Estimation and Inference for Spatial and Spatio-Temporal Mixed Effects Models

Casey M. Jelsema
Western Michigan University, jelsema.casey@gmail.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Statistical Models Commons

Recommended Citation
https://scholarworks.wmich.edu/dissertations/209

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
ESTIMATION AND INFERENCE FOR SPATIAL AND SPATIO-TEMPORAL MIXED EFFECTS MODELS

by

Casey M. Jelsema

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Department of Statistics
Western Michigan University
December 2013

Doctoral Committee:
Rajib Paul, Ph.D., Chair
Magdalena Niewiadomska-Bugaj, Ph.D.
Joseph McKean, Ph.D.
Kathleen Baker, Ph.D.
One of the most common goals of geostatistical analysis is that of spatial prediction, in other words: filling in the blank areas of the map. There are two popular methods for accomplishing spatial prediction. Either kriging, or Bayesian hierarchical models. Both methods require the inverse of the spatial covariance matrix of the data. As the sample size, n, becomes large, both of these methods become impractical. Reduced rank spatial models (RRSM) allow prediction on massive datasets without compromising the complexity of the spatial process. This dissertation focuses on RRSMs, particularly situations where the data follow non-Gaussian distributions.

The manner in which data can be non-Gaussian varies, and we address multiple such situations. We begin by developing multivariate log-normal kriging and block kriging equations, and explain how to implement them for data that are compositional. We also propose a robust non-parametric approach to parameter estimation for kriging.

Moving towards hierarchical models, we develop an empirical Bayes method to estimate parameters for heavy-tailed distributions. After, we turn to the problem of knot selection. The topic is under-represented in literature despite its importance to reduced rank spatial models.

Finally, we explore fully Bayesian models for non-Gaussian data based on scale mixtures of Gaussians. The method is flexible enough to model a variety of distributional forms. Furthermore, we discuss how our models can be extended to consider dependencies in space as well as time.
ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Rajib Paul. For the past five years he has filled the roles of professor, mentor, colleague, and friend. His endless support, encouragement, and dedication are responsible for making me into the Statistician I am today.

I would also like to thank my dissertation committee members, Drs. Magdalena Niewiadomska-Bugaj, Joseph W. McKean, and Kathleen Baker. Their excellent teaching and collaboration with my work have been invaluable during my studies at Western Michigan University and .

If not for the professors of Statistics at Grand Valley State University, I would probably not have changed my major to Statistics. So I would like to thank Drs. Soon Hong, Phyllis Curtiss, Kirk Anderson, and John Gabrosek (among others) for ensuring that my first experiences with Statistics were positive and exciting.

My family and friends also deserve my thanks here. Most notably my parents, Tom and Micky; my brothers, Timothy, Derek, and Jared; and my closest friend of more than 20 years, Jousha J. Kuipers. Their constant support and encouragement, while moral instead of academic, was integral to helping me make it through. I can assure them that I am indeed, finally, done with school.

And a special thanks to my grandfather, ‘Opa,’ the Rev. Gerard Bouma. If anyone is more excited for me to finish this process, it is him. He has been waiting for years to call me ‘Dr. C’ like his father, my namesake, who as he likes to say ‘had an honest doctorate.’

Casey M. Jelsema
Contents

List of Figures

1 Introduction

2 Multivariate Log-normal Kriging and Block Kriging with Applications to Coal Geology
   2.1 Introduction ...................................................... 7
   2.2 Multivariate Log-normal Kriging ............................. 12
   2.3 Multivariate Log-normal Block Kriging ...................... 20
   2.4 Implementation for Large Datasets .......................... 23
   2.5 Application to Coal Data ................................. 30
   2.6 Conclusions and Discussion .................................. 41

3 Robust Estimation and Fitting for Reduced Rank Spatial Models
   3.1 Introduction ..................................................... 43
CONTENTS

3.2 Robust Estimation and Fitting ........................................ 46
   3.2.1 Estimation Stage .............................................. 47
   3.2.2 Fitting Stage ................................................ 49
   3.2.3 Summary of Proposed Method .................................. 51
3.3 Asymptotic Properties .................................................. 52
3.4 Simulation Study ...................................................... 57
3.5 Application to NASA Data ............................................ 61
3.6 Conclusions and Discussion ........................................... 70

4 An Empirical Bayesian Approach for Selection of Knot Locations in Reduced Rank Spatial Models ........................................... 73
   4.1 Introduction ....................................................... 73
   4.2 Ascent Based MCEM ............................................... 78
   4.3 Knot Selection using Bayesian LASSO ............................ 86
   4.4 Applications ....................................................... 94
   4.5 Conclusions and Discussion ....................................... 99

5 A Flexible Class of Reduced Rank Spatial Models for Non-Gaussian Tail Behaviors ...................................................... 102
   5.1 Introduction ....................................................... 102
   5.2 Flexible Reduced Rank Spatial Model ............................. 104
   5.3 MCMC Algorithm and its Implementation ........................ 113
   5.4 Verifying the Conditions of MCCLT .............................. 118
   5.5 Application: Simulated and Real Data ............................ 122
CONTENTS

  5.5.1 Simulation Study ........................................... 122
  5.5.2 Application to Ozone Data ................................. 126
  5.6 Conclusions and Discussion ................................. 131

6 Spatio-Temporal Models ........................................ 133
  6.1 Introduction ................................................... 133
  6.2 A Flexible Non-Gaussian Spatio-temporal Model .......... 135

Appendix .......................................................... 140

Bibliography ....................................................... 143
List of Figures

2.1 Observed ilr coordinates (a) \( \ln \left( \frac{U_2}{U_1} \right)^{1/\sqrt{2}} \), (b) \( \ln \left( \frac{U_3}{U_1 U_2} \right)^{1/\sqrt{6}} \), and (c) \( \ln \left( \frac{U_4}{U_1 U_2 U_3} \right)^{1/\sqrt{12}} \). The horizontal axes are degrees longitude and the vertical axes are degrees latitude. The colorbar represents the unit free ilr coordinates. .................................................. 31

2.2 Histograms of the three ratios (a) \( \left( \frac{U_2}{U_1} \right)^{1/\sqrt{2}} \), (b) \( \left( \frac{U_3}{U_1 U_2} \right)^{1/\sqrt{6}} \), and (c) \( \left( \frac{U_4}{U_1 U_2 U_3} \right)^{1/\sqrt{12}} \). The respective natural logarithms (i.e. ilr coordinates) are in (d)-(f). .................................................. 33

2.3 Normal Probability Plots of detailed residuals after removing trend from (a) first, (b) second, and (c) third ilr coordinates using ordinary least-squares estimates. Dash-dotted lines represent the Normal distribution reference line. ......................... 34
2.4 Knot locations for the three ilr coordinates, we use $r_1 = 37, r_2 = 42, r_3 = 46$ knots for these three ratios, respectively. Asterisks represent the 37 knot locations, plusses represent the 42 knot locations, and circles represent the 46 knot locations. Note that we use overlapping knot locations. . . . . . . . . . . 35

2.5 Pointwise predictions (a)-(c) of the three isometric log-ratios over block 1 shown in Figure (2.6). The ilr transformation is given in (2.6). The respective RMSPEs are shown in (d) - (f). The horizontal axes are degrees longitude and the vertical axes are degrees latitude. The colorbar represents the unit free ilr coordinates (and corresponding unit free RMSPEs). . . . . . . 37

2.6 Map of Illinois with the dots showing the observed locations. The boxes represent the successive blocking regions where we perform pointwise and block kriging. The outermost level of block 1 (lower left) is used to illustrate pointwise predictions. . 38

2.7 Block estimates of the (a) first, (b) second, and (c) third ilr coordinates over the the blocks. Block 1 is represented with dashed line, Block 2 is represented with solid line. Horizontal axis is $|B|$, the area of the block in degrees squared. . . . . . . 39
2.8 Block RMSPEs over (a) block 1 and (b) block 2. Horizontal axis is $|B|$, the area of the block in degrees squared. Dashed lines represent the first ilr coordinate, dotted lines represent the second ilr coordinate, and solid lines represent the third ilr coordinate. .......................................................... 40

3.1 Simulation results comparing CJ (o) and Robust (△) estimates of binned covariance matrix $\hat{\Sigma}_M$ using Frobenius fitting. Plotted value is natural logarithm of expression in (3.11)........ 60

3.2 Simulation results comparing Frobenius (—) and Robust (- - -) fitting when using the Robust estimate of binned covariance matrix. Plotted value is natural logarithm of expression in (3.11).................. 61

3.3 Simulation results comparing CJ estimate with Frobenius fitting (—o—) to the Robust estimate with Robust (---△---) fitting when using the Robust estimate of binned covariance matrix. Plotted value is natural logarithm of expression in (3.11)...... 62

3.4 Plot of observed Cloud Water Path over the spatial domain. ........ 64

3.5 Normal probability plot of the detailed residuals. ................. 65

3.6 Plot of the knot locations of the basis functions over the spatial domain. Asterisks represent the 38 knot locations of the first resolution, and circles represent the 97 knot locations of the second resolution. ................................................. 66
3.7 Plot of (a) predictions and (b) RMSPEs when using the CJ estimate and the Frobenius fit. 67
3.8 Plot of (a) predictions and (b) RMSPEs when using the Robust estimate and the Frobenius fit. 68
3.9 Plot of (a) predictions and (b) RMSPEs when using the Robust estimate and the Robust fit. 69
3.10 Histograms comparing the RMSPE for each of the three combinations of estimation and fitting. (a) CJ estimate, Frobenius fit, (b) Robust Estimate, Frobenius fit, and (c) Robust Estimate, Robust fit. 70
4.1 Plots of the three simulated datasets. 76
4.2 Probability density functions of the (a) Bivariate Normal and (b) Bivariate Laplace distributions. 79
4.3 Diagnostic results from the MCEM procedure. 96
4.4 Shapes of basis functions for four arbitrary knot locations from the coarse simulation. Knots are blue circles, the chosen knot is a red triangle. Locations associated with the chosen knot are small black circles. 98
4.5 Plots of $\hat{Y}$ vs. $Y$ using only the selected knots for the the (a) coarse, (c) medium, and (e) smooth simulated processes. Panels (b), (d), and (f) show the same, when using all of the knots. 100
5.1 Probability plot of detailed residuals of log-transformed daily total ozone maximum data. ........................................... 106

5.2 Normal QQ plot of simulated datasets: (a) $\lambda = 13, \theta^2 = 4, \nu^2 = 0.0001$ (b) $\lambda = 1, \theta^2 = 1, \nu^2 = 0.0001$, (c) $\lambda = 4, \theta^2 = 4, \nu^2 = 1$, (d) $\lambda = 9, \theta^2 = 4, \nu^2 = 1$, (e) $\lambda = 100, \theta^2 = 1, \nu^2 = 1$, and (d) $\lambda = 1, \theta^2 = 4, \nu^2 = 0.0001$. ........................................... 109

5.3 Results of the simulations. Panels (a) and (b) correspond to $W_1$, panels (c) and (d) correspond to $W_2$, and panels (e) and (f) correspond to $W_3$. The left column are histograms of $Y - \hat{Y}$ for test sample. The right column are scatterplots of $Y$ vs $\hat{Y}$. 125

5.4 Map of observed data (log TOM). ................................. 127

5.5 Maps of (a) predictions using Bayesian posterior predictive distribution and (b) batch-means standard errors using the proposed RRSM. ............................... 128

5.6 Maps of (a) predictions using Bayesian posterior predictive distribution and (b) batch-means standard errors using a purely Gaussian model. ............................... 129

5.7 Map of $\nu_G^2 - \nu_W^2$, the location-wise difference of Monte Carlo Standard Errors. ............................... 130
Chapter 1

Introduction

Spatial statistics is a branch of statistics in which data are considered to be spatially correlated. Spatial data may arise from many and varied contexts, including natural resource assessment, environmental and climate monitoring, epidemiology, and public health. This dissertation considers geostatistics in particular, wherein the coordinates of an observation are continuous over some spatial domain, $D$ (as opposed to being fixed regions such as political units, as is the case with lattice data). The theory of geostatistics was laid out by Matheron (1962, 1963 a,b), hence the field is still somewhat recent. Cressie (1993) brought together much of the literature on spatial statistics into a single text which has served as a ubiquitous reference since that time.

One of the most common goals of geostatistical analysis is that of spatial
CHAPTER 1. INTRODUCTION

Measurements of a spatial process

\[ Z = \{ Z(s_i); s_i \in \mathcal{D}, i = 1, \ldots, n \} \]

are taken at point locations \( s_i \), termed observed locations, across \( \mathcal{D} \), and researchers are interested in predicting the value of \( Z(\cdot) \) on unobserved locations. In other words, the geostatistical modeler desires to fill in the blank areas of the map. There are multiple strategies by which to accomplish spatial prediction; two popular choices are kriging, and Bayesian hierarchical models. In a Bayesian analysis, one puts prior distributions on model parameters. Parameter estimation and spatial prediction are generally obtained through Monte Carlo Markov Chain (MCMC) sampling and posterior predictive distributions. On the other hand, kriging is defined as the best linear unbiased predictor. Kriging equations are derived by minimizing some measure of error, typically the mean squared prediction error (MSPE). Any necessary parameters are then estimated and substituted into the kriging equations to obtain spatial predictions.

Using either kriging or Bayesian methods, spatial prediction requires the inverse of \( \Sigma_Z \), the spatial covariance matrix of the data. For \( n \) observed locations, \( \Sigma_Z \) is an \( n \times n \) matrix, and hence \( \Sigma_Z^{-1} \) requires computations of order \( O(n^3) \). When \( n \) becomes greater than several thousands, computation becomes impractical. To address this, reduced rank spatial models (RRSM) have been developed, allowing spatial predictions using massive
datasets without compromising the complexity of the process under analysis. In a nutshell, RRSM is an approach where spatial covariance is modeled in terms of basis functions and a reduced dimensional process defined on a selected number of knot locations. Some of the seminal papers on the topic of RRSM are Banerjee et al. (2008), who approximate the original spatial process with a reduced dimension process defined over a fixed number of knot locations, and Cressie & Johannesson (2008), who provide an approach for exact kriging on large datasets. The same basic model is used in each of the chapters is a mixed effects model where the random component is defined to capture spatial variability (hence Cressie & Johannesson (2008) refer to it as \textit{Spatial Mixed Effects}). The model is described here for reference, each chapter will only describe it to the extent necessary to motivate the problem being addressed.

\[ Z(s) = Y(s) + \epsilon(s), \]  

(1.1)

where \( \epsilon(s) \) are measurement errors, a zero-mean white noise process with variance \( \sigma^2 \). Two mutually independent Gaussian distributions are typically assumed for \( Y(s) \) and \( \epsilon(s) \). Therefore for \( n \) observed locations, \( Z(s) \equiv \{Z(s_1), \ldots, Z(s_n)\} \) is an \( n \)-Gaussian process with mean \( E(Y(s)) = \mu_Y \) and covariance matrix expressed as \( \Sigma_Z = \Sigma_Y + \sigma^2 I_n \), where \( \Sigma_Y \) is the covariance matrix of \( Y(s) \equiv \{Y(s_1), \ldots, Y(s_n)\} \) and \( I_n \) is the identity matrix of rank \( n \).
CHAPTER 1. INTRODUCTION

The fundamental idea of SME is to represent $Y(s)$ as,

$$Y(s) = X(s)\beta + S(s)\eta + \delta(s). \quad (1.2)$$

In this model, $X(s)\beta$ is called large-scale variation; $X(s)$ is a matrix of known covariates and $\beta$ is the associated vector of regression coefficients. Then $S(s)$ is a sparse $n \times r$ matrix of spatially varying basis functions which are also considered known, where $r \ll n$. Various classes of basis functions may be used, for example Shi & Cressie (2007) used W-wavelet basis functions, while Cressie & Johannesson (2008) used bisquare basis functions.

Whatever the choice of basis functions, they should provide a relationship between the observed locations and selected knot locations $S = (S_1, \ldots, S_r)$. The latent process $\eta$ is a zero-mean $r$-dimensional Gaussian process defined on these knot locations, with covariance matrix $V$. Finally, $\delta(s)$, the process error, is an iid zero-mean Gaussian process with variance $\tau^2$. Process error takes into account the variations unexplained by the large scale variations $X(s)\beta$ and spatial process $S(s)\eta$, and uncertainties arising from the dimension reduction. The process and measurement errors are usually assumed to be independent. Note that often the $(s)$ notation is omitted for convenience and readability.

When there is only one observation at each spatial location, $\tau^2$ and $\sigma^2$ are non-identifiable. Instead they are combined as $\nu^2 = \sigma^2 + \tau^2$, called the nugget variance (though indirect means exist to estimate these separately,
Katzfuss & Cressie (2011 a)). Using this modeling framework, $\Sigma_z^{-1}$ can be accessed analytically using the much less computationally-demanding $V^{-1}$ and $1/\nu^2$. This dissertation considers the improves the existing methodologies involved with spatial mixed effects models by providing novel solutions to some commonly encountered problems which have not been addressed before.

In Chapter 2 we develop kriging equations for multivariate log-normal spatial processes of compositional data. We also develop block kriging equations for multivariate log-normal compositional data. Block kriging predicts the average of a spatial process over a selected region of finite area. We apply these equations to a large dataset of coal composition, that is, the variables add to a constant sum. Such data require special consideration, and are best analyzed in isometric log-ratio (ilr) coordinates. Failure to use ilr (or similar transformations) will result in biases to the covariances. Due to the size of the dataset we implement our equations through a RRSM.

In Chapter 3 we investigate the model introduced by Cressie & Johannesson (2008) and note that their suggested method of estimating model parameters and kriging is susceptible to outliers. As an alternative we propose robust, non-parametric estimates of model parameters. Our estimates are shown to possess desirable asymptotic properties. We apply our methods on remote sensing data on cloud properties obtained from NASA. These data cover the whole earth, over 40,000 observations, so a RRSM is necessary. Typically a log-normal distribution is assumed for these data. However, because the variance of a log-normal distribution increases with the mean,
we illustrate our proposed estimates on the log-scale only. The results show that our robust estimates perform uniformly better than those of Cressie & Johannesson (2008).

We also develop Bayesian methods for non-Gaussian data. In Chapter 4, we develop an empirical Bayes method based on the multivariate Laplace distribution. This allows us to model variables with heavier tails than the normal distribution. We implement a Monte Carlo Expectation-Maximization (MCEM) algorithm to estimate model parameters. In addition to this, we use a Bayesian penalized regression to select necessary knots for the spatial domain under analysis.

In Chapter 5, we develop a fully Bayesian model for data exhibiting non-Gaussian tails. Instead of assuming a Gaussian distribution on the reduced dimensional process, we use scale mixtures of Gaussian distributions, where the scale parameters follow an exponential distribution with parameter 1. This modification adds enough flexibility for the model to handle data with several non-Gaussian tail behaviors. Predictions are obtained through an MCMC algorithm.

Finally, in Chapter 6 we discuss how our models (from chapter 5 especially) can be extended to consider dependencies in space and time instead of only in space.
Chapter 2

Multivariate Log-normal Kriging and Block Kriging with Applications to Coal Geology

2.1 Introduction

Coal is an important natural resource for many industrial sectors; in 2011 in the United States, 42% of electricity was generated by coal power plants (USEIA, 2012). In addition to generating electricity, coal is also used for making steel, and a variety of other common items including fertilizer, tar, and plastics. Coal by-products are used by many pharmaceutical companies. Coal is classified into four common ranks: anthracite, bituminous, sub-bituminous, and lignite. Bituminous and sub-bituminous are the most
common rank of coal found in the US, accounting for over 90% of its production. As the rank of coal increases, carbon content increases and moisture decreases. Anthracite coal has the highest carbon content, and lignite coal has the lowest.

To assess the rank of coal over a region, samples of coal are obtained and several measurements are made, including percent moisture, percent fixed carbon, percent volatile material, and percent ash. Samples are usually obtained at point locations, so we would like to develop procedures which can characterize the rank of coal over a region of interest by predicting at unobserved locations from the available observations. These measurements are referred in literature as *compositional data*, which is to say that they are percents of a whole and thus inherently related (Aitchison, 2003). Compositional data require special consideration because the sum of the $K$ components is a defined constant (i.e. they are in a $K$–part simplex). When data are compositional, traditional correlations are often spurious because each component carries only relative information about the composition. This feature has been extensively investigated (e.g. Chayes, 1960, 1962, 1971, 1983; Pawloswsky, 1984). To address this, compositional data are analyzed through an Aitchison transformation such as the additive log-ratio (alr), centered log-ratio (clr), or isometric log-ratio (ilr). If desired, the results are easily back-transformed to the simplex. Much research has been done regarding compositional data analysis to address the constant-sum constraint and develop the methodology for properly analyzing data of this sort. For
Kriging is very popular in the field of geostatistics for spatial prediction at the unobserved locations based on the available data. Cressie (1993), Chapter 3 proved that kriging is a best linear unbiased predictor (BLUP). In addition to prediction, kriging equations also provide root mean squared prediction errors (MSPE) as a measure of precision. See Cressie (1993), Stein (1999), and Webster & Oliver (2007) for discussion on kriging. For compositional data, analyses are carried out in log-ratio coordinates. Simply taking log-transformation of the ratios and transforming the kriging estimates back to original scale will not provide unbiased estimates with respect to the probability measure of a log-normal distribution. Cressie (2006) provides an useful discussion on the optimal log-normal kriging equations presented in Matheron (1974). The kriging estimates obtained from optimal log-normal kriging equations are unbiased. While punctual (pointwise) kriging enables one to complete the map by filling unobserved locations, block kriging seeks average prediction over a region of finite area. The theory of log-normal block kriging was first established in an unpublished article by Matheron (1974). Further details can be found in Rivoirard (1990), Cressie (1993) pp. 136, and Cressie (2006). Optimal log-normal block kriging requires numerical quadrature; due to this heavy computational overhead, log-normal block kriging was carried out in the past under the assumption of “permanence
of log-normality” by adjusting for the bias using a corrector term. Cressie (2006) revisited the topic and showed that the estimates obtained under this assumption are inferior. Now, in modern computing environment, it is easy to solve integrals numerically. Thus obtaining optimal block predictions for log-normal processes is feasible. The kriging estimates obtained from optimal log-normal block kriging equations are unbiased and do not require any corrective term. But the existing papers only address issues on univariate spatial processes. In this chapter, we extend the idea of optimal log-normal block kriging equations to obtain the estimates and the associated root mean squared errors for multivariate spatial processes of compositional data. Our equations are flexible enough to handle imprecise observations, i.e., when the observations come with some measurement errors.

In geostatistics, a common strategy of specifying the spatial dependence is through variogram (Cressie, 1993). Modeling multivariate spatial processes also requires the specification of spatial dependence across variables. Cross-variograms can be used though are not recommended due to their limitations in capturing cross-dependence (Verhoef, Cressie, & Barry, 2004). Our kriging equations are derived in terms of covariance matrices and do not require fitting cross-variograms to data. There is an extensive body of literature which addresses how to use spatial moving average method for modeling flexible cross-covariance functions for multivariate spatial processes (Verhoef, Cressie, & Barry, 2004; Majumdar & Gelfand, 2007; Higdon, 2002). We use the similar idea while applying our log-normal kriging equations to mu-
tivariate spatial data.

In our analysis we consider four geochemical components of coal: percent moisture, percent volatile material, percent fixed carbon, and percent ash. The data were obtained from the Illinois State Geological Survey (ISGS) which consist of observations at 3957 locations (after data cleaning) throughout the state of Illinois, USA. Due to this large sample size (3957 for each of three ilr coordinates), a straightforward kriging will demand inverse of a large covariance matrix. Cressie & Johannesson (2008) and Shi & Cressie (2007) overcome this difficulty for univariate spatial variables by spatial mixed effects (SME) model through spatially varying basis function matrix. We extend their method under the situation of block predictions for multivariate spatial processes by estimating the spatial and cross-spatial dependence.

Our model predicts the isometric log-ratios at specific point locations, as well as their average over a specified region of finite area (termed blocks). We also calculate measures of precision to assess the goodness of the predictions on point locations and also over blocks. Figure (2.6) shows the map of the state of Illinois with the black dots denoting the locations from where data on the geochemical makeup of coal were obtained. Naturally, mining cannot occur at every point in the spatial domain (e.g. in a city) and so our goal is not to predict over the entire state of Illinois, but on particular regions of interest. As an illustration of our methodology, we will predict at point locations inside the largest level of block 1 (lower left) in Figure (2.6). To illustrate block kriging, we predict over all four levels of each block and show
that the precision of estimation increases with the size of the block.

