Empirical Hierarchical Modeling and Predictive Inference for Big, Spatial, Discrete, and Continuous Data

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of the Ohio State University

By
Aritra Sengupta, M.Sc (India)
Graduate Program in Statistics

The Ohio State University
2012

Dissertation Committee:
Professor Noel Cressie, Advisor
Professor Radu Herbei
Professor Desheng Liu
ABSTRACT

This dissertation is comprised of an introductory chapter and three stand-alone chapters. The three main chapters are tied together by a common theme: empirical hierarchical spatial-statistical modeling of non-Gaussian datasets. Such non-Gaussian datasets arise in a variety of disciplines, for example, in health studies, econometrics, ecological studies, and remote sensing of the Earth by satellites, and they are often very-large-to-massive. When analyzing “big data,” traditional spatial statistical methods are computationally intensive and sometimes not feasible, even in supercomputing environments. In addition, these datasets are often observed over extensive spatial domains, which make the assumption of spatial stationarity unrealistic.

In this dissertation, we address these issues by using dimension-reduction techniques based on the Spatial Random Effects (SRE) model. We consider a hierarchical spatial statistical model consisting of a conditional exponential-family model for the observed data (which we call the data model), and an underlying (hidden) geostatistical process for some transformation of the (conditional) mean of the data model. Within the hierarchical model, dimension reduction is achieved by modeling the geostatistical process as a linear combination of a fixed number of basis functions, which results in substantial computational speed-ups. These models do not rely on specifying a spatial weights matrix, and no assumptions of homogeneity, stationarity, or isotropy are made.

Another focus of the research presented in this dissertation is to properly account for
spatial heterogeneity that often exists in these datasets. For example, with county-level health data, the population at risk is different for different counties and is typically a source of heterogeneity. This type of heterogeneity, whenever it exists, needs to be incorporated into the hierarchical model. We address this through the use of an offset term and by properly weighting the SRE model (e.g., Chapter 2), and through data-model specifications (e.g., Chapter 3).

Following the introductory chapter, in Chapter 2 we consider spatial data in the form of counts. We consider a Poisson data model for the counts, and develop maximum likelihood (ML) estimates for the unknown parameters using an expectation-maximization (EM) algorithm. We illustrate the hierarchical nature of our approach to the spatial modeling of counts, through the analysis of a spatial dataset of Sudden Infant Death Syndrome (SIDS) counts for the counties of North Carolina.

Then, in Chapter 3, we extend the empirical hierarchical modeling framework of Chapter 2, which was developed for counts, to the exponential family of distributions. The data model is a conditionally independent exponential-family model. A process model is specified for some transformation of the (conditional) mean of the data model. We present the EM algorithm to obtain ML estimates of the unknown parameters in the empirical-hierarchical-modeling framework introduced in this chapter. The methodological results are illustrated and compared to some other approaches using a simulation study. We then apply our methodology to analyze a remote sensing dataset on aerosol optical depth (AOD).

Finally, in Chapter 4, we use the methodology developed in Chapter 3, to analyze a remote sensing dataset on clouds. We analyze the cloud data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) instrument, which is on board the Terra satellite that was launched in December 1999. Clouds play an important role in climate studies, and hence an accurate quantification of the spatial distribution of clouds is necessary. In this chapter, we build a spatial statistical model for the underlying clear-sky-probability (or
conversely, the cloud-probability) process, and we quantify the uncertainty in our predictions. We consider a hierarchical statistical model for analyzing the cloud data, where we postulate a hidden process for the probability of clear sky using a transformed SRE model. Its advantages are considerable: It can represent many types of spatial behavior, it permits fast computations when datasets are very large, and it has attractive change-of-support properties.
Dedicated to my Mom, Dad, Sister, and Grandpa.
ACKNOWLEDGMENTS

I would like to express my deepest thanks to my advisor, Dr. Noel Cressie. His guidance, encouragement, patience, and support have been the largest factor behind the success of the research that is presented in this dissertation. His passion for scientific research, and his attention to minor details, has always inspired me.

I would like to thank: my candidacy examination committee members (Peter Craigmile, Radu Herbei, and Desheng Liu) for their general comments and suggestions on the research presented in this dissertation; Amy Braverman for her feedback on different aspects of the research; Christopher K. Wikle for bringing up the issue of spatial confounding that is presented in Chapter 3; Brian Kahn, Richard Frey, Robert Pincus, and Mathias Schreier, for their help and input with regard to the research that is presented in Chapter 4; the computer support team at the Department of Statistics, The Ohio State University, for helping me with my computing needs; Jackie Miller for being such a helpful course coordinator with whom I had the pleasure of working for several academic terms, as a Teaching Assistant; Jon, Lili, and Matthias for being such wonderful office-mates; and the Department of Statistics, The Ohio State University, for giving me the opportunity to pursue a doctoral degree.

