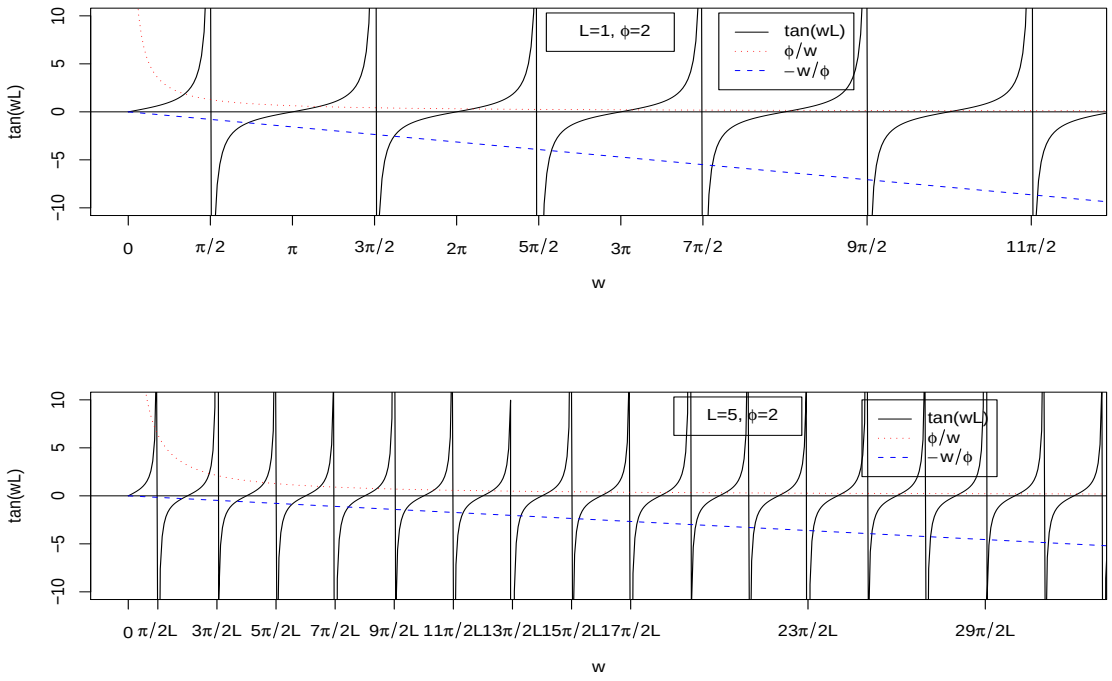

Figure 3.4: Solutions for (3.32) and (3.33)

spectral representation of $\varrho(s-t) = \int_{-\infty}^{\infty} f(\omega)\exp\{i\omega(s-t)\}\,\mathrm{d}\omega$. Then we have

$$\begin{aligned}
\int_{-L/2}^{L/2} \varrho(s-t)e^{i\pi nt/L}\,\mathrm{d}t &= \int_{-L/2}^{L/2}\Big[\int_{-\infty}^{\infty} f(\omega)e^{i\omega(s-t)}\,\mathrm{d}\omega\Big]e^{i\pi nt/L}\,\mathrm{d}t \\
&= \int_{-\infty}^{\infty} f(\omega)e^{i\omega s}\Big[\int_{-L/2}^{L/2} e^{i(\pi n/L-\omega)t}\,\mathrm{d}t\Big]\,\mathrm{d}\omega \\
&= \int_{-\infty}^{\infty} f(\omega)e^{i\omega s}\Big[\frac{2\sin(\pi n/2-\omega L/2)}{\pi n/L-\omega}\Big]\,\mathrm{d}\omega \\
&\approx \int_{-\infty}^{\infty} f(\omega)e^{i\omega s}\delta(\omega-\pi n/L)\,\mathrm{d}\omega \\
&= f(\pi n/L)e^{i\pi ns/L}. \qquad (3.36)
\end{aligned}$$

If you plot the function $\frac{2\sin(\pi n/2-\omega L/2)}{\pi n/L-\omega}$, you will see $\frac{2\sin(\pi n/2-\omega L/2)}{\pi n/L-\omega} \approx \delta(\omega-\pi n/L)$ for large L. Therefore, we find a "cheap" way to find the eigenvalues and correspond-

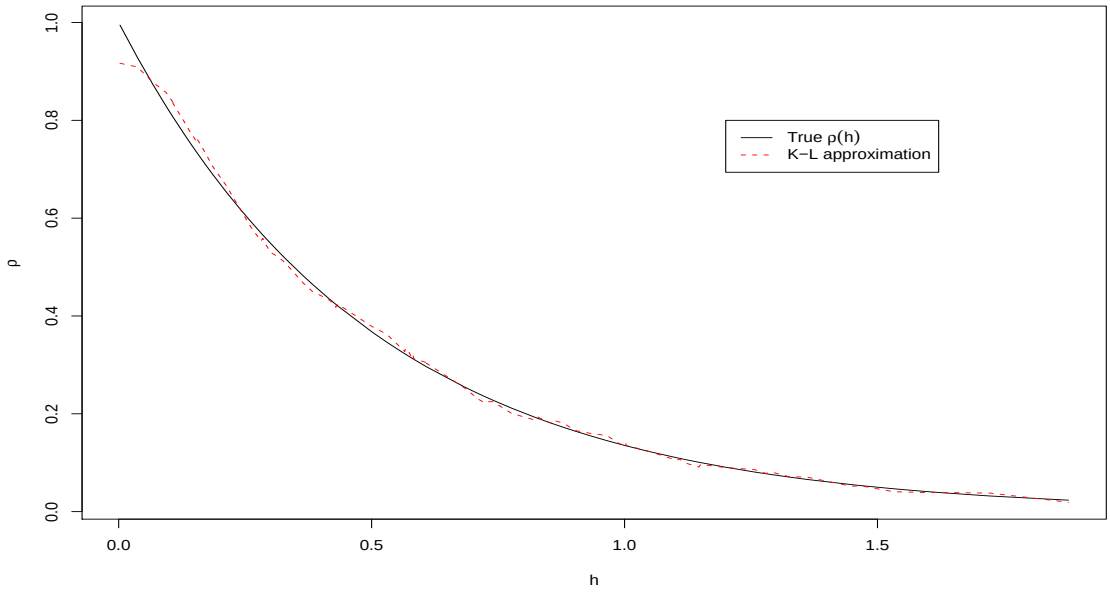

Figure 3.5: Karhunen-Loève expansion approximation for the exponential correlation function on $[-1, 1]$: $\varrho(s, t) = \exp(-2|s - t|)$.

ing eigenfunctions

$$\lambda_j = f(j\pi/(2L)) \quad \text{and} \quad \psi_j(t) = c\, e^{ij\pi t/(2L)}, \tag{3.37}$$

where $f(\cdot)$ is the spectral density associated with ϱ and c is a normalizing constant. The approximation is convenient since many common correlation functions have closed-form spectral densities. For example, for the exponential correlation function on $[-L, L]$, the approximated eigenvalues are $\lambda_j = f(j\pi/(2L)) = \frac{2\phi}{\phi^2+(0.5j\pi/L)^2}$ (compare with (3.30) and (3.31)). It can be seen from Figure 3.4 that $w_n \approx \frac{n\pi}{2L}$ when w_n is large. When L converges to infinity, w is more and more dense on $\mathbb{R}$ and eventually the eigenvalues become $f(w)$ for all $w \in \mathbb{R}$. For the Matérn covariance function on $[-L, L]$, the approximated eigenvalues are

$$\lambda_j = \frac{2\sqrt{\pi}\Gamma(\nu + 1/2)\phi^{2\nu}}{\Gamma(\nu)\{\phi^2 + (0.5j\pi/L)^2\}^{\nu+1/2}}. \tag{3.38}$$

3.4.6 Extension to $\mu(s) = X^T(s)\boldsymbol{\beta}$

For illustration we consider the model $Y(s) = \beta_0 + x(s)\beta_1 + W(s)$ (but the results below are routinely extended to handle a general covariate vector $X(s)$). The information matrix and the inverse information matrix for the mean parameters are

$$I(\beta_0, \beta_1) = \begin{pmatrix} \mathbf{1}^T R_n^{-1}\mathbf{1} & \mathbf{x}^T R_n^{-1}\mathbf{1} \\ \mathbf{x}^T R_n^{-1}\mathbf{1} & \mathbf{x}^T R_n^{-1}\mathbf{x} \end{pmatrix} \tag{3.39}$$

and

$$I^{-1}(\beta_0, \beta_1) = \frac{1}{|(\mathbf{1}^T R_n^{-1}\mathbf{1})(\mathbf{x}^T R_n^{-1}\mathbf{x}) - (\mathbf{x}^T R_n^{-1}\mathbf{1})^2|} \begin{pmatrix} \mathbf{x}^T R_n^{-1}\mathbf{x} & -\mathbf{x}^T R_n^{-1}\mathbf{1} \\ -\mathbf{x}^T R_n^{-1}\mathbf{1} & \mathbf{1}^T R_n^{-1}\mathbf{1} \end{pmatrix}, \tag{3.40}$$

where $\mathbf{x} = (x(s_1), ..., x(s_n))^T$. So

$$\mathrm{Var}(\hat{\beta}_0) \geq \frac{1}{|\mathbf{1}^T R_n^{-1}\mathbf{1} - (\mathbf{x}^T R_n^{-1}\mathbf{1})^2/\mathbf{x}^T R_n^{-1}\mathbf{x}|}.$$

By the Cauchy-Schwarz inequality, $\mathbf{1}^T R_n^{-1}\mathbf{1} \geq (\mathbf{x}^T R_n^{-1}\mathbf{1})^2/\mathbf{x}^T R_n^{-1}\mathbf{x}$. We may view

$$1/I(\beta_0, \beta_1)^{1,1} = \mathbf{1}^T R_n^{-1}\mathbf{1} - (\mathbf{x}^T R_n^{-1}\mathbf{1})^2/\mathbf{x}^T R_n^{-1}\mathbf{x}$$

as the "information" for β_0 when β_1 is unknown, where $I^{i,i}$ means the i^{th} diagonal element of the inverse information matrix I^{-1}. $0 < \mathbf{1}^T R_n^{-1}\mathbf{1} - (\mathbf{x}^T R_n^{-1}\mathbf{1})^2/\mathbf{x}^T R_n^{-1}\mathbf{x} < \mathbf{1}^T R_n^{-1}\mathbf{1} = I(\beta_0|\beta_1)$, in agreement with our intuition that the information for estimating β_0 decreases if we have a *non-orthgonal* nuisance parameter β_1. Thus we can conclude that $I(\beta_0)$ (when β_1 unknown) is bounded if $I(\beta_0|\beta_1) = \mathbf{1}^T R_n^{-1}\mathbf{1}$ is bounded.

Remark 5. *It can be shown that the conditional information is greater than unconditional information, i.e. $I_{i,i} \geq 1/I^{i,i}$. This follows since the information matrix I is the covariance matrix associated with the score vector. I^{-1} is the matrix such that the reciprocals of the diagonal entries are conditional variances for one score given all of the others. Since the conditional variance is always less than or equal to the marginal variance, we immediately have the inequality.*

Remark 6. *If $x(t)$ is bounded for $t \in D$ and $\mathbf{1}^T R_n^{-1}\mathbf{1}$ is bounded, $x(t)^T R_n^{-1} x(t)$ is bounded.*

3.4.7 Under expansion and middle-ground asymptotics

For estimating the mean, we have an *either-or* situation. From the discussion in Section 3.4.1, in general, we can consistently estimate the mean as long as the study region D increases without bound.

3.5 Unknown center and scale

3.5.1 Information and MLEs

Let us next investigate the model where the correlation function is assumed known for (3.1)-(3.2), but with unknown center μ and scale σ^2. Data therefore follow the model $\mathbf{Y} \sim \mathrm{N}_n(\mu\mathbf{1}, \sigma^2 R_n)$, with log-likelihood

$$\ell_n(\mu,\sigma) = -n\log\sigma - \tfrac{1}{2}\log|R_n| - \tfrac{1}{2}(\mathbf{y}-\mu\mathbf{1})^T R_n^{-1}(\mathbf{y}-\mu\mathbf{1})\sigma^2, \tag{3.41}$$

with R_n known. One finds

$$\frac{\partial \ell_n}{\partial \mu} = (1/\sigma^2)(\mathbf{1}^T R_n^{-1}\mathbf{y} - A_n\mu),$$
$$\frac{\partial \ell_n}{\partial \sigma} = -n/\sigma + (1/\sigma^3)(\mathbf{y}-\mu\mathbf{1})^T R_n^{-1}(\mathbf{y}-\mu\mathbf{1})$$

yielding the ML estimators as expected, namely

$$\widehat{\mu} = \frac{\mathbf{1}^T R_n^{-1}\mathbf{y}}{\mathbf{1}^T R_n^{-1}\mathbf{1}} \quad \text{and} \quad \widehat{\sigma} = \{(\mathbf{y}-\widehat{\mu}\mathbf{1})^T R_n^{-1}(\mathbf{y}-\widehat{\mu}\mathbf{1})/n\}^{1/2}. \tag{3.42}$$

Also, the information matrix becomes

$$I_n(\mu,\sigma) = \begin{pmatrix} A_n(\mu)/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{pmatrix}, \tag{3.43}$$

revealing that information is bounded in μ but *not* in σ. The ML estimators match the Cramér-Rao lower bound $I_n(\mu,\sigma)^{-1}$ (modulo the usual $(n-1)/n$ factor for $\widehat{\sigma}^2$). To see this, let us write

$$\mathbf{Y} = \mu\mathbf{1} + \sigma R_n^{1/2}\boldsymbol{\epsilon} \quad \text{in terms of } \boldsymbol{\epsilon} \sim \mathrm{N}_n(0, \mathrm{I}_n), \text{ where } \mathrm{I}_n \text{ is an } n \times n \text{ identity matrix.}$$

Then

$$\widehat{\mu} = \mu + \sigma\frac{\mathbf{1}^T R_n^{-1/2}\boldsymbol{\epsilon}}{A_n} \quad \text{and} \quad \widehat{\mu}\mathbf{1} = \mu\mathbf{1} + \sigma\frac{\mathbf{1}\mathbf{1}^T R_n^{-1/2}}{A_n}\boldsymbol{\epsilon}, \tag{3.44}$$

leading to

$$Y - \widehat{\mu}\mathbf{1} = \sigma(R_n^{1/2} - \mathbf{1}\mathbf{1}^T R_n^{-1/2}/A_n)\boldsymbol{\epsilon}$$

and

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{\sigma^2}{n}\boldsymbol{\epsilon}^T\Big(R_n^{1/2} - \frac{R_n^{-1/2}\mathbf{1}\mathbf{1}^T}{A_n}\Big)R_n^{-1}\Big(R_n^{1/2} - \frac{\mathbf{1}\mathbf{1}^T R_n^{-1/2}}{A_n}\Big)\boldsymbol{\epsilon} \\
&= \frac{\sigma^2}{n}\boldsymbol{\epsilon}^T\Big(\mathrm{I}_n - \frac{R_n^{-1/2}\mathbf{1}\mathbf{1}^T R_n^{-1/2}}{A_n}\Big)\boldsymbol{\epsilon}.
\end{aligned}$$

This implies independence between the two estimators and

$$\widehat{\mu} \sim N(\mu, \sigma^2/A_n), \tag{3.45}$$

$$\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-1}/n. \tag{3.46}$$

3.5.2 Bayesian caveat

Again there is a warning for Bayesians coming out of this; parts of any prior for (μ,σ) will be retained in the posterior distribution, even with infinitely many data points on a bounded study domain.

Suppose we specify a *normal inverse-gamma* prior $\pi(\mu, \sigma^2) = \mathrm{N}(\mu_0, 1/A_0)\mathrm{IG}(\alpha_0, \beta_0)$ for mean and scale parameters. Then the posterior is

$$\begin{aligned}\pi(\mu, \sigma^2|\text{data}) &\propto e^{-\frac{A_0}{2}(\mu-\mu_0)^2}(\sigma^2)^{-(\alpha_0+1)}e^{-\beta_0/\sigma^2}|2\pi\sigma^2 R_n|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{y}-\mu\mathbf{1})^T R_n^{-1}(\mathbf{y}-\mu\mathbf{1})/\sigma^2} \\ &\propto e^{-\frac{A^*}{2}[(\mu-\mu^*)^2-\mu^{*2}+\mathbf{y}^T R_n^{-1}\mathbf{y}/(A^*\sigma^2)+\mu_0 A_0/A^*]-\beta_0/\sigma^2}(\sigma^2)^{-(\alpha_0+1+n/2)},\end{aligned}$$

where $\mu^* = \frac{\mathbf{y}^T R_n^{-1}\mathbf{1}/\sigma^2+\mu_0 A_0}{\mathbf{1}^T R_n^{-1}\mathbf{1}/\sigma^2+A_0}$ and $A^* = \mathbf{1}^T R_n^{-1}\mathbf{1}/\sigma^2 + A_0$. The conditional densities have closed form expressions:

$$\mu|\sigma^2, \text{data} \sim \mathrm{N}(\mu^*, 1/A^*), \tag{3.47}$$

$$\sigma^2|\mu, \text{data} \sim \mathrm{IG}(\alpha_0 + n/2, \beta_0 + \frac{1}{2}(\mathbf{y} - \mu\mathbf{1})^T R_n^{-1}(\mathbf{y} - \mu\mathbf{1})). \tag{3.48}$$

This leads

$$\mathrm{E}(\mu|\sigma^2) = (\mu_0 A_0 + \hat{\mu}_n A_n/\sigma^2)/(A_0 + A_n/\sigma^2) \text{ and} \tag{3.49}$$

$$\mathrm{E}(\sigma^2|\mu) = \frac{\beta_0 + (\mathbf{y} - \mu\mathbf{1})^T R_n^{-1}(\mathbf{y} - \mu\mathbf{1})/2}{\alpha_0 + n/2 - 1} \approx \frac{(\mathbf{y} - \mu\mathbf{1})^T R_n^{-1}(\mathbf{y} - \mu\mathbf{1})}{n} \tag{3.50}$$

for large n. It is easy to see that for large n, σ forgets where it comes from, but not μ.

3.6 Known μ, σ: estimating ϕ only

Suppose both the mean and standard deviation are known in our standard model. How well can we estimate the parameters of the correlation function? Without loss of generality, we can take the model to be of the form

$$\mathbf{Y} \sim \mathrm{N}_n(0, R_n(\boldsymbol{\phi})), \tag{3.51}$$

with $R_n(\boldsymbol{\phi})$ having elements $\varrho(|t_i - t_j|; \boldsymbol{\phi})$, corresponding to the process Y having been observed at positions $t_1, \ldots, t_n$.

3.6.1 Information calculations

In the following, let $\boldsymbol{\phi}$ be a scalar for now and let

$$B_n = \mathrm{Tr}(R_n^{-1}(\phi)\bar{R}_n(\phi)), \tag{3.52}$$

$$E_n = \frac{1}{2}\mathrm{Tr}(R_n^{-1}(\phi)\bar{R}_n(\phi)R_n^{-1}(\phi)\bar{R}_n(\phi)), \tag{3.53}$$

where $\bar{R}_n(\phi)$ is the matrix with entries, $\partial R_{n,i,j}(\phi)/\partial\phi$.

For the OU case with equidistance sampling on the interval $[0, L]$ (data are observed at $t_i = iL/n$ for $i = 1, ..., n$), we have (see Appendix A.1)

$$\begin{aligned} B_n &= 2(n-1)\frac{\rho^2}{1-\rho^2}\Delta, \\ E_n &= (n-1)\frac{\rho^2(1+\rho^2)}{(1-\rho^2)^2}\Delta^2, \end{aligned} \tag{3.54}$$

where $\rho = \exp(-\phi\Delta)$ and $\Delta = L/n$.

The log-likelihood function for model (3.51) is

$$\ell_n(\phi) = -\tfrac{1}{2}\log|R_n(\phi)| - \tfrac{1}{2}\mathbf{y}^T R_n(\phi)^{-1}\mathbf{y}, \tag{3.55}$$

with consequent

$$V_n = (\partial/\partial\phi)\ell_n = -\tfrac{1}{2}B_n + \tfrac{1}{2}\mathbf{y}^T R_n^{-1}(\phi)\bar{R}_n(\phi)R_n^{-1}(\phi)\mathbf{y}.$$

The ML estimator $\widehat{\phi}$ solves the equation

$$\mathbf{y}^T R_n(\phi)^{-1}\bar{R}_n(\phi)R_n(\phi)^{-1}\mathbf{y} = B_n(\phi).$$

Writing $\mathbf{Y} = R_n^{1/2}(\phi)\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}_n(0, \mathrm{I}_n)$, the random score function may be expressed as

$$V_n = \tfrac{1}{2}(\boldsymbol{\epsilon}^T G_n \boldsymbol{\epsilon} - B_n), \text{ with } G_n = R_n^{-1/2}(\phi)\bar{R}_n(\phi)R_n^{-1/2}(\phi). \tag{3.56}$$

It has mean zero and variance $I_n(\phi) = E_n$.

3.6.2 Asymptotic normality?

Suppose that information increases linearly with n, and that $n^{-1}E_n \to E$, say. The usual approximation argument, when it works, produces

$$\sqrt{n}(\widehat{\phi} - \phi) \doteq (n^{-1}\mathrm{I}_n(\phi))^{-1} n^{-1/2} V_n \to_d \mathrm{N}(0, 1/E), \tag{3.57}$$

where $\to_d$ means convergence in distribution.

For the OU case, for example,

$$\begin{aligned} n^{-1}E_n &= \frac{n-1}{n}\rho^2(1+\rho^2)\frac{\Delta^2}{(1-\rho^2)^2} \\ &\doteq \rho^2(1+\rho^2)\Big(\frac{\Delta}{2\phi\Delta}\Big)^2 \to \frac{1}{2\phi^2}. \end{aligned}$$

This holds, interestingly, under (and only under) the minimal requirement $\Delta \to 0$, i.e. both for infill and for middle-ground asymptotics. Thus we expect $\sqrt{n}(\widehat{\phi} - \phi) \to_d \mathrm{N}(0, 2\phi^2)$ for the OU process, as long as the sampling distance $\Delta \to 0$, regardless of the speed with which this happens.

A crucial ingredient in (3.57) is that $n^{-1/2}V_n \to_d \mathrm{N}(0, E)$. Let Q_n be a unitary matrix with $Q_n G_n Q_n^T = \Lambda_n$, the diagonal matrix with entries $\lambda_1, \ldots, \lambda_n$ equal to the eigenvalues of G_n. Then $G_n = Q_n^T \Lambda_n Q_n$ and $V_n = \frac{1}{2}\sum_{i=1}^n \lambda_i(N_i^2 - 1)$, in terms of independent standard normals $N_1, N_2, \ldots$. Note that $B_n = \sum_{i=1}^n \lambda_i$ and $E_n = \frac{1}{2}\sum_{i=1}^n \lambda_i^2$. It is now not difficult to demonstrate, via characteristic functions, that $n^{-1/2}V_n \to_d \mathrm{N}(0, E)$ if and only if

$$n^{-1}E_n = \frac{1}{2}n^{-1}\sum_{i=1}^n \lambda_i^2 \to E \quad \text{and} \quad n^{-1/2}\max_{i \le n}|\lambda_i| \to 0. \tag{3.58}$$

This is a version of the Lindeberg-Feller conditions applied to an infill setting.

3.6.3 Decay parameters versus smoothness parameters

Inference for the decay parameter in a covariance function differs dramatically from that for a smoothness parameter. Consider the power exponential correlation function

$$\varrho(h;\phi) = \exp(-\phi h^{\gamma}) \quad \text{for } h \geq 0, \tag{3.59}$$

where $\gamma \in (0,2)$ is viewed as a fixed parameter, reflecting the amount of smoothness in our random curves. The derivative of ϱ, used to compute $\bar{R}_n(\phi)$ and thus, G_n and its eigenvalues, is $\varrho'(h;\phi) = -h^{\gamma}\exp(-\phi h^{\gamma})$. Numerical work does suggest that (3.58) and hence (3.57) holds, for each $\gamma \in (0,2)$.

Suppose, in (3.59) that we consider ϕ fixed (and equal to 1 without loss of generality) with γ unknown in $(0,2)$, i.e.,

$$\varrho(h;\gamma) = \exp(-h^{\gamma}) \quad \text{for } h \geq 0. \tag{3.60}$$

Here B_n and in particular E_n explode very rapidly with n, seemingly with each given $\gamma \in (0,2)$, and the eigenvalues are not evenly enough distributed for limiting normality to hold. It is possible that $k_n(\widehat{\gamma} - \gamma) \to_d$ some limit for some k_n sequence that is much more rapid than $\sqrt{n}$. Note that $\varrho'(h;\gamma) = -\gamma h^{\gamma-1}\exp(-h^{\gamma})$, but that the diagonal elements of $\bar{R}_n$ are defined as 0 in general, also here, when $\gamma < 1$. We expect similar non-traditional behavior of information and estimators when one attempts to estimate the Matérn smoothness parameter.

One might also consider the more ambitious task of estimating both the decay and smoothness parameters in the Matérn correlation function or ϕ and γ in (3.59).

We find a two-dimensional score vector

$$V_n = \frac{\partial \ell_n(\phi)}{\partial \phi} = \begin{pmatrix} -\frac{1}{2}B_{n,1} + \frac{1}{2}\mathbf{y}^T R_n^{-1}\bar{R}_n^{(1)}R_n^{-1}\mathbf{y} \\ -\frac{1}{2}B_{n,2} + \frac{1}{2}\mathbf{y}^T R_n^{-1}\bar{R}_n^{(2)}R_n^{-1}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\boldsymbol{\epsilon}^T R_n^{-1/2}\bar{R}_n^{(1)}R_n^{-1/2}\boldsymbol{\epsilon} - B_{n,1}) \\ \frac{1}{2}(\boldsymbol{\epsilon}^T R_n^{-1/2}\bar{R}_n^{(2)}R_n^{-1/2}\boldsymbol{\epsilon} - B_{n,2}) \end{pmatrix},$$

where

$$B_{n,j} = \mathrm{Tr}(R_n^{-1}\bar{R}_n^{(j)}) \quad \text{and} \quad \bar{R}_n^{(j)} = (\partial/\partial\phi_j)R_n.$$

The information matrix may be expressed as

$$I_n(\phi) = \operatorname{Var} V_n = E_n = \begin{pmatrix} E_{n,1,1} & E_{n,1,2} \\ E_{n,1,2} & E_{n,2,2} \end{pmatrix},$$

where

$$E_{n,i,j} = \operatorname{Tr}(R_n^{-1}\bar{R}_n^{(i)}R_n^{-1}\bar{R}_n^{(j)}) \quad \text{for } i,j = 1,2.$$

Limiting behaviour of both information and distribution of estimations would essentially be determined by the eigenvalues of respectively

$$G_n^{(1)} = R_n^{-1/2}\bar{R}_n^{(1)}R_n^{-1/2} \quad \text{and} \quad G_n^{(2)} = R_n^{-1/2}\bar{R}_n^{(2)}R_n^{-1/2}.$$

3.7 All parameters unknown

Now let us return to the general model

$$\mathbf{Y} \sim \mathrm{N}_n(\mu\mathbf{1}, \sigma^2 R_n(\boldsymbol{\phi})) \tag{3.61}$$

with one or more unknown parameters in the correlation function as well as unknown mean μ and variance σ^2.

3.7.1 Basic information calculations

The log-likelihood function is as in (3.41), but with ϕ (assume $\boldsymbol{\phi}$ is one dimensional for now) present in R_n. One finds a score function with components

$$\begin{aligned}
\frac{\partial \ell_n}{\partial \mu} &= U_n = (1/\sigma^2)(\mathbf{1}^T R_n^{-1}\mathbf{y} - A_n\mu), \\
\frac{\partial \ell_n}{\partial \sigma} &= V_n = -n/\sigma + (1/\sigma^3)(\mathbf{y}-\mu\mathbf{1})^T R_n^{-1}(\mathbf{y}-\mu\mathbf{1}), \\
\frac{\partial \ell_n}{\partial \phi} &= W_n = -\tfrac{1}{2}\operatorname{Tr}(R_n^{-1}\bar{R}_n) + \tfrac{1}{2}(\mathbf{y}-\mu\mathbf{1})^T R_n^{-1}\bar{R}_n R_n^{-1}(\mathbf{y}-\mu\mathbf{1})/\sigma^2,
\end{aligned}$$

with $\bar{R}_n(\phi)$ as above.

To find ML estimators numerically, one maximizes the profile log-likelihood function

$$\ell_{n,\mathrm{prof}}(\phi) = -n \log \widehat{\sigma}(\phi) - \tfrac{1}{2} \log |R_n(\phi)| \tag{3.62}$$

with respect to ϕ, where

$$\widehat{\sigma}(\phi) = \left[\{\mathbf{y} - \widehat{\mu}(\phi)\mathbf{1}\}^T R_n(\phi)^{-1} \{\mathbf{y} - \widehat{\mu}(\phi)\mathbf{1}\}/n \right]^{1/2} \quad \text{and} \quad \widehat{\mu}(\phi) = \frac{\mathbf{1}^T R_n(\phi)^{-1} \mathbf{y}}{\mathbf{1}^T R_n(\phi)^{-1} \mathbf{1}}. \tag{3.63}$$

We write $\mathbf{Y} = \mu\mathbf{1} + \sigma R_n(\phi)^{1/2} \boldsymbol{\epsilon}$ in terms of a vector $\boldsymbol{\epsilon}$ of independent standard normals. One finds

$$\begin{aligned} U_n &= \mathbf{1}^T R_n^{-1} \boldsymbol{\epsilon}/\sigma, \\ V_n &= (-n + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon})/\sigma, \\ W_n &= -\tfrac{1}{2} \mathrm{Tr}(R_n^{-1} \dot{R}_n) + \tfrac{1}{2} \boldsymbol{\epsilon}^T R_n^{-1/2} \dot{R}_n R_n^{-1/2} \boldsymbol{\epsilon}. \end{aligned}$$

As a special case of (3.5), the information matrix is

$$I_n(\mu, \sigma, \phi) = \begin{pmatrix} A_n/\sigma^2 & 0 & 0 \\ 0 & 2n/\sigma^2 & B_n/\sigma \\ 0 & B_n/\sigma & E_n \end{pmatrix} \tag{3.64}$$

where A_n, B_n, and E_n are as before (with ϕ suppressed)

The inverse information matrix is

$$I^{-1}(\mu, \sigma, \phi) = \begin{pmatrix} \frac{\sigma^2}{A_n} & 0 & 0 \\ 0 & \frac{E_n \sigma^2}{2nE_n - B_n^2} & -\frac{B_n \sigma}{2nE_n - B_n^2} \\ 0 & -\frac{B_n \sigma}{2nE_n - B_n^2} & \frac{2n}{2nE_n - B_n^2} \end{pmatrix}. \tag{3.65}$$

From the Cramér-Rao information inequality, which applies here,

$$\mathrm{Var}\, \widehat{\sigma} \geq \frac{E_n}{2nE_n - B_n^2} \sigma^2 \quad \text{and} \quad \mathrm{Var}\, \widehat{\phi} \geq \frac{2n}{2nE_n - B_n^2}. \tag{3.66}$$

3.7.2 Information for the OU model

For the OU process on $[0, L]$ with equidistance sampling (i.e., Y are observed at at $t_i = iL/n$ for $i = 1, ..., n$), we have (see Appendix A.1):

$$A_n = \frac{(n-1)(1-\rho)}{1+\rho} + 1 \doteq 1 + \frac{L\phi}{2}, \tag{3.67}$$

$$B_n = \frac{2(n-1)\rho^2}{1-\rho^2}\Delta \doteq \frac{n-1}{\phi} - 2L, \tag{3.68}$$

$$E_n = \frac{(n-1)\rho^2(1+\rho^2)}{(1-\rho^2)^2}\Delta^2 \doteq \frac{n-1}{2\phi^2} - \frac{3L}{2\phi}. \tag{3.69}$$

We also can compute the entries in the inverse information matrix (3.65):

$$\frac{E_n\sigma^2}{2nE_n - B_n^2} = \frac{(1+\rho^2)}{2n - 2(n-2)\rho^2}\sigma^2 \doteq \frac{\sigma^2}{2(1+L\phi)}, \tag{3.70}$$

$$-\frac{B_n\sigma}{2nE_n - B_n^2} = -\frac{1-\rho^2}{\Delta(n - n\rho^2 + 2\rho^2)}\sigma \doteq -\frac{\sigma\phi}{1+L\phi}, \tag{3.71}$$

$$\frac{2n}{2nE_n - B_n^2} = \frac{n(1-\rho^2)^2}{\Delta^2(n-1)(n\rho^2 - n\rho^4 + 2\rho^4)} \doteq \frac{2\phi^2}{1+L\phi}. \tag{3.72}$$

So we have

$$I(\mu, \sigma, \phi) \doteq \begin{pmatrix} \frac{1+L\phi/2}{\sigma^2} & 0 & 0 \\ 0 & \frac{2n}{\sigma^2} & \frac{n-2L\phi}{\phi\sigma} \\ 0 & \frac{n-2L\phi}{\phi\sigma} & \frac{n-3L\phi}{2\phi^2} \end{pmatrix} \text{ and} \tag{3.73}$$

$$I^{-1}(\mu, \sigma, \phi) \doteq \begin{pmatrix} \frac{\sigma^2}{1+L\phi/2} & 0 & 0 \\ 0 & \frac{\sigma^2}{2(1+L\phi)} & -\frac{\sigma\phi}{1+L\phi} \\ 0 & -\frac{\sigma\phi}{1+L\phi} & \frac{2\phi^2}{1+L\phi} \end{pmatrix}. \tag{3.74}$$

By the Cramér-Rao information inequality and the inverse information matrix

(3.74),

$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &\geq \frac{\sigma^2}{1+L\phi/2} > 0, \quad \mathrm{Var}(\hat{\sigma}) \geq \frac{\sigma^2}{2(1+L\phi)} > 0 \quad \text{and} \\ \mathrm{Var}(\hat{\phi}) &\geq \frac{2\phi^2}{1+L\phi} > 0. \end{aligned} \tag{3.75}$$

This indicates that none of (μ, ϕ, σ) can be estimated consistently.

