2013

Non-parametric and semi-parametric estimation of spatial covariance function

Yang Li
Iowa State University

Non-parametric and semi-parametric estimation of spatial covariance function

by

Yang Li

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Zhengyuan Zhu, Major Professor

Wayne A. Fuller

Mark S. Kaiser

Daniel J. Nordman

Huaiqing Wu

Iowa State University

Ames, Iowa

2013

To Katie and Aizhen

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First I would like to acknowledge the support I have received from my advisor Dr. Zhengyuan Zhu. He has provided me countless advice, given me his wisdom, and shared with me his knowledge. His office door has always been open for me for questions and discussions. Without him this dissertation would not have been completed. I am grateful to the opportunities that Dr. Zhu provided me to meet with the statistics community.

I would also like to thank my dissertation committee, Dr. Wayne A. Fuller, Dr. Mark S. Kaiser, Dr. Daniel J. Nordman, and Dr. Huaiqing Wu, for agreeing to be on my committee, and their suggestions to improve this dissertation.

I wish to thank Dr. Wayne A. Fuller, Dr. Sarah M. Nusser, Dr. Emily Berg, Dr. Jae-Kwang Kim, and Dr. Cindy L. Yu for their help when I was a research assistant at Center for Survey Statistics and Methodology.

I also want to thank my friends and colleagues in CSSM and in the Department of Statistics at Iowa State University. Thank you for giving me a lot of fun time during the past five years.

My family is always my motivation for study and work. My parents are always there to support me no matter what. I would like to dedicate this dissertation to my wife Aizhen. Without her support I would not have been able to complete this work. I would also like to thank my daughter Katie who makes my life a little hard in the past year, but in the meanwhile brings endless happiness and smile to the family.

ABSTRACT

I will present three projects that are related to the modeling of covariance structures on the Euclidean space and on sphere.

Firstly, we propose a method to model isotropic random field on sphere, where a tapered Matérn covariance function is used to capture the local behavior while a nonparametric expansion controls the behavior at large distances. A model selection procedure based on residual sum of squares with penalization is used to reduce over fitting.

Secondly, we address the issue of modeling axially symmetric spatial random fields on sphere with a kernel convolution approach. The observed random field is generated by convolving a latent uncorrelated random field with a class of Matérn type kernel functions. By allowing the parameters in the kernel functions to vary with locations, we are able to generate a flexible class of covariance functions and capture the nonstationary properties. We use pre-computation tables to speed up the computation. For regular grid data on sphere, the block circulant property of the covariance matrix enables us to use Fast Fourier Transform (FFT) to get the determinant and the inverse efficiently.

Thirdly, we proposed a semiparametric variogram estimating method through its spectral representation to model the intrinsically stationary random on $\mathbb{R}^2$. The low frequency part of the spectral density is estimated by solving a regularized inverse problem through quadratic programming. The behavior at high frequencies, however, is modeled via a parametric tail in the form of a power decaying function. The power parameter in the tail is estimated by a log likelihood method.

All proposed methodologies are supplemented with simulation studies and real data analyses.

CHAPTER 1. INTRODUCTION

In this chapter, an introduction to basic concepts of spatial geostatistics which are necessary for the remaining part of the dissertation is presented. These topics include the definitions of stationarity and isotropy, the validity of covariance function and variogram, examples of parametric forms, continuity and differentiability, definition of spectral density and Bochner's theorem, and basic formulas of universal kriging.

1.1 Spatial Random Field

In geostatistics, a spatial random field, or spatial stochastic process Z is a spatial random variable that varies over a continuous subset $\mathcal{D}$. For any fixed, finite set of spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, the random vector $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ is a random vector whose distribution is given by the associated finite-dimensional joint distributions

$$F(z_1, \ldots, z_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) = P(Z(\mathbf{s}_1) \leq z_1, \ldots, Z(\mathbf{s}_n) \leq z_n)\,. \tag{1.1}$$

1.1.1 Stationary and isotropic random field

A spatial realization is actually an incomplete observation of sample size 1. To do any kinds of inference, we need assumptions to introduce replication information. A spatial random field is called *strictly stationary* if (1.1) is invariant under translation, that is,

$$F(z_1, \ldots, z_n; \mathbf{s}_1 + \mathbf{h}, \ldots, \mathbf{s}_n + \mathbf{h}) = F(z_1, \ldots, z_n; \mathbf{s}_1, \ldots, \mathbf{s}_n) \tag{1.2}$$

for any vector $\mathbf{h} \in \mathbb{R}^d$. Strict stationarity is a very stringent condition and a weaker condition suffices in many cases since most statistical methods are based on the moments

rather than the distribution itself. A weaker assumption is the *weakly/second-order stationary* if the first two moments are invariant under spatial shifts. In other words,

$$\begin{aligned} E(Z(\mathbf{s})) &= \mu\,, \\ \mathrm{Cov}(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})) &= C(\mathbf{h})\,, \end{aligned} \tag{1.3}$$

where μ is a constant and $C(\mathbf{h})$ is the covariance function. The second order structure between two locations only depends on their relative difference, rather than the locations themselves. For a Gaussian random field, strict stationarity is equivalent to weak stationarity since the first two moments uniquely determine the distribution.

The covariance function $C(\mathbf{h})$ of a second-order stationary random field has the following properties,

(i) $C(\mathbf{0}) \geq 0$;

(ii) C is an even function such that $C(\mathbf{h}) = C(-\mathbf{h})$;

(iii) $C(\mathbf{0}) \geq |C(\mathbf{h})|$;

(iv) If $C_i(\mathbf{h})$ is a series of valid covariance functions, then

  (a) $\sum_{i=1}^{m} b_i C_i(\mathbf{h})$ is also a valid covariance function if all $b_i \geq 0$;

  (b) $\prod_{i=1}^{m} C_i(\mathbf{h})$ is also a valid covariance function;

  (c) $C(\mathbf{h}) = \lim_{i\to\infty} C_i(\mathbf{h})$ provided that limit exists for all $\mathbf{h}$.

Matheron proposed the use of (semi)variogram as an alternative to the covariance function. For an intrinsically stationary random field,

$$\begin{aligned} E(Z(\mathbf{s})) &= \mu\,, \\ \gamma(\mathbf{h}) &= \frac{1}{2}\mathrm{Var}(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))\,, \end{aligned} \tag{1.4}$$

where γ is the semivariogram. For weakly stationary random process, we have $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$. Intrinsic stationarity is more general than weak stationarity. One-dimensional

Brownian motion, which has semivariogram $\gamma(h) = |h|$, is clearly intrinsically stationary but not weakly stationary.

A special case of weakly stationary random field is when its covariance function $C(\mathbf{h})$ depends on the lag vector $\mathbf{h}$ only through its Euclidean norm $h = \|\mathbf{h}\|$, that is $C(\mathbf{h}) = C^*(\|\mathbf{h}\|)$. We call a process with such a covariance function *isotropic*.

1.1.2 Validity of covariance function and variogram

A real continuous function $C(\cdot)$ is a valid covariance function on $\mathbb{R}^d$ if and only if it is non-negative definite

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0 \tag{1.5}$$

for any integer n, any vector $\mathbf{a} = (a_1, \ldots, a_n)$ and any locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n) \in \mathbb{R}^d$. Similarly, the variogram has to be conditionally negative definite

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0 \tag{1.6}$$

for any integer n, any locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n) \in \mathbb{R}^d$, and any vector $\mathbf{a} = (a_1, \ldots, a_n)$ satisfying $\sum_{k=1}^{n} a_k = 0$.

If $C(\mathbf{h})$ or $\gamma(\mathbf{h})$ is valid in $\mathbb{R}^d$, it is also valid in $\mathbb{R}^p$ for $p < d$. It is general practice to use parametric forms as covariance function (or variogram function). The most popular families of isotropic covariance functions are summarized in Table 1.1 and plotted in Figure 1.1.

Name	$C(h)$	$\gamma(h)$	Parameters
Matérn	$\frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)}(\frac{h}{\phi})^\kappa K_\kappa(\frac{h}{\phi})$	$\sigma^2\left(1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)}(\frac{h}{\phi})^\kappa K_\kappa(\frac{h}{\phi})\right)$	$\kappa, \phi, \sigma^2 > 0$
Spherical	$\sigma^2(1 - \frac{3}{2}\frac{h}{\phi} + \frac{1}{2}(\frac{h}{\phi})^3)\mathbb{I}_{0\leq h\leq\phi}$	$\sigma^2(\frac{3}{2}\frac{h}{\phi} - \frac{1}{2}(\frac{h}{\phi})^3)\mathbb{I}_{0\leq h\leq\phi}$	$\phi, \sigma^2 > 0$
Exponential	$\sigma^2 \exp\{-(h/\phi)\}$	$\sigma^2(1 - \exp\{-(h/\phi)\})$	$\phi, \sigma^2 > 0$
Wave	$\sigma^2\frac{\phi}{h}\sin(\frac{h}{\phi})$	$\sigma^2(1 - \frac{\phi}{h}\sin(\frac{h}{\phi}))$	$\phi, \sigma^2 > 0$
Gaussian	$\sigma^2 \exp\{-(h/\phi)^2\}$	$\sigma^2(1 - \exp\{-(h/\phi)^2\})$	$\phi, \sigma^2 > 0$

Table 1.1: Some commonly used families of covariance functions.

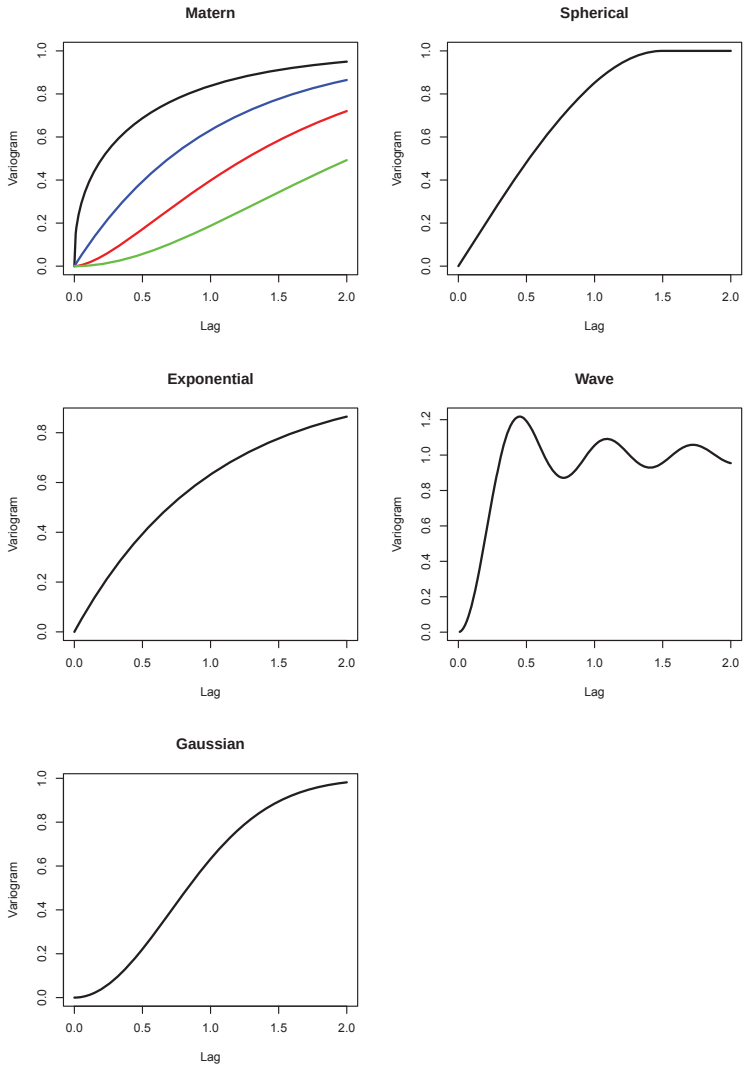


Figure 1.1: Variogram plots for some commonly used parametric models.

1.1.3 Continuity and differentiability

A spatial random field $Z(s)$ is called *mean square continuous* if

$$E(Z(s) - Z(s+h))^2 \to 0 \quad \text{as } h \to 0\,. \tag{1.7}$$

Mean square continuity is equivalent to the continuity of the covariance function at the origin since $E(Z(s) - Z(s+h))^2 = 2(C(0) - C(h))$ for a stationary random field. Additionally, $Z(s)$ is called *mean square differentiable* with mean square derivative $Z'(s)$

if

$$E\left(\frac{Z(s+h)-Z(s)}{h}-Z'(s)\right)^2 \to 0 \quad \text{as } h \to 0\,. \tag{1.8}$$

A stationary random field with correlation function $\rho(h)$ is k times mean-square differentiable if and only if $\rho(h)$ is $2k$ times differentiable at $h = 0$ [Stein (1999)].

1.1.4 Bochner's theorem

Bochner realized that a real-valued continuous function C is nonnegative definite if and only if it can be expressed as

$$C(\mathbf{h}) = \int_{\mathbb{R}^d} \exp(i\mathbf{h}^T\mathbf{x})dF(\mathbf{x}) = \int_{\mathbb{R}^d} \cos(\mathbf{h}^T\mathbf{x})dF(\mathbf{x})\,, \tag{1.9}$$

where F is a symmetric and nonnegative measure on $\mathbb{R}^d$. If F has a Lebesgue density f, then we have

$$C(\mathbf{h}) = \int_{\mathbb{R}^d} \exp(i\mathbf{h}^T\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^d} \cos(\mathbf{h}^T\mathbf{x})\,, f(\mathbf{x})d\mathbf{x} \tag{1.10}$$

and f is called the spectral density of the random field. If the covariance function C is integrable over $\mathbb{R}^d$, the inverse Fourier transform gives the relationship

$$f(\mathbf{x}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \cos(\mathbf{h}^T\mathbf{x})C(\mathbf{h})d\mathbf{h}\,. \tag{1.11}$$

The spectral density of a valid covariance function is required to be nonnegative. The conditions for valid covariance functions, (1.5), are not easy to check in the spatial domain. Sometimes, it is easier to work in the spectral domain to verify the validity of covariance functions.

1.1.5 Spatial prediction

In general, a common problem in geostatistics is to predict the values of the random field at an unobserved location $\mathbf{s}_0 \in \mathbb{R}^d$. A usual way to find the predictor is to minimize

the expected squared prediction error. For a stationary random field, the optimal predictor $\hat{Z}(\mathbf{s}_0)$ is the conditional expectation given the observations at locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, that is,

$$\hat{Z}(\mathbf{s}_0) = E(Z(\mathbf{s}_0)|Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)) \,, \tag{1.12}$$

which is generally difficult to evaluate in closed form since the $(n+1)$-dimensional joint distribution is involved. However, in the special case of Gaussian random field, the conditional distribution of $Z(\mathbf{s}_0)|Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ has a closed form and can be evaluated analytically.

Suppose the random field can be decomposed as a deterministic linear regression part and a stationary random error. The form of the model is then

$$\begin{aligned} \mathbf{Z}(\mathbf{s}) &= \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{e}(\mathbf{s}) \,, \\ Z(\mathbf{s}_0) &= \mathbf{x}(\mathbf{s}_0)^T\boldsymbol{\beta} + e(\mathbf{s}_0) \,, \end{aligned}$$

where $\mathbf{X}$ is the covariate matrix and $\boldsymbol{\beta}$ is the unknown vector of parameters. $\mathbf{e}(\mathbf{s})$ is a zero-mean stationary random process with covariance function $C(\mathbf{h})$. Their covariance structure is $\mathrm{Var}(\mathbf{Z}(\mathbf{s})) = \boldsymbol{\Sigma}$, $\mathrm{Cov}(\mathbf{Z}(\mathbf{s}), Z(\mathbf{s}_0)) = \boldsymbol{\sigma}$, and $\mathrm{Var}(Z(\mathbf{s}_0)) = \sigma_0$, where $\boldsymbol{\Sigma}$, $\boldsymbol{\sigma}$, and σ_0 are known. To find the optimal linear predictor, we intend to minimize the mean-squared prediction error $E(\mathbf{a}^T\mathbf{Z}(\mathbf{s}) - Z(\mathbf{s}_0))^2$ to solve for $\mathbf{a}$. Using Lagrangian multiplier, the result is

$$\hat{Z}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^T\hat{\boldsymbol{\beta}}_{gls} + \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z}(\mathbf{s}) - \mathbf{X}(\mathbf{s})\hat{\boldsymbol{\beta}}_{gls}) \,, \tag{1.13}$$

where $\hat{\boldsymbol{\beta}}_{gls} = (\mathbf{X}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{s}))^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{s})$ is the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$. This procedure is called *universal kriging*. The corresponding universal kriging variance is

$$\sigma^2_{UK}(\mathbf{s}_0) = \sigma_0^2 - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + (\mathbf{x}_0^T - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}_0^T - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^T \,. \tag{1.14}$$

CHAPTER 2. NONPARAMETRIC MODELING OF COVARIANCE FUNCTIONS ON SPHERE

Spatial analysis of large data sets on spheres has drawn attention recently. To better quantify the uncertainty in spatial prediction and estimation, it is often necessary to have a good estimate of the covariance structure of the underlying process. Conventional full likelihood approaches require full specification of parametric models and face the computational obstacle of getting the inverse and determinant of a covariance matrix. Alternatively, nonparametric methods which do not require subjectively specifying a parametric covariance function can be utilized. A valid covariance function on spheres can be written as a constrained expansion of Legendre polynomials. However, the truncation of the expansion introduces too much smoothness. We propose to add a tapered Matérn covariance function to capture the local behavior while the nonparametric expansion controls the behavior at large distances. A model selection procedure based on residual sum of squares with penalization is used to reduce over fitting. Simulation studies show that our method greatly improves the kriging performance. Under the framework of infill asymptotics, our prediction performance becomes comparable to the parametric approach with least squares. Our new method is then applied to the Total Ozone Mapping Spectrometer data which are observed over the entire globe.

2.1 Introduction

Spatial statistics is concerned with statistical models and methods of spatially-indexed data. The main complexity in spatial analysis stems from the fact that observations at

different locations are correlated, which can be quantified by statistical concepts such as covariance function or variogram [Cressie (1993)]. In many cases the number of observations in spatial problems is large, especially in geophysical and environmental applications. For example, the Level 3 TOMS (Total Ozone Mapping Spectrometer) data set that is widely used in the literature has more than 10^4 observations in just one day [Cressie and Johannesson (2008); Jun and Stein (2008); Stein (2008)]. The best linear unbiased predictor, often called kriging predictor in geostatistics, requires the evaluation of the inverse and the determinant of the covariance matrix. This typically leads to computational operations of the order $O(n^3)$ for a likelihood-based approach where n is the number of observations in the study. For large data sets (for example, $n \sim 10^5$ for satellite imaging data), the number of computations is out of reach for the currently available computational resources. This vast amount of data makes it impossible to assess the full likelihood function and people are seeking help from approximations to the full likelihood to carry out the analysis. For random fields on $\mathbb{R}^d$, Vecchia proposed an approximation based on the fact that joint densities can be represented as a product of sequential conditional densities. These conditional densities can be further approximated by conditioning on a subset of the conditioning locations [Vecchia (1988)]. Stein *et al* extended Vecchia's idea by adapting to restricted likelihood [Stein *et al* (2004)]. Stein *et al* included some distant observations in the conditional densities instead of only the nearest neighbors and claimed to be able to gain considerable benefit compared to Vecchia's original approach. Approximations to the full likelihood thus could greatly reduce the computational burden and still maintain a reasonable approximation.

Many large spatial data sets are actually global data in which the curvature of the Earth cannot be ignored. Many of the models and methods developed in $\mathbb{R}^d$ cannot be transported to the sphere without modifications [Gneiting (2012); Huang *et al* (2011)]. There are already some analysis methods designed particularly to handle large spherical data, such as the TOMS data mentioned above. Cressie and Johannesson expressed the

covariance matrix in terms of a diagonal matrix plus a fixed low rank matrix, which makes it possible to compute the likelihood function exactly with massive spatial data [Cressie and Johannesson (2008)]. Stein (2008) further replaced the diagonal matrix with a sparse matrix hoping to capture both the small-scale and large-scale spatial dependence structures. Jun and Stein (2007) proposed an approach to producing space-time covariance functions on spheres by applying differential operators to fully symmetric processes. In this way, nonstationary spatial random fields can be produced with a closed form on sphere $\times$ time. Jun and Stein (2008) applied this method to the analysis of TOMS data in an axially symmetric modeling framework. With the aid of Discrete Fourier Transform (DFT), they were able to calculate the exact likelihood for large data sets on regular grids.

In this paper, we propose a different method of modeling the covariance structure of an isotropic random field on sphere which is computationally efficient. Our approach incorporates both the flexibility of a nonparametric method without having to specify the parametric form of the covariance model, and the contribution from a compactly supported Matérn model whose full log likelihood function can be easily computed. The Matérn model controls the local behavior of the field and contributes most for interpolation purposes. In section 2.2, a characterization of valid covariance functions and variograms on spheres is given. Section 2.3 shows that a main issue associated with the nonparametric approach is the smoothness of the fitted covariance model at small lags. Therefore, in section 2.4, we introduce our modeling approach in which the covariance function has contributions from a compactly supported Matérn covariance structure for local behavior and a variogram fitting part for spatial structure at large distances. The local behavior is estimated with maximum likelihood while the variogram fitting is done with the help of quadratic programming. In section 2.5, we illustrate our method with a simulation study where a comparison with conventional parametric approach is made in terms of kriging performance. The parametric approach is carried

out with both maximum likelihood and weighted least squares. In section 2.6, the new methodology is applied to the TOMS data which are observed over the entire globe. In section 2.7, a discussion of possible future work is presented.

2.2 Covariance Structure on Spheres

Suppose $Z(\cdot)$ is a random field on a sphere S^2 with radius r. The sphere is the Earth in many environmental and geophysical applications. It is customary to specify a location $s \in S^2$ by its latitude L and longitude l, where $-\pi/2 \leq L \leq \pi/2$ and $0 \leq l < 2\pi$. Random field $Z(\cdot)$ is called isotropic (sometimes also referred to as homogeneous) if its first two moments are invariant under any rotations on the sphere. In other words, $E(Z(s)) = \mu$ is a constant for any $s \in S^2$ and its covariance function $\mathrm{Cov}(Z(s_1), Z(s_2))$ depends only on the spherical angle $\theta(s_1, s_2)$ between the two locations, where

$$\theta(s_1, s_2) = \arccos(\sin L_1 \sin L_2 + \cos L_1 \cos L_2 \cos(l_1 - l_2)) \,. \tag{2.1}$$

It is equivalent to write the covariance function as a function of the great circle distance ($\mathrm{gc}(s_1, s_2) = r\theta(s_1, s_2)$) or chordal distance ($\mathrm{cd}(s_1, s_2) = 2r \sin(\theta(s_1, s_2)/2)$) since they have one-to-one correspondence if the radius is fixed.

Similar to the case in $\mathbb{R}^d$, a real continuous function $C(\cdot)$ is said to be a valid covariance function on the sphere S^2 if and only if it is non-negative definite, that is

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j C(\theta(s_i, s_j)) \geq 0 \tag{2.2}$$

for any integer n, any vector $\mathbf{a} = (a_1, \ldots, a_n)$ and any locations $(s_1, \ldots, s_n) \in S^2$.