The main contributions of this chapter are the derivation of kriging and block kriging estimates and MSPEs for multivariate log-normal processes, with special attention to compositional data. By pursuing a log-normal approach, we obtain estimates which are unbiased for both the original scale of the data, and for the log-scale (as we use in this paper). The block kriging allows estimation over a finite region with greater precision than at point locations. In our multivariate SME modeling, spatial and cross-spatial dependencies are characterized using a flexible class of basis functions which can account for non-stationarity and anisotropy.

Section 2.2 describes the geostatistical model and multivariate log-normal kriging equations in presence of measurement errors, and how to obtain the estimates. Section 2.3 extends these equations to log-normal block kriging. In Section 2.4, we address the computational issues, explaining in detail the spatial mixed effects model and how to implement log-normal block kriging on a large multivariate spatial dataset by estimating spatial and cross-spatial dependence. In Section 2.5, we apply our technique to the coal data. Section 2.6 provides conclusions and a brief discussion. Note that this chapter is a author accepted manuscript version of Jelsema & Paul (2013).

2.2 Multivariate Log-normal Kriging

Pawloswsky-Glahn & Olea (2004), Chapter 5 discussed briefly multivariate
log-normal kriging. Here we derive the log-normal kriging equations using empirical Bayesian approach. This approach will help us later to obtain the log-normal block kriging estimates and mean squared prediction errors explicitly. First we will develop the log-normal kriging equations, and explain how to apply these equations to compositional data. For clarity, in our notation vectors and matrices are represented in boldface.

We start with some notations. Since we will assume a log-normal distribution, these change slightly from that presented in equations (1.1) and (1.2). Let \( D \) be a spatial domain in \( \mathbb{R}^2 \) with positive area \( |D| \) and let \( Z(s) = [Z_1(s), \ldots, Z_K(s)]' \) be a \( K \)-variate log-normal spatial process over \( D \). Then \( Y(s) \), where \( Z(s) = \exp\{Y(s)\} \), is a multivariate Gaussian process. We model \( Y(s) \) as:

\[
Y_k(s) = W_k(s) + \epsilon_k(s); \quad k = 1, \ldots, K, \tag{2.1}
\]

where \( \epsilon_k(s) \) is a zero-mean Gaussian white noise (measurement errors) with variance \( \sigma_k^2 \), and \( W_k(s) \) is the measurement-error-free latent process (truth), which is independent of measurement errors and follows a Gaussian distribution with mean \( \mu_{W_k}(s) \) and variance \( C_{W_k}(s,s) \). Hence, we can write \( Z_k(s) \equiv T_k(s) \exp\{\epsilon_k(s)\} \), where \( T_k(s) = \exp\{W_k(s)\} \). In our interpretation, \( Z_k(s) \) is the observed \( k^{th} \) process and \( T_k(s) \) is the true \( k^{th} \) process at location \( s \) with multiplicative error \( \exp\{\epsilon_k(s)\} \). We want to predict \( T(s) \equiv \{T_k(s), k = 1, \ldots K\} \), the measurement-error-free processes us-
CHAPTER 2. LOG-NORMAL KRIGING AND BLOCK KRIGING

ing available data. Note that, $T(s)$ is a $K$-variate log-normal distribution with mean vector $E\{T_k(s)\} = [\mu'_{T_1}(s), \ldots, \mu'_{T_K}(s)]'$, where

$$
\mu_{T_k}(s) = \exp \{\mu_{W_k}(s) + 0.5 C_{W_k}(s, s)\}.
$$

The covariance matrix of $K$ processes on two locations $u$ and $v$ is a $2K \times 2K$ positive-definite matrix with the following entries:

$$
C_{T_i T_j} (u, v) = \text{cov} \left( T_i(u), T_j(v) \right) \text{ for } i, j = 1, \ldots, K
$$

$$
= \mu_{T_i}(u) \mu_{T_j}(v) \left[ \exp \left\{ C_{W_i W_j} (u, v) \right\} - 1 \right],
$$

where $C_{W_i W_j} (u, v) = \text{cov} \left( W_i(u), W_j(v) \right)$.

Let us assume here that for the $k^{th}$ process we have observations from $n_k$ locations, $(s_{k1}, \ldots, s_{kn_k})$, in domain $D$. In total, we have $N = \sum_{k=1}^{K} n_k$ observations. Thus our observations constitute an $N \times 1$ vector $Z$, or $Y$ in log-scale. $Y$ follows a multivariate Gaussian distribution with mean $\mu_W = E(W)$ and covariance matrix $\Sigma_Y$, which can be partitioned in blocks as:

$$
\begin{pmatrix}
\Sigma_{n_1 \times n_1} + \sigma_1^2 I_{n_1} & \Sigma_{n_1 \times n_2} & \ldots & \Sigma_{n_1 \times n_K} \\
\ldots & \ldots & \ldots & \ldots \\
\Sigma_{n_K \times n_1} & \Sigma_{n_K \times n_2} & \ldots & \Sigma_{n_K \times n_K} + \sigma_K^2 I_{n_K}
\end{pmatrix}.
$$

In the above matrix, $\Sigma_{n_i \times n_j} = \text{cov}(W_i, W_j)$, an $n_i \times n_j$ matrix and $I_r$ is an identity matrix of rank $r$. Log-normal kriging equations provide the un-
biased estimators, \( p(T(s_0); Z) \) of \( K \) processes at unobserved location \( s_0 \) by minimizing the trace of conditional mean squared error loss matrix

\[
E((T(s_0) - p(T(s_0)))(T(s_0) - p(T(s_0)))' | Z). \]

Here, the minimum mean squared predictor will be the conditional expectation of \( T(s_0) \) given data \( Z \), \( E(T(s_0) | Z) \). We can find an explicit expression for this conditional expectation as:

\[
E(T(s_0) | Z) = \begin{bmatrix}
E(\exp\{W_1(s_0)\} | Y) \\
\vdots \\
E(\exp\{W_K(s_0)\} | Y)
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\exp\{\mu_{W_1}(s_0) + C_1'(s_0)\Sigma^{-1}_Y(Y - \mu_W) + 0.5 C_{W_1}(s_0, s_0) - 0.5 C_1'(s_0)\Sigma^{-1}_Y C_1(s_0)\} \\
\vdots \\
\exp\{\mu_{W_K}(s_0) + C_K'(s_0)\Sigma^{-1}_Y(Y - \mu_W) + 0.5 C_{W_K}(s_0, s_0) - 0.5 C_K'(s_0)\Sigma^{-1}_Y C_K(s_0)\}
\end{bmatrix}, \quad (2.2)
\]

where \( C_k(s_0) \) is an \( N \times 1 \) vector of \( \text{cov}(W_k(s_0), Y) \). In practice, the mean and covariance functions are unknown and need to be estimated from data. In geostatistics, it is a common practice to model \( \mu_{W_k}(s) \) as a linear function of \( p_k \) covariates, \( x_k(s) \), and regression coefficients \( \beta_k \) as, \( x_k(s)\beta_k \). If we plug-in the generalized least square estimates of \( \mu_{W_k}(s_0) \) and \( \mu_W \) in (2.2), we get the
log-normal kriging estimate $\hat{T}(s_0)$ as:

$$
\hat{T}(s_0) = \begin{bmatrix}
\exp \{ A_1(s_0)Y + B_1(s_0) \} \\
\vdots & \vdots & \vdots \\
\exp \{ A_K(s_0)Y + B_K(s_0) \}
\end{bmatrix},
$$

(2.3)

where $A_k(s_0) = \{ [0_{1\times p_1}, \ldots, x_k(s_0)', \ldots, 0_{1\times p_K}] - C_k'(s_0)\Sigma_Y^{-1}X \} (X'\Sigma_Y^{-1}X)^{-1} X'\Sigma_Y^{-1} + C_k'(s_0)\Sigma_Y^{-1}$ and $B_k(s_0) = 0.5 C_W(s_0, s_0) - 0.5 C_k(s_0)\Sigma_Y^{-1} C_k(s_0)$. $0_{l\times q}$ is a $l \times q$ matrix of of zeros. $X$ denotes the matrix of all $\sum_{k=1}^K p_k$ covariates for all $N$ observed locations. We estimate $C_k(\cdot)$ and $\Sigma_Y$ from data, but while deriving the minimized trace of MSPE matrix, we will assume that they are known. Thus, our kriging estimates will be empirical Bayes estimates. The MSPE matrix is a $K \times K$ matrix with the diagonal terms

$$
E \left( T_k(s_0) - \hat{T}_k(s_0) \right)^2,
$$

and off-diagonal terms,

$$
E \left( (T_i(s_0) - \hat{T}_i(s_0))(T_j(s_0) - \hat{T}_j(s_0)) \right).
$$

Explicit expressions for these two terms are derived below:

$$
E \left( T_k(s_0) - \hat{T}_k(s_0) \right)^2 =
$$
\[ = \exp \left\{ 2\mu_{W_k}(s_0) + 2C_{W_k}(s_0, s_0) \right\} + \exp \left\{ 2A_k(s_0)\mu_W + 2A_k(s_0)\Sigma_Y A_k(s_0)' \right. \]
\[ + 2B_k(s_0) \} - 2 \exp \left\{ \mu_{W_k}(s_0) + 0.5C_{W_k}(s_0, s_0) + A_k(s_0)\mu_W + 0.5A_k(s_0)\Sigma_Y A_k(s_0)' + A_k(s_0)C_k(s_0) + B_k(s_0) \right\} \]
\[ = \exp \left\{ \mu_{W_i}(s_0) + 0.5C_{W_i}(s_0, s_0) + \mu_{W_j}(s_0) + 0.5C_{W_j}(s_0, s_0) + C_{W_iW_j}(s_0, s_0) \right\} \]
\[ - \exp \left\{ \mu_{W_i}(s_0) + 0.5C_{W_i}(s_0, s_0) + A_i(s_0)\mu_W + 0.5A_i(s_0)\Sigma_Y A_i(s_0)' + A_i(s_0)C_i(s_0) + B_i(s_0) \right\} - \exp \left\{ \mu_{W_j}(s_0) + 0.5C_{W_j}(s_0, s_0) + A_j(s_0)\mu_W + 0.5A_j(s_0)\Sigma_Y A_j(s_0)' + A_j(s_0)C_j(s_0) + B_j(s_0) \right\} + \exp \left\{ (A_i(s_0) + A_j(s_0))\mu_W + 0.5(A_i(s_0) + A_j(s_0))\Sigma_Y (A_i(s_0) + A_j(s_0))' \right. \]
\[ + B_i(s_0) + B_j(s_0) \right\} \]

In practice, estimates of (2.4) and (2.5) can be obtained by replacing \( \mu_{W_k}(s) \) and \( \mu_W \) by their generalized least square estimates.

**Multivariate Log-normal Kriging for Compositional Data**

Let \( \{U(s), s \in D\} \) denote a vector of compositional data with \( K+1 \) components, so that for every location \( s \in D, \sum_{k=1}^{K+1} U_k(s) = \kappa \). For convenience we take \( \kappa = 1 \), but compositional data may sum to any positive constant. Due to this constant-sum constraint, each of the \( (K+1) \) components carry
only relative information. Analysis of compositional data typically considers isometric log-ratio (ilr) coordinates (Egozcue et al., 2003). The ilr transformation and its inverse are defined by:

\[
\text{ilr}(x) = \ln(x)E \\
\text{ilr}^{-1}(x) = \exp \left(xE'\right).
\]

The matrix \(E\) is a \((K + 1) \times K\) orthonormal basis matrix satisfying \(E'E = I_{K-1}\) and \(EE' = I_K - K^{-1}1_{K \times K}\), where \(1_{K \times K}\) is a \(K \times K\) matrix of ones.

For example, a 4-part composition may have the basis matrix,

\[
E = \begin{bmatrix}
\frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \\
\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{12}} \\
0 & \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{12}} \\
0 & 0 & \frac{3}{\sqrt{12}}
\end{bmatrix}, \tag{2.6}
\]

and ilr coordinates given by,

\[
\text{ilr}(U)' = \begin{bmatrix}
\frac{1}{\sqrt{2}} (\ln U_2 - \ln U_1) \\
\frac{1}{\sqrt{6}} (2 \ln U_3 - \ln U_1 - \ln U_2) \\
\frac{1}{\sqrt{12}} (3 \ln U_4 - \ln U_1 - \ln U_2 - \ln U_3)
\end{bmatrix}.'
Alternatively, we may consider the ratios,

\[
\left[ \left( \frac{U_2}{U_1} \right)^{1/\sqrt{2}}, \left( \frac{U_3}{U_1U_2} \right)^{1/\sqrt{6}}, \left( \frac{U_4}{U_1U_2U_3} \right)^{1/\sqrt{12}} \right],
\]  

(2.7)

and express the ilr (\(U\)) in terms of the componentwise natural log of the ratios,

\[
\left[ \ln \left( \frac{U_2}{U_1} \right)^{1/\sqrt{2}}, \ln \left( \frac{U_3}{U_1U_2} \right)^{1/\sqrt{6}}, \ln \left( \frac{U_4}{U_1U_2U_3} \right)^{1/\sqrt{12}} \right].
\]  

(2.8)

The ratios in (2.7) are positive real numbers on \(\mathbb{R}^+_K\) and the log-ratios (ilr) are real numbers on \(\mathbb{R}^K\). If we assume a multivariate Gaussian distribution on (2.8), then that would induce a multivariate log-normal distribution on (2.7). The methodology described earlier can be used to obtain predictions and MSPEs for both the ratios and the ilr coordinates.

The general definition for unbiasedness of a composition \(\hat{U}\) is

\[
\ilr^{-1} \left( E \left[ \ilr(\hat{U}) - \ilr(U) \right] \right) = \mathcal{C} \left( 1, 1, \ldots, 1 \right),
\]

where \(\mathcal{C}\) is the closure operation, \(\mathcal{C}(x) = \kappa x / \sum_{i=1}^{K+1} x_i\). By this definition, for a prediction \(\hat{U}\), if \(E \left[ \ilr(\hat{U}) - \ilr(U) \right] = [0, 0, \ldots, 0]\) then \(\hat{U}\) is unbiased. Then we note that, since our multivariate log-normal kriging estimates are unbiased in the log-scale, this expectation is a vector of zeros. Therefore our predictions are unbiased by the compositional definition.
2.3 Multivariate Log-normal Block Kriging

Recall the notations from previous section, $T(\cdot)$ is the true $K$-variate log-normal spatial process. Our interest is in predicting $T(\cdot)$ and then $U(\cdot)$ over a block $B$ of finite area $|B|$, where $B \in \mathcal{D}$, i.e. to find an optimal predictor $\hat{T}(B)$ for:

$$T(B) = \frac{1}{|B|} \begin{bmatrix} \int_B T_1(u) du \\ \vdots \\ \int_B T_K(u) du \end{bmatrix}.$$  \hspace{1cm} (2.9)

We obtain $\hat{T}(B)$ by minimizing the trace of conditional MSPE matrix,

$$E \left( (T(B) - \hat{T}(B))(T(B) - \hat{T}(B))' | \mathbf{Z} \right).$$

This predictor can be computed by calculating the integral:

$$\hat{T}(B) = \frac{1}{|B|} \begin{bmatrix} \int_B \hat{T}_1(u) du \\ \vdots \\ \int_B \hat{T}_K(u) du \end{bmatrix},$$  \hspace{1cm} (2.10)

where $\hat{T}_k(u)$ is the optimal predictor of $T_K(u)$, given in Equation (2.3). The MSPE matrix is a $K \times K$ matrix with the diagonal terms $E \left( T_k(B) - \hat{T}_k(B) \right)^2$ and off-diagonal terms $E \left( (T_i(B) - \hat{T}_i(B))(T_j(B) - \hat{T}_j(B)) \right)$. We obtain ex-
plicit expressions for these two terms below.

\[ E\left( T_k(B) - \hat{T}_k(B) \right)^2 = \frac{1}{|B|^2} \int_B \int_B E\left( (T_k(u) - \hat{T}_k(u))(T_k(v) - \hat{T}_k(v)) \right) du dv. \]

(2.11)

The integrand in (2.11) can be written as:

\[ E\left( T_k(u)T_k(v) \right) - E\left( T_k(u)\hat{T}_k(v) \right) - E\left( \hat{T}_k(u)T_k(v) \right) + E\left( \hat{T}_k(u)\hat{T}_k(v) \right), \]

(2.12)

where

\[
E\left( T_k(u)T_k(v) \right) = E\left( \exp\left\{ W_k(u) + W_k(v) \right\} \right) \\
= \exp\left\{ \mu_{W_k}(u) + \mu_{W_k}(v) + 0.5C_{W_k}(u,u) + 0.5C_{W_k}(v,v) \\
+ C_{W_kW_k}(u,v) \right\},
\]

(2.13)

\[
E\left( \hat{T}_k(u)T_k(v) \right) = E\left( \exp\left\{ A_k(u)Y + B_k(u) + W_k(v) \right\} \right) \\
= \exp\left\{ A_k(u)\mu_Y + \mu_{W_k}(v) + 0.5A_k(u)\Sigma_Y A_k(u)' + 0.5C_{W_k}(v,v) \\
+ A_k(u)C_k(v) + B_k(u) \right\},
\]

(2.14)

and

\[
E\left( \hat{T}_k(u)\hat{T}_k(v) \right) = E\left( \exp\left\{ A_k(u)Y + B_k(u) + A_k(v)Y + B_k(v) \right\} \right)
\]
\begin{equation}
\exp \{ (A_k(u) + A_k(v)) \mu_w + 0.5(A_k(u) + A_k(v)) \Sigma_Y (A_k(u) + A_k(v))' \\
+ B_k(u) + B_k(v) \}.
\end{equation}

The off-diagonal terms in the MSPE matrix will be:

\begin{equation}
E \left( (T_i(B) - \hat{T}_i(B))(T_j(B) - \hat{T}_j(B)) \right) \\
= \frac{1}{|B|^2} \int_B \int_B E \left( (T_i(u) - \hat{T}_i(u))(T_j(v) - \hat{T}_j(v)) \right) dudv.
\end{equation}

The integrand in (2.16) can be written as:

\begin{equation}
E(T_i(u)T_j(v)) - E(T_i(u)\hat{T}_j(v)) - E(\hat{T}_i(u)T_j(v)) + E(\hat{T}_i(u)\hat{T}_j(v)),
\end{equation}

where

\begin{align}
E(T_i(u)T_j(v)) &= E(\exp \{ W_i(u) + W_j(v) \}) \\
&= \exp \{ \mu_{W_i}(u) + \mu_{W_j}(v) + 0.5C_{W_i}(u,u) + 0.5C_{W_j}(v,v) \\
&\quad + C_{W_iW_j}(u,v) \},
\end{align}

\begin{equation}
E(\hat{T}_i(u)T_j(v)) = E(\exp \{ A_i(u)Y + B_i(u) + W_j(v) \})
\end{equation}

\begin{align}
&= \exp \{ A_i(u)\mu_w + \mu_{W_j}(v) + 0.5A_i(u)\Sigma_Y A_i(u)' + 0.5C_{W_j}(v,v) \\
&\quad + A_i(u)C_j(v) + B_i(u) \},
\end{align}
and

\[
E\left( \hat{T}_i(u)\hat{T}_j(v) \right) = E\left( \exp \{ A_i(u)Y + B_i(u) + A_j(v)Y + B_j(v) \} \right) \\
= \exp \{ (A_i(u) + A_j(v))\mu_w + 0.5(A_i(u) + A_j(v))\Sigma_Y (A_i(u) + A_j(v))' + B_i(u) + B_j(v) \}. 
\] (2.20)

The integrals in Equations (2.10), (2.11), and (2.16) will be computed by approximating with finite sums. To the best of our knowledge, nobody before us derived these expressions for multivariate log-normal block kriging.

For compositional data in ilr coordinates, block predictions are obtained in a similar way. First we compute ilr coordinates for every element in the approximating sums, and then may obtain the average of each ilr coordinate. For block MSPEs, we likewise obtain the MSPEs in ilr coordinates before approximating the integrals with sums.

### 2.4 Implementation for Large Datasets

The point-wise kriging estimates in Equations (2.2) and the block-kriging estimates in Equations (2.10) require inversion of \( N \times N \) covariance matrix, \( \Sigma_Y \). This is a computation of complexity \( O(N^3) \). When \( N \) is large, this will be a demanding computational overhead. Cressie & Johannesson (2008) recommends *spatial mixed effects* (SME) model to overcome this hurdle. They
analyze univariate spatial process. We extend their method to multivariate log-normal spatial processes in this section. We write $W_k$ as a sum of a deterministic and a stochastic term as:

$$W_k = \mu_k + S_k \eta_k + \delta_k; \quad k = 1, \ldots, K,$$

where each term in $\mu_k$ is $\mu_k(s) = x_k(s)\beta_k$, a deterministic function of covariates. $S_k$ is an $n_k \times r_k$ matrix of spatially varying basis/kernel functions, $S_k(s)$. $\eta_k$ is a zero-mean $r_k$-dimensional Gaussian process on selected knot locations, $S_k = (S_{k1}, \ldots, S_{kr_k})$, with covariance matrices $\text{cov}(\eta_k, \eta_k) = V_{kk}$ and $\text{cov}(\eta_k, \eta_j) = V_{kj}$. $\delta_k(\cdot)$ is another iid zero-mean Gaussian process with variance $\tau_k^2$. $\delta_k(\cdot)$ can be interpreted as process error which takes into account the errors due to dimension reduction.

In their analysis of global ozone data, Cressie & Johannesson (2008) assumed that the measurement-error variance is known and only estimate this process-error variance. Estimating both $\sigma_k^2$ and $\tau_k^2$ is not possible as they are not identifiable. What we can estimate is $\nu_k^2 = \sigma_k^2 + \tau_k^2$. $\nu_k^2$ is called nugget effect. If we have more than one observation on certain or on all locations, then we can estimate $\sigma_k^2$ easily (e.g., Paul & Cressie, 2011). But in most situations, data come without specific knowledge of the measurement-error variance and we have only one observation per location. Katzfuss & Cressie (2011a) suggested an empirical method. But their method underestimates the actual measurement error variance. In fact, one can treat the signal-to-
noise ratio, $\tau_k^2/\sigma_k^2$, as a tuning parameter and monitor the effect of its value on predictions and root mean squared prediction errors.

Selection of knot locations is crucial for the success of SME models. In fact, this is an ongoing investigation in two dimensional spatial design. While there is no specific rule for selecting $r_k$ and the knot locations $S_k$, Banerjee et al (2008) offer some suggestions. If the dataset is spatially dense, then usually one selects the knot locations that cover uniformly the entire area under study. If observed locations are not dense, like our datasets in this study, then one should keep in mind few things: First, select more knots near the observed locations. Second, select fewer knots where the process exhibits spatial smoothness and the spatial correlation is strong. Third, select more knots in the coarser regions where spatial correlation is weak. The number of knots, $r_k$ is often selected by monitoring Akaike Information Criterion (AIC), see Fuentes (2001). Using too few knot locations is not recommended, as model may fail to capture the small-scale variations efficiently (e.g., Rodrigues & Diggle, 2010). Our recommendation is to start with a fewer number of knots and increase them until there is no significant improvement in mean squared prediction errors. Under this representation of SME model, the spatial and cross-spatial covariances will be:

\[
\text{cov}(W_k(u), W_k(v)) = S_k(u)V_{kk}S_k(v)' + \tau_k^2 I(u = v)
\]

\[
\text{cov}(W_i(u), W_j(v)) = S_i(u)V_{ij}S_j(v)'
\]
Here $I(A)$ is an indicator function which takes the value 1 if $A$ occurs and 0 otherwise.

Our multivariate SME model is flexible enough to combine different spatial processes of varying degrees of smoothness. If $K$ processes under consideration exhibit different degrees of smoothness, then one can vary the number and the locations of the knots for each process individually.