The asymptotic correlation between $\hat{\sigma}$ and $\hat{\phi}$ is

$$\begin{aligned} \mathrm{Corr}(\hat{\sigma}, \hat{\phi}) &= -\frac{2\sigma B_n/(2nE_n - B_n^2)}{\sqrt{\frac{\sigma^2 E_n}{(2nE_n - B_n^2)}\frac{2n}{(2nE_n - B_n^2)}}} \\ &= -B_n/\sqrt{2nE_n} \\ &= -\sqrt{2}\rho/\sqrt{n(1+\rho^2)/(n-1)} \doteq -1. \end{aligned} \tag{3.76}$$

The implications of this result are well-known more generally in terms of essentially a *ridge* in the likelihood surface and *drifting* of MCMC algorithms due to the weak identifiability.

Although ϕ and σ cannot be consistently estimated separately, it might be that some new parameter $\eta(\sigma, \phi)$ could be estimated consistently. In particular, consider the information matrix for the parametrization $\boldsymbol{\lambda} = (\mu, \sigma, \sigma\sqrt{\phi})^T$. Recall that, if $\boldsymbol{\lambda} = g^{-1}(\boldsymbol{\theta})$ with $\dim(\boldsymbol{\lambda}) = s \leq \dim(\boldsymbol{\theta}) = p$, then,

$$I(\boldsymbol{\lambda}) = \begin{pmatrix} \frac{\partial g_1(\boldsymbol{\lambda})}{\partial \lambda_1} & \cdots & \frac{\partial g_p(\boldsymbol{\lambda})}{\partial \lambda_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_1(\boldsymbol{\lambda})}{\partial \lambda_s} & \cdots & \frac{\partial g_p(\boldsymbol{\lambda})}{\partial \lambda_s} \end{pmatrix} \Big(I(g(\boldsymbol{\lambda})) \Big) \begin{pmatrix} \frac{\partial g_1(\boldsymbol{\lambda})}{\partial \lambda_1} & \cdots & \frac{\partial g_p(\boldsymbol{\lambda})}{\partial \lambda_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_1(\boldsymbol{\lambda})}{\partial \lambda_s} & \cdots & \frac{\partial g_p(\boldsymbol{\lambda})}{\partial \lambda_s} \end{pmatrix}^T . \tag{3.77}$$

So,

$$
\begin{aligned}
I(\mu,\sigma,\sigma\sqrt{\phi}) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -2\lambda_3^2\lambda_2^{-3} \\ 0 & 0 & 2\lambda_3\lambda_2^{-2} \end{pmatrix} \begin{pmatrix} \frac{A_n}{\sigma^2} & 0 & 0 \\ 0 & \frac{2n}{\sigma^2} & \frac{B_n}{\sigma} \\ 0 & \frac{B_n}{\sigma} & E_n \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -2\lambda_3^2\lambda_2^{-3} \\ 0 & 0 & 2\lambda_3\lambda_2^{-2} \end{pmatrix}^T \\
&= \begin{pmatrix} \frac{A_n}{\sigma^2} & 0 & 0 \\ 0 & \frac{2n-4B_n\phi+4E_n\phi^2}{\sigma^2} & \frac{2B_n\sqrt{\phi}-4E_n\phi^{\frac{3}{2}}}{\sigma^2} \\ 0 & \frac{2B_n\sqrt{\phi}-4E_n\phi^{\frac{3}{2}}}{\sigma^2} & \frac{4E_n\phi}{\sigma^2} \end{pmatrix} \\
&\doteq \begin{pmatrix} \frac{1+L\phi/2}{\sigma^2} & 0 & 0 \\ 0 & \frac{2L\phi+2}{\sigma^2} & \frac{2L\sqrt{\phi}}{\sigma^2} \\ 0 & \frac{2L\sqrt{\phi}}{\sigma^2} & \frac{2(n+1)/\phi-6L}{\sigma^2} \end{pmatrix}. \qquad (3.78)
\end{aligned}
$$

Note that the information for estimating σ is bounded given $\sigma\sqrt{\phi}$. It is different from the situation of the information for σ given ϕ which, from formula (3.73), is unbounded. The inverse of the new information matrix is

$$
\begin{aligned}
I^{-1}(\mu,\sigma,\sigma\sqrt{\phi}) &= \begin{pmatrix} \frac{\sigma^2}{A_n} & 0 & 0 \\ 0 & \frac{\sigma^2 E_n}{2nE_n-B_n^2} & \frac{\sigma^2}{2\sqrt{\phi}}\frac{2E_n\phi-B_n}{2nE_n-B_n^2} \\ 0 & \frac{\sigma^2}{2\sqrt{\phi}}\frac{2E_n\phi-B_n}{2nE_n-B_n^2} & \frac{\sigma^2(n-2B_n\phi+2E_n\phi^2)}{2\phi(2nE_n-B_n^2)} \end{pmatrix} \\
&\doteq \begin{pmatrix} \frac{\sigma^2}{1+L\phi/2} & 0 & 0 \\ 0 & \frac{\sigma^2}{2(1+L\phi)} & O(\frac{1}{n}) \\ 0 & O(\frac{1}{n}) & O(\frac{1}{n}) \end{pmatrix} \qquad (3.79)
\end{aligned}
$$

which indicates that $\sigma\sqrt{\phi}$ can be estimated consistently.

For the OU model, in a general sampling setting, say, on $[0, L_n]$ with $L_n = Ln^{1-\gamma}$ and $\Delta_n = L/n^\gamma$ for $\gamma \in (0,1)$, $\mathrm{Corr}_n(\hat{\sigma},\hat{\phi})^2 = 2(n-1)\rho_n^2/(n+n\rho_n^2)$ which goes to 1 if and only if $\Delta_n \to 0$. So the bivariate distribution of $(\hat{\sigma},\hat{\phi})$ collapses to one dimension under both infill and middle-ground asymptotics. In this regard, the approximation $\mathrm{Corr}_n^2 \doteq 1-\phi\Delta_n$ for small Δ_n suggests potential problems for autoregressive time series models when Δ_n is small and adjacent correlation is high.

Next,

$$nE_n - B_n^2 = 2n(n-1)\frac{\rho_n^2}{(1-\rho_n^2)^2}\Delta_n^2\{1-\rho_n^2+(2/n)\rho_n^2\}.$$

For $\Delta_n \to 0$, $nE_n - B_n^2 \doteq n^2\phi^{-1}\Delta_n$. Along with $B_n \doteq n/\phi$ and $E_n \doteq n/(2\phi^2)$, valid as good approximations as long as $\Delta_n \to 0$, we obtain

$$I_n^{-1}(\sigma,\phi) \doteq \frac{1}{n\Delta_n}\begin{pmatrix} \sigma^2/(2\phi) & -\sigma \\ -\sigma & 2\phi \end{pmatrix}.$$

This means that for infill asymptotics, where $n\Delta_n$ is bounded away from zero and infinity, neither σ nor ϕ can be estimated consistently.

3.7.3 Beyond OU

Abt and Welch (1998) studied the behavior of the information matrix under the triangular, exponential, and Gaussian covariance functions. For a general covariance function, closed form expressions for A_n, B_n, and E_n will not be available. By simulation, Figures 3.6-3.9 reveal the behavior of the information quantities (see formulas (3.64) and (3.65)) for several customary covariance functions, including the power exponential, Matérn, Gaussian, and Cauchy.

Again, the challenge is to characterize exactly which parametric functions $\eta(\sigma^2, \boldsymbol{\phi})$ have the property of increasing information under infill. Here, we have the known results using equivalence/orthogonality of Gaussian measures. For example, two Gaussian measures $P(0, C_i)$ $(i = 1, 2)$ with $C_i(s,t) = \sigma_i^2 \exp(-\phi_i|s-t|)$ $(s, t \in [0, L])$ are equivalent if and only $\sigma_1\sqrt{\phi_1} = \sigma_2\sqrt{\phi_2}$ (Ibragimov and Rozanov, 1978, Chapter 3). So in the exponential covariance function case, the new parameter $\sigma\sqrt{\phi}$ determines the equivalence of the measures so it can be consistently estimated.

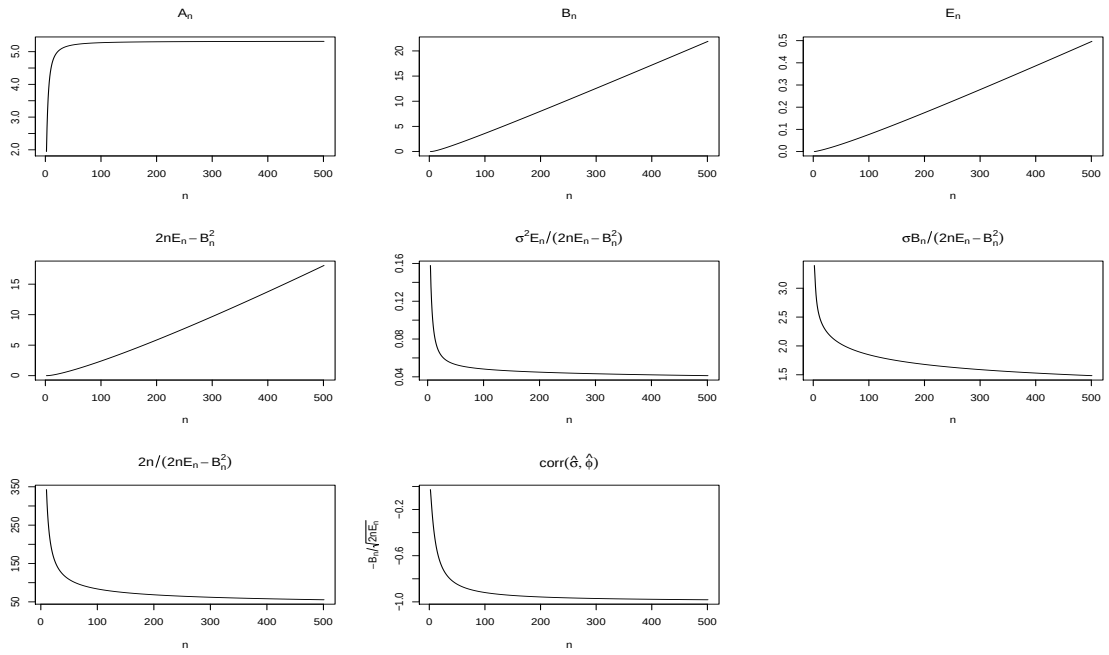

Figure 3.6: Information for power exponential covariance model: $\sigma^2 \exp(-|\phi h|^{0.5})$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 10)$.

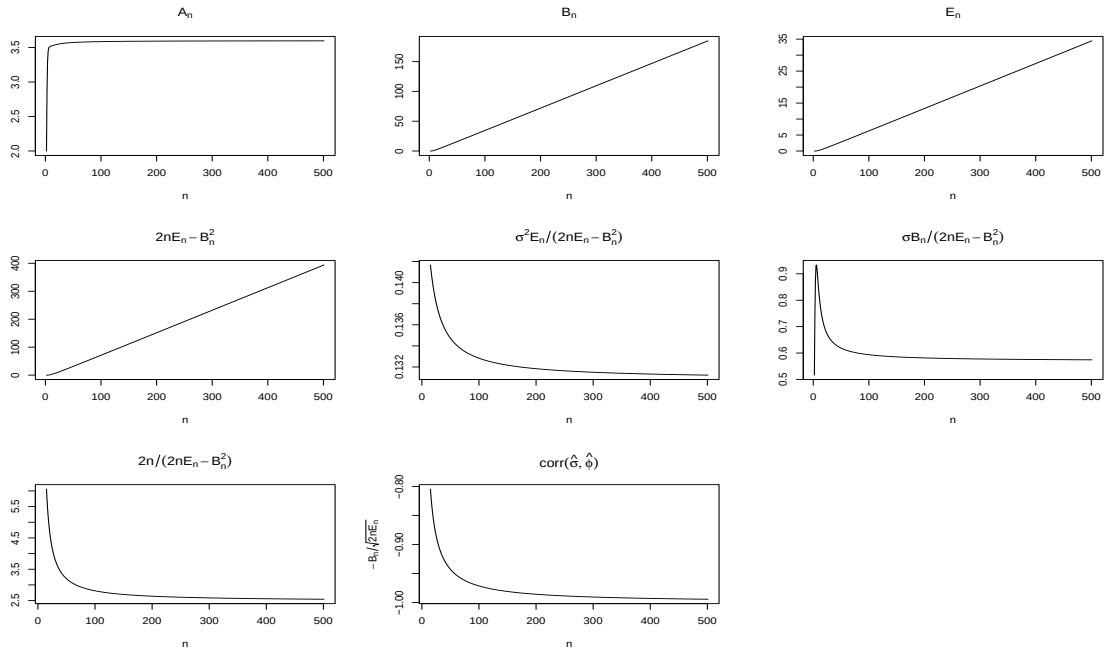

Figure 3.7: Information for Matérn covariance model when $\nu = 1.5$: $\sigma^2(1 + \phi|h|)\exp(-\phi|h|)$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 8)$.

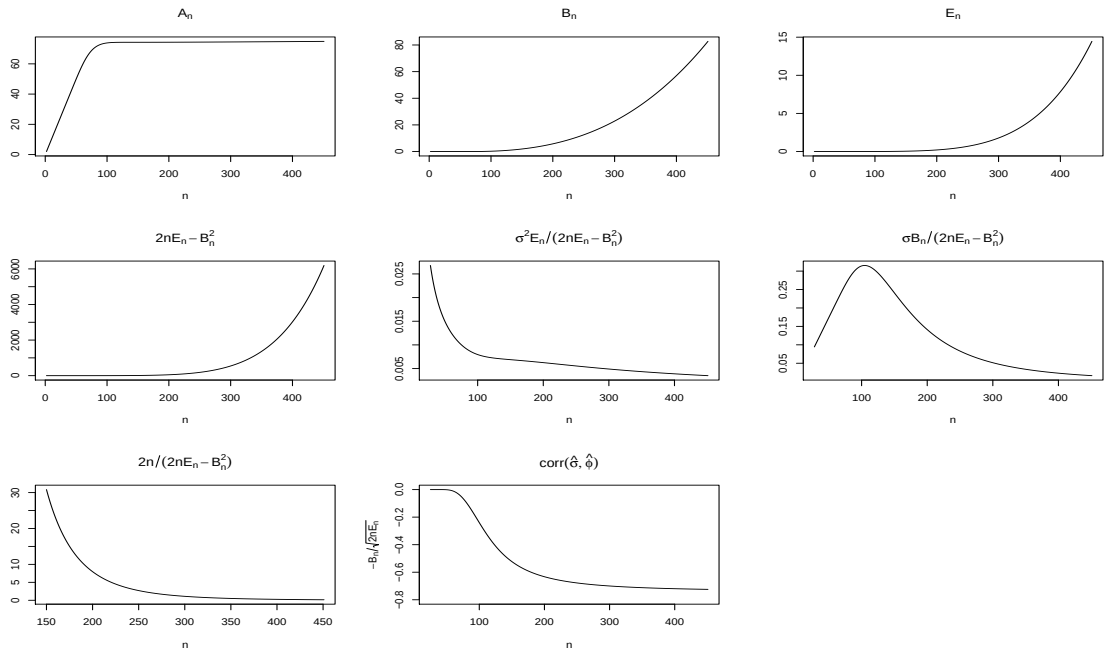

Figure 3.8: Information for Gaussian covariance model: $\sigma^2 \exp\{-(\phi h)^2\}$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 100)$.

More generally, there are well established explicit conditions expressed in terms of the spectral densities (indexed by same parameters as in the corresponding covariance functions) for us to check equivalent Gaussian measures (see e.g. Yadrenko, 1983, Chapter 3). From those results, we can see which reparametrization will make the new parameters to be estimated consistently. One of the easy-to-verify conditions is provided by Stein (2004) which is a modified version of Skorokhod and Yadrenko (1973, Theorem 4).

Theorem 6 (Stein, 2004, Theorem A.1)**.** *$P(0, K_1) \equiv P(0, K_2)$ on any bounded $D \subset$*

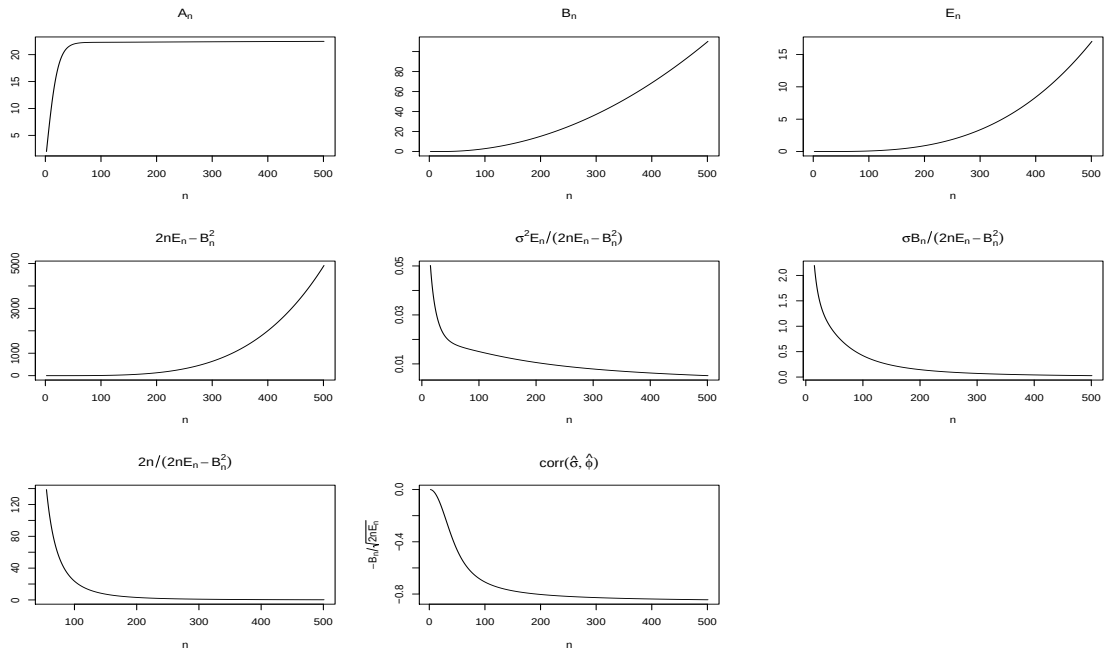

Figure 3.9: Information for Cauchy covariance function: $\sigma^2/\{1 + (\phi h)^2\}$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 50)$.

$\mathbb{R}^d$ *if for some finite* c, $\int_{|\omega|>c}(\frac{f_2(\omega)-f_1(\omega)}{f_1(\omega)})^2 < \infty$ *and* $f_1(\omega)|\omega|^\alpha$ *is bounded away from 0 and* ∞ *as* $|\omega| \to \infty$ *for some* $\alpha > d$*, where* f_i *is spectral density of* K_i*.*

Using the above theorem, Zhang (2004) shows that for Matérn class

$$C(h; \sigma^2, \phi, \nu) = \sigma^2(\alpha h)^\nu \kappa_\nu(\alpha h)/(\Gamma(\nu)2^{\nu-1}),$$

two measures corresponding to $(\sigma_1^2, \phi_1, \nu)$ and $(\sigma_2^2, \phi_2, \nu)$ are equivalent if and only if $\sigma_1\phi_1^\nu = \sigma_2\phi_2^\nu$.

As an *ad hoc* strategy, from the likelihood perspective, if some $\eta(\sigma^2, \boldsymbol{\phi})$ can be consistently estimated under infill, the MLE's will be clustered around a contour of some constant η. By simulation, we may attempt to deduce a functional form for $\eta(\sigma^2, \boldsymbol{\phi})$. Figure 3.10 provides an example to illustrate this idea. In this case, the data are generated from a Gaussian process with an exponential covariance function. For each simulation, we find the MLEs for the parameters and we find the functional

form $\eta = \sigma^2\phi^{0.96}$ which is close to the theoretical form $\sigma^2\phi$. A Matérn example is shown in Figure 3.11 and we find $\eta = \sigma^2\phi^{2.76}$, not too far from $\sigma^2\phi^3$.

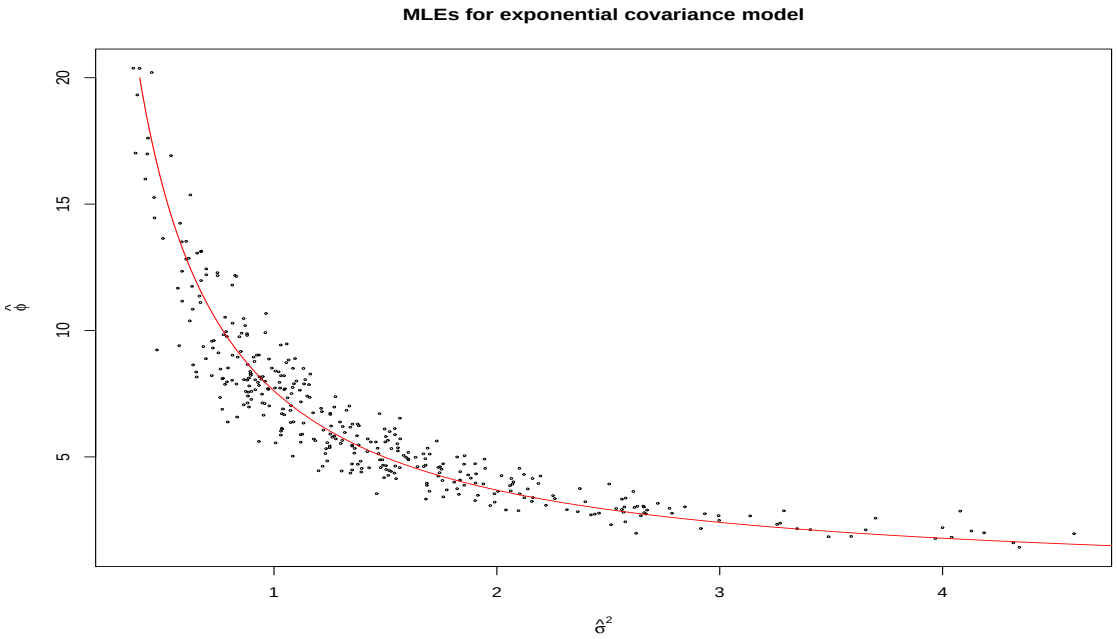

Figure 3.10: MLEs for the exponential covariance model ($C(h) = \sigma^2 \exp(-\phi|h|)$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 5)$), with n=100. We fit a linear model of $\log(\hat{\sigma}^2)$ on $\log(\hat{\phi})$ and obtain the curve $\sigma^2\phi^{0.96} = 6.95$.

3.8 Middle-ground asymptotics

Here, we consider the three-parameter OU process with equidistance sampling on a general interval $[0, L_n]$, where L_n could be allowed to vary with n. Explicit formula for the information matrix for (μ, σ, ϕ) enable investigation of the delicate balancing between infill and expansion.

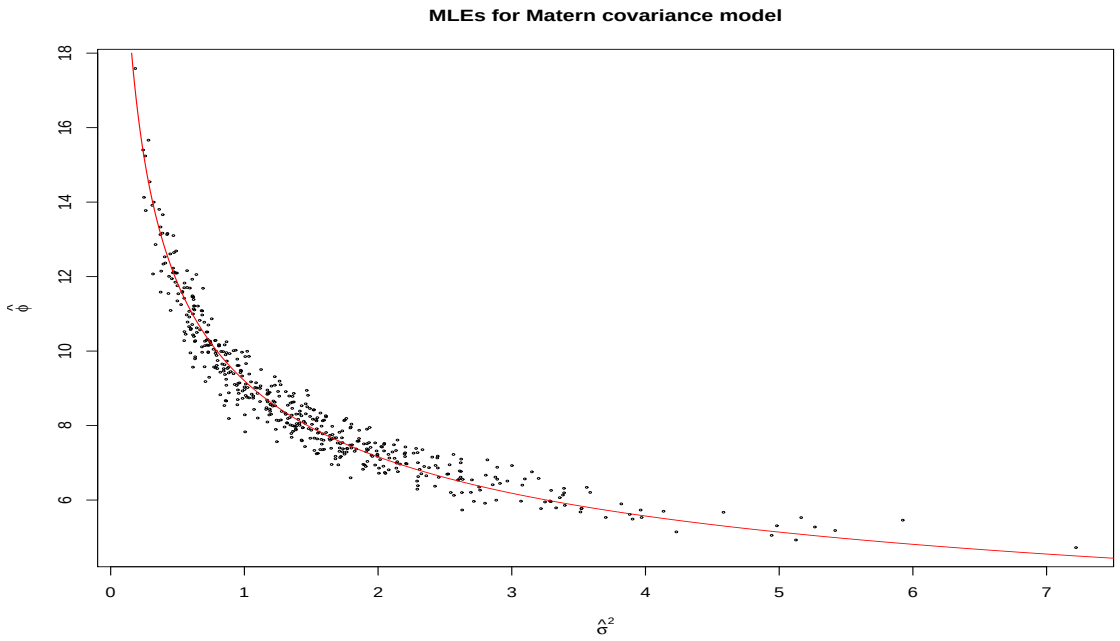

Figure 3.11: MLEs for the Matérn covariance model with ν=1.5 ($C(h) = \sigma^2(1 + \phi|h|)\exp(-\phi|h|)$ on $[0, L]$, where $(L, \sigma^2, \phi) = (1.3, 1.5, 8)$), n=100. We fit a linear model of $\log(\hat{\sigma}^2)$ on $\log(\hat{\phi})$ and plot the curve $\sigma^2\phi^{2.76} = 460.14$.

3.8.1 Information formula

Our data are $Y(L_n i/n)$ for $i = 1, \ldots, n$, giving a vector $\mathbf{Y} \sim \mathrm{N}_n(\mu\mathbf{1}, \sigma^2 R_n)$ where R_n has elements $\exp(-L_n\phi d_{i,j}) = \exp(-\phi L_n|j - i|/n)$. The sampling distance and basic neighbor correlation are now $\Delta_n = L_n/n$ and $\rho = \exp(-\phi\Delta_n)$, i.e., data are sampled with inter-distance $d_{i,j} = |j - i|\Delta_n$. Our previous calculations of Section 3.7.2 (done for infill asymptotics) remain valid. Writing $R^0_n, A^0_n, B^0_n, E^0_n$ for the old formula, valid for the constant interval $L_n = L$, as functions of ϕ, we now need to work with the correlation matrix $R_n = R^0_n(L_n\phi)$. This leads in particular to

$$\bar{R}_{n,i,j} = (\partial/\partial\phi)R^0_{n,i,j}(L_n\phi) = L_n\bar{R}^0_{n,i,j}(L_n\phi) = -\exp(-\phi d_{i,j})d_{i,j}. \tag{3.80}$$

With care to detail we find

$$A_n = A_n^0(L_n\phi) = 1 + (n-1)\frac{1-\rho_n}{1+\rho_n} = 1 + (n-1)\frac{1-\exp(-\phi L_n/n)}{1+\exp(-\phi L_n/n)}, \quad (3.81)$$

$$B_n = B_n^0(L_n\phi) = 2(n-1)\Delta_n\frac{\rho_n^2}{1-\rho_n^2} = 2L_n\frac{n-1}{n}\frac{\exp(-2\phi L_n/n)}{1-\exp(-2\phi L_n/n)}, \quad (3.82)$$

$$\begin{aligned} E_n &= E_n^0(L_n\phi) = (n-1)\Delta_n^2\rho_n^2\frac{1+\rho_n^2}{(1-\rho_n^2)^2} \\ &= (n-1)\frac{(L_n/n)^2\exp(-2\phi L_n/n)\{1+\exp(-2\phi L_n/n)\}}{\{1-\exp(-2\phi L_n/n)\}^2}. \end{aligned} \quad (3.83)$$

3.8.2 Infill, expansion, and middle ground

Let us consider three different scenarios.

Case 1: Infill

Similar to our asymptotics above, assume that L_n is small compared to a growing n. Then $\rho_n \doteq 1 - \phi L_n/n$, leading to

$$A_n \doteq 1 + \tfrac{1}{2}L_n\phi, \quad B_n \doteq n/\phi, \quad E_n \doteq n/(2\phi^2),$$

and therefore to

$$I_n(\sigma,\phi) \doteq n\begin{pmatrix} 2/\sigma^2 & 1/(\phi\sigma) \\ 1/(\phi\sigma) & 1/(2\phi^2) \end{pmatrix}$$

for the lower right block of the information matrix. This implies limited information for μ and trouble for (σ,ϕ), since the matrix approaches singularity.

Case 2: Expansion

Assume L_n is proportional to n, say $L_n/n = \Delta_n \to d$ whence $\rho = \exp(-\phi d)$. Then

$$A_n \doteq n\frac{1-\rho}{1+\rho}, \quad B_n \doteq 2n\Delta\frac{\rho^2}{1-\rho^2}, \quad E_n \doteq n\Delta^2\rho^2\frac{1+\rho^2}{(1-\rho^2)^2},$$

with consequent customary large-sample behavior for the information matrix. This case places us in the domain of time series methodology. Here ML estimators are

consistent and asymptotically normal, with variances

$$\begin{aligned}
\operatorname{Var}\widehat{\mu} &\doteq \frac{1}{n}\frac{1+\rho}{1-\rho}, \\
\operatorname{Var}\widehat{\sigma} &\doteq \tfrac{1}{2}\frac{\sigma^2}{n}\frac{1+\rho^2}{1-\rho^2}, \\
\operatorname{Var}\widehat{\phi} &\doteq \frac{1}{n}\frac{1-\rho^2}{\Delta^2\rho^2}.
\end{aligned}$$

Case 3: Middle ground

We consider a middle-ground scenario here.

(i) The information becomes unlimited for μ, i.e., it can be consistently estimated, if and only if $L_n \to \infty$. This is intuitive; we need an unlimited stretch on the time axis to learn about μ consistently. The conclusion follows from the fact that

$$n\{1-\exp(-\phi L_n/n)\} \to \infty \quad \text{if and only if } L_n \to \infty.$$

(ii) The estimator precision bounds for (unbiased) estimators of $\widehat{\sigma}$ and $\widehat{\phi}$ are

$$\operatorname{Var}\widehat{\sigma} \geq \frac{\sigma^2}{2n}\frac{E_n}{E_n - B_n^2/(2n)} \quad \text{and} \quad \operatorname{Var}\widehat{\phi} \geq \frac{1}{E_n - B_n^2/(2n)},$$

from results of Section 3.7.2. These are exact bounds, valid for any n. From previous calculations, we find

$$E_n - B_n^2/(2n) = \frac{(n-1)\Delta_n^2\rho_n^2}{1-\rho_n^2}\frac{1-\rho_n^2+(2/n)\rho_n^2}{1-\rho_n^2}$$

and

$$\frac{E_n}{E_n - B_n^2/(2n)} = \frac{1+\rho_n^2}{1-\rho_n^2+(2/n)\rho_n^2},$$

leading essentially to the precision bounds

$$\operatorname{Var}\widehat{\sigma} \geq \frac{\sigma^2}{2n}\frac{1+\rho_n^2}{1-\rho_n^2} \quad \text{and} \quad \operatorname{Var}\widehat{\phi} \geq \frac{1-\rho_n^2}{n\Delta_n^2\rho_n^2}. \tag{3.84}$$

To formalize our notion of middle-ground asymptotics, let $\Delta_n = L_n/n = L/\sqrt{n}$ so that $L_n = L\sqrt{n}$. Then the bounds say

$$\mathrm{Var}\,\widehat{\sigma} \geq \frac{1}{\sqrt{n}}\frac{\sigma^2}{2\phi L}[1 + O(1/\sqrt{n})] \quad \text{and} \quad \mathrm{Var}\,\widehat{\phi} \geq \frac{1}{\sqrt{n}}\frac{2\phi}{L}[1 + O(1/\sqrt{n})]. \tag{3.85}$$

We shall show that the ML estimators are consistent, under the $\Delta = L/\sqrt{n}$ sampling plan, and achieve these bounds. We *do* expect

$$\begin{aligned} n^{1/4}(\widehat{\mu} - \mu) &\to_d \mathrm{N}(0, 2\sigma^2/(\phi L)), \\ n^{1/4}(\widehat{\phi} - \phi) &\to_d \mathrm{N}(0, 2\phi/L), \\ n^{1/4}(\widehat{\sigma} - \sigma) &\to_d \mathrm{N}(0, \sigma^2/(2\phi L)). \end{aligned} \tag{3.86}$$

There would also be limiting binormality here, in fact

$$n^{1/4}\begin{pmatrix} \widehat{\sigma} - \sigma \\ \widehat{\phi} - \phi \end{pmatrix} \to_d \mathrm{N}_2\left(0, \frac{1}{L}\begin{pmatrix} \sigma^2/(2\phi) & -\sigma \\ -\sigma & 2\phi \end{pmatrix}\right). \tag{3.87}$$

Furthermore, $\hat{\sigma}\hat{\phi}^{1/2}$ is $\sqrt{n}$ consistent and

$$\widehat{\sigma}\hat{\phi}^{1/2} - \sigma\phi^{1/2} \to_d \mathrm{N}(0, \sigma^2\phi/(2L)). \tag{3.88}$$

Figure 3.12 gives credence to (3.86), (3.87), and (3.88). The bottom right plot displays 500 simulated version of $(\widehat{\sigma}, \widehat{\phi})$ for the OU case, in a situation where $(\sigma, \phi) = (1, 1)$, $\Delta_n = 1/\sqrt{n}$, and n=4000. The observed scaled standard deviations $n^{1/4}\mathrm{sd}(\widehat{\sigma})$, $n^{1/4}\mathrm{sd}(\widehat{\phi})$, and $n^{1/4}\mathrm{corr}(\widehat{\sigma}, \widehat{\phi})$ are (0.671, 1.414, -0.977), agreeing well with the theoretical middle-ground asymptotics values (0.707, 1.414, -1). $n^{1/2}\mathrm{sd}(\hat{\sigma}\hat{\phi}^{1/2}) = 0.736$ also agrees with the theoretical value 0.707 (see Table 3.1 for the simulation results for n=100, 500, 1000, 4000).