Schoenberg provided a nonparametric characterization of a valid covariance function for isotropic process on the sphere [Schoenberg (1942)]. A real continuous function $C(\cdot)$ is a valid isotropic covariance function on the sphere if and only if it can be expressed in the form

$$C(\theta) = \sum_{k=0}^{\infty} b_k P_k(\cos\theta) \,, \tag{2.3}$$

where coefficients b_k are non-negative real numbers satisfying $\sum_{k=0}^{\infty} b_k < \infty$, and $P_k(\cdot)$ are the Legendre polynomials [Abramowitz and Stegun (1965)]. By the orthogonal properties of the Legendre polynomials $\int_{-1}^{1} P_m(x)P_n(x) = \frac{2}{2n+1}\delta_{mn}$ (where δ_{mn} denotes the Kronecker delta which equals 1 if $m = n$ and 0 otherwise), the coefficients b_k can be calculated explicitly as

$$b_k = \frac{2n+1}{2} \int_0^{\pi} C(\theta) P_k(\cos\theta) \sin\theta d\theta \,. \tag{2.4}$$

For some form of covariance function, it is straightforward to check if it is a valid covariance function on the sphere by evaluating the above integral. Huang *et al* (2011) showed that many valid covariance functions in $\mathbb{R}^d$ are no longer valid in S^2, including Gaussian and Matérn models. Gneiting (2012) further proved that a Matérn covariance function is valid on the sphere if and only if its smoothness parameter is no greater than 1/2. Generally, valid covariance functions on the sphere are obtained by restricting covariance functions in $\mathbb{R}^3$ in S^2. That is, if a function $C_0(h)$ is a valid covariance function in $\mathbb{R}^3$, a new function defined as $C(\theta) = C_0(2\sin(\theta/2))$ is a valid covariance function on the unit sphere.

Parallel to the intrinsically stationary process in $\mathbb{R}^d$, it is straightforward to define an intrinsically stationary process in S^2 [Huang *et al* (2011)]. Suppose a random field $Z(s)$ satisfies $E(Z(s)) = \mu$ and $\mathrm{Var}(Z(s_1) - Z(s_2)) = 2\gamma(\theta(s_1, s_2))$ for all $s, s_1, s_2 \in S^2$, then $Z(s)$ is said to be intrinsically stationary in S^2 with $2\gamma(\cdot)$ the variogram. The variogram has to be conditionally negative definite,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\theta(s_i, s_j)) \le 0 \tag{2.5}$$

for any integer n, any locations $(s_1, \ldots, s_n) \in S^2$, and any vector $\mathbf{a} = (a_1, \ldots, a_n)$ satisfying $\sum_{k=1}^{n} a_k = 0$. Similar to (2.3), a continuous function $2\gamma(\cdot)$ satisfying $\gamma(0) = 0$ is conditionally negative definite if and only if

$$2\gamma(\theta) = \sum_{k=1}^{\infty} b_k (1 - P_k(\cos\theta)) \,, \tag{2.6}$$

where coefficients b_k are non-negative real numbers and satisfy $\sum_{k=1}^{\infty} b_k < \infty$, and $P_k(\cdot)$ are the Legendre polynomials. Yaglom noted that the space of valid variograms coincides with the space of valid covariances if the random fields are on the sphere [Yaglom (1987)]. For any valid variogram $2\gamma(\theta)$, it is always possible to construct a covariance function $C(\theta) = c_0 - \gamma(\theta)$ with $c_0 \geq \int_0^{\pi} \gamma(\theta) \sin\theta d\theta$ [Huang *et al* (2011)]. Therefore, it is equivalent to work with covariance function or variogram function in S^2.

2.3 Nonparametric Approach

To guarantee the required positive definiteness of the covariance function, it is convenient to fit a parametric model with a small number of parameters. A variety of methods can be used for the fitting, including likelihood-based approaches and least squares methods. It is known that the likelihood-based estimation has some nice asymptotic properties [Mardia and Marshall (1984)]. However, the choice of the model is subjective and requires additional assumptions about the distribution of the underlying random process. Sometimes it is more desirable to use nonparametric methods to estimate covariance functions or variograms, and (2.3) and (2.6) give us a good starting point. Nonparametric variogram estimation in $\mathbb{R}^d$ has been discussed in [Shapiro and Botha (1991); Genton and Gorsich (2002); Cherry *et al* (1996); Ecker and Gelfand (1997); Huang *et al* (2011)]. In $\mathbb{R}^d$, a valid isotropic covariance function takes the form of $C_d(h) = \int_0^{\infty} \Omega_d(ht) F(dt)$, where $\Omega_d(x) = (2/x)^{(d-2)/2}\Gamma(d/2)J_{(d-2)/2}(x)$ form a basis for functions in $\mathbb{R}^d$. Here $F(\cdot)$ is a non-decreasing function, $\Gamma(\cdot)$ is the gamma function, and $J_\nu(\cdot)$ is the Bessel function of the first kind with order ν. By taking F a step function with jump points t_k [Shapiro and Botha (1991); Cherry *et al* (1996)], a nonparametric covariogram estimator has the form

$$\hat{C}_d(h) = \sum_{k=1}^{m} p_k \Omega_d(t_k h)\,. \tag{2.7}$$

Notice the similarity between (2.7) and (2.3). However, in the expansion (2.7) in $\mathbb{R}^d$, the locations of the jumps t_k have to be determined beforehand, usually in an *ad hoc* way. Genton and Gorsich (2002) argued that t_k should be taken as the root of the Bessel functions $J_\nu(t_k) = 0$. The expansion on spheres (2.3), however, requires only one input value, the maximum order of the Legendre polynomials. As we will shown in section 2.4.3, this input can be determined by model selection.

One way to get a nonparametric fit of the covariance function or variogram is to first compute an empirical covariogram/variogram, then use a least squares approach to get the coefficients in (2.3) or (2.6) by minimizing the L^2 distance between empirical estimates and fitted curve. However, this seemingly straightforward approach has a couple of significant drawbacks. Firstly, the fitted model is too smooth in terms of differentiability. The following theorem clarifies this statement.

Theorem 2.3.1. *Let $C(\theta)$ be an isotropic covariance function on sphere with expansion $C(\theta) = \sum_{k=0}^{\infty} b_k P_k(\cos\theta)$. If the coefficients b_k are of the order $O(k^{-2m-1-\delta})$ as k goes to infinity where $\delta > 0$, the corresponding random field $Z(s)$ is m times mean square differentiable.*

The proof of this theorem requires iterative properties of Legendre polynomials and is shown in Appendix A.

Corollary 2.3.2. *If the coefficients b_k only have finite nonzero terms, that is, $b_k = 0$ for k bigger than a finite integer, the corresponding random field $Z(s)$ is infinitely mean square differentiable.*

In applications, a common practice is to use only a finite number of terms in the expansion (2.3) and the coefficient is estimated by fitting to empirical variogram estimates through least squares type methods. Such an approach has to potential problems. First, as shown in the corollary above, the obtained random field will be infinitely mean square differentiable which is generally considered as unrealistic and should be avoided

in real physical processes [Stein (1999)]. The second issue is associated with the positive definiteness of the expansion with finite number of terms. A necessary condition for the covariance function to be strictly positive definite on a sphere is that the coefficient b_k in (2.3) are positive with at most finite number of zeros [Xu and Cheney (1992); Schreiner (1997)]. If only a finite number of Legendre polynomials are used in (2.3), the resulting covariance function is positive definite but not strictly positive definite. We propose a method which incorporates both the flexibility of a nonparametric estimation without having to specify the parametric forms of the covariance model, and the contribution from a compactly supported Matérn model whose full loglikelihood function can be easily computed. The tapered Matérn model controls the local behavior of the field and contributes most for interpolation purposes. Such a model is guaranteed to be positive definite, and the smoothness of the random process is determined by the local Matérn model which can be estimated from the data.

2.4 Methodology

2.4.1 Compactly supported Matérn covariance function

There are many families of commonly used covariance functions, including spherical, exponential, and Matérn families [Stein (1999)]. The first two classes do not have a parameter that controls the differentiability around the origin. Their regularities at the origin are fixed which in turn control the quadratic mean differentiability of the underlying process. Matérn covariance functions, on the other hand, do have a parameter that is related to the mean square differentiability of the random process and thus are the preferred model for many statisticians in kriging. A Matérn covariance function takes the form of

$$C(h;\sigma^2,\psi,\nu) = \sigma^2 \frac{2}{\Gamma(\nu)} \left(\frac{h}{2\psi}\right)^{\nu} K_\nu \left(\frac{h}{\psi}\right) , \tag{2.8}$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind, $\Gamma(\cdot)$ is the gamma function, σ^2 is the variance of the process, ψ is the scale parameter, and ν is the smoothness parameter which is related to the differentiability of the process. The larger ν is, the smoother the random field. In particular, the random field will be k-times mean square differentiable if and only if $\nu > k$. When the value of ν is an integer plus a half $\nu = p + \frac{1}{2}$ where p is a non-negative integer, the covariance function is the product of an exponential function and a polynomial of degree p. If $\nu = \frac{1}{2}$, Matérn covariance function reduces to the exponential covariance function [Stein (1999)].

The idea of tapering covariance function has been used to ease the computational burden in likelihood approaches [Furrer *et al* (2006); Kaufman *et al* (2008)]. A tapered covariance function is exactly zero when the distance between two observations is bigger than a threshold distance. Suppose the original covariance function is $C_0(x; \boldsymbol{\alpha}_0)$ and $C_{\text{taper}}(x; \boldsymbol{\alpha}_{\text{taper}}, \xi)$ is the tapering function which is an isotropic function being identically 0 whenever $x > \xi$. The tapered compactly supported covariance function is then defined as

$$C(x; \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_{\text{taper}}, \xi) = C_0(x; \boldsymbol{\alpha}_0) C_{\text{taper}}(x; \boldsymbol{\alpha}_{\text{taper}}, \xi)\,. \tag{2.9}$$

In this way, the problem is transformed to a sparse approximate linear system and can be efficiently solved using sparse matrix algorithms [Furrer *et al* (2006)].

In our approach, the tapered covariance function is mainly used to get the local behavior of the underlying random process. Gneiting proposed the following compactly supported correlation function Gneiting (2002)

$$\phi(t) = \left[(1-t)\frac{\sin(2\pi t)}{2\pi t} - \frac{1}{\pi}\frac{1-\cos(2\pi t)}{2\pi t}\right] I(0 \le t \le 1)\,, \tag{2.10}$$

where $I(\cdot)$ is an indicator function so that $\phi(t)$ is identically 0 for $t > 1$. This compactly supported covariance function is three times differentiable for $t > 0$. It has both statistical and computational advantages over the once differentiable spherical model and the twice differentiable pentaspherical model. We use (2.10) as our tapering function.

2.4.2 Model

Variograms are defined for intrinsically stationary processes which are more general than the second-order stationarity. Moreover, the method of moments estimator of variogram is unbiased while the estimator of covariance function is not. Therefore, we choose to write our model in terms of variograms. The model we propose has the following form

$$\gamma(\theta; \alpha_p, \alpha_l) = \gamma_{\text{poly}}(\theta; \alpha_p) + \gamma_{\text{local}}(\theta; \alpha_l)\,, \tag{2.11}$$

where $\gamma_{\text{poly}}(\cdot)$ is the expansion in terms of Legendre polynomials as in (2.6) and $\gamma_{\text{local}}(\theta) = \gamma_M(\theta)\gamma_C(\theta)$, where $\gamma_M(\cdot)$ is a Matérn variogram function and $\gamma_C(\cdot)$ is the compactly supported variogram function in (2.10). α_p are the parameters in the polynomial expansion (2.6) which consists of the maximum order M in the polynomial expansion and the coefficients of the polynomials up to order M. α_l includes all parameters in the Matérn covariance function and the compactly supported covariance function. Since the sum and product of two valid variograms are also valid, (2.11) is then a valid variogram on the sphere S^2. Model (2.11) takes the following explicit form

$$\begin{aligned}
\gamma(\theta; \alpha_p, \alpha_l) &= \gamma(\theta; \mathbf{b}_k, M, \sigma^2, \psi, \nu, s_r) \\
&= \sum_{k=1}^{M} b_k (1 - P_k(\cos\theta)) \\
&+ \sigma^2 \left(1 - \frac{2}{\Gamma(\nu)} \left(\frac{1}{\psi} \sin\frac{\theta}{2}\right)^{\nu} K_\nu \left(\frac{2}{\psi} \sin\frac{\theta}{2}\right)\right) \left(1 - \phi\left(\frac{2\sin\frac{\theta}{2}}{s_r}\right)\right)
\end{aligned} \tag{2.12}$$

where $\phi(\cdot)$ is given in (2.10). M is the maximum order of the Legendre polynomials which will be determined via AIC/BIC-based model selection in section 2.4.3. s_r is the range of the tapering function which is set to constrain the impact range of the tapered Matérn covariance function.

2.4.3 Estimation procedure

The first step in a nonparametric estimation of the variogram is to get an empirical estimate. The classical estimator of the variogram based on the method of moments was proposed by Matheron (1962), which takes the form of

$$2\gamma(\theta) = \frac{1}{N_\theta} \sum_{N_\theta} (Z(s_i) - Z(s_j))^2 \,, \tag{2.13}$$

where the set N_θ consists of the location pairs such that $\|s_i - s_j\| \in \epsilon(\theta)$ and $\epsilon(\theta)$ is a tolerance neighborhood at distance θ if the observations are irregularly located. This intuitive empirical variogram estimator has some nice properties such as unbiasedness, evenness, and being zero at lag zero. On the other hand, it faces the difficulty of being sensitive to outlier observations. Cressie and Hawkins (1980) suggested an estimator that alleviates the effect of outliers and proposed

$$2\tilde{\gamma}(\theta) = \frac{\left(\frac{1}{N_\theta} \sum_{N_\theta} |Z(s_i) - Z(s_j)|^{1/2}\right)^4}{0.457 + \frac{0.494}{N_\theta}} \,. \tag{2.14}$$

This estimator is not unbiased, but the denominator is chosen to achieve approximate unbiasedness.

To get estimates of local behavior, it is necessary that the minimum spacing between observations is smaller than the range of the tapering function s_r. Under that condition, the local variogram γ_{local} is fitted first by maximizing the local likelihood of observations with small spacings. The loglikelihood for the tapered covariance function is

$$l(\alpha_l; Z) = -\frac{1}{2} \log |\det \Sigma_{\alpha_l}| - \frac{1}{2} (Z - \mu)^T \Sigma_{\alpha_l}^{-1} (Z - \mu) \tag{2.15}$$

where Z are the observations within that small range. It is also possible to maximize the log likelihood for several approximately independent regions simultaneously as shown in section 2.6. In that case, $l(\alpha_l, Z) = \sum_{i=1}^{I} l(\alpha_l, Z_i)$ where $l(\alpha_l, Z_i)$ is the log likelihood function within the i^{th} region. The parameter estimate for the local Matérn covariance function is

$$\hat{\alpha}_l = \arg\max_{\alpha_l} l(\alpha_l; Z) \,. \tag{2.16}$$

Evaluation of (2.15) requires inverting the covariance matrix which is difficult if the number of observations in the small region is large. If the observations get denser and denser, the dimension of Σ_{α_l} will increase. In that case, we can adjust the tapering range s_r so that the likelihood evaluation is at a manageable level while still capturing the local behavior of the random process.

The contribution of the Legendre polynomials in variogram estimate is

$$\tilde{\gamma}_{\text{poly}}(\theta) = \tilde{\gamma}(\theta) - \gamma_{\text{local}}(\theta, \hat{\alpha}_l)\,. \tag{2.17}$$

The nonnegative coefficients $\mathbf{b}$ can be obtained by minimizing a constrained objective function

$$S(\mathbf{b}) = (\tilde{\gamma}_{\text{poly}}(\theta) - H\mathbf{b})^T W(\tilde{\gamma}_{\text{poly}}(\theta) - H\mathbf{b})\,, \tag{2.18}$$

while requiring $\mathbf{b} \geq \mathbf{0}$. W is an $N \times N$ weight matrix approximating the structure of the estimated variogram where N is the number of bins in the empirical variogram estimate. In many cases W is taken as an identity matrix for the sake of computational simplicity by ignoring all the correlation between variogram estimates at different lags. Cressie (1985) proposed a weighted least squares method for variogram estimation in which the variance structure can be approximated by

$$\text{Var}(\gamma(\theta_j)) \approx \frac{2\gamma(\theta_j)^2}{|N(\theta_j)|}\,. \tag{2.19}$$

In (2.18), H is an $N \times m$ matrix with entries $H_{ij} = 1 - P_i(\cos\theta_j)$ where m is the number of Legendre polynomial terms in the expansion. To fit a valid covariance function of the form (2.6), the coefficients b_k are required to be non-negative. The linear constraints $\mathbf{b} \geq \mathbf{0}$ make the optimization problem a quadratic programming problem with linear constraints, which can be solved by standard numerical methods [Nocedal and Wright (2006)].

To implement our procedure in practice, a data-driven choice of the maximum order M of the Legendre polynomials will be needed to avoid overfitting. Akaike's information

criterion (AIC) and Bayesian information criterion (BIC) can be used for model selection purposes which penalize for model complexity. For each integer m, a (weighted) least squares estimate $\hat{\alpha}_p(m) = (m, \mathbf{b}(m))$ is obtained from (2.18). A quantity which takes into account both the goodness-of-fit and the model complexity is defined as

$$\Lambda(m) = \sum_{i=1}^{n} [\tilde{\gamma}(\theta_i) - \hat{\gamma}_{\text{poly}}(\theta_i, \hat{\alpha}_p(m))]^2 + \lambda m \,, \tag{2.20}$$

where λ equals to 2 for AIC and $\log n$ for BIC. The plot of $\Lambda(m)$ versus m will have a V-shape pattern and the integer M which minimizes $\Lambda(m)$ is selected. The final fitted variogram will be

$$\hat{\gamma}(\theta; \hat{\alpha}) = \gamma_{\text{local}}(\theta, \hat{\alpha}_l) + \gamma_{\text{poly}}(\theta, \hat{\alpha}_p(M)) \,. \tag{2.21}$$

In the simulation studies, we find that these two criteria do not make much difference in terms of the selected values for M.

After the fitted model is obtained in (2.21), it is straightforward to do interpolation at unobserved sites.

2.5 Simulations

In the simulation study, the observation locations are regularly distributed on a unit sphere covering 2π longitude and half of the latitude, that is, the region $[0, 2\pi) \times [-\pi/4, \pi/4]$. The spacings between neighboring sites are 10° on latitude and 20° on longitude. We also include a small patch in our simulation locations which has spacing 1.3° between neighboring sites on both directions. Since our proposed model (2.11) includes a term from compactly supported covariance function, it is necessary to have data with small spacing to be able to have a reliable estimate of parameters α_l. See Figure 2.1for details.

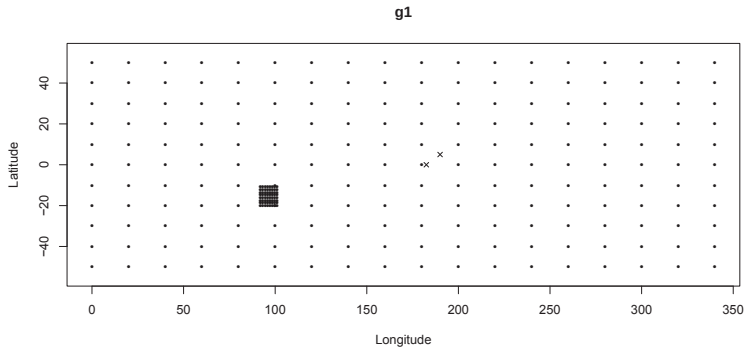
It is known that there are two distinct asymptotic frameworks in spatial statistics: increasing domain asymptotics where the observation domain is increased while keeping a constant spacing, and infill domain asymptotics where more data are collected by

sampling more densely in a fixed region [Stein (1999)]. For random fields on a fixed sphere, the latter is the relevant one. To get an idea of the performance in the infill asymptotics, we reduce the spacing of the regular grid by half each time, and get two more sampling configurations as shown in Figure 2.1 (b) and (c). These three configurations are labeled g1, g2, and g3, with numbers of simulated observations 262, 799, and 2893, respectively. The simulations are carried out using Cholesky decomposition. Suppose the simulation locations are $\mathbf{s} = (s_1, \ldots, s_n)$ and the true covariance matrix is K. Cholesky decomposition gives us $K = L^T L$ where L is an upper-triangular matrix. Then $Z = L^T Y$ has the desired distribution on $\mathbf{s}$, where $Y \sim N(0, I)$. Cholesky decomposition method works perfectly fine in this simulation. However, if the number of sampling locations is sufficiently large, Cholesky decomposition of the relevant covariance matrix is not feasible and other methods such as circulant embedding should be utilized instead [Wood and Chan (1994)].

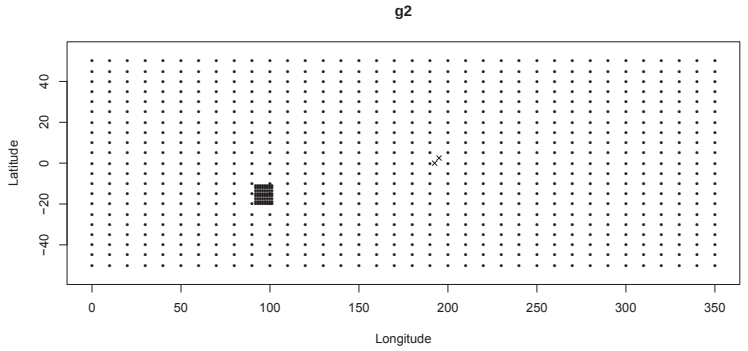
We choose two kriging locations for each setup. One is in the center of a lattice square and the other one is at a fixed distance from its nearest lattice point which is suitable for comparison among three configurations. These points are marks with $\times$ on Figure 2.1. Following [Stein (1999); Im *et al* (2007)], we define a prediction performance measure that is more appropriate for interpolation purposes. Suppose that $\hat{Z}_0(s)$ is the predicted value at location s using the true covariance function C_0 and $\hat{Z}_i(s)$ is the predicted value with covariance function C_i (which may be misspecified). Let $e_i(s) = Z(s) - \hat{Z}_i(s)$ be the prediction error. E_0 is the expectation under the true covariance function C_0. Then $E_0 e_0^2$ is the mean squared prediction error (MSPE) of the best linear unbiased predictor (BLUP) or the kriging variance. Im *et al* (2007) defined a quantity IPE(s) which indicates the increase in prediction error at location s:

$$\text{IPE}(s) = \frac{E_0 e_i^2(s)}{E_0 e_0^2(s)} - 1 = \frac{E_0(\hat{Z}_i(s) - \hat{Z}_0(s))^2}{E_0 e_0^2(s)}\,. \tag{2.22}$$

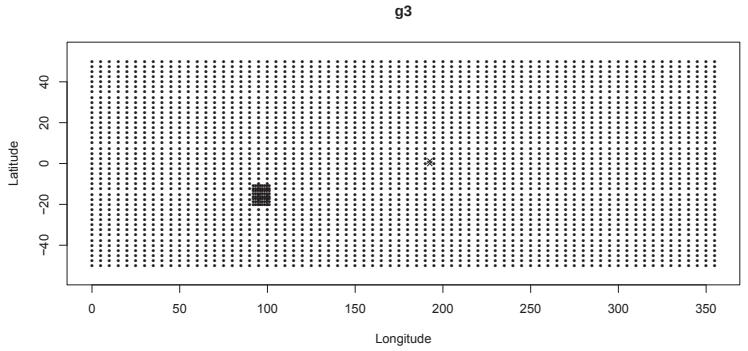
This quantity represents the extra mean squared prediction error introduced by predicting with an estimated (possibly misspecified) covariance function instead of the true



(a)



(b)



(c)

Figure 2.1: Simulation locations are on regular grids within $[0, 2\pi) \times [-\pi/4, \pi/4]$. The small patch in the lower left region is used for local behavior estimation. Two kriging locations are marked with ($\times$) symbol. One of them is in the center of a lattice square and the other one is at a fixed distance from its nearest lattice point.

one. Smaller IPE value indicates a better kriging performance for the corresponding covariance function. The numerator on the right-hand side of (2.22) can be estimated by computing the sample mean of the squared difference between the predicted values using the true covariance function and the predicted values using covariance function C_i, at location s,

$$\frac{1}{N}\sum_{l=1}^{N}(\hat{Z}_i(s,l)-\hat{Z}_0(s,l))^2\,, \tag{2.23}$$

where N is the total number of simulations and l is the index for simulation which runs from 1 to N.