The covariance matrix, $\Sigma_Y$ can be written as:

$$\Sigma_Y = SVS' + D,$$  \hspace{1cm} (2.22)

where

$$V = \sum_{k=1}^{K} r_k \times \sum_{k=1}^{K} r_k$$

is a positive definite covariance matrix of the vector $\eta = [\eta_1', \ldots, \eta_k']$. $D$ is a block-diagonal matrix:

$$D = \begin{bmatrix}
\sigma_1^2 I_{n_1} & 0_{n_1 \times n_2} & \cdots & 0_{n_1 \times n_K} \\
0_{n_2 \times n_1} & \sigma_2^2 I_{n_2} & \cdots & 0_{n_2 \times n_K} \\
\vdots & \ddots & \ddots & \vdots \\
0_{n_K \times n_1} & \cdots & 0_{n_K \times n_{K-1}} & \sigma_K^2 I_{n_K}
\end{bmatrix}.$$  \hspace{1cm} (2.24)

Cressie & Johannesson (2008) noted the inverse of $\Sigma_Y$ can be obtained
by Sherman-Woodbury-Morrison formula:

\[
\Sigma_Y^{-1} = D^{-1} - D^{-1} S \{ V^{-1} + S'D^{-1}S \}^{-1} S'D^{-1}.
\] (2.25)

Note that (2.25) requires the inversion of much smaller dimensional matrices compared to \( \Sigma_Y \) and inverse of \( D \), which is a diagonal matrix and its inverse is available analytically.

There are various choices for the matrix \( S \). One can also select different basis functions for different processes while analyzing multivariate spatial datasets. Which one fits the data best is still an ongoing research, but interested readers can see Johannesson & Cressie (2004) for a discussion. One can also monitor the spatial covariance structures by fitting local variogram models to the subset of the data and select the basis function, see Verhoef, Cressie, & Barry (2004). For fitting the model to data, we consider that \( S \) matrix is completely known. Now the task is to estimate \( V \) and \( D \) from data. There are two common approaches for this estimation - (1) Method of Moments (MOM) and (2) EM Algorithm (Dempster, Laird, & Rubin, 1977). Cressie & Johannesson (2008) used method of moments (MOM) for univariate spatial process. But the MOM estimates may not be positive definite and Kang, Cressie, & Shi (2010) adjusted by lifting the eigenvalues of the matrices while preserving the total variability. Katzfuss & Cressie (2011 a) and Katzfuss & Cressie (2011 b) used EM algorithm for univariate spatial and spatio-temporal processes. A drawback of the EM algorithm is that it
depends heavily upon the goodness of the likelihood assumption. If the data deviate from the specified probabilistic model (in this case, normality), EM can lead to very poor estimates of the model parameters. But departures from normality do not harm the fact that kriging estimates are BLUE, thus MOM estimation has its own importance as a distribution-free estimation technique. For this reason we chose to use MOM estimates in our analysis. For details on both these methods, see the discussion on univariate spatial process in Katzfuss & Cressie (2011 a).

**Method of Moments:** First we compute the detailed residuals, $\tilde{R}_k = Y_k - X_k \tilde{\beta}_k$, where $\tilde{\beta}_k$ is ordinary least squares estimate of $\beta_k$, $(X_k'X_k)^{-1}X_k'Y_k$. We divide the entire region of interest in $M$ bins, where $M$ is larger than $\max(r_k)$ but much smaller than $\min(N_k)$. Katzfuss & Cressie (2011 a) recommends that for numerical stability, one should select the bins such that each bin has at least 15 observations. Then, we compute the averages of detailed residuals and the averages of squared of detailed residuals falling in each bin. An empirical estimate of the binned covariance matrix, $\hat{\Sigma}_M$, is obtained by using these averages. We can write $\hat{\Sigma}_M$ in blocks as:

$$
\hat{\Sigma}_M = \begin{bmatrix}
\hat{\Sigma}_{11} & \hat{\Sigma}_{12} & \cdots & \hat{\Sigma}_{1K} \\
\vdots & \ddots & \vdots \\
\hat{\Sigma}_{K1} & \hat{\Sigma}_{K2} & \cdots & \hat{\Sigma}_{KK}
\end{bmatrix}
$$

(2.26)

Note that here $\hat{\Sigma}_M$ is a $KM \times KM$ matrix and each of the block matrices on right hand side of (2.26) is $M \times M$ matrix. For $\hat{\Sigma}_{kk}$, the $m^{th}$ diagonal
element is the average of the squared detailed residuals for \( k^{th} \) process falling in \( m^{th} \) bin and the \((m, m')\) off-diagonal element is the product of the averages of the detailed residuals for \( k^{th} \) process falling in \( m^{th} \) and \( m'^{th} \) bin. For the off-diagonal block matrices \( \hat{\Sigma}_{kk'} \), the \((m, m')\) element is the product of the averages of the detailed residuals for \( k^{th} \) and \( k'^{th} \) processes falling in \( m^{th} \) and \( m'^{th} \) bin, respectively. A binned version of \( S \) matrix is computed by taking the column averages of the rows of \( S \) matrix correspond to observations falling in each bin. We denote this by \( \overline{S} \). Following the similar notation as (2.23) we can write \( \overline{S} = diag(\overline{S}_1, \ldots, \overline{S}_K) \). Q-R decomposition of \( \overline{S} \) matrix will be: \( \text{diag}(Q_1L_1, \ldots, Q_KL_K) \), where \( Q_k \) is \( M \times r_k \) orthonormal matrix and \( L_k \) is \( r_k \times r_k \) non-singular upper triangular matrix. Similarly to \( \overline{S} \), we compute a binned version of \( D \) matrix as \( \overline{D} = \text{diag}(\nu^2_1I_M, \ldots, \nu^2_KI_M) \). We estimate \( V \) and \( \nu^2_k \)'s in two steps by minimizing the Frobenius norm between \( \hat{\Sigma}_M \) and \( \Sigma_M = \overline{S}V\overline{S}' + \overline{D} \). Note that \( \Sigma_M \) is a function of the unknown quantities which we want to estimate. \( \nu^2_k \) can be estimated by minimizing \( \sum^M_{i,j} \left\{ (\hat{\Sigma}_{kk} - Q_kQ_k'\hat{\Sigma}_{kk}Q_kQ_k')_{i,j} - \nu^2_k(I_M - Q_kQ_k'I_MQ_kQ_k')_{i,j} \right\} \) with respect to \( \nu^2_k \). Computationally, this is just a linear regression with zero intercept and slope \( \nu^2_k \). Once we estimate all the \( \nu^2_k \)'s, we estimate \( V \) by:

\[
\hat{V} = L^{-1}Q'(\hat{\Sigma}_M - \text{diag}(\hat{\nu}^2_1I_M, \ldots, \hat{\nu}^2_KI_M))Q(L^{-1})',
\]

where \( L^{-1} = \text{diag}(L_1^{-1}, \ldots, L_K^{-1}) \) and \( Q = \text{diag}(Q_1, \ldots, Q_K) \). The positive-
definiteness of $\hat{V}$ depends on the positive-definiteness of

$$G = (\hat{\Sigma}_M - \text{diag}(\hat{\nu}_1^2 I_M, \ldots, \hat{\nu}_K^2 I_M)).$$

If $G$ is not positive-definite, then we have to “lift” the eigenvalues of $G$, at the same time preserving the total variability, $\text{trace}(G)$. For details, see Kang and Cressie (2010).

### 2.5 Application to Coal Data

By modeling the geochemical properties of coal, the rank of coal may be assessed over a broad region. The dataset under analysis consists of observations at 3957 locations throughout the state of Illinois, obtained through the Illinois State Geological Survey (ISGS). $U_1(s), \ldots, U_4(s)$ denote the percents moisture, volatile material, fixed carbon, and ash, respectively. From these we compute the ilr coordinates, which can be seen in figure (2.1).

We obtained $E$, through the R package compositions. The function $\text{ilrBase}$, based on the work of Egozcue et al (2003), computes a triangular Helmert matrix. Recall that the basis matrix we used is,

$$E = \begin{bmatrix}
-\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{12}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{12}} \\
0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{12}} \\
0 & 0 & \frac{3}{\sqrt{12}}
\end{bmatrix}.$$
Figure 2.1: Observed ilr coordinates (a) $\ln\left(\frac{U_2}{U_1}\right)^{1/\sqrt{2}}$, (b) $\ln\left(\frac{U_2^2}{U_1U_2}\right)^{1/\sqrt{6}}$, and (c) $\ln\left(\frac{U_3}{U_1U_2U_3}\right)^{1/\sqrt{12}}$. The horizontal axes are degrees longitude and the vertical axes are degrees latitude. The colorbar represents the unit free ilr coordinates.

We define ratios as described before, $\left[\left(\frac{U_2}{U_1}\right)^{1/\sqrt{2}} ; \left(\frac{U_2^2}{U_1U_2}\right)^{1/\sqrt{6}} ; \left(\frac{U_3}{U_1U_2U_3}\right)^{1/\sqrt{12}}\right]$.
and refer to the componentwise natural logarithms as the first, second, and third ilr coordinate, respectively. By inspection, we see that larger values of the second ilr coordinate corresponds to larger values of fixed carbon content. Figure (2.2) shows the histograms of the ratios in the left panels and the log-ratios in the right panels. The left panels exhibit varying degrees of right-skewness and approximate symmetry is achieved by taking log-transformation.

Based on our exploratory data analysis, the deterministic large scale variations of the three log-ratios were modeled with a linear trend, where the design matrix consists of ones for an intercept, longitude, latitude, and interaction between longitude and latitude. Recall that the detailed residuals are obtained by subtracting the fitted trend (using ordinary least-squares) from the data. Figure (2.3) shows probability plots of the the detailed residuals.

These detailed residuals deviate from the normal reference line at the tails, a phenomenon common to large datasets. This was a deciding factor to use the distribution-free MOM estimation as EM algorithm failed to estimate the small scale variation appropriately due to its heavy dependence on the likelihood assumptions. However, as mentioned in Section (2.4), the violation of normality will not harm our estimates from being BLUP.

We choose different numbers of knots for the three ilr coordinates, specifically, \( r_1 = 37 \), \( r_2 = 42 \), and \( r_3 = 46 \). This illustrates how spatial processes may be simultaneously modeled in a non-homogeneous manner. Figure (2.4) shows the knot locations from these three schemes, asterisks represent 37
Figure 2.2: Histograms of the three ratios (a) $\left( \frac{U_2}{U_1} \right)^{1/\sqrt{2}}$, (b) $\left( \frac{U_3}{U_1 U_2} \right)^{1/\sqrt{6}}$, and (c) $\left( \frac{U_3}{U_1 U_2 U_3} \right)^{1/\sqrt{12}}$. The respective natural logarithms (i.e. ilr coordinates) are in (d)-(f).

Knot locations, plusses represent the 42 knot locations, and circles represent the 46 knot locations (note that there are overlapping knots locations). Our $V$ matrix then is of dimension $125 \times 125$.

Cressie & Johannesson (2008) and Katzfuss & Cressie (2011 a,b) used
local bisquare basis functions while analyzing total column ozone and CO$_2$ on the entire globe. We use modified bisquare basis function for the geochemical components of coal which allows us to handle the geometric anisotropy (Cressie, 1993). The modified bisquare basis function takes the form:

$$
\begin{align*}
&\begin{cases} 
1 - 0.25d^2(s, u_{j(l)}) & \text{for } d(s, u_{j(l)}) \leq 2 \\
0 & \text{otherwise},
\end{cases}
\end{align*}
$$

Figure 2.3: Normal Probability Plots of detailed residuals after removing trend from (a) first, (b) second, and (c) third ilr coordinates using ordinary least-squares estimates. Dash-dotted lines represent the Normal distribution reference line.
Figure 2.4: Knot locations for the three ilr coordinates, we use $r_1 = 37$, $r_2 = 42$, $r_3 = 46$ knots for these three ratios, respectively. Asterisks represent the 37 knot locations, plusses represent the 42 knot locations, and circles represent the 46 knot locations. Note that we use overlapping knot locations.

where $u_j$ is the $j^{th}$ knot location, $s$ are the observed locations, and

$$d(s, u_j) = \sqrt{(s(lon) - u_j(lon))^2/r_{lon}^2 + (s(lat) - u_j(lat))^2/r_{lat}^2}.$$
s(lon) and s(lat) denote the longitude and latitude of the location s. r_{lon(l)} and r_{lat(l)} control the maximum distance between an observation and a knot such that there is non-zero weight associated between the two. Due to some regions where data are sparse, we chose these values to be r_{lon} = r_{lat} = 0.25, so the result was that all points within 0.5° of a knot have non-zero modified bisquare values.

When estimating V using method of moments we consider M = 58 bins (total 174 when considering all three ilr coordinates). The estimated nugget effects are: \( \hat{\nu}_1^2 = 0.0330 \), \( \hat{\nu}_2^2 = 0.0155 \), and \( \hat{\nu}_3^2 = 0.0950 \). To get an idea of measurement-error variances, we bin the log-transformed observations in small regions and compute the variances for each region. We consider the average of these variances as estimated measurement-error variances \( \hat{\sigma}_i^2 \) and compute \( \hat{\tau}_i^2 = \hat{\nu}_i^2 - \hat{\sigma}_i^2 \), \( i = 1, \ldots, 3 \). Our results are \( \hat{\sigma}_1^2 = 0.0179 \), \( \hat{\sigma}_2^2 = 0.0059 \), and \( \hat{\sigma}_3^2 = 0.0606 \). Thus \( \hat{\tau}_1^2 = 0.0152 \), \( \hat{\tau}_2^2 = 0.0096 \), and \( \hat{\tau}_3^2 = 0.0344 \).

Figures (2.5)(a)-(c) show the pointwise kriging predictions of the ilr coordinates. Because the second ilr coordinate is an increasing function of carbon content, the larger values in the bottom and right sides of Figure (2.5)(b) indicate higher carbon content in those areas. The halo patterns effects are a facet of the observed ilr coordinates over the block. The model successfully captured and smoothed the spatial pattern for all three ilr coordinates.

The associated root mean squared prediction (RMSPE) errors are shown in Figures (2.5) (d)-(f). Note that these are the RMSPEs from universal kriging (see Cressie, 1993, pg. 34, eqn. 3.4.16). As we anticipate, Fig-
Figure 2.5: Pointwise predictions (a)-(c) of the three isometric log-ratios over block 1 shown in Figure (2.6). The ilr transformation is given in (2.6). The respective RMSPEs are shown in (d) - (f). The horizontal axes are degrees longitude and the vertical axes are degrees latitude. The colorbar represents the unit free ilr coordinates (and corresponding unit free RMSPEs).

Figure (2.5)(d)-(f) shows comparatively smaller RMSPE in areas where the observed locations are dense, and larger RMSPE where the observed locations are sparse. We emphasize the fact that, while we are illustrating pointwise
predictions and associated RMSPEs over a particular area, these predictions and RMSPEs were obtained using all the data points.

For blocks we chose two regions, as shown in Figure (2.6).

Figure 2.6: Map of Illinois with the dots showing the observed locations. The boxes represent the successive blocking regions where we perform pointwise and block kriging. The outermost level of block 1 (lower left) is used to illustrate pointwise predictions.
Each block had four levels, the smallest a $0.28^\circ \times 0.28^\circ$ square and the largest a $0.7^\circ \times 0.7^\circ$ square. The block estimates, in ilr coordinates, over the two blocks are found in Figure (2.7).

![Graph of Block Estimates](image)

Figure 2.7: Block estimates of the (a) first, (b) second, and (c) third ilr coordinates over the the blocks. Block 1 is represented with dashed line, Block 2 is represented with solid line. Horizontal axis is $|B|$, the area of the block in degrees squared.

Since block 1 has a larger mean for the second ilr coordinate, the coal in that region appears to have typically higher concentration of carbon. Further, the block prediction is stable as the block size increases. For the second block, the second ilr coordinate is decreasing as the block size increases, indicating
that concentration of carbon diminishes. This could be indicative of the block being centered on a ‘hotspot’ of carbon content, or near a region with much less concentrated carbon. This information could be used, for instance, to assess the rank of coal over a region, and determine which is a more attractive site for mining. Figure (2.8) highlights the inverse relationship of the block RMPSE to $|B|$, the area of the block.

Figure 2.8: Block RMSPEs over (a) block 1 and (b) block 2. Horizontal axis is $|B|$, the area of the block in degrees squared. Dashed lines represent the first ilr coordinate, dotted lines represent the second ilr coordinate, and solid lines represent the third ilr coordinate.

We note that the RMSPEs of the block averages are less than the point-
wise RMSPEs. This is in compliance with the fact that there is more precision in estimating the average than predicting at individual locations. In fact, even on the smallest block size used, where \(|B| = 0.28^\circ \times 0.28^\circ\) the RMSPEs are an order of magnitude smaller than the RMSPEs of point predictions.

### 2.6 Conclusions and Discussion

In the above sections we demonstrate how one can apply log-normal punctual and block kriging to large multivariate datasets after successfully estimating spatial and cross-spatial dependence. We use log-normal methodology to obtain unbiased estimates and their variability. In our application, due to the practice in compositional data analysis, we focus on the log-scale (ilr). But we note that the expressions provide unbiased estimates for both the log-scale and the original scale, and so they are appropriate for any log-normal distributed spatial processes. *Multivariate Spatial Mixed Effects* model provides enough flexibility to handle various processes exhibiting different degrees of smoothness and spatial nonstationarity. Our computing codes were written using Matlab software and implemented on a Windows machine with 8 GB memory. Making the pointwise predictions on a 0.007\(^\circ\) \times 0.007\(^\circ\) resolution covering an area of 0.7\(^\circ\) \times 0.7\(^\circ\) (total 10201 point predictions) took 170 seconds.

In our analysis we did not parametrize or impose any condition on \(V\) matrix, except the positive-definiteness. The resulting spatial and cross-spatial
covariances are anisotropic and nonstationary. The size of $V$ matrix increases with number of knots and number of spatial processes under consideration. To reduce the number of estimable parameters, one can parametrize $V$ matrix by monitoring local variograms or establishing some conditional distributions from an effective exploratory data analysis, e. g. Royle & Berliner (1999). A Bayesian formulation of the multivariate SME model will require prior distributions for $r(r - 1)/2$ parameters of $r \times r$ $V$ matrix, or via a matrix prior (e.g. Inverse-Wishart). Kang & Cressie (2011) provides Bayesian analysis of *Spatial Random Effects* model for univariate spatial process. They decompose $V$ matrix in terms of Givens angles and eigenvalues and put priors on them. But, if one can parametrize $V$ effectively, then it will simplify the analysis without much compromise.

We use the method of moments for estimating $V$. While EM algorithm depends on likelihood assumption, MOM is a distribution free technique. In our experience, for large datasets that show departure from normality, MOM turns out to be more effective than EM algorithm.
Chapter 3

Robust Estimation and Fitting for Reduced Rank Spatial Models

3.1 Introduction

A common method of modeling spatial covariance is the variogram, the variance of the difference of a spatial process on two locations with a specified distance $h$. Typically an empirical variogram is calculated and then a parametric model (e.g. exponential, Gaussian) is fitted to this estimate. Matheron (1962) introduced a MOM estimate of the variogram. However, this estimate is sensitive to outliers. To combat this, Cressie & Hawkins (1980) proposed empirical estimates of the variogram which, by using the fourth
roots of observations at lag $h$, possess the property of robustness (that is, resisting the effect of outliers). However, like other traditional geostatistical methods, this estimate is infeasible for large datasets. In this chapter, we develop an estimation strategy for SME models which is robust to contaminated data.

Recall the model presented in equations (1.1) and (1.2). The two equations may be combined so that for a spatial process $Z$, we write,

$$Z = X\beta + S\eta + \delta + \epsilon.$$  

In this chapter, we assume $X\beta$ to be a fixed (non-random) effect. Recall that with this model the covariance matrix $\Sigma_Z$ can be written as:

$$\Sigma_Z = SVS' + \nu^2I_n,$$  \hspace{1cm} (3.1)

where, as noted, $\nu^2$ is the combined variance of $\delta$ and $\epsilon$. The objective is to estimate the model parameters, $\beta, V$ and $\sigma^2$. Once this has been done one may obtain the inverse of $\Sigma_Z$ easily using the Sherman-Morrison-Woodbury matrix identity as described in section 2.4. This model has a lot of flexibility. The only restriction on $V$ is the positive-definiteness, hence the resulting covariance matrix is anisotropic and nonstationary.

Katzfuss & Cressie (2011 a) estimated the model parameters using EM algorithm. While EM provides a powerful technique for estimating model parameters, it is sensitive to the model specifications and leads to poor pre-
predictions when datasets depart from normality. Instead of EM algorithm, one may use the method of moments estimation introduced by Cressie & Johannesson (2008). MOM is distribution-free and hence has the advantage over EM algorithm in situations where the data depart from normality. However, the MOM estimation minimizes the Frobenius norm and is sensitive to outliers, providing poor estimates in the presence of contaminated data. We develop a distribution free robust estimation method that is robust to contamination, has desirable asymptotic properties, and performs uniformly better than the method of Cressie & Johannesson (2008).

MOM estimation was described in section 2.4. There are two important components of MOM estimation to point out:

- $\hat{\Sigma}_M$ is constructed using bin means of the detailed residuals.

- Estimates of $V$ and $\nu^2$ are obtained by minimizing the Frobenius norm. It is well-known that the mean is not robust to outliers, so the estimation stage is influenced by outliers. Also, for a matrix $A$ with elements $a_{ij}$, the Frobenius norm is $||A||_F = \sqrt{\sum |a_{ij}|^2}$. Large values among the elements of $A$ will inflate this value. Since the Frobenius norm minimizes $||\hat{\Sigma}_M - \Sigma_M||_F$, poor estimation of $\Sigma_M$ will result in a poor fit.

Because both stages are susceptible to outliers, the MOM as suggested by Cressie & Johannesson (2008) breaks down in the presence of contaminated data. In this chapter we propose robust alternatives to both the estimation and fitting stages. In the estimation stage we use the median absolute devia-
tion to construct the empirical binned covariance matrix. In the fitting stage we use quantile regression to fit model parameters. Our proposed robust estimates out-perform the current MOM methods both in simulation and when applied to a large remote sensing data obtained through NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS) on the Terra satellite.

Section 3.2 introduces our proposed robust alternatives for obtaining parameter estimates. In section 3.3 we prove the consistency of our proposed estimate. Section 3.4 provides a simulation study investigating the performance of the proposed methods compared to the recommended methods. Section 3.5 applies both the proposed and traditional methods to real data. Finally, section 3.6 provides a discussion about these estimates based on application to the NASA data.

### 3.2 Robust Estimation and Fitting

For notation we denote the empirical estimate of the binned covariance matrix recommended by Cressie & Johannesson (2008) as $\hat{\Sigma}^{(CJ)}_M$. The fitting stage minimizing the Frobenius norm is termed as *Frobenius fit*. In this section we provide robust alternatives to both the estimation stage and fitting stage of MOM estimation for the FRK model. First we define $\hat{\Sigma}^{(rob)}_M$, an empirical binned covariance matrix which is robust to contamination. Then we describe a robust strategy to fit the model parameters, which we call the
3.2.1 Estimation Stage

Use of the mean causes $\hat{\Sigma}^{(CJ)}_M$ to be sensitive to outliers. We define $\hat{\Sigma}^{(rob)}_M$ using the median absolute deviation, MAD$(X) = \text{med}(|X - \text{med}(X)|)$, a robust statistic which quantifies variability. A constant scale factor is applied to the MAD which causes it to be a consistent estimate for the standard deviation (see Hettmansperger & McKean, 2011, Eqn 3.9.27). For the normal distribution this scale factor is $1/\Phi^{-1}(0.75)$, so a consistent estimate of standard deviation is $\text{MAD}(X)/\Phi^{-1}(0.75)$. Due to the normality supposed by the SME model, we use this scale factor when calculating the MAD. Other scale factors may be used if the distribution of the detailed residuals $\tilde{R}$ follows a distribution other than Gaussian.

Recall that $\Sigma_M$ is the covariance matrix over the $M$ bins, therefore the diagonal elements represent variability within a bin. To estimate the variability of the $m^{th}$ bin, the diagonal elements of $\hat{\Sigma}^{(rob)}_M$ are given by,

$$\hat{\Sigma}^{(rob)}_M(m, m) = \text{MAD}^2\left(\tilde{R}_m\right), m = 1, \ldots, M,$$  \hspace{1cm} (3.2)

the square of the median absolute deviation of the detailed residuals falling into the $m^{th}$ bin. We square the MAD to obtain an estimate of the variance instead of the standard deviation.