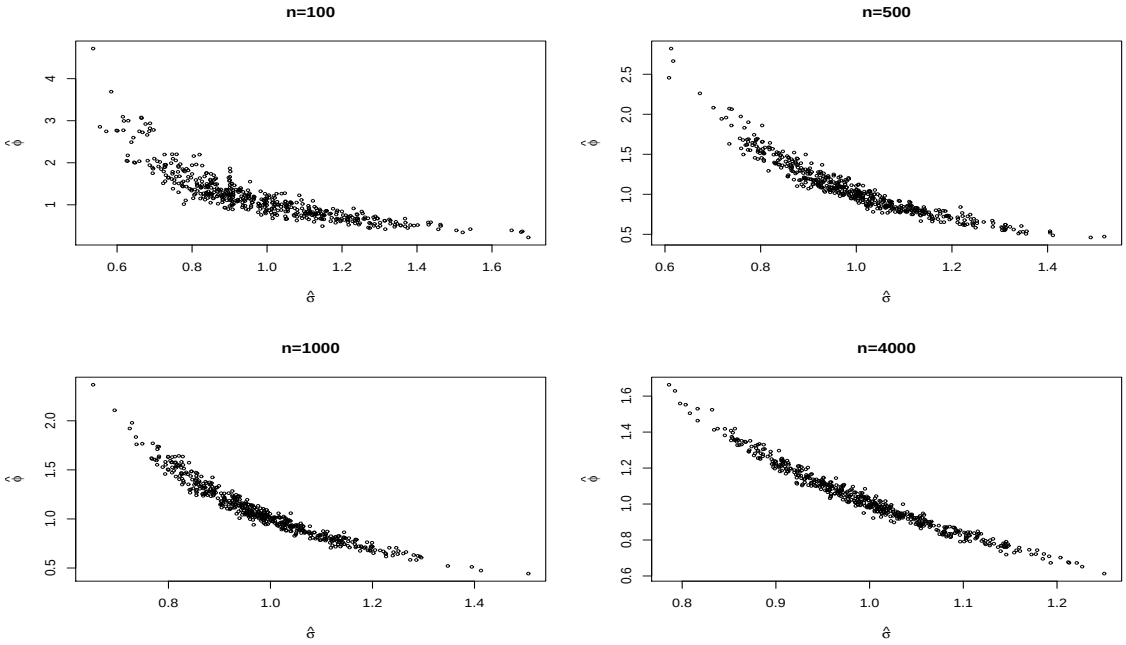

Figure 3.12: MLEs of (σ, ϕ) from simulation runs. The true exponential model parameters $(\sigma, \phi) = (1, 1)$ with $L_n = \sqrt{n}$ and $\Delta = 1/\sqrt{n}$ and $n = 100, 500, 1000, 4000$.

Table 3.1: Simulation results for the middle-ground exponential model.

n	$n^{1/4}\mathrm{sd}(\widehat{\sigma})$	$n^{1/4}\mathrm{sd}(\widehat{\phi})$	$n^{1/4}\mathrm{corr}(\widehat{\sigma}, \widehat{\phi})$	$n^{1/2}\mathrm{sd}(\hat{\sigma}\hat{\phi}^{1/2})$
100	0.645	1.791	-0.842	0.766
500	0.699	1.588	-0.923	0.730
1000	0.710	1.570	0.949	0.706
4000	0.671	1.414	-0.977	0.736
asy. val.	0.707	1.414	-1	0.707
$(\sigma, \phi) = (1, 1)$, $L_n = \sqrt{n}$, and $\Delta_n = 1/\sqrt{n}$				

Remark 7. The ML estimators are defined (and computed) as follows, for a fixed sampling window size Δ, e.g. $\Delta = L/\sqrt{n}$ in what could be our favourite example. Let

$$\widehat{\sigma}(\phi) = \{\mathbf{y}^T R_n(\phi)^{-1}\mathbf{y}/n\}^{1/2}.$$

Then $\widehat{\phi}$ is the minimiser of

$$Q_n(\phi) = \log \widehat{\sigma}(\phi) + \tfrac{1}{2}(1 - 1/n)\log\{1 - \exp(-2\phi\Delta)\},$$

and $\widehat{\sigma} = \widehat{\sigma}(\widehat{\phi})$. This is easy to implement. Note also that

$$\mathbf{y}^T R_n^{-1} \mathbf{y} = y_1^2 + \sum_{i=2}^{n} \frac{(y_i - \rho y_{i-1})^2}{1 - \rho^2},$$

where $\rho = \exp(-\phi\Delta)$.

Our challenge is now to prove that $(\widehat{\sigma}, \widehat{\phi})$ is consistent and asymptotically normal, as envisaged above. Perhaps the log-likelihood $\ell_n(\sigma, \phi)$ will more and more tend to exhibit a ridge at the maximum. The correlation between the two estimators seems to converge to -1.

We have stumbled upon a situation that resembles the following constructed example, which could be helpful for both intuition and proofs: one observes i.i.d. pairs

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} 1 & -\rho_n \\ -\rho_n & 1 \end{pmatrix} \right),$$

where $\rho_n = 1 - L/\sqrt{n}$ is on its way to 1, and taken as known in this construction. To match this situation even more closely to some of the problems examined in this section, one could assume that one observes $\sqrt{n}$ such pairs. Then,

$$n^{1/4} \begin{pmatrix} \widehat{a} - a \\ \widehat{b} - b \end{pmatrix} \to_d \mathrm{N}_2\left(0, \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\right).$$

The log-likelihood surface would have a strongly visible ridge where $a + b = \bar{x} + \bar{y}$, and the particular parameter $a + b$ is estimable with much better precision, namely $O(1/n)$ variance, than all other parameters.

Remark 8. We have focussed here on $\Delta = L/\sqrt{n}$ as middle ground, partly for pedagogical reasons, but there is really a full bridge here, spanned by $\Delta = L/n^\gamma$, for $\gamma \in (0, 1)$, where $\gamma = 0$ is time series (classic expanding domain) and $\gamma = 1$ is infill.

Remark 9. Everyone does time series with AR and ARMA etc., but there is no reason not to use the Matérn class for correlation. As Δ is fixed positive, in principle both ϕ and ν may be estimated well from the data, along with (μ, σ).

Remark 10. One could also indicate that the $n^{1/4}$ middle ground phenomenon would hold for slowly expanding circles in dimension two.

Remark 11. We have of course been generously helped in our analysis of the OU process by the fully explicit expressions for the information matrix. We may speculate that the same type of phenomena take place also for the full Matérn class of correlation functions, i.e. not only for $\nu = \frac{1}{2}$. This might be harder to prove mathematically, but we could 'inspect' and verify, if required, from computing the required A_n, B_n, E_n numerically (see Figure 3.7).

We simply anticipate that the $n^{1/4}$ asymptotics result (3.86) must be true, based on the information calculus and simulation study. The part relating to $n^{1/4}(\widehat{\mu} - \mu)$ is easy, coming from normality etc.; the challenge is to prove $n^{1/4}$ results for $(\widehat{\sigma}, \widehat{\phi})$, under $L_n = L\sqrt{n}$, or $\Delta = L/\sqrt{n}$, circumstances. If we succeed in really proving this, it would presumably not be more difficult to reach results for the full bridge, stretching from infill to time series, where the sampling range is $L_n = Ln^{1-\gamma}$ and the sampling distance is $\Delta = L/n^{\gamma}$.

Under the $\Delta = L/\sqrt{n}$ assumption, let us start with the log-likelihood function

$$\ell_n(\sigma, \phi) = -n \log \sigma - \tfrac{1}{2}(n-1) \log(1-\rho^2) - \tfrac{1}{2} \frac{1}{\sigma^2} Q_n(\rho),$$

where

$$Q_n(\rho) = \frac{y_1^2 + \sum_{i=2}^n (y_i - \rho y_{i-1})^2}{1-\rho^2}.$$

Here $\rho = \exp(-\phi L/\sqrt{n})$ tends slowly to 1 as n grows. Our first exercise is to derive the limit of $n^{-1}\ell_n(\sigma, \phi)$, under the assumption of a true model with (σ_0, ϕ_0). Note

that $y_i - \rho y_{i-1}$ has mean zero and variance

$$\sigma_0^2(1+\rho^2-2\rho_0\rho) = \sigma_0^2\big[1+\exp(-2\phi c/\sqrt{n}) - 2\exp\{-(\phi+\phi_0)/\sqrt{n}\}\big] \doteq \sigma_0^2 2\phi_0 L/\sqrt{n},$$

and, in view of the denominator $1-\rho^2 \doteq 2\phi L/\sqrt{n}$, we have

$$Q_n(\rho)/n \to_p \sigma_0^2\phi_0/\phi.$$

Furthermore,

$$\log(1-\rho^2) \doteq \log(2\phi L/\sqrt{n}) = \log\phi + \text{const.}.$$

Ignoring terms that do not depend on the parameters, therefore,

$$n^{-1}\ell_n(\sigma,\phi) \to_p -\log\sigma - \tfrac{1}{2}\log\phi - \tfrac{1}{2}\frac{\sigma_0^2\phi_0}{\sigma^2\phi} = -\log\eta - \tfrac{1}{2}\frac{\eta_0^2}{\eta^2},$$

in terms of

$$\eta_0 = \sigma_0\sqrt{\phi_0} \quad \text{and} \quad \eta = \sigma\sqrt{\phi}.$$

We anticipate two things from this result. (i) The limit function is uniquely maximised for $\eta = \eta_0$, so when maximising the likelihood, $\widehat{\eta} = \widehat{\sigma}\widehat{\phi}^{1/2}$ should converge to the maximiser of the limit, i.e. η; (ii) we cannot estimate σ and ϕ well, only their inferred parameter η. As (3.86) claims (and that we wish to prove rigorously), not all hope is lost for (σ,ϕ), but in the large-sample limit η is rather more advantaged (MLE for η is $\sqrt{n}$ consistent). Also, we expect from the above that the log-likelihood surface will tend to exhibit a strong ridge where $\sigma\rho^{1/2} \doteq \widehat{\eta}$.

It appears necessary to prove three things now. The first is that $n^{1/4}(\widehat{\phi}-\phi)$ has a limit distribution, where $\widehat{\phi}$ maximises the profile

$$\pi_n(\phi) = \max\{\ell_n(\sigma,\phi)\colon \sigma>0\} = -n\log\widehat{\sigma}(\phi) - \tfrac{1}{2}(n-1)\log\{1-\rho(\phi)^2\} - \tfrac{1}{2}n,$$

where

$$\widehat{\sigma}(\phi) = \{Q_n(\rho(\phi))/n\}^{1/2}.$$

The second would be to prove the corresponding result for $\widehat{\sigma}$; the profile for σ is a bit more cumbersome to work with, though. And the third is that $\sqrt{n}(\widehat{\eta}-\eta)$ should have a normal limit, a precision order of magnitude better than for the two other parameters. This would be a proper generalisation and strengthening of the results of Ying (1991) and Zhang (2004); they work only with infill, whereas here we think we can reach a $\sqrt{n}$-result valid also with middle-ground asymptotics.

Remark 12. Assume we manage to prove the simultaneous limit distribution results anticipated just after (3.86). Then, for any smooth parameter $\lambda = h(\sigma, \phi)$, $n^{1/4}(\widehat{\lambda}-\lambda)$ has a limit distribution, with variance equal to

$$\Big\{\Big(\frac{\partial h}{\partial \sigma}\Big)^2\frac{\sigma^2}{2\phi}+\Big(\frac{\partial h}{\partial \phi}\Big)^2 2\phi - 2\frac{\partial h}{\partial \sigma}\frac{\partial h}{\partial \phi}\sigma\Big\}\frac{1}{L}.$$

But this is equal to zero for $\lambda = \sigma\phi^{1/2}$, and, only for functions of this η parameter. All other estimands need to be satisfied with $n^{1/4}$ convergence, whereas η can brag of $n^{1/2}$. Equivalent measure results can explain why only $h(\sigma, \phi) = g(\sigma\phi^{1/2})$ functions have this property.

Remark 13. The point of asymptotics is to provide good approximations for finite n. What we could do, presumably, in order to compare infill methods, expansion methods, and middle-ground methods, is to reach mathematical results for the scenario with $\Delta = L/n^\gamma$, and then phrase approximations based on the limit distributions in terms of Δ. Then for actual applications one inserts the Δ, the range of sites divided by sample size.

3.9 Including a nugget

For a model including a nugget component, i.e., $Y(s) = \mu(s) + W(s) + \epsilon(s)$, where $\epsilon(s) \sim \mathrm{N}(0, \tau^2)$, what can we conclude with regard to asymptotic behavior? The

following theorem indicates that, under infill asymptotics, the parameter τ^2 can be consistently estimated with the usual rate of convergence and the results for the covariance parameters remain same as in Section 3.6-3.8.

Theorem 7 (Stein, 1999, p.122–123)**.** *$P(\mu_1, C_1 + \tau_1^2 I) \equiv P(\mu_2, C_2 + \tau_2^2 I)$ if and only if $\tau_1^2 = \tau_2^2$ and $P(\mu_1, C_1) \equiv P(\mu_2, C_2)$.*

For information regarding the mean parameter, one would expect that $\mathbf{1}^T(\sigma^2 R_n + \tau^2 \mathrm{I}_n)^{-1}\mathbf{1} < \mathbf{1}^T(\sigma^2 R_n)^{-1}\mathbf{1}$. The intuitive explanation is that the information for mean parameter μ will *decrease* if there is some extra noise (e.g. measurement error) in the model. In fact, it is true in general and can be proved easily. If $\mathbf{1}^T(\sigma^2 R_n)^{-1}\mathbf{1}$ is bounded and convergent, then $\mathbf{1}^T(\sigma^2 R_n + \tau^2 \mathrm{I}_n)^{-1}\mathbf{1}$ must be bounded.

Lemma 1. *$\mathbf{1}^T(\sigma^2 R_n)^{-1}\mathbf{1} > \mathbf{1}^T(\sigma^2 R_n + \tau^2 I_n)^{-1}\mathbf{1}$ for each n and for all valid covariance functions.*

Proof. It is sufficient to show $\mathbf{1}^T(aR_n)^{-1}\mathbf{1} > \mathbf{1}^T(aR_n + \mathrm{I}_n)^{-1}\mathbf{1}$ for a positive number a. Since R_n is an $n \times n$ positive definite matrix, it has a spectral decomposition form $R_n = U_n \Lambda_n U_n^T$, where $U_n U_n^T = \mathrm{I}_n$ and $\Lambda_n = \mathrm{diag}(\lambda_1, ..., \lambda_n), \lambda_i > 0$. Note $(aR_n + \mathrm{I}_n)^{-1} = U_n(a\Lambda_n + \mathrm{I}_n)^{-1}U_n^T$. So

$$\begin{aligned}
\mathbf{1}^T(aR_n)^{-1}\mathbf{1} - \mathbf{1}^T(aR_n + \mathrm{I}_n)^{-1}\mathbf{1} &= \mathbf{1}^T U_n[(a\Lambda_n)^{-1} - (a\Lambda_n + \mathrm{I}_n)^{-1}]U_n^T\mathbf{1} \\
&= \frac{1}{a^2}\mathbf{1}^T U_n \mathrm{diag}\left(\frac{1}{\lambda_1^2 + \lambda_1/a}, ..., \frac{1}{\lambda_n^2 + \lambda_n/a}\right)U_n^T\mathbf{1} > 0
\end{aligned}$$

□

Chen, Simpson and Ying (2000) address infill asymptotics for a one-dimensional OU process with a nugget term. They show that if ϕ is known, then

$$\hat{\tau}^2 \to \tau_0^2 \quad \text{a.s.} \quad \text{but} \quad \hat{\sigma}^2 \to \sigma_0^2 \text{ in probability,}$$

where $\hat{\tau}^2$ and $\hat{\sigma}^2$ are MLEs and τ_0^2 and σ_0^2 are true values. Furthermore,

$$\begin{pmatrix} n^{1/4}(\hat{\sigma}^2 - \sigma_0^2) \\ n^{1/2}(\hat{\tau}^2 - \tau_0^2) \end{pmatrix} \rightarrow_d \mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\sqrt{2}\tau_0\phi^{-1/2}\sigma_0^3 & 0 \\ 0 & 2\tau_0^4 \end{pmatrix} \right).$$

If ϕ is unknown, then

$$\hat{\tau}^2 \rightarrow \tau_0^2 \quad \text{a.s.} \quad \text{and} \quad \hat{\phi}\hat{\sigma}^2 \rightarrow \phi_0\sigma_0^2 \text{ in probability,}$$

where $\hat{\phi}$ is the MLE for ϕ and ϕ_0 is the true value. And

$$\begin{pmatrix} n^{1/4}(\hat{\phi}\hat{\sigma}^2 - \phi_0\sigma_0^2) \\ n^{1/2}(\hat{\tau}^2 - \tau_0^2) \end{pmatrix} \rightarrow_d \mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\sqrt{2}\tau_0(\phi\sigma_0^2)^{3/2} & 0 \\ 0 & 2\tau_0^4 \end{pmatrix} \right).$$

3.10 Prediction

So far we focused on the estimation problem. In other words, we wish to learn how well we can estimate the model parameters. In various geostatistical applications, the primary goal might be prediction and interpolation, rather than super-accurate estimation of the model parameters. Stein (1999) provides a mathematical treatment for the prediction (kriging) problem when the covariance structure of the random process is unknown. The basic message regarding prediction from Stein's book is that even if we do not mange to estimate the model parameters well, or even if we estimate the covariance structure non-negligibly wrong, it is possible to have nearly optimal linear predictors. The underlying mathematical reason is related to the equivalent measure ideas we discuss in Section 3.3. If a large amount of data cannot help us distinguish two measures with high probability, then the prediction based on these two measures won't differ much.

In fact, there is a more general statement about the principles of prediction or forecasting. Dawid (1984, p.285) says "things we shall never find much out about

cannot be very important for prediction" and he calls it Jeffreys's law based on a similar statement by Jeffreys (1938, p.718). See Stein (1999, p.140–143) for more discussion.

3.11 Discussion

In this chapter, we discussed various topics, including information, consistency, and asymptotic normality for spatial models under three asymptotics settings. We provided many new results; however, we are still unable to give a general theory for the questions we would like to answer. In the future, we shall prove limiting normality of estimators for the Matérn model under infill and middle-ground asymptotics. We shall prove similar asymptotic results for the generalized linear spatial model, i.e. Y in the first stage is non-Gaussian. We shall also study the asymptotic properties of the spatial predictive distributions.

Chapter 4

Analysis for Large Spatial Datasets

4.1 Introduction

As we have discussed in Chapter 1, for spatial data analysis exercises, we usually build hierarchical models with spatial structure described though random effects using Gaussian processes. If the sample size n is very large, exact likelihood based inference becomes unstable and, eventually, infeasible since it involves computing the inverse and determinant of a large covariance matrix (the computation cost of matrix inversion and determinant is $O(n^3)$). If we wish to fit a Bayesian model, implementing a suitable MCMC algorithm, the large matrix will make repeated calculations impractical. In this chapter, we review a number of strategies for handling large spatial datasets. We propose a finite sum process approximation model which is conceptually simple and routine to implement. Simulated and real data examples are provided to illustrate the method.

For a point-referenced dataset, a random surface is assumed over a region D in $\mathbb{R}^2$. More specifically, observation $Y(s_i)$ at location $s_i \in D$ for $i = 1, ..., n$, is assumed to have an additive structure

$$Y(s_i) = \mu(s_i) + W(s_i) + \epsilon(s_i), \tag{4.1}$$

where $\mu(s)$ is a mean function; $W(s)$ is a mean zero Gaussian process; and $\epsilon(s)$ is white noise process with variance τ^2.

As the number of observations n grows large, evaluation of the likelihood will become very slow and, eventually, impractical. Likelihood calculation requires evaluating quadratic forms involving the inverse of covariance matrix of size $n \times n$ as well as the determinant of that matrix. If we wish to fit a Bayesian model, we will need to implement an MCMC algorithm that will include a Metropolis-Hastings step to update the parameters in the covariance function. In turn, this will require repeated likelihood evaluation both within and across iterations.

We refer to this computational difficulty informally as the *"large n problem"* in spatial data analysis. Banerjee, Carlin and Gelfand (2004) summarize several possible approaches to the "large n problem".

- *Subsampling strategy*

  As indicated by its name, one could take a subset of size n_0 from the n sampled locations, resulting in a computationally tractable sample size. For independent data, usually the variances of the maximum likelihood estimators decrease at a rate of $\sqrt{n}$. But because of the dependence in spatial data, the convergence rate could be much slower than $\sqrt{n}$ (see Chapter 3). Intuitively, adding a new observation very close to an existing observation only helps to estimate the noise term τ^2 and gains little information in terms of learning the spatial structure. So if the spatial correlation is strong and noise is weak, one would hope that it will not lose much information by ignoring some of the available data. For subsampling method, we need to determine the subsample size n_0 and the subsampling locations. Usually we choose the value of n_0 as large as the computation power allows. For a fixed n_0, we wish to select an optimal set of n_0 locations from a total n locations. This is not an easy task when n is

large. In Chapter 5, we address this optimal spatial design problem in detail.

- *Likelihood approximation*

  The second method is to approximate the likelihoods following Vecchia's (1988) idea and reduce the size of the covariance matrix. Stein et al. (2004) develop this likelihood approximation in more detail. The basic idea is to approximate the likelihood based on the fact that any joint density can be written as a product of conditional densities. Specifically, suppose that $\mathbf{Y} = (Y(s_1), ..., Y(s_n))^T$ has joint density $f(\mathbf{y}; \boldsymbol{\theta})$, where $\mathbf{y} = (y(s_1), ..., y(s_n))^T$ and $\boldsymbol{\theta}$ is a vector of parameters. Then we can write

$$f(\mathbf{y}; \boldsymbol{\theta}) = f(y(s_1)) \prod_{i=2}^{n} f\{y(s_i)|y(s_1), ..., y(s_{i-1})\}. \tag{4.2}$$

  Next, we approximate $f\{y(s_i)|y(s_1), ..., y(s_{i-1})\}$ by $f\{y(s_i)|\partial s_i\}$, where ∂s_i defines a set of "neighborhood" observations of s_i among $\{y(s_1), ..., y(s_{i-1})\}$. For instance, ∂s_i might be the $\min(m, i-1)$ nearest observations to s_i in Euclidean distance for a fixed integer $m \ll n$ ($\ll$ means much smaller). By this approximation, the evaluation of the likelihood involves at most 'n' $m \times m$ matrices, instead of a single $n \times n$ matrix. The performance of this likelihood approximation depends on the value of m and the ordering of the observations. The large value of m helps increase the accuracy of likelihood approximation but makes likelihood computation less efficient. Vecchia (1988) studies the ordering effect and finds this effect to be small in his examples.

  More generally, if we partition $\mathbf{Y}$ into subvectors $\{\mathbf{Y}_1, ..., \mathbf{Y}_p\}$ ($\mathbf{Y}_i$ may have different length), we can write the joint density $f(\mathbf{y}; \boldsymbol{\theta})$ as a product of conditional densities of these subvectors. That is, $f(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}_1) \prod_{i=2}^{p} f(\mathbf{y}_i|\mathbf{y}_1, ..., \mathbf{y}_{i-1})$. Then we can approximate each conditional density $f(\mathbf{y}_i|\mathbf{y}_1, ..., \mathbf{y}_{i-1})$ in a similar way as above.

It is a challenging task to determine a right value of m and define the optimal "neighborhood" (neighborhood need not mean nearest points; it may be useful to include some distant observations in the neighborhood). See Stein et al. (2004) for more details.

- *Spectral methods*

  Another approach is to use spectral methods to approximate the likelihood. The basic idea is to work in the *spectral domain*, develop a *periodogram*, and then utilize the *Whittle likelihood* (Whittle, 1954). The Whittle likelihood in the spectral domain is an approximation to the exact likelihood in the spatial domain. The word "periodogram" was coined by Arthur Schuster in 1898 and is a nonparametric estimator of the spectral density of a random field. Suppose we observe the process $W(s)$ in a regular two-dimensional $n_1 \times n_2$ grid so the sample size is $n = n_1 n_2$. Periodogram at frequency ω is defined as follows (Priestley, 1981, Chapter 6):

$$I_n(\omega) = (2\pi)^{-2} n^{-1} |\sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} W(s) e^{-is^T\omega}|^2, \tag{4.3}$$

  where $s = (s_1, s_2)$. $I_n(\omega)$ is an asymptotically unbiased estimator of the spectral density $f(\omega)$ of W. Although $I_n(\omega)$ is defined for all ω but we are able to evaluate it numerically only at a discrete set of frequencies. In particular, this set of frequencies is $2\pi\mathbf{m}/\mathbf{n}$, where $\mathbf{m}/\mathbf{n} = (m_1/n_1, m_2/n_2)$, $(m_1, m_2) \in B_\mathbf{n}$ and $B_\mathbf{n} = \{-\lfloor n_1 - 1 \rfloor, ..., n_1 - \lfloor n_1/2 \rfloor\} \times \{-\lfloor n_2 - 1 \rfloor, ..., n_2 - \lfloor n_1/2 \rfloor\}$ ($\lfloor x \rfloor$ means the greatest integer less than or equal to the real number x). The Gaussian negative log likelihood can be approximated by (Whittle, 1954)

$$\sum_{\mathbf{m} \in B_\mathbf{n}} \{\log f(2\pi\mathbf{m}/\mathbf{n}) + I_n(2\pi\mathbf{m}/\mathbf{n}) f^{-1}(2\pi\mathbf{m}/\mathbf{n})\}. \tag{4.4}$$

The calculation of (4.4) is rapid by using the fast Fourier transform and we can make inference based on this approximated likelihood (e.g. we can find the maximum likelihood estimators of the model parameters $\boldsymbol{\theta}$.) This approach must be used with caution. The performance of Whittle likelihood is unclear.

- *Gaussian Markov random field methods*

  Gaussian Markov random fields (GMRF) are often used to model areal unit data and lattice data. In particular, *simultaneously autoregressive* (SAR) and *conditionally autoregressive* (CAR) models are two very popular GMRF models. For a point-referenced dataset, we can use inverse distance to create a proximity matrix for the observations at a finite set of locations. The joint distribution is determined through its full conditional distributions (by Brook's Lemma) and the inverse matrix emerges explicitly (Besag, 1974). Note that there is no notion of a spatial process if we work with a Gaussian Markov random field. In other words, we can not specify the finite-dimensional distribution for an arbitrary set of locations in the study region.

  It is possible to use a Markov random field to approximate a Gaussian process. One can choose a Markov random field which has joint density "close" to the joint density of Y. However, the approximation is not transparent and requires extrapolation to a regular grid (see e.g. Rue and Tjelmeland, 2002).

- *Covariance tapering*

  Furrer et al. (2005) propose using a covariance tapering to produce a sparse covariance matrix. The idea is to truncate the covariance function to zero with a certain range using an appropriate compactly supported positive definite function. They use a conjugate gradient algorithm to speed up calculation.

We shall first propose what we call *"multilevel process model"*. Suppose we partition the study region D into m subregions, denoted as D_j, $j = 1, ..., m$. There are n_j points

in subregion j so that $n = \sum_{j=1}^{m} n_j$. We assume the spatial process on each subregion to be stationary. We assign random effects $\boldsymbol{\gamma}_{\mathbf{s}_j^*} = (\gamma(s_{j,1}^*), \gamma(s_{j,2}^*), ..., \gamma(s_{j,p}^*))^T$ for the j^{th} subregion, where $\mathbf{s}_j^* = (s_{j,1}^*, ..., s_{j,p}^*)$ is a set of "representative" points for subregion j and p is a small integer numer. Conditioning on the random effects, $Y(s_i)$ and $Y(s_j)$ are assumed to be independent, given s_i lies in subregion i and s_j is in subregion j. The random effects $\boldsymbol{\gamma}_{\mathbf{s}_j^*}$ itself follows a spatial process that can be envisioned as a *hyper* process. We build a *multilevel* spatial model as follows:

$$Y(s) = X^T(s)\boldsymbol{\beta} + W_j(s) + h^T(s, \mathbf{s}_j^*)\boldsymbol{\gamma}(\mathbf{s}_j^*) + \epsilon(s), \tag{4.5}$$

where $s \in D_j$; $W_j(s)$ is a Gaussian process on D_j; $h(s, \mathbf{s}_j^*)$ is a vector which relates s with $\mathbf{s}_j^*$ (its l^{th} element may take the form $\exp(-\phi|s - s_l^*|)$); and $\gamma(s^*)$ is a hyper Gaussian process.

The above model can be viewed as the following hierarchical model:

$$\textbf{Stage 1:} \quad Y(s)|\boldsymbol{\beta}, W_j(s), \tau^2, \boldsymbol{\gamma}_{\mathbf{s}_j^*} \sim \mathrm{N}(X^T(s)\boldsymbol{\beta} + W_j(s) + h^T(s, \mathbf{s}_j^*)\boldsymbol{\gamma}(\mathbf{s}_j^*), \tau^2), \tag{4.6}$$

$$\textbf{Stage 2:} \quad W_j(s)|\sigma^2, \boldsymbol{\phi} \sim \mathrm{N}_{n_j}(0, \sigma^2 R_{n_j}(\boldsymbol{\phi})), \tag{4.7}$$

$$\gamma(s^*)|\sigma^2, \boldsymbol{\phi} \sim \mathrm{N}_{pm}(0, \sigma^2 R_{pm}(\boldsymbol{\phi})). \tag{4.8}$$

The covariance matrix of the vector $\mathbf{W} = (W(s_1), ..., W(s_n))$ is

$$K_{\mathbf{W}} = \begin{pmatrix} (\sigma^2 R_{n_1}(\boldsymbol{\phi})) & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & (\sigma^2 R_{n_m}(\boldsymbol{\phi})) \end{pmatrix} + \sigma^2 H R_{pm}(\boldsymbol{\phi}) H^T, \tag{4.9}$$

where $H = (h(s_1, \cdot), ..., h(s_n, \cdot))^T$ ($h(s_1, \cdot) = h(s_1, \mathbf{s}_j^*)$ if $s_1 \in D_j$).

It is natural to fit the multilevel process model (4.5) in a Bayesian framework and the model implementation is straightforward. The computation is fast if the values of $\max\{n_j\}$, m, and p are not large. This model may be flexible enough in terms of fitting the data. The W process associated with the model is *nonstationary*. The

multilevel process model cannot approximate a desired stationary model. In other words, it is easy to see from (4.9) that the resulting covariance matrix $K_{\mathbf{W}}$ cannot approximate the covariance matrix of a stationary process. We do not pursue this model further here.

In this chapter, our focus is to investigate the performance of an attack on the "large n problem" using a finite sum process approximation model. More specifically, a spatial random process can be approximately described as a linear combination of, say, m random variables. Thus, no matter how large n is, we only need to handle an $m \times m$ matrix. We propose three methods to approximate the spatial process, namely, i) approximation based on the *kernel mixing representation* of the process; ii) approximation based on the projection ideas; and iii) approximation based on *Karhunen-Loève expansion*.

The kernel mixing approximation has a similar looking form as in Higdon et al. (1998) but with a perspective different from ours. They seek a flexible modeling approach, confined to Gaussian kernels, with focus on achieving certain association structure while we propose a process approximation tool with focus on approximating a desired stationary model. Furthermore, we clarify the theoretical justification for the use of the approximation. We formalize the mechanics of the approximation for various classes of covariance functions.

After approximating the spatial process using a finite sum (no matter which method you choose to develop the approximation), we implement the approximate model in a Bayesian framework. We illustrate our method though several examples.

4.2 Theoretical preliminaries

In order to develop the approximate model, we first recall Karhunen's theorem on the generalized orthogonal representation of random functions (Yaglom, 1987, Section 26)

which asserts that a random function $W(s)$ is representable as $W(s) = \int_A \psi(s,t)Z(\mathrm{d}t)$ where A is a given measurable set, ψ is a complex valued function of two arguments and $Z(\cdot)$ is an orthogonal random measure on A (see below).

4.2.1 Representations for stationary process

In the case of a stationary process several representations are available. These include the spectral representation, kernel mixing (moving average) representation, and the Karhunen-Loève expansion representation.

Spectral representation

The spectral representation theorem asserts that for every stationary process $W(s)$ on $\mathbb{R}^d$ with mean 0 and finite variance there can be assigned a random measure $Z(d\omega)$ with orthogonal increments such that for each s we have the representation (see Section 2.1.4):

$$W(s) = \int_{\mathbb{R}^d} e^{i\omega^T s} Z(\mathrm{d}\omega). \tag{4.10}$$

Orthogonal increments of Z imply that for disjoint measurable sets A and B: (i) $\mathrm{E}(Z(A)) = 0$, (ii) $\mathrm{E}(Z(A)\overline{Z(B)}) = 0$, (iii) $Z(A \cup B) = Z(A) + Z(B)$. If we define $\mathrm{E}|Z(A)|^2 = \nu(A)$ where $\nu(\cdot)$ is a σ-finite measure, this representation produces a spectral representation for the covariance function $C(h)$ associated with $W(s)$:

$$C(h) = \int_{\mathbb{R}^d} e^{i\omega^T h} \nu(\mathrm{d}\omega). \tag{4.11}$$

If we abandon the orthogonality for the random measure $Z(\cdot)$, we can obtain a *nonstationary* process.