We compare four different estimation methods: a nonparametric Legendre polynomial expansion as in (2.6), the model we proposed which includes a compactly supported Matérn covariance function and a polynomial expansion in (2.11), a full maximum likelihood method with parametric Matérn covariance structure, and a parametric Matérn model using least squares. These four methods are labeled as LPE, CM+PE, MaMLE, and MaLS, respectively. The true covariance function in the simulation study is taken to be a Matérn type with parameters $\sigma^2 = 2$, $\nu = 2$, $\eta = 0.5$, and the Handcock-Wallis range parameter $\alpha = 0.3$ [Handcock and Wallis (1994)]. The simulations are carried over in the densest configuration "g3", out of which the other two configurations are subsets. The total number of simulations is 100.

The full maximum likelihood estimation method is very computationally intensive. Therefore, we only use it for the most sparse sampling setup "g1", Figure 2.1 (a). For the other two configurations, we cannot get the estimation done in a reasonable time period. Possible alternatives to approximate the full likelihood include the methods described in [Vecchia (1988); Stein *et al* (2004)].

For all other three estimating methods, the first step is to get the empirical variogram estimate with the robust estimator (2.14). The weighted least square estimation minimizes the weighted sum of squares $(\tilde{\gamma}(\theta)-\gamma(\theta,\alpha))^T W(\alpha)^{-1}(\tilde{\gamma}(\theta)-\gamma(\theta,\alpha))$ where $W(\alpha)$ is a diagonal matrix with diagonal elements given by (2.19). Zimmerman and Zimmer-

man (1991) found that the ordinary least squares has similar performance compared to the weighted least squares estimators. We also find similar properties between the two in this study.

In CM+PE method, the next step is to get the estimates of the local Matérn covariance function $\gamma_{\text{local}}(\theta, \hat{\alpha}_l)$ by maximizing the local loglikelihood. Instead of being estimated from the data, the range of the compactly supported covariance function is fixed to be 0.2 for the simulation study which is the maximum spherical angles between the sampling locations in the densely-spaced patch. Afterwards, its contribution is subtracted from the empirical variogram and the remaining partial variogram is fitted with Legendre polynomials with model selection. For each integer m ranging from 1 to 20, the quantity $\Lambda(m)$ defined in (2.20) is plotted against m and the integer M corresponding to the minimum value is selected as the order of the Legendre polynomials for the nonparametric variogram fitting used in (2.6). As shown in (2.21), the model selection procedure can be carried out with AIC or BIC criteria. It is found that these two choices differ little in terms of the selected optimal values M. In the following, we only present the results with BIC selection criteria since BIC generally has better asymptotic properties. Figure 2.2 shows the plot for one of the realizations which shows a V-shaped pattern with selected order of 10.

For LPE method, the empirical variograms are directly fitted with nonparametric Legendre polynomials in (2.6). Note that the same model selection procedure is also performed for LPE.

To push our results to the large sample limit and get a taste of the asymptotic performance of the CM+PE method, the local Matérn likelihood estimation is based on all 100 realizations. That is, the log likelihood function which needs to be maximized is actually the summation of 100 individual likelihood functions since all realizations are simulated independently. Thus, we have just one set of estimates for the local Matérn covariance function. Similarly, the weighted least squares estimation is also based on the

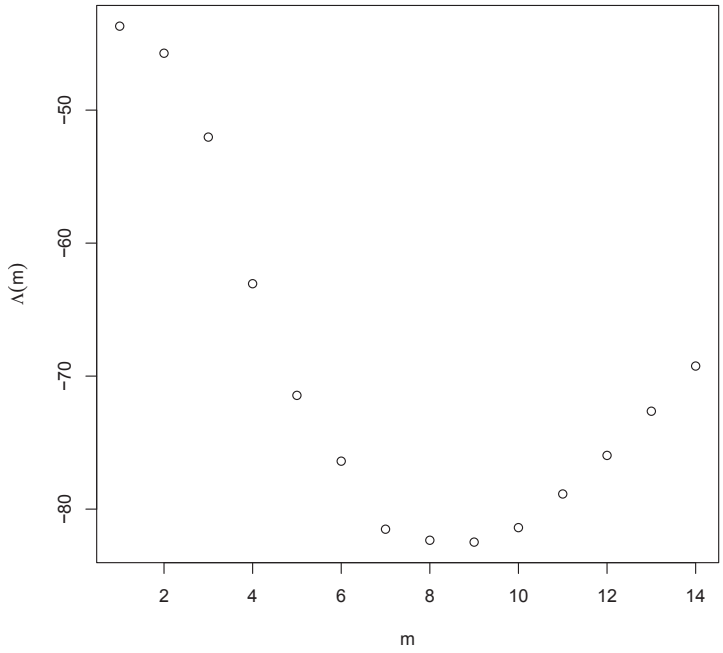


Figure 2.2: Model selection with BIC criterion. The vertical axis is the quantity $\Lambda(m)$ defined in (2.20), and the horizontal axis is the order of the Legendre polynomials. For this plot, an order of 10 is selected.

summation of weighted sum of squares for all 100 simulation replicates.

The estimated IPE values based on (2.23) are given in Table 2.1 and 2.2 for two prediction locations. As can be seen from the tables, although the prediction performance of our CM+PE method is worse than the maximum likelihood or the least squares method for small sample size, it performs roughly equally well as the sample becomes denser and denser. On the other hand, the LPE method which fits the empirical variograms directly does a much worse job for all three sampling configurations. The inclusion of tapered local Matérn covariance function greatly enhances the kriging performance of the nonparametric approach. This improvement is mainly due to the data-driven estimation of the local smoothness which plays an prominent role in interpolations.

Fixed rank kriging (FRK) method proposed in [Cressie and Johannesson (2008)] can also be used to model the covariance structure of global data. The covariance matrix

Table 2.1: Estimated IPE (defined in (2.23)) values evaluated at the first kriging location. LPE: nonparametric Legendre polynomial expansion; CM+PE: our model including a compactly supported Matérn covariance function and a polynomial expansion; MaMLE: full maximum likelihood method with parametric Matérn covariance structure; MaLS: parametric Matérn model using least squares.

	LPE	CM+PE	MaMLE	MaLS
$\text{IPE}(s)_{\text{g1}}$	1.518	0.121	0.00152	0.00217
$\text{IPE}(s)_{\text{g2}}$	1.100	0.0719	NA	0.00089
$\text{IPE}(s)_{\text{g3}}$	0.071	3.58×10^{-5}	NA	4.09×10^{-5}

Table 2.2: IPE values evaluated at the second kriging location.

	LPE	CM+PE	MaMLE	MaLS
$\text{IPE}(s)_{\text{g1}}$	1.887	1.582	0.00166	0.00227
$\text{IPE}(s)_{\text{g2}}$	1.346	0.246	NA	0.00065
$\text{IPE}(s)_{\text{g3}}$	0.089	2.84×10^{-5}	NA	3.57×10^{-5}

can be reparameterized as a quadratic form of the base vector and a positive definite matrix of fixed rank plus a nugget effect. Accordingly, a closed form of the inverse of the covariance matrix can be given which only involves taking inverses of much smaller matrices with fixed rank. FRK can be utilized in both stationary and nonstationary spatial processes in $\mathbb{R}^2$ and S^2. As an application, authors applied the method to the global TOMS data but did not have a simulation study [Cressie and Johannesson (2008)].

We did a simulation study to compare the performance of our approach with FRK. The main focus is again the prediction performance at different kriging sites. The comparison is carried on for the configuration "g2" in Figure 2.1 (b). We choose 196 (14×14) bins which are equally spaced in terms of latitude and longitude. Those bins partitions the study region. We also choose 81 (9×9) basis locations which are also equally spaced on the sphere. The basis function in the model is the local bisquare function used [Cressie and Johannesson (2008)]. Here two prediction sites are the same as described in the previous paragraphs. For each of the 100 simulated realizations, the best fitted model in (2.21) and the best FRK model are used for prediction. Table 2.3 shows the summary statistics of the squared prediction error $(\hat{Z}_i(x) - \hat{Z}_0(x))^2$ at one of the prediction lo-

cations. It can be seen that the CM+PE model can significantly reduce the squared prediction error compared to FRK method as well.

Table 2.3: Summary statistics of the squared prediction error for CM+PE and FRK based on 100 simulations. The prediction location is the one in the middle of the grid.

CM+PE		FRK	
Median	IQR	Median	IQR
0.237	0.744	0.504	1.519

2.6 Real Data Analysis

As an application, we apply our estimation method to total column ozone level data on a global scale, which have been widely used as an example for global covariance modeling [Cressie and Johannesson (2008); Jun and Stein (2008); Stein (2008)]. We will only present a short introduction to the data here and refer interested readers to the cited papers for details.

Total Ozone Mapping Spectrometer (TOMS) is a satellite instrument for measuring ozone values on a global scale. During the period of November 1978 to December 1994, several TOMS instruments were carried by NASA-satellites into the outer space including Nimbus-7 and Meteor-3 and provided global measurements of total column ozone on a daily basis. The data are either Level 2 or Level 3 versions. Level 2 data give spatial and temporal irregular measurements following the satellite scanning tracks [Jun and Stein (2008)]. Since the instrument relies on backscattered light, there are a lot of missing observations in Level 2 data. On the other hand, Level 3 data are post-processed from Level 2 data. They are obtained by averaging Level 2 data pixel by pixel and are on a spatially regular lattice with spacing 1° in latitude and 1.25° in longitude. The number of missing values is thus greatly reduced.

The data set used in this analysis is the original TOMS Level 3 data for May 15, 1990, which is the same as [Jun and Stein (2008)]. There are 288 longitude points and

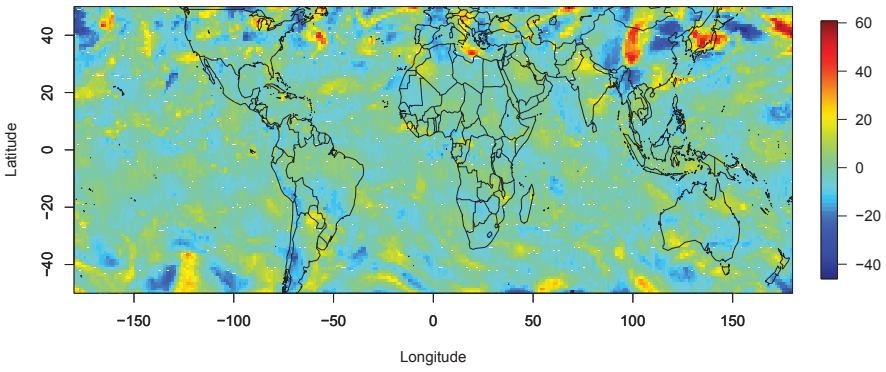
100 latitude points. The longitude points are evenly spaced over $-180°$ to $180°$ and the latitude points are evenly spaced over $50°$ South to $50°$ North. 71 out of the total 28,800 observations have missing values. We follow the same strategy in [Jun and Stein (2008)] and simply impute these missing values with the average of their 8 nearest neighboring sites. Different imputation methods can certainly be used. However, because of the small fraction of missing values, the choice of the imputing methods would not affect the results significantly.

The statistical model describing the data is assumed to be

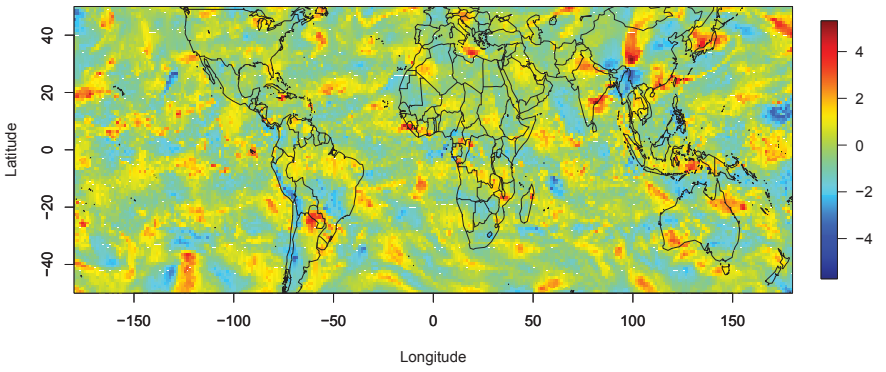
$$Y(s) = \mu(s) + \sigma(L)R(s)\,, \tag{2.24}$$

where L is the latitude of location s. $\mu(s)$ is the mean structure and $R(s)$ is an isotropic random field on Earth (possibly with measurement errors), and $\sigma^2(L)$ is the variance parameter as a function of latitude L. This particular model form will be explained in detail in subsequent context. First, it is observed that there is a mean structure in the data. The observed values range from 227 to 432 with noticeable bands along different latitudes. See Figure 2 in [Jun and Stein (2008)]. It turns out that subtracting the monthly average of May from the data does not remove the mean structure in a satisfactory way. There are hot spots at the high and low latitudes. Since spherical harmonics provide a natural basis for functions on the sphere, we regress the ozone level on $\{Y_n^m(\sin L, l)|n = 0, 1, \ldots, 12, m = -n, \ldots, n\}$ which better captures the large-scale mean structure in the data [Jun and Stein (2008)]. Afterwards the estimated average $\hat{\mu}(s)$ is subtracted from the data to get the residuals shown in Figure 2.3 (a). It can be seen that the spherical harmonics do a good job of modeling the mean structure and the residuals do not have any noticeable patterns such as hot spots.

Since our proposed model (2.21) works for isotropic random fields on the sphere, we have to make sure the data have this property. For atmospheric processes on the global scale, the assumption of isotropy is often too strong since the rotation of the Earth about its axis affect the processes differently at different latitudes. However, the pattern



(a)



(b)

Figure 2.3: (a) Residuals of TOMS column ozone level on May 15, 1990, after subtracting the mean structure estimated from the regression on spherical harmonics. (b) The approximately stationary process obtained by scaling (a) with the estimated Matérn variance at each latitude.

of *axial symmetry* [Jones (1962)] can often be assumed to be approximately true. For an axially symmetric process, the first two moments are invariant to rotations with respect to the Earth's axis. Authors in [Stein (2007); Jun and Stein (2008); Huang *et al* (2012)] discusses how to model an axially symmetric process with differential operators.

For an axially symmetric process, the process at each latitude is actually stationary. This motivates us to manipulate the obtained anisotropic residuals and transform them to an approximately isotropic one. At each latitude, we fit a Matérn covariance function

based on likelihood and get the parameter estimates. Because the ozone observations are on a regular lattice, the covariance matrix at each latitude is a circulant matrix and it can be diagonalized through Discrete Fourier Transform (DFT) [Davis (1979)]. In this application, the small number of observations at each latitude makes it feasible to take the inverse of the covariance matrix directly or with the aid of eigen-decomposition (Chapter 4.5 in [Brockwell and Davis (1991)]) when optimizing the likelihood function. The obtained Matérn parameter estimates are plotted as a function of the latitude in Figure 2.4. The most striking feature is the big variation in the variance σ^2, while other estimates stay roughly constants. This feature justifies the operation of scaling the residuals by the estimated variance at each latitude as shown in (2.24). In other words, we take

$$R(s) = (Y(s) - \mu(s))/\sigma(L)\,, \tag{2.25}$$

as an isotropic random process on the sphere, where $\sigma^2(L)$ is the covariance at latitude L. The "data" used in this analysis is $\hat{R}(s) = (Y(s) - \hat{\mu}(s))/\hat{\sigma}(L)$, where $\hat{\mu}(s)$ is the estimated trend obtained from regressing on spherical harmonics and $\hat{\sigma}^2(L)$ is the estimated Matérn covariance at latitude L. Figure 2.3 (b) shows the projected map of $\hat{R}(s)$ which does not show any obvious departure from an isotropic assumption.

The analysis is actually based on a subset of the total 28800 observations with two reasons. First, we want to assess the prediction performance of our proposed method. The subset serves as the training data for model fitting and the remaining observations are for model evaluation. Second, the total sample size of 28800 is not easy to handle to evaluate the inverse of the estimated covariance matrix for kriging purposes. It is possible, however, to utilize its special property of block circulant and reduce the computational burden [De Mazancourt and Gerlic (1983)]. In this study, the training data set is chosen so that the analysis can be done in a feasible time window with the available computing powers at hand. It is composed of every other point along the latitude and every other two points along the longitude from the original data set. We also include four densely spaced

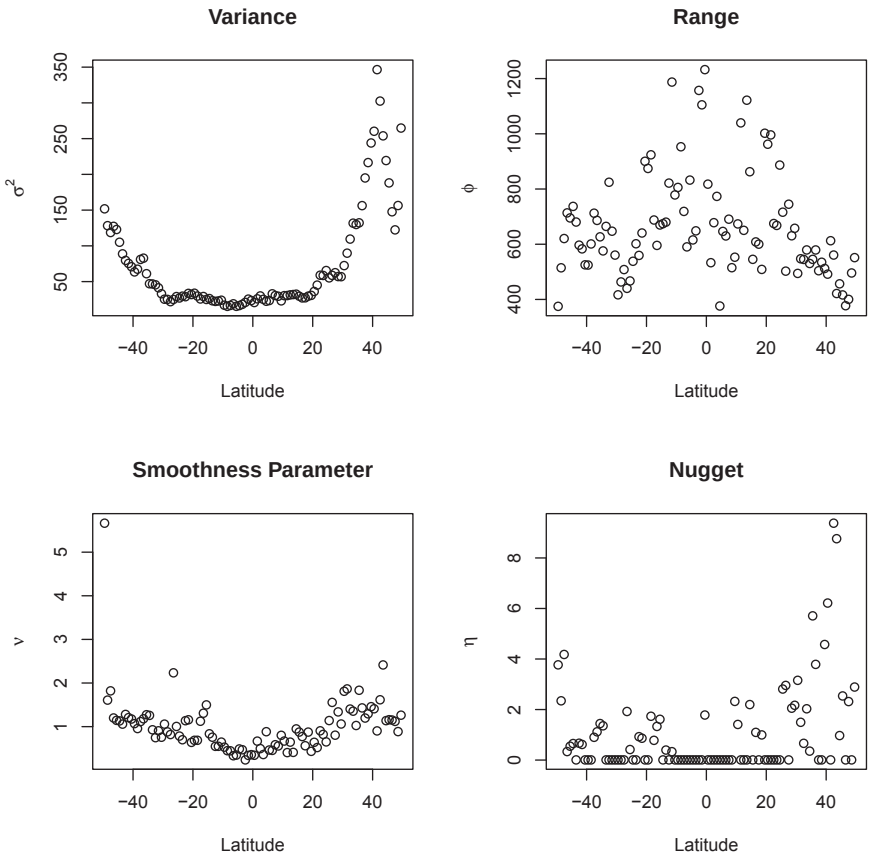


Figure 2.4: Matérn parameter estimates at different latitudes. There is a huge variation in variance σ^2 while smoothness and range parameters do not vary much.

patches for the sake of estimating the compactly supported Matérn covariance structure. These four dense patches are selected randomly. Figure 2.5 shows the detailed locations. Each of these patches includes 160 observations and the total number of observations is 5330. This data set serves as the "actual" data set from an isotropic random field on the sphere which we will apply our estimating method on.

Table 2.4: Averaged parameter estimate for the tapered Matérn variogram.

	σ^2	α	ν	η
Local parameter estimate	0.712	0.185	0.602	0.001

In the first step, we estimated the local Matérn covariance function from the four dense regions. The compact range is set to be 0.15 which is similar to the spherical size

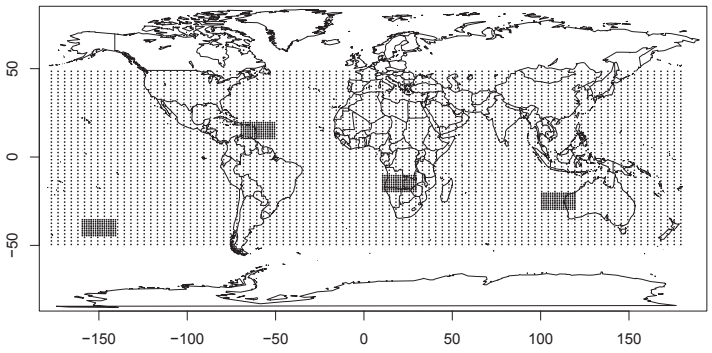


Figure 2.5: Observation locations used for model fitting, which consists of a regular lattice and four randomly selected small dense patches.

of each patch. Maximum likelihood method is applied to each of them and the estimated values are quite close among each other. Alternatively, we can apply maximum likelihood method to four patches simultaneously and get the parameter estimates as shown in Table 2.4. It can be seen that smoothness parameter is close to 0.5 which means that at small distances the covariance structure is approximately exponential. The nugget effect is very close to zero which is confirmed from estimations with other randomly selected patches (not shown here). The spatial range parameter α is higher than the values estimated in [Jun and Stein (2008)]. Our data were first scaled in (2.25) while [Jun and Stein (2008)] worked on it directly. So our parameter estimates are not directly comparable to theirs. The nonparametric Legendre polynomials fitting procedure is carried out after subtracting the contribution from the local Matérn variogram. Model selection based on BIC gives us a fit with an order of 15 Legendre polynomial terms. The final fitted variogram is given in Figure 2.6 where circles are empirical variogram estimates from robust estimator (2.14) and the solid line is the fitted curve. It can be seen that our method captures both the short-range smoothness behavior and the long-range waving structure quite well. Additionally, the procedure ensures that the curve is a valid variogram satisfying the conditionally non-negative conditions, thus alleviating

the obvious difficulty in a nonparametric smoothing method of the empirical variogram.

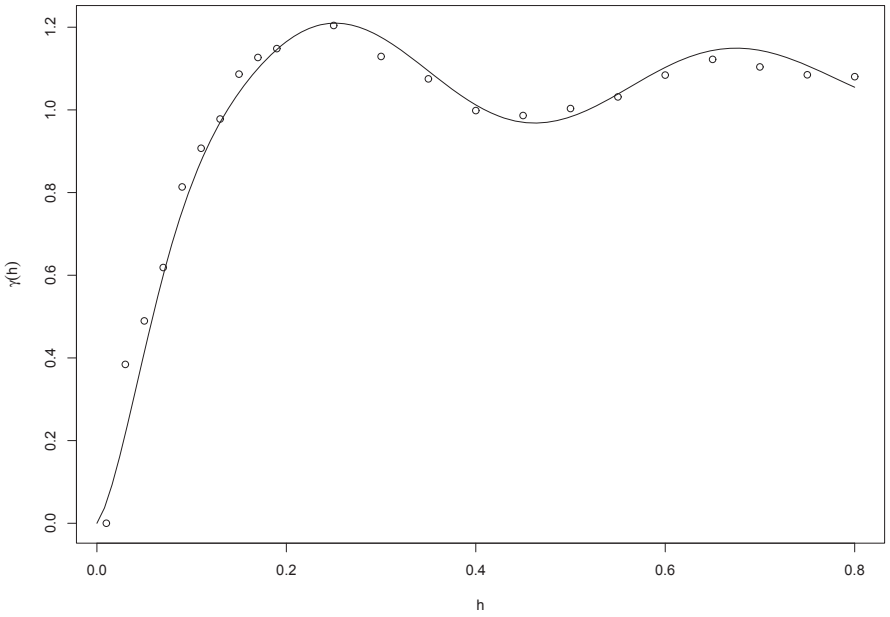


Figure 2.6: Variogram of the data $\hat{R}(s)$ and the corresponding CM+PE fit. Circles are empirical variogram estimate using robust estimator (2.14). Solid line is the CM+PE fit. Both small and large distance structures are fitted very well by this method.