The off-diagonal elements of $\hat{\Sigma}^{(rob)}_M$ measure the covariance between two
bins. Using the identities:

\[ V(A + B) = V(A) + V(B) + 2Cov(A, B), \]
\[ V(A - B) = V(A) + V(B) - 2Cov(A, B), \]

it follows that,

\[ Cov(A, B) = \frac{V(A + B) - V(A - B)}{4}. \]

However, the \( m^{th} \) and \( m'^{th} \) bins do not necessarily contain the same number of observed locations, so \( \tilde{R}_m \pm \tilde{R}_{m'} \) may not be defined due to the differing dimensions of \( \tilde{R}_m \) and \( \tilde{R}_{m'} \). Furthermore, even if the number of observed locations contained in each bin was equal, there is no unique association between the points. That is, there is no clear way to determine which locations from bin \( m \) are associated with which locations in bin \( m' \). Lacking such a relationship between the locations in two bins, any calculation (if even possible) of \( \tilde{R}_m \pm \tilde{R}_{m'} \) is arbitrary, based simply on the order in which the data appear.

To overcome this problem we estimate the covariance between the \( m^{th} \) and \( m'^{th} \) bins with

\[
\Sigma_M^{(rob)}(m, m') = \frac{\text{MAD}^2 \left( \tilde{R}_m \oplus \tilde{R}_{m'} \right) - \text{MAD}^2 \left( \tilde{R}_m \ominus \tilde{R}_{m'} \right)}{4},
\]

where \( \oplus \) indicates the pairwise sums, and \( \ominus \) the pairwise differences.
3.2.2 Fitting Stage

Given an empirical covariance matrix $\hat{\Sigma}_M$, the task is to fit the covariance matrix $V$ by minimizing some norm between $\hat{\Sigma}_M$ and $\Sigma_M$. As noted, the Frobenius norm minimizes $||\hat{\Sigma}_M - \Sigma_M||_F$, so poor estimation of $\Sigma_M$ will inflate the Frobenius norm and result in a poor fit. To develop a robust fitting stage, we start from equation (3.1),

$$
\Sigma_M = \mathbf{S} \mathbf{V} \mathbf{S}' + \nu^2 \mathbf{I}_M
$$

$$
\Sigma_M \mathbf{S} \left( \mathbf{S}' \mathbf{S} \right)^{-1} = \mathbf{S} \mathbf{V} + \nu^2 \mathbf{S} \left( \mathbf{S}' \mathbf{S} \right)^{-1}
$$

$$
\left( \Sigma_M - \nu^2 \mathbf{I}_M \right) \mathbf{S} \left( \mathbf{S}' \mathbf{S} \right)^{-1} = \mathbf{S} \mathbf{V} \quad (3.4)
$$

Note that while $\mathbf{S}$ is sparse, it is full column rank and therefore $\left( \mathbf{S}' \mathbf{S} \right)$ is invertible. Then, substituting in estimates of $\Sigma_M$ and $\nu^2$, we may see equation (3.4) as a multivariate regression problem with $\mathbf{S}$ as the design matrix and $\mathbf{V}$ as the matrix of regression coefficients. Any method of robust regression may then be implemented to obtain an estimate of $\mathbf{V}$. We use quantile regression as described by Koenker & Bassett (1978). For this, let $x_t, t = 1, \ldots, T$, be a sequence of $K-$dimension row vectors of the design matrix, and let $y_t$ be a random sample from the regression process $u_t = y_t - x_t \gamma$ having distribution function $F$. Then quantile regression seeks to minimize,

$$
\lim_{b \in \mathbb{R}^K} \left[ \sum_{t \in \{ t : y_t \geq x_t b \}} \theta |y_t - x_t b| + \sum_{t \in \{ t : y_t < x_t b \}} (1 - \theta) |y_t - x_t b| \right],
$$
where $\theta \in (0, 1)$ is a constant which determines the regression quantile modeled. In our application, we model the median ($\theta = 0.5$), and the rows of $\overline{S}$ represent $x_t$. Each column of $(\Sigma_M - \nu^2 I_M) \overline{S} \left( \overline{S}' \overline{S} \right)^{-1}$ is used as the response $y_t$ in a separate estimation. There are therefore $r$ estimates of $\gamma$ to obtain, each of which corresponds to a column of $V$. Due to rounding by the software package, the final estimate $V$ may not be numerically symmetric, so we symmetrize $\hat{V}$ by taking $\hat{V} = 0.5 \left( \hat{V} + \hat{V}' \right)$.

Estimation of $V$ requires an estimate of $\nu^2$. By substituting the left side of (3.4) for $S V$ in (3.1) we obtain:

$$
\Sigma_M = \left( \Sigma_M - \nu^2 I_M \right) \overline{S} \left( \overline{S}' \overline{S} \right)^{-1} \overline{S}' + \nu^2 I_M
$$

$$
\Sigma_M = \Sigma_M \overline{S} \left( \overline{S}' \overline{S} \right)^{-1} \overline{S}' + \nu^2 \left( I_M - \overline{S} \left( \overline{S}' \overline{S} \right)^{-1} \overline{S}' \right)
$$

$$
\Sigma_M \left( I_M - \overline{S} \left( \overline{S}' \overline{S} \right)^{-1} \overline{S}' \right) = \nu^2 \left( I_M - \overline{S} \left( \overline{S}' \overline{S} \right)^{-1} \overline{S}' \right).
$$

(3.5)

We then stack the columns of $\Sigma_M (I_M - \overline{S} (\overline{S}' \overline{S})^{-1} \overline{S}')$ and the columns of $(I_M - \overline{S} (\overline{S}' \overline{S})^{-1} \overline{S}')$. Doing this, we again cast the problem as a zero-intercept robust regression, where $\nu^2$ is the slope. This estimate may then be substituted into equation (3.4) to obtain an estimate of $V$.

As before, we may need to lift the eigenvalues to ensure that the resulting matrix is positive definite, while preserving the total variability. We lift the eigenvalues by:

1. Obtain the eigenvalues, $\lambda_i, i = 1, \ldots, M$ and their sum, $\Lambda = \sum_{i=1}^{M} \lambda_i$. 
2. If \( \min(\lambda_i) \) is negative, compute \( \lambda_i^* = \lambda_i + |\min(\lambda_i)| \) and \( \Lambda^* = \sum_{i=1}^{M} \lambda_i^* \).

3. Use the \( \lambda_i^* \)'s as weights to compute \( \hat{\lambda} = \Lambda (\lambda_i^*/\Lambda^*) \).

4. The \( \hat{\lambda}_i \)'s are the lifted eigenvalues.

In this manner, we approximate the eigenvalues of \( V \) with lifted eigenvalues which are necessarily non-negative and proportional to the original eigenvalues. That is, the larger eigenvalues of \( \tilde{V} \) will receive a larger portion of \( \Lambda \), the sum of the original eigenvalues. This method is a different approximation of the eigenvalues than that by Kang, Cressie, & Shi (2010), but also allowed for the process to be automated, e.g. for simulations. Their method may still be used for individual analyses.

### 3.2.3 Summary of Proposed Method

A geostatistical analysis using the methods described above can be summarized in the following steps.

1. Using some design matrix \( X \), obtain \( \tilde{R} \) through OLS regression.

2. Select \( r \) knots over the spatial domain (these can be multi-resolutional) and define \( M \) bins, where \( r \leq M \).

3. Using the observed locations and selected knots, define \( S \), the matrix of basis functions.

4. Construct \( \Sigma^{(rob)}_M \) by \( \hat{\Sigma}^{(rob)}_M (m, m) = \text{MAD}^2 \left( \tilde{R}_m \right) \) and \( \hat{\Sigma}^{(rob)}_M (m, m') = \frac{1}{4} \left( \text{MAD}^2 \left( \tilde{R}_m \oplus \tilde{R}_{m'} \right) - \text{MAD}^2 \left( \tilde{R}_m \ominus \tilde{R}_{m'} \right) \right) \).
CHAPTER 3. ROBUST ESTIMATION AND FITTING FOR RRSM

5. Use quantile regression to obtain $\hat{\nu}^2$ via as shown in equation (3.5).

6. Substitute $\hat{\nu}^2$ into equation (3.4); using $\bar{S}$ as the design matrix, perform quantile regression on the columns of $(\Sigma_M - \nu^2 I_M) \bar{S} \left( \bar{S}' \bar{S} \right)^{-1}$ to obtain each column of $\hat{V}$.

7. Substitute $\hat{V}$ and $\hat{\nu^2}$ into appropriate kriging equations to make spatial predictions.

3.3 Asymptotic Properties

Here we discuss some of the infill asymptotic properties of our proposed estimator, $\hat{\Sigma}_M^{(rob)}$. Infill asymptotics is a common method of considering asymptotics related to geostatistical methodology in which the domain, $\mathcal{D}$, remains fixed but the density of observed locations is increased.

Recall that we obtain $\hat{V}$ by minimizing some norm $|| \cdot ||$,

$$\hat{V} = ||\hat{\Sigma}_M - \Sigma_M||,$$

Hence, once $\hat{\Sigma}_M$ is known, $\hat{V}$ is fully determined by the fitting method. Therefore, a desirable property of the empirical binned covariance matrix $\hat{\Sigma}_M^{(rob)}$ is that it be consistent for $\Sigma_M$.

We show consistency for $\hat{\Sigma}_M^{(rob)}$ element-wise, i.e. $a' \hat{\Sigma}_M^{(rob)} a \to a' \Sigma_M a$ for any $M \times 1$ vector $a$. Under model assumptions the detailed residuals $R$ obtained from ordinary least-squares regression with mean vector $0$ and co-
variance matrix $\Sigma_{R_{m}}$ within bin $m$. Wendler (2011) provides useful results on convergence rates of Bahadur representations of sample quantiles and U-quantiles for dependent data. They provide in particular two theorems showing the error terms of the Bahadur representations for sample quantiles and $U$-quantiles are consistent for zero. These theorems and several important definitions are reproduced in the appendix, we refer to these as theorem 1 and theorem 2. There are three conditions we must assume for these theorems to apply. First, both assume stationary processes. While the overall spatial process $R$ may be non-stationary, this assumption is not problematic, because we consider the detailed residuals within a bin, denoted $\tilde{R}_{m}$. As the bins are considered to be small areas, $\tilde{R}_{m}$ may be assumed to be stationary.

The theorems also require some form of mixing. For a stationary process $W_{n}$, the strong mixing coefficients are defined as:

$$\alpha(k) := \sup \{|P(AB) - P(A)P(B)| : A \in F_{1}^{n}, B \in F_{n+k}^{\infty}\}$$

where $F_{p}^{q}$ is the $\sigma$-field generated by random variables $W_{p}, \ldots, W_{q}$. The process is strongly mixing if $\lim_{k \to \infty} \alpha(k) = 0$.

In the context of geostatistics, strong mixing implies an exponential decay in the spatial correlation between two locations. Such decay in the correlation is a common feature in spatial modeling (e.g. the Matérn class of autocorrelation models). For our proof, let $R_{m}$ denote the random detailed residual process within the $m^{th}$ bin, and let $\tilde{R}_{m} = \{\tilde{R}_{m_{1}}, \ldots, \tilde{R}_{m_{k}}\}$ be the
CHAPTER 3. ROBUST ESTIMATION AND FITTING FOR RRSM

$k$ observed detail residuals from that bin. We assume that $R_m$ and, as will be seen, $|R_m|$, exhibit strong mixing as described in theorems 1 and 2. Finally, the theorems of Wendler (2011) have a condition on the distribution function, $F$ (or $U$, for the distribution function of $U$-quantiles). For our purposes, $F$ (and $U$) being twice differentiable at the median is sufficient. Since $R_m$ is normally distributed, and $|R_m|$ is folded normal, this condition will hold. For our proof, we make use of this standard inequality:

$$|a| = |a - b + b| \leq |a - b| + |b|.$$

Recall the diagonal elements of $\hat{\Sigma}^{(rob)}_M$ are given by MAD$(R_m)$. The statistic and its functional are:

$$\text{MAD}\{\tilde{R}_m\} = \text{med} |\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}|$$

$$\xi_m = \text{med} |R_m - \text{med}\{R_m\}|$$

Without loss of generality let $\text{med}\{R_m\} = 0$. Since $\text{med}\{\tilde{R}_{m_i}\} \xrightarrow{P} 0$, in probability, choose $N_0$ sufficiently large so that, given $\epsilon > 0$,

$$k \geq N_0 \Rightarrow |\text{med}_{1 \leq i \leq k}\{\tilde{R}_{m_i}\}| < \epsilon,$$  \hspace{1cm} (3.6)

with probability greater than $(1 - (\epsilon/2))$. Let $A_n$ denote the event where
(3.6) occurs. Then on $A_n$ we have,

$$|\tilde{R}_{m_i}| = |\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\} + \text{med}\{\tilde{R}_{m_j}\}|$$

$$\leq |\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}| + |\text{med}\{\tilde{R}_{m_j}\}|$$

$$< |\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}| + \epsilon$$

Hence on $A_n$,

$$\text{med}_i|\tilde{R}_{m_i}| < \text{med}_i|\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}| + \epsilon. \quad (3.7)$$

Also on $A_n$ we have,

$$|\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}| = |\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\} - \tilde{R}_{m_i} + \tilde{R}_{m_i}|$$

$$\leq |\text{med}\{\tilde{R}_{m_j}\}| + |\tilde{R}_{m_i}| < |\tilde{R}_{m_i}| + \epsilon.$$

Hence on $A_n$,

$$\text{med}_i|\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}| < \text{med}_i|\tilde{R}_{m_j}| + \epsilon. \quad (3.8)$$

Putting (3.7) and (3.8) together, we have on $A_n$,

$$\left|\text{med}_i|\tilde{R}_{m_i} - \text{med}\{\tilde{R}_{m_j}\}| - \text{med}_i|\tilde{R}_{m_i}|\right| < \epsilon. \quad (3.9)$$

Since this occurs with probability of at least $(1 - (\epsilon/2))$, the difference on the
left-side goes to 0 in probability. Because med$_i |\tilde{R}_{m_i} | \xrightarrow{P} \xi$, so does med$_i |\tilde{R}_{m_i} - \med \{R_{m_j} \}|$. QED.

Now recall that the off-diagonal elements of $\Sigma^{(rob)}$ are given by equation (3.3), which can be expressed as the sum of $U-$quantiles as follows:

$$
\Sigma^{(rob)}_M (m, m') = \left( \text{MAD} \left\{ \frac{R_m \oplus R_{m'}}{2} \right\} \right)^2 - \left( \text{MAD} \left\{ \frac{R_m \ominus R_{m'}}{2} \right\} \right)^2
$$

(3.10)

For two sequences $x_n$ and $y_n$, if $x_n \xrightarrow{P} x$ and $y_n \xrightarrow{P} y$ then $(x_n - y_n) \xrightarrow{P} (x - y)$. Also, letting $q(u) = u^2$, if $x_n \xrightarrow{P} x$ then $q(x_n) \xrightarrow{P} q(x)$. Therefore, we may consider the $U-$quantiles only, and do so separately.

Theorem 2 of Wendler (2011) provides analogous convergence results for $U-$quantiles, given strong mixing, stationarity, and appropriate kernal functions $h(x, y, t)$. For example, the kernal for the Hodges-Lehmann (HL) estimator (described in Hettmansperger & McKean, 2011, chapter 1) is the identity function $h(x, y, t) = I(\frac{1}{2} (x + y) < t)$.

Define $T = \frac{1}{2} (R_m \oplus R_{m'})$, and let $\tilde{T} = \frac{1}{2} (\tilde{R}_{m_i} \oplus \tilde{R}_{m_j})$ for $i = 1, \ldots, k$ and $j = 1, \ldots, l$. Then the statistic and its functional, respectively, for the off-diagonal elements are:

$$
\text{MAD}\{\tilde{T}\} = \med |\tilde{T}_p - \med \{T_q\}|
$$

$$
\xi_{m, m'} = \med |T - \med \{T\}|
$$
Without loss of generality let $\text{med}\{T\} = 0$. Our assumptions ensure that $\text{med}\{	ilde{T}_i\} \xrightarrow{P} 0$, and then the proof follows in the same manner as for the diagonal elements. So each of the $U-$quantiles in equation (3.10) are consistent, and therefore entire expression is consistent. Hence we have shown $a' \hat{\Sigma}^{(\text{rob})}_M a \rightarrow a' \Sigma^{(\text{rob})}_M a$ for any $M \times 1$ vector $a$.

Throughout we have treated the number of bins, $M$, as fixed, and do not consider limits over that quantity. This is analogous to the work of Bliznyuk et al (2012). They consider $m$ as a radius to determine “adjacency” of locations, where $m$ does not depend on $n$, and do not limit over $m$. The only restriction on $M$ is that it be large enough to ensure that the assumption of stationary within bins holds for practical implementation.

### 3.4 Simulation Study

To compare our proposed methods with the existing methods using simulated data, we generate a spatial process $Z$ according to the SME model:

$$Z = X\beta + S\eta + \epsilon.$$ 

We set $\beta = 0$, (i.e. we simulate the residual process directly) and then select $n$ observed locations equidistant over a $100 \times 100$ grid. A fixed rank of $r$ was chosen, with knot locations equidistant over the grid. Then we define $V$ such that $V_{ij} = \exp(-\theta ||S_i - S_j||)$ where $||S_i - S_j||$ is the Euclidean distance...
between the two knot locations \( S_i \) and \( S_j \), and \( \theta \) is some positive constant. Note that as the distance increases, the correlation decreases, as is to be expected with spatial correlation. We construct \( S \) using the bisquare basis functions defined as,

\[
S_{i,j} = \begin{cases} 
\left( 1 - \left( \frac{1}{r_u} ||s_i - u_j|| \right)^2 \right)^2 & \text{for } ||s_i - u_j|| \leq a \\
0 & \text{otherwise,}
\end{cases}
\]

where \( r_u \) is 1.5 times the minimum distance between knots. Note that we do not use multi-resolution basis functions for the simulation, but for our data analysis in section (3.5) we demonstrate that multi-resolutional may be used with our method.

For simulating datasets we first generate a \( r \)–dimensional process \( \eta \) from a zero-mean multivariate normal with covariance \( V \). To induce outliers, the measurement error process \( \epsilon \) is generated from a contaminated normal distribution, \( \alpha \mathcal{N}(0, \nu_c^2) + (1 - \alpha) \mathcal{N}(0, \nu^2) \), where \( \alpha \in (0, 1) \). Finally, we obtain the simulated detail residuals by \( R = S\eta + \epsilon \).

We varied \( n \in \{10000, 19881, 29929\} \) and \( r \in \{25, 100, 225\} \). The contamination level for \( \epsilon \) was varied within \( \alpha \in \{0.00, 0.05, 0.10, 0.15\} \). Values of \( \theta = 1, \nu^2 = 0.1 \) and \( \nu_c^2 = 1 \) were all held constant. These choices are not sensitive to our estimation technique. We use 5 replications for each combination of simulation parameters values. For each replication we compute both binned covariance matrices, \( \hat{\Sigma}_M^{(CJ)} \) and \( \hat{\Sigma}_M^{(rob)} \). For \( \hat{\Sigma}_M^{(CJ)} \) we obtain \( \hat{V} \).
using only the Frobenius fitting, while $\hat{\Sigma}_M^{(\text{rob})}$ for we obtain $\hat{V}$ using both the Frobenius fitting and the proposed robust fitting. Thus each replication produces three different estimates of $V$. We do not consider the combination of $\hat{\Sigma}_M^{(CJ)}$ with the robust fitting because the purpose of the robust fit is to protect against poor estimation, and if the estimation is a concern, then $\hat{\Sigma}_M^{(CJ)}$ should not have been used in the first place.

For a given estimate $\hat{V}$, we measure the closeness of $\hat{V}$ and $V$ by

$$||\hat{V} - V|| = \text{MAD} \left( \hat{V}_{ij} - V_{ij} \right),$$

(3.11)

the MAD of the elementwise differences. Based on this metric, smaller values indicate a more accurate fit to the true value of $V$. Simulation results are shown in Figures (3.1)-(3.3), though we plot the natural logarithm of (3.11).

Figure (3.1) shows the value of equation (3.11) when the Frobenius fitting is used. The figure displays use of both the CJ estimate and the Robust estimate. We note that in all cases, the Robust estimate provides a closer estimate to $V$ than the CJ estimate.

Figure (3.2) shows a comparison of the Frobenius fitting and Robust fitting when the Robust estimate is being used. Note that the values on the y-axis are all negative, hence both methods have produced excellent results.

Figure (3.3) compares the CJ estimate with the Frobenius fit (i.e. the method of Cressie & Johannesson (2008)) against the Robust estimate with Robust fit. The Robust estimation and fitting provide uniformly better esti-
mates of $V$, even when there is no contamination.

Figure 3.1: Simulation results comparing CJ (o) and Robust (△) estimates of binned covariance matrix $\hat{\Sigma}_M$ using Frobenius fitting. Plotted value is natural logarithm of expression in (3.11).
3.5 Application to NASA Data

We use remote sensing data on daily cloud liquid water path (CWP), obtained through NASA’s Moderate Resolution Imaging Spectroradiometer

Figure 3.2: Simulation results comparing Frobenius (——) and Robust (---) fitting when using the Robust estimate of binned covariance matrix. Plotted value is natural logarithm of expression in (3.11).
(MODIS) on the Terra satellite. The data were collected on April 22, 2012. Because the dataset is extremely large \((n = 48552)\), a reduced rank model would be a reasonable choice for inference. These datasets are typically right-skewed, we take the log-transformation before finding kriging estimates and

![Simulation results comparing CJ estimate with Frobenius fitting (○○○) to the Robust estimate with Robust (-△-) fitting when using the Robust estimate of binned covariance matrix. Plotted value is natural logarithm of expression in (3.11).](image)

Figure 3.3: Simulation results comparing CJ estimate with Frobenius fitting (○○○) to the Robust estimate with Robust (-△-) fitting when using the Robust estimate of binned covariance matrix. Plotted value is natural logarithm of expression in (3.11).
associated predictions and root mean square prediction errors (RMSPE). As observed in chapter 2, the MSPE of a log-normal process is proportional to the expected value. Therefore, we restrict our focus to log-scale to more directly compare the CJ and robust models in terms of their predictions and MSPEs.

The observed data are plotted in Figure (3.4). Note that there is a distinct north-south trend present, showing generally smaller values near the equator, and generally larger values nearing the poles. Consequently, we model the large-scale variation using Legendre polynomials (Stein, 2007) \( P_n^m(\sin(L)) \) of degree \( n = 80 \) and order \( m = 0, 1, \ldots, n \). The design matrix consists of 81 regressors of spherical harmonics.

For the MOM estimation described in the preceding sections we first need to compute the detailed residuals, \( \tilde{\mathbf{R}} = \mathbf{Z} - \mathbf{X}\hat{\mathbf{\beta}} \). The normal probability plot of the detailed residuals in Figure (3.5) shows contamination in \( \tilde{\mathbf{R}} \). This motivates the use of the robust techniques described earlier in Section (3.2).

As recommended by Cressie & Johannesson (2008), we use a multi-resolution model for CWP (see Nychka, Wikle, & Royle, 2002), to capture multiple scales of variation. We choose \( r_1 = 38 \) knot locations for the first resolution, and \( r_2 = 97 \) knot locations for the second resolution. Therefore the estimate of \( \mathbf{V} \) is a \( 135 \times 135 \) matrix. A map of these knot locations is given in Figure (3.6).

To construct the \( \mathbf{S} \) matrix, we used the modified bisquare function, de-
Figure 3.4: Plot of observed Cloud Water Path over the spatial domain.

Defined as:

\[
S_{i,j(l)} = \begin{cases} 
1 - 0.25d^2(s_i, u_{j(l)}) & \text{for } d(s_i, u_{j(l)}) \leq 2 \\
0 & \text{otherwise},
\end{cases}
\]

where \(u_{j(l)}\) is the \(j^{th}\) knot location of the \(l^{th}\) resolution, \(s_i\) are the observed locations. The distance is given by:

\[
d(s_i, u_{j(l)}) = \sqrt{d_{\text{long}}^2(s_i, u_{j(l)})/r_{\text{long}}^2(l) + d_{\text{lat}}^2(s_i, u_{j(l)})/r_{\text{lat}}^2(l)},
\]

where \(d_{\text{long}}(s_i, u_{j(l)})\) and \(d_{\text{lat}}(s_i, u_{j(l)})\) denote the longitude (east-west) and
CHAPTER 3. ROBUST ESTIMATION AND FITTING FOR RRSM

Figure 3.5: Normal probability plot of the detailed residuals.

latitude (north-south) distances, respectively, between the location \( s \) and the knot location \( u_j(l) \). The values \( r_{\text{long}(l)} \) and \( r_{\text{lat}(l)} \) control the maximum distance between an observation and a knot such that there is non-zero weight associated between the two. We chose these to be the minimum east-west distance and minimum north-south distance between two knot locations of the same resolution.