Kernel mixing representation

An alternative representation for a large class of mean 0 stationary process on $\mathbb{R}^d$

takes the form:

$$W(s) = \int_{\mathbb{R}^d} k(s-t)Z(\mathrm{d}t), \tag{4.12}$$

where $Z(\cdot)$ is an orthogonal random measure and $k(\cdot)$ is a non-random kernel function. It is straightforward to show that (4.12) produces a stationary random process (see Section 4.2.2 for details). However, not all stationary random process can be written using kernel mixing (again, see examples in Section 4.2.2).

Karhunen-Loève expansion representation

For a random process $W(s)$ defined on some compact set $D \in \mathbb{R}^d$ with covariance function $C(s,t)$ for $s,t \in D$, under certain conditions for $C(\cdot,\cdot)$, the Karhunen-Loève expansion decomposes $W(s)$ into a countable orthogonal series,

$$W(s) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\phi_i(s)Z_i, \tag{4.13}$$

where λ_i are the eigenvalues for the process, $\phi_i(s)$ are orthonormal eigenfunctions associated with λ_i, and

$$Z_i = \frac{1}{\sqrt{\lambda_i}} \int_D W(s)\overline{\phi_i(s)}\,\mathrm{d}s. \tag{4.14}$$

We solve

$$\int_D C(s,t)\phi_i(t)\,\mathrm{d}t = \lambda_i\phi_i(s) \quad \text{and} \quad \int_D \phi_i(s)\overline{\phi_j(s)}\,\mathrm{d}s = \delta_{ij}, \tag{4.15}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise, to obtain λ_i and $\phi_i(s)$. The covariance function has the representation

$$C(s,t) = \sum_{i=1}^{\infty} \lambda_i\phi_i(s)\phi_i(t). \tag{4.16}$$

The Karhunen-Loève expansion can produce both stationary and nonstationary processes and offers the possibility of an approximation when a process can not be represented using kernel mixing.

4.2.2 Kernel mixing in detail

Here, we elaborate some of the details involved in kernel mixing with a random measure. Let $(X, \mathcal{F}, \nu)$ be a σ-finite measure space and $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space. Define

$$W(s) = \int_X k(s-t)Z(\mathrm{d}t), \tag{4.17}$$

where $k(\cdot)$ is a non-random function in $L^2(X, \mathcal{F}, \nu) \equiv L^2_\nu$, the space of square integrable functions defined on $(X, \mathcal{F}, \nu)$. $Z(dt)$ is a random orthogonal measure as defined below (4.10).

Consider the stochastic integral $Z[f] = \int_X f(t)Z(\mathrm{d}t)$, where we think of $Z[f]$ as a mapping from a function $f \in L^2_\nu$ to a random variable in $L^2(\Omega, \mathcal{B}, \mathbb{P}) \equiv L^2_{\mathbb{P}}$, the space of random variables with finite second moment. We define this stochastic integral in the usual fashion, building from simple functions. For a simple function $f(t) = \sum_{i=1}^n a_i 1_{A_i}(t)$, where $A_1, ..., A_n$ are measurable and disjoint, we define $Z[f] = \sum_{i=1}^n a_i Z(A_i)$. Then $\mathrm{E}(Z[f]) = 0$ and $\mathrm{Var}(Z[f]) = \sum_{i=1}^n a_i^2 \nu(A_i)$. Omitting details (see, e.g. Gikhman and Skorokhod, 1974), since the class of simple functions is dense in L^2_ν, for each $f \in L^2_\nu$, there exists a sequence of simple function $\{f_n\}$ such that $\|f_n - f\|_{L^2_\nu} \to 0$. Thus we can define $Z[f]$ as a mean square limit of $Z[f_n]$. This limit exists and is independent of the sequence $\{f_n\}$. It also follows that $\mathrm{E}(Z[f]Z[g]) = \int_X f(t)g(t)\nu(\mathrm{d}t)$ for all $f, g \in L^2_\nu$.

Under this definition, suppose that we specify $Z(A)$ to be a Gaussian measure by assuming $Z(A) \sim \mathrm{N}(0, \nu(A))$, where $\mathrm{N}(a, b)$ is the normal distribution with mean a,

variance b. For the measure space $(\mathbb{R}^2, \mathcal{F}, \nu)$, we can see that $W(s) \equiv Z[k(s-\cdot)] = \int_{\mathbb{R}^2} k(s-t)Z(\mathrm{d}t)$ defines a Gaussian random variable for each s since the L^2 limit of Gaussian random variables is also Gaussian. In fact, the collection of $W(s)$ is a Gaussian field because its finite dimensional distributions are multivariate normal. For completeness, a direct proof is given in Lemma B.1 of the Appendix. Moreover, it is easy to show that $W(s)$ is stationary with covariance function

$$C(h) = \int_{\mathbb{R}^2} k(h+s)k(r)\nu(\mathrm{d}s) = \int_{\mathbb{R}^2} k(u-h)k(u)\nu(\mathrm{d}u). \tag{4.18}$$

Define the Fourier transform with respect to measure ν (we assume ν is well behaved to ensure the existence of the integral) as $\hat{f}(\omega) = \int_{\mathbb{R}^2} e^{i\omega^T x} f(x)\nu(\mathrm{d}x)$. Under usual regularity conditions, we have

$$\begin{aligned} \hat{C}(\omega) &= \int_{\mathbb{R}^2} \Big[\int_{\mathbb{R}^2} e^{i\omega^T h} k(u-h)\nu(\mathrm{d}h) \Big] k(u)\nu(du) \\ &= \hat{k}(-\omega) \int_{\mathbb{R}^2} e^{i\omega^T u} k(u)\nu(\mathrm{d}u) \\ &= \overline{\hat{k}(\omega)}\hat{k}(\omega) = |\hat{k}(\omega)|^2. \end{aligned} \tag{4.19}$$

The essence of our kernel approximation approach below is based upon the fact that, using (4.18) and (4.19), for certain covariance functions $C(\cdot)$, we can obtain an associated kernel $k(\cdot)$ that produces $C(\cdot)$. Assume $\nu(dt) = dt$, i.e., Lebesgue measure, and $\overline{\hat{k}(\omega)} = \int_{\mathbb{R}^2} e^{i\omega^T u} k(-u)\mathrm{d}u = \int_{\mathbb{R}^2} e^{i\omega^T u} k(u)\,\mathrm{d}u = \hat{k}(\omega)$ for each ω (if and only if $k(u) = k(-u)$). We have

$$\hat{C}(\omega) = \overline{\hat{k}(\omega)}\hat{k}(\omega) = \hat{k}^2(\omega), \tag{4.20}$$

which implies that

$$k(u) = (2\pi)^{-2} \int_{\mathbb{R}^2} e^{-i\omega^T u} \sqrt{\hat{C}(\omega)}\,\mathrm{d}\omega. \tag{4.21}$$

Remark 14. *Though our primary interest is in $\mathbb{R}^2$, all the above discussion can be easily generalized to the $\mathbb{R}^d$ case. For example, in $\mathbb{R}^d$, (4.21) becomes $k(u) = (2\pi)^{-d} \int e^{-i\omega^T u} \sqrt{\hat{C}(\omega)}\, d\omega$.*

Remark 15. *A rich class of stationary processes can be defined in the kernel mixing form of (4.17). Specifically, if a stationary random process can be represented by a kernel mixing form, then this process has a spectral density $\hat{C}(\omega) = |\hat{k}(\omega)|^2$. On the other hand, suppose a process has a spectral density $\hat{C}(\omega)$ and letting $\hat{k}(\omega)$ be any function satisfying $|\hat{k}(\omega)|^2 = \hat{C}(\omega)$. If $\hat{k}(\omega)$ is in L_1, then the inverse Fourier transform of $\hat{k}(\omega)$ exists hence the process has a kernel mixing form representation (see e.g. Yaglom, 1987, Section 26).*

Note that if $k(u)$ depends only on $\|u\|$, where $\|\cdot\|$ indicates the Euclidean distance, then the covariance function is isotropic and there is a one-to-one relationship between $C(\cdot)$ and $k(\cdot)$. In general, formula (4.19) shows that $C(\cdot)$ does not uniquely determine $k(\cdot)$, but it is possible to seek a $k(\cdot)$ which satisfies equation (4.19) as the following examples illustrate.

Example 1: Gaussian covariance function on $\mathbb{R}^2$: $C(h) = \sigma^2 e^{-\|h\|^2/\tau^2}$. It is easy to calculate $\hat{C}(\omega) = \sigma^2 \pi \tau^2 e^{-\tau^2 \omega^T \omega/4}$. Then $k(u) = 2\sigma\pi^{-1/2}\tau^{-1} e^{-2\|u\|^2/\tau^2}$.

Example 2: Matérn covariance function on $\mathbb{R}^d$:

$$C_{\phi,\alpha,\nu}(h) = \frac{\pi^{d/2}\phi}{2^{\nu-1}\Gamma(\nu + d/2)\alpha^{2\nu}} (\alpha\|h\|)^\nu \kappa_\nu(\alpha\|h\|), \tag{4.22}$$

where α is correlation decay parameter, ν is the smoothness parameter, ϕ is the variance parameter, i.e., the process variance $\sigma^2 = \frac{\pi^{d/2}\Gamma(\nu)\phi}{\Gamma(\nu+d/2)\alpha^{2\nu}}$, and κ_ν is the modified Bessel function of the second kind of order ν. The Fourier transform of the Matérn

covariance function is given by, e.g., Stein (1999, p.31),

$$\hat{C}_{\phi,\alpha,\nu}(\omega) = (2\pi)^d \phi(\alpha^2 + \|\omega\|^2)^{-\nu-d/2}. \tag{4.23}$$

Thus,

$$\hat{k}_{\phi,\alpha,\nu}(\omega) = (2\pi)^{\frac{d}{2}} \phi^{1/2}(\alpha^2 + \|\omega\|^2)^{-\nu/2-d/4}. \tag{4.24}$$

Hence

$$k_{\phi,\alpha,\nu}(u) = (2\pi)^{-\frac{d}{2}} C_{\phi^{1/2},\alpha,\nu/2-d/4}(u), \tag{4.25}$$

with the restriction $\nu > d/2$. So, when $d = 2$ we must have $\nu > 1$. We can not apply our kernel solving approach to covariance function that are less than mean square differentiable (see Stein, 1999, Banerjee and Gelfand, 2003). In particular, we can not handle the familiar exponential covariance function.

Example 3: Suppose $C^*(\cdot)$ is a covariance function associated with *geometric anisotropy*, i.e., $C^*(h) = C(h^T B h)$ where B is positive definite and $C(\cdot)$ is a valid isotropic covariance function in $\mathbb{R}^d$. Then straightforward calculation shows that the associated kernel, $k^*(u)$, takes the form $k^*(u) = k(B^{\frac{1}{2}}u)$, where $k(\cdot)$ is the kernel associated with $C(\cdot)$.

Example 4: If $C(\cdot)$ is a valid covariance function in $\mathbb{R}^d$, then convolution of $C(\cdot)$ with itself produces a valid covariance function in $\mathbb{R}^d$. That is, $C \star C(h) = \int_{\mathbb{R}^d} C(h-u)C(u)\,\mathrm{d}u$ is valid (see, e.g., Majumdar and Gelfand, 2005). But then, following (4.18), immediately we can associate $C \star C(h)$ with the process $W(s) = \int_{\mathbb{R}^d} C(s-t)Z(\mathrm{d}t)$, i.e., $k(u) = C(u)$.

Example 5: An example in which a stationary Gaussian random process cannot be written as a kernel mixing form can be constructed as follows (R. Wolpert, personal

communication). In $\mathbb{R}^1$, let $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$ and set $X(t) = Z_1\cos(t) + Z_2\sin(t)$. It is easy to see that $\mathrm{E}\{X(t)\} = 0$ and $\mathrm{Cov}\{X(t_1), X(t_2)\} = \cos(t_1 - t_2)$. So $X(t)$ is a stationary Gaussian process. But $C(h) = \cos(h) = \int_{-\infty}^{\infty} e^{i\omega h}\frac{1}{2}[\delta(\omega-1)+\delta(\omega+1)]\,\mathrm{d}\omega$. So its spectral density is a dirac delta function and it does not have a square root. Thus it does not admit a kernel mixing representation.

4.3 Finite sum approximation

The basic idea behind our approach is to approximate the spatial process $W(s)$ using a linear combination of a set of random variables $\{Z_1, ..., Z_m\}$. The generic form of approximating the vector $\mathbf{W} = (W(s_1), ..., W(s_n))^T$ is

$$\tilde{\mathbf{W}} = G\mathbf{Z}, \tag{4.26}$$

where $\mathbf{Z} = (Z_1, ..., Z_m)^T$ and G is an $n \times m$ matrix (the form of G will become apparent later). We consider three ways of constructing G and $\mathbf{Z}$, namely, i) kernel mixing approximation; ii) projection process approximation; and iii) Karhunen-Loève approximation.

4.3.1 Kernel mixing process approximation

The two subsections here explicitly detail the finite sum kernel mixing process approximation as well as approaches to assess its accuracy.

In practice, once we have associated $k(\cdot)$ with $C(\cdot)$, we introduce a *two-step* approximation to work with the process $W(s) = \int_{\mathbb{R}^2} k(s-t)Z(\mathrm{d}t)$. We restrict the integration to a bounded region D_r and replace the integral by a finite sum. We look at the details a bit more closely. In particular, we consider m grid locations ("knots") $\{t_{rj}, j = 1, ..., m\}$, regardless of the number of sampled locations $s_1, ..., s_n$. As shown in the rest of the chapter, random effects at a relatively small number of

those grid locations will capture most of the variation and association structure of the process $W(s)$. Let D_r be a square region (i.e. $[a,b]\times[a,b]$ for $a,b\in\mathbb{R}$) in $\mathbb{R}^2$ and $D_1\subset D_2...\subset D_r\subset ...\to\mathbb{R}^2$ as $r\to\infty$. Let B_{rj}, $j=1,...,m$ be an equally spaced partition of D_r, i.e., each is a square with edge length $\sqrt{\frac{|D_r|}{m}}$, where $|D_r|$ is the area of D_r. We proceed with the following approximation:

$$\begin{aligned}
\int_{\mathbb{R}^2} k(s-t)Z(\mathrm{d}t) &= \lim_{r\to\infty}\int_{D_r} k(s-t)Z(\mathrm{d}t) \\
&= \lim_{r\to\infty}\sum_{j=1}^{m}\int_{B_{rj}} k(s-t)Z(\mathrm{d}t) \\
&\approx \lim_{r\to\infty}\sum_{j=1}^{m} k(s-t_{rj})\int_{B_{rj}} Z(\mathrm{d}t) \\
&= \lim_{r\to\infty}\sum_{j=1}^{m} k(s-t_{rj})Z(B_{rj}) \\
&\approx \sum_{j=1}^{m} k(s-t_{rj})V_j\sqrt{|B_{rj}|}. \qquad (4.27)
\end{aligned}$$

Here, we envision the t_{rj} as grid points, e.g., the centroids of B_{rj}. The number of the grid points is denoted by 'm' and the V_j are independent normal random variables with mean 0 and variance σ^2. Introducing σ^2 as the variance for the V's implies that k becomes the kernel associated with the *correlation* function of the process.

Remark 16. Note that letting D_r be a square, making B_{rj} an equally spaced partition of D_r, and choosing t_{rj} as the centroid of B_{rj} are just convenient choices to make the limiting argument easier. In practice, one can determine D_r and B_{rj} according to the distribution of locations where the data were collected. For example, if a subset of sample locations $\{s_i, i=1,...,n\}$ is densely clustered in some area, one might want to make B_{rj} small in that area in order to better approximate $W(s)$ for s in that subset.

We define the approximation process $\tilde{W}$ as follows:

$$\tilde{W}_{r,m}(s) \equiv \sum_{j=1}^{m} \sqrt{|B_{rj}|} k(s - t_{rj}) V_j. \tag{4.28}$$

It can be shown that $\tilde{W}_{r,m}(s)$ converges to $W(s)$ in the L^2 sense for every s as $m \to \infty$ and $r \to \infty$. (See Appendix B.2 for a proof.) The covariance matrix of $\tilde{W}_{r,m}(s)$ also converges to the covariance matrix of $W(s)$. To assess the accuracy of the approximation, we have to specify some distance notion for two measures corresponding to the exact and the approximate processes respectively. See Section 4.3.2 for further discussion.

It is easy to see that under kernel mixing approximation, G in (4.26) takes the form $(g(s_1)^T, ..., g(s_n)^T)^T$, where $g(s_i) = (\sqrt{B_{r1}} k(s_i - t_{r1}), ..., \sqrt{B_{r1}} k(s_i - t_{rm}))^T$, and $\mathbf{Z} = \{V_1, ..., V_m\}$ is a vector of independent normal random variables. $\mathbf{Z}$ is associated with "knots" $\{t_{rj}\}$ and G connects locations in $\mathbf{W}$ with locations in $\mathbf{Z}$.

Remark 17. For a process that can not be written as a kernel mixing form, we could attempt to use a partial sum approximation to the Karhunen-Loève expansion or a projection process approximation (see below two sections). Note that here Karhunen-Loève expansion approximation only involves one step of approximation while the kernel mixing approximation involves two. However, solving the integral equation (4.15) for the eigenfunctions $\phi_i(s)$ is far more challenging than implementing the approximation in (4.28).

Remark 18. If the process $\tilde{W}_{r,m}(\cdot)$ is "close" to the process $W(\cdot)$, then one would expect that $f(W(\cdot))$ and $f(\tilde{W}_{r,m}(\cdot))$ should also be "close" for "nice" functions $f(\cdot)$. For example, based on Proposition B.2 in the appendix, we immediately obtain $\mathrm{E}[Y(s) - \tilde{Y}(s)]^2 \to 0$ as $m, r \to \infty$ for every s, where $Y(s) = X^T(s)\boldsymbol{\beta} + W(s) + \epsilon(s)$ and $\tilde{Y}(s) = X^T(s)\boldsymbol{\beta} + \tilde{W}_{r,m}(s) + \epsilon(s)$.

4.3.2 Choice of D_r and m

For the finite sum approximation to the spatial process as given in the formula (4.28), assuming equally spaced partitioning and centroid approximation, how shall we choose the "cover" region D_r and the number of grid points m? Informally, D_r should *cover* the study region while m should be as large as is computationally feasible. Conceptually, we want to encourage the approximate process to be close to the exact process. For example, after defining a suitable distance (or closeness measure) between the true process and its approximation, we may seek to make that distance small under some constraints. Specifically, suppose $d(W, \tilde{W}_{r,m})$ is some distance (or closeness) between two processes W and $\tilde{W}_{r,m}$. If we fix m, we can find

$$\arg \sup_{D_r \in \mathcal{D}} d(W, \tilde{W}_{r,m})$$

for some class of $\mathcal{D}$ (for example, the class of all square areas). Or if we want to control the accuracy of the approximation, for a fixed small positive number ϵ, we can find a $D_r \in \mathcal{D}$ with m as small as possible such that $d(W, \tilde{W}_{r,m}) < \epsilon$.

Let us now consider a few examples of $d(W, \tilde{W}_{r,m})$. The natural choices would be

$$\sup_{s \in S} (\mathrm{E}(|W(s) - \tilde{W}_r(s)|^p))^{\frac{1}{p}}, \tag{4.29}$$

or

$$\int_{s \in S} (\mathrm{E}(|W(s) - \tilde{W}_r(s)|^p))^{\frac{1}{p}} \, \mathrm{d}s, \tag{4.30}$$

where $p > 0$ and S is some bounded region in $\mathbb{R}^2$ (e.g. the study region D). But the calculations of (4.29) and (4.30) are difficult.

Another idea is to choose D_r and m to make the following probability to be small for a positive number ϵ,

$$\Pr(\sup_{s \in S} |W(s) - \tilde{W}_{r,m}(s)| > \epsilon). \tag{4.31}$$

Note that the set

$$\{\sup_{s\in S}|W(s)-\tilde{W}_{r,m}(s)|>\epsilon\}=\bigcup_{s\in S}\{|W(s)-\tilde{W}_{r,m}(s)|>\epsilon\}$$

in (4.31) is the union of uncountably many measurable sets, it is measurable if $W(s)$ and $\tilde{W}_{r,m}(s)$ have continuous sample paths. The reason is that $\bigcup_{s\in S}\{|W(s)-\tilde{W}_{r,m}(s)|>\epsilon\}=\bigcup_{s_i\in S}\{|W(s_i)-\tilde{W}_{r,m}(s_i)|>\epsilon\}$ is a countable union of measurable sets, where two coordinate values of s_i are rational. Again, (4.31) is not easy to compute.

From practical point of view, since the $\{V_j\}$ in (4.28) are Gaussian, the Kullback-Leibler divergence between the joint distribution f_W of a set of random variables $\{W(s_i), i=1,...,n\}$ and the joint distribution $f_{\tilde{\mathbf{W}}}$ of the set of random variables $\{\tilde{W}_{r,m}(s_i)\}$ for the same set of s_i is easy to compute. We have

$$\mathrm{KL}(f_{\mathbf{W}},f_{\tilde{\mathbf{W}}})=-\frac{1}{2}\log|K_{\tilde{\mathbf{W}}}^{-1}K_{\mathbf{W}}|+\frac{1}{2}\mathrm{Tr}(K_{\tilde{\mathbf{W}}}^{-1}K_{\mathbf{W}}-\mathrm{I}_n), \tag{4.32}$$

where $K_{\mathbf{W}}$ and $K_{\tilde{\mathbf{W}}}$ are the covariance matrices of $W(s_i)$ and $\tilde{W}_{r,m}(s_i)$, respectively, I_n is an $n\times n$ identity matrix. One could imagine increasingly dense s_i and attempt to compute the limiting Kullback-Leibler divergence. Instead, for a fixed m and n, we can find (at least empirically) the D_r such that $\mathrm{KL}(f_{\mathbf{W}},f_{\tilde{\mathbf{W}}})$ is minimized.

Another easy calculation is to find the covariance matrix $K_{\mathbf{W}-\tilde{\mathbf{W}}}$ for the vector of $W(s_i)-\tilde{W}_{r,m}(s_i), i=1,...,n$, and roughly minimize its determinant or its trace. A natural choice for the set of $\{s_i\}$ would be a subset of the n data locations. The

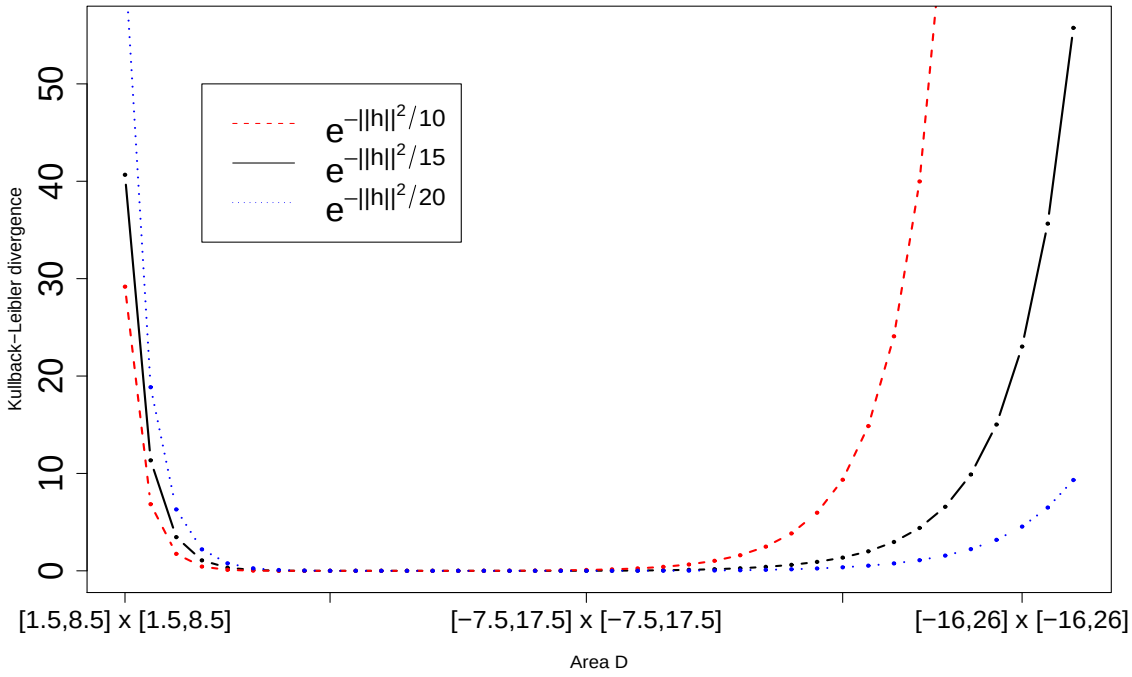

Figure 4.1: The effect of choice of D_r.

diagonal terms of $K_{\mathbf{W}-\tilde{\mathbf{W}}}$ can be calculated as follows,

$$\begin{aligned} & \mathrm{Var}(W(s) - \tilde{W}_{r,m}(s)) \\ = \; & \sigma^2 \Big[\int_{\mathbb{R}^2} k^2(s-t)\,\mathrm{d}t + \frac{|D_r|}{m} \sum_{j=1}^{m} k^2(s-t_{rj}) - 2\sum_{j=1}^{m} (k(s-t_{rj}) \int_{B_{rj}} k(s-t)\,\mathrm{d}t) \Big]. \end{aligned} \quad (4.33)$$

To illustrate, we consider the simulation example in Section 4.5 below. Observations are taken on a $[0,10] \times [0,10]$ square and then we fix the set of data locations at the "center" of D_r while we slowly expand D_r and adjust t_{rj} accordingly (fix m=100). Figure 4.1 shows how the Kullback-Leibler divergence changes as we change the cover region (three expansions are marked on the x-axis) using three different range parameters. The good news is that to some extent the approximation performance seems fairly "robust" over a range of choices for D_r. We also see that the stronger the spa-

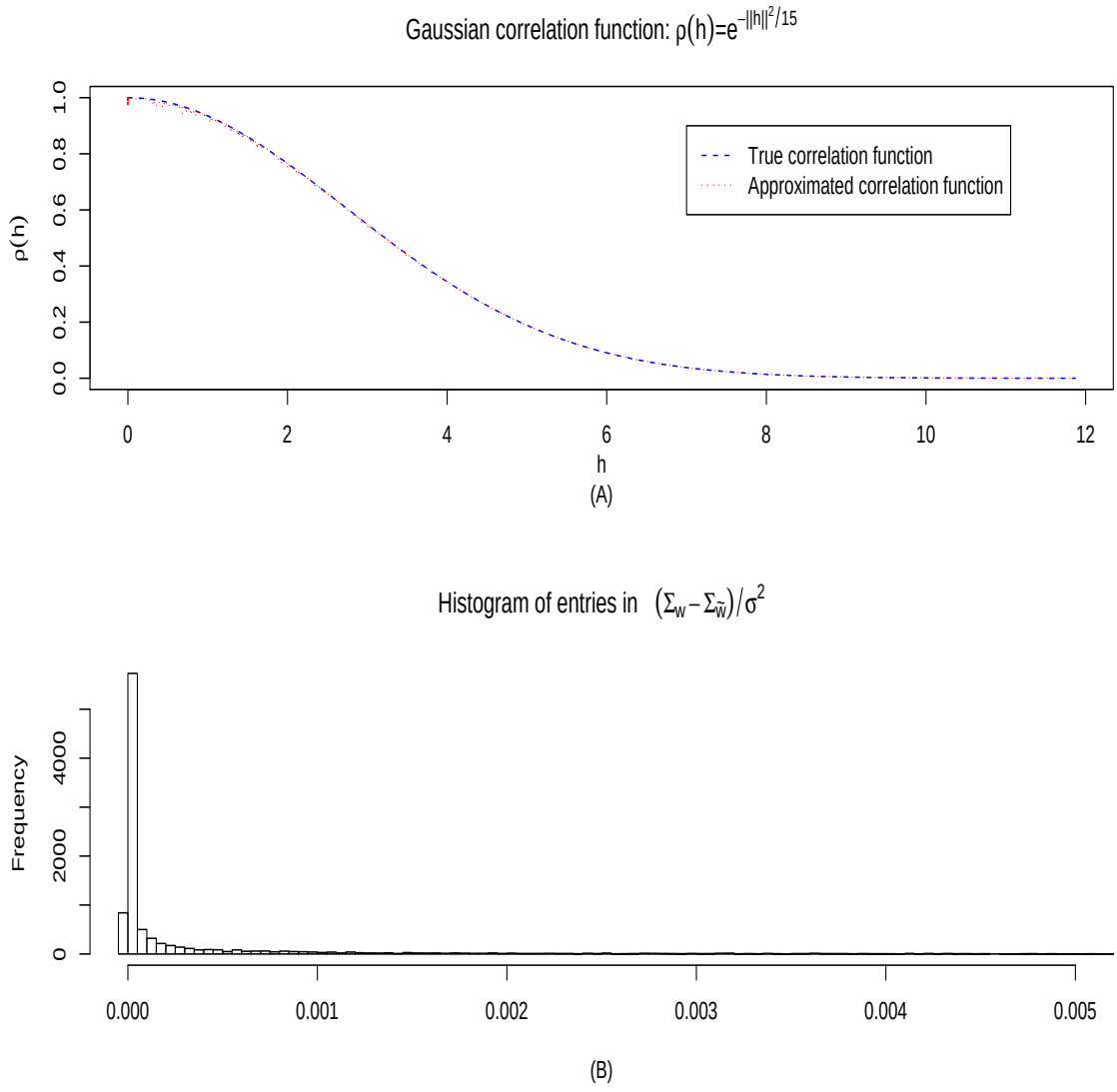

Figure 4.2: Performance of approximating Gaussian covariance function $C(h) = \sigma^2 \exp(-\|h\|^2/\phi)$ (σ^2=1, ϕ=15) based on kernel mixing.

tial correlation is, the less sensitive the performance of the approximation is to the choice of D_r. In Figure 4.2 the upper graph (A) shows the true Gaussian correlation function as well as the approximated correlation function. The lower graph (B) shows the histogram of the entries in $(K_{\mathbf{W}} - K_{\tilde{\mathbf{W}}})/\sigma^2$ for $D_r = [-2.5, 12.5] \times [-2.5, 12.5]$. The performance of the approximation is satisfying.

4.3.3 Projection process approximation

Given a study region D, suppose we observe a spatial process W at $s_1^*, ..., s_m^*$ and wish to predict W at a new location $s_0 \in D$. Assume $W(s)$ to be weakly stationary with mean zero and covariance function $C(\cdot)$. It is not too hard to show that the best linear prediction for $W(s_0)$ given $\mathbf{W}^* = \mathbf{w}^*$, where $\mathbf{W}^*=(W(s_1^*), ..., W(s_m^*))^T$ and $\mathbf{w}^*=(w(s_1^*), ..., w(s_m^*))^T$, is

$$\tilde{W}(s_0) = h_0^T K^{*-1} \mathbf{w}^* \tag{4.34}$$

with the mean square error

$$C(0) - h_0^T K^{*-1} h_0, \tag{4.35}$$

where h_0=Cov$\{\mathbf{W}^*, W(s_0)\}$, K^*=Cov(W^*, W^*), and $C(0)$=Cov$\{W(s_0), W(s_0)\}$.

If we assume $W(s)$ to be a Gaussian process, then the prediction of $W(s_0)$ should be based on the conditional distribution of $W(s_0)$ given $\mathbf{W}^*=\mathbf{w}^*$. That is,

$$W(s_0)|\mathbf{W}^* \sim \mathrm{N}(h_0^T K^{-1} \mathbf{w}^*, C(0) - h_0^T K^{-1} h_0). \tag{4.36}$$

Mathematically speaking, the prediction of $W(s_0)$ given $\mathbf{W}^*$ is simply the *projection* of $W(s_0)$ on a *linear manifold spanned* by $\{W(s_1^*), ..., W(s_n^*)\}$. From a theoretical point of view, we may consider a more general problem of predicting $W(s_0)$ based on having observed the process W on some set $\mathcal{A} \subset D$. $\mathcal{A}$ could be finite, countable, or uncountable. Now the linear manifold is generated by $\{W(s), s \in \mathcal{A}\}$ and linear algebra is not adequate to study prediction if the set $\mathcal{A}$ has infinite elements. Hilbert space is naturally introduced to study projection in this situation. For our purpose, the Hilbert space here is generated by the process $W(s)$ on D. More specifically, it is the closure of the linear manifold of $\{W(s), s \in D\}$, denoted by $\mathcal{H}_D(0, C)$ (C is the covariance function), with the inner product $(h_1, h_2) = \mathrm{E}(h_1 h_2)$, where $h_1, h_2 \in \mathcal{H}_D(0, C)$.