To compare the kriging performance, we first randomly select 200 test locations which are left out of the training data set. Let $R(s_i)$ be the observed (true) value at location s_i and $\hat{R}(s_i)$ be the predicted value and $d(s_i)$ denotes the squared difference $d(s_i) = (R(s_i) - \hat{R}(s_i))^2$. Then we compute the median and interquartile range of $d(s_i)$ as shown in Table 2.5. It can be seen that the CM+PE method greatly outperforms the purely nonparametric method. The mean squared prediction error for two methods are 0.235 and 0.487, respectively. Introducing a tapered Matérn covariance function reduces the MSPE by 52%. As a comparison, we also try kriging only with the tapered Matérn covariance function. Its performance is quite close to CM+PE. This is not surprising since the local behaviors of the covariance structure contribute most for interpolation [Fuentes (2001)]. However, the inclusion of a nonparametric expansion allows us to

capture the large lag dependence which will become critical if we want to simulate from the estimated covariance structure and to impute data using spatial interpolation when there are large patches of missing values as was in the level 2 data.

Table 2.5: Summary statistics of the squared difference $d(s_i)$ between the true values and the predicted values at location s_i. 200 sites are randomly selected for prediction.

LPE		CM+PE	
Median	IQR	Median	IQR
0.170	0.441	0.059	0.227

2.7 Discussions

In this paper, we introduce a hybrid method for estimating the covariance function/variogram of an isotropic random field on spheres. The method is based on Legendre polynomial expansion plus a compactly supported Matérn covariance function for local behavior. The motivation for adding a Matérn component is to avoid the excessive smoothness of the purely nonparametric model. Our approach has the advantage of not requiring a subjective parametric form and allow the flexibility of estimating the local smoothness. Moreover, it also requires less computation. The empirical variogram estimate poses no special difficulties for estimation and the tapering range of the local Matérn covariance function can be chosen to maintain a reasonable computational burden when evaluating its full likelihood.

The tapering range is chosen in an *ad hoc* way in this study. We fix it to be comparable to the range of the densely sampled patches and assume the small scale component of the observations separated farther than this range do not correlate with each other. We can certainly regard it as an extra parameter and estimate it from the data. The size of the patches, on the other hand, can be chosen to be close to the range parameter from fitting an empirical variogram (for example, a Matérn variogram) of the whole data set. Issues related to maximum likelihood estimation are reported including slow convergence and

uncertainty about if a global maximum or even a local maximum is reached [Jun and Stein (2008); Stein (2008)]. The main reason is the large number of parameters in their models. We did not find it a big concern in our approach. By restricting the likelihood on small patches and using Matérn covariance function, we do not have a large volume of parameter space. In most cases, we obtained the same estimation with different starting points.

Our method requires the random field to be stationary and isotropic. It is known that most random fields observed on Earth are nonstationary. Some of them can be transformed (for example, by detrending and scaling) to an approximately homogeneous one, such as the TOMS data used in this paper. For a general *axially symmetric* random field, the form of Legendre polynomial expansion has many more parameters and is more difficult to fit. We plan to address this issue in a separate paper.

CHAPTER 3. MODELING NONSTATIONARY COVARIANCE FUNCTION WITH CONVOLUTION ON SPHERE

We address the issue of modeling axially symmetric spatial random fields on sphere with a kernel convolution approach. The observed random field is generated by convolving a latent uncorrelated random field with a class of Matérn type kernel functions. By allowing the parameters in the kernel functions to vary with locations, we are able to generate a flexible class of covariance functions and capture the nonstationary properties. Since the corresponding covariance functions do not generally have a closed-form expression, numerical evaluations are necessary and a pre-computation table is used to speed up the computation. For regular grid data on sphere, the block circulant property of the covariance matrix enables us to use Fast Fourier Transform (FFT) to get the determinant and the inverse efficiently. The methodology is applied to the Total Ozone Mapping Spectrometer ozone data for illustration.

3.1 Introduction

The need to model a large-scale spatial data has been increasing in the past decades. Due to the wide use of high-tech instruments and accumulation of observed data over time, it is not uncommon to have large data sets which have nonstationary dependence structure, especially for global data. As an example, the Level 3 TOMS (Total Ozone Mapping Spectrometer) data, a satellite measurement on the global ozone level, have more than 10^4 daily observations and the spatial structure is far from being stationary

[Cressie and Johannesson (2008); Jun and Stein (2008); Stein (2008)].

Statisticians have recognized the necessity to model nonstationary spatial random processes and have proposed different methodologies on this topic. Haas (1990) used a moving window approach to model acid deposition, where only the data in a local window were used in both estimation and prediction. Sampson and Guttorp (1992) used a smooth deformation of the spatial space, which is equivalent to a nonlinear transformation to generate nonstationarity. Fuentes (2001) proposed a method where the random field is represented locally as stationary and isotropic, but allowing the parameters to vary across space. Paciorek and Schervish (2006) introduced a new class of nonstationary covariance functions with closed forms.

Another approach that can easily implement nonstationarity is the process convolution approach introduced by [Higdon *et al* (1999)]. In this approach, let M be a random measure defined on $\mathbb{R}^d$ such that $E(M(A)) = 0$ and $E|M(A)|^2 = F(A)$ for some positive finite measure F for any measurable set $A \subset \mathbb{R}^d$. For any two disjoint measurable set A and B, we have $E(M(A)\bar{M}(B)) = 0$. The random process $Z(s)$ can be defined as a kernel convolution of the underlying excitation field as

$$Z(s) = \int_{\mathbb{R}^d} K(s-u;\xi_s)M(du)\,, \tag{3.1}$$

where $K(s;\xi_s)$ is a nonrandom, square-integrable kernel function with ξ_s being the parameters at location s. It is easy to see that $Z(s)$ has a constant mean zero and its covariance function $C(u,v) \equiv \mathrm{Cov}(Z(u), Z(v))$ is

$$C(u,v) = \int_{\mathbb{R}^d} K(u-w;\xi_u)K(v-w;\xi_v)F(dw)\,. \tag{3.2}$$

The convolution approach can be generalized to model a nonstationary process. By allowing ξ_s to vary at different locations, it is possible to generate a nonstationary field on $\mathbb{R}^d$. Convolution-based methods have the appealing features in nonparametric modeling since one only need to model the smoothing kernel $K(t)$ instead of the covariance function which is restricted to be non-negative definite. It is also not difficult to augment the space

with time so that the kernel function and the excitation field are both spatio-temporally related.

The choice of the kernel function is important in the modeling since it controls the properties of the resulting covariance structure, including the range, variance, and smoothness. An intuitive choice would be the Gaussian kernel [Higdon *et al* (1999)]. It has the advantage of being evaluated analytically since the covariance function also has a Gaussian form. However, as described in [Stein (1999)], the Gaussian covariance function is infinitely differentiable which may not be realistic for physical processes. It is known that Matérn covariance function has a smoothness parameter which can be estimated from the data. It has been shown that if the kernel function is chosen to be a modified Bessel function as [Zhu and Wu (2010); Xia and Gelfand (2006)]

$$K(x;\theta) = \frac{2\Gamma(\nu+d/2)^{1/2}\nu^{\nu/4+d/8}\sigma^{1/2}|x|^{\nu/2-d/4}}{\pi^{d/4}\Gamma(\nu/2+d/4)\Gamma(\nu)^{1/2}\rho^{\nu/2+d/4}}\mathcal{K}_{\nu/2-d/4}\left(\frac{2\nu^{1/2}|x|}{\rho}\right), \tag{3.3}$$

the corresponding covariance function takes the familiar Matérn form

$$C_{\sigma,\rho,\nu}(u) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}u}{\rho}\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{2\nu^{1/2}u}{\rho}\right), \tag{3.4}$$

where $\mathcal{K}_\nu$ is the modified Bessel function of order ν [Abramowitz and Stegun (1965)].

All the methods mentioned up to now are for Euclidean space $\mathbb{R}^d$ where the covariance structure is based on the Euclidean distance. Recently, however, more and more large-scale data in climatology and environtology are collected where the curvature of the Earth cannot be simply neglected. The aforementioned Level 3 TOMS data are observed globally along satellite tracks. It is apparently not appropriate to use Euclidean distance if the two locations are far apart from each other.

Some analysis methods designed specifically to handle global data, such as the TOMS data, are already available. Cressie and Johannesson (2008) expressed the covariance matrix in terms of a diagonal matrix plus a fixed low rank matrix, which makes it possible to compute the likelihood function exactly with massive spatial data. Stein (2008) further

replaced the diagonal matrix with a sparse matrix hoping to capture both the small-scale and large-scale spatial dependence structures. Jun and Stein (2007) proposed an approach to producing space-time covariance functions on sphere by applying differential operators to fully symmetric processes. In this way, nonstationary spatial random fields can be produced with a closed form on sphere and time. Jun and Stein (2008) applied this method to the analysis of TOMS data in an axially symmetric modeling framework. With the aid of Discrete Fourier Transform (DFT), they were able to calculate the exact likelihood for large data sets on regular grids.

Spatial random fields observed within a local region can often be approximated to be stationary or isotropic. On the other hand, large scale or global processes usually show the pattern of nonstationarity since the factors driving the characteristics of the random field typically vary at different locations. For example, the varying temperatures at high and low latitude have different influence on climatological properties. A special kind of nonstationarity is the axial symmetry as described in [Jones (1962)]. For an axially symmetric process, the first two moments are invariant to rotations with respect to the Earth's axis. Their covariance function depends on longitude only through their difference. Stein (2007); Jun and Stein (2007, 2008) applied this approach to model the ozone data described above on a global scale, where they consider the axially symmetric process by applying differential operators to an isotropic process. If we can further assume longitudinal reversibility [Stein (2007); Huang *et al* (2012)] so that the covariance function between two locations (L_1, l_1) and (L_2, l_2) satisfies

$$C(L_1, L_2, l_1 - l_2) = C(L_1, L_2, l_2 - l_1)\,, \tag{3.5}$$

we can consider the process at each latitude as an isotropic process. The parameters of the random field at each latitude are homogeneous.

Some of the methodologies described above require a large number of parameters and pose a challenge on computing. In this article, we present a more intuitive approach to handle a nonstationary spatial process on a sphere without intense computational

burden. The spatial random field is modeled as a kernel convolution of a latent uncorrelated random field. The kernel function is chosen similar to the Matérn covariance function on $\mathbb{R}^d$ which has an extra parameter controlling the smoothness. The random field is first modeled at each latitude and the raw estimates are smoothed with a local linear smoothing method across different latitudes. A precomputation table is used to approximate the covariance structure. The computational issue of getting the inverse of a large covariance matrix can be alleviated for regularly spaced observations on sphere. This paper is organized as follows. In Section 3.2, we briefly introduce the covariance structure on sphere and the concept of axial symmetry of nonstationary process. A kernel convolution procedure which is analog to the Euclidean space is proposed and its properties are also presented. Section 3.3 discusses the methodology to model an axially symmetric random field on sphere, where raw estimates are obtained at each latitude and are smoothed across latitudes. A procedure of using a pre-computation table is introduced to overcome the computational difficulty. For regular grids on sphere, we present a computation-friendly way of calculating the inverse and determinant of the covariance matrix of a large data set. In Section 3.5, a simulation study is carried out to illustrate the proposed methodology and a real data analysis using the TOMS data is presented in Section 3.6. At last we conclude and discuss future work directions in Section 3.7.

3.2 Covariance Structure on Spheres

Suppose $Z(\cdot)$ is a random field on sphere $\mathbb{S}^2$ with radius R. The sphere is usually the Earth in many environmental and geophysical applications. It is customary to specify a location $s \in \mathbb{S}^2$ by its latitude L and longitude l, where $-\pi/2 \leq L \leq \pi/2$ and $-\pi \leq l < \pi$. Random field $Z(\cdot)$ is called isotropic (sometimes referred to as homogeneous) if its first two moments are invariant under any rotations of the sphere. In other words, $E(Z(s)) = \mu$ is a constant for any $s \in \mathbb{S}^2$ and its covariance function $\mathrm{Cov}(Z(s_1), Z(s_2))$

depends only on the spherical angle $\theta(s_1, s_2)$ between the two locations, where

$$\theta(s_1, s_2) = \arccos(\sin L_1 \sin L_2 + \cos L_1 \cos L_2 \cos(l_1 - l_2)) .$$

It is equivalent to write the covariance function as a function of the great circle distance ($\text{gc}(s_1, s_2) = R\theta(s_1, s_2)$) or chordal distance ($\text{cd}(s_1, s_2) = 2R\sin(\theta(s_1, s_2)/2)$) between the pair of locations s_1 and s_2 since they have one-to-one correspondence if the radius is fixed.

Similar to the case in $\mathbb{R}^d$, a real continuous function $C(\cdot)$ is said to be a valid covariance function on $\mathbb{S}^2$ if and only if it is non-negative definite, that is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\theta(s_i, s_j)) \geq 0 \tag{3.6}$$

for any integer n, any vector $\mathbf{a} = (a_1, \ldots, a_n)$ and any locations $\{s_1, \ldots, s_n\} \in \mathbb{S}^2$. For intrinsically stationary processes, a valid variogram has to be conditionally negative definite so that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\theta(s_i, s_j)) \leq 0 \tag{3.7}$$

for any integer n, any locations $(s_1, \ldots, s_n) \in \mathbb{S}^2$, and any vector $\mathbf{a} = (a_1, \ldots, a_n)$ satisfying $\sum_{k=1}^{n} a_k = 0$.

Huang *et al* (2011) showed that many valid covariance functions in $\mathbb{R}^d$ are no longer valid in $\mathbb{S}^2$, including Gaussian and some Matérn models. Gneiting (2012) further proved that a Matérn covariance function is valid on sphere if and only if its smoothness parameter is no greater than 1/2. Generally, valid covariance functions on sphere are obtained by restricting covariance functions in $\mathbb{R}^3$ on sphere. That is, if a function $C_0(h)$ is a valid covariance function in $\mathbb{R}^3$, a new function defined as $C(\theta) = C_0(2\sin(\theta/2))$ is a valid covariance function on the unit sphere.

3.2.1 Kernel convolution on spheres

A spatial process $Z(s)$ on sphere can be constructed as a kernel convolution as

$$Z(s) = \int_{u \in \mathbb{S}^2} k(u - s|\eta_s) X(u) du \,, \tag{3.8}$$

where η_s are the parameters of kernel k which may depend on location s due to nonstationarity. $X(u)$ is an infinitely dense Gaussian white noise process at $u \in \mathbb{S}^2$ (continuous white noise process) with the properties of

$$\begin{aligned} EX(u) &= 0 \,, \\ \int_{\mathcal{A} \subset \mathbb{S}^2} X(u) du &\sim N(0, \sigma_\omega^2 \times \text{Area}(\mathcal{A})) \,. \end{aligned} \tag{3.9}$$

It is easy to verify that $EZ(s) = 0$ and

$$\text{Cov}(Z(s_1), Z(s_2)) = \sigma_\omega^2 \int_{u \in \mathbb{S}^2} k(u - s_1|\eta_{s_1}) k(u - s_2|\eta_{s_2}) du \,. \tag{3.10}$$

Such a covariance function is positive definite since

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j C(s_i, s_j) &= \sigma_\omega^2 \int_{u \in \mathbb{S}^2} k(u|\eta_{s_1}) k(u|\eta_{s_2}) du \\ &= \sigma_\omega^2 \int_{u \in \mathbb{S}^2} \left(\sum_{i=1}^{n} a_i k(u|\eta_{s_i}) \right)^2 du \geq 0 \,. \end{aligned} \tag{3.11}$$

Inspired by the convolution formulas in [Zhu and Wu (2010)], we propose a kernel function with the form of

$$f(\mathbf{x}; \boldsymbol{\mu}, \rho, \nu) = \sigma \alpha(\rho, \nu) \left[\sqrt{2\rho\nu} d(\mathbf{x}, \boldsymbol{\mu}) \right]^{\nu} K_\nu(\sqrt{2\rho\nu} d(\mathbf{x}, \boldsymbol{\mu})) \,, \tag{3.12}$$

where K_ν is the modified Bessel function of the second kind with order ν; ρ is the concentration factor which is related to the range of the resulting random field, and ν is the smoothness parameter. The distance between two locations $\mathbf{x}$ and $\boldsymbol{\mu}$ is defined as $d(\mathbf{x}, \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\| = \sqrt{2(1 - \mathbf{x} \cdot \boldsymbol{\mu})} = \sqrt{2[1 - \cos\theta(\mathbf{x}, \boldsymbol{\mu})]}$. If we let θ denote the spherical angle between $\mathbf{x}$ and $\boldsymbol{\mu}$, $d(\mathbf{x}, \boldsymbol{\mu}) = 2\sin(\theta/2)$ which can be approximated by θ if θ is small.

The coefficient $\alpha(\rho,\nu)$ is the normalization factor such that $\int_{\mathbb{S}^2} f(\mathbf{x};\boldsymbol{\mu},\rho,\nu)=1$ which can be computed from the condition that

$$\int_0^{2\pi}\int_0^{\pi}\sigma\alpha(\rho,\nu)\left[2\sqrt{2\rho\nu}\sin(\theta/2)\right]^{\nu}K_{\nu}(2\sqrt{2\rho\nu}\sin(\theta/2))\sin\theta d\theta d\varphi=1\,, \tag{3.13}$$

which has a closed-form solution

$$\alpha(\rho,\nu)=\frac{1}{2\pi\sigma}\frac{\rho\nu}{2^{\nu-1}\Gamma(\nu+1)-2^{\frac{3\nu-1}{2}}(\rho\nu)^{\frac{\nu+1}{2}}K_{\nu+1}(2\sqrt{2\rho\nu})}\,. \tag{3.14}$$

The resulting covariance function is obtained as

$$\begin{aligned}C(Z(s_1),Z(s_2)) &= \sigma_1\sigma_2\int_{\mathbf{u}\in\mathbb{S}^2}f(\mathbf{u}|\eta_{s_1})f(\mathbf{u}|\eta_{s_2})d\mathbf{u}\\ &= \sigma_1\sigma_2\alpha(\nu_1,\rho_1)\alpha(\nu_2,\rho_2)\int_{u\in\mathbb{S}^2}[\sqrt{2\rho_1\nu_1}d(\mathbf{u},\mathbf{s}_1)]^{\nu_1}[\sqrt{2\rho_2\nu_2}d(\mathbf{u},\mathbf{s}_2)]^{\nu_2}\\ &\times\ K_{\nu_1}(\sqrt{2\rho_1\nu_1}d(\mathbf{u},\mathbf{s}_1))K_{\nu_2}(\sqrt{2\rho_2\nu_2}d(\mathbf{u},\mathbf{s}_2))d\mathbf{u}\,.\end{aligned} \tag{3.15}$$

For a homogeneous random field on the sphere where ρ, σ^2, and ν are constant, we have

$$\begin{aligned}C(Z(s_1),Z(s_2)) &= \sigma^2\alpha^2(\nu,\rho)(2\rho\nu)^{\nu}\int_{\mathbf{u}\in\mathbb{S}^2}[d(\mathbf{u},\mathbf{s}_1)d(\mathbf{u},\mathbf{s}_2)]^{\nu}K_{\nu}(\sqrt{2\rho\nu}d(\mathbf{u},\mathbf{s}_1))\\ &\times\ K_{\nu}(\sqrt{2\rho\nu}d(\mathbf{u},\mathbf{s}_2))d\mathbf{u}\,,\end{aligned} \tag{3.16}$$

which intuitively should only depend on the spherical angle $\theta(\mathbf{s}_1,\mathbf{s}_2)$, or equivalently the great arc distance $d(\mathbf{s}_1,\mathbf{s}_2)$ between $\mathbf{s}_1$ and $\mathbf{s}_2$. This result is proved in theorem 3.2.1. Figures 3.1 and 3.2 show the correlation functions with different ρ and ν values. It is shown that ρ controls the concentration of the random field which has effect similar to the range parameter in $\mathbb{R}^d$. Large values of ρ will result in a correlation with small range. Note that in Figures 3.1 and 3.2, ν is the order of the Bessel function in the kernel function.

Theorem 3.2.1. *The random field $Z(\mathbf{s})$ generated by convolving a homogeneous kernel function on the sphere is isotropic. That is, $C(Z(\mathbf{s}_1),Z(\mathbf{s}_2))=C(\theta(\mathbf{s}_1,\mathbf{s}_2))$.*

Proof. Here we are proving for a general case, that is,

$$I(\mathbf{s}_1,\mathbf{s}_2)=\int_{\mathbb{S}}f(\mathbf{x}\cdot\mathbf{s}_1)f(\mathbf{x}\cdot\mathbf{s}_2)d\mathbf{x} \tag{3.17}$$

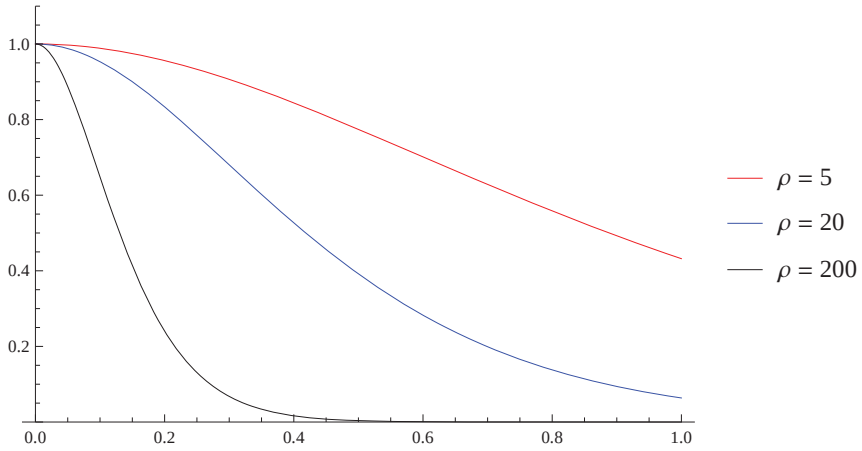


Figure 3.1: Correlation functions with different ρ and fixed $\nu = 1$.

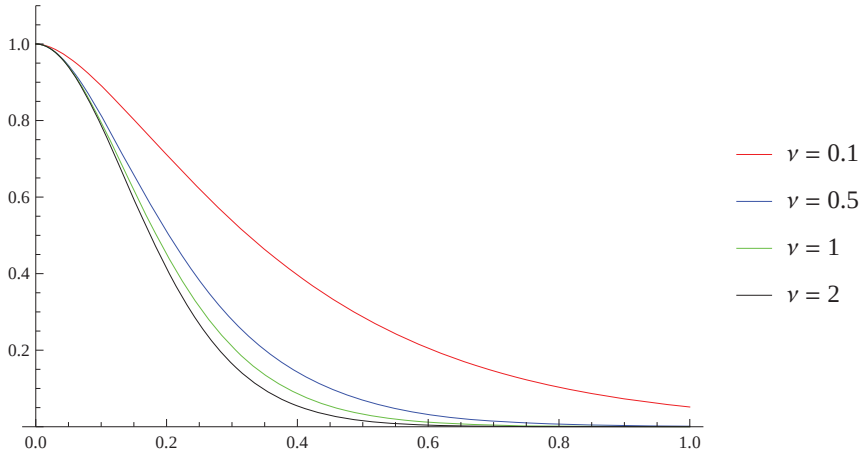


Figure 3.2: Correlation functions with different ν and fixed $\rho = 100$.

depends only on the relative angle between $\mathbf{s}_1$ and $\mathbf{s}_2$, or $I(\mathbf{s}_1, \mathbf{s}_2)$ is a function of $\mathbf{s}_1 \cdot \mathbf{s}_2$.