Terminology is important in describing results, because there are two methods of estimating \( \Sigma_M \) and two methods of fitting \( V \) from \( \Sigma_M \). Remember that we have termed the MOM estimation and fitting described in
Figure 3.6: Plot of the knot locations of the basis functions over the spatial domain. Asterisks represent the 38 knot locations of the first resolution, and circles represent the 97 knot locations of the second resolution.

Cressie & Johannesson (2008) as the CJ estimate and the Frobenius fit, respectively. We name estimate in Eqns (3.2) and (3.3) as the robust estimate and the method of fitting described in equations (3.4) and (3.5) as the robust fit. Further, either estimate can be used with either fitting scheme, resulting in four combinations of estimation and fitting. The predictions of log CWP and the associated MSPEs are shown in Figures (3.7)-(3.9). Figure (3.7) shows the CJ estimate of $\Sigma_M$ with the Frobenius fit. Figure (3.8) shows the robust estimate with the Frobenius fit. Figure (3.9) shows both the robust estimate with the robust fit.
Figure 3.7: Plot of (a) predictions and (b) RMSPEs when using the CJ estimate and the Frobenius fit.

The predictions are similar for the three combinations, however the RMSPE show large differences between these methods. Figure (3.10) shows histograms of the RMSPEs for the three combinations. Panel (a) shows the CJ estimate with the Frobenius fit, panel (b) shows the robust estimate with the Frobenius fit, and panel (c) shows both the robust estimate with the robust fit.

Clearly, using the robust estimate (panels (b) and (c)) result in uniformly
smaller RMSPEs than using the CJ estimate. When using the robust estimate, the Frobenius fit appears to provide the smaller RMSPEs. However, in another analysis using a single-resolution knot scheme, the combination of robust estimation and robust fitting provided the smallest RMSPEs. It should be noted that we observe these results despite the fact that, as shown in the quantile plot in Fig (3.5), the dataset displays only a moderate degree of non-normality and contains few outliers.
Figure 3.9: Plot of (a) predictions and (b) RMSPEs when using the Robust estimate and the Robust fit.

We artificially contaminated the log CWP data by replacing the 2% of observed values $Z_i(s)$ with $1.5Z_i(s)$ (inspection of the normal probability plot revealed significant deviation from normality). The results followed the same pattern as those described above.
Figure 3.10: Histograms comparing the RMSPE for each of the three combinations of estimation and fitting. (a) CJ estimate, Frobenius fit, (b) Robust Estimate, Frobenius fit, and (c) Robust Estimate, Robust fit.

### 3.6 Conclusions and Discussion

The Method of Moments is a flexible and powerful tool for estimating the parameters of a FRK model. Bayesian methods are more accurate than kriging (Kang & Cressie, 2011), but they are also more time-consuming, and often rely on assumption of normality. Kriging is typically a faster process, and kriging estimates are BLUP even in the face of non-normality, so kriging presents benefits of its own. However the typical MOM estimates
are susceptible to contaminated data. In this work we have provided robust alternatives to both stages of the MOM estimation. Our results indicate that the proposed estimate and fitting scheme successfully capture the spatial covariance.

The robust estimate has been shown to perform uniformly better than the CJ estimate. This is visible both through the simulations and the data application. Even with no contamination the robust estimate out-performed the CJ estimate.

At a glance the simulations (particularly Figure (3.2)) and data application may seem to indicate that the robust fit is not always beneficial. There is good reason for this: the fitting scheme only considers the estimate, $\hat{\Sigma}_M$. Extreme values among the elements of this matrix are what cause the Frobenius fitting to break down. Using the robust estimate will largely prevent such extreme values among the elements of $\hat{\Sigma}_M$ in the first place, enabling the Frobenius fit to be successful. As shown through the simulations (particularly Figure (3.2) and as observed in the data application, when the Robust estimate is used, the difference between the Frobenius fit and Robust fit is slight.

Besides the $L_1$-fit, other robust fits can be used. For example, the Wilcoxon fit is a robust fit that minimizes the sum of the absolute differences of the residuals (see Hettmansperger & McKean, 2011, Section 3.8). The Wilcoxon fit is generally more efficient than the $L_1$-fit and it generalizes to fits for skewed-error distributions.
It should be noted that the kriging equations were derived by minimizing the mean square prediction error. These predictions are then simply functions of $V$ and $\nu$. In our work, we have provided robust methods of estimating these same parameters. Yet when using robust techniques, it may be desirable to derive predictions and measures of precision using a different loss function than the squared error loss, or such that the predictions are robust in addition to the parameter estimates (Cressie & Hawkins, 1984). We emphasize that the robust estimates perform well in spite of this.
Chapter 4

An Empirical Bayesian Approach for Selection of Knot Locations in Reduced Rank Spatial Models

4.1 Introduction

Success of SRE model to characterize the spatial variability depends on careful selection of number \( r \) and locations of knots, \( \mathcal{S} \). In the existing literature, there is a very limited discussion on selecting \( r \) and the knot locations. Guhaniyyogi et al. (2011) and Katzfuss (2013) are among the few to address this topic in particular, and both of their methods are computationally
intense. This chapter is an attempt to supplement this lack of literature by developing an objective method for knot selection which is less computationally demanding. Usually one selects the knot locations that cover uniformly the entire area under study. Higdon (1998) and Calder (2007) used uniform knot locations for process convolution models and similar approach was used by Banerjee et al (2008) for spatial coregionalization models. The papers on reduced rank spatial and spatio-temporal models written by Cressie and his coauthors used multi-resolution approaches for designing the knot locations. They used three resolution levels and the knot locations were uniformly distributed over the area under study in each scheme. Using too few knots will result in an overly smoothed model that may fail to capture the small-scale variations adequately (see Rodrigues & Diggle, 2010). On the other hand, too many knots will slow down the fitting and the prediction as inversion of $V$ matrix is required several times in these steps. Rodrigues & Diggle (2010) recommended that one should put more knots where sample locations are dense. More knots are required in the regions of weak spatial dependence. While the former is easy to detect by viewing data maps, a strictly visual assessment of spatial dependence is neither easy, nor objective. Fuentes (2001) assessed the number of knots by monitoring Akaike Information Criterion (AIC) by comparing only a handful of models. Use of information criteria (such as, AIC and Bayes Information Criterion (BIC)) is not effective as the number of candidate models to compare is quite large ($2^r$). Crainiceanu, Diggle, & Rowlingson (2008) used space-filling design
proposed by Nychka & Saltzman (1998) while analyzing Loa loa prevalence in tropical Africa using bivariate binomial spatial modeling. But Banerjee et al. (2008) pointed out that space-filling designs are based on geometric criteria and independent of the assumed covariance function.

Usually selection of knot locations is done in the modeling stage and an important requirement is that the procedure should not be expensive in terms of time. Keeping this in mind, we develop a two step empirical Bayes procedure. We recast the problem of knot selection as a variable selection problem in linear regression. In our proposed method we start with a sufficiently large number of equispaced knots, say \( r \). Then based on our selection criteria, we exclude those knots that are unnecessary. In what follows, we treat \( \eta \) as regression coefficients. In Bayesian least absolute shrinkage operator (LASSO), Park & Casella (2008) used independent Laplace priors on regression coefficients. But here, \( \eta \) are spatially correlated. Hence, some modifications are required. We use a multivariate symmetric Laplace (see Eltoft, Kim, & Lee, 2006) prior for \( \eta \). In the first step, we estimate the parameters in multivariate Laplace distribution using an ascent based Monte Carlo Expectation Maximization (MCEM) algorithm, see Caffo, Jank, & Jones (2005). The samples required for MCEM are obtained through a split-chain MCMC algorithm (Mykland, Tierney, & Yu, 1995). In the second step, we select the knot locations using Bayesian LASSO, similar to the methodologies described in Park & Casella (2008) and Lykou & Ntzoufras (2012).

As an illustration, we simulate random numbers from a zero mean mul-
Figure 4.1: Plots of the three simulated datasets.
CHAPTER 4. EMPIRICAL BAYESIAN KNOT SELECTION

tivariate Gaussian distribution with exponential covariance function \( c(h) = \exp\{-\theta h\} \) on an unit square. Here \( h \) denotes Euclidean distance between two locations and \( \theta > 0 \) is a range parameter that controls the strength of spatial correlation. Smaller values of \( \theta \) will lead to smoother process and strong spatial dependence and larger values will result in coarser spatial process with weak spatial dependence. We generate three realizations of the spatial domain, using progressively fewer knots in each. The details of simulations are provided in section 4.4. The generated spatial processes are shown in Figure (4.1)(a)-(c). As anticipated, simulation 1 has the most course spatial process and simulation 3 has the smoothest spatial process.

Multi-resolution is an added interesting feature of reduced rank spatial models, not essential to model the spatial covariance of large datasets. Our proposed method does not preserve the multi-resolution property. In fact, in multi-resolution models, one compares a handful of models and can use AIC or BIC. But, multi-resolution models select knot locations that are uniformly distributed over the entire study region and do not use any optimal criteria for knot selection. In this research our main goal is to find an objective method for minimal number of knots along with its locations that are sufficient to model the spatial correlation adequately and lead to more accurate, as well as faster predictions.

The rest of the chapter is organized as follows: In Section (4.2), we develop an ascent based MCEM algorithm for estimating parameters of multivariate Laplace distribution that we impose on \( \eta \). Section (4.3) provides the details
CHAPTER 4. EMPIRICAL BAYESIAN KNOT SELECTION

of our proposed knot selection procedure. We apply our method on the simulated datasets in Section (4.4). Section (4.5) concludes the chapter with a discussion on the several aspects of the proposed methodology.

4.2 Ascent Based MCEM

We start this section with the specification of multivariate Laplace distribution on $\eta$ of Equation (1.2). Laplace-like (double-exponential) priors have heavier tails and higher spike near zero compared to Gaussian distributions. Thus, unimportant $\eta$’s approach to zero faster and less shrinkage is applied on important ones. This property of Laplace-like priors will help in knot selection. In Bayesian LASSO problem, these types of priors were used by many researchers including Bae & Mallick (2004) and Yuan & Lin (2005).

We are using multivariate Laplace distribution as $\eta$’s are spatially correlated. We formulate $\eta$ as: $\eta = \sqrt{W}U$, where $W$ follows an exponential distribution with mean 1 and $U$ follows a zero-mean multivariate Gaussian distribution with covariance matrix $V$. $W$ and $U$ are independently distributed. Under this formulation, $\eta$ follows a symmetric multivariate Laplace distribution with pdf:

$$f(\eta|V) = \frac{1}{(2\pi)^{r/2}} \frac{2K_{r/2-1} \left( \sqrt{2q(\eta)} \right)}{\left( \sqrt{(1/2)q(\eta)} \right)^{r/2-1}}, \eta \in \mathbb{R}, \quad (4.1)$$

where $K_m(\cdot)$ is modified Bessel function of the second kind and order $m$, and $q(\eta) = \eta'V^{-1}\eta$. On principle, one can assume an Exponential distribution
on $W$ with mean $\lambda^{-1}$, but in that case $\lambda$ would be nonidentifiable. Details on the properties and parameter estimation of (4.1) can be found in Eltoft, Kim, & Lee (2006) and Kotz, Kozubowski, & Podg (2003).

Figure (4.2a) shows the pdf of a zero mean bivariate Gaussian distribution with covariance matrix
\[
\begin{bmatrix}
1 & 0.5 \\
0.5 & 1
\end{bmatrix}
\]. Figure (4.2b) shows the pdf of a bivariate Laplace distribution that is symmetric about zero and the parameter $V$ is specified by the same covariance matrix we used to plot the bivariate Gaussian. It can be easily seen that the later has higher peak near the origin and have a slower decay rate compared to bivariate Gaussian distribution.
Note that multivariate Laplace distribution is a scale mixture of Gaussian random variables and thus provides extra flexibility to the model. An added benefit of multivariate Laplace is that it models datasets with non-Gaussian tails better than a Gaussian distribution.

Under the above formulation, conditioning on $W$ and $\eta$, $Z$ follows a multivariate Gaussian distribution with mean vector, $\mu + S\eta$, and covariance matrix $\nu^2I_n$. From now on, we specify $\mu$ by a linear function of covariates as $X\beta$. In what follows, we develop an ascent based MCEM algorithm for the parameters $\Theta \equiv (\beta, V, \nu^2)$ treating $W$ and $\eta$ as missing variables. We denote the conditional pdf of $Z$ by $f(Z|W, \eta, \Theta)$. Note that replacing the distribution of $\eta$ by multivariate Laplace does not increase the number of parameters of the model to be estimated. In MCEM algorithm, when the E-step is analytically intractable, it is instead evaluated numerically using samples of the missing variables. For details on this method, see Levine & Casella (2001) and Casella (2001). Monte Carlo sample size is very crucial for successful implementation of MCEM. Usually, a very large sample size is required for acceptable accuracy of the estimates. Ascent based MCEM developed by Caffo, Jank, & Jones (2005) is a data driven method that controls the Monte Carlo sample size in such a way so that the property of increasing likelihood remains preserved in M-step. Our discussion on ascent based MCEM is problem specific and we encourage the readers to see Caffo, Jank, & Jones (2005) for further details.

In our MCEM algorithm, we treat $\eta$ and $W$ as missing observations and
the $t^{th}$ iteration comprises of two steps, E-step and M-step. Basics of EM algorithm can be found in Dempster, Laird, & Rubin (1977). In E-step, we evaluate:

$$Q(\Theta, \Theta^{(t-1)}) = E_{f(\eta, W | Z, \Theta^{(t-1)})} \{ \log(f(Z, \eta, W | \Theta)) \}, \quad (4.2)$$

where $f(Z, \eta, W | \Theta)$ is the full data likelihood as:

$$f(Z, \eta, W | \Theta) = f(Z | \eta, W, \Theta) f(\eta | W, \Theta) f(W). \quad (4.3)$$

Recall that $f(Z|\eta, W, \Theta)$ is a multivariate Gaussian distribution described in the previous paragraph. $f(\eta|\Theta)$ is also a zero mean multivariate Gaussian distribution with covariance matrix $W \Sigma$. $f(W)$ is an exponential distribution with mean parameter 1. Here, $\Theta$ is the generic notation of all unknown parameters we want to estimate and $\Theta^{(t-1)}$ is the estimated value from $(t - 1)^{th}$ iteration. The expectation in Equation (4.2) is analytically intractable and we will evaluate it numerically using Markov Chain Monte Carlo samples of $\eta, W$ given the data $Z$ and parameter estimate $\Theta^{(t-1)}$. Specifically, we use a Gibbs Sampler that updates $W$ and $\eta$ in two steps:

- Generate $W$ from $f(W|\eta, Z, \Theta^{(t-1)})$,
- Generate $\eta$ from $f(\eta|W, Z, \Theta^{(t-1)})$. 
$f(W|\eta, Z, \Theta^{(t-1)})$ is a generalized inverse Gaussian distribution with pdf:

$$f(W|\eta, Z, \Theta^{(t-1)}) = \left(\frac{\psi}{\lambda}\right)^{\lambda} \frac{w^{\lambda-1}}{2\kappa_\lambda(\sqrt{\chi \psi})} \exp\left\{-0.5 \left(\frac{\chi}{w} + \psi w\right)\right\}, \ w > 0, \quad (4.4)$$

where $\psi = 2, \chi = \eta'(V^{(t-1)})^{-1}\eta, \lambda = 1 - r/2$, and $\kappa_\lambda(\cdot)$ is a modified Bessel function of third kind and order $\lambda$. $f(W|\eta, Z, \Theta^{(t-1)})$ is a multivariate Gaussian distribution with mean = $\bar{G}S'Z - X\beta^{(t-1)}/(\nu^{(t-1)})^2$ and variance, $\bar{G} = (S'S/(\nu^{(t-1)})^2 + (V^{(t-1)})^{-1}/W)^{-1}$. $V^{(t-1)}, \beta^{(t-1)}$, and $(\nu^{(t-1)})^2$ are estimates of $V, \beta$, and $\nu^2$ from $(t-1)^{th}$ iteration, respectively.

We generate $\tilde{m}_t$ MCMC samples of $(W, \eta)$ at the $t^{th}$ iteration of MCEM and proceed to the M-step. In M-step, we maximize $Q(\Theta, \Theta^{(t-1)})$ with respect to $\Theta$. The estimates of the parameters, $\tilde{\Theta}^{(t)}$, are obtained as:

$$\tilde{V}^{(t)} = \frac{1}{\tilde{m}_t} \sum_{i=1}^{\tilde{m}_t} \eta_i^i/W_i. \quad (4.5)$$

$$\tilde{\beta}^{(t)} = \frac{1}{\tilde{m}_t} \sum_{i=1}^{\tilde{m}_t} (X'X)^{-1} X'(Z - S\eta_i). \quad (4.6)$$

$$(\tilde{\nu}^{(t)})^2 = \frac{1}{\tilde{m}_t} \sum_{i=1}^{\tilde{m}_t} (1/n)(Z - X\tilde{\beta}^{(t)})' (Z - X\tilde{\beta}^{(t)}) - S\eta_i'. \quad (4.7)$$

Here $\{W_i, \eta_i, i = 1, \ldots, \tilde{m}_t\}$ denotes $\tilde{m}_t$ MCMC samples. Next, we compute a lower bound:

$$\text{LB}^{\tilde{m}_t} = \tilde{Q}(\tilde{\Theta}^{(t)}, \Theta^{(t-1)}) - \tilde{Q}(\Theta^{(t-1)}, \Theta^{(t-1)}) - z_\alpha\text{ASE}, \quad (4.8)$$
where \( \tilde{Q}(\cdot, \cdot) \) is a Monte Carlo estimate of \( Q(\cdot, \cdot) \) (Equation 4.2) using \( \{W_i, \eta_i, i = 1, \ldots, \tilde{m}_t\} \). ASE is an estimate of asymptotic standard error of

\[
\Delta \tilde{Q} = \tilde{Q}(\tilde{\Theta}^{(t)}, \Theta^{(t-1)}) - \tilde{Q}(\Theta^{(t-1)}, \Theta^{(t-1)})
\]

and \( z_\alpha \) is the upper \( \alpha^{th} \) percentile of the standard normal distribution. Usually, one selects \( \alpha = 0.25 \). We accept \( \tilde{\Theta}^{(t)} \) as the estimate of \( \Theta \) at the \( t^{th} \) iteration of MCEM and proceed to \( (t+1)^{th} \) iteration if \( \text{LB}^{\tilde{m}_t} > 0 \). Otherwise, we stay at the \( t^{th} \) iteration stage, generate more MCMC samples of \( (W, \eta) \), append them to the previously generated \( \tilde{m}_t \) samples, and repeat the steps from Equations (4.5).

As noted by Caffo, Jank, & Jones (2005), the ascent property may fail a finite number of times. Therefore, if \( \Delta \tilde{Q} < 0 \) (which naturally implies \( \text{LB}^{\tilde{m}_t} < 0 \)), we allow the algorithm to proceed to the \( (t+1)^{th} \) at most 4 iterations in a row. If the fifth iteration also failed to satisfy the ascent property, the iteration would be reset (throwing out the \( \tilde{m}_t \) samples and obtaining new samples). If the iteration was reset more than four times in a row, the algorithm would return to the \( (t - 1)^{th} \) iteration parameter values. In practice we observed that the early iterations more often failed to satisfy the ascent property, and rarely did an iteration need to be reset.

We denote by \( m_t \) the final sample size after \( t^{th} \) iteration, before proceeding to \( (t+1)^{th} \) iteration. In our simulation studies, we see that \( \beta \) and \( \nu^2 \) converge faster than \( V \). Hence, the MCEM algorithm is stopped when the average
Frobenius norm $\|V^{(t-1)} - V^{(t)}\|/r^2$ is sufficiently small and stabilizes. $V$ is a positive-definite matrix of rank $r$, and its estimation requires a sample size more than $r$. Within the $t^{th}$ iteration of the MCEM algorithm, we keep on increasing Monte Carlo sample size at an increment of $r$ until the lower bound in Equation (4.8) becomes positive before moving to the $(t + 1)^{th}$ iteration. Between the $t^{th}$ and $(t + 1)^{th}$ iteration, Monte Carlo sample size is enforced to increase by the similar technique described in Section 2.3 of Caffo, Jank, & Jones (2005).

We obtain an ASE using the regenerative simulation as described in Jones et al (2006). ASE is obtained from split chain MCMC using the formulas given in Section 2.2.2 of Caffo, Jank, & Jones (2005). The tours of a Markov Chain between two regenerations are independent. This feature is exploited while calculating ASE. If we use the Gibbs sampling described earlier, then we will not be able to detect the regenerations. Instead, we develop a split chain MCMC following the basic idea given in Mykland, Tierney, & Yu (1995).

Recall that in $(q+1)^{th}$ iteration of Gibbs sampling of $\{W, \eta\}$, we generate $W_{q+1}$ from $f(W|\eta_q, \text{rest})$ and $U_{q+1}$ is generated from $f(\eta|W_{q+1}, \text{rest})$. $W_{q+1}$ depends on the previous iteration through $\chi(\eta_q) = \eta_q V \eta_q$. After generating $\{W_{q+1}, \eta_{q+1}\}$, we generate an additional Bernoulli random variable $B_q$ with
success probability:

\[
p(q, q+1) = \begin{cases} 
\exp \left\{ (\tilde{\chi} - \chi(\eta_q)) \left( 1/d(\chi(\eta_q)) - 1/W_{q+1} \right) \right\} & \text{for } d_1 < W < d_2 \\
0 & \text{otherwise.} 
\end{cases}
\]  

(4.9)

where \( d(\chi(\eta_q)) = d_2 \) if \( \tilde{\chi} > \chi(\eta_q) \) and \( d(\chi(\eta_q)) = d_1 \) if \( \tilde{\chi} \leq \chi(\eta_q) \). Based on a short preliminary run of the Gibbs sampler (about 300 iterations), we set \( \tilde{\chi} \) at the posterior mean of \( \chi(\eta) \). \( d_i \)'s are chosen as \( \mu_W \pm 1.1 s_W \), where \( \mu_W \) and \( s_W \) are posterior mean and posterior standard deviation of \( W \) obtained from this preliminary run of the Gibbs sampler. The cut-off value in the previously mentioned intervals has been chosen to be 1.1 and that provides us with sufficient regenerations to obtain a reasonable estimate of ASE. If \( d_1 \leq 0 \), we set \( d_1 = \frac{1}{2} \min(W) \). Regeneration occurs at the \( q^{th} \) iteration if \( B_q = 1 \). Note that this split chain method makes use of the closed form full conditionals we obtained before and we can generate \( B_q \) right after generating \( \{W_{q+1}, \eta_{q+1}\} \). The joint posterior distribution \( f(W, \eta|Z, \Theta) \) is log-concave and our Gibbs sampling for \( (W, \eta) \) converges almost immediately. While computing ASE in each iteration of MCEM algorithm, Caffo, Jank, & Jones (2005) fixed the number of regenerations and ran the MCMC algorithm until that fixed number of regenerations occurred. In our Gibbs sampling for \( (W, \eta) \), we notice that average length of regeneration tours is between 5 to 10 iterations, pretty small. Hence, while implementing our MCEM algorithm, we do not need to fix the number of regenerations. If we run the chain for \( m_t \) times,
then number of regenerations we get is about $m_t/10$ to $m_t/5$, which is large enough to obtain a reasonable estimate of ASE. For example, if we use 3000 Monte Carlo samples in a particular iteration of our MCEM algorithm, then number of regenerations we get is around 30 to 60.

In general, starting values are crucial for faster convergence of EM algorithm. But ascent based MCEM is more robust with respect to its starting values. In all our applications we use the ordinary least squares estimate as starting values for $\beta$, $\frac{1}{n} \sum_{i=1}^{n} Z_i^2$ as starting value for $\nu^2$, and $(\frac{1}{r} \hat{\eta}' \hat{\eta})I_r$ as starting value for $V$, where $I_r$ is the identity matrix of rank $r$ and $\hat{\eta} = (S' S)^{-1} S (Y - X\beta)$.