Now the prediction problem becomes clear: $W(s_0)$ is an element in $\mathcal{H}_D(0, C)$ and its prediction given observing W on $\mathcal{A}$ is simply the projection of $W(s_0)$ onto the subspace $\mathcal{H}_{\mathcal{A}}(0, C)$. The projection theorem of Hilbert space tells us that there exists a unique element $\tilde{W}(s_0)$ in $\mathcal{H}_{\mathcal{A}}(0, C)$ such that

$$\|W(s_0) - \tilde{W}(s_0)\| = \inf_{V \in \mathcal{H}_{\mathcal{A}}(0,C)} \|W(s_0) - V\|. \tag{4.37}$$

Now it should be apparent how to approximate $W(s)$ based on projection ideas. First we define a set of *"knots"* (or grid points) $\{s_1^*, ..., s_m^*\}$, then we define a process $\tilde{W}(s)$ to approximate $W(s)$ as

$$\tilde{W}(s) = h(s)^T K^{*-1} \mathbf{W}^*, \tag{4.38}$$

where $h(s) = \text{Cov}\{\mathbf{W}^*, W(s)\}$, $\mathbf{W}^* = (W(s_1^*), ..., W(s_m^*))^T$ and $K^* = \text{Cov}(\mathbf{W}^*, \mathbf{W}^*)$. $\tilde{W}(s)$ is the "projection" of $W(s)$ onto the manifold spanned by $\mathbf{W}^*$. Loosely speaking, if $\mathbf{W}^*$ represents "good" amount of information for the process W on D, $\tilde{W}(s)$ would be a good approximation. The $(i, j)^{th}$ element of the covariance matrix $K_{\tilde{\mathbf{W}}}$ of the vector $\tilde{\mathbf{W}}$=$(\tilde{W}(s_1), ..., \tilde{W}(s_n))$ is

$$\text{Cov}\{\tilde{W}(s_i), \tilde{W}(s_j)\} = h(s_i)^T K^{*-1} h(s_j). \tag{4.39}$$

The above projection calculation does not depend on the probability law of the process. If we assume $W(s)$ is a Gaussian process, we know from (4.36) that,

$$\tilde{W}(s) = \text{E}(W(s)|\mathbf{W}^*). \tag{4.40}$$

We wish the approximation process $\tilde{W}(s)$ to be "close" to $W(s)$. Intuitively, as m tends to ∞, we expect $K_{\tilde{\mathbf{W}}} - K_{\mathbf{W}}$ (or $K_{\tilde{\mathbf{W}}|\mathbf{W}}$) goes to the zero matrix. More precisely, if $\{W(s^*), s^* \in \mathcal{A}\}$ is a basis of $\mathcal{H}_D(0, C)$, perfect prediction is possible, i.e. $K_{\tilde{\mathbf{W}}|\mathbf{W}} = 0$.

Projection approximation provides an easy way to approximate a desired spatial process (both stationary and nonstationary), not requiring to find the kernel function (which is needed in the kernel mixing approximation) or the eigenvalues and eigenfunctions (which is needed in the Karhunen-Loève approximation). G in (4.26) takes the form $(h(s_1)^T K^{*-1}, ..., h(s_n)^T K^{*-1})^T$ and $\mathbf{W}^*$ is the $\mathbf{Z}$ vector in (4.26). $\mathbf{W}^*$ is associated with "knots" and G relates locations in $\mathbf{W}$ with locations in $\mathbf{W}^*$. It is worth noting that the components of $\mathbf{W}^*$ are dependent normal random variables, while $\{Z_1, ..., Z_m\}$ in the kernel mixing process approximation are independent normal random variables.

Let us consider a concrete example. Let $D = [0, 1]$ and $C(h) = \sigma^2 \exp(\phi|h|)$. For computational simplicity, suppose we observe W at an even number of equally spaced knots $\{0, \Delta, 2\Delta, ..., (m/2-1)\Delta, (m/2)\Delta, ..., (m-2)\Delta, 1\}$ and we want to approximate $W(1/2)$, where $\Delta = 1/(m-1)$ and m is the number of knots. Let $\rho = \exp(-\phi\Delta)$. So, $h(1/2) = (\rho^{(m-1)/2}, \rho^{(m-3)/2}, ..., \rho^{1/2}, \rho^{1/2}, ..., \rho^{(m-1)/2})^T$. From Appendix A.1, we can compute K^{*-1} explicitly and after some algebra we get

$$\mathrm{Var}\{\tilde{W}(1/2)\} = h(1/2)^T K^{*-1} h(1/2) = \sigma^2 \frac{2\rho}{1+\rho} \to \sigma^2, \text{as } m \to \infty \ . \qquad (4.41)$$

In other words, $\mathrm{Var}(W(s)|\mathbf{W}^*) \to 0$ as $m \to \infty$. So for the one dimensional OU model, we can do prefect prediction (kriging) if we know the model parameters. Similarly, we can show $\mathrm{Cov}\{\tilde{W}(s_i), \tilde{W}(s_j)\} \to \sigma^2 \exp(-\phi|s_i - s_j|)$ as $m \to \infty$. In this example, as $m \to \infty$, $\{W(s), s \in \{0, \Delta, ..., (m-2)\Delta, 1\}\}$ becomes a basis for the Hilbert space $\mathcal{H}_{[0,1]}(0, C)$, where $C(h) = \sigma^2 \exp(\phi|h|)$.

An natural question to ask is how to choose the number and locations of the knots. Large m increases the performance of approximation but makes the computation less efficient. Generally, we choose m as large as the computation power allows. For a fixed m, we shall choose the set of knots $\{s_1^*, ..., s_m^*\}$ such that $K_{\tilde{\mathbf{W}}} = G K_{\mathbf{W}^*} G^T$ is

as close to $K_{\mathbf{W}}$ as possible. See Section 4.3.2 for more discussion. In general, the inverse of the covariance matrix K^* is not available in a closed form. We may study the performance of the approximation through numerical simulations. Figure 4.3 shows an approximated Matérn correlation function based on the forgoing projection ideas.

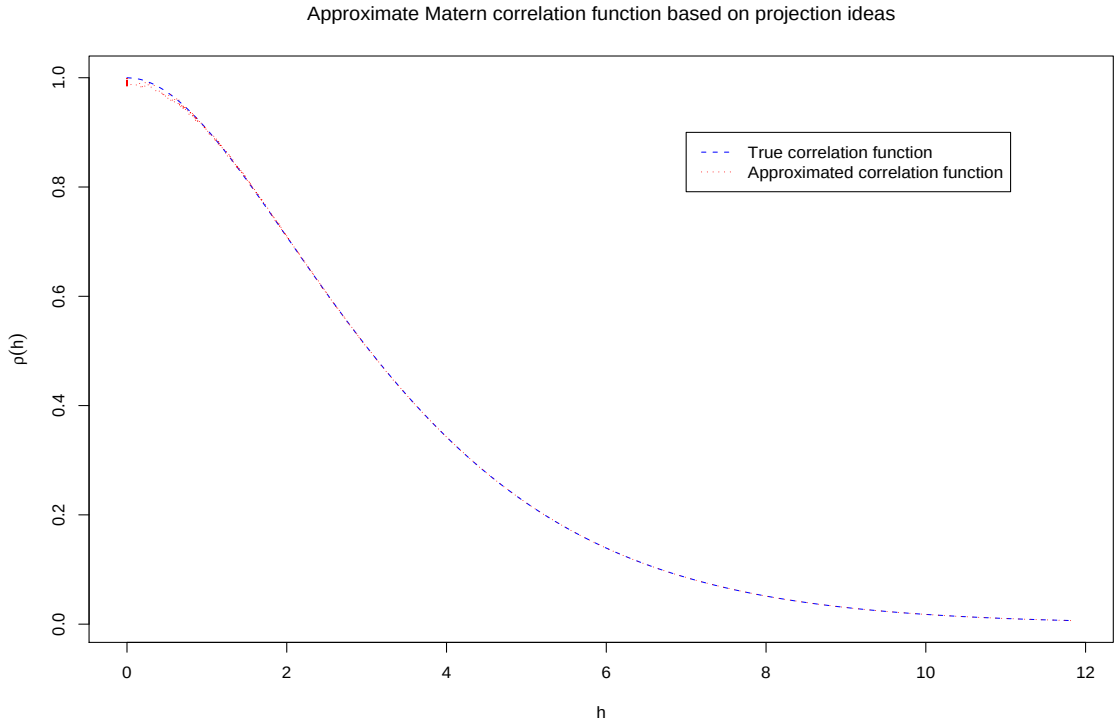

Figure 4.3: Approximate Matérn covariance function $C(h) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)}(\phi\|h\|)^{\nu}\kappa_{\nu}(\phi\|h\|)$ (σ^2=1, ϕ=2/3, ν=2) using projection process approximation.

4.3.4 Karhunen-Loève approximation

From Section 4.2.1, we know the Gaussian process $W(s)$ admits the following Karhunen-Loève expansion representation.

$$W(s) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\phi_i(s)Z_i$$

So we can approximate $W(s)$ by truncating the infinite series, i.e.,

$$\tilde{W}(s) = \sum_{i=1}^{m} \sqrt{\lambda_i}\phi_i(s)Z_i, \tag{4.42}$$

where $\{\lambda_i, i = 1, ..., m\}$ are first m biggest eigenvalues and $\{\phi_i(\cdot)\}$ are associated eigenvectors.

Now G in (4.26) takes the form $(g(s_1)^T, ..., g(s_n)^T)^T$, where

$$g(s_i) = (\sqrt{\lambda_1}\phi_1(s_i), ..., \sqrt{\lambda_m}\phi_m(s_i))^T.$$

$\mathbf{Z} = \{Z_1, ..., Z_m\}$ in (4.26) are independent normal random variables. Note that in Karhunen-Loève approximation, there is no notion of "knots" any more. A row vector of G is a vector of basis functions and $\{Z_1, ..., Z_m\}$ are corresponding random coefficients. Karhunen-Loève expansion is one way of finding basis functions (they are orthogonal). Polynomials, wavelets, Fouriers, and splines can also serve as basis functions.

Performing Karhunen-Loève expansion approximation depends on the ability to solve the integral equation (4.15). Typically it is a hard task and no closed-form solutions are available except for a few special cases. In Chapter 3, we have seen how to find the eigenvalues and eigenfunctions for the one-dimensional exponential covariance function case (see Section 3.2 and Appendix A.1). We also discuss an easy approach to find approximate eigenvalues and eigenfunctions. There are also

numerical methods, for example, the "Galerkin method", to find approximate eigenvalues and eigenfunctions. The Karhunen-Loève expansion offers the possibility of an approximation when a process can not be represented using kernel mixing.

4.4 Formalizing the analysis in a Bayesian framework

Returning to the model in (4.1), we have $W(s)$ a mean zero Gaussian process with covariance function $\sigma^2\varrho(\cdot;\boldsymbol{\phi})$ where $\boldsymbol{\phi}$ denotes the parameters in ϱ, hence in the kernel $k(\cdot)$. For example, if we started with a Matérn correlation function, $\boldsymbol{\phi}$ would include a scale parameter α and a smoothness parameter ν. With samples from n locations we obtain

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{W} + \boldsymbol{\epsilon}, \tag{4.43}$$

where $\mathbf{W} \sim \mathrm{N}_n(0, \sigma^2 R(\boldsymbol{\phi}))$ with $R(\boldsymbol{\phi})_{ij} = \varrho(s_i - s_j; \boldsymbol{\phi})$.

We shall use $\tilde{\mathbf{W}}$ in (4.26) to approximate $\mathbf{W}$. Kernel mixing approximation form (4.28), or projection process approximation form (4.38), or Karhunen-Loève approximation form (4.42) (if available) can be chosen. The dependence in the $\mathbf{Z}$ vector (e.g. in projection process approximation) may introduce more computation demand.

For illustration, we use $\tilde{W}_{r,m}(s)$ as in (4.28) to approximate $W(s)$, the approximate model becomes

$$Y = X\boldsymbol{\beta} + \tilde{\mathbf{W}} + \epsilon, \tag{4.44}$$

where, suppressing the subscript r and m, $\tilde{\mathbf{W}} = G(\boldsymbol{\phi})\mathbf{V}$; $G(\boldsymbol{\phi})_{ij} = k(s_i - t_j; \boldsymbol{\phi})\sqrt{|B_j|}$ $(i = 1, ..., n,\ j = 1, ..., m)$; and $\mathbf{V}$=$(V_1, ..., V_m) \sim \mathrm{N}_m(0, \sigma^2\mathrm{I})$. We have $\tilde{\mathbf{W}} \sim \mathrm{N}_n(0, \tilde{K})$

and $\tilde{K} = \sigma^2 G(\boldsymbol{\phi}) G(\boldsymbol{\phi})^T$ where, for clarity,

$$\tilde{K}_{ij} = \sigma^2 \sum_{l=1}^{m} k(s_i - t_l; \boldsymbol{\phi}) k(s_j - t_l; \boldsymbol{\phi}) |B_l|. \tag{4.45}$$

The parameters are $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2)$. Fitting the approximate model in a Bayesian framework is straightforward using MCMC methods. In particular, we work in the space of the latent $\mathbf{V}$'s updating $\mathbf{V}$ in addition to $\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\phi}$.

Finally, to implement spatial prediction (kriging) (see Section 2.2.5), we would again use the approximate model in (4.44). That is, even if we have fitted the original model, fully Bayesian kriging requires conditional distributions for new locations given the sampled locations, again bringing in the "large n problem" (see, e.g., Banerjee et al., 2004, Chapter 5 for details). Based on the approximate model, the prediction of the response $Y(s_0)$ at a new location s_0 is

$$\begin{aligned} & f(y(s_0)|\mathbf{Y}) \\ = & \int f(y(s_0)|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{V}, \boldsymbol{\phi}, \tau^2) f(\boldsymbol{\beta}, \mathbf{V}, \boldsymbol{\phi}, \tau^2|\mathbf{Y}) \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\mathbf{V} \, \mathrm{d}\tau^2 \, \mathrm{d}\boldsymbol{\phi}. \end{aligned} \tag{4.46}$$

Obviously we can extend (4.46) to simultaneously prediction at a collection of new locations. In practice, Monte Carlo methods are used to obtain estimates of (4.46) based on the posterior draws of the parameters. Note that, conditioned on $\mathbf{V}$, $Y(s_0)$ is independent of $\mathbf{Y}$. So, under the approximation model, there is no "large n problem" in prediction.

4.5 Examples

In order to appreciate the computational advantage of our approximate model and examine its performance, we illustrate with a simulated dataset and a real dataset of single family house prices.

4.5.1 Simulation example

In this simulation study, we generated 10050 data points in a $[0, 10] \times [0, 10]$ square from a Gaussian random field with, without loss of generality, constant mean μ and covariance matrix $\sigma^2 R(\phi^2) + \tau^2 \mathrm{I}$, where $R(\phi^2) = e^{-\|h\|^2/\phi^2}$, resulting in a Gaussian kernel $k(\cdot)$. The true values of the parameters are $\mu = 0$, $\sigma^2 = 5$, $\phi^2 = 15$, and $\tau^2 = 1$. Figure 4.4 gives the plot of 10050 data locations with an illustrative choice of 100 grid locations (empty circles).

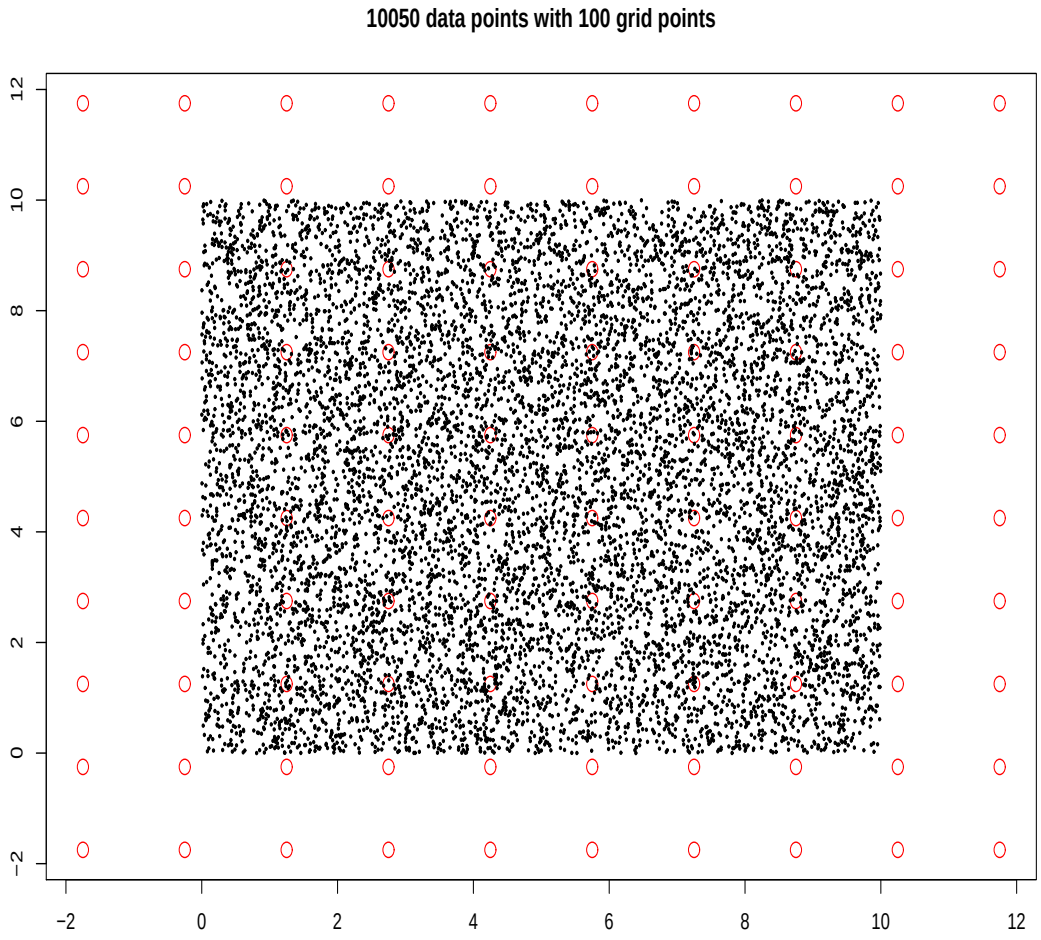

Figure 4.4: 10050 data points in a $[0, 10] \times [0, 10]$ square with 100 grid points($\circ$).

We left out 50 data points for the purpose of prediction. Then we fitted the finite sum approximation model using 100 grid points to the remaining 10000 data points. We choose $[-2.5, 12.5] \times [-2.5, 12.5]$ as the cover region. Also, for comparison, we randomly subsampled n_0=100 data points from the set of 10000 data points and estimated the parameter values by fitting the exact model without approximation to these 100 points. (One could attempt to introduce sampling design consideration in the selection of this subsample of locations but we have not pursued this here.) The results are shown in Table 4.1, where $\hat{\mu}_{m=100}$ denotes the estimation for μ under the approximation model with $m = 100$ and $\hat{\mu}_{n_0=100}$ is under the subsampling model with n_0=100. It can be seen that the approximate model recovers the parameters very well. The posterior mean and median of the parameters are fairly close to the true values. On the other hand, the estimation based on the exact model with subsampling is not as good as that based on the full data with the approximate model. For the latter, the point estimators are closer to the true values and the credible intervals tend to be shorter.

Table 4.1: Posterior summaries for the simulated data example.

Parameter	Median	Mean	95% interval
$\hat{\mu}_{m=100}$	0.36	0.44	(-0.54 , 1.60)
$\hat{\mu}_{n=100}$	0.70	0.74	(-0.88 , 2.58)
$\hat{\phi}^2_{m=100}$	14.71	14.75	(12.04 , 17.98)
$\hat{\phi}^2_{n=100}$	11.60	12.05	(6.24 , 20.36)
$\hat{\sigma}^2_{m=100}$	5.23	5.65	(2.95 , 10.01)
$\hat{\sigma}^2_{n=100}$	3.46	3.92	(1.67 , 9.06)
$\hat{\tau}^2_{m=100}$	0.996	0.997	(0.97 , 1.02)
$\hat{\tau}^2_{n=100}$	0.994	0.995	(0.79 , 1.48)
N=10000, $\mu = 0$, $\phi^2 = 15$, $\sigma^2 = 5$, $\tau^2 = 1$			

In this regard, we note that the 95% credible intervals for μ, for ϕ^2, and for σ^2 are quite wide even when the sample size is relatively large. In fact, the parameter μ cannot be consistently estimated as long as the sampling locations are restricted to a bounded region. See Xia et al. (2006) for details. Also it is well known that, for customary covariance functions, σ^2 and the decay parameters ϕ^2 are not well-identified. See Stein (1999, Chapter 4), Zhang (2004), and Xia et al. (2006) for extensive discussions.

Figure 4.5 turns to the 50 held out points and shows the predictive median and the 95% predictive intervals using these two analysis. Most of the predictive intervals based on the approximate model are tighter than those based on the subsampling model. In both cases, 2 out of 50 true values fall out of the 95% predictive intervals. We also computed the value of mean square predictive error (MSPE) $\sum_{i=1}^{50}(\hat{Y}_{i,l} - Y_i^{obs})^2$, $l = 1, 2$, where l=1 refers to the approximate model and l=2 to the subsampling model. Y_i^{obs} are true values of the left out data and $\hat{Y}_{i,l}$ are their predictions under the respective models. The MSPE under l=1 is 50.1 while under l=2 the MSPE is 58.0. In both cases, we have 100 latent random effects but the approximation strategy is 15% better with respect to MSPE.

4.5.2 Baton Rouge housing data example

There is expected to be spatial pattern in selling prices of houses adjusting for various physical characteristics such as square feet of the living area, age of the house, number of bathrooms, etc. So, spatial modeling for house price data is natural (see e.g. Gelfand et al. 2004). Here we consider a real estate dataset with observations at 8774 locations in the city of Baton Rouge, Louisiana. Figure 4.6 shows the data locations (defined by latitude and longitude) as well as 96 grid points used for the

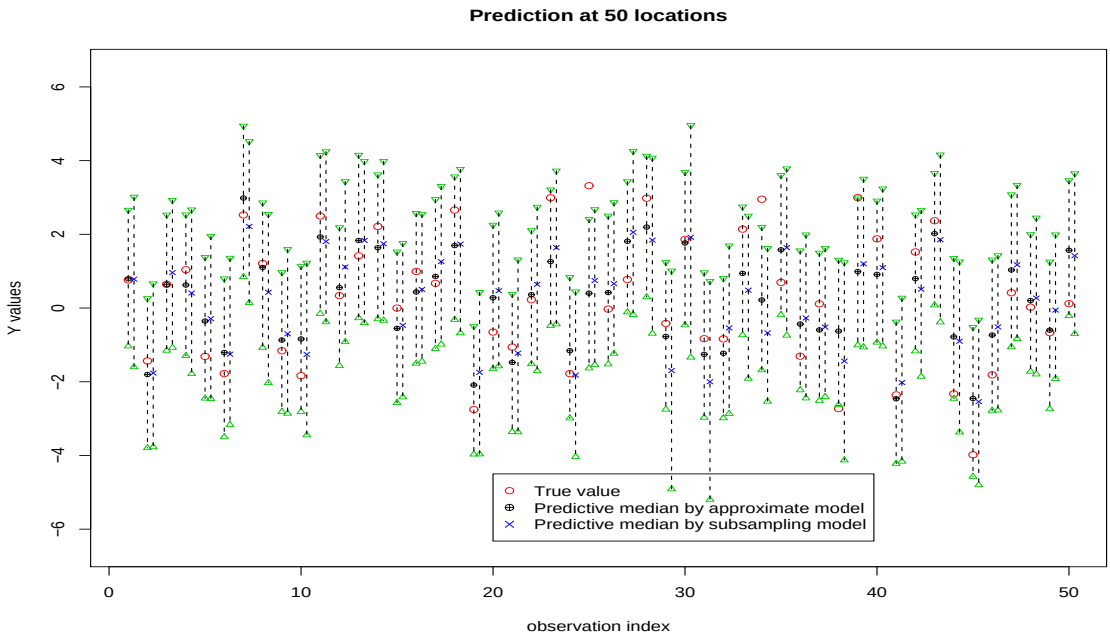

Figure 4.5: Prediction at 50 locations.

approximate spatial model. It is customary to model the response $Y(s)$ as log-selling price of the property at location s. Illustratively, we use age, living area, other area (e.g. patios, garages, and carports), and number of bathrooms as covariates in our analysis.

We fit the Bayesian model described in Section 4.4 for this dataset. We use a Matérn covariance function specified in (4.22), resulting in a four-parameter variance-covariance specification. With nearly 9000 sales, fitting the exact model in (4.43) is infeasible. Using the approximate model in (4.44), the run time is about 1.2 hours per 1000 iterations using C code on a Linux machine with a 3.4 GHz Intel Pentium 4 processor (the evaluation of Bessel function in the Matérn class is computational costly). Table 4.2 provides the posterior summaries for the coefficients of house characteristics as well as the parameters in the covariance structure. It can be seen

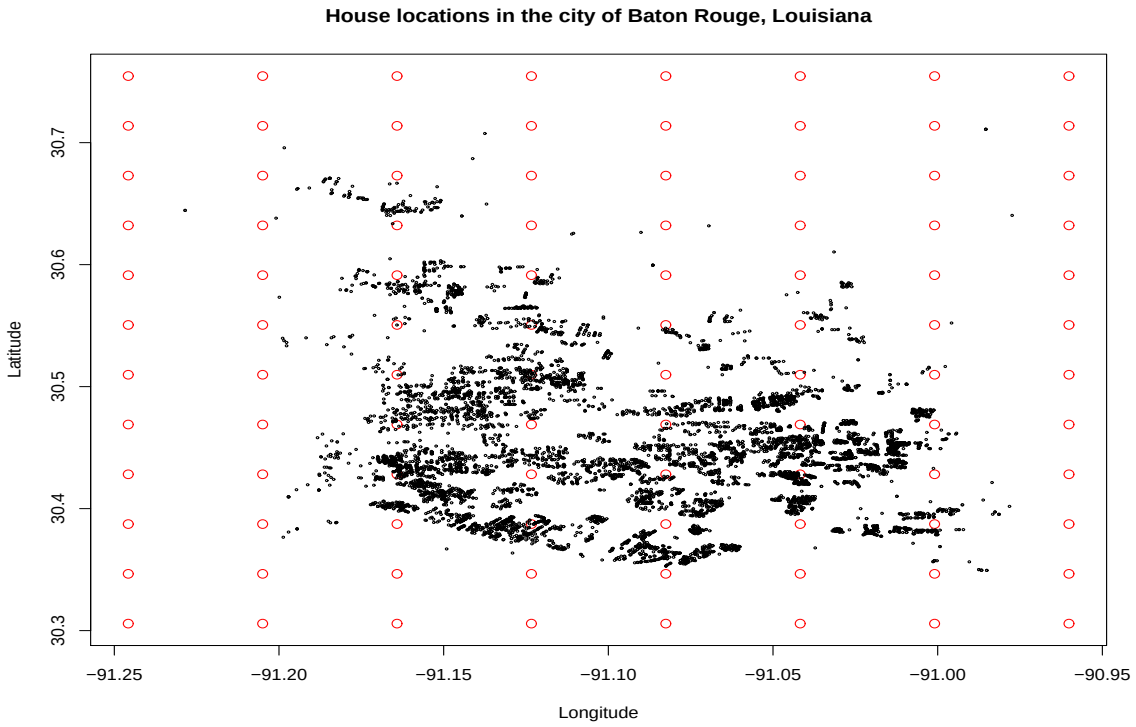

Figure 4.6: Baton Rouge house locations (∘ denotes the grid points).

that all covariates are significent. The proportion of spatial variation (σ^2) to total variation ($\sigma^2+\tau^2$) is around 74%. The range is around 23% of the maximum distance.

Table 4.2: Posterior summaries for Baton Rouge housing data.

Variable	Parameter	Median	Mean	95% interval
Intercept	$\hat{\beta}_0$	10.18	10.18	(10.08 , 10.29)
Age	$\hat{\beta}_1$	-0.0020	-0.0020	(-0.0024 , -0.0015)
Bathrooms	$\hat{\beta}_2$	0.069	0.069	(0.054 , 0.085)
Living area	$\hat{\beta}_3$	0.00038	0.00038	(0.00036 , 0.00039)
Other area	$\hat{\beta}_4$	0.00023	0.00023	(0.00021 , 0.00025)
Smoothness parameter	$\hat{\nu}$	1.90	1.89	(1.80 , 1.99)
Decay parameter	$\hat{\alpha}$	0.50	0.49	(0.44 , 0.54)
Spatial variance parameter	$\hat{\phi}$	0.0086	0.0087	(0.006 , 0.014)
Nugget parameter	$\hat{\tau}^2$	0.068	0.068	(0.066 , 0.070)

4.6 Extensions

In this chapter, we have developed a conceptually simple and computationally straightforward process approximation approach for fitting spatial models to data observed at a very large number of locations. We have implemented the approximation within a Bayesian framework, achieving relatively short run times and satisfying performance. It is important to note the generality of our process approximation idea. Specifically, with a finite sum approximation for the process $W(s)$ as in (4.26), all of the discussion in Section 4.4 will apply. For example, the process $W(s)$ also could be represented and approximated using other forms of basis functions, e.g., polynomials, Fouriers, wavelets, splines, etc.

Developing analogues of our approximation strategy for spatio-temporal (and multivariate response) data is a natural extension. We can envision dimension reduction in both space and time. For a very large number of spatial locations and time points, an additive approximation in space and time may provide the only feasible implementation. Extension to a spatial model with a nonGaussian first stage (i.e. the response Y may be binary or a count) is obvious. For example, if Y takes only 0 or 1 values, the first stage of the model might be that $Y(s_i)$ are conditionally independent Bernoulli random variables given $W(s_i)$, covariates and other parameters. In the second stage, we could approximate $W(s_i)$ using (4.26).

Chapter 5

Spatial Design

5.1 Introduction

Spatial sampling design is an important area in spatial statistics. The basic problem is to decide how to choose a set of sample locations so that a maximum amount of information can be obtained. For example, how shall we sample house locations to learn the real estate market in a particular area? Or, how shall we sample individuals within a region to measure lead level in the blood?

Our focus in this chapter is on approximately optimal spatial design in the case of one-time sampling at a large number of spatial locations. We address the design problem in the context of environmental health research. But the application of spatial design in other areas can be easily envisioned.

Environmental health research considers the relationship between exposure to environmental contaminants and particular health endpoints. Many environmental health issues are characterized by spatial structure in either the contaminant surfaces or the pattern of observed cases. Thus, spatial modeling is making rapid inroads in environmental health. For exposure, which is our focus, models that explicitly include spatial structure provide better explanation of contaminant surfaces both with regard

to estimation of levels and the uncertainty in this estimation.

By definition, if exposure surfaces are envisioned as conceptually measurable at every (point) location in a study region, then such surfaces are inherently spatial in nature. Anticipating spatial association in contaminant levels, with an uncountable collection of locations, we naturally turn to point-referenced association models, i.e., spatial process models. In this chapter, our attention is to a particular aspect of sampling design: How shall we choose locations to sample exposure levels (possibly ambient or deposition) that are anticipated to be essentially static? For example, how shall we sample individuals within a region to measure contaminant levels in the blood? Or, how shall we sample locations to learn about ambient levels of air toxics or perhaps arsenic levels in the water table? We are focusing on one-time sampling at a large number of locations rather than designing long-term typically sparse monitoring networks. Thus, we are not considering the costs for installing, operating and maintaining a network but rather the cost of collecting a single observation. If we plan to use spatial processes in building models to analyze such data, it seems equally appropriate to use such models in developing the sampling design.

The criteria we focus on are developed from the Fisher information matrix with the goal of learning not only about the regression structure in the model but also about the dependence structure. Under a criterion that attempts to maximize information gain, we consider three strategies to develop an approximately optimal design: sequential sampling, block sampling, and stochastic search. We also discuss utility-based modification of these strategies to achieve oversampling with regard to specified objectives. We present some theoretical and empirical properties and relationships among these strategies and provide an illustrative implementation for a simulated dataset. We also describe a real application in the context of the Toxics Release Inventory (TRI).

5.2 Background and literature review

A brief review of the history of spatial modeling for environmental health may be useful. Two broad paths have been followed. The first views the surface as a random realization of a spatial process above two-dimensional space. Measurements are taken at point-referenced (geo-coded) spatial locations. Inference involves fitting an explanatory process model using these measurements. In some cases, exposure levels are essentially stable and static modeling based upon single measurements at individual locations is the objective. In other cases, the locations are monitoring stations whence data collection is dynamic and a temporal component is added to the modeling to capture evolution of the contaminant surface over time. The literature on spatial and spatio-temporal process modeling in environmental health is substantial. Noteworthy examples for the static case include Le and Zidek (1992); Brown, Le and Zidek (1994); Shaddick and Wakefield (2002); and Schmidt and Gelfand (2003). Examples in the dynamic setting include Guttorp, Meiring and Sampson (1994); Huerta, Sanso and Stroud (2004) and Sahu, Gelfand and Holland (2005). Gelfand, Banerjee and Gamerman (2005) provide a general dynamic modeling development for univariate and multivariate spatial data settings.

The second path has focused on areal partitions of the study region into, for example, census units, zip codes, or counties. Typically, counts of some adverse health outcome are aggregated to these units (usually for purposes of confidentiality). Environmental risk factors are supplied for these areal units to explain the counts. Spatially structured random effects are introduced to provide spatial smoothing of the counts. Work here dates to Clayton and Kaldor (1987). See also Bernardinelli and Montomoli (1992), Knorr-Held (2002), and Zhu, Carlin and Gelfand (2003). More flexible regression settings are discussed in Assunção (2003).

With regard to sampling for point-referenced data, we first note that optimal ex-

perimental design has a long statistical history. See the book of Pukelsheim (1993) for a review. The dominant path has focused on design for independent data collection. There is history for the case of correlated data dating to 1966 (Sacks and Ylvisaker, 1966). More recently, there has been attention directed at accommodating data with structured dependence. For spatial data this has been expressed through random fields. See the review paper of Fedorov (1996) and the book of Müller (2001). Generally, designs are classified as either probability or model-based. The former includes widely-used simple random sampling without replacement. They tend to be robust in that they make no population assumptions regarding, for example, mean structure or dependence structure.