Addition theorem in spherical harmonics states that

$$P_l(\mathbf{x} \cdot \mathbf{y}) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_{lm}^*(\theta', \varphi') Y_{lm}(\theta, \varphi) \,,$$

where $\mathbf{x}$ and $\mathbf{y}$ are two unit vectors with spherical coordinates (θ, φ) and (θ', φ'), respectively.

Due to the completeness of the Legendre polynomials, a function on the unit sphere can be written as

$$f(x) = \sum_{l=0}^{\infty} f_l P_l(x) \,,$$

where $f_l = \frac{2l+1}{2}\int_{-1}^{1} f(x)P_l(x)dx$. This is also called Fourier-Legendre series or generalized Fourier series.

Without loss of generality, we choose $\mathbf{s}_1$ to be along the z-axis and $\mathbf{s}_2$ in the xz plane, then

$$\begin{aligned} I(\mathbf{s}_1, \mathbf{s}_2) &= \int f(\cos\theta_{\mathbf{x}}) f(\mathbf{x}\cdot\mathbf{s}_2) d\Omega_{\mathbf{x}} \\ &= \sum_{l,l'=0}^{\infty} f_l f_{l'} \int P_l(\cos\theta_{\mathbf{x}}) P_{l'}(\mathbf{x}\cdot\mathbf{s}_2) d\Omega_{\mathbf{x}}\,. \end{aligned} \tag{3.18}$$

Using the fact that

$$\begin{aligned} P_l(\cos\theta_{\mathbf{x}}) &= P_l^0(\cos\theta_{\mathbf{x}}) = \sqrt{\frac{4\pi}{2l+1}} Y_{l0}(\theta_{\mathbf{x}}, 0)\,, \\ P_{l'}(\mathbf{x}\cdot\mathbf{s}_2) &= \frac{4\pi}{2l'+1} \sum_{m'=-l'}^{l'} Y^*_{l'm'}(\theta_{\mathbf{x}}, \varphi_{\mathbf{x}}) Y_{l'm'}(\theta_{\mathbf{s}_2}, 0)\,, \end{aligned}$$

we have

$$\begin{aligned} I(\mathbf{s}_1, \mathbf{s}_2) &= \sum_{l,l'=0}^{\infty} f_l f_{l'} \sqrt{\frac{4\pi}{2l+1}} \frac{4\pi}{2l'+1} Y_{l'm'}(\theta_{\mathbf{s}_2,0}) \int Y_{l0}(\theta_{\mathbf{x}}, 0) Y^*_{l'm'}(\theta_{\mathbf{x}}, \varphi_{\mathbf{x}}) d\Omega_{\mathbf{x}} \\ &= \sum_{l=0}^{\infty} f_l^2 \left(\frac{4\pi}{2l+1}\right)^{3/2} Y_{l0}(\theta_{\mathbf{s}_2}, 0) \\ &= \sum_{l=0}^{\infty} f_l^2 \frac{4\pi}{2l+1} P_l(\cos\theta_{\mathbf{s}_2}) \\ &= \sum_{l=0}^{\infty} f_l^2 \frac{4\pi}{2l+1} P_l(\mathbf{s}_1\cdot\mathbf{s}_2)\,, \end{aligned} \tag{3.19}$$

since $\int Y_{lm}(\theta_{\mathbf{x}}, 0) Y^*_{l'm'}(\theta_{\mathbf{x}}, \varphi_{\mathbf{x}}) d\Omega_{\mathbf{x}} = \frac{4\pi}{2l+1}\delta_{mm'}\delta_{ll'}$. □

3.2.2 Special cases

Similar to the Matérn covariance function in the Euclidean space, different choices of smoothness parameter in the kernel function will lead to covariance functions with different properties. In the Euclidean space, the Matérn covariance function approaches a Gaussian covariance function if the smoothness parameter goes to infinity. Similarly, the

Bessel kernel defined in (3.12) becomes the well-known von Mises-Fisher distribution on sphere if $\nu \to \infty$ [Heaton (2013)]. The von Mises-Fisher distribution has the probability density function

$$f_p(\mathbf{x}; \boldsymbol{\mu}, \rho) = C_p(\rho) \exp(\rho \boldsymbol{\mu}^T \mathbf{x}) \,, \tag{3.20}$$

where $\rho \geq 0$, $\|\boldsymbol{\mu}\| = 1$ and the normalization constant $C_p(\rho)$ is

$$C_p(\rho) = \frac{\rho^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\rho)} \,,$$

where I_v denotes the modified Bessel function of the first kind with order v. For three-dimensional case where $p = 3$, we have

$$C_3(\rho) = \frac{\rho}{2\pi(e^{\rho} - e^{-\rho})} \,.$$

The parameters $\boldsymbol{\mu}$ and ρ are called the mean direction and concentration parameter, respectively. The greater the value of ρ, the higher the concentration of the distribution around the mean direction $\boldsymbol{\mu}$. The distribution is unimodal for $\rho > 0$, and is uniform on the sphere for $\rho = 0$.

For the isotropic case where ρ is a constant for all locations on the sphere, we can compute the covariance function as

$$\begin{aligned} \mathrm{Cov}(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) &= \frac{\sigma_\omega^2 \rho^2}{4\pi^2 (e^{\rho} - e^{-\rho})^2} \int_{u \in \mathbb{S}^2} \exp\{\rho \mathbf{s}_1^T \cdot \mathbf{u}\} \exp\{\rho \mathbf{s}_2^T \cdot \mathbf{u}\} d\mathbf{u} \\ &= \frac{\sigma_\omega^2 k}{4\pi |\bar{\mathbf{s}}|} \frac{e^{2\rho|\bar{\mathbf{s}}|} - e^{-2\rho|\bar{\mathbf{s}}|}}{(e^{\rho} - e^{-\rho})^2} \,, \end{aligned} \tag{3.21}$$

where $\bar{\mathbf{s}} = (\mathbf{s}_1 + \mathbf{s}_2)/2$. The variance of the random field Z is computed when $\mathbf{s}_1 = \mathbf{s}_2$ which gives

$$\mathrm{Var}(Z(\mathbf{s})) = \frac{\sigma_\omega^2 \rho}{4\pi} \frac{e^{2\rho} - e^{-2\rho}}{(e^{\rho} - e^{-\rho})^2} \,.$$

Since $|\bar{\mathbf{s}}| = \cos(\theta/2)$ where θ is the spherical angle between $\mathbf{s}_1$ and $\mathbf{s}_2$, we can get

$$\mathrm{Cov}(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = \frac{\sigma_\omega^2 \rho}{4\pi \cos\frac{\theta}{2}} \frac{e^{2\rho\cos\frac{\theta}{2}} - e^{-2\rho\cos\frac{\theta}{2}}}{(e^{\rho} - e^{-\rho})^2} \,.$$

This covariance function is infinitely differentiable since $\cos(\theta/2)$ is an even function of θ and can be represented as $\sum_{i=0}^{\infty} \frac{(-1)^i}{2^i(2i)!}|\theta|^{2i}$ so the Taylor expansion only involves the even power of $|\theta|$. From Theorem 2 on page 29 of [Stein (1999)], $C(\theta)$ has infinitely many derivatives. Since von Mises-Fisher distribution is the analogue of the bivariate normal distribution on the two-dimensional unit sphere, this infinite smoothness is expected.

3.3 Estimation

If we assume longitudinal reversibility, the process at each latitude is actually an isotropic process. We assume that ν is a constant across the whole sphere. Additionally we assume that σ^2 and ρ are smooth functions of latitude solely so that they are homogeneous at each latitude. For a fixed value of ν, we first obtain raw estimates of σ^2 and ρ by fitting isotropic covariance functions (3.16) to observations at each latitude l. The next step is to construct the functions $\sigma^2(l)$ and $\rho(l)$ using local linear smoothing method. The bandwidth is chosen by cross-validation. This procedure is repeated for a range of ν and the value $\hat{\nu}$ which maximizes the log likelihood is chosen to be the estimated value of ν.

3.3.1 Raw estimates at each latitude

We assume σ^2 and ρ are smooth functions of latitude. At each latitude l, they are constants but are allowed to vary across different latitudes due to nonstationarity. At each latitude, we propose to use likelihood method to get the estimates of σ^2 and ρ which stand for the variance and concentration of the process. Let $\boldsymbol{\eta} = (\sigma^2, \rho)$ be the parameters in the covariance structure, the loglikelihood function is

$$l(\boldsymbol{\eta}) = -\frac{1}{2}\log|\det\Sigma_{\boldsymbol{\eta}}| - \frac{1}{2}(Z_l - \mu)^T\Sigma_{\boldsymbol{\eta}}^{-1}(Z_l - \mu)\,, \tag{3.22}$$

where Z_l are the observations at the current latitude l and the elements in $\Sigma_{\boldsymbol{\eta}}$ are the homogeneous covariance function in (3.16) between any pair of these data points. For

a regular grid on sphere, the observations are also equally spaced at each latitude. The covariance matrix is a circular matrix and (3.22) can be easily optimized on a one-dimensional circle using DFT [Jun and Stein (2008)]. If the observations are made at latitudes $\{l_1, \ldots, l_N\}$, we will obtain a list of raw estimates $\{(\sigma_i^2, \rho_i) : i = l_1, \ldots, l_N\}$.

3.3.2 Smoothing across latitudes

The raw estimates in Section 3.3.1 are only available at discrete latitudes. Spatial prediction, however, generally requires estimates at any possible latitudes. Thus we need to smooth out the raw estimates to get a continuous function of latitude. For each of the estimates, a nonparametric smoothing method can be applied to get the estimates at any latitude. The local linear smoothing estimator of ρ_l at any given latitude l is given by minimizing

$$\sum_{i=1}^{n} K\left(\frac{l_i - l}{h}\right)(\rho_i - a_0 - a_1(l_i - l))^2\,, \tag{3.23}$$

where $K(\cdot)$ is a univariate kernel function and h is the smoothing bandwidth [Wand and Jones (1994)]. The local linear smoothing estimator ρ_l is just $\hat{a}_0$. The bandwidth h can be selected through leave-one-out cross validation score

$$\mathrm{CV}(h) = \sum_{i=1}^{n}(\rho_i - \hat{\rho}_{(-i)}(h))^2\,, \tag{3.24}$$

where $\hat{\rho}_{(-i)}$ is the estimator obtained by omitting the i^{th} pair (l_i, ρ_i). The h value which minimizes $\mathrm{CV}(h)$ is chosen as the smoothing bandwidth to be used in (3.23). The same technique can be applied to σ^2 and a complete estimate of the parameter space can thus be obtained.

3.4 Spatial Prediction

Estimation is rarely the final goal of a spatial analysis. In practice, predictions at unobserved locations are often desired. Suppose the observed data are $Y(s_1), \ldots, Y(s_n)$

at spatial locations $s_1, \ldots, s_n$ on the sphere and we want to predict $Y(s_0)$ at a new location s_0 where no observation was made. Further suppose the model is

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{Z}(\mathbf{s})\,, \tag{3.25}$$

where $\mathbf{Z}(\mathbf{s}) \sim (\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is assumed to be known. Here $\mathbf{X}$ is an $n \times p$ matrix with the i^{th} row, $\mathbf{x}_i$, corresponding to the p-dimensional explanatory variables at location s_i. The data and the unobservables are spatially correlated as $\text{Cov}[\mathbf{Y}(\mathbf{s}), Y(s_0)] = \boldsymbol{\sigma}$ and $\text{Var}[Y(s_0)] = \sigma_0$. Here, $\boldsymbol{\Sigma}, \boldsymbol{\sigma}$, and σ_0 are all functions of the covariance parameters $\boldsymbol{\theta} = (\sigma^2, \rho, \nu)$. If $\boldsymbol{\theta}$ is known, the Best Linear Unbiased Predictor (BLUP) of $Y(s_0)$ has the form of

$$\hat{Y}_{UK}(s_0) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\sigma} - \mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} - \mathbf{x}_0))\mathbf{Y}\,. \tag{3.26}$$

The kriging variance is

$$\sigma^2_{UK}(s_0) = C(s_0, s_0) - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + (\mathbf{x}_0^T - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}_0^T - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^T\,. \tag{3.27}$$

These are the universal kriging formulas in geostatistics. A special case is the ordinary kriging where the random field $\mathbf{Y}$ is assumed to have a constant mean so that $\mathbf{X}\boldsymbol{\beta}$ is replaced by $\mu\mathbf{1}$. The resulting optimal predictor is

$$\hat{Y}_{OK}(s_0) = \left(\boldsymbol{\sigma} + \mathbf{1}\frac{1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}\right)^T \boldsymbol{\Sigma}^{-1}\mathbf{Y} \tag{3.28}$$

and the kriging variance is

$$\sigma^2_{OK} = C(s_0, s_0) - \boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma} + \frac{(1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma})^2}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}\,. \tag{3.29}$$

In practice, the covariance parameters $\boldsymbol{\theta}$ are generally unknown and has to be estimated from the data. After we have estimator $\hat{\boldsymbol{\theta}} = (\hat{\sigma}^2, \hat{\rho}, \hat{\nu})$, we can compute all elements in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\sigma}(\boldsymbol{\theta})$ as $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$, and compute the empirical BLUP (EBLUP) of $\hat{Y}_0$ instead. However, there are some technical issues associated with the procedure of kriging, most of which are related to the large size of the data set and thus the difficulty of inverting the covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.

3.4.1 Precomputation

In principle, we can compute all elements in $\boldsymbol{\Sigma}$ and $\boldsymbol{\sigma}$ once we know the estimates $\hat{\boldsymbol{\theta}}$ at any locations. However, direct implementation of this approach is very computationally intensive.

Suppose we have a convolution rule on the sphere

$$C(r, \rho_1, \rho_2) = \int_{\mathbb{S}^2} K(s; \rho_1) K(r + s; \rho_2) F(ds) \,, \tag{3.30}$$

where K is the kernel function. For an $n \times n$ covariance matrix, the integration (3.30) has to be numerically evaluated $n(n+1)/2$ times, which is not feasible for a large sample. Moreover, in the process of maximizing the likelihood, this integration need to be evaluated repeatedly. Instead, we can resort to the precomputation technique that is widely used in applied mathematics [Zhu and Wu (2010)]. First we compute (3.30) for a three-dimensional grid over the possible region of the parameters (r, ρ_1, ρ_2). The range of this lattice grid can be specified from the context of the problem at hand. In the spherical problem we considered, r is in the range of $[0, \pi R]$ where R is the radius of the sphere. Other parameter bounds can come from some pilot data analysis or be chosen in a convenient way. In the current study, we choose r in the range of $[0, \pi]$ with step size 0.05, and for parameter ρ we choose the range of $[100, 2000]$ with step size 50. The maximum approximation error is less than 0.01 when compared with numerical integrated values with precision of 10^{-5}.

Afterwards, covariance function between any two locations on sphere can be computed by the following weighted sum

$$C(r, \rho_1, \rho_2) = \sum_{ijk} C(r_i, \rho_{1j}, \rho_{2k}) f_i(r) f_j(\rho_1) f_k(\rho_2) \,, \tag{3.31}$$

where the summation is over all 2^3 neighboring lattice points around the parameter vector (r, ρ_1, ρ_2) at which one wants to interpolate the value $C(r, \rho_1, \rho_2)$. Here $f_i(x)$ is equal to $1 - \left|\frac{x_i - x}{\delta x}\right|$ for $(i = -, +)$. Here x_- is the biggest grid coordinate in the x direction in the

precomputation table that is smaller than or equal to x, and $x_+ = x_- + \delta x$ where δx is the spacing of the grid in the x direction.

3.4.2 Inverting the covariance matrix

After getting all elements in $\boldsymbol{\Sigma}$ and $\boldsymbol{\sigma}$, it is straightforward to carry out the kriging procedure as shown in (3.26) through (3.29). However, the number of observations in a global data set is generally very large. For example, the TOMS data that we considered in this paper have more than $20,000$ observations in just one day. In the kriging formulae, it is necessary to evaluate the inverse of the corresponding covariance matrix. It is well known that the floating operations of inverting an $n \times n$ matrix is in the order of $O(n^3)$. Such a brute-force inverting method seems hopeless even for this moderately large data set. It is desirable to utilize some special properties of the covariance matrix that can ease the computational burden. For a data set on regular spherical grids with axially symmetric property, it can be shown that the covariance matrix is actually *block circulant*. This is a special property for random field on sphere which is not present on the Euclidean space. That is, the covariance matrix takes the form of

$$C = \begin{pmatrix} C_1 & C_2 & C_3 & \dots & C_{n-1} & C_n \\ C_n & C_1 & C_2 & \dots & C_{n-2} & C_{n-1} \\ C_{n-1} & C_n & C_1 & \dots & C_{n-3} & C_{n-2} \\ & & & \dots & & \\ C_2 & C_3 & C_4 & \dots & C_n & C_1 \end{pmatrix}_{np \times np} \tag{3.32}$$

where each C_i is a $p \times p$ matrix of complex or real-valued elements itself. Here C_1 is the covariance matrix of observations on longitude 1 with itself; C_2 is the covariance matrix of observations on longitude 1 and observations on longitude 2, and so on. Vescovo (1997) gives an algorithm of computing the inverse of such a block circulant matrix which only involves inverting $p \times p$ matrices instead of the $np \times np$ one. Define $\omega = \exp\{2i\pi/n\}$ being a complex root of unity. Further define the Fourier transform of the C_k matrices

as

$$S_q = \frac{1}{n}\sum_{k=1}^{n}(\omega^*)^{(q-1)(k-1)}C_k\,, \tag{3.33}$$

for $q = 1, \ldots, n$ where ω^* denotes the complex conjugate of ω. For each $S_q(q = 1, \ldots, n)$, compute the p eigenvalues of nS_q and let them be $(\lambda_{q1}, \ldots, \lambda_{qp})$. If we go through all $q = 1, \ldots, n$, we will have a list of np eigenvalues

$$\{\underbrace{\lambda_{11}, \lambda_{21}, \ldots, \lambda_{1p}}_{\text{eigenvalues of } nS_1}, \underbrace{\lambda_{21}, \lambda_{22}, \ldots, \lambda_{2p}}_{\text{eigenvalues of } nS_2}, \ldots, \underbrace{\lambda_{n1}, \lambda_{n2}, \ldots, \lambda_{np}}_{\text{eigenvalues of } nS_n}\}\,. \tag{3.34}$$

The determinant of the original matrix C is the product of these np eigenvalues such that

$$\det C = \prod_{i=1}^{n}\prod_{j=1}^{p}\lambda_{ij}\,. \tag{3.35}$$

The inverse of a block circulant matrix is also block circulant, that is,

$$B \equiv C^{-1} = \begin{pmatrix} B_1 & B_2 & B_3 & \ldots & B_{n-1} & B_n \\ B_n & B_1 & B_2 & \ldots & B_{n-2} & B_{n-1} \\ B_{n-1} & B_n & B_1 & \ldots & B_{n-3} & B_{n-2} \\ & & & \ldots & & \\ B_2 & B_3 & B_4 & \ldots & B_n & B_1 \end{pmatrix}, \tag{3.36}$$

where each B_i is a $p \times p$ matrix. It is shown that

$$B_k = \frac{1}{n^2}\sum_{j=1}^{n}\omega^{(j-1)(k-1)}S_j^{-1}\,, \tag{3.37}$$

where S_j^{-1} stands for the inverse of matrix S_j. B_k is the inverse Fourier transform of the S_j^{-1}. The computational requirement for this algorithm is $O(np^3)$ instead of $O(n^3p^3)$ if inverting it directly. Details can be found in [De Mazancourt and Gerlic (1983); Vescovo (1997); Tee (2007)].

3.5 A Simple Simulation Study

A simple Monte Carlo simulation study is carried out on a unit sphere for illustration of how this methodology works. The random field is observed on a regular grid with

longitudes $\{-180°, -170°, \ldots, 160°, 170°\}$ and latitudes $\{-62.5°, -55°, \ldots, 55°, 62.5°\}$. The total number of observations is 1872. For an axially symmetric random field, ρ_s and σ^2 depends only on the latitude and they are parameterized by functions $\rho(s) = 100 + 0.4l^2$ and $\sigma^2(s) = 0.1 + 0.001l^2$ where l is the latitude at location s. We also add a small nugget effect of 0.1 to the simulated field. The realization of Z is plotted in Figure 3.3. The pattern of greater variation and smaller range at high latitudes is easily to spot from the plot.

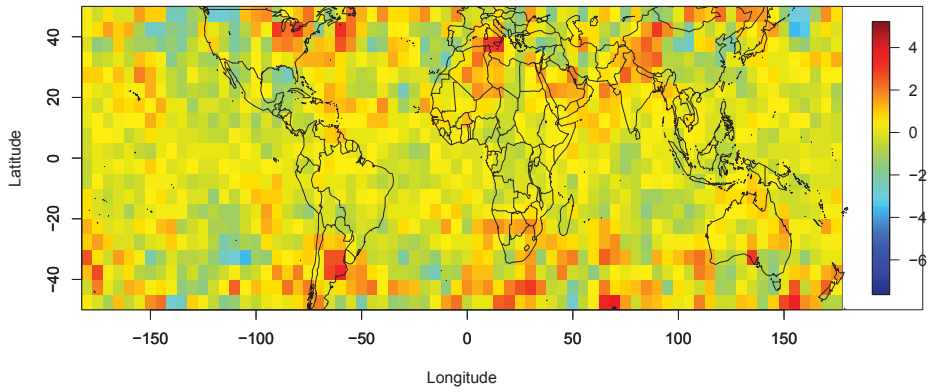


Figure 3.3: The simulated random field on sphere.

At each of the 26 latitudes, we use MLE to compute the estimated ρ and σ^2 as described in Section 3.3.1. Another option is to fit to the empirical variogram with least squares at each latitude. After getting the raw estimates, we apply local linear smoothing to get a smooth continuous function of both ρ and σ^2 as a function of the latitude. The result is shown in Figure 3.4. It is shown that the methodology is able to recover the true functional form reasonably well, especially for σ^2 where the estimated function is very close to the true one.

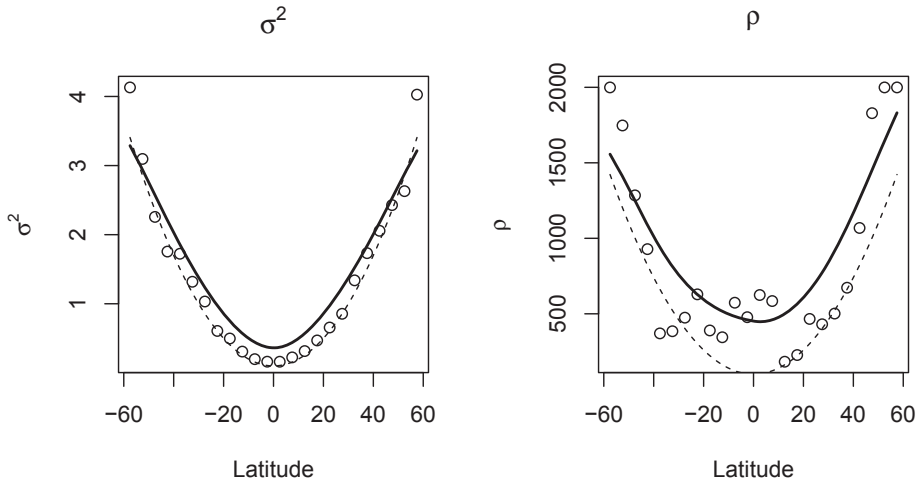


Figure 3.4: The estimated parameter at different latitudes. Open circles are raw estimates obtained by fitting MLE at each single latitude. The solid lines are from local linear smoothing. The dashed lines are the true parameter functions.