### 4.3 Knot Selection using Bayesian LASSO

Once we get the estimates of $\Theta = \{\beta, V, \sigma^2\}$ from the algorithm described in the previous section, model fitting is completed. But our goal here is not to fit a reduced rank spatial model using multivariate Laplace distribution, rather developing an objective method for selecting few important knots from $r$ equispaced knots that will lead to modeling spatial covariance with adequate accuracy. Rewriting Equations (1.1) and (1.2), we get:

$$Z = X\beta + SV^{1/2} \sqrt{W} \hat{U} + e,$$

(4.10)
where \( e = \delta + \epsilon \), zero mean iid Gaussian random variables with variance \( \nu^2 \).

Recall that \( \nu^2 \) can be referred as nugget variance. \( \tilde{U} \equiv \{ \tilde{U}_i, i = 1, \ldots, r \} \) are independent standard Gaussian random variables. Denoting \( \sqrt{W} \tilde{U}_i \) by \( \gamma_i \), we get \( \mathbb{E}\{\gamma_i\} = 0 \), \( \text{Var}\{\gamma_i\} = 1 \), and \( \text{Cov}(\gamma_i, \gamma_j) = 0 \). In fact, \( \gamma \equiv \{\gamma_1, \ldots, \gamma_r\} \) are uncorrelated Laplace (or double-exponential) random variables. Our Laplace model definitely provides some shrinkage on \( \eta \), but we do not have control over smoothness of our model. Hence, for knot selection, we impose a conditional Laplace distribution on \( \gamma \), conditioning on \( \nu^2 \) of the form:

\[
    f_\gamma(\gamma|\nu^2) = \prod_{j=1}^{r} \frac{\lambda}{2\sqrt{\nu^2}} \exp\{-\lambda|\gamma_j|/\sqrt{\nu^2}\}, -\infty < \gamma_j < \infty. \tag{4.11}
\]

Here, \( \lambda > 0 \) is the smoothing parameter, often referred as penalization parameter. Large values of \( \lambda \) will lead to smoother models (fewer knots) and small values will result in coarser model (more knots). This type of distribution is widely used in Bayesian LASSO regression, see Park & Casella (2008), Hans (2009), Hans (2010), and Yuan & Lin (2005) for details and discussion. Conditioning on \( \nu^2 \) guards the posterior distribution of \( \gamma \) against being multimodal. A multimodal posterior distribution can slow down MCMC, and at the same time complicates the interpretability of the model, see Park & Casella (2008).

LASSO is very popular for variable selection in linear regression and was first proposed by Tibshirani (1996). Later the methodology has been extended to logistic regression (Meier, van de Geer, & Buhlmann (2008)).
and Cox proportional hazard model (Tibshirani (1997), Gui % Li (2005)). In Bayesian paradigm, LASSO estimates are obtained as posterior modes of regression coefficients by imposing Laplace prior distributions (Equation 4.11) on them. Since using Laplace distributions directly may make the MCMC algorithm cumbersome, Park & Casella (2008) instead used the scale mixtures of Gaussian distributions, where the scale parameters follow exponential distributions. After the scale parameters are integrated out, the resulting marginal distributions become Laplace. We use a similar strategy. Hans (2009) used orthant-truncated Gaussian distributions and orthant integrals to specify prior distributions and developed a Gibbs sampling algorithm for obtaining posterior predictive distributions for predicting future observations. Laplace-like (double-exponential) priors have heavier tails and a higher spike near zero compared to Gaussian distributions. Thus, unimportant coefficients approach to zero faster and less shrinkage is applied on larger coefficients. Lykou & Ntzoufras (2012) pointed out that using Laplace-like priors on regression coefficients facilitates the shrinkage but fails to provide a precise decision on variable selection. In regular LASSO, coefficients for unimportant covariates are zeroed out, where as in Bayesian LASSO, MCMC algorithms are constructed to obtain samples from the target posterior, from which Bayesian credible intervals are constructed for each of the regression coefficients. However, decisions based upon these credible intervals are prone to subjectivity. Realizing this problem, Yuan & Lin (2005) and Hans (2010) used mixture priors of double-exponential distribution and degenerate dis-
distribution at zero. Following Kuo & Mallick (1998), Lykou & Ntzoufras (2012) developed a more objective method for Bayesian variable selection by introducing a set of indicator variables that follow Bernoulli distributions. Based on the posterior distributions of these indicator variables, one decides which variables to include in the model. We use similar techniques here for selecting knot locations in our reduced rank SRE model. In what follows, we rewrite Equation (4.10) as:

\[ Z = X\beta + SV^{1/2}\Gamma\gamma + e, \] (4.12)

where \( \Gamma \) is a diagonal matrix of the form \( \text{diag}(\alpha_1, \ldots, \alpha_p) \) and \( \alpha_1, \ldots, \alpha_r \) are independent and identical Bernoulli random variables with success probability 0.5. Now, denoting \( \alpha_j \gamma_j \) by \( \theta_j \), a priori distribution of \( \gamma_j \) is mixtures of point mass at zero and Laplace distribution with mixing probability 0.5. Following Kuo & Mallick (1998), this prior specification will avoid evaluation of posterior probabilities of all \( 2^r \) possible models and identifying the important knot locations. As Kuo & Mallick (1998) pointed out that the concern about the identifiability of \( \alpha_j \) and \( \gamma_j \) can be eased by noting that for knot selection, we are interested in the posterior summaries of \( \theta_j \), which are always identifiable. We will make our decision by monitoring the posterior inclusion probabilities (i.e. the posterior distributions of \( \alpha'_j \)'s) and posterior medians of \( \theta_j \). If the posterior median is zero, then the \( j^{\text{th}} \) knot will not be selected. Lykou & Ntzoufras (2012) used similar techniques for Gaussian
linear regression, but one important difference in our approach is that we impose Laplace priors in two hierarchies - Gaussian distribution and Exponential distribution. Where as, Lykou & Ntzoufras (2012) imposed Laplace priors directly on regression coefficients and in their Gibbs Sampling they updated each regression coefficient one at a time. In our MCMC algorithm, we block update $\gamma$. Since, $\gamma_j's$ are linearly related in Equation (4.12), block updating provides better mixing and convergence of the MCMC algorithm by taking into account the cross-correlation efficiently. Also, block updating speeds up the Gibbs sampling.

In SRE model, $X\beta$ characterizes the large scale variation of the process under analysis. We impose zero-mean independent Gaussian prior distributions on $\beta$ with variance $10^6$. An uniform prior distribution is assumed for $\nu^2$ on the domain $(0, a)$, where the upper bound $a$ can be selected to be sufficiently large (e.g., 100) to make the prior non-informative. Another choice for prior on $\nu^2$ could be inverse-gamma prior. However, Gelman (2006) cautioned that using small shape and scale parameters, like, (0.001 and 0.001), puts more prior weight near zero and is not diffuse.

In a full Bayesian analysis, one would also impose prior distributions on the smoothness parameters $\lambda$. While in a linear regression problem on LASSO, a common practice is to impose gamma priors on the tuning parameters, see Park & Casella (2008) and Lykou & Ntzoufras (2012). Lykou & Ntzoufras (2012) pointed out that noninformative gamma priors can lead to the Lindley-Bartlett paradox or ‘over-shrink’ model. Under the Lindley-
Bartlett paradox, posterior inclusion probabilities degenerate to zero, where as under an over-shrink model, these probabilities converge to 0.5. Lykou & Ntzoufras (2012) suggested monitoring Bayes factors to select appropriate hyper-parameters for the gamma distribution. Park & Casella (2008) specified these parameters based on a Monte Carlo Expectation Maximization (MCEM) estimate of the smoothness parameter. For probit regression, Bae & Mallick (2004) did not impose any prior distribution on the the tuning parameter, but fixed the value in such a way that the prior variances of the regression coefficients were sufficiently large. For linear regression problem, in the seminal work on LASSO Tibshirani (1996) selected the tuning parameter using cross validation, generalized cross validation, and Stein’s unbiased estimate of risk. All three methods require a grid search for finding the optimal tuning parameter. The most important message here is that one should adopt a method for specifying $\lambda$ (or the prior distribution of $\lambda$) that will guard against over-smoothing. If we assume a Gamma prior on $\lambda^2$ with shape parameter $c$ and scale parameter $d$, then the unconditional variance of $\gamma_j$ becomes:

$$\text{Var}\{\gamma_j\} = E\{\text{Var}\{\gamma_j|\lambda\}\} + \text{Var}\{E\{\gamma_j|\lambda\}\} = \frac{2\nu^2}{(c-1)d}. \quad (4.13)$$

Note that $\gamma_j$'s in the SRE model that we fitted in the previous section using ascent based MCEM algorithm have unit variance. Now, if we set $(2\hat{\nu}^2)/((c-1)d) = 1$, where $\hat{\nu}^2$ is the estimated $\nu^2$ from MCEM algorithm, then the mode
of prior distribution of $\lambda^2$ becomes $2n\hat{u}^2$. We still have two unknowns ($c$ and $d$) to solve from this equation. For better mixing of MCMC samples, we recommend that the shape and scale parameters should be chosen in such a way so that the prior of $\lambda^2$ approaches zero relatively fast as $\lambda$ goes to $\infty$. This will also guard against the selected model being over-smoothed.

We derive the Gibbs sampling algorithm for $\beta, \gamma, \alpha, \nu^2$, and $\lambda$, assuming $V$ to be known and we set it at its MCEM estimate (say $\hat{V}$) from the final iteration of our algorithm. For ease of Gibbs sampling, Laplace distribution on $\gamma$ is specified in hierarchies of Gaussian and exponential distributions as:

$$
\gamma | \nu^2, \rho^2_1, \ldots, \rho^2_r \sim \text{Gau} (0, \nu^2 \text{diag}\{\rho^2_1, \ldots, \rho^2_r\}) \quad (4.14)
$$

$$
\rho^2_1, \ldots, \rho^2_r \sim \prod_{j=1}^r \frac{\lambda^2}{2} \exp\{-\lambda^2 \rho^2_j / 2\} \quad (4.15)
$$

Once the augmented variables $\rho \equiv \{\rho^2_1, \ldots, \rho^2_r\}$ are integrated out from the hierarchical distributions in Equation (4.14), the resulting distributions of $\gamma_j$'s are Laplace of the form given in Equation (4.11).

In the derived Gibbs sampling algorithm, the full conditional distribution for $\beta$ is multivariate Gaussian with covariance matrix $\Sigma_{\beta|\cdot} = (X'X/\nu^2 + 10^{-6}I_r)^{-1}$ and mean vector $\mu_{\beta|\cdot} = \Sigma_{\beta|\cdot} X'(Z - \hat{S}\Gamma\gamma)/nu^2$, where $\hat{S} = \hat{S}\hat{V}^{-1/2}$. The full conditional distribution for $\gamma$ is also multivariate Gaussian with covariance matrix $\Sigma_{\gamma|\cdot} = \nu^2 \left(\Gamma\hat{S}'\hat{S}\Gamma + \text{diag}(\rho^{-2}_1, \ldots, \rho^{-2}_r)\right)^{-1}$ and mean vector $\mu_{\gamma|\cdot} = \Sigma_{\gamma|\cdot} \Gamma\hat{S}'(Z - X\beta))/\nu^2$. The full conditional distribution for $\nu^2$ is Truncated Inverse Gamma distribution over the range $(0, a)$ with shape parameter
\( n/2 + r/2 - 1 \) and scale parameter \( (\mathbf{Z} - \mathbf{X}\beta - \hat{\mathbf{S}}\mathbf{\gamma})'(\mathbf{Z} - \mathbf{X}\beta - \hat{\mathbf{S}}\mathbf{\gamma})/2 + \sum_{j=1}^{r} \gamma_j^2/\rho_j^2 \). The full conditional distributions of augmented variables \( \rho_j^{-2} \) are inverse Gaussian distribution of the form:

\[
f(x|\theta_1, \theta_2) = \sqrt{\frac{\theta_1}{2\pi x^{-3/2}}} \exp\left\{ -\frac{\theta_1(x - \theta_2)^2}{2\theta_2^2 x} \right\}, \quad x > 0,
\]

with \( \theta_1 = \lambda^2 \) and \( \theta_2 = \sqrt{\lambda^2 \nu^2 / \gamma_j^2} \). The full conditional distributions of \( \alpha_j \)'s are also conditionally independent Bernoulli distributions with probability of success \( \pi_j \) where

\[
\pi_j = \frac{f(\mathbf{Z}|\beta, \gamma, \nu^2, \alpha_{-j}, \alpha_j = 1)}{f(\mathbf{Z}|\beta, \gamma, \nu^2, \alpha_{-j}, \alpha_j = 1) + f(\mathbf{Z}|\beta, \gamma, \nu^2, \alpha_{-j}, \alpha_j = 0)},
\]

where \( f(\mathbf{Z}|\beta, \gamma, \nu^2, \alpha_{-j}, \alpha_j = 1) \) is pdf of a multivariate Gaussian distribution with mean vector \( \mathbf{X}\beta + \hat{\mathbf{S}}\mathbf{\gamma} \) setting \( \alpha_j = 1 \) in \( \mathbf{\Gamma} \) matrix, and covariance matrix \( \nu^2 \mathbf{I}_n \). Similarly, \( f(\mathbf{Z}|\beta, \gamma, \nu^2, \alpha_{-j}, \alpha_j = 0) \) denotes the same setting \( \alpha_j = 0 \). The notation \( \alpha_{-j} \) denotes \( \alpha \) vector excluding \( \alpha_j \).

Finally, the full conditional distribution for \( \lambda^2 \) is Gamma distribution with shape parameter \( r + c \) and scale parameter \( \left(1/d + \sum_{j=1}^{r} \rho_j^2/2\right)^{-1} \).

We implement our ascent based MCEM and Gibbs sampling for knot selection in R software version 2.14.2. Convergence of our Gibbs sampling is very fast and assessed using standard techniques in \texttt{CODA} package.
4.4 Applications

To assess the performance of our ascent based MCEM algorithm, we apply it on simulated datasets based on the model specified in Equations (1.1) and (1.2) of size \( n = 10000 \) on a unit square. 121 equispaced knots are selected on this area in an \( 11 \times 11 \) grid. A modified bisquare function of the following form is used to specify \( S \) matrix, where the \((i,j)\)th element is:

\[
S(i,j) = \begin{cases} 
1 - d_{ij} & \text{for } d_{ij} \leq 2 \\
0 & \text{otherwise,}
\end{cases}
\]  

\[ (4.18) \]

\[
d_{ij} = \sqrt{\left(\frac{v_{i,1} - u_{j,1}}{a_1^2}\right)^2 + \left(\frac{v_{i,2} - u_{j,2}}{a_2^2}\right)^2},
\]

\[ (4.19) \]

where \( v_{i,1} \) and \( u_{j,1} \) are horizontal coordinates of the \( i \)th observation and \( j \)th knot, respectively, and \( v_{i,2} \) and \( u_{j,2} \) are vertical coordinates of the \( i \)th observation and \( j \)th knot, respectively. \( a_1 \) and \( a_2 \) are the shortest distances along horizontal and vertical directions between knots. Here we are using equispaced knots, hence \( a_1 = 1/11 \) and \( a_2 = 1/11 \). This specification of \( S \) matrix allows one to model geometric anisotropy along different directions. For the three simulated datasets, we selected 121, 25, and 4 of these knot locations, respectively. The \( S \) matrix for each was obtained by removing the columns of non-selected knots.

The \( V \) matrix is specified by exponential covariance function of the form
CHAPTER 4. EMPirical BayesIn Kn ot Sel ection

\[ \exp\{-\theta h_{jj'}\} \]

where \( h \) is the Euclidian distance between \( j^{th} \) and \( j'^{th} \) knot locations. The large scale variation is specified in terms of linear regression on horizontal coordinates and vertical coordinates as \( \beta_1 v_{i,1} + \beta_2 v_{i,2} \). In all three simulations, we set \( \beta_1 = 2, \beta_2 = 1, \) and \( \nu^2 = 0.001 \).

For the first (coarsest) simulation, our MCEM algorithm converged in 25 iterations with an average Frobenius norm of 0.0002 between the estimates of \( V \) in final two iterations. Figure (4.4)(a) shows the average Frobenius norm versus the number of iterations. From this plot we see initial erratic behavior (which is very common in MCEM algorithms) and after 10\( ^{th} \) iteration the plot stabilizes. For the first simulation, the estimated values for \( \beta_1 \) and \( \beta_2 \) from the final iteration are \( \hat{\beta} = (0.854, 0.796) \), and the value of \( \nu^2 \) was 0.221. For the second simulation, \( \hat{\beta} = (1.869, 0.831) \) and \( \hat{\nu^2} = 0.0311 \). And for the third simulation, \( \hat{\beta} = (2.045, 1.023) \) and \( \hat{\nu^2} = 0.0144 \). Note that these values become progressively closer to the true values. This is to be expected, as the first simulation was the most coarse, and the third simulation was the most smooth. The coarser the spatial process, the more the estimates will deviate from the true values. This is also evident in the results.

Figure (4.4)(b) plots the final Monte Carlo sample size of each iteration. From this plot, we see that the algorithm takes longer time at 10\( ^{th} \) and at 20\( ^{th} \) iterations to satisfy ascent property. The initial sample size used is 300 and the final Monte Carlo sample size is 7400 at 25\( ^{th} \) iteration. Figure (4.4)(c) plots average number of regenerations at each iteration. This plot clearly indicates that we have sufficiently large number of regenerations to
Figure 4.3: Diagnostic results from the MCEM procedure.
estimate ASE adequately. Figure (4.4)(d) shows histogram of average tour length between two regenerations within an iteration. From this plot we see that a typical tour length is between 5 to 10 runs of our Gibbs sampler.

Following knot selection, the knots were no longer equispaced. As a result, it is possible for some locations to not be associated with any knots, and be modeled exclusively by the large scale variations. We redesign the S matrix following knot selection to allow knots to take on irregular shapes (more than the modified bisquare function is able to produce). Letting $r^*$ be the selected number of knots, the new $S$ is defined by:

1. Calculate $d_{i,j} = ||s_i - u_j||; i \in 1, \ldots, n, j \in 1, \ldots, r^*$ the distances between each location $s_i$ and each knot $u_j$.

2. For each location $s_i$, determine the closest knot $u_j$ and obtain that distance, $\min(d_i)$. For that location, set any $d_{i,j} > 1.5 \min(d_i)$ to zero.

3. For each knot $u_j$, obtain $\max(d_j)$ and scale the distances by $d_{i,j}^* = d_{i,j} / \max(d_j)$.

4. $S_{i,j} = (1 - d_{i,j}^*)^2 \times I(d_{i,j}^* > 0)$, where $I(\cdot)$ is the indicator function.

Note that based on the locations of selected knots, the shape of the basis function may exhibit ‘panhandles.’ The condition in step 2, defining which locations are associated with which knots, may be made more restrictive to eliminate these. Figure 4.4(a-d) shows a plot of the selected knots (o) for the coarse simulation, as well as the area associated with some selected knots (▲).
Figure 4.4: Shapes of basis functions for four arbitrary knot locations from the coarse simulation. Knots are blue circles, the chosen knot is a red triangle. Locations associated with the chosen knot are small black circles.
After redefining $S$, final estimates of $V$ and $\nu^2$ were obtained using MOM as described in section 2.4 and kriging was performed. Figure 4.4 shows scatterplots of the predicted values vs. the true values for each of the three simulated spatial processes shown in Figure 4.1. The left column (panels (a), (c), and (e)) are results from using the knot-selection algorithm described in this chapter on the coarse, medium, and smooth processes, respectively. We expect to see the increasing accuracy observed, as smooth processes are more easily modeled. The right-hand column corresponds to using all of the original knots. These are more accurate predictions, but again this is an expected result; using more knots will permit the model to capture more variability. Note how, as the process becomes more smooth, the reduced number of knots becomes nearly equivalent to using the full number of knots. Furthermore, the right-hand columns uses the same knots as were used to generate the data. In practice, the ‘true’ knots and basis function of the spatial process will not be known, so all models will perform less efficiently.

4.5 Conclusions and Discussion

In this chapter we have proposed a modification of the distribution typically assumed for the random component of the SME model. Instead of assuming a multivariate normal distribution, we assume a multivariate Laplace with the introduction of $\sqrt{W}$. This permits the model to account for heavier tails in the distribution. We derive and implement an MCEM algorithm to success-
Figure 4.5: Plots of $\hat{Y}$ vs. $Y$ using only the selected knots for the the (a) coarse, (c) medium, and (e) smooth simulated processes. Panels (b), (d), and (f) show the same, when using all of the knots.
fully estimate the large scale variation slopes, the reduced rank covariance matrix, and the nugget variance.

Following this we interpret knot selection as a variable selection technique, and apply Bayesian LASSO to select only those knots that are necessary for predicting the spatial process. As we observed, prediction accuracy will diminish as fewer knots are selected. However, the real question of importance is the degree of uncertainty or error permissible. We allow a smoothness parameters $\lambda$ which can control over or under-smoothing. Over-smoothing will lead to greater prediction errors while under-smoothing will lead to greater computational burden.
Chapter 5

A Flexible Class of Reduced Rank Spatial Models for Non-Gaussian Tail Behaviors

5.1 Introduction

While a significant amount of research has been done on Gaussian RRSM, very limited research was conducted on non-Gaussian datasets. For example, Paciorek (2007) conducted research on developing computational techniques for spatial logistic regression for large datasets, Sengupta & Cressie (2013) developed RRSM for Poisson count data, and Sengupta & Cressie (2013) developed a hierarchical model that comprises of a conditional exponential family model for the data and a latent geostatistical process model for some
Datasets exhibiting non-Gaussian tails are very common in environmental studies. Both tails can be heavier or lighter than normal, or one tail can be lighter and the other heavier. If one uses Gaussian reduced rank spatial model, then likelihood based approaches, such as EM algorithm, will suffer due to wrong distributional assumption, and can lead to poor estimation and predictions. Instead of assuming Gaussian distribution on the reduced dimensional latent process, we propose a model where scale mixtures of Gaussian distributions are imposed and the scale parameters are exponentially distributed random variables with mean 1. This model is flexible enough to handle non-Gaussian tail behaviors. The proposed model belongs to exponential family. Estimation of model parameters and predictions at unobserved locations are done using a MCMC algorithm. Full conditional distributions for all processes under consideration and model parameters are not available in closed form. Hence, a hybrid MCMC algorithm is constructed. Markov Chain Central Limit Theorem (MCCLT) based posterior summaries alleviate the storing problem of large MCMC generated samples to a great extent. The conditions required for MCCLT are verified theoretically. See Jones et al (2006) for more general discussion on MCCLT and their applications. The proposed method is applied on several simulated datasets and on real datasets on atmospheric properties that are publicly available from NASA’s (National Aeronautic and Space Administration) Terra satellite.

The rest of the chapter is organized as follows: Section (5.2) provides the...
details of our modeling approach. Section (5.3) discusses the derived MCMC algorithm and its implementation. In Section (5.4), we verify the conditions required for MCCLT. In Section (5.5), we apply our proposed model to several simulated dataset and a real dataset on ozone. Section (5.6) concludes the paper with discussion on more general applicability of our proposed method.

5.2 Flexible Reduced Rank Spatial Model

As mentioned, there is a vast amount of literature that addressed issues with model formulation and model fitting for Gaussian spatial processes, not much work has been done to improve accuracy for datasets that depart from Gaussian assumptions. One important issue is the tail behavior. Simple QQ plots of detailed residuals can easily detect departures from Gaussian distribution. Detailed residuals are defined as: $Z - X \hat{\beta}_{OLS}$, where $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$, the ordinary least squares estimate of $\beta$. In (Katzfuss & Cressie, 2011 a), while analyzing CO2 data, the authors showed through a QQ plot and density estimation plot that the detailed residuals exhibit right skewness. They also discussed that simple transformations were unable to provide any improvement. Datasets were still analyzed with Gaussian models arguing that kriging estimates are still best linear unbiased predictor (BLUP), conditioning on the estimates of $V$ and $\nu^2$ obtained from EM algorithm. When distributional assumptions are wrong, likelihood based approaches such as EM algorithm suffer, and do not provide accurate estimates and make the predic-
tions unreliable. One can use distribution free estimation techniques, such as binned method of moments (MOM) estimation (see (Cressie & Johannes-son, 2008) and for more a robust version chapter 3 of this dissertation). But binned estimation does not provide a ready positive-definite estimate of $\mathbf{V}$ and requires eigenvalue adjustment. While analyzing total ozone maximum (TOM) data for April 22, 2012 obtained by NASA’s Terra satellite, we observed that EM algorithm and Gaussian reduced rank spatial model performed poorly. Figure (5.1) shows the probability plot of detailed residuals of log-transformed TOM data. One can clearly see from this plot that the upper tail contains many outliers. This is just one example, while analyzing datasets on environmental processes very often one faces situations where one tail is lighter and the other is heavier than Gaussian distribution, or both tails are lighter/heavier than Gaussian distribution. In what follows, a flexible class of reduced rank spatial models will be proposed. The proposed model can handle these different tail behaviors with lots of flexibility. Instead of Gaussian distribution on $\eta$, scale mixtures of Gaussian distributions will be imposed, where the scale parameters follow exponential distributions with mean 1.