Model-based design has followed a regression model path or a random field model path. Under regression modeling with independent data, optimality is defined with regard to efficiency of the estimates of the regression coefficients. An optimization criterion that is a function of the design matrix is specified and then the "best" design optimizes this criterion over all design matrices. Again, see Pukelsheim (1993) or Müller (2001) for details. This theory is not directly extensible to spatial design but approximately optimal solutions based upon information-theoretic measures have emerged, most notably the recursion in Brimkulov, Krug and Savanov (1986). (See Fedorov, 1996, in this regard.) This recursion is elaborated in Section 5.4 below. Its focus is exclusively on gaining information regarding the regression structure or model mean. A different type of approximation in the context of anisotropic dependence is proposed in recent work of Arbia and Lafratta (2002).

Model-based design, motivated by a random field specification, has been strongly advocated in Le and Zidek (1992) and Zidek, Sun and Le (2000) as well as references therein. The proposal is an entropy-based design where the selection of the next site to be added will be the one with the largest entropy where entropy can be viewed

as uncertainty. Under a Gaussian field assumption, the criterion that emerges is the conditional variance of an observation at a new location given the locations already selected. (This conditional variance depends only upon the previously selected locations but not on the data already collected at those locations.) The site with the most uncertainty is the one with the largest conditional variance given the selected sites. Extension to multivariate data at a location converts the criterion to a conditional covariance matrix. This approach has no interest in mean structure. In fact, quoting Zidek, Sun and Le (2000, p.66), "[I]t avoids the need to specify objectives like parameter estimation." Implementations have been in the area of network monitoring design and thus, initial preparation and operating costs are built into the adopted optimization criterion.

For a stationary Gaussian process with regression structure, two types of design questions can be asked: What is the optimal sampling design for prediction at an unobserved set of locations? What is the optimal sampling design for estimation of the parameters in the covariance function? Because prediction is often the primary use for the model, the first question has received much attention. See, for example, McBratney et al. (1981), Su and Cambanis (1993), Ritter (1996), and Zhu (2002). For the latter question, with a constant mean, the classical procedure for estimating the covariance structure is based upon the variogram. See, for example, Warrick and Myers (1987), Bogaert and Russo (1999), and Müller and Zimmerman (1999). Very recent work by Zhu and Stein (2006) focuses on designs based upon optimization using the likelihood. They suggest working with the Fisher information as a measure in the form of a ratio of determinants and implement the optimization using a simulated annealing algorithm. Finally, there is some literature on Bayesian experimental design (see e.g. Clyde, 1993; Clyde, Müller and Parmigiani, 1995, 1996; Chaloner and Verdinelli, 1995). This methodology is based on decision theory and useful for spatial

design. We do not pursue this direction further because Bayesian design usually is very computational intensive.

Our approach is to also be model-based, working with the likelihood and focusing on the information matrix as well. As noted above, our perspective is primarily pragmatic. We conceive a sampling setting in which we envision hundreds to thousand of sites being sampled and seek to make the required sampling design easily understood and computationally manageable for the practitioners who wish to implement it. We take as our design objectives learning about the mean structure as well as the covariance function, noting that these objectives are usually in conflict. We also introduce a further utility notion, providing an additional objective of sampling for say, large values (as with contaminant surfaces). We consider the situation where we already have a partial sample and we wish to augment the available data. After clarifying that obtaining the optimal solution is a combinatorially complex computation problem, we consider three approaches toward achieving approximate optimization - sequential selection, block selection, and tuned stochastic search.

5.3 An overview of the issues

Our objective is, for a given study region, to develop an approximately optimal sampling strategy to learn about the spatial distribution of a contaminant across the region. Optimal design is intractable working with the continuum of locations so, as is customary, we presume that the region has been gridded (not necessarily a regular grid) to high resolution. For instance, in the context of sampling childhood blood lead levels, the tax parcel level (equivalently the residential property on the parcel) provides a natural discretization for sampling locations. In the ensuing development, we assume that the parcels can be viewed as points in the region but, ultimately, with regard to design, we have only a finite set of locations to select from.

We will work within the model-based framework for developing designs. The two types of criteria we might consider are:

(1) An information criterion that arises from the regression perspective (Section 5.4) but incorporates learning about strength of spatial dependence as well as the regression component.

(2) An entropy criterion that focuses on uncertainty, yielding a conditional covariance. We emphasize the information criterion in this chapter for reasons we elaborate in Section 5.6.2. However, we reserve Section 5.6.2 for some comparison.

We also note that in the multi-parameter case (almost certainly the case of interest in applications), both criteria emerge as matrices. So, to achieve a single number summary for a design, we will have to summarize the resulting matrix either through a determinant or a (possibly weighted) trace.

We further assume that sampling is not *ab initio* or "preposterior". Rather, we assume that a collection of n locations have already been sampled and that we have this data available to us. Such collection may have been implemented by simple random sampling or perhaps, through *ad hoc* methods. If not, how should the initial set of n points be selected? A convenient approach referred to as space-filling designs, has been discussed in Nychka and Saltzman (1998). Such designs are based upon geometric measures of how well a given set of points covers the study region, independent of the assumed covariance function. Such designs are not optimal but for an initial selection will work nearly as well as optimal ones. In any event, as the number of sampling sites grows, effects of the initial selection dissipate.

Based upon the data from these n sites, we can implement a preliminary fit of the model to obtain preliminary parameter estimates. This is crucial since our design criteria emerge as parametric functions. To evaluate a criterion for a given set of locations, we insert the parameter estimates into the function as well as the locations.

We recognize that this fails to account for the uncertainty in the parameter estimation and that averaging over a suitable distribution for these parameters would enable us to attach uncertainty to the criterion value. However, with interest in design rather than inference (which would come later, after all of the data collection) we adopt the pragmatic "plug in" approach. It is also computationally much more convenient and avoids the need for prior specification at the time of sampling.

Thus the formal goal is, given the n locations already chosen and sampled, and given that we want to choose m additional sites to sample, how shall we choose these m locations? Even given an explicit, evaluable criterion and a finite collection of N sites to choose from, obtaining the optimal choice is not a tractable problem. In our setting it would be referred to as an "N-n choose m" combinatorially hard problem. So, we will have to consider approximate solutions to this problem. We examine three strategies: (i) sequential selection, (ii) block selection; and (iii) stochastic search (including a modified procedure).

5.4 The information criterion

In presenting the information criterion, rather than elaborating the formal optimal design machinery (as described, for instance, in Pukelsheim 1993 or in Müller 2001), we offer an intuitive development built from the well-established Fisher information measure (see, e.g., Rao, 1973; Cox and Hinkley, 1974). The Fisher information arises from expectation of second derivatives of the log likelihood. In the multiparameter case, it becomes the expectation of a matrix of mixed partial derivatives (the Hessian) associated with the log likelihood. Under normality and a linear mean form (in the coefficients) it emerges as a parametric function of the dependence structure. The matrix is reduced to a scalar criterion either through the trace or determinant.

More precisely, suppose we consider the widely used spatial model:

$$Y(s_i) = \mu(s_i) + W(s_i) + \epsilon(s_i), \tag{5.1}$$

where $Y(s_i)$ $(i = 1, ..., n)$ are observations from a spatial process over a region D in $\mathbb{R}^2$ and $\mu(s_i)$ is the linear mean form, $X^T(s_i)\boldsymbol{\beta}$. $W(s_i)$ is a mean 0 spatial process (typically a stationary Gaussian process) and $\epsilon(s_i)$ is a pure error process with mean 0 and variance τ^2. W and ϵ are independent. Written in vector form $\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{W} + \boldsymbol{\epsilon}$, where $\mathbf{Y} = (Y(s_1), ..., Y(s_n))^T$, $\mathbf{W} = (W(s_1), ..., W(s_n))^T$, $\boldsymbol{\epsilon} = (\epsilon(s_1), ..., \epsilon(s_n))^T$, and

$$\begin{pmatrix} \mathbf{W} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim \mathrm{N}_{2n}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 R(\boldsymbol{\phi}) & 0 \\ 0 & \tau^2 \mathrm{I}_n \end{pmatrix} \right). \tag{5.2}$$

Here, $R(\cdot)$ is the correlation matrix associated with the n locations and $\boldsymbol{\phi}$ indexes the parameters of the correlation function, for example, in the Matérn case (see Section 2.1.3), a smoothness parameter and a decay parameter.

Let $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi}, \tau^2)^T$ with $\Sigma_{\boldsymbol{\theta}} = \sigma^2 R(\boldsymbol{\phi}) + \tau^2 \mathrm{I}_n$. The log likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\boldsymbol{\theta}}| - \frac{1}{2}(Y - X\boldsymbol{\beta})^T \Sigma_{\boldsymbol{\theta}}^{-1}(Y - X\boldsymbol{\beta}). \tag{5.3}$$

The expected information matrix for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ has the block diagonal form (see Section 3.2)

$$I(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{pmatrix} X^T \Sigma_{\boldsymbol{\theta}}^{-1} X & 0 \\ 0 & I(\boldsymbol{\theta}) \end{pmatrix}, \tag{5.4}$$

where the $(i, j)^{th}$ element of $I(\boldsymbol{\theta})$ is $\frac{1}{2}\mathrm{Tr}[\Sigma_{\boldsymbol{\theta}}^{-1}\frac{\partial \Sigma_{\boldsymbol{\theta}}}{\partial \theta_i}\Sigma_{\boldsymbol{\theta}}^{-1}\frac{\partial \Sigma_{\boldsymbol{\theta}}}{\partial \theta_j}]$. The block diagonal form in (5.4) shows that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are *orthogonal* parameters (Cox and Reid, 1987). Informally, this means that an information criterion for design will "separate" information regarding $\boldsymbol{\beta}$ from information regarding $\boldsymbol{\theta}$.

As it stands, (5.4) is not a *criterion*. We need to reduce it to a univariate summary which we will then seek to maximize. Such optimization will correspond to maximizing information gain, as we detail in the next section. To achieve such reduction, we

introduce a mapping from information matrices to scalars. Customary approaches work with either $\mathrm{Tr}(I(\boldsymbol{\beta},\boldsymbol{\theta}))$ or $|I(\boldsymbol{\beta},\boldsymbol{\theta})|$. The former emerges as $\mathrm{Tr}(X^T\Sigma_{\boldsymbol{\theta}}^{-1}X)$ + $\sum_i \frac{1}{2}\mathrm{Tr}[\Sigma_{\boldsymbol{\theta}}^{-1}\frac{\partial\Sigma}{\partial\theta_i}\boldsymbol{\theta}\Sigma_{\boldsymbol{\theta}}^{-1}\frac{\partial\Sigma}{\partial\theta_i}\boldsymbol{\theta}]$, the latter as $|X^T\Sigma_{\boldsymbol{\theta}}^{-1}X| \times |I(\boldsymbol{\theta})|$. In either case, we see the separation mentioned above. The forms also reveal that there can be tension between the component terms. That is, for a fixed m, the set of points which maximizes our information gain about $\boldsymbol{\beta}$ will be different from that for $\boldsymbol{\theta}$. Also, the trace criterion suggests the possibility of weighted components (see Section 5.9).

Here, and in the sequel, we work with the trace of the information matrix to provide our criterion rather than the determinant. The former is more intuitive in appreciating the components in the information gain but requires standardization of the covariates since it is not independent of the scale of the covariates (as the form $X^T\Sigma_{\boldsymbol{\theta}}^{-1}X$ reveals). The latter avoids that problem since any scaling emerges as a constant multiple of the determinant but at the expense of ease of interpretation.

Lastly, the response can be modeled on a suitably transformed scale in order to make the Gaussian assumption more comfortable. Moreover, in what follows we work with the foregoing modeling assumptions because they yield convenient computational expressions. We are not restricted to this setting; with additional computational effort, we can accommodate non-Gaussian data, e.g., categorical outcomes or counts and/or nonlinear means.

5.5 Approaches for approximately optimal design

To address the "N-n choose m" combinatorially hard design problem noted at the end of Section 5.3, we consider three approximate solutions, sequential selection, block selection and (modified) stochastic search. We consider them individually here though one could readily envision hybrid versions.

The sequential approach would require us to (i) identify, as the $(n+1)^{st}$ parcel,

the one which provides the maximum increase in information; (ii) sample it and add its data to the data already collected; (iii) revise our current information, now based upon $n+1$ parcels; (iv) reorder the remaining parcels; and (v) select the $(n+2)^{nd}$. In fact, we can make a modest compromise (which is appropriate for the way that the data collection would likely proceed) by sequentially ordering the parcels but only assuming we have data about the underlying process model from the first n locations. In this fashion we can order the next m parcels to be selected. Then, if additional sampling were sought after these new m locations are sampled, we would refit the model and revise our knowledge about the model parameters in order to further sample.

The block selection approach would order all of the remaining parcels, given the n already selected and then choose the m parcels with the m largest values of the criterion. Evidently, it offers computational savings. Again, after these new m parcels were sampled, we would update our parameter estimates before further sampling. Hence, either scheme provides an ordering to all of the unsampled parcels. However, as we clarify below, these two approaches provide dramatically different sampling designs and, though the sequential scheme emerges as generally preferable, we can not assert that for any N, n and m, it will always ensure greater information gain.

The third approach introduces stochasticity into the selection process. The most naive stochastic selection algorithm would choose m points at random and would be simple random sampling. Stochastic search is introduced if we make, say, b random selections, calculate the information gain for each, and adopt the one yielding the largest gain. Of course, the choice of b is unclear. The larger b is the closer we must get to the optimal design; however, computation cost increases linearly in b. Refinement of the stochastic search is possible. For instance, consider any location s which was selected in at least one of the b searches. We can compute the average information

gain for this location over all of the searches in which it was included. We could then propose to sample the m locations providing the largest average information gains.

We do note that, though it is not feasible to obtain the optimal design, the fact that we are dealing with a finite set of locations N does enable us to compute an upper bound on the information gain, i.e., the information gain associated with sampling all remaining N-n locations. Evidently, the information gain for choosing m points will tend to this bound as m increases. In fact, this raises an important theoretical point that we discuss in detail in Chapter 3. What can we say about $I(\boldsymbol{\beta}, \boldsymbol{\theta})$ as N grows large? What can we say about $I(\boldsymbol{\beta}, \boldsymbol{\theta})$ for fixed N as m increases? For the former, the key point is whether information tends to ∞ as $N \to \infty$ or remains bounded. For the latter, typically the information gain increases rapidly over smaller m with diminishing returns from there on. Hence, the upper bound not only provides a measure of what proportion of potential information we will gain from our sample of size m but also, if we see an "elbow" in the information gain as a function of m, we might conclude that there will be little value in spending for additional sampling.

5.6 Information gain and comparison to the entropy criterion

5.6.1 Calculation of the information gain

Returning to the model in (5.1), recall that we seek to learn about the importance of the covariates (the X's) in explaining the responses (the Y's) as well as the nature and strength of spatial dependence. Assume the Gaussian process has stationary covariance function $\sigma^2 \varrho(s_i - s_j; \phi)$. (This is not required but does simplify the ensuing presentation.) Here, we assume $\varrho \geq 0$ and that ϱ strictly decreases from 1 to 0 as ϕ goes from 0 to ∞. A typical example is the so-called exponential covariance function with $\varrho(s_i, s_j) = \exp(-\phi \|s_i - s_j\|)$. Also included are the powered exponential and

Matérn families of covariance functions of which the exponential is a special case. We also assume, for the moment, that $\tau^2 = 0$, i.e., that there is no nugget.

From information formula (5.4), if the Y's all have a common mean say μ, up to a constant, the information in the sample about μ given spatial dependence measured by ϕ is the scalar $I_n(\mu) = \mathbf{1}^T R_n^{-1}(\phi)\mathbf{1} \equiv A_n$ where $R_n(\phi)$ is the $n \times n$ matrix with (i, j) entry $\varrho(s_i - s_j; \phi)$. (So, we can ignore the unknown σ^2 in comparing designs.)

Despite its innocuous form, general behavior for A_n is not easy to prove except in very special cases (see Section 3.4). However, we can explicitly compute the information gain in sampling location s_0. We have

$$A_{n+1} - A_n = \frac{(1 - \mathbf{1}^T R_n^{-1}(\phi)\mathbf{r}_{n0}(\phi))^2}{1 - \mathbf{r}_{n0}^T(\phi)R_n^{-1}(\phi)\mathbf{r}_{n0}(\phi)}. \tag{5.5}$$

In this expression, $\mathbf{r}_{n0}(\phi)$ is an $n \times 1$ vector with i^{th} entry $\varrho(s_i - s_0; \phi)$. So, the s_0 that maximizes this difference is the location that maximizes information gain. The maximization is easy to carry out since we only have a finite number of sites and since $\mathbf{r}_{n0}(\phi)$ changes with s_0 but $R_n^{-1}(\phi)$ does not. In fact, we suggest the creation of a GIS display in the form of a choropleth map or a contour plot to reveal where in the region information gain is high and where it is low. (See the illustrative example in Section 5.9.) Evaluation of the criterion requires knowing the covariance function, i.e., requires estimating ϕ. As discussed above, this will be done using the n data points already collected. That is, the initial data provides our starting knowledge regarding spatial structure. As we collect additional data, we use it to revise our learning about this structure.

As noted above, A_n calculates the information in the sample about the mean μ. There is also information in the sample about σ^2 and ϕ. In particular, $I(\mu, \sigma^2, \phi)$, as

a special case of (5.4), takes the form, for sample size n,

$$I_n(\mu, \sigma^2, \phi) = \begin{pmatrix} A_n/\sigma^2 & 0 & 0 \\ 0 & n/\sigma^2 & B_n/\sigma \\ 0 & B_n/\sigma & E_n \end{pmatrix}, \tag{5.6}$$

where A_n is as above, $B_n = \mathrm{Tr}(R_n^{-1}\frac{\partial R_n}{\partial \phi})$ and $E_n = \frac{1}{2}\mathrm{Tr}(R_n^{-1}\frac{\partial R_n}{\partial \phi}R_n^{-1}\frac{\partial R_n}{\partial \phi})$. Hence $|I_n(\beta, \sigma^2, \phi)|$ has the simple form $\sigma^{-4}A_n(nE_n - B_n^2)$ explicitly revealing the separation in information contributions. Suppose that our interest focuses on the information gain for both μ and ϕ (i.e., we pretend that σ^2 is known). We can simplify (5.6) to

$$I_n(\mu, \phi) = \frac{1}{\sigma^2}\begin{pmatrix} A_n & 0 \\ 0 & \sigma^2 E_n \end{pmatrix} \tag{5.7}$$

Taking the trace of this matrix revises the criterion to $A_n/\sigma^2 + E_n =$

$$\mathbf{1}^T R_n^{-1}(\phi)\mathbf{1}/\sigma^2 + \frac{1}{2}\mathrm{Tr}\Big(R_n^{-1}(\phi)\frac{\partial R_n(\phi)}{\partial \phi}R_n^{-1}(\phi)\frac{\partial R_n(\phi)}{\partial \phi}\Big). \tag{5.8}$$

Evaluation of (5.8) requires estimating σ^2 and ϕ. Again, this will be done using the n data points already collected. Using convenient matrix identities (e.g., Harville, 1997), computational methods for the rapid calculation of the analogue of (5.5) are available; we omit details. In fact, it may be of interest to compare the approximately optimal designs for just the first term in (5.8) or just the second term in (5.8) (see Section 5.9). However, for the remainder of this section we omit the contribution of ϕ to the information gain.

For each site we typically have available covariate information, e.g., for tax parcels, the age of the house on the parcel might be assumed to provide explanation regarding the presence of biologically available lead at the location. Suppose in subsequent analysis, once the data is collected, we anticipate using such information in the mean specification, say in the form of a linear regression, $\beta_0 + \beta_1 X(s)$ where $X(s)$ is the

age associated with parcel s. Then, we might seek to choose the parcels to maximize the information about the linear function in age, i.e., in β_0 and β_1. (Note that this is not the same objective as choosing sites to encourage $Y(s)$ to be large. See Section 5.8 below.)

More generally, with a $p \times 1$ vector of covariates $X(s)$, including the intercept, we obtain a $p \times p$ information matrix (the upper left matrix in (5.4)). Using the trace, we now obtain $\sum_{l=1}^{p} X_l^T R_n^{-1}(\phi) X_l$ where X_l is the l^{th} column vector of the design matrix $X = (X_1, ..., X_p)$ (the i^{th} entry of X_l is $X_l(s_i)$). Again we can compute the information gain explicitly in selecting parcel s_0. In fact, we obtain

$$\frac{\sum_{l=1}^{p}(X_l(s_0) - X_l^T R_n^{-1}(\phi)\mathbf{r}_{n0}(\phi))^2}{1 - \mathbf{r}_{n0}^T(\phi) R_n^{-1}(\phi)\mathbf{r}_{n0}(\phi)}. \tag{5.9}$$

This is the recursion of Brimkulov, Krug and Savanov (1986).

5.6.2 Comparison of the information criterion and the entropy criterion

From Section 5.1, the entropy criterion is phrased in terms of extent of uncertainty and is motivated by work in pollution-monitoring network design as summarized in Zidek, Sun and Le (2000). A scalar arises in the univariate case, the determinant of a matrix in the multivariate case. Again, with normally distributed, dependent data both the scalar and the determinant will be parametric functions of the dependence structure. In the design setting it is intuitively easiest to interpret entropy as uncertainty. Sites with high entropy, given those that we have already sampled, would be desirable choices to select. That is, from the remaining sites, we would seek to learn about those for which we are most uncertain. Hence, the criterion computes the entropy given the current set of sites and adds next the site with the largest conditional entropy. With the assumptions and notation above, it is straightforward to show that

the conditional entropy associated with site s_0 is $1 - \mathbf{r}_{n0}^T(\phi)R_n^{-1}(\phi)\mathbf{r}_{n0}(\phi)$ (Zidek, Sun and Le, 2000). As a conditional variance, this quantity is obviously nonnegative and so, we choose s_0 to maximize this. Computation is straightforward. A choropleth map or contour plot of the values of this criterion over the collection of parcels would provide a useful display.

It is interesting to note that the entropy criterion is the denominator of the information criterion. This appears paradoxical since we are proposing to maximize both criteria. In fact, the square in the numerator of the information criterion offsets the denominator to remove the paradox. We can clarify by looking at the $n = 1$ case. The information criterion becomes $(1-\rho)/(1+\rho)$ while the entropy criterion becomes $1 - \rho^2$. Both decrease from 1 to 0 as ρ increases from 0 to 1. However, the functions are quite different; for instance, the former is convex while the latter is concave.

The entropy criterion can be extended to accommodate pure error as well, replacing $R_n(\phi)$ with $\sigma^2 R_n(\phi) + \tau^2 \mathrm{I}_n$ as in Section 5.4. The resulting form for the criterion is $\sigma^2 + \tau^2 - \sigma^4 \mathbf{r}_{n0}(\phi)^T(\sigma^2 R_n(\phi) + \tau^2 \mathrm{I}_n)^{-1}\mathbf{r}_{n0}(\phi)$. The criterion can also be extended to multivariate measurements in the form of the determinant of the conditional covariance matrix associated with $\mathbf{Y}(s_0)$. With a separable specification for the error structure, an argument similar to that for the information criterion enables us to use the same entropy criterion as above.

Finally, the criterion would not be affected by the introduction of covariate information for each site. The entropy measure focuses only on uncertainty arising from spatial structure. The conditional variance is not affected by the mean specification. As Zidek, Sun and Le (2000) note, the criterion avoids issues like parameter estimation and hypothesis testing. In our context, we would not view this as advantageous since we want to learn about the nature of the regression relationship between the level of the response and the proposed explanatory variables. So, for our purposes,

the information criterion emerges as preferred. In particular, it will use the available $X(s)$ vectors in the determination of the selection order.

5.7 Some remarks

Remark 19. Consider the model $Y(s) = \mu + W(s)$ and suppose that n locations $s_1, ..., s_n$ have already been sampled. Expression 5.5 shows that the information gain at s_0, $\Delta I_{s_0}(\mu) > 0$ for a new s_0. In fact, $\Delta I_{s_0}(\mu) \to 0$ when $\|s_0 - s_i\| \to 0$, where $s_i (i = 1, ...n)$ is any of the n samples and $\|s_0 - s_i\|$ is the Euclidean distance between s_0 and s_i. So, since $A_1 = 1$, we have $A_n > 1$ and A_n increases in n. If $R_n \to \mathrm{I}_{n\times n}$, the identity matrix, then $A_n \to n$. Is $A_n \leq n$? Since we showed in the previous section that $A_2 - A_1 = (1 - \rho)/(1 + \rho)$, if $\rho < 0, A_2 > 2$ and, in fact, $A_2 \to \infty$ as $\rho \to -1$.

Remark 20. Assuming that all the covariance parameters are given, the general behavior of $\mathbf{1}^T R^{-1}(s_1, ...s_n)1$ is surprisingly difficult to investigate (see Section 3.4). Results depend upon the form of ϱ and the nature of the asymptotics. For example, under infill asymptotics and a separable covariance function that is a product of one-dimensional exponential covariance functions, we can compute A_n explicitly as well as its limit which is finite. If we allow the size of the region to grow as n grows, then the relative rates of growth determine the behavior of A_n. Detailed discussion is presented in Chapter 3.

Remark 21. If one were to think in terms of choosing a distribution to randomly sample the locations from, intuition might suggest that the uniform distribution produces the maximum expected information for μ. In fact, if sampling is for $s \in D$, $\mathrm{E}[\mathbf{1}^T R^{-1}(s_1, ...s_n)\mathbf{1}|s_i \sim \mathrm{unif}(D)]$ will not maximize $\mathrm{E}[\mathbf{1}^T R^{-1}(s_1, ...s_n)\mathbf{1}|s_i \sim f(D)]$ for all distributions f over D. Intuitively, appropriate systematic selection of points will provide greater information than the average under random selection. Consider the following simple example. Suppose we sample 4 points (s_1, s_2, s_3, s_4) uniformly on

[0,1]. With the exponential correlation function $R(s_i, s_j) = e^{-7|s_i - s_j|}$, we can obtain, by Monte Carlo integration, $\mathrm{E}[\mathbf{1}^T R^{-1}(s_1, ...s_4)\mathbf{1}|s_i \sim \mathrm{unif}(0,1)] = 2.52$ (Figure 5.1 shows the density function for the information in this case). However, if we choose $(s_1, s_2, s_3, s_4) = (0, 0.33, 0.67, 1)$, $\mathbf{1}^T R^{-1}(s_1, ...s_4)\mathbf{1} = 3.47$. Hence, a nondegenerate distribution that is not far from this *degenerate* choice will achieve a larger expectation than under uniform selection.

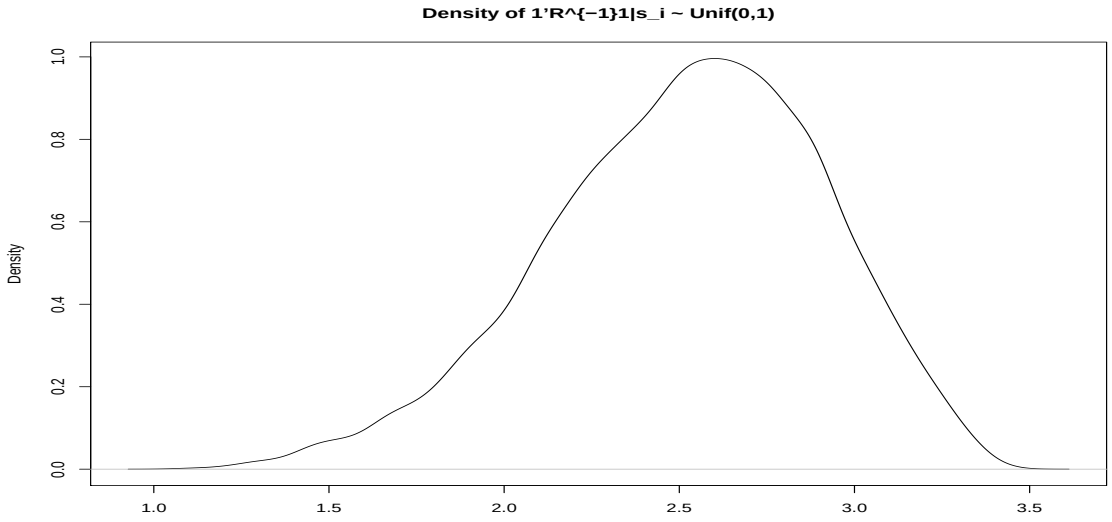

Figure 5.1: Density of $\mathbf{1}^T R^{-1}\mathbf{1}$ given $s_1, ..., s_4 \sim \mathrm{unif}(0,1)$

Remark 22. In considering say, $I_n(\mu, \sigma^2, \phi)$, the trace cumulates what we would define as the conditional information, e.g., the information in the sample about ϕ given μ and σ^2 are known. We could also calculate *unconditional* information. We can show that the sum of the reciprocals of the diagonal elements of $I_n^{-1}(\mu, \sigma^2, \phi)$ cumulates this unconditional information. Furthermore, the asymptotic behavior of unconditional information need not agree with that of conditional information.

Remark 23. We hope that it is clear that the sequential approach need not produce the optimal choice of m points. The useful analogy here is to variable selection in

multivariate linear regression. A forward stepwise procedure is not guaranteed to produce the subset of variables of a fixed size which maximizes R^2.

Remark 24. As an example to illustrate a case where sequential design will be worse than block design, suppose we have s_1 at the origin. We want to select three additional points to learn about the mean from $s_2, ..., s_5$ as shown in Figure 5.2. The block design will select (s_1, s_4, s_2, s_5) while the sequential design will select (s_1, s_4, s_3, s_2). The corresponding sequence of information values is (1, 1.915, 2.728, 3.539) and (1, 1.915, 2.736, 3.521), respectively; the four points selected by the block design produce greater information than those selected by the sequential design.

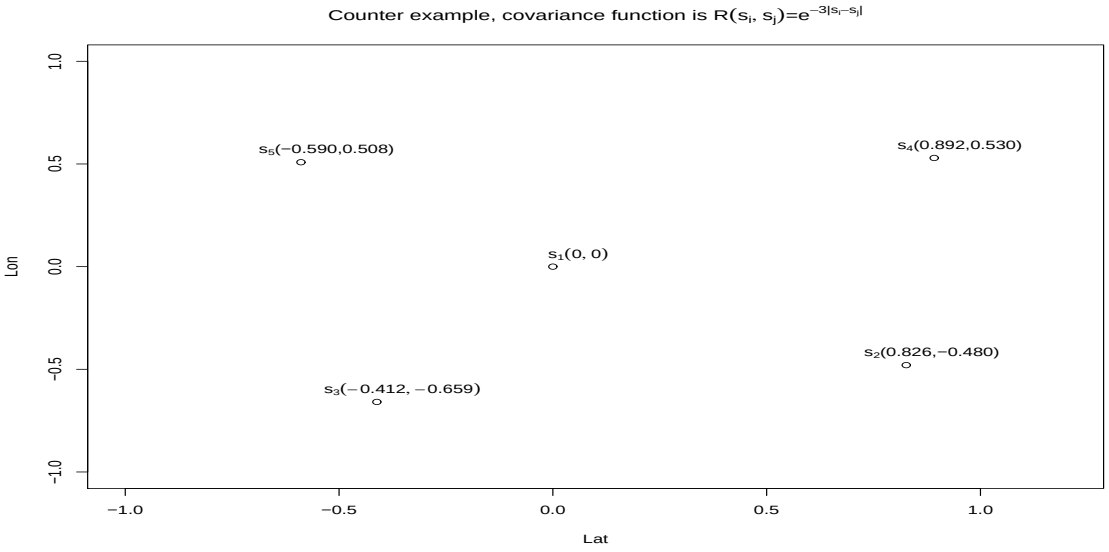

Figure 5.2: Sequential selection vs. block selection

Remark 25. We conclude with an illustrative comparison among the sequential scheme, the block scheme, the stochastic search scheme ($b = 500$) and the refined stochastic search scheme (again, $b = 500$). In Figure 5.3 (which arises from the simulation illustration in Section 5.9), we plot the information growth for the intercept.

The information for the intercept is bounded as noted in Remark 19 and the upper bound is given. The sequential design scheme is clearly the best, as it will be generally except for pathological examples such as in Remark 24. With 40 sites already sampled and 960 that could still potentially be sampled, more than 95% of the upper bound is achieved with only 20 additional observations. For the block scheme and the modified stochastic search schemes roughly 70 additional observations are needed to do as well. The inferior performance of the simple stochastic search scheme is evident. We also plotted (Figure 5.4) the information growth for $I(\phi)$, calculated through the lower right entry in (5.4). Note the striking difference in the information scales between Figures 5.3 and 5.4. Also, we see that, with σ^2 fixed (known), information growth for ϕ is not bounded. See, e.g., Xia, Hjort and Gelfand (2006) in this regard.