3.6 Real Data Analysis

As an application, we apply our estimation method to total column ozone level data on a global scale, which have been widely used as an example for global covariance modeling with axial symmetry [Cressie and Johannesson (2008); Jun and Stein (2008); Stein (2008)].

Total Ozone Mapping Spectrometer (TOMS) is a satellite instrument for measuring ozone values on a global scale. During the period of November 1978 to December 1994, several TOMS instruments were carried by NASA-satellites into the outer space including Nimbus-7 and Meteor-3 and provided global measurements of total column ozone on a daily basis. The data are either Level 2 or Level 3 versions. Level 2 data give spatial and temporal irregular measurements following the satellite scanning tracks [Jun and Stein (2008)]. Since the instrument relies on backscattered light, there are a lot of missing observations in Level 2 data. On the other hand, Level 3 data are post-processed from Level 2 data. They are obtained by averaging Level 2 data pixel by pixel and are on a spatially regular lattice with spacing 1° in latitude and 1.25° in longitude. The number

of missing values are thus greatly reduced.

The data set used in this analysis is the original TOMS Level 3 data for May 15, 1990, which is the same as [Jun and Stein (2008)]. There are 288 longitude points and 100 latitude points. The longitude points are evenly spaced over $-180°$ to $180°$ and the latitude points are evenly spaced over $50°$ South to $50°$ North. 71 out of the total 28,800 observations have missing values. We follow the same strategy as [Jun and Stein (2008)] and simply impute them with the average of their 8 queen contiguity neighboring sites. Different imputation methods can certainly be used. However, because of the small fraction of missing values, the choice of the imputing methods would not affect the results significantly.

First, it is observed that there is a mean structure in the data. The observed values range from 227 to 432 with noticeable bands along different latitudes. See Figure 2 in [Jun and Stein (2008)]. It turns out that subtracting the monthly average of May from the data does not remove the mean structure in a satisfactory way. There are hot spots at the high and low latitudes. Since spherical harmonics provide a natural basis for functions on the sphere, we regress the ozone level with $\{Y_n^m(\sin L, l)|n = 0, 1, \ldots, 12, m = -n, \ldots, n\}$ which better capture the large-scale mean structure in the data [Jun and Stein (2008)]. Afterwards the estimated average $\hat{\mu}(s)$ is subtracted from the data to get the residuals as shown in Figure 3.5. It can be seen that the spherical harmonics do a good job of modeling the mean structure and the residuals do not have any noticeable patterns such as hot spots.

We use one eighth of the original data for this data analysis. The latitudes are from $-46.5°$ to $49.5°$ with interval of $4°$ while the longitudes are from $-179.375°$ to $178.125°$ with interval of $2.5°$. The total number of observations is 3600. The number of observations at each latitude is 144.

It is easily seen from Figure 3.5 that the variations are bigger at high latitudes than at small latitude. Big positive and negative values are present at the top and bottom

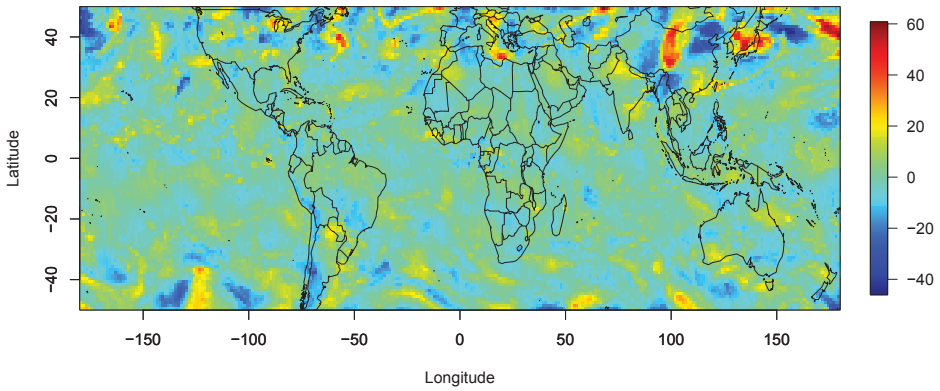


Figure 3.5: Residuals of TOMS column ozone level on May 15, 1990, after subtracting the mean structure estimated from the regression on spherical harmonics.

portion of the map, while the variability is small close to the equator. This feature is extracted in Figure 3.6 where we plot observations at three latitudes. The pattern of increasing magnitude of oscillations with increasing latitude is obvious from the plot.

We assume the smoothness parameter ν is a constant across the Earth. Its estimated value is obtained by maximizing the likelihood function value. In this analysis, we choose ν from a list of $\{0.1, 0.5, 1.0, 1.5, 2.0\}$. For each value of ν in this list, we compute the log likelihood as follows. We first use maximum likelihood in (3.22) to estimate parameters ρ and σ^2 at each latitude. Figure 3.7 shows the estimated variograms and the empirical method-of-moments variogram at the same latitudes as in Figure 3.6. At high latitudes, variogram has a larger value of sill but a smaller range than at low latitudes. Figure 3.8 shows the raw estimates at different latitudes and their smoothed lines using local linear smoothing. Figure 3.5 shows that the variability is larger at high latitude where most of the high and low residuals are present. This is confirmed by the estimates of σ^2 as shown in the left panel of Figure 3.8. At high latitudes, the estimates of σ^2 can be as high as 200 while it is around 15 at low latitudes. Similarly the estimates of ρ (right panel of Figure 3.8) also have a V-shaped pattern although it is not as strong as $\hat{\sigma}^2$. The estimated values of ρ range from 500 to 3500 at low and high latitudes, respectively.

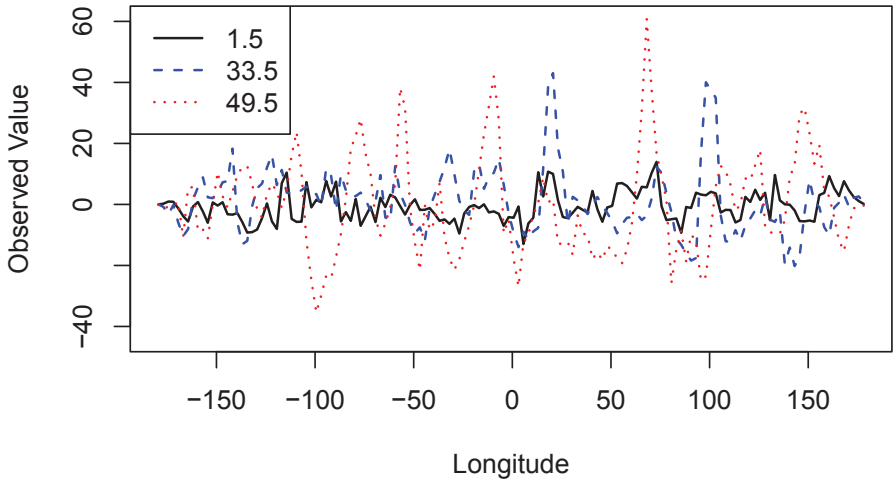


Figure 3.6: Observations at three different latitudes. At high latitude ($49.5°N$), observations have larger variation than those at low latitude.

The solid lines in Figure 3.8 are from the local linear smoothing with bandwidth selected from leave-one-out cross validation.

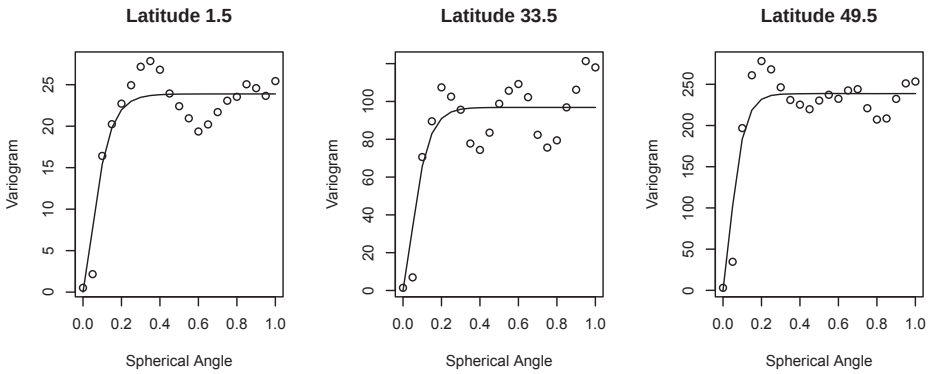


Figure 3.7: Empirical variogram and the estimated variogram function at different latitudes.

After getting the smoothed curves of estimation for both ρ and σ^2, we can get the estimated covariance function between any two locations on the sphere using (3.30). It is straightforward to compute all elements in $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ and the estimated covariance vector $\boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$ in the ordinary kriging formulae from the pre-computation table. The log likelihood

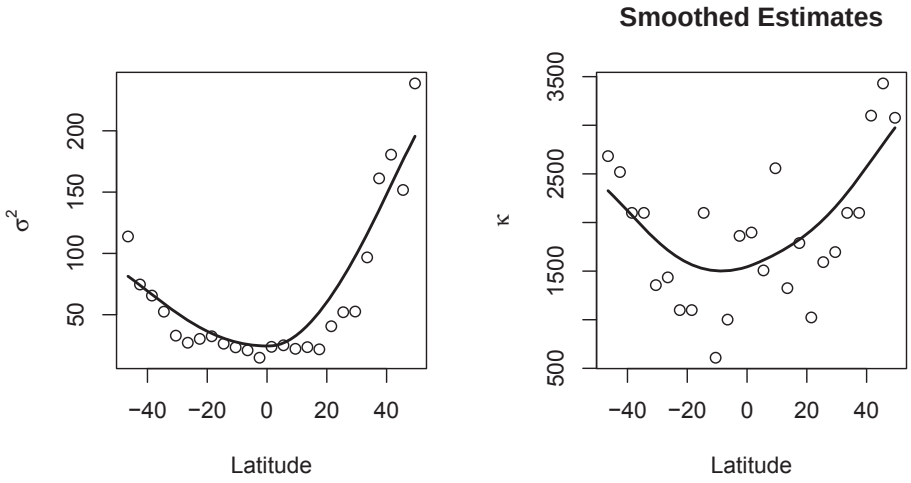


Figure 3.8: Smoothed parameter estimates across latitudes.

function can then be computed as

$$l(\nu) = -\frac{1}{2}\log|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})_\nu| - \frac{1}{2}(\mathbf{Z}-\hat{\boldsymbol{\mu}})^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})_\nu^{-1}(\mathbf{Z}-\hat{\boldsymbol{\mu}})\,. \tag{3.38}$$

Here $|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})_\nu|$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})_\nu^{-1}$ are computed from (3.35) and (3.37) using FFT. The log likelihood function at different values of ν is compared and it turns out that $\nu = 0.1$ has the maximum value of $l(\nu)$ so $\hat{\nu} = 0.1$. The computational time to get the inverse and determinant of the 3600×3600 covariance matrix is about 1 minute using the brute force method. As a comparison, the operation time is reduced to about 4 seconds using the algorithm described in section 3.4.2 taking into account of the block circulant property.

We carry out the prediction procedure at all 28800 locations in the original TOMS level 3 data set. The obtained maps of predicted values are shown in Figure 3.9 which is quite similar to the original data from Figure 3.5.

3.7 Conclusion and Discussion

In this article, we consider an approach of modeling nonstationary random field on sphere which has the property of axial symmetry. The spatial random field is modeled as a kernel convolution of a latent uncorrelated random field. The kernel function is

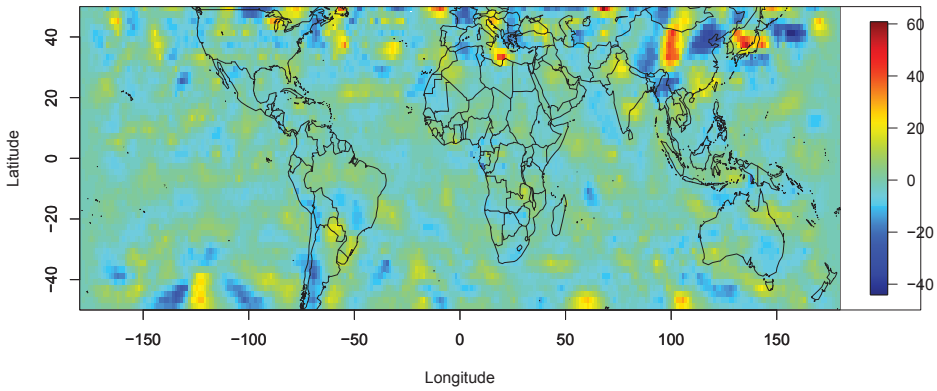


Figure 3.9: The map of predicted ozone value.

chosen similar to the Matérn covariance function on $\mathbb{R}^d$ which has an extra parameter controlling the smoothness. One advantage of using a convolution approach is that it automatically lead to a flexible and valid covariance function on sphere while many valid covariance functions on $\mathbb{R}^d$ are not valid on sphere anymore. If we assume the random field to be axially symmetric and longitudinally reversible, the process at each latitude is isotropic. We are able to estimate at discrete latitudes and smooth across latitudes using local linear smoothing method if we assume the parameters are smooth functions of latitude.

There are several computational issues associated with the large sample size. We use precomputation tables and linear interpolation to approximate the covariance functions between any two locations. For regular grids on sphere, the covariance matrix is block circulant whose determinant and inverse can be obtained from FFT to alleviate the computational burden. In this analysis we are dealing with a covariance matrix of size 3600, which is about the maximum size we can handle on our computer using brute force approach without having to wait a too long period of time. On the other hand, using the FFT method described in Section 3.4.2, we can easily handle a much bigger covariance matrix in a fraction of time.

One assumption in this study is the longitudinal reversibility (3.5). However, some

evidence indicates that this may not be true so that $C(L_1, L_2, l) \neq C(L_1, L_2, -l)$ for some L_1, L_2, and l [Jun and Stein (2008)]. Instead of fit an isotropic variogram, we might model the directional variograms along several directions at each latitude. Another assumption is that observations are on regular grids. For observations on irregular grids, it is intuitive to group data with similar latitudes. Instead of fit the isotropic variogram at a single latitude, we can aggregate the data within a latitude band. We can first select a list of discrete latitudes $\{l_i : i = 1, \ldots, n\}$. To get the raw estimates at latitude l_i, we will be using all observation from $(l_i - \delta l, l_i + \delta l)$. If parameters are smooth functions of latitude, it is safe to use this binning method for estimation.

CHAPTER 4. SEMIPARAMETRIC ESTIMATION OF SPECTRAL DENSITY AND VARIOGRAM WITH IRREGULAR OBSERVATIONS

In the study of intrinsically stationary spatial processes, we proposed a semiparametric variogram estimating method through its spectral representation. The spectral representation of the isotropic variogram alleviates the problem of negative definiteness in the spatial representation. The low frequency part of the spectral density is estimated by solving a regularized inverse problem through quadratic programming. The behavior at high frequencies, which is more important in spatial kriging, is modeled via a parametric tail in the form of a power decaying function. The power parameter in the tail is estimated by a log likelihood method. A simulation study is carried out to compare the estimation and prediction performance with the nonparametric approach proposed in [Huang *et al* (2011)].

4.1 Introduction

In geostatistics, a random field Z is a collection of random variables that are indexed by a continuous subset $D \subset \mathbb{R}^d$. Its short range structure is often modeled through covariance function or variogram [Cressie (1993)]. For inference purposes, $Z(s)$ is usually assumed to bear certain stationary properties. For example, a common assumption is the intrinsic stationarity where

$$E(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = 0\,, \tag{4.1}$$

$$\mathrm{Var}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = 2\gamma(\mathbf{s}_i - \mathbf{s}_j)\,, \tag{4.2}$$

for any locations $\mathbf{s}_i, \mathbf{s}_j \in D$. The variance of the increments only depend on their relative displacement. If we further assume that the variance only depends on the magnitude of the relative displacement, the random field is said to be isotropic, that is, $2\gamma(\mathbf{s}_i - \mathbf{s}_j) = 2\gamma^*(\|\mathbf{s}_i - \mathbf{s}_j\|)$.

A valid variogram function 2γ must satisfy the conditionally negative definiteness condition,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0\,, \tag{4.3}$$

for any finite number n of spatial locations $\{s_1, \ldots, s_n\} \subset D$, and real numbers $\{a_1, \ldots, a_n\}$ which satisfy $\sum_{i=1}^{n} a_i = 0$. A common practice to get a valid isotropic variogram satisfying (4.3) is to develop parametric variogram models. These parametric models are obtained either by fitting to empirical method-of-moments variogram estimates by weighted least squares [Cressie (1985)], or by likelihood-based approach [Mardia and Marshall (1984); Stein *et al* (2004)]. However, the choice of the parametric forms is subjective and an inappropriate model sometimes will result in bad prediction performance.

Some work has been done considering a broad class of conditionally negative definite functions based on the spectral representation of variogram functions. Schoenberg (1938) showed that a spectral representation of an isotropic variogram in $\mathbb{R}^d$ is

$$2\gamma(h) = \int_0^\infty (1 - \Omega_d(\omega h))dF(\omega)\,, \tag{4.4}$$

where $\Omega_d(t) = (2/t)^{(d-2)/2}\Gamma(d/2)J_{(d-2)/2}(t)$ and $J_v(t)$ is the Bessel function of the first kind of order ν and F is a nondecreasing function $(0, \infty)$ with $\int_0^\infty (1/r(\omega))dF(\omega) < \infty$, where $r(\omega) = (1 + \omega^2)/\omega^2$ is a regularization function [Huang *et al* (2011)]. For the planar case with $d = 2$, it simplifies to

$$2\gamma(h) = \int_0^\infty (1 - J_0(\omega h))dF(\omega)\,. \tag{4.5}$$

In the cases where F has a Lebesgue density f, called the spectral density, (4.5) can be rewritten as

$$2\gamma(h) = \int_0^\infty (1 - J_0(\omega h))r(\omega)f(\omega)d\omega\,. \tag{4.6}$$

The advantage of the above spectral representation in estimating the variogram compared to the spatial representation is that it is easier to construct a valid variogram function in the former. A valid variogram function has to be conditionally negative definite which is not generally easy to verify for nonparametric estimations. However, in the spectral representation, the only requirement is a non-negative spectral density function which is much easier to construct.

Shapiro and Botha (1991) proposed to use a finite discrete measure $F(\omega)$ in (4.5) with positive jumps $p_1, \ldots, p_m$ at nodes $t_1, \ldots, t_m$, and (4.5) reduces to

$$2\gamma(h) = \sum_{j=1}^{m} p_j(1 - J_0(\omega t_j)) \,. \tag{4.7}$$

The jump values can be computed by minimizing the mean squared difference between the raw variogram estimates and the estimator. In addition to reinforce the positivity constraints with quadratic programming, they also considered different conditions regarding the smoothness, monotonicity, and convexity of the computed estimators. Shapiro and Botha's method requires a subjective choice of the locations of nodes $t_1, \ldots, t_m$. A common choice of equispaced nodes may convey to spurious oscillations of the estimator. Genton and Gorsich (2002) proposed to use the root of some Bessel functions where fewer nodes are needed and computation becomes simpler.

Hall *et al* (1994) proposed a method of kernel estimator where the kernel smoothed version of the preliminary empirical covariogram estimates is first Fourier-tranformed to the spectral domain. They then truncated it by setting the negative values to 0 and Fourier-transformed back to the spatial domain to ensure a valid covariance function estimator.

García-Soidán *et al* (2004) proposed a similar nonparametric Nadaraya-Watson estimator as [Hall *et al* (1994)], but adapted to isotropic setting. They proved the properties of asymptotic unbiasedness and consistency of their estimator, and utilized an adaptation of Shapiro and Botha's fit to guarantee a valid variogram estimator.

Im *et al* (2007) proposed a semiparametric method for estimating spectral densities of isotropic Gaussian processes with scattered data. They used a linear combination of B-splines up to a cutoff frequency and a truncated algebraic tail for higher frequencies. They showed that the prediction performance is superior considering the high frequency behavior which is very important for interpolation purposes.

Huang *et al* (2011) proposed a nonparametric variogram estimator through its spectral representation by solving a regularized inverse problem. They used smoothing splines to fit the spectral density up to a threshold frequency and compute the estimated variogram from a discrete Riemann sum. We will brief describe their estimating methodology in the next section.

In this paper, we propose a semiparametric method for estimating spectral densities of isotropic random process, which is a generalization of [Huang *et al* (2011)]. The spectral density function is modeled by smoothing splines for low frequencies up to a cutoff, and by a truncated algebraic tail for high frequencies.. In section 4.2, we brief review the nonparametric estimation method proposed by [Huang *et al* (2011)] and propose our generalization of adding a parametric tail. In section 4.3 we discuss the numerical implementation. By discretization, the problem at hand can be solved with an iterative quadratic programming. Section 4.4 describes the simulation studies, comparing estimation performance of the spectral density, the variogram function, and the nugget effect, and the prediction performance of irregularly distributed observations. Section 4.5 summarizes and discusses possible future work.

4.2 Methodology

We assume that the observations come from an isotropic spatial random field on $\mathbb{R}^2$ which takes the form of

$$Z(s) = Y(s) + \epsilon(s)\,, \quad s \in D \tag{4.8}$$

where Y is an intrinsically stationary spatial process with isotropic variogram 2γ and spectral density $f(\omega)$, and $\epsilon(\cdot)$ is a mean zero, white noise measurement error process that is independent of the process $Y(\cdot)$ with $\mathrm{Var}(\epsilon(s)) = \sigma_\epsilon^2$. Suppose the process is observed at spatial locations $s_1, s_2, \ldots, s_N$ with N observational sites. We can see that

$$\begin{aligned} E[Z(s_i) - Z(s_j)]^2 &= E[Y(s_i) - Y(s_j) + \epsilon(s_i) - \epsilon(s_j)]^2 \\ &= \mathrm{Var}[Y(s_i) - Y(s_j)] + E[\epsilon(s_i) - \epsilon(s_j)]^2 , \end{aligned}$$

which is an unbiased estimator of $2\gamma(\|s_i - s_j\|) + 2\sigma_\epsilon^2$ with

$$2\gamma(\|s_i - s_j\|) + 2\sigma_\epsilon^2 = \int_0^\infty [1 - J_0(\omega\|s_i - s_j\|)]r(\omega)f(\omega)d\omega + 2\sigma_\epsilon^2 . \tag{4.9}$$

Intuitively, if a function g is close to f and a positive number c is close to $2\sigma_\epsilon^2$, the following sum of squares will be small

$$\sum_{i\neq j}^{N} \left[z_{i,j} - \int_0^\infty [1 - J_0(\omega\|s_i - s_j\|)]r(\omega)g(\omega)d\omega - c \right]^2 , \tag{4.10}$$

where $z_{ij} = (Z(s_i) - Z(s_j))^2$. However, an straightforward nonparametric estimation approach will typically result in a spurious oscillation. To balance the goodness-of-fit and smoothness of the fit, a regularization term is added to (4.10) to penalize for the roughness of the fitted function. An effective way to address the penalized-least-squares problem is through the smoothing spline approach [Wahba (1990)]. Then the target function that we want to minimize becomes

$$\sum_{i\neq j}^{N} \left[z_{i,j} - \int_0^\infty [1 - J_0(\omega\|s_i - s_j\|)]r(\omega)g(\omega)d\omega - c \right]^2 + \lambda J(g) . \tag{4.11}$$

Here, $\lambda > 0$ is the smoothing parameter of the regularization and the penalty term $J(g) = \int_0^\infty [g^{(2)}(\omega)]^2 d\omega$.