**PROPOSED RRSM:** We start from the model defined by (1.1) and (1.2). Instead of a Gaussian distribution on $\eta$, we propose to model, $\eta = A^{1/2}U$, where $U$ is a zero mean $r$-dimensional Gaussian random variable with co-variance matrix $\mathbf{V}$. $A$ is a diagonal matrix with diagonal elements as $W = \{W_1, \ldots, W_r\}$, where the scale parameters $W_1, \ldots, W_r$ are identically and in-
dependent on exponentially distributed random variables with mean 1. Other distributional assumptions in Equation (1.2) remain same. This distribution is motivated by multivariate symmetric Laplace distribution. Note that this model specification does not increase the number of unknown parameters, it still remains the same, $\beta, V,$ and $\tau^2$. The joint distribution of $\{Z, U, W\}$ belongs to exponential family and is log-concave. Large scale variation of the spatial process under analysis remains same as before, $X\beta$. To illustrate the flexibility of the proposed model, several datasets are generated on a spatial domain of unit square, setting $\beta = 0$. In the simulation study, $S$ is specified by Gaussian kernel function with location parameter 0 and scale pa-
rameter 0.08. \( V \) matrix is constructed using exponential covariance function 
\((\theta^2 \exp\{-\lambda h\})\) with range parameter \( \lambda^{-1} \) and variance parameter \( \theta^2 \). Figure (5.2) represents several QQ plots for different values of \( \lambda, \theta^2, \) and \( \sigma^2 \). From these plots one can easily see that the proposed model is flexible enough to handle several types of tail behaviors. Just for illustration purpose, exponential covariance function for \( V \) is assumed. Actual analysis will be performed for any positive definite \( V \) without imposing any specific structure. Hence, the flexibility and broader applicability of the proposed model can be easily perceived from these QQ plots.

Under our model assumptions, the latent spatial process \( \eta \) has zero mean and closed form expression for covariance matrix. Note that,

\[
\mathbb{E}(\eta) = \mathbb{E}_{W,U}(A^{1/2}U) = \mathbb{E}_W(A^{1/2})\mathbb{E}_U(U)
\]

\[
= \begin{bmatrix}
\mathbb{E}_{W_1}(\sqrt{W_1})\mathbb{E}_{U_1}(U_1), \ldots, \mathbb{E}_{W_r}(\sqrt{W_r})\mathbb{E}_{U_r}(U_r)
\end{bmatrix}'
\]

\[
= \begin{bmatrix}
\mathbb{E}_{W_1}(\sqrt{W_1}) \times 0, \ldots, \mathbb{E}_{W_r}(\sqrt{W_r}) \times 0
\end{bmatrix}'
\]

\[
= [0, \ldots, 0]'
\]

The second equality in Equation (5.1) comes from the independence of \( W_j \) and \( U_j \), and the third equality comes from the fact that means of \( U_j \) are all zero.

The covariance matrix of \( \eta \) is same as \( \mathbb{E}_{W,U}(\eta\eta') \) and can be simplified
as follows:

\[
E(\eta') = E_{W,U}(A^{\frac{1}{2}}UU'A^{\frac{1}{2}}) = E_{W}\left(E_{U|W}(A^{\frac{1}{2}}UU'A^{\frac{1}{2}}|W)\right) = E_{W}\left(A^{\frac{1}{2}}VA^{\frac{1}{2}}\right)
\]

\[
= E_{W} \begin{bmatrix}
W_{1}V(1,1) & W_{1}^{\frac{3}{2}}V(1,2)W_{2}^{\frac{1}{2}} & \ldots & W_{1}^{\frac{3}{2}}V(1,r)W_{r}^{\frac{1}{2}} \\
\vdots & \vdots & \ddots & \vdots \\
W_{r}^{\frac{3}{2}}V(r,1)W_{1}^{\frac{1}{2}} & W_{r}^{\frac{3}{2}}V(r,2)W_{2}^{\frac{1}{2}} & \ldots & W_{r}V(r,r)
\end{bmatrix}
= \begin{bmatrix}
V(1,1) & \frac{\pi}{4}V(1,2) & \ldots & \frac{\pi}{4}V(1,r) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\pi}{4}V(r,1) & \frac{\pi}{4}V(r,2) & \ldots & V(r,r)
\end{bmatrix}
\]

In our notation, \(V(i,j)\) denotes the \((i,j)\)th element in \(V\) matrix. The last equality in Equation (5.2) follows from the fact that \(W_i\) and \(W_j\) are independent, \(E(W_i) = 1\), and \(E(\sqrt{W_i}) = \sqrt{\pi}/2\).

**Prior Distributions:** While specifying prior distributions on model parameters, we ensure that they are noninformative or weakly informative, so that the posterior results are not sensitive to the choice of hyperparameters.

We assume zero mean *Gaussian* prior for \(\beta\) with diagonal covariance matrix having all of them large variance as \(10^6\).

Half-t distribution is imposed on the square root of \(\tau^2\) with hyperparameters 2 and \(10^5\). Gelman (2006) and Huang & Wand (2013) discuss that this
prior distribution is highly noninformative and has better behavior near zero compared to Inverse-Gamma distribution with very small shape and scale parameters that puts unnecessary prior weight near small values. Half-t prior is usually cumbersome to deal with in MCMC implementations, but it can
be specified in hierarchies of two Inverse-Gamma distributions, see (Huang & Wand, 2013). We assume:

\[
\tau^2|a \sim \text{Inverse-Gamma} \left(1, a/2 \right), \tag{5.3}
\]
\[
a \sim \text{Inverse-Gamma} \left(1/2, 10^{-10} \right), \tag{5.4}
\]

where $\text{Inverse-Gamma}(b_1, b_2)$ denotes an Inverse-Gamma distribution with shape parameter $b_1$ and scale parameter $b_2$. Under this model specification, the marginal distribution (integrating over $a$) of $\tau$ is $\text{Half-t}(2, 10^5)$. This prior distribution achieves a high degree of noninformativity and better than Inverse-Gamma or Uniform prior. See Huang & Wand (2013) for more details. This two-level prior specification preserves conjugacy and the full conditional distributions of $\tau^2$ and $a$ are Inverse-Gamma distributions.

Specifying priors for covariance matrices poses multiple challenges (see Yang & Berger, 1994). It is common in geostatistics to have only one single realization of the spatial process, hence consistent estimation of the covariance matrix is impossible. One instead hopes to obtain a covariance matrix which, while it may not be the same as the true covariance matrix, will still allow for faithful predictions, as well as reasonable interpretations. Another problem is that of specifying a noninformative prior on covariance matrices.

In the existing literature, one can find at least three ways of specifying priors on the covariance matrix of low rank latent process, in our notation $\mathbf{V}$. 
In their Gaussian process model, Higdon (1998) assumed that the underlying latent process is identically and independently distributed zero mean Gaussian distribution with a common variance parameter. This assumption compromises flexibility as the spatial covariance of the process under analysis is modeled only by basis function matrices, $S$. In the Gaussian multivariate spatial model, (Banerjee et al., 2008) imposed an inverse-Wishart prior for modeling cross covariances and spatial covariance is specified by stationary Matérn covariance functions. They considered parameters in the Matérn model unknown and put priors on them. While Matérn models reduce the number of parameters, the range and smoothness parameters are somewhat delicate to deal with under Bayesian estimation as they are weakly identifiable. (Stein, 2005) pointed out that data cannot detect the smoothness parameter in Matérn covariance of order higher than 3. Further, Matérn class can result in an overly smooth model. (Sang & Huang, 2012) overcame the smoothness problem by amending the Matérn covariance with another compactly supported covariance through tapering. (Kang & Cressie, 2011) considered more general covariance function and imposed priors on eigenvalues and Givens angles of $V$, the technique proposed by (Daniels & Kass, 1999). Their hyperparameters for priors on eigenvalues and Givens angles were specified by method of moments estimator of $V$. The main issue with this prior specification is that Givens angles priors can slow down MCMC implementation and their priors are data dependent.

We use the two level hierarchical priors on the $V$ matrix proposed by
Our prior is not data dependent and marginally noninformative, and our MCMC implementation is comparatively faster.

(Gelman, 2006) noted that inverse-Gamma priors for variance parameters with small shape and scale hyperparameters give too large of weight to small values, and are not noninformative. They recommended instead the Half-$t$ prior which we used for $\tau^2$. (Huang & Wand, 2013) extend this two-level hierarchy to provide a marginally non-informative prior for covariance matrices. We specify the prior on $V$ as,

$$V|a_1, \ldots, a_r \sim \text{Inv-Wishart} \left( \lambda + r - 1, 2\lambda \text{diag} \left( \frac{1}{a_1}, \ldots, \frac{1}{a_r} \right) \right) \quad (5.5)$$

$$a_k \overset{\text{ind}}{\sim} \text{Inv-Gamma} \left( 1/2, 1/B_k^2 \right), \ k = 1, \ldots, r \quad (5.6)$$

When $B_k$ takes on arbitrarily large values (e.g. $B_k = 10^5$), and $\lambda = 2$, this specification leads to the marginally noninformative prior where all variance parameters in $V$ are Half-$t(2, 10^5)$, and all correlation parameters are uniform on $(-1, 1)$. Therefore this is a highly noninformative prior marginally, though joint noninformativity is hard to assess as the closed form expressions are not available. (Huang & Wand, 2013) showed that any sub-matrice of $V$ belongs to the same family of distributions under this prior specification, and like Givens angles priors, multi-resolution feature is preserved. Like Givens angles priors of (Kang & Cressie, 2011) our two stage inverse Wishart prior takes into account anisotropy and nonstationarity and
they are more general in nature, bringing in a nonparametric flavor in co-
variance estimation.

Finally, our parameter set $\Omega$ consists of $(\beta, \tau^2, V)$ and latent processes $U, W, \delta$. The posterior distribution is unavailable in closed form and in
the following section, we develop a MCMC algorithm for model fitting and
predictions at unsampled locations.

## 5.3 MCMC Algorithm and its Implementation

We derive the full conditional distributions for all our unknown parameters
in the model and latent processes. We block update several parameters for
better mixing and convergence, as block updating handles the correlations of
Markov Chains better than individual updating. We implement our derived
algorithm in R software and use the packages MASS, MCMCpack, and HI.

We assume that the measurement error variance $\sigma^2$, is known, in fact, we
estimate it offline using a data analytic approach. First we select $n_0$ of the
observed locations, and define a small bin ($2^\circ$ distance in all directions from
each point). Then we calculate $\sigma^2_M = \{\sigma^2_1, \ldots, \sigma^2_{n_0}\}$, the variance of $Z(s)$
within each of these bins, and compute as our estimate of the measurement
error, $\hat{\sigma}^2 = \text{med}(\sigma^2_M)$.

For hyperparameters, we set $B_k = 10^5, k = 1, \ldots, r$ and $\lambda = 2$. The
full conditional distribution for $V$ is inverse-Wishart distribution with shape
parameter $\lambda + r$ and scale matrix

$$2\lambda \text{diag}(1/a_1, \ldots, 1/a_r) + U'U.$$ 

The parameters $\{a_k\}$ are sampled from its full conditional distributions, specifically, Inverse-Gamma distributions with scale parameter $3/2$ and rate parameter $\lambda V^{-1}(k,k) + 1/B_k^2$, where $V^{-1}(k,k)$ is the $k^{th}$ diagonal element of $V^{-1}$ matrix.

The latent process, $\{W_i, i = 1, \ldots, r\}$, are linearly related, hence we block update them. The full conditional distribution is unavailable analytically and takes the form:

$$[W|\text{rest}] \propto \exp\left\{-0.5R'R\right\} \times \exp\left\{-\sum_{i=1}^r W_i\right\},$$

where $R = Z - X\beta - S A^{\frac{1}{2}} \eta$. This full conditional distribution is log-concave and we use adaptive rejection metropolis sampling (ARMS) to obtain samples for $W$. The ARS algorithm allows samples to be obtained even from complicated densities that may not be log-concave. This is accomplished through a piecewise linear function which is an envelope for the log of the target density, $\ln f(x)$. Samples are drawn and either accepted, or used to update and tighten the envelope. For details see (Gilks, Best, & Tan, 1995).

$\beta$ and $U$ are linearly related in our model and we sample them together from a multivariate Gaussian distribution of dimension $p + r$, with mean
vector $\mathbf{\mu}_{\mathbf{\beta,U}}$ and covariance matrix $\Sigma_{\mathbf{\beta,U}}$, where:

$$
\Sigma_{\mathbf{\beta,U}} = \begin{bmatrix}
\frac{1}{\nu^2} \mathbf{X}' \mathbf{X} + 10^{-6} \mathbf{I}_p & \frac{1}{\nu^2} \mathbf{X}' \mathbf{S} \mathbf{A}^{1/2} \\
\frac{1}{\nu^2} \mathbf{A}^{1/2} \mathbf{S}' \mathbf{X} & \frac{1}{\nu^2} \mathbf{S}' \mathbf{A} \mathbf{S} + \mathbf{V}^{-1}
\end{bmatrix}^{-1}
$$

(5.8)

$$
\mathbf{\mu}_{\mathbf{\beta,U}} = \Sigma_{\mathbf{\beta,U}} \begin{bmatrix}
\mathbf{X}' \mathbf{Y} \\
\mathbf{A}^{1/2} \mathbf{S}' \mathbf{Y}
\end{bmatrix}.
$$

(5.9)

Recall that $\nu^2 = \sigma^2 + \tau^2$ is the nugget variance.

The process error variance $\tau^2$ is sampled from Inverse-Gamma distribution with shape parameter $n/2 + 1$ and rate parameter $\frac{1}{\nu^2} \mathbf{\delta}' \mathbf{\delta} + \frac{2}{a}$. The parameter $a$ is sampled from another Inverse-Gamma distribution with shape parameter $3/2$ and rate parameter $\frac{2}{\tau^2} + 10^{-10}$.

The process errors, $\mathbf{\delta}$, for $n$ observed locations are sampled from a multivariate Gaussian distribution with covariance matrix and mean vectors as:

$$
\Sigma_{\mathbf{\delta}} = (\sigma^{-2} + \tau^{-2})^{-1} \mathbf{I}_n
$$

(5.11)

$$
\mathbf{\mu}_{\mathbf{\delta}} = \Sigma_{\mathbf{\delta}} (\mathbf{Z} - \mathbf{X} \mathbf{\beta} - \mathbf{S} \mathbf{A}^{1/2} \mathbf{U}).
$$

(5.12)

For each unobserved location, the process error is sampled from a zero mean Gaussian distribution with variance $\tau^2$.

Once, we have sampled all the parameters and latent processes mentioned above, $\mathbf{Y}$ becomes deterministic and are obtained as:
CHAPTER 5. RRSM FOR NON-GAUSSIAN TAILS

\[ Y_i = X_i \beta + S_i A^{1/2} U + \delta_i, \]  \hspace{1cm} (5.13)

where \( X_i \) and \( S_i \) denote the rows of \( X \) and \( S \) matrices correspond to \( Y_i \).

The convergence of our Markov Chains are assessed using standard techniques, trace-plots, Gelman and Rubin statistics, (Gelman & Rubin, 1992).

Storing simulations from MCMC run on model parameters does not require large memory. Posterior means and standard deviations of the latent processes can be obtained sequentially within MCMC loop. But problem arises when one wants to compute prediction intervals. Quantile based prediction intervals require storing realizations of all \( Y(s_i) \)'s after burn-in. Instead, if one can apply Markov Chain Central Limit Theorem (MCCLT), then Monte Carlo standard error can be computed without burdening computer memory and hence credible intervals of posterior means under normality assumptions. Application of MCCLT on ergodic averages is not immediate, one needs to verify some nontrivial conditions, one of them is geometric ergodicity, i.e., the chains converge fast enough to the target posterior at a exponential rate. For more discussion and insights on MCCLT see (Jones et al., 2006). Most popular posterior summary that is often used is the posterior expectation:

\[ E(Y(s_i)|Z) = X E(\beta|Z) + S_i E(A^{1/2}U|Z) + E(\delta(s_i)|Z). \]  \hspace{1cm} (5.14)

Now, note that the posterior mean is a function \( g_i(\beta, W, U, \sigma^2, \tau^2) \) of the
parameters \( \{ \beta, W, U, \sigma^2, \tau^2 \} \). Since the full conditional distribution of \( \delta \) is a function of \( (\beta, W, U, \sigma^2, \tau^2) \) only and the measurement error variance \( \sigma^2 \) is assumed to be known, it is sufficient to establish the geometric ergodicity of the chain \( (\beta, W, U, \tau^2) \). Ergodic average based on MCMC samples can be written as

\[
\overline{g}_i = \frac{1}{T} \sum_{t=1}^{T} g_i(\beta(t), W(t), U(t), \tau^2(t)).
\]

where \( T \) denotes the number of MCMC samples that are used to compute these averages and the index \( (t) \) denotes the generated sample at \( t^{th} \) MCMC iteration. The version of the CLT that we will use is:

**Theorem 5.3.1** \( \{ \beta(t), W(t), U(t), \tau^2(t), t = 0, 1, 2, \ldots \} \) is a Harris ergodic Markov Chain with invariant probability distribution \( f(\cdot) \) on state space \( \mathcal{X} \) and \( E_f |g_i| < \infty, E_f |g_i|^{2+\epsilon} < \infty \) for \( \epsilon > 0 \). If this Markov chain is geometrically ergodic, then:

\[
\sqrt{T}(\overline{g}_i - E(\overline{g}_i)) \rightarrow \mathcal{N}(0, \sigma^2_{g_i}),
\]

where \( \sigma_{g_i} \) is asymptotic standard deviation.

\( \sigma_{g_i}/\sqrt{T} \) can be considered as a measure of precision of the posterior means obtained from \( T \) MCMC samples. \( \sigma^2_{g_i} \) is unknown and it is estimated using batch mean method, (see (Jones et al., 2006)) by splitting the generated MCMC samples into \( K \) batches, each of size \( J \) as:

\[
\hat{\sigma}^2_{g_i} = \frac{K}{J-1} \sum_{k=1}^{K} (g_i^{(k)} - \overline{g}_i)^2.
\]  

(5.15)
$g_i^{(k)}$ is the $k^{th}$ batch mean obtained by computing

$$g_i^{(k)} = \frac{\sum_{j=J(k-1)+1}^{kJ} g_i((\beta^{(j)}, W^{(j)}, U^{(j)}, \tau^{2(j)})_J,}{J},$$

and $T = KJ$.

To verify the conditions for MCCLT, first we prove that the Markov chain $\beta, W, U, \tau^2$ generated from from the derived MCMC algorithm is *Harris ergodic*. Next geometric ergodicity of this chain is established using *drift* and *minorization* conditions, see (Jones & Hobart, 2001). We prove these results in Section (5.4).

## 5.4 Verifying the Conditions of MCCLT

In this section we verify the conditions of MCCLT on the Markov chain $\{\beta^{(t)}, W^{(t)}, U^{(t)}, \tau^{2(t)}\}_{t=0}^\infty$ stated in Theorem (5.3.1). From the full conditional distributions that we derived in previous section, we can clearly see that all of them are *continuous*. Further, Markov chains generated from these full conditionals are *aperiodic, Harris recurrent*, and $\phi$–irreducible. Now, we will establish the *geometric ergodicity* of this Markov chain by *drift* and *minorization* conditions.

Drift condition will be established by finding a function $V(\cdot) : \mathbb{R}^p \times \mathbb{R}^+ \times \mathbb{R}^r \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ of $\{\beta, W, U, \tau^2\} \in \mathbb{R}^p \times \mathbb{R}^+ \times \mathbb{R}^r \times \mathbb{R}^+$, such that:
\[(PV)(\beta, W, U, \tau^2) \leq \rho V(\beta, W, U, \tau^2) + L,\]

for all \(\{\beta, W, U, \tau^2\} \in \mathbb{R}^p \times \mathbb{R}^{+r} \times \mathbb{R}^{r} \times \mathbb{R}^{+}\) \hspace{1cm} (5.16)

where \(\rho \in [0, 1)\) and \(L \in \mathbb{R}\); and

\[(PV)(\beta, W, U, \tau^2) = \int \int \int \int V(\beta, W, U, \tau^2)
\Gamma(\beta, W, U, \tau^2|\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau^2})d\{\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau^2}\},\]

\hspace{1cm} (5.17)

where \(\Gamma(\beta, W, U, \tau^2|\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau^2})\) is one-step transition kernel from \((\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau^2})\) to \((\beta, W, U, \tau^2)\).

We consider \(V(\beta, W, U, \tau^2)\) to be \((Y - X\beta - SA^{1/2}U)'(Y - X\beta - SA^{1/2}U)/\tau^2\). Recall from our MCMC algorithm described in the previous section that the latent variable \(Y\) is just a function of \(\{\beta, W, U, \tau^2, \sigma^2\}\), where the measurement error variance \(\sigma^2\) is known. We apply Fubini’s Theorem for interchanging integrals and Equation (5.17) becomes:
\[
\int_{\Theta^{-}} \int_{\mathbb{R}^p \times \mathbb{R}^r} \int_{\mathbb{R}^r \times \mathbb{R}^r} V(\cdot) \Gamma(\beta, W, U, \tau^2 | \Sigma^-, \tilde{\beta}, \tilde{W}, \tilde{U}, \tau^2) \\
\Gamma(\Omega^{-} | \tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau}^2) d\{\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau}^2\} d\Omega^{-}
\]
\[
= \int_{\Theta^{-}} \int_{\mathbb{R}^p \times \mathbb{R}^r} \int_{\mathbb{R}^r \times \mathbb{R}^r} V(\cdot) \Gamma(\tau^2 | \beta, W, U, \Sigma^-, \tilde{\beta}, \tilde{W}, \tilde{U}, \tau^2) \\
\Gamma(\beta, W, U | \Sigma^-, \tilde{\beta}, \tilde{W}, \tilde{U}, \tau^2) \Gamma(\Omega^{-} | \tilde{\beta}) \\
d\{\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau}^2\} d\Omega^{-} \tag{5.18}
\]
\[
\leq n + 2 \\
\leq \rho V(\cdot) + n + 2 \tag{5.19}
\]

We denote the remaining parameters in the model except \(\{\beta, W, U, \tau^2\}\) by \(\Sigma^-\) and \(\Theta^{-}\) is the parameter space for \(\Sigma^-\). The full conditional distribution \(\Gamma(\tau^2 | \beta, W, U, \Sigma^-, \tilde{\beta}, \tilde{W}, \tilde{U}, \tau^2)\) is Inverse-Gamma distribution with shape parameter \(n/2 + 1\) and rate parameter \(\frac{(Y - X\beta - SA^{1/2}U)'(Y - X\beta - SA^{1/2}U)}{2 + 2/a}\).

The innermost integral in (5.18) with respect to \(\tau^2\) becomes:

\[
\frac{(n/2 + 1)(Y - X\beta - SA^{1/2}U)'(Y - X\beta - SA^{1/2}U)}{(Y - X\beta - SA^{1/2}U)'(Y - X\beta - SA^{1/2}U)/2 + 2/a} \tag{5.20}
\]
\[
\leq n + 2 \tag{5.21}
\]
The inequality in (5.20) follows from the fact that $2/a > 0$ and the last inequality in (5.18) holds as $V(\cdot) \geq 0$.