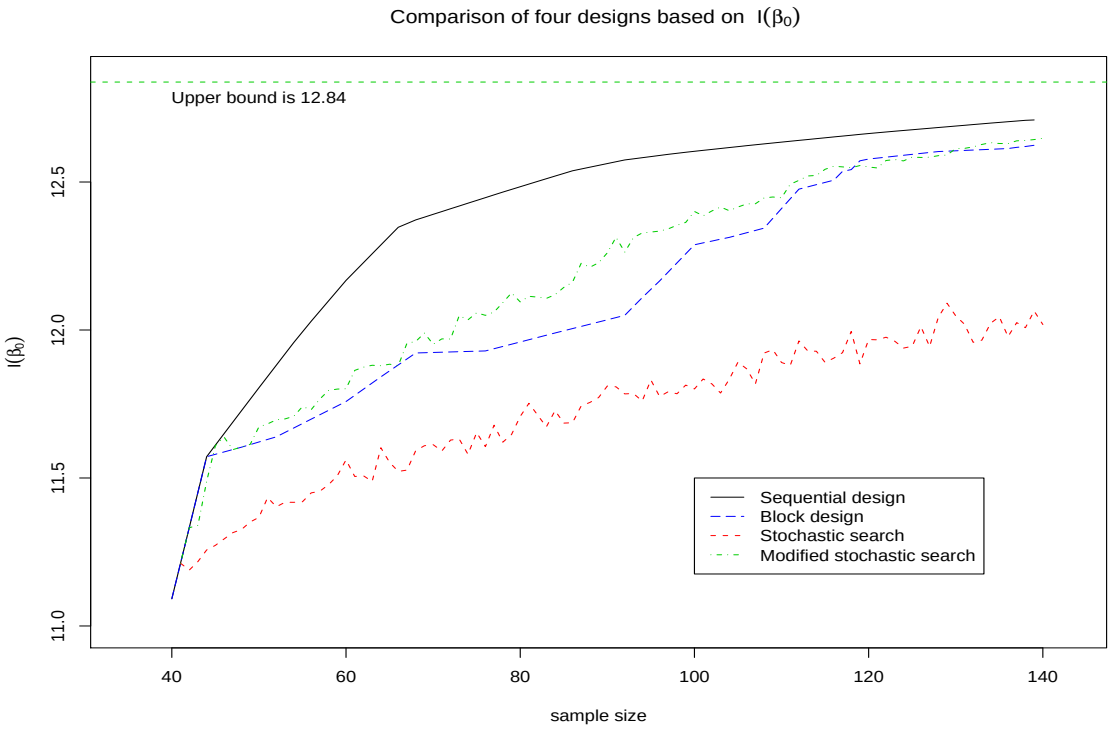

Figure 5.3: Information ($I(\beta_0)$) growth in sample size for the four sampling schemes

5.8 Modified utility spatial designs

The goal here is to propose the use of overlays of either (estimated) mean response or covariate data layers to achieve specific objectives, e.g., to separate essentially equivalent locations under the foregoing criteria or to modify utility for point selection resulting in revised approximate optimization for the spatial sampling design.

For instance, the goal may be to learn about the regression relationship but this does not imply selection of sites where the response is expected to be high, e.g., high levels of biologically available lead or of arsenic contamination. One way to achieve this is a model-based strategy, obtaining the estimated spatial surface based upon the data collected thus far, i.e., based upon $Y(s_1), Y(s_2), ..., Y(s_n)$. Overlay of this surface on the selection surface will reveal parcels where both layers achieve high values in order to determine selection. Alternatively, we could multiply the surfaces to upweight/downweight the selection surface. One might also work not with the fitted model layer but, instead, a different data layer, perhaps external to those used in the model fitting. Such layers might reflect established geographic gradients with regard to say, the contaminant or distance from a site that is a known source for high contaminant levels.

This strategy would also address the matter of locations having essentially equivalent values under the criterion. They can be distinguished by using the second weighting layer, yielding a weighted criterion. For instance, one could upweight parcels that are expected to exhibit high levels of the contaminant being sampled.

A somewhat different objective that could be used to distinguish parcels which are essentially equivalent under the information criterion would be to work with demographic data layers. In this case, the second objective would be to oversample

parcels with certain demographic features, e.g., in low socio-economic status areas or high crime rate areas. A spatial surface reflecting such a layer would be created. Again, overlaying or multiplication provides upweighting or downweighting of the selection surface. Ultimately, the issue is one of utility for the data collection. If we seek to learn not only about the exposure surface but also to achieve certain expected features in our samples, then we need to specify a utility function that reflects this objective.

5.9 Computational issues and a simulation illustration

In providing a simulation illustration, we focus on sequential design and block design to select an additional collection of m parcels from N-n parcels given n have already been selected. We adopt the model $Y(s) = X(s)^T\boldsymbol{\beta} + W(s) + \epsilon(s)$, and work with $I(\boldsymbol{\beta}, \boldsymbol{\theta})$ as in (5.4). For the sequential design we do not update the parameter estimates after each new location is selected. We only use the parameter estimation based upon the original n samples.

We generalize the trace of $I(\boldsymbol{\beta}, \boldsymbol{\theta})$ to define $I(\boldsymbol{\beta}) = \sum_{i=1}^{p} w_i I(\beta_i)$ for vector $\boldsymbol{\beta}$ of length p , $I(\boldsymbol{\theta}) = \sum_{j=1}^{q} v_j I(\theta_j)$ for vector $\boldsymbol{\theta}$ of length q and finally, the combined information as $I(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=0}^{p} w_i I(\beta_i) + \sum_{j=1}^{q} v_j I(\theta_j)$. The weights allow us to rescale the components of the trace to reflect the fact that the information is affected by scale. For example, the information value would change if we convert distance from say kilometers to miles. Also the scale of information values for mean structure parameters $\boldsymbol{\beta}$ and for dependence structure parameters $\boldsymbol{\theta}$ can be very different (recall Figures 5.3 and 5.4). However, here we do not pursue standardizations (choices of v's and w's) since, as in the previous section, they too reflect utility for the design. Instead, in the illustration below, we show approximately optimal designs for $\boldsymbol{\beta}$,

for $\boldsymbol{\theta}$ and for utility-weighted versions of these following Section 5.8. Furthermore, motivated by Figure 5.3, we work only with the sequential sampling scheme.

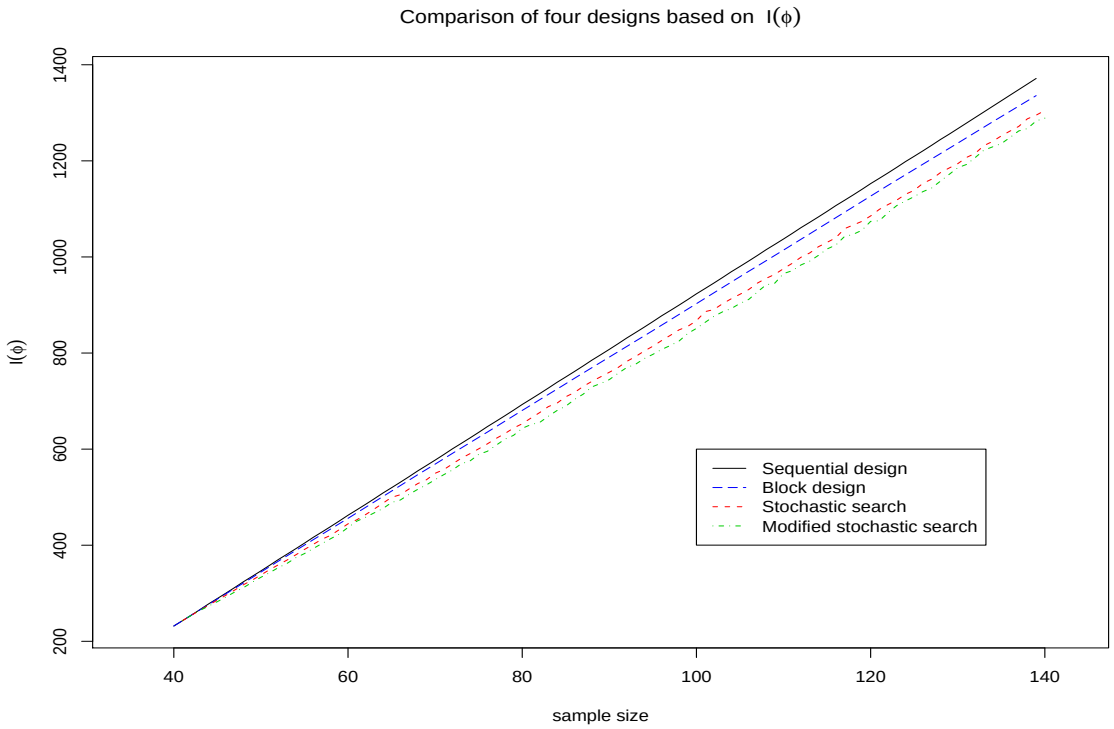

Figure 5.4: Information ($I(\phi)$) growth in sample size for the four sampling schemes

In particular, we turn to a simulation example where we conduct the spatial design based upon a grid of 40×25 parcels as Figure 5.5 shows.

We first sampled 40 of these parcels according to, for example, space-filling design at locations indicated by + on the grid. We generated a random realization of a Gaussian process of the form $Y(s) = \beta_0 + \beta_1 X_1(s) + \beta_2 X_2(s) + W(s)$, ignoring the pure error term $\epsilon(s)$, for convenience. $X_1(s)$ denotes the distance of location s from

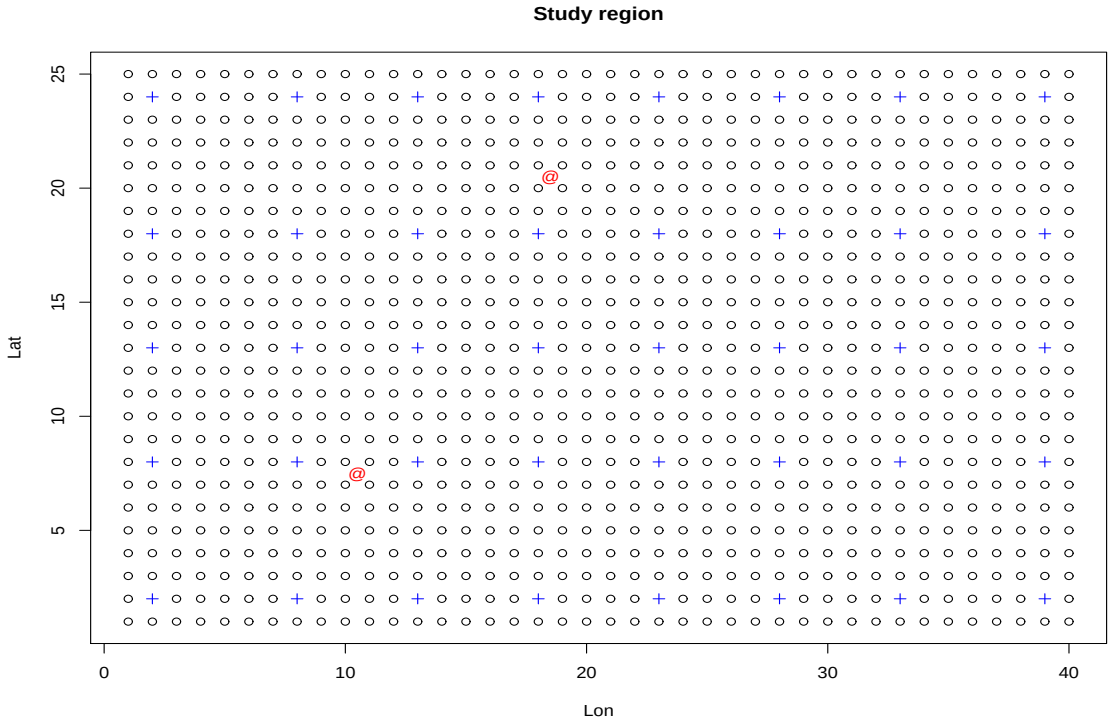

Figure 5.5: Study region

a pollution source located at $(10.5, 7.5)$ while $X_2(s)$ denotes the distance of s from a different pollution source located at $(18.5, 20.5)$ (these are indicated by @ on the grid). The true $\beta_0 = 2$, the true $\beta_1 = 0.5$ and the true $\beta_2 = 1$. The spatial variability σ^2 is set to 1. We use the exponential covariance function with decay parameter $\phi = 0.2$, resulting in a spatial range of 31.75% of the maximum distance in the region. Figure 5.6 is a three dimensional perspective plot of the true mean surface.

Figure 5.7 shows 100 selected locations (indicated by •) based on the information gain of β_0. Though the selection sequence is not numbered (the figure becomes too cluttered if we do), adding "less dependent", i.e., "most isolated" locations will most increase information. Notice also the striking edge effects which are inherent in sampling from a bounded region.

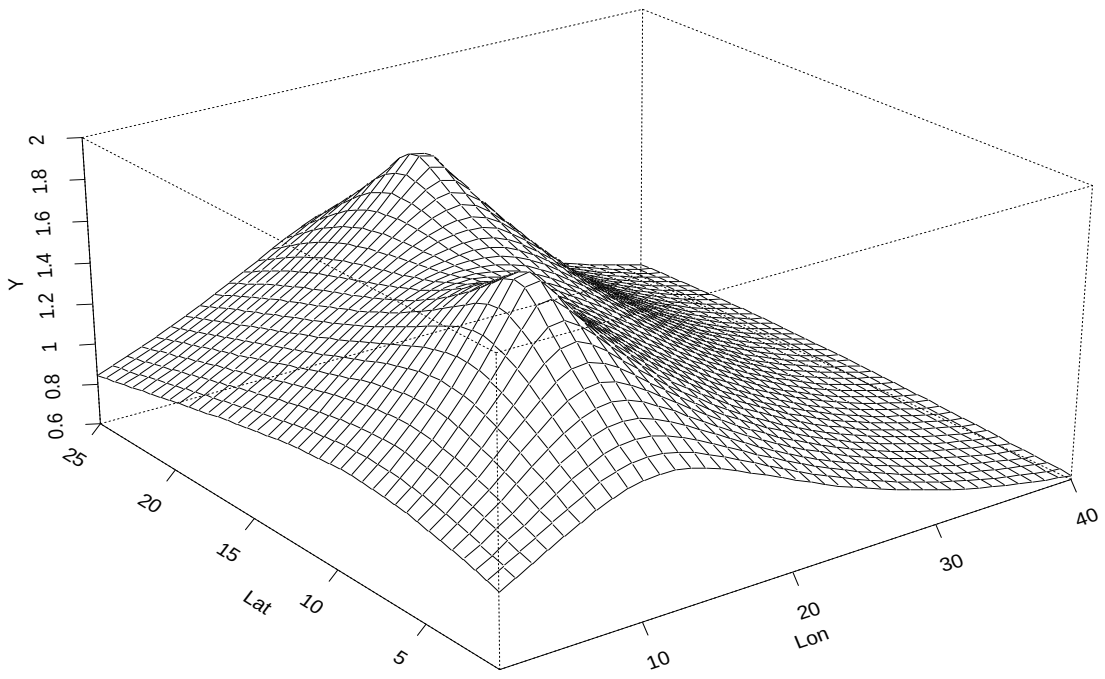

Figure 5.6: Mean surface

Figure 5.8 provides 100 selected points based on the information gain of $\beta_0, \beta_1, \beta_2$. It can be seen that, in addition to isolated locations, we also choose locations that are clustered around the two pollution sources. In this case, the information gain by adding a particular point depends on the distance from that point to the pollution sources as well as the dependence between that point and all the existing samples. Again, edge effects are strong.

Figure 5.9 shows the design that results from the criterion which attempts to maximize $I(\phi)$, the gain in information about the spatial dependence. The choice of points is dramatically different from that for $I(\beta_0)$. To learn about decay in spatial association, we need points near to each other. Edge effects are not an issue here.

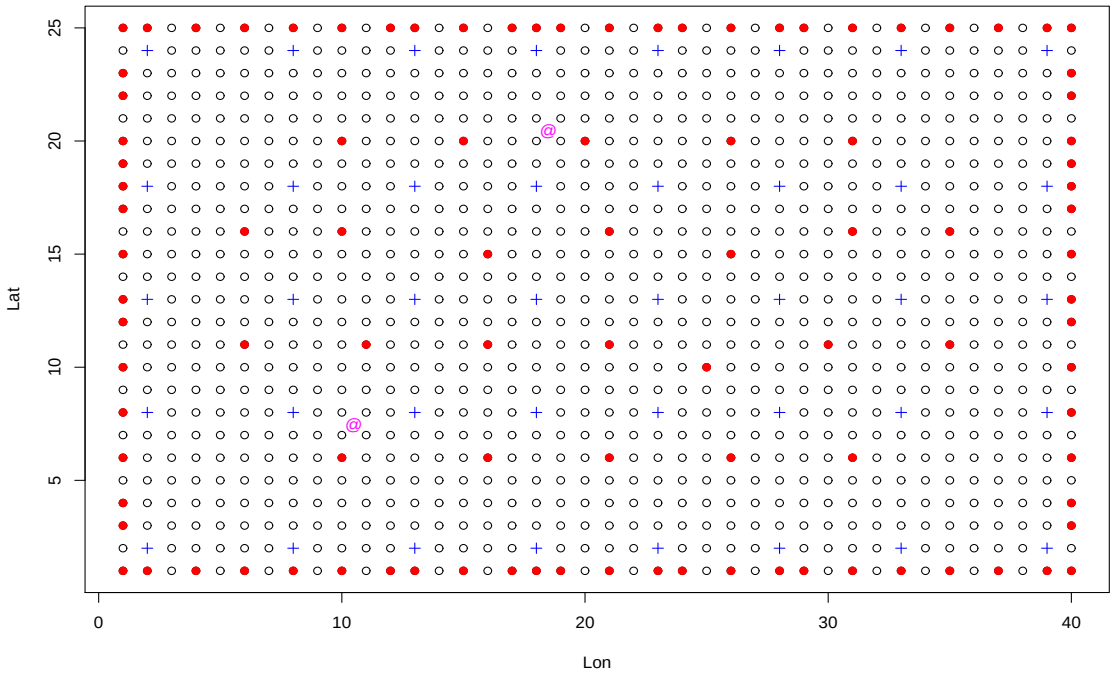

Figure 5.7: Sequential design based on $I(\beta_0)$

Figures 5.10 and 5.11 provide the analogues of Figure 5.7 and Figure 5.8 using the design points based upon a weighted criterion, in particular, weighted by the estimated mean at each location. Note that in the present case, Figure 5.10 changes dramatically from Figure 5.7 while Figure 5.11 is nearly the same as Figure 5.8.

5.10 A prospective real application

In 1984, a Union Carbide plant in Bhopal, India released methyl isocyanate into the air at levels high enough to kill several thousand people in the immediate surrounding

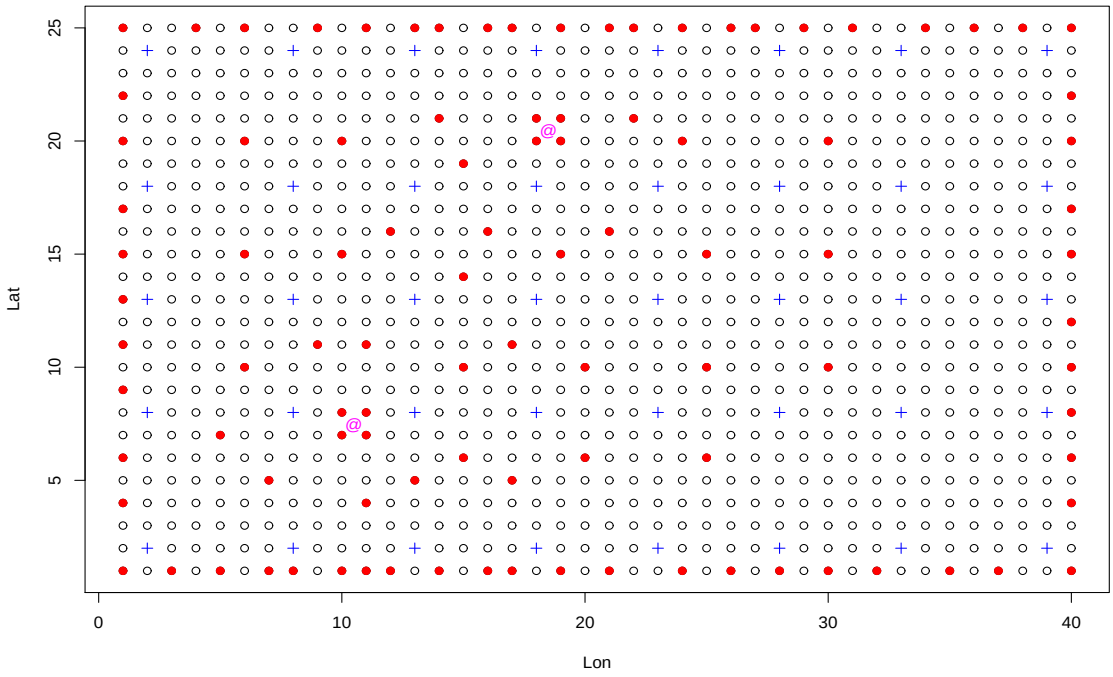

Figure 5.8: Sequential design based on $I(\beta_0, \beta_1, \beta_2)$

area. Not long after, Union Carbide's sister plant in South Charleston, West Virginia also experienced a significant release of acetone and mesityl oxide. Concerned for their safety, both industrial workers and local communities called for freely available information on the chemicals being used in and released from industrial facilities. In response to strong public demand, in 1986, the United States enacted the Emergency Planning and Community Right to Know Act.

Among other things, under Section 313 the Act established the Toxics Releases Inventory (TRI). In its original form, TRI required all businesses in Standard Industrial Classification (SIC) codes 20-39 that employed ten or more employees and released into the air, water, or ground either 10,000 pounds or more of any one of the 350 chemicals on the TRI list or 25,000 pounds or more of any combination of the 350 chemicals to report this information to the USEPA. SIC codes 20-39 cover the follow-

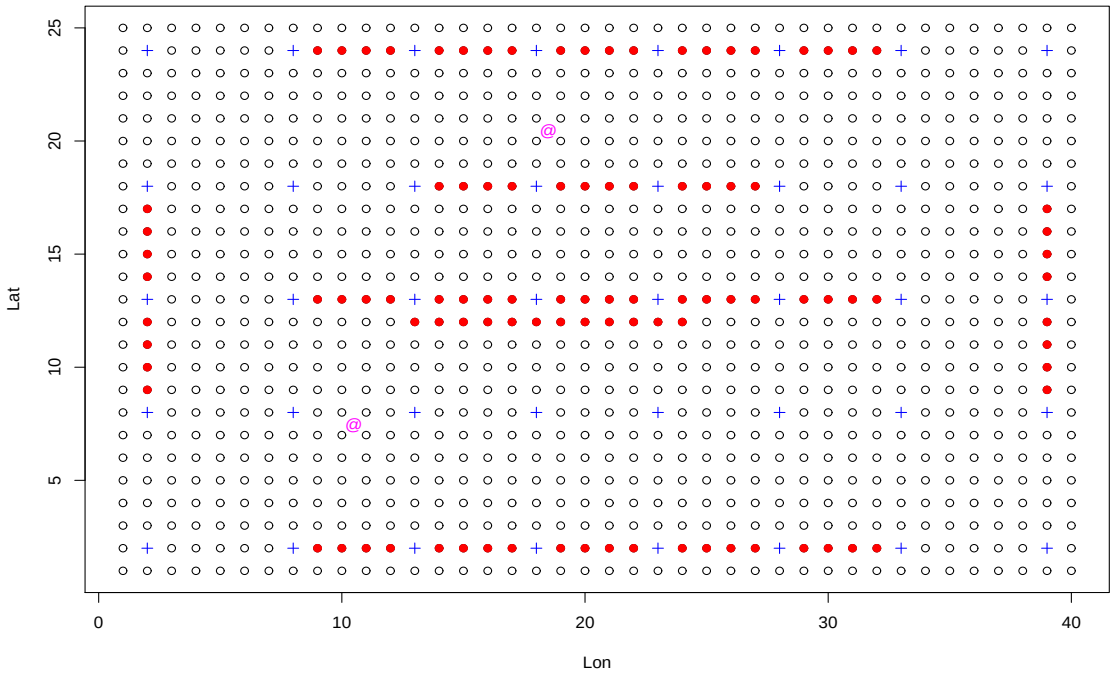

Figure 5.9: Sequential design based on $I(\phi)$

ing industries: food, tobacco, textiles, apparel, lumber and wood, furniture, paper, printing and publishing, chemical, petroleum and coal, rubber and plastics, leathers, stone clay and glass, primary metals, machinery, electrical and electronic equipment, transportation equipment, instruments, and miscellaneous manufacturing. The TRI has subsequently been expanded to include metal mining, coal mining, coal- and oil-fired electric utilities, hazardous waste treatment and disposal facilities, chemicals and allied products wholesale distributors, petroleum bulk plants and terminals and solvent recovery services, reflecting a total of 667 chemicals. In addition, the USEPA has recently reduced the reporting threshold for several chemicals that are considered either persistent or especially toxic, including hexachlorobenzene, mercury, and lead.

Reconsideration of the TRI reporting requirements includes questions regarding: 1) whether smaller facilities (fewer than 10 employees or lower chemical use levels)

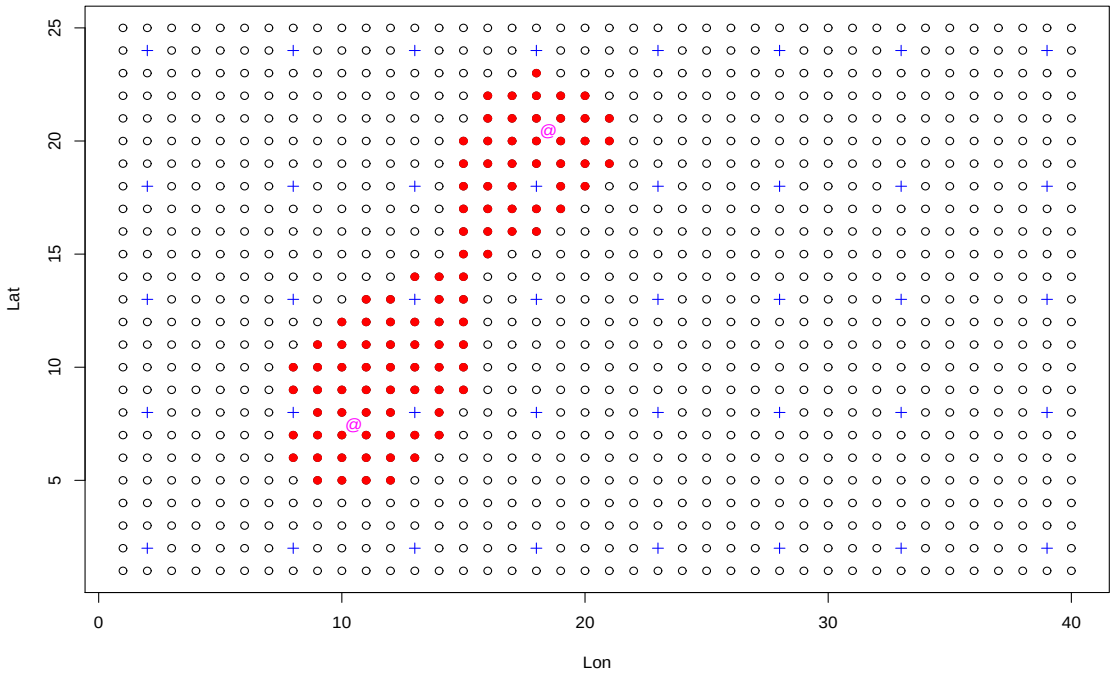

Figure 5.10: Sequential design based on $Y * I(\beta_0)$

in TRI-reporting SIC codes should be required to report their emissions; 2) whether additional SIC codes should be required to report; 3) whether additional chemicals should be added to the TRI list; 4) whether reporting thresholds should be lowered on particular compounds (as was done for hexachlorobenzene, mercury, and lead; 5) whether facilities should be required to report both use and emissions; and 6) whether facilities that previously reported, but do not report currently, should be required to provide an explanation for this change in status. All of these policy questions are substantially hampered by the lack of systematic data on ambient levels of air toxics. Given the paucity of existing data and the cost of collecting new data, an optimized method for sampling design is essential. Take, for example, question (1) above regarding whether smaller facilities should be required to report their emissions. Previous research where emissions are imputed to smaller facilities indicates that

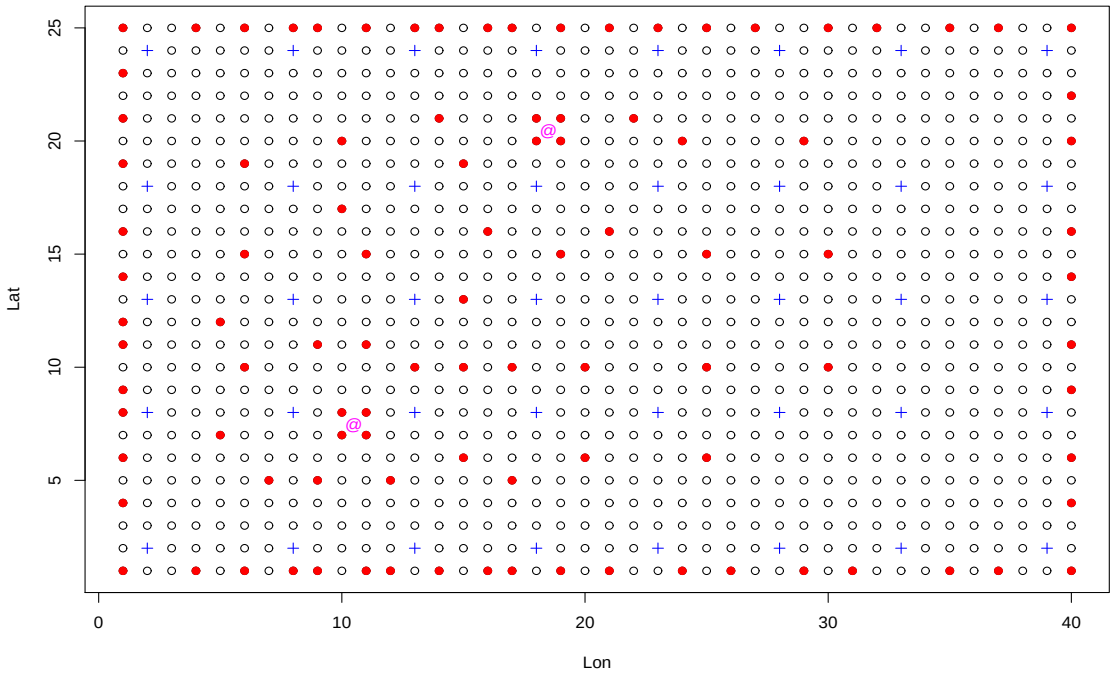

Figure 5.11: Sequential design based on $Y * I(\beta_0, \beta_1, \beta_2)$

including smaller facilities has a substantial impact on the spatial distribution of modeled ambient air concentrations of contaminants (Dolinoy and Miranda 2004). This work, however, necessarily relies on model-based estimates of ambient levels that result from dispersion models. Alternatively, the facilities that already report to TRI would be analogous to the two point sources delineated in the simulation presented here. The smaller facilities represent known point sources with unknown emissions levels. Since the emissions from the smaller facilities are unknown, they are not available for considering the question of whether these smaller facilities should be required to report. Determining how important the smaller facilities are to ambient concentrations can be much more efficiently accomplished through optimization of sampling design.

The optimized sampling design approach described can incorporate multiple sources

of emissions and multiple chemicals emitted. The sampling design can also be shaped to specifically assess exposures to specific sub-populations whose geographic distribution can be characterized. Thus the approach holds great potential for helping scientists, agencies, and communities understand the distribution of TRI chemicals released into the environment.

5.11 Extensions

We have considered approximately optimal sampling design for the setting where we expect intensive one-time sampling rather than sparse continuous monitoring. We have adopted information-based performance criteria and suggested a sequential implementation. We have shown that such a strategy is straightforward to implement with computational demand that is not excessive. We have also suggested utility-weighting as a mechanism for oversampling to achieve specific objectives.

We have developed the approaches in the setting of data from a Gaussian process. However, we can work with non-Gaussian models for the data. In fact, we can also handle discrete data, e.g., binary or count data by representing our model in a hierarchical fashion with the process specification moved to the second stage. (See Section 2.2.3). In these cases we merely replace the Gaussian likelihood with a different first stage likelihood before calculating the information.

A longer view of the exposure data collection might introduce a temporal component in the sense that we may seek to revisit locations that have been previously sampled at a future point in time. If we introduce suitable dependence into our modeling, we can extend our information-based sampling approaches to accommodate this setting as well.

Lastly, the foregoing development is described in terms of the spatial surface of levels for a single contaminant. A broader experiment may consider multiple contam-

inants. If so, we can optimize location selection when levels of several contaminants are sampled at a given location. (As a variant, we may have multiple types of sampling, e.g., ambient sampling, ground deposition sampling, or organism sampling. Similar to the above, we can optimize sampling when multiple types of sampling will be carried out at a location.) In particular, suppose at each parcel we measure levels of say r contaminants. Now, we replace $Y(s_i)$ with an $r \times 1$ vector $\mathbf{Y}(s_i)$. The resulting information gain now depends on both the spatial dependence across locations as well as the dependence between the measurements within each location. A simplified form arises under a separable specification for this error structure (see, e.g., Banerjee, Carlin and Gelfand, 2004 and references therein). The resulting form is the above information multiplied by the within location covariance matrix. Since the latter is free of n, we can use the same criteria as above in this case.

Appendix A

Details in Chapter 3

A.1 The Ornstein-Uhlenbeck process calculation

For the Ornstein-Uhlenbeck process, with ϕ as correlation decay parameter, suppose we observe Y at $t_1, ..., t_n$, the correlation matrix is $R_{n,i,j} = \exp(-\phi|t_i - t_j|)$ and $\bar{R}_{n,i,j} = -|t_i - t_j|R_{n,i,j}$. A Gaussian process $Y(t)$ is Markovian when

$$f\{Y(u) \,|\, Y(s), Y(t)\} = f\{Y(u) \,|\, Y(t)\}, \tag{A.1}$$

for time points $s < t < u$. This shows that the necessary and sufficient condition is

$$\varrho(s,t)\varrho(t,u) = \varrho(s,u)\varrho(t,t). \tag{A.2}$$

One sees from this that of all correlation functions $\exp\{-\phi|s-t|^\alpha\}$, where $\alpha \in (0, 2]$, only the Ornstein-Uhlenbeck yields a Markov process.