4.2.1 Semiparametric estimation

In [Huang *et al* (2011)], g is estimated with smoothing spline method where a high frequency cutoff ω_c is introduced as the integral upper limit. In this way, frequencies

higher than ω_c is simply set to be zero. It is known that the kriging performance is mostly governed by the properties of the variogram at small distance lags which in turn correspond to high frequencies in the spectral density [Stein (1999); Fuentes (2001)]. Therefore, the information that is ignored could be important for interpolation purposes. Im *et al* (2007) proposed the following semiparametric form for the spectral density

$$f_\theta(\omega) = \sigma^2 \sum_{i=-1}^{l+1} b_i B_i(\omega) \mathbb{I}_{[0,\omega_c]}(\omega) + C_f \left(\frac{\omega_c}{\omega}\right)^\gamma \mathbb{I}_{[\omega_c,\infty)}(\omega)\,, \tag{4.12}$$

where B_is are B-splines of order 4 on $[0, \omega_c]$. The coefficient C_f is chosen to achieve continuity at ω_c so that $C_f = \sigma^2 \sum_{i=-1}^{l+1} b_i B_i(\omega_c)$. Inspired by this form, we propose a semiparametric estimator of the spectral density $\tilde{g}$ which is a summation of nonparametric smoothing splines at low frequencies and a parametric tail for high frequencies. The cutoff frequency is ω_c. In other words, $\tilde{g}(\omega)$ has the following semiparametric form

$$\tilde{g}(\omega) = g(\omega) + C_t \left(\frac{\omega_c}{\omega}\right)^\gamma \mathbb{I}_{[\omega_c,\infty)}(\omega)\,, \tag{4.13}$$

where $g(\omega)$ is the nonparametric spline estimator of the spectral density within the range of $[0, \omega_c]$. The power γ is essentially the smoothness parameter of the variogram at small lags since $\gamma = 2\nu + d$ where ν is the Matérn smoothness parameter. We assume $\tilde{g}$ is a continuous function on the real line which imposes the constraint on the coefficient C_t that $C_t = g(\omega_c)$.

By plugging (4.13) into (4.11), we get the following

$$\sum_{i \neq j}^{N} \left[z_{i,j} - L_{i,j}(g) - g(\omega_c) a(\omega_c, \gamma, h_{i,j}) - c\right]^2 + \lambda J(g)\,, \tag{4.14}$$

where

$$L_{i,j}(g) = \int_0^{\omega_c} (1 - J_0(\omega \|s_i - s_j\|)) r(\omega) g(\omega) d\omega \tag{4.15}$$

$$a(\omega_c, \gamma, h_{i,j}) = \int_{\omega_c}^{\infty} (1 - J_0(\omega \|s_i - s_j\|)) r(\omega) \left(\frac{\omega_c}{\omega}\right)^\gamma d\omega\,. \tag{4.16}$$

Since the tail already has a parametric form, it is redundant to penalize on its smoothness. Therefore, the penalty term $J(g)$ only depends on g.

Function $a(\omega_c, \nu, h_{i,j})$ can be evaluated through the Hankel transform $\int_{\omega_c}^{\infty}(1-J_0(\omega\|s_i-s_j\|))r(\omega)\omega^{-\gamma}d\omega$ [Im *et al* (2007)], which can be analytically computed as

$$\begin{aligned}&\int_{\omega_c}^{\infty}(1-J_0(\omega h_{ij}))\omega^{-\gamma}d\omega\\ =\;& \frac{\omega_c^{1-\gamma}}{\gamma-1}-\frac{h^{\gamma-1}}{2^{\gamma}}\frac{\Gamma\left(\frac{1-\gamma}{2}\right)}{\Gamma\left(\frac{1+\gamma}{2}\right)}-\frac{\omega_c^{1-\gamma}F_2\left(\frac{1-\gamma}{2},\frac{3-\gamma}{2},-\frac{1}{4}h_{i,j}^2\omega_c^2\right)}{\gamma-1}\end{aligned} \tag{4.17}$$

for $\omega_c \gg 1$, where $F_2(a;b,c;z)$ is a generalized hypergeometric function which can be represented by the series $\sum_{k=0}^{\infty}\frac{(a)_k}{(b)_k(c)_k}\frac{z^k}{k!}$. Here $(\cdot)_k$ represents Pochhammer's symbol [Abramowitz and Stegun (1965)], which is defined by $(a)_0 = 1$ and $(a)_k = a(a+1)(a+2)\cdots(a+n-1) = \Gamma(a+n)/\Gamma(a)$. Although the Hankel transform has an analytical form, it is hard to evaluate since it involves summing over a large number of terms in its series expansion. Numerical issues in evaluating Hankel transform are discussed in details in [Im *et al* (2007)].

The valid estimator $g \in W_m[0,\omega_c]$ where $W_m[0,\omega_c]$ is the Sobolev space of order m, consisting of functions on $[0,\omega_c]$ that are m-times differentiable with square integrable mth derivative. $W_m[0,\omega_c]$ has the structure of $\mathcal{H}_0 \oplus \mathcal{H}_1$ where

$$\begin{aligned}\mathcal{H}_0 &= \text{span}\{1, x, \ldots, x^{m-1}/(m-1)!\}\\ \mathcal{H}_1 &= \{f : f^{(s)}(0) = 0, s = 0, \ldots, m-1, \int_0^{\nu}(f^{(m)})^2dx < \infty\}.\end{aligned} \tag{4.18}$$

$W_m[0,\omega_c]$ is a reproducing kernel Hilbert space (RKHS) with corresponding reproducing kernels

$$\begin{aligned}R_0(x,z) &= \sum_{s=1}^{m}\frac{x^{s-1}}{(s-1)!}\frac{z^{s-1}}{(s-1)!}\\ R_1(x,z) &= \int_0^{\nu}\frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(z-u)_+^{m-1}}{(m-1)!}du\end{aligned} \tag{4.19}$$

with function $(\cdot)_+ = \max\{\cdot, 0\}$.

To ensure a valid variogram, its spectral density has to be non-negative on the real line. In the smoothing splines framework, this translates into the requirement that $g \geq 0$

which can be approximated by $g(\omega_1) \geq 0, \ldots, g(\omega_L) \geq 0$ for a set of densely distributed frequencies $\{\omega_1, \ldots, \omega_L\}$ on $[0, \omega_c]$. This constraint of non-negativity of $g(\omega_l)$ can be translated to $\langle g, R_{\omega_l} \rangle \geq 0$ for $l \in \{1, \ldots, L\}$ by the reproducing property, where $\langle \cdot, \cdot \rangle$ is the inner product of $W_m[0, \omega_c]$. Let $\eta_{i,j}$ be the representer of $L_{i,j}$ such that $\langle \eta_{i,j}, g \rangle = L_{i,j} g$ for $\forall g \in W_m[0, \omega_c]$ where L is the linear functional defined in (4.15). Moreover, define $\mathcal{P}$ to be the orthogonal projection operator onto $\mathcal{R}_1$, then the unique minimizer to (4.14) is

$$\hat{g} = \sum_{i \neq j}^{N} c_{i,j} \xi_{i,j} + \sum_{\tau=1}^{m} d_\tau \phi_\tau + \sum_{j=1}^{L} b_j \rho_j \,, \tag{4.20}$$

where $\xi_{i,j} = \mathcal{P}\eta_{i,j}$, and $\rho_j = \mathcal{P}R_{\omega_j}$.

4.2.2 Selection of the cutoff frequency

In [Huang *et al* (2011)], the cutoff frequency ω_c is taken to be π divided by the grid size when data are equally spaced. For irregularly distributed observations in this study, ω_c can be taken to be π divided by the average distance. In the simulation study, we try several different values of ω_c. The results do not depend heavily on the choices of ω_c which will be discussed in detail in Section 4.4.

4.2.3 Estimation of the decay rate

For the purpose of spatial prediction, the behavior of the process at high frequencies, which corresponds to the properties of the covariance function at small spatial lags, is more relevant. It is possible to obtain asymptotically optimal prediction when the spectral density at low frequencies is misspecified [Stein (1999)]. Fuentes (2001) proposed a procedure for interpolation that uses an expression of spectral density at high frequencies. For a random field with Matérn covariance function, the spectral density takes the form of $f(\omega) = \phi(\alpha^2 + \omega^2)^{-\nu - d/2}$. At high frequencies, it is approximated by $\phi\omega^{-2\nu-d}$. The power of the tail γ in our proposed method is related to ν by $\gamma = 2\nu + d$.

We propose to estimate γ (or ν) by comparing likelihood values. For each ν_i in a list of candidate power values $\{\nu_1, \ldots, \nu_m\}$, we compute the corresponding log-likelihood value

$$l(\nu_i) = -\frac{1}{2}\log|\Sigma(\nu_i)| - \frac{1}{2}(Z-\mu)^T\Sigma^{-1}(\nu_i)(Z-\mu)\,, \tag{4.21}$$

where the approach of computing covariance matrix $\Sigma(\nu_i)$ is described in Section 4.3. We then choose the ν_i value corresponding to the maximum value as the estimated power.

4.2.4 Selection of smoothing parameter

A data-driven method of choosing the smoothing parameter λ was discussed in [Huang *et al* (2011)] where a generalized cross validation approach for smoothing splines proposed by Villalobos and Wahba [Villalobos and Wahba (1987)] is utilized. For each given smoothing parameter λ, the quadratic programming solution in (4.28) can be obtained as $\hat{\mathbf{u}}(\lambda)$ and the fitted value is $\hat{\mathbf{y}}(\lambda) = B\hat{\mathbf{u}}(\lambda)$. Define $\mathrm{RSS}(\lambda) = \sum_{i=1}^{n_0} w_i(y_i - \hat{y}_i(\lambda))^2$ and $A(\lambda) = \tilde{B}(\tilde{B}^T W\tilde{B} + \lambda\tilde{\Psi})^{-1}\tilde{B}^T W$. The modified generalized cross validation function which takes into account the correlation in the spatial data is [Wang (1998)]

$$\frac{\mathrm{RSS}(\lambda)}{[1-(1/p_0)\mathrm{Tr}(\Xi^{-1}A(\lambda))]^2}\,, \tag{4.22}$$

where Ξ is the covariance matrix of y and $p_0 = \mathrm{Tr}(\Xi^{-1})$. Since Ξ needs to be estimated, Ξ is replaced by W^{-1} and p_0 by p and the following function is to be minimized with respect to λ,

$$V(\lambda) = \frac{\mathrm{RSS}(\lambda)}{[1-(1/p)\mathrm{Tr}(WA(\lambda))]^2}\,, \tag{4.23}$$

where $p = \mathrm{Tr}(WA(0)) = \mathrm{Tr}(W\tilde{B}(\tilde{B}^T W\tilde{B})^{-}\tilde{B}W)$. The value of λ which minimizes $V(\lambda)$ will used as the selected smoothing parameter in (4.28).

4.3 Numerical Evaluation

Since the closed-form expressions for $\xi_{i,j}$, $L_{i,j}\phi_\tau$ and $\langle\xi_{i,j}, \xi_{i',j'}\rangle$ in (4.20) is not known, some form of the numerical approximation is necessary to estimate g.

Following [Huang *et al* (2011)], we first choose $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_L\}$ to be a dense grid on $[0, \omega_c]$ with equal spacing with $\omega_1 = 0$ and $\omega_L = \omega_c$. The integration in (4.14) will be replaced by a discrete Riemann summation. We further define

$$\mathbf{l}_{i,j} = \frac{\omega_c}{L} \left[(1 - J_0(\omega_1 h_{i,j}))r(\omega_1), \ldots, (1 - J_0(\omega_L h_{i,j}))r(\omega_L) \right] \tag{4.24}$$

and $\mathbf{g}_\omega = \{g(\omega_1), \ldots, g(\omega_L)\}$. The roughness penalty term $J(g)$ can be expressed as $\mathbf{g}_\omega^T K \mathbf{g}_\omega$ where K is an $L \times L$ matrix which is determined by the locations of the knots [Green and Silverman (1994)]. $\tilde{L}_{i,j}(g)$ is now approximated by the Riemann sum

$$\tilde{L}_{i,j}(g) \approx l_{i,j}^T g_\omega + g(\omega_L) a(\omega_L, \nu, h_{i,j}) = \tilde{\mathbf{l}}_{i,j}^T \mathbf{g}_\omega \,, \tag{4.25}$$

where vector $\tilde{\mathbf{l}}_{i,j}$ differs from $\mathbf{l}_{i,j}$ only in the last element, that is, $(1 - J_0(\omega_L h_{i,j}))r(\omega_L) + a(\omega_L, \nu, h_{i,j})$ which contains the contribution for the parametric tail.

Solutions $\hat{g}$ is the unique natural spline that interpolates $(\omega_l, \hat{g}_{\omega_l})$ where $\hat{\mathbf{g}}_\omega = (\hat{g}_{\omega_1}, \ldots, \hat{g}_{\omega_L})^T$ is the minimizer of

$$\underset{c \geq 0, g_\omega \in \mathbb{R}^L, \mathbf{g}_\omega \geq \mathbf{0}}{\operatorname{argmin}} \left\{ \sum_{i \neq j}^{N} (x_{i,j} - \tilde{\mathbf{l}}_{i,j}^T \mathbf{g}_\omega - c)^2 + \lambda \mathbf{g}_\omega^T K \mathbf{g}_\omega \right\} . \tag{4.26}$$

Huang *et al* (2011) proposed a simplification of (4.26) when the data are observed on a $N = N_0 \times N_0$ regular grid. In that case, each distance $h_{i,j}$ will be duplicated a large number of times and the double summation can be replaced by a single-index summation.

Irregularly located data, on the other hand, rarely have the same distance between a pair of observations. Therefore, the dimension of B will be $N(N-1)/2 \times (L+1)$ where $N = N_0^2$. For large data set, the computation is overwhelming and quadratic programming can fail to work in practice. The method-of-moments variograms estimates for irregular data usually involves using tolerance regions [Cressie (1993)]. It is also reasonable to use tolerance region to get the empirical estimate of $E[Z(s_i) - Z(s_j)]^2$ here. Specifically, for a given spatial lag h_m, we define a tolerance region N_m which

includes all pairs with $\|s_i - s_j\| \leq \delta$ where δ is a pre-specified tolerance size. Let

$$u_m = \frac{1}{w_m} \sum_{\substack{i,j=1 \\ h_{i,j} \in N_m}}^{N} z_{i,j}, \quad w_m = |N_m| \,. \tag{4.27}$$

After going through all spatial lags $h_m : m = 1, \ldots, M$, we obtain a sequence $\{(h_m, u_m, w_m) : m = 1, \ldots, M\}$, which stands for the spatial lag, empirical variogram estimate, and the number of pairs at this lag. The double summation in (4.26) reduces to a single weighted summation of

$$\underset{c \geq 0, g_\omega \in \mathbb{R}^L, \mathbf{g}_\omega \geq \mathbf{0}}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} w_m (u_m - \tilde{\mathbf{l}}_m^T \mathbf{g}_\omega - c)^2 + \lambda \mathbf{g}_\omega^T K \mathbf{g}_\omega \right\} . \tag{4.28}$$

Let $\mathbf{u} = (u_1, \ldots, u_M)^T$ and $W = \operatorname{diag}(w_1, \ldots, w_M)$, (4.28) can be written as

$$\underset{\mathbf{v} \in \mathbb{R}^{L+1}, \mathbf{v} \geq \mathbf{0}}{\operatorname{argmin}} \left[\mathbf{v}^T (B^T W B + \lambda \Psi) \mathbf{v} - 2 \mathbf{u}^T W B \mathbf{v} \right] , \tag{4.29}$$

where

$$B = \begin{pmatrix} \tilde{\mathbf{l}}_1^T & 1 \\ \vdots & \vdots \\ \tilde{\mathbf{l}}_M^T & 1 \end{pmatrix}, \quad \Psi = \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix} . \tag{4.30}$$

It is further observed that (4.28) has a similar form as the weighted least squares approach of estimating variogram function with a parametric form. Cressie (1985) showed that the approximated variance of the variogram estimate is

$$\operatorname{Var}[\hat{\gamma}(h)] \approx \frac{2\gamma^2(h, \theta)}{|N(h)|} , \tag{4.31}$$

where $N(h)$ is the number of pairs used to estimate the variogram at spatial lag h. In (4.28), the weighting factor is w_m, the pair of observations at lag h_m, which does not full consider the correlation of the variogram estimates at different spatial lags. We propose to replace the weighting factor w_m in (4.28) by $w_m / 2\hat{\gamma}^2(h_m)$. Therefore, we will be evaluating

$$\underset{c \geq 0, g_\omega \in \mathbb{R}^L, \mathbf{g}_\omega \geq \mathbf{0}}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} \frac{w_m}{2\hat{\gamma}^2(h_m)} (u_m - \tilde{\mathbf{l}}_m^T \mathbf{g}_\omega - c)^2 + \lambda \mathbf{g}_\omega^T K \mathbf{g}_\omega \right\} , \tag{4.32}$$

which can be evaluated in an iterative procedure.

Equation (4.29) can be solved numerically using standard quadratic programming techniques. The first L elements in the solution $\hat{\mathbf{v}}$ from (4.28) are the estimate of the values of the spectrum at the knots, that is, $\hat{f}(\omega_i) = \hat{v}_i$. The entire function of f is then estimated by the natural spline that interpolates $\{\omega_l, \hat{f}(\omega_l)\}$. The last element in $\hat{\mathbf{v}}$ is the estimate of the measurement error variance such that $\hat{\sigma}_\epsilon^2 = \hat{v}_{L+1}/2$.

The variogram can be estimated via (4.5)

$$2\hat{\gamma}(h) = \int_0^\infty [1 - J_0(\omega h)] r(\omega) \hat{f}(\omega) d\omega \tag{4.33}$$

which can be approximated by Riemann sum

$$2\hat{\gamma}(h) = \frac{\omega_c}{L} \sum_{l=1}^{L} [1 - J_0(\omega_l h)] r(\omega_l) \hat{f}(\omega_l)\,. \tag{4.34}$$

For a stationary random field, the covariance function can be estimated by

$$\hat{C}(h) = \frac{\omega_c}{L} \sum_{l=1}^{L} J_0(\omega_l h) r(\omega_l) \hat{f}(\omega_l)\,. \tag{4.35}$$

Since all $\{\hat{f}(\omega_l)\}$ in the summation are nonnegative, the above estimate of the variogram or the covariance function is indeed valid.

4.4 Simulation

We simulate an irregularly distributed random field on the domain $[0, 10] \times [0, 10] \subset \mathbb{R}^2$. The results are based on 100 Monte Carlo simulations. The true covariance function is a Matérn function with parameter $\sigma^2 = 1$, $\phi = 2$, $\sigma_\epsilon^2 = 0.16$, and $\kappa = 1$. The true variogram function is thus

$$\gamma(h) = \sigma_0^2 \left(1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{a} \right)^\kappa K_\kappa \left(\frac{h}{a} \right) \right) + \sigma_\epsilon^2\,, \tag{4.36}$$

and the corresponding measurement error free spectral density is

$$f(\omega) = \sigma_0^2 \frac{2\kappa\omega^3}{a^{2\kappa}(\omega^2 + 1/a^2)^{\kappa+1}(1+\omega^2)}\,. \tag{4.37}$$

We use different sample sizes of $\{200, 500, 1000, 2000\}$ in this study. We first select a sample of 2000 observations which are uniformly distributed in that domain. Samples of smaller sizes are nested within bigger samples. For example, sample of size 1000 is randomly selected from the initial size-2000 sample, and so on. We then fix these locations so that all 100 simulations share the same sampling design. In each Monte Carlo simulation, we use Cholesky decomposition to simulate the random field.

We compare two estimating methods; one is the smoothing spline approach as proposed in [Huang *et al* (2011)] denoted as SS, and the other one is our proposed method including a smoothing spline plus a parametric tail which is denoted as SS+T. SS is fitted on the frequency interval of $[0, \omega_c]$. In SS+T, we use smoothing splines to fit on the interval of $[0, \omega_c]$ and fit the parametric tail for frequencies higher than ω_c. We use two cutoff frequencies $\omega_c = \{2, 4\}$ to study its effect on the estimation and prediction performance. The candidate power values discussed in Section 4.2.3 are from 0.5 to 2 with an interval of 0.25. For each value of ν in this list, we follow the steps in Section 4.3 and get the estimated covariance function from (4.35). We then compute the loglikelihood function from (4.21) and our estimate of ν is chosen to the one which maximizes the loglikelihood.

4.4.1 Estimation

We compare the integrated squared error (ISE) [Yu *et al* (2007)] for two methods

$$\text{ISE}(f) = \int_0^{\omega_c} [\hat{f}(\omega) - f(\omega)]^2 d\omega \,, \tag{4.38}$$

$$\text{ISE}(\gamma) = \int_0^{h_c} [\hat{\gamma}(h) - \gamma(h)]^2 dh \,, \tag{4.39}$$

which characterizes the fitting performance of the spectral density and variogram. Here h_c is the largest spatial lags used in the estimation of the variogram which is 5 in this simulation. The means and standard deviations of the ISEs for spectral density and variogram function are shown in Tables 4.1 and 4.2, respectively. The rows stand for

different sample size in the range of 200 and 2000. The numbers in parentheses are the standard deviations of 100 ISEs.

In Table 4.1, the means and standard deviations of SS and SS+T with different cutoff frequencies ω_c are shown. It is observed that the estimation of spectral density with SS+T is consistently better than using a strict cutoff. The improvement is larger with smaller cutoff frequencies since more information is lost by throwing away tail information higher than the cutoff. For large cutoff $\omega_c = 4$, the results are becoming similar between SS and SS+T. This discrepancy will eventually disappear for a sufficiently high ω_c.

Table 4.1: Entries in the table show the mean and standard error of the ISE for the spectral density for different sample sizes and cutoff frequencies. The numbers in parentheses are standard deviations. SS stands for the smoothing splines method in [Huang *et al* (2011)] and SS+T is our method. The results are based on 100 Monte Carlo simulations.

	$\omega_c = 2$				$\omega_c = 4$			
	SS		SS+T		SS		SS+T	
n	Mean	SE	Mean	SE	Mean	SE	Mean	SE
200	0.0165	0.0007	0.0147	0.008	0.0166	0.0006	0.0142	0.0005
500	0.0153	0.0005	0.0138	0.011	0.0157	0.0005	0.0141	0.0005
1000	0.0142	0.0004	0.0132	0.012	0.0146	0.0004	0.0140	0.0003
2000	0.0167	0.0003	0.0141	0.016	0.0144	0.0004	0.0134	0.0003

Table 4.2 summarizes the estimation result of the variogram function. Since the objective is to find the optimal fit to the empirical variogram estimates, the ISE values are close in the table no matter the sample size and the cutoff frequency. However, we still observe a slight improvement with increasing ω_c. At $\omega_c = 4$, SS+T is slightly better than SS. With a low cutoff frequency $\omega_c = 2$, SS+T has a larger ISE which is mainly due to the fact that we put more weight on the estimates at small lags (see (4.31)) so the fit at large lags tend to deviates more from the empirical estimates. The increase in sample sizes does not help much here after $n = 500$.

The comparison of the nugget estimates are shown in Table 4.3. SS+T has a much better estimating performance than SS across all sample sizes and cutoff frequencies. For example, at $n = 200$ and $\omega_c = 4$, the mean squared errors of SS+T is half of that of SS.

Table 4.2: Entries in the table show the mean and standard error of the ISE for the variogram function for different sample sizes and cutoff frequencies. The numbers in parentheses are standard deviations. SS stands for the smoothing splines method in [Huang *et al* (2011)] and SS+T is our method. The results are based on 100 Monte Carlo simulations.