We establish the minorization condition by showing that on a set $C = \{ (\beta, W, U, \tau^2) : V(\cdot) < l \}$ for $l > 2L/(1 - \rho)$, there exists a density $q$ and $\varepsilon > 0$ such that for $(\tilde{\beta}, \tilde{W}, \tilde{U}, \tilde{\tau}^2) \in C$ and $(\beta, W, U, \tau^2) \in \mathbb{R}^p \times \mathbb{R}^r \times \mathbb{R}^r \times \mathbb{R}^r$,

$$\Gamma(\beta, W, U, \tau^2|\beta, W, U, \tau^2) \geq \varepsilon q(\beta, W, U, \tau^2). \quad (5.22)$$

The right-hand side of Equation (5.22) is:

$$= \int_{\Theta^-} \Gamma(\beta, W, U, \tau^2|\beta, W, U, \tau^2, \Sigma^-) \Gamma(\Omega^-|\beta, W, U, \tau^2) d\Omega^- \quad (5.23)$$
$$= \int_{\Theta^-} \Gamma(\beta, W, U|\tau^2, \beta, W, U, \tau^2, \Sigma^-) \Gamma(\tau^2|\beta, W, U, \tau^2, \Sigma^-) \Gamma(\Omega^-|\beta, W, U, \tau^2) d\Omega^- \quad (5.24)$$
$$= \int_{\Theta^-} \Gamma(\beta, W, U|\tau^2, \beta, W, U, \tau^2, \Sigma^-) \Gamma(\Omega^-|\beta, W, U, \tau^2, \Sigma^-) \Gamma(\Omega^-|\beta, W, U, \tau^2) d\Omega^- \quad (5.25)$$
$$\geq \varepsilon \int_{\Theta^-} \Gamma(\beta, W, U|\tau^2, \beta, W, U, \tau^2, \Sigma^-) \Gamma(\Omega^-|\beta, W, U, \tau^2) d\Omega^- \quad (5.26)$$

IG$(a, b; x)$ represents Inverse-Gamma distribution with shape $a$ and rate $b$, evaluated at $x$. In $C$, $(Y - X\bar{\beta} - S\bar{A}^{1/2}U)'(Y - X\bar{\beta} - S\bar{A}^{1/2}U) \leq l\bar{\tau}^2$ and the exponent term in Inverse-Gamma distribution is a decreasing function of its rate parameter, hence the inequality in (5.23) holds. $\varepsilon$ is given by:
and the remaining terms constitute \( q(\cdot) \), which is a \textit{bona fide} probability distribution. This completes the proof. For general discussion and insight on MCMC theory and the techniques and theorems that are used for our proof, readers are encouraged to see (Jones & Hobart, 2001), (Meyn & Tweedie, 1993), (Rosenthal, 1995), and (Jones, 2004) and the references there in.

5.5 Application: Simulated and Real Data

In this section we apply our data to real and simulated data. We use the simulation study to assess the validity of our model through cross-validation. Afterward, we model a large dataset of atmospheric properties and obtain predictions over the entire earth.

5.5.1 Simulation Study

To assess the performance of our class of models, we first apply it on simulated datasets. We simulate data based on the model specified in Equations (1.1) and (1.2) but setting, as mentioned, \( \eta = \mathbf{A}^{1/2} \mathbf{U} \). For observed locations we select \( n = 25,600 \) locations on a regular grid over the unit square. Two resolutions of knot locations were selected. The first resolution had \( r_1 = 16 \) knots
laid out in a $4 \times 4$ grid, the second resolution had $r_2 = 64$ knots in an $8 \times 8$ grid. Within each resolution, the knots were equidistant over the unit square.

For the large-scale variation, we set $\mathbf{X}$ to be an $n \times 2$ matrix where the first column is the $X$–coordinate and the second column is the $Y$–coordinate, with corresponding $\mathbf{\beta} = (5, 2)'$. The $\mathbf{S}$ matrix was specified as using modified bisquare function as described in section 3.5 (with the $X$-coordinate as longitude and the $Y$-coordinate as latitude). Since the knots are equidistant on a regular grid, we had $r_{lat(1)} = r_{long(1)} = 0.25$ and $r_{lat(2)} = r_{long(2)} = 0.125$. The fixed rank covariance matrix $\mathbf{V}$ we generate using exponential covariance, where the $(i,j)^{th}$ element is, $\mathbf{V}(i,j) = \exp\{-\theta||\mathbf{u}_i - \mathbf{u}_j||\}$, where $||\mathbf{u}_i - \mathbf{u}_j||$ is the Euclidean distance between $\mathbf{u}_i$ and $\mathbf{u}_j$ and $\mathbf{u}_i$ denotes the knot location for $i^{th}$ knot. The parameter $\theta$ impacts the range, large values cause the covariance to decay quickly (implying a more coarse process) and small values lead to a more smooth process. We set $\theta = 0.05$.

We draw a sample of $\mathbf{U}$ from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{V}$, specified in the previous paragraph. To test the flexibility of the model, we generate three samples for $\mathbf{W}$ as follows,

$$\begin{align*}
\mathbf{W}_1 & \sim \text{Expon}(1), \\
\mathbf{W}_2 & \sim \text{Gamma}(5, 1), \\
\mathbf{W}_3 & \sim \text{GIG}(1, 2, 2), 
\end{align*}$$

where $\text{GIG}(\chi, \psi, \lambda)$ is the Generalized Inverse Gaussian distribution with
density,
\[ f(x) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\psi\chi})} x^{\lambda-1} \exp \left\{ -\frac{1}{2} \left( \frac{\chi}{x} + \psi x \right) \right\}. \]

The Gamma and GIG distributions were chosen in part because each contains the Exponential distribution as a special case. As the model uses the Exponential distribution simulating \( W \), good performance on the simulations with the Gamma and GIG distributions on \( W \) are indications of the flexibility of our model. Then \( A = \text{diag} (W) \).

Finally, we generate the measurement error \( \epsilon \) from a normal distribution with variance 0.0001 and calculate the observed data \( Y = X\beta + SA^2U + \epsilon \). Quantile plots for all three simulations showed significant non-Gaussian tail behavior.

We randomly select 5,000 observed locations and keep them aside as test case. Remaining 20,600 observations are used for estimation and prediction (i.e. the training sample). Results are shown in figure (5.3). The left column are histograms of \( Y - \hat{Y} \) for the test data under three different simulations of \( W \). Note that these are centered about 0, and have very little spread. In the right column are scatterplots of \( Y \) vs \( \hat{Y} \). Superimposed (although obscured by the points themselves) is a red line plotting \( y = x \); points falling on this line are perfect predictions. All of the points lie very close to the \( y = x \) line, meaning that our proposed model provides excellent predictions. And note that these are predictions for the unobserved test case locations. These results show that the proposed method is able to successfully model datasets with non-Gaussian tails. Also, note that we use the same model described
in Section (5.2) to all three simulated datasets. Hence, the model is flexible enough to handle varying forms of the scale factor $W$ even with just an Expon(1) prior distribution on $W$.

Figure 5.3: Results of the simulations. Panels (a) and (b) correspond to $W_1$, panels (c) and (d) correspond to $W_2$, and panels (e) and (f) correspond to $W_3$. The left column are histograms of $Y - \hat{Y}$ for test sample. The right column are scatterplots of $Y$ vs $\hat{Y}$.

Results for running a purely Gaussian model (that is, not modeling $\eta$ as
were similar. Both models made highly accurate predictions. However, it should be noted that while the simulated datasets deviated from the Gaussian distribution, all were unimodal and none showed extreme outliers. In a sense, then, this simulation illustrates non-inferiority of the proposed model. We intend to proceed with additional simulations which deviate from the Gaussian distribution to a greater degree.

5.5.2 Application to Ozone Data

From the Moderate Resolution Imaging Spectroradiometer (MODIS) on board NASA’s Terra satellite, we obtain a large dataset of atmospheric properties. We model the daily total ozone maximum. These data are heavily right-skewed, and even after taking log-transformation still exhibit strong deviation from the Gaussian distribution. For our purposes, we consider the log-transformed TOM. Recall that the probability plot of log(TOM) showed many outliers on the upper tail (e.g. figure (5.1)).

Many atmospheric processes display a north-south trend, but comparatively little east-west trend. The plot of log TOM in figure (5.4) shows such a trend, so we model the large-scale variation using spherical harmonics. The design matrix $X$ is specified by Legendre polynomials (see (Stein, 2007)) $P_m^n(\sin(L))$ of degree $n = 80$ and order $m = 0, 1, \ldots, n$, where $L$ is the latitude of a location.

We employ a two-resolution scheme for $S$. This allows the model to capture multiple scales of spatial variation Nychka, Wikle, & Royle (2002);
Cressie & Johannesson (2008). The first resolution has \( r_1 = 68 \) knot locations, the second resolution has \( r_2 = 200 \) knot locations. Both resolutions are approximately equidistant over the entire globe. The \( \mathbf{S} \) matrix is defined as in section 3.5. In particular, distances between locations and knots, \( d(\mathbf{s}_i, \mathbf{u}_{j(l)}) \), is defined using great-arc distances based on Vincenty’s formula, which is a more accurate means to compute distances on an ellipsoid than Euclidean distances. The distance 'scaling factors' we use are, for the first resolution, \( r_{\text{lon}}(1) = 2725.1, r_{\text{lat}}(1) = 2700.4 \), and for the second resolution, \( r_{\text{lon}}(2) = 1546.4, r_{\text{lat}}(2) = 1667.9 \).

As described in Section 5.3, a data analytic approach is used to estimate
the measurement error variance. In our analysis, we randomly select 20 observed locations from the dataset and calculate the variance of all points within a $4 \times 4$ bin centered about each of the selected locations. The estimate of $\sigma^2$ is the median of these variances, $\hat{\sigma}^2 = 0.1$ The predictions of log TOM using Bayesian posterior predictive distribution and the batch-means standard errors are shown in figure (5.5)(a),(b). As with the simulated data, we also ran a Gaussian model. The analogous plots to figure (5.5) are shown in figure (5.6)(a),(b)

![Maps](image)

Figure 5.5: Maps of (a) predictions using Bayesian posterior predictive distribution and (b) batch-means standard errors using the proposed RRSM.
Unsurprisingly, the predictions show a large degree of similarity. Even with non-Gaussian data, reduced rank models based on the assumption of a Gaussian distribution are able to produce accurate predictions. Slightly more of interest is how well the Gaussian model was able to match the proposed model in terms of the Monte Carlo Standard Errors. As shown in chapter 3, even in the face of a slight deviation from Gaussian distribution, the robust methods performed better than Gaussian-based techniques.

Here, we attribute this to the prior specification of the covariance matrix.
The non-informative inverse-Wishart prior of Huang & Wand (2013) provides a highly flexible framework, allowing the model to adapt to the data in spite of the distribution. However, in spite of the added flexibility this brings to purely Gaussian models, we still observe the benefit of using the proposed model. Figure (5.7) shows a map of the difference in Monte Carlo Standard Errors (MCSE) for each location. Differences were taken as \((\text{Gaussian Model} - \text{Proposed Model})\), so small values indicate the Gaussian model had smaller standard errors, and large values indicate the proposed model had smaller standard errors.

Figure 5.7: Map of \(\nu_G^2 - \nu_W^2\), the location-wise difference of Monte Carlo Standard Errors.

We observe what appears to be a spatial dependence among these differ-
ences. Notably, the region over the Pacific Ocean seems to have a concentration of small values, while much of the rest of the map has larger values. Looking back to figure (5.4), the Pacific Ocean was where there were greater quantities of missing data. In geostatistics, areas of missing data are expected to have larger standard errors, because no or little data to provide information in the area. Hence, our model can be seen as respecting this facet of spatial models. It has reduced the MCSEs where there is data, and allowed the MCSEs to increase where there is no data, which is expected. On the other hand, the Gaussian model appears to be smoothing out the MCSEs across the map, contrary to what one expects from a geostatistical model.

5.6 Conclusions and Discussion

In this chapter we have proposed a new class of reduced rank spatial models. Instead of modeling the latent spatial component as a multivariate Gaussian distribution, we have replaced it with a scale mixture of Gaussian distributions, where the scale parameters are iid Exponential. This specification enables the model to account for a wide variety of tail behaviors, including light or heavy tails.

In addition, we have added a great degree of flexibility to reduced rank spatial models as a whole through the inverse-Wishart prior of Huang & Wand (2013). As seen in both the simulation and the data analysis, the
Gaussian model was able to be successful in spite of deviations from the Gaussian distribution. Simulations for the proposed model took approximately 2.68 hours each, while for the Gaussian model it was 1.64 hours each. Data analysis took 19.59 hours for the proposed model, and 3.14 hours for the Gaussian model. While this is a significant increase in computational resources, the added flexibility of the model by using the scale-mixture of Gaussians cannot be discounted. Moving forward, we intend to test our model on other forms of non-Gaussian distributions, such as bi-modal distributions. It is unlikely that a Gaussian model will be successful in such a situation, while we anticipate our model will handle it without difficulty.
Chapter 6

Spatio-Temporal Models

6.1 Introduction

Predicting a spatial process at unobserved locations is an important component of geostatistics. However large datasets lead to heavy computational overhead. Once the sample size exceeds several thousand dimension-reduction techniques are required for efficient computation. Banerjee et al (2008) addressed this by approximating the spatial process with a reduced dimension process which they define over a fixed number of knot locations.

Going further, a method for exact kriging estimates is provided by Cressie & Johannesson (2008), who term their model Fixed Rank Kriging (FRK). They implement a 2-stage method of moments (MOM) means of estimating the spatial covariance matrix to perform kriging. This approach has been further studied and elaborated. Cressie & Kang (2010) used a gen-
eral, nonstationary covariance matrix for the reduced rank process, leaded to added flexibility in the model. Katzfuss & Cressie (2011a) implement an expectation-maximization (EM) algorithm to estimate the parameters of a FRK model. EM is a powerful tool for estimating model parameters based on likelihood assumption, but suffers (or fails) when data do not follow the specified distribution.

The key idea in all of these is to model the spatial covariance as a function of known basis functions defined over a fixed (and relatively small) number of knot locations, and the covariance of a reduced rank spatial process. Various classes of basis functions may be used, Shi & Cressie (2007) used W-wavelet basis functions, while Cressie & Johannesson (2008) used bisquare basis functions. Under this approach, the model is separated into the data model and the process model, which may be viewed as a hidden state model.

An additional challenge arises when considering spatial and temporal dependencies between observations. One of the main challenges in this area is that of specifying the form temporal evolution. Often this is assumed to be linear and first-order Markov, so the temporal dependence is modeled through a single propagator matrix. Some have assumed this to be the identity (Stroud, Müller & Sans, 2001) or a diagonal matrix (Lopes, Salazar & Gamerman, 2008). But such an assumption limits the flexibility of the model. Kalman filters (Kalman, 1960) have been developed for spatio-temporal models, see Cressie & Wikle (2002) for a summary of the development of such models. Yet these were not developed for large-data
problems. Cressie et al. (2010) extended the FRK model, they used a Kalman filter for modeling temporal dependence. Katzfuss & Cressie (2011 b) developed an empirical Bayesian model for space-time smoothing of large datasets.

While (Katzfuss & Cressie, 2012) provided a fully Bayesian space-time smoothing model, the assumption of Gaussianity is still present. In particular, their reduced-rank or hidden state process was assumed to be Gaussian. This assumption is not always valid, as many environmental datasets show non-Gaussian tail behaviors. We model the state variable as a scale mixture of Gaussian distributions, where the scale variables follow independent Exponential distributions with mean parameter 1. This adds extra flexibility to capture more accurately a variety of non-Gaussian tail behaviors.

6.2 A Flexible Non-Gaussian Spatio-temporal Model

The spatio-temporal random effects (STRE) model described in Cressie et al. (2010) and Katzfuss & Cressie (2012) is a very flexible class of geostatistical models. For a time $t$, let $Z_t \equiv \{Z(s_{1,t}), \ldots, Z(s_{n_t,t}), s_1, \ldots, s_{n_t} \in \mathcal{D}\}$ be a vector of $n_t$ observed locations $s_t$ on domain $\mathcal{D} \in \mathbb{R}^d$, and at time points $t = 1, 2, \ldots, T$. In most applications, $d = 2$. Then we model $Z_t$ as:

$$Z(s_t) = Y(s_t) + \epsilon(s_t); \quad t = 1, \ldots, T \quad (6.1)$$
CHAPTER 6. NON-GAUSSIAN SPATIO-TEMPORAL MODELS

where \( \epsilon(s_t) \) is a zero-mean Gaussian white noise (measurement errors) with variance \( \sigma^2 \nu_{\epsilon,t}(\cdot) \) for some known function \( \nu_{\epsilon,t}(\cdot) > 0 \). Then \( Y(s_t) \) is the measurement-error-free latent process (truth), which is independent of measurement errors and follows a Gaussian distribution with mean \( \mu(s_t) \) and covariance matrix \( \Sigma_Y \). Predictions of \( Y \) in space and time with generally require the inversion of \( \Sigma_Y \). If \( n_t \) is too large (exceeding several thousands), this is a computationally infeasible (or impossible) task. Because of this, we impose a *spatio-temporal random effects* model (STRE) on \( Y(s_t) \), decomposing it as:

\[
Y(s_t) = X_t \beta_t + S_t \eta_t + \delta; \quad t = 1, \ldots, T \tag{6.2}
\]

In this expression, we represent the mean term previously denoted \( \mu(s_t) \) with a linear function of covariates and their slopes, \( X_t \beta_t \). Then we choose \( r_t \) knot locations over which we define the basis functions. The reduced rank process is \( \eta_t \), and \( S_t \) is the \( n_t \times r_t \) matrix of basis functions which maps the \( \eta_t \) to \( Y_t \). The main goal of STRE models is to characterize the spatio-temporal evolution of \( \eta_t \). Typically, \( \eta_t \) is assumed to be vector auto-regressive (VAR);

\[
\eta_t|\eta_{t-1} \sim \text{MVN}_{r_t}(H_t \eta_{t-1}, K_t) ; \quad t = 1, \ldots, T, \tag{6.3}
\]

where \( H_t \) is known as the propagator matrix, and \( K_t \) is known as the innovation matrix. Finally, \( \delta_t \) is known as the *process error*, which accounts for uncertainties due to the dimension reduction. Similarly to \( \epsilon \), we assume \( \delta \) to be a zero-mean Gaussian process, with variance \( \tau^2 \nu_{\delta,t}(\cdot) \).
Now we diverge from the fully Bayesian STRE model described by Katzfuss & Cressie (2012) by replacing $\eta_t$ with $A_t^{\frac{1}{2}} \eta_t$. The matrix $A_t$ is an $r_t \times r_t$ diagonal matrix with diagonal elements $W_{i,t} \sim \text{Expon}(1)$, $i = i, \ldots, r_t$, $t = 1, \ldots, T$. In this way, instead of a Gaussian distribution, we assume a more general scale mixture of Gaussians on $\eta_t$. Hence we have,

$$\eta_t | \eta_{t-1}, W_t \sim \text{MVN}_{r_t}\left(H_t A_t^{-\frac{1}{2}} \eta_{t-1}, K_t\right): t = 1, \ldots, T. \quad (6.4)$$

Next, we will put prior distributions on all of the unknown terms so that we may implement a Monte Carlo Markov Chain (MCMC) algorithm to estimate parameters and obtain predictions in space and time. Further, given the complexity of the models and the large number of parameters to be samples in each iteration of MCMC, this can be computationally prohibitive. Hence, one can proceed with sequential procedures, such as, Sequential Monte Carlo (SMC). For a basic foundation of SMC, see Doucet, Freitas & Gordon (2001). SMC has been successfully applied on several spatial models and Bayesian learning of model parameters have been implemented with sequential updating of state variables, some notable recent works are Stroud, Miller & Sans (2001) and Polson, Stroud & Miller (2008), Stroud et al. (2010), Wikle & Hooten (2010), and the references there in.

The modeling approach discussed in the previous paragraphs assumes that the underlying process is defined on discrete time points. But there are several spatial processes (such as ocean variables) that are observed through
continuous monitoring systems, for example, satellites, ship-tracks, or buoys. Sahu & Challenor (2008) developed a joint Bayesian space-time model for temperature and salinity based on Atlantic Ocean data collected by Argo floats. They used kernel based approach for modeling space-time correlations. One can start the model building from the general formulation of the space-time model described in Equation (6.8) Cressie & Wikle (2011), Chapter 6, where a space-time process is factored into a large scale trend due to space and time, a purely spatial random effect, a purely temporal random effect, a random effect due to space-time interaction that is unexplained by the large scale variation, and a nugget effect. Carefully chosen kernel functions and statistical distributions on hidden processes will lead to nonseparable and nonstationary covariance functions.
Appendix

Consistency of Dependent Quantiles
In this appendix we reproduce from Wendler (2011) the two theorems referenced in section 3.3, along with related definitions. In these, $X_1, \ldots, X_n$ is a stationary sequence of dependent random variables which have distribution function $F$. The distribution function of a $U$—quantile is denoted by $U$.

**Definition 1.** (variation condition for quantile functions)

1. A function $g: \mathbb{R} \to \mathbb{R}$ satisfies the variation condition, if there is a constant $L$ such that,

$$E \left[ \sup_{|x - x_0| \leq \epsilon} |g(x) - g(x_0)| \right] \leq L \epsilon.$$

2. A function $g: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfies the uniform variation condition on $B \subset \mathbb{R}$, if there is a constant $L$ such that the above holds for all functions $g(\cdot, t), t \in B$.

**Theorem 1.** Let $g: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a nonnegative, bounded, measurable function which is nondecreasing in the second argument, let $F(t) := E[g(X_1, t)]$ be differentiable in $t_p \in \mathbb{R}$ with $F'(t_p) = f(t_p) > 0$ and

$$|F(t) - F(t_p) - f(t_p)(t - t_p)| = o \left( |t - t_p|^{3/2} \right) \text{ as } t \to t_p. \quad (6.5)$$

Assume that one of the following two conditions holds:
1. \((X_n)_{n \in \mathbb{Z}}\) is strongly mixing with \(\alpha(n) = O(n^{-\beta})\) for some \(\beta \geq 3\). Let \(\gamma := \frac{\beta-2}{\beta}\).

2. \((X_n)_{n \in \mathbb{Z}}\) is a 1-approximating functional of an absolutely regular process \((Z_n)_{n \in \mathbb{Z}}\) with mixing coefficients \((\beta(n))_{n \in \mathbb{N}}\) and approximation constants \((a_n)_{n \in \mathbb{N}}\) such that \(\beta(n) = (n^{-\beta})\) and \(a_n = (n^{-(\beta+3)})\) for some \(\beta > 3\). Let \(g\) satisfy the variation condition uniformly in some neighborhood of \(t_p\) and let \(\gamma := \frac{\beta-3}{\beta+1}\).

**Definition 2.** (variation condition for \(U\)-quantile kernels)

1. The kernel \(h\) satisfies the variation condition for \(t \in \mathbb{R}\), if there is a constant \(L\) such that,

\[
E \left[ \sup_{|| (x,y) - (X,Y) || \leq \epsilon} |h(x,y,t) - h(X,Y,t)| \right] \leq L \epsilon,
\]

where \(X,Y\) are independent with the same distribution as \(X_1\) and \(||(x_1, x_2)|| = (x_1^2 + x_2^2)^{\frac{1}{2}}\) denotes the Euclidean norm.

2. The kernel \(h\) satisfies the uniform variation condition on \(B \subset \mathbb{R}\), if there is a constant \(L\) such that the above holds for all functions \(g(\cdot, t), t \in B\).

**Theorem 2.** Let \(h : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}\) be a bounded kernel function that satisfied the uniform variation condition in some neighborhood of \(t_p\). Let
CHAPTER 6. NON-GAUSSIAN SPATIO-TEMPORAL MODELS

\( U(t) := E[h(X,Y,t) \) be differentiable in \( t_p \in \mathbb{R} \) with \( U'(t_p) = u(t_p) > 0 \) and

\[
|U(t) - U(t_p) - u(t_p)(t - t_p)| = o \left( |t - t_p|^{\frac{3}{2}} \right) \text{ as } t \to t_p. \quad (6.6)
\]

Assume that one of the following two conditions holds:

1. \( ||X||_1 < \infty \) is strongly mixing and the mixing coefficients satisfy \( \alpha(n) = O \left( n^{-\beta} \right) \) for some \( \beta \geq \frac{13}{4} \). Let \( \gamma := \frac{\beta-2}{\beta} \).

2. \( (X_n)_{n \in \mathbb{Z}} \) is a \( 1- \) approximating functional of an absolutely regular process \( (Z_n)_{n \in \mathbb{Z}} \) with mixing coefficients \( (\beta(n))_{n \in \mathbb{N}} \) and approximation constants \( (a_n)_{n \in \mathbb{N}} \) such that \( \beta(n) = (n^{-\beta}) \) and \( a_n = (n^{-(\beta+3)}) \) for some \( \beta > 3 \). Let \( \gamma := \frac{\beta-3}{\beta+1} \).
Bibliography


BIBLIOGRAPHY


positive variables revisited: Sample space and scale considerations. *Mathematical Geology*, 39, 529-558.