For this process, therefore, when sampled at points $t_1 < ... < t_n$, the likelihood may be written in two ways:

$$(2\pi)^{-n/2}\sigma^{-n}|R_n|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{y} - \mu\mathbf{1})^T R_n^{-1}(\mathbf{y} - \mu\mathbf{1})\sigma^2\}, \tag{A.3}$$

where R_n has elements $\exp(-\phi|t_i - t_j|)$, and

$$f(y_1)\prod_{i=2}^{n} f\{y_i \,|\, y_{i-1}\}, \tag{A.4}$$

in terms of the conditional densities

$$Y(t_i) \,|\, \text{past} \sim \mathrm{N}(\mu + \rho_i(y_{i-1} - \mu), \sigma^2(1-\rho_i^2)), \tag{A.5}$$

where $\rho_i = \exp(-\phi|t_i - t_{i-1}|)$. Equating these two expressions gives simplifications for $|R_n|$ and for R_n^{-1}. For the case of equidistant sampling, with $t_i - t_{i-1} = \Delta$ and $\rho = \exp(-\phi\Delta)$, say, where the expression (A.4) becomes

$$\frac{1}{(2\pi)^{1/2}}\frac{1}{\sigma}\exp\Big\{-\frac{(y_1-\mu)^2}{2\sigma^2}\Big\}\prod_{i=2}^{n}\frac{1}{(2\pi)^{1/2}}\frac{1}{\sigma(1-\rho^2)^{1/2}}\exp\Big[-\frac{\{(y_i-\mu)-\rho(y_{i-1}-\mu)\}^2}{2\sigma^2(1-\rho^2)}\Big],$$

So, compared with formula (A.3), we have

$$|R_n| = (1-\rho^2)^{n-1}, \quad \text{or} \quad \log|R_n| = (n-1)\log(1-\rho^2), \tag{A.6}$$

and that R_n^{-1} is a bandmatrix with

$$\begin{aligned} R_n^{i,i} &= \frac{1+\rho^2}{1-\rho^2} \text{ for } 2 \le i \le n-1, \quad R_n^{1,1} = R_n^{n,n} = \frac{1}{1-\rho^2}, \quad \text{and} \\ R_n^{i-1,i} &= R_n^{i,i-1} = -\frac{\rho}{1-\rho^2} \text{ for } 2 \le i \le n. \end{aligned} \tag{A.7}$$

Summing these, the μ information number has the exact formula

$$A_n = 1 + (n-1)\frac{1-\rho}{1+\rho}. \tag{A.8}$$

For $\Delta = L/n$ and $\rho = \exp(-\phi\Delta)$ for dense infill of the $[0, L]$ interval, this leads to

$$A_n \to 1 + \tfrac{1}{2}L\phi. \tag{A.9}$$

We may also utilize the information reached here about R_n^{-1} to work out a formula for B_n of (3.64):

$$\begin{aligned} B_n &= \mathrm{Tr}(R_n^{-1}\bar{R}_n) = \sum_{i,j} R_n^{i,j}\bar{R}_{n,i,j} \\ &= \sum_{|j-i|=1} R_n^{i,j}\bar{R}_{n,i,j} = 2(n-1)\frac{\rho^2}{1-\rho^2}\Delta. \end{aligned} \tag{A.10}$$

For large n, in view of $\Delta = 1/n$ here, $B_n \doteq n/\phi$, growing linearly with n.

Turning to E_n, write

$$\begin{aligned} \mathrm{diag}(R_n^{-1}) &= (a_0, a \ldots, a, a_0), \\ \mathrm{sidediag}(R_n^{-1}) &= (b, b, \ldots, b, b), \end{aligned}$$

where $a_0 = \frac{1}{1-\rho^2}$, $\quad a = \frac{1+\rho^2}{1-\rho^2}$, $\quad b = -\frac{\rho}{1-\rho^2}$. Some work reveals that $R_n^{-1}\bar{R}_n R_n^{-1}$ is also a bandmatrix, say with

$$\begin{aligned} \mathrm{diag}(R_n^{-1}\bar{R}_n R_n^{-1}) &= (c_0, c \ldots, c, c_0), \\ \mathrm{sidediag}(R_n^{-1}\bar{R}_n R_n^{-1}) &= (d, d, \ldots, d, d), \end{aligned}$$

and zeroes outside. Hence

$$\mathrm{Tr}(R_n^{-1}\bar{R}_n R_n^{-1}\bar{R}_n) = \sum_{i,j} (R_n^{-1}\bar{R}_n R_n^{-1})_{i,j}\bar{R}_{n,i,j} = 2(n-1)d\bar{R}_{n,1,2},$$

in that $\bar{R}_n$ has zeroes on its diagonal. Therefore, we only need find a formula for d. In fact,

$$\begin{aligned} (R_n^{-1}\bar{R}_n R_n^{-1})_{1,2} &= \sum_{j,k} R_n^{1,j}\bar{R}_{n,j,k}R_n^{k,2} \\ &= R_n^{1,1}\sum_k \bar{R}_{n,1,k}R_n^{k,2} + R_n^{1,2}\sum_k \bar{R}_{n,2,k}R_n^{k,2} \\ &= a_0(\bar{R}_{n,1,2}R_n^{2,2} + \bar{R}_{n,1,3}R_n^{3,2}) + b(\bar{R}_{n,2,1}R_n^{1,2} + \bar{R}_{n,2,3}R_n^{3,2}) \\ &= -\rho\Delta\frac{1+\rho^2}{(1-\rho^2)^2}. \end{aligned}$$

This leads finally to the explicit formula

$$E_n = (n-1)d\bar{R}_{n,1,2} = (n-1)\Delta^2\rho^2\frac{1+\rho^2}{(1-\rho^2)^2}. \tag{A.11}$$

For general sampling at $t_1 < ... < t_n$, $R_n(i,j) = \rho_{i,j} = \exp(-\phi|t_i - t_j|)$ and the joint distribution for $Y(t_i)$, $i = 1, ..., n$, is:

$$\begin{aligned} f(y_1, ..., y_n) &= \mathrm{N}(\mu, 1) \prod_{i=2}^{n} \mathrm{N}(\mu + \rho_{i-1,i}(y_{i-1} - \mu), 1 - \rho_{i-1,i}^2) \\ &= (2\pi)^{-\frac{n}{2}} \left[\prod_{i=2}^{n} (1 - \rho_{i-1,i}^2)^{-\frac{1}{2}} \right] \left[\exp\big(-\frac{1}{2} \{ (y_1 - \mu)^2 \right. \\ &+ \left. \sum_{i=2}^{n} \frac{1}{1 - \rho_{i-1,i}^2} [y_i - u - \rho_{i-1,i}(y_{i-1} - \mu)]^2 \} \big) \right]. \end{aligned} \tag{A.12}$$

From the above equation, we can find $|R_n|$ and R_n^{-1}. In particular, R_n^{-1} is

$$\begin{pmatrix} \frac{1}{1-\rho_{1,2}^2} & \frac{-\rho_{1,2}}{1-\rho_{1,2}^2} & 0 & 0 & 0 & \dots \\ \frac{-\rho_{1,2}}{1-\rho_{1,2}^2} & \frac{1}{1-\rho_{1,2}^2} + \frac{\rho_{2,3}^2}{1-\rho_{2,3}^2} & \frac{-\rho_{2,3}}{1-\rho_{2,3}^2} & 0 & 0 & \vdots \\ 0 & \frac{-\rho_{2,3}}{1-\rho_{2,3}^2} & \frac{1}{1-\rho_{2,3}^2} + \frac{\rho_{3,4}^2}{1-\rho_{3,4}^2} & \frac{-\rho_{3,4}}{1-\rho_{3,4}^2} & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{-\rho_{n-2,n-1}}{1-\rho_{n-2,n-1}^2} & \frac{1}{1-\rho_{n-2,n-1}^2} + \frac{\rho_{n-1,n}^2}{1-\rho_{n-1,n}^2} & \frac{-\rho_{n-1,n}}{1-\rho_{n-1,n}^2} \\ 0 & 0 & 0 & 0 & \frac{-\rho_{n-1,n}}{1-\rho_{n-1,n}^2} & \frac{1}{1-\rho_{n-1,n}^2} \end{pmatrix} \tag{A.13}$$

So,

$$A_n = \mathbf{1}^T R_n^{-1} \mathbf{1} = -\mathrm{E}\left(\frac{\partial^2 \log f(y_1, ..., y_n)}{\partial \mu^2}\right) = 1 + \sum_{i=2}^{n} \frac{1 - \rho_{i-1,i}}{1 + \rho_{i-1,i}}. \tag{A.14}$$

Result 1: For an OU process on $[0, L]$, $\mathbf{1}^T R_n^{-1}(\phi)\mathbf{1}$ increases as ϕ increases, i.e. $\mathbf{1}^T R_n^{-1}(\phi_1)\mathbf{1} > \mathbf{1}^T R_n^{-1}(\phi_2)\mathbf{1}$ if $\phi_1 > \phi_2 > 0$.

Proof. If $\phi_1 > \phi_2 > 0$, $0 < \rho_{i-1,i}(\phi_1) < \rho_{i-1,i}(\phi_2)$. Thus $\frac{1-\rho_{i-1,i}(\phi_1)}{1+\rho_{i-1,i}(\phi_1)} > \frac{1-\rho_{i-1,i}(\phi_2)}{1+\rho_{i-1,i}(\phi_2)}$. □

Result 2: For an OU process on $[0, L]$, $\lim_{\phi \to 0} \mathbf{1}^T R_n^{-1} \mathbf{1} = 1$ and $\lim_{\phi \to \infty} \mathbf{1}^T R_n^{-1} \mathbf{1} = n$.

Proof. $\rho_{i-1,i}(\phi) \to 1$ as $\phi \to 0$. Thus $\mathbf{1}^T R_n^{-1}(\phi)\mathbf{1} = 1 + \sum_{i=2}^{n} \frac{1-\rho_{i-1,i}}{1+\rho_{i-1,i}} \to 1$. Similarly, $\rho_{i-1,i}(\phi) \to 0$ as $\phi \to \infty$, thus $\mathbf{1}^T R_n^{-1}(\phi)\mathbf{1} \to n$. □

A.2 Proof of Theorem 3

Proof. Let us first consider the finite dimensional case. Correlation matrix R_n has a spectral decomposition $R_n = U\Lambda U^T$, where $U = (u_1, ..., u_n)$ with $UU^T = \sum_{i=1}^{n} u_i u_i^T = \mathrm{I}_n$, and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ such that $\lambda_i \geq \lambda_{i+1} > 0$. So $R_n^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{n} \frac{u_i u_i^T}{\lambda_i}$. This leads to

$$\begin{aligned} \mathbf{x}^T R_n^{-1}\mathbf{x} &= \sum_{i=1}^{n} \frac{(\mathbf{x}^T u_i)^2}{\lambda_i} \\ &= \int \frac{|\psi(\lambda)|^2}{\lambda}\, \mathrm{d}E_\lambda, \end{aligned} \tag{A.15}$$

where $E_\lambda = \sum_{i=1}^{n} \delta(\lambda - \lambda_i)$ is a sum of Dirac measures; $\psi(\lambda) = \int x(t)u_\lambda(t)\, \mathrm{d}H_t$; $H_t = \sum_{i=1}^{n} \delta(t - t_i)$; and $u_\lambda(t)$ is the eigenvector associated with λ.

For the infinite dimensional (continuous) case, we need some spectral theory results for the correlation operator $\mathcal{K}$ corresponding to ϱ. In our case, the domain D is compact and $\mathcal{K}$ is symmetric, bounded and of trace type (i.e. $\mathrm{Tr}(\mathcal{K}) = \int_D \varrho(t,t)\, \mathrm{d}t$). So $\mathcal{K}$ has a spectral representation in terms of a countable set of eigenvalues and eigenvectors which looks similar to the finite dimensional case.

$$\mathcal{K} = \sum_{i=1}^{\infty} \lambda_i P_i,$$

where λ_i is a summable series of positive eigenvalues and P_i is the corresponding projection operators. The projection operator P_i can be represented by a unit vector $\psi_i(t)$. $(\psi_i(\cdot), x(\cdot)) = \int_D \psi_i(t)x(t)\, \mathrm{d}t$ is the Fourier coefficients of $x(\cdot)$ and $x(t) =$

$\sum_{i=1}^{\infty}(\psi_i(\cdot), x(\cdot))\psi_i(t)$. Thus

$$(x, \mathcal{K}^{-1}x) = (\mathcal{K}^{-\frac{1}{2}}x, \mathcal{K}^{-\frac{1}{2}}x) = \sum_{i=1}^{\infty} \frac{|(\psi_i(\cdot), x(\cdot))|^2}{\lambda_i}. \tag{A.16}$$

□

A.3 Karhunen-Loève expansion for the exponential correlation function

For $\varrho(t,s) = e^{-\phi|t-s|}$ on $D = [-L, L]$, we need to solve the integral equation $\lambda\psi(t) = \int_{-L}^{L} e^{-\phi|t-s|}\psi(s)\,\mathrm{d}s$ with $\int_{-L}^{L} \psi_i(s)\psi_j(s)\,\mathrm{d}s = \delta_{ij}$. Note that

$$\lambda\psi(t) = \int_{-L}^{t} e^{\phi(s-t)}\psi(s)\,\mathrm{d}s + \int_{t}^{L} e^{\phi(t-s)}\psi(s)\,\mathrm{d}s. \tag{A.17}$$

Differentiation with respect to t leads to

$$\lambda\psi'(t) = -\phi\int_{-L}^{t} e^{\phi(s-t)}\psi(s)\,\mathrm{d}s + \phi\int_{t}^{L} e^{\phi(t-s)}\psi(s)\,\mathrm{d}s.$$

Differentiation with respect to t again yields

$$\lambda\psi''(t) = -2\phi\psi(t) + \phi^2\int_{-L}^{t} e^{\phi(s-t)}\psi(s)\,\mathrm{d}s + \phi^2\int_{t}^{L} e^{\phi(t-s)}\psi(s)\,\mathrm{d}s = (-2\phi + \phi^2\lambda)\psi(t).$$

So we only need to solve the differential equation

$$\psi''(t) + (2\phi - \phi^2\lambda)\psi(t)/\lambda = 0 \tag{A.18}$$

with the boundary conditions

$$\begin{cases} \phi\psi(L) + \psi'(L) = 0 \\ \phi\psi(-L) - \psi'(-L) = 0 \end{cases}.$$

So the eigenfunctions have the following form:

$$\psi(t) = c_1\cos(wt) + c_2\sin(wt),\ c_1 \text{ and } c_2 \text{ are constants} \tag{A.19}$$

where $w = \sqrt{(2\phi - \phi^2\lambda)/\lambda}$ and $\lambda = 2\phi/(w^2 + \phi^2)$. (It can be shown that $(2\phi - \phi^2\lambda)/\lambda > 0$ in order to make the differential equation solvable.) The boundary conditions become

$$\begin{cases} c_1(\phi\cos(wL) - w\sin(wL)) + c_2(w\cos(wL) + \phi\sin(wL)) = 0 \\ c_1(\phi\cos(wL) - w\sin(wL)) - c_2(w\cos(wL) + \phi\sin(wL)) = 0 \end{cases}.$$

So the solutions are satisfying the following equations:

$$\tan(wL) = \phi/w, \quad c_2 = 0 \quad \text{or} \tag{A.20}$$

$$\tan(wL) = -w/\phi, \quad c_1 = 0. \tag{A.21}$$

Denoting the solutions of (A.20) by $w_{1,i}$ and the solutions of (A.21) by $w_{2,i}$. The corresponding eigenvalues and eigenfunctions are

$$\lambda_{1,i} = \frac{2\phi}{w_{1,i}^2 + \phi^2}, \quad \psi_{1,i}(t) = \frac{\cos(w_{1,i}t)}{\sqrt{L + \sin(2w_{1,i}L)/(2w_{1,i})}} \tag{A.22}$$

and

$$\lambda_{2,i} = \frac{2\phi}{w_{2,i}^2 + \phi^2}, \quad \psi_{2,i}(t) = \frac{\sin(w_{2,i}t)}{\sqrt{L - \sin(2w_{2,i}L)/(2w_{2,i})}}. \tag{A.23}$$

Finally, we have the Karhunen-Loève expansion for the exponential correlation function:

$$\varrho(s,t) = \sum_{i=1}^{\infty} \Big(\lambda_{1,i}\psi_{1,i}(s)\psi_{1,i}(t) + \lambda_{2,i}\psi_{2,i}(s)\psi_{2,i}(t)\Big). \tag{A.24}$$

Appendix B

Details in Chapter 4

Lemma B.1

Define

$$W(s) \equiv Z[K(s-\cdot)] = \int_{\mathbb{R}^2} K(s-t)Z(dt), \tag{B.1}$$

where $Z(dt)$ is an orthogonal random measure satisfying the three conditions below formula (4.10). Assume $Z(dt) \sim N(0, \nu(dt))$. Then $W(s)$ defines a Gaussian process with mean 0 and covariance function in (4.18).

Proof. We first assume $K(\cdot - t)$ is a simple function. We will prove that the distribution of $W(s_1)$ and $W(s_2)$ is jointly normal. Then we generalize two dimensional distribution case to the finite dimensional distribution case. Finally, we relax the simple function assumption. For a simple function $K(\cdot - t)$, according to the definition of $Z[K]$ for simple functions K in Section 4.2.2, suppose $W(s_1) = \sum_{i=1}^{I} a_i Z(A_i)$ and $W(s_2) = \sum_{j=1}^{J} b_j Z(B_j)$, we can make a finer partition C_k $(k = 1, \cdots, r)$ such that $W(s_1) = \sum_{i=1}^{I'} a'_i Z(C_{n_i})$ and $W(s_2) = \sum_{j=1}^{J'} b'_j Z(C_{m_j})$, where $C_{n_i}, C_{m_j} \in \{C_k\}$, $(\bigcup_{i=1}^{I'} C_{n_i}) \bigcup (\bigcup_{j=1}^{J'} C_{m_j}) = \bigcup_{k=1}^{r} C_k$, and C_k are disjoint to each other. By our construction, $Z(C_k)$ are independent normal with mean zero and variance $\nu(C_k)$. Denote

the vector $Z(C_k)$ as $\mathbf{Z}$, then the distribution of $\mathbf{Z}$ is a multivariate normal with mean zero and covariance matrix Σ ($\Sigma_{ij} = \nu(C_i)1_{i=j}$) since any linear combinations of $Z(C_k)$ are univariate normal random variables. Standard normal theory says that $H\mathbf{Z}+\mu$, where H is any $q\times r$ matrix and μ is any non-random q-vector, has a q-variate normal distribution with mean μ and the covariance matrix $H\Sigma H^T$. Here we choose an appropriate $2 \times r$ matrix H such that $H\mathbf{Z}$ is $\begin{pmatrix} W(s_1) \\ W(s_2) \end{pmatrix}$. This shows that $W(s_1)$ and $W(s_2)$ are bivariate normal. The same argument applies when finding the joint distribution of $W(s_1), \cdots, W(s_n)$. As a result, it is an n-variate normal distribution. Finally, for a general $K(\cdot) \in L^2(\nu)$, there exists a sequence of simple functions $\{K_1, K_2, \cdots\}$ which converges to $K(\cdot)$. Then we have a sequence of associated normal vectors $\{\mathbf{W}_1, \mathbf{W}_2, \cdots\}$ with corresponding distribution functions $F_1, F_2, \cdots$ and characteristic functions $\phi_1(t), \phi_2(t), \cdots$. It is easy to see that $\lim_{i\to\infty} \phi_i(t)$ exists, denoted as $\phi(t)$, for all t. $\phi(t)$ is continuous at 0 and it is the characteristic function of a multivariate normal vectors. Thus by the continuity theorem, the distribution function F for the limiting random vector $\mathbf{W}$ is also multivariate normal.

□

Proposition B.2

Assume the conditions about D_r, B_{rj} and t_{rj} described in Section 4.3.1 are satisfied, then $\tilde{W}_{r,m}(s) \xrightarrow{L^2} W(s)$ as $m \to \infty$ and $r \to \infty$ for ever s. Moreover, $Cov(\tilde{W}_{r,m}(s), \tilde{W}_{r,m}(s')) \to Cov(W(s), W(s'))$ as $m \to \infty$ and $r \to \infty$ for ever s and s'.

Proof. Note $\mathrm{E}(W(s)) = 0$ and $\mathrm{E}(\tilde{W}_{r,m}(s)) = 0$. We have

$$\begin{aligned}\mathrm{E}(W(s) - \tilde{W}_{r,m}(s))^2 &= \sigma^2\Big[\int_{\mathbb{R}^2} K^2(s-t)\,\mathrm{d}t + \sum_{j=1}^{m} K^2(s-t_{rj})|B_{rj}| \\ &\quad -2\sum_{j=1}^{m}\big(K(s-t_{rj})\int_{B_{rj}} K(s-t)\,\mathrm{d}t\big)\Big]. \qquad \text{(B.2)}\end{aligned}$$

We first consider the middle term of (B.2). Since $K(\cdot)$ is "nice" (i.e. square integrable on $\mathbb{R}^2$ and it decays to 0 at ∞) and B_{rj} is a square, using familiar arguments as in the Riemann integration theory, for a fixed r,

$$\lim_{m\to\infty}\sum_{j=1}^{m} K^2(s-t_{rj})|B_{rj}| = \int_{D_r} K^2(s-t)\,\mathrm{d}t.$$

Also it is easy to see $\lim_{r\to\infty}\int_{D_r} K^2(s-t)\,\mathrm{d}t = \int_{\mathbb{R}^2} K^2(s-t)\,\mathrm{d}t$.

For the third term in (B.2), similarly,

$$\lim_{m\to\infty} 2\sum_{j=1}^{m}\big(K(s-t_{rj})\int_{B_{rj}} K(s-t)\,\mathrm{d}t\big) = 2\sum_{j=1}^{m}\int_{B_{rj}} K^2(s-t)\,\mathrm{d}t = \int_{D_r} K^2(s-t)\,\mathrm{d}t.$$

So $\lim_{r\to\infty}\lim_{m\to\infty}\mathrm{E}(W(s) - \tilde{W}_{r,m}(s))^2 = 0$.

Next, it is easy to calculate that

$$\mathrm{Cov}(\tilde{W}_{r,m}(s), \tilde{W}_{r,m}(s')) = \sigma^2\sum_{j=1}^{m} K(s-t_{rj})K(s'-t_{rj})|B_{rj}|,$$

which converges to $\int_{D_r} K(s-t)K(s'-t)\,\mathrm{d}t$ for a fixed r when $m\to\infty$. So

$$\lim_{r\to\infty}\lim_{m\to\infty}\mathrm{Cov}(\tilde{W}_{r,m}(s), \tilde{W}_{r,m}(s')) = \int_{\mathbb{R}^2} K(s-t)K(s'-t)\,\mathrm{d}t = \mathrm{Cov}(W(s), W(s')).$$

□

References

Abramowitz, M. and Stegun, I.A. (1965) *Handbook of Mathematical Functions*. New York: Dover Publications.

Abt, M. and Welch, W.J. (1998) Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes. *Canadian Journal of Statistics*, **26,** 127–137.

Adler, R.J. (1981) *The Geometry of Random Fields*. New York: Wiley.

Arbia, G. and Lafratta, G. (2002) Anisotropic spatial sampling designs for urban pollution. *Applied Statistics*, **51,** 223–234.

Assunção, R.M. (2003) Space varying coefficients models for small area data. *Environmetrics*, **14,** 453–473.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

Banerjee, S. and Gelfand, A.E. (2003) On smoothness properties of spatial processes. *Journal of Multivariate Analysis*, **84,** 85–100.

Bernardinelli, L. and Montomoli, C. (1992) Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, **11,** 983–1007.

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

Billingsley, P. (1995) *Probability and Measure*, Third edn. New York: Wiley.

Blackwell, D. and Dubins, L.E. (1962) Merging of opinions with increasing information. *Annals of Mathematical Statistics*, **33,** 882–886.

Bogaert, P. and Russo, D. (1999) Optimal spatial sampling design for the estimation of the variogram based on a least-squares approach. *Water Resources Research*, **35,** 1275–1289.

Brimkulov, U., Krug, G. and Savanov, V. (1986) *Design of Experiments for Random Fields*. Moscow: Nauka.

Brown, P.J., Le, N.D. and Zidek, J.V. (1994) Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics*, **22,** 489–509.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, **10**.

Chen, H.-Sh., Simpson, D.S. and Ying, Z. (2000) Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, **10,** 141–156.

Clayton, D.G. and Kaldor, J.M. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43,** 671–681.

Clyde, M. (1993) Bayesian optimal design for approximate normality. Ph.D. Thesis. University of Minnesota.

Clyde, M.A., Müller, P. and Parmigiani, G. (1995) Optimal design for heart defibrillators. In *Bayesian Statistics in Science and Engineering: Case Studies II* (eds C. Gatsonis, J. S. Hodges, R. E. Kass and N. D. Singpurwalla), pp. 278–292. Springer–Verlag.

Clyde, M.A., Müller, P. and Parmigiani, G. (1996) Inference and design strategies for a hierarchical logistic regression model. In *Bayesian Biostatistics* (eds D.A. Berry and D. Stangl), pp. 297–320.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. London: Chapman-Hall.

Cox, D.R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical (Society B)*, **49,** 1–39.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*. Wiley-Interscience.

Dawid, A.P. (1984) Statistical theory: the prequential approach. *Journal of the Royal Statistical Society (Series A)*, **147,** 278–292.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) Model-based geostatistics (with discussion). *Applied statistics*, **47,** 299–350.

Fedorov, V.V. (1996) *Design of Spatial Experiments, Model Fitting and Prediction*. Oak Ridge National Laboratory.

Furrer, R., Genton, M.G. and Nychka, D. (2005) Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics (under revision)*.

Gelfand, A.E., Banerjee, S. and Gamerman, D. (2005) Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, **16,** 1–15.

Gelfand, A.E., Ecker, M.D., Knight, J.R. and Sirmans, C.F. (2004) The dynamics

of location in home price. *Journal of Real Estate and Financial Economics*, **29,** 149–166.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85,** 398–409.

Ghanem, R.G. and Spanos, P.D. (1991) *Stochastic Finite Elements: A Spectral Approach*. New York: Springer.

Gikhman, I.I. and Skorokhod, A.V. (1974) *The Theory of Stochastic Processes I.* New York: Springer.

Grenander, U. (1981) *Abstract Inference*. New York: Wiley.

Guttorp, P., Meiring, W. and Sampson, P. (1994) A space-time analysis of ground-level ozone data. *Environmetrics*, **5,** 241–254.

Handcock, M.S. and Stein, M.L. (1993) A Bayesian analysis of kriging. *Technometrics*, **35 (4),** 403–410.

Harville, D.A. (1997) *Matrix Algebra for a Statistician's Perspective*. New York: Springer.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57,** 97–109.

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

Huerta, G., Sanso, B. and Stroud, J.R. (2004) A spatio-temporal model for Mexico city ozone levels. *Applied Statistics*, **53,** 231–248.

Ibragimov, I.A. and Rozanov, Y.A. (1978) *Gaussian Random Processes*. New York: Springer.

Ickstadt, K. and Wolpert, R.L. (1999) Spatial regression for marked point processes (with discussion). In *Bayesian Statistics 6, Oxford, UK* (eds J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), pp. 323–341. Oxford University Press.

Jeffreys, H. (1938) Science, logic and philosophy. *Nature*, **141,** 716–719.

Knorr-Held, L. (2002) Some remarks on Gaussian Markov Random Field models for disease mapping. In *Highly Structured Stochastic Systems* (eds Green P.J., Hjort N.L. and Richardson S.). Oxford University Press.

Krige, D.G. (1951) A statistical approach to some basic mine evaluation problems on the witwatersrand. *Journal of Chemical Metallurgical Mining Society of South Africa*, **52,** 119–139.

Le, N. and Zidek, J.V. (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal of Multivariate Analysis*, **43,** 351–374.

Loh, W-H. (2005) Fixed-domain asymptotics for a subclass of Matérn-type gaussian random fields. *Annals of Statistics*, **33,** 2344–2394.

Majumdar, A. and Gelfand, A.E. (2005) Multivariate spatial modeling using convolved covariance functions. *Mathematical Geology*, to appear.

Mardia, K.V. and Marshall, R.J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71,** 135–146.

Matheron, G. (1963) Principles of geostatistics. *Economic Geology*, **58,** 1246–1266.

McBratney, A.B., Webster, R. and Burgess, T.M. (1981) The design of optimal sampling schemes for local estimation and mapping of ragionalized variables. i. theory and method. *Computer & Geosciences*, **7(4),** 331–334.

Metropolis, N., Rosenbluth, A.E., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21,** 1087–1092.

Müller, W.G. (2001) *Collecting Data: Optimum Design of Experiments for Random Fields*. New York: Springer.

Müller, W.G. and Zimmerman, D.L. (1999) Optimal designs for variogram estimation. *Environmetrics*, **10,** 23–37.

Nychka, D. and Saltzman, N. (1998) Design of air-quality monitoring networks. In *Case studies in Environmental Statistics, Lecture Notes in Statistics* (eds Cox L. Nychka D. and Piegorsch W.). Springer Verlag: New York.

Priestley, M.B. (1981) *Spectral Analysis and Time Series*. London: Academic Press.

Pukelsheim, F. (1993) *Optimum Design of Experiments.* New York: Wiley.

Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.

Ritter, K. (1996) Asymptotic optimality of regular sequence designs. *Annals of Statistics*, **24,** 2081–2096.

Rue, H. and Tjelmeland, H. (2002) Fitting gaussian markov random fields to gaussian fields. *Scandinavian Journal of Statistics*, **29,** 31–49.

Sacks, J. and Ylvisaker, D. (1966) Design for regression problems with correlated errors. *Annals of Mathematical Statistics*, **37,** 66–89.

Sahu, S., Gelfand, A.E. and Holland, D.M. (2005) Spatio-temporal modeling of fine particulant matter. *Journal of Agricultural, Biological and Environmental Statistic*, To appear.

Schmidt, A.M. and Gelfand, A.E. (2003) Bayesian coregionalization approach to multivariate pollutant data. *Journal of Geophysical Research–Atmosphere*, **108 (D24),** 8783.

Shaddick, G. and Wakefiled, J. (2002) Modelling multivariate pollutant data at multiple sites. *Applied Statistics*, **51,** 351–372.

Skorokhod, A. and Yadrenko, M. (1973) On absolute continuity of measures corresponding to homogeneous Gaussian fields. *Theory of Probability and its Applications*, **18,** 27–40.

Stein, M.L. (1995) Fixed-domain asymptotics for spatial periodograms. *Journal of the American Statistical Association*, **90,** 1277–1288.

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

Stein, M.L. (2004) Equivalence of gaussian measures for some nonstationary random fields. *Journal of Statistical Planning and Inference*, **123,** 1–11.

Stein, M.L., Chi, Z. and Welty, L.J. (2004) Approximating likelihoods for large spatial datasets. *Journal of the Royal Statistical Society (Series B)*, **66,** 275–296.

Su, Y. and Cambanis, S. (1993) Sampling design for estimation of a random process. *Stochastic Process Appl.*, **46,** 47–89.

Van Trees, H.L. (1968) *Detection, Estimations, and Modulation Theory I*. New York: John Wiley & Sons.

Vecchia, A.V. (1988) Estimation and identification for continuous spatial process. *Journal of the Royal Statistical Society (Series B)*, **50,** 297–312.

Warrick, A.W. and Myers, D.E. (1987) Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23,** 496–500.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41,** 434–449.

Wolpert, R.L. and Ickstadt, K. (1998) Poisson/gamma random field models for spatial statistics. *Biometrika*, **85,** no. 2, 251–267.

Xia, G., Hjort, N.L. and Gelfand, A.E. (2006) Information growth and asymptotic inference under stochastic process models with structured dependence. Technical Report. ISDS, Duke University.

Yadrenko, M. (1983) *Spectral theory of random fields*. New York: Optimization software Inc.

Yaglom, A. M. (1987) *Correlation Theory of Stationary and Related Random Functions: Basic Results*. New York: Springer-Verlag.

Ying, Z. (1991) Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis*, **36,** 280–296.

Ying, Z. (1993) Maximum likelihood estimation of parameters under a spatial sampling plan. *Annals of Statistics*, **21,** 1567–1590.

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

Zhu, L., Carlin, B.P. and Gelfand, A.E. (2003) Hierarchical regression with misaligned spatiotemporal data: relating ambient ozone and pediatric asthma visits in atlanta. *Environmetrics*, **14,** 537–557.

Zhu, Z. (2002) Optimal sampling design and parameter estimation of gaussian random fields. Ph.D. Thesis. Department of Statistics, University of Chicago.

Zhu, Z. and Stein, M.L. (2006) Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference (to appear)*.

Zidek, J.V., Sun, W. and Le, N.D. (2000) Designing and integrating composite networks for monitoring multivariate gaussian pollution fields. *Applied Statistics and Computing*, **49,** 63–79.

Biography

Gangqiang Xia was born on February 8, 1979 in Jiangyin, Jiangsu, China. He earned his B.S. in Statistics and Operations Research in 2001 from Fudan University in Shanghai, China. He has co-authored the following articles:

1. Yi, K., Yu, H., Yang, J., Xia, G. and Chen, Y. (2003). Efficient Maintenance of Materialized Top-k Views. *Proceedings of the 19th International Conference on Data Engineering,* 189-200

2. Xia, G., Miranda, M.L. and Gelfand, A.E. (2006). Approximately Optimal Spatial Design Approaches for Environmental Data. *Environmetrics*, to appear.

3. Xia, G. and Gelfand, A.E. (2005). Stationary Process Approximation for the Analysis of Large Spatial Datasets. Submitted.
   `http://ftp.stat.duke.edu/WorkingPapers/05-24.html`
   Technical Report, ISDS, Duke University.

4. Xia, G., Hjort, N.L. and Gelfand, A.E. (2006). Information Growth and Asymptotic Inference Under Stochastic Process Models with Structured Dependence. (In progress.)

5. Xia, G. and Gelfand, A.E. (2006). Spatial Analysis for Amyotrophic Lateral Sclerosis Disease Among Gulf War Veterans. (In progress.)