	$\omega_c = 2$				$\omega_c = 4$			
	SS		SS+T		SS		SS+T	
n	Mean	SE	Mean	SE	Mean	SE	Mean	SE
200	0.188	0.018	0.194	0.019	0.186	0.019	0.176	0.017
500	0.173	0.010	0.178	0.012	0.172	0.011	0.169	0.010
1000	0.163	0.007	0.166	0.008	0.163	0.008	0.163	0.007
2000	0.166	0.005	0.175	0.006	0.160	0.005	0.165	0.005

Such improvement can also be seen from Figure 4.2.

Table 4.3: Bias and mean squared error of nugget estimates $\hat{\sigma}^2_\epsilon$ for different sample sizes and cutoff frequencies. The true value of nugget is $\sigma^2_\epsilon = 0.16$. The results are based on 100 Monte Carlo simulations.

	$\omega_c = 2$				$\omega_c = 4$			
	SS		SS+T		SS		SS+T	
n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
200	0.038	0.042	0.008	0.030	-0.044	0.058	-0.017	0.029
500	0.042	0.027	0.008	0.019	-0.012	0.033	-0.005	0.019
1000	0.046	0.017	0.007	0.012	0.004	0.015	0.000	0.011
2000	0.048	0.011	0.004	0.008	0.009	0.008	0.000	0.007

Figure 4.1 shows the spectral densities for different cutoff frequencies and Figure 4.2 shows the corresponding variogram functions for one simulation. The cutoff frequencies are 2 and 4, respectively. The thick solid lines are the true spectral density. The thin solid lines are the estimated spectral densities with SS+T with different colors standing for different sample sizes, and the dashed lines are estimates from SS. It can be seen that the estimates of the spectral density for small ω_c are not very satisfactory, especially for small sample sizes. The objective function is the weighted squared difference between the empirical variogram and the fit. Even though the spectral density estimates are not good, the corresponding variogram function estimates are satisfactory. This is a common issue in an ill-posed inverse problem. As the cutoff frequency ω_c increases, the estimated curves are getting closer to the true spectral function. If sample sizes are increased for a

fixed ω_c, the estimates are also getting closer to the truth. An observation is that SS+T tends to bring down the estimates closer to the cutoff frequency because of the continuity constraint we imposed on the estimation of the spectral density. In Figure 4.2, the most striking feature is the improvement in the estimation of the nugget effect, especially for low cutoff frequencies. SS tends to overestimate the nugget effect as noted from Table 4.3. The inclusion of a tail generally will help in the estimation at small lags which is more important in kriging. For large cutoff frequencies, the difference between SS and SS+T diminishes.

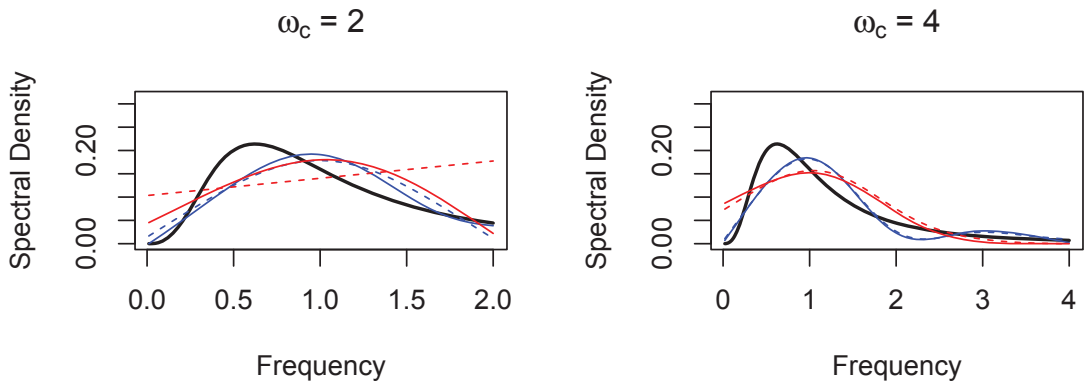


Figure 4.1: Plots of spectral densities for different cutoff frequencies ω_c. Thick solid lines are true spectral density. Thin solid lines are estimates with SS+T and dashed lines are from SS. Different colors stands for different sample sizes. Blue: $n = 2000$ and Red: $n = 500$.

4.4.2 Prediction

Similar to Chapter 2, we define a prediction performance measure that is more appropriate for interpolation purposes. Suppose that $\hat{Z}_0(s)$ is the predicted value at location s using the true covariance function C_0 and $\hat{Z}_i(s)$ is the predicted value with covariance function C_i (which may be misspecified). Let $e_i(s) = Z(s) - \hat{Z}_i(s)$ be the prediction error. E_0 is the expectation under the true covariance function C_0. Then $E_0 e_0^2$ is the mean squared prediction error (MSPE) of the best linear unbiased predictor (BLUP)

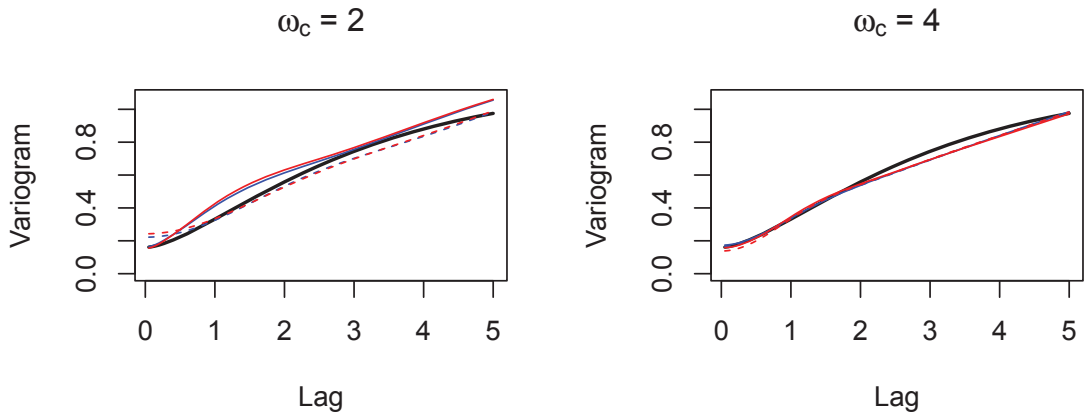


Figure 4.2: Plots of variogram functions for different cutoff frequencies ω_c. Thick solid lines are true spectral density. Thin solid lines are estimates with SS+T and dashed lines are from SS. Different colors stands for different sample sizes. Blue: $n = 2000$ and Red: $n = 500$.

or the kriging variance. Im *et al* (2007) defined a quantity IPE(s) which indicates the increase in prediction error at location s:

$$\text{IPE}(s) = \frac{E_0 e_i^2(s)}{E_0 e_0^2(s)} - 1 = \frac{E_0(\hat{Z}_i(s) - \hat{Z}_0(s))^2}{E_0 e_0^2(s)} \,. \tag{4.40}$$

This quantity represents the extra mean squared prediction error introduced by predicting with an estimated (possibly misspecified) covariance function instead of the true one. Smaller IPE value indicates a better kriging performance for the corresponding covariance function.

In this study we consider 81 interpolation sites $\{1, \ldots, 9\} \times \{1, \ldots, 9\}$ inside the observation region and 40 sites on the edge for extrapolation purposes as shown in Figure 4.3. For each simulation $l = 1, \ldots, 100$, we compute the median IPE (mIPE) as

$$\overline{\text{mIPE}}_{\text{in},l} = \text{median}\left\{ [\hat{Z}_i(s,l) - \hat{Z}_0(s,l)]^2 \middle| i = 1, \ldots, N_{\text{in}}, s = 1, \ldots, 81 \right\} , \tag{4.41}$$

and

$$\overline{\text{mIPE}}_{\text{ex},l} = \text{median}\left\{ [\hat{Z}_i(s,l) - \hat{Z}_0(s,l)]^2 \middle| i = 1, \ldots, N_{\text{ex}}, s = 1, \ldots, 40 \right\} , \tag{4.42}$$

for $N_{\text{in}} = 81$ interpolation sites and N_{ex} extrapolation sites, respectively. The mean and standard deviation from 100 Monte Carlo simulations are shown in Tables 4.4 and 4.5.

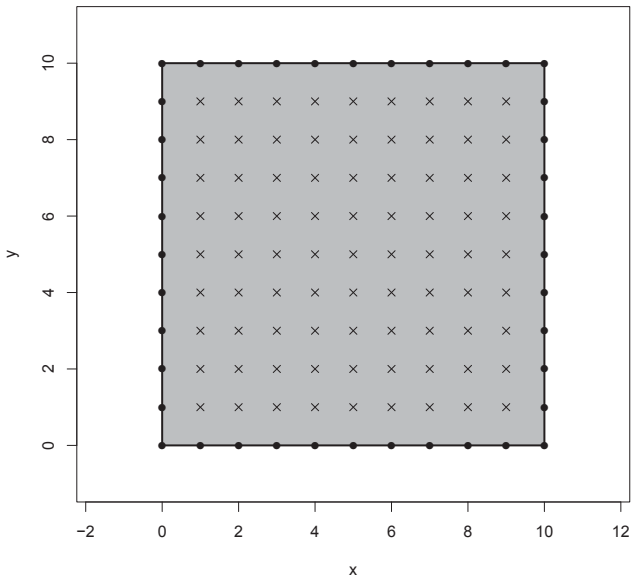


Figure 4.3: The grey area is the sampling domain $[0, 10] \times [0, 10]$. Crosses are interpolation sites $\{1, \ldots, 9\} \times \{1, \ldots, 9\}$, and solid circles are extrapolation sites on the edge of the domain.

For both tables, the prediction performance is improved with increasing sample size. For interpolation, SS+T is superior to SS at small sample sizes. The difference becomes smaller with increasing sample sizes. SS+T is consistently better than SS at $\omega_c = 4$. For extrapolation where predictions are made on the edge of the sampling domain (Table 4.5), SS+T has better prediction performance across all sample sizes and cutoff frequencies. This indicates the importance of high frequency information in spatial prediction.

Table 4.4: Mean and standard error of the median IPE at 81 interpolation locations for different sample sizes and cutoff frequencies. The results are based on 100 Monte Carlo simulations. The entries are in units of 10^{-3}.

	$\omega_c = 2$				$\omega_c = 4$			
	SS		SS+T		SS		SS+T	
n	Mean	SE	Mean	SE	Mean	SE	Mean	SE
200	1.691	0.047	1.463	0.041	1.942	0.053	1.391	0.049
500	0.972	0.015	0.892	0.014	1.061	0.016	0.882	0.014
1000	0.745	0.007	0.811	0.008	0.8227	0.010	0.793	0.008
2000	0.741	0.005	0.852	0.006	0.809	0.005	0.802	0.004

Table 4.5: Mean and standard error of the median IPE at 40 extrapolation locations for different sample sizes and cutoff frequencies. The results are based on 100 Monte Carlo simulations. The entries are in units of 10^{-3}.

	$\omega_c = 2$				$\omega_c = 4$			
	SS		SS+T		SS		SS+T	
n	Mean	SE	Mean	SE	Mean	SE	Mean	SE
200	4.643	0.173	4.214	0.191	4.575	0.205	4.052	0.202
500	3.285	0.117	2.922	0.109	3.036	0.121	2.865	0.118
1000	2.151	0.033	2.031	0.035	1.992	0.038	1.996	0.032
2000	1.765	0.021	1.666	0.019	1.633	0.022	1.601	0.020

4.5 Summary and Discussion

In this article we use a semiparametric approach to model the variogram function and its spectral density for irregularly distributed spatial data. Using the spectral representation of the random field, we circumvent the difficulty of getting a valid variogram function. The spectral density is modeled by a smoothing splines for low frequencies and a parametric tail at frequencies higher than a threshold value. This regularized inverse problem can be solved through an iterative quadratic programming procedure. The simulation study shows that the inclusion of a parametric tail has the advantages of better estimation of the spectral density and the nugget effect.

In this article the cutoff frequency ω_c is chosen in an *ad hoc* way. By allowing ω_c to vary at several values, we could study its effect on estimation and prediction. For a small ω_c value, the spectral density may not start to decay as a parametric tail. On the other hand, if ω_c is taken to be too large, the contribution from the tail may be too small. Therefore a data-driven method of selecting ω_c would be useful in future studies.

The power of the tail function is chosen to maximize the log likelihood function. For an intrinsically stationary random field which is not weakly stationary, it is possible to fit a linear model for high frequencies using OLS, $\log(f(\omega)) = \beta_0 + \beta_1 \log(\omega)$ and the power of the tail can be obtained from the fitted value $\hat{\beta}_1$ which is an estimate of $-2\nu - d$.

APPENDIX A. ADDITIONAL MATERIAL

Proof of Theorem 2.3.1

Lemma 1. *The k^{th} derivative of Legendre polynomial with order n evaluated at 1 is a polynomial of n with order $2k$. In other words, $P_n^{(k)}(1) = \sum_{j=1}^{2k} a_j n^j$ with a_j the coefficients and $a_{2k} > 0$.*

Proof. Prove by induction. From the iterative property of Legendre polynomials

$$(2n+1)P_n(x) = \frac{d}{dx}[P_{n+1}(x) - P_{n-1}(x)]\,, \tag{A.1}$$

we have

$$P_n'(1) = P_{n-2}'(1) + (2n-1)\,, \tag{A.2}$$

Combining with the initial conditions $P_0'(1) = 0$ and $P_1'(1) = 1$, it gives us the following iterative formula for $P_n'(1)$

$$\begin{aligned} P_n'(1) &= \begin{cases} \sum_{j=1}^{n/2}(4j-1) & \text{even } n \\ \sum_{j=0}^{(n-1)/2}(4j+1) & \text{odd } n \end{cases} \\ &= \frac{n(n+1)}{2}\,. \end{aligned} \tag{A.3}$$

Now suppose $P_n^{(k)}(1) = \sum_{j=1}^{2k} a_j n^j$ and $a_{2k} > 0$. Using $P_n^{(k+1)}(1) = P_{n-2}^{(k+1)}(1) + (2n -$

$1)P_{n-1}^{(k)}(1)$, we have

$$\begin{aligned} P_n^{(k+1)} &= \begin{cases} \sum_{j=1}^{2k} a_j \left[\sum_{l=1}^{n/2}(4l-1)(2l-1)^j\right] & \text{even } n \\ \sum_{j=1}^{2k} a_j \left[\sum_{l=1}^{(n+1)/2}(4l-3)(2l-2)^j\right] & \text{odd } n \end{cases} \\ &= \sum_{j=1}^{2k} a_j \sum_{l=1}^{j+2} b_{jl} n^l \,, \text{ where } b_{j,j+2} > 0 \text{ for any } j \in \{1,\ldots,2k\} \\ &= \sum_{l=1}^{2(k+1)} \left(\sum_{j=1}^{2k} a_j b_{jl}\right) n^l \,, \end{aligned} \tag{A.4}$$

where the coefficient of n^{2k+2} term is $a_{2k}b_{2k,2k+2} > 0$. □

Lemma 2. *$P_n^{(k)}(\cos\theta)|_{\theta=0} = \sum_{j=1}^{k} c_j n^j$ for even k with $c_k \neq 0$; $P_n^{(k)}(\cos\theta)|_{\theta=0} = 0$ for odd k.*

Proof. It is easy to show that a general form for $P_n^{(k)}(\cos\theta)$ is

$$P_n^{(k)}(\cos\theta) = \sum d_j (\sin\theta)^{k_1} (\cos\theta)^{k_2} P_n^{(k_3)}(\cos\theta) \,, \tag{A.5}$$

where $k_1 + k_2 = k_3$ and the sum is over all finite number of terms. By taking derivative of $P_n(\cos\theta)$ with respect to θ k_3 times, we get $(\sin\theta)^{k_3}$ in the coefficient. The remaining $k - k_3$ derivatives with respect to θ must be taken on $\sin\theta$ and $\cos\theta$. In order for k_1 to be zero, we must have $k - k_3 \geq k_3$ so that $k_3 \leq k/2$. Therefore, using Lemma 1 we have

$$P_n^{(k)}(\cos\theta)|_{\theta=0} = \sum d_j P_n^{(k/2)}(1) = \sum_{j=1}^{k} c_j n^j \,. \tag{A.6}$$

If k is an odd integer, all terms in $\sum d_j (\sin\theta)^{k_1} (\cos\theta)^{k_2} P_n^{(k_3)}(\cos\theta)$ have $k_1 > 0$ so that $P_n^{(k)}(\cos\theta)|_{\theta=0} = 0$. □

Proof. (**Theorem 3.2.1**)

A random process Z is m-times mean square differentiable if and only if $C^{(2m)}(0)$ exists and is finite [Stein (1999)]. From the form $C(\theta) = \sum_{n=0}^{\infty} b_n P_n(\cos\theta)$ and the result in Lemma 2, we have

$$C^{(2m)}(0) = \sum_{n=0}^{\infty} b_n P_n^{(2m)}(\cos\theta)|_{\theta=0} = \sum_{n=0}^{\infty} b_n \sum_{j=1}^{2m} c_j n^j \,. \tag{A.7}$$

If $b_n \sim O(n^{-2m-1-\delta})$ with $\delta > 0$, $C^{(2m)}(0) = \sum_{n=0}^{\infty} O(n^{-1-\delta}) < \infty$ so that Z is m-times mean square differentiable. It is easy to see that Corollary 2.3.2 follows. $\square$

BIBLIOGRAPHY

Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics*, *39*, 811–841.

Schoenberg, I. J. (1942). Positive definite functions on spheres. *Duke Mathematical Journal*, *9*, 96–108.

Jones, R. H. (1962). Stochastic processes on a sphere. *The Annals of Mathematical Statistics*, *34*, 213–218.

Matheron, G. (1962). *Traite de geostatistique appliquee, Tome I.* Paris: Memoires du Bureau de Recherches Geologiques et Minieres, no. 14, Editions Technip.

Abramowitz, M. and Stegun, I. (1965). *Handbook of mathematical functions (9ed).* New York: Dover.

Davis, P. J. (1979). *Circulant matrices.* New York: John Wiley & Sons.

Cressie, N. and Hawkins, D. M. (1980). Robust estimation of variogram: I. *Mathematical Geology*, *12*, 115–125.

De Mazancourt, T. and Gerlic, D. (1983). The inverse of a block-circulant matrix. *IEEE Transactions on Antennas and Propagation*, *31*, 808–810.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, *71*, 135–146.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, *17*, 563–586.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions, Vol. I & II*. New York: Springer.

Villalobos, M. and Wahba, G. (1987). Inequality constrained multivariate smoothing splines with application to the estima- tion of posterior probabilities. *Journal of the American Statistical Association*, *82*, 239–248.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B*, *50*, 297–312.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

Haas, T. C. (1990). Lognormal and Moving Window Methods of Estimating Acid Deposition. *Journal of the American Statistical Association*, *85*, 950–963.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods (2ed)*. New York: Springer.

Zimmerman, D. L. and Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, *33*, 77–91.

Shapiro, A. and Botha, J. D. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis*, *11*, 87–96.

Xu, Y. and Cheney E. W. (1992). Strictly positive definite functions on spheres. *Proceedings of the American Mathematical Society*, *116*, 977–981.

Sampson, P. D., and Guttorp, P. (1992). Nonparametric Estimation of Nonstationary Spatial Covariance Structure. *Journal of the American Statistical Association*, *87*, 108–119.

Cressie, N. (1993). *Statisics for spatial data*. New York: Wiley.

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, *89*, 368–378.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. New York: Chapman & Hall.

Wood, A. T. A. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics*, *3*, 409–432.

Wand, M. P. and Jones, M. C. (1994) *Kernel smoothing*. London: Chapman & Hall.

Peter H., Fisher, N. I., and Hoffmann B. (1994). On the Nonparametric Estimation of Covariance Functions. *Annals of Statistics*, *22*, 2115-2134.

Cherry, S., Banfield, J., and Quimby, W. F. (1996). An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes. *Journal of Applied Statistics*, *23*, 435–449.

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

Vescovo, R. (1997). Inversion of block-circulant matrices and circular array approach. *IEEE Transactions on Antennas and Propagation*, *45*, 1565–1567.

Schreiner, M. (1997). On a new condition for strictly positive definite functions on spheres. *Proceedings of the American Mathematical Society*, *125*, 531–539.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, *93*, 341–348.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. New York: Springer.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian statistics 6*, Oxford University Press, 761–768.

Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, *12*, 469-483.

Genton, M. G. and Gorsich, D. J. (2002). Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices. *Computational Statistics & Data Analysis*, *41*, 47–57.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, *83*, 493–508.

Gorsich, D. J. and Genton, M. G. (2004). On the discretization of nonparametric isotropic covariogram estimators. *Statistics and Computing*, *14*, 99–108.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, *66*, 275–296.

García-Soidán, P. H., Febrero-Bande, M., and González-Manteiga, W. (2004). Nonparametric kernel estimation of an isotropic variogram. *Journal of Statistical Planning and Inference*, *121*, 65–92.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Boca Raton: Chapman & Hall/CRC.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization (2ed)*. New York: Springer.

Xia, G. and Gelfand, A. E. (2006). Stationary process approximation for the analysis of large spatial datasets. *Technical report*, Duke University.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, *15*, 502–523.

Paciorek, C. J., and Schervish, M.J. (2006). Spatial Modelling Using a New Class of Nonstationary Covariance Functions. *Environmetrics*, *17*, 483-506.

Im, H. K., Stein, M. L., and Zhu, Z. (2007). Semiparametric estimation of spectral density with irregular observations. *Journal of the American Statistical Association*, *102*, 726–735.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based geostatistics*. New York: Springer.

Tee, G. J. (2007). Eigenvectors of block circulant and alternating circulant matrices. *New Zealand Journal of Mathematics*, *36*, 195–211.

Jun, M. and Stein, M. L. (2007). An approach to producing space-time covariance functions on spheres. *Technometrics*, *49*, 468–479.

Yu, K., Mateu, J., and Porcu, E. (2007). A kernel-based method for nonparametric estimation of variogram. *Statistica Neerlandica*, *61*, 173–197.

Stein, M. L. (2007). Spatial variation of total column ozone on a global scale. *The Annals of Applied Statistics*, *1*, 191–210.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, *103*, 1545–1555.

Jun, M. and Stein, M. L. (2008). Nonstationary covariance models for global data. *The Annals of Applied Statistics*, *2*, 1271–1289.

Stein, M. L. (2008). A modeling approach for large spatial datasets. *Journal of Korean Statistical Society*, *37*, 3–10.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, *70*, 209–226.

Zhu, Z. and Wu, Y. (2010). Estimation and prediction of a class of convolution-based spatial nonstationary models for large spatial data. *Journal of Computational and Graphical Statisics*, *19*, 74–95.

Huang, C., Hsing, T., and Cressie, N. (2011). Nonparametric estimation of the variogram and its spectrum. *Biometrika*, *98*, 775–789.

Huang, C., Zhang, H., and Robeson, S. M. (2011). On the Validity of commonly used covariance and variogram functions on the sphere. *Mathematical Geosciences*, *43*, 721–733.

Huang, C., Hsing, T., and Cressie, N. (2011). Spectral density estimation through a regularized inverse problem. *Statistica Sinica*, *21*, 1115–1144.

Gneiting, T. (2012). Strictly and non-strictly positive definite functions on spheres. *Preprint*.

Huang, C., Zhang, H., and Robeson, S. M. (2012). A simplified representation of the covariance structure of axially symmetric processes on the sphere. *Statistics and Probability Letters*, *82*, 1346–1351.

Heaton, M. (2013). Private communication.