The research presented in this dissertation was partially supported by NASA’s Earth Science Technology Office through its Advanced Information Systems Technology Program, ROSES 2008.

For my parents, Maa and Baba, I cannot express my gratitude towards them in a few
sentences. They have always been there with me through all my whims and fancies. Whatever I am today is the result of hardships and sacrifices they have made in bringing me up, and I cannot ever imagine being able to repay that. Also, a special thanks to my sister for all her support and encouraging words, and to my grandpa for always motivating me towards my degree.

Finally, I would like to thank all my friends, especially those from Columbus, OH, for the wonderful memories I share with them. The five years in Columbus, away from home and family, would not have been the same without my long list of friends. I’ll always cherish the time I spent with them, in coffee shops, over lunch or dinner, at birthday parties, at potlucks, or on a holiday trip.
VITA

1984 ................................. Born - Kolkata, India

2005 ................................. B. Sc. in Statistics, Calcutta University, Kolkata, India.


2011-2012 .......................... Graduate Research Assistant, The Ohio State University, Columbus, Ohio.

2007-2012 .......................... Graduate Teaching Assistant, The Ohio State University, Columbus, Ohio.

PUBLICATIONS


FIELDS OF STUDY

Major Field: Statistics

Specialization: Spatial statistics, Bayesian methods, Hierarchical statistical models
# TABLE OF CONTENTS

Abstract ................................................................. ii
Dedication ................................................................. v
Acknowledgments ............................................................ vi
Vita .................................................................................. viii
List of Figures ................................................................. xiii
List of Tables ................................................................. xvii

## CHAPTER

<table>
<thead>
<tr>
<th>PAGE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>4</td>
</tr>
<tr>
<td>1.1</td>
<td>4</td>
</tr>
<tr>
<td>1.1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2</td>
<td>13</td>
</tr>
<tr>
<td>2.2.3</td>
<td>14</td>
</tr>
<tr>
<td>2.2.4</td>
<td>16</td>
</tr>
<tr>
<td>2.3</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2</td>
<td>25</td>
</tr>
<tr>
<td>2.4</td>
<td>26</td>
</tr>
<tr>
<td>2.4.1</td>
<td>27</td>
</tr>
<tr>
<td>2.4.2</td>
<td>31</td>
</tr>
<tr>
<td>2.4.3</td>
<td>33</td>
</tr>
<tr>
<td>2.4.4</td>
<td>35</td>
</tr>
<tr>
<td>2.5</td>
<td>38</td>
</tr>
</tbody>
</table>

x
Hierarchical Statistical Modeling of Big Spatial Datasets Using the Exponential Family of Distributions

3.1 Introduction
3.2 Hierarchical Statistical Model
  3.2.1 Components of the Hierarchical Statistical Model
  3.2.2 Spatial Confounding of Fixed and Random Effects
3.3 Empirical-Bayesian Inference
3.4 EM Estimation of Parameters
  3.4.1 The E-step
  3.4.2 The M-step
  3.4.3 Starting Values for the EM Algorithm
  3.4.4 Properties of the Resulting EM Algorithm
3.5 A Simulation Study
  3.5.1 Simulation Set-Up
  3.5.2 Assessment of the EM Estimates
  3.5.3 Predictive Properties
  3.5.4 Computational Time: EHM versus BHM
  3.5.5 Sensitivity Study Comparing EHM to BHM
3.6 Analysis of Aerosol Optical Depth from the MISR Instrument
  3.6.1 Background to the Dataset
  3.6.2 Hierarchical Spatial Statistical Modeling of AOD
  3.6.3 Parameter Estimation and Optimal Spatial Mapping of AOD
3.7 Discussion and Conclusions

4 Predictive Inference for MODIS Cloud Data: Hierarchical Spatial Statistical Model and Massive Datasets

4.1 Introduction
4.2 MODIS Cloud-Mask Product
4.3 Hierarchical Model for the Cloud Data
  4.3.1 Data Model
  4.3.2 Process Model
4.4 Empirical Bayesian Inference
4.5 EM Estimation of Parameters
  4.5.1 E-step
  4.5.2 M-step
  4.5.3 Starting Values for the EM Algorithm
4.6 Spatial Statistical Analysis of MODIS Cloud Data
  4.6.1 Spatial Basis Functions
  4.6.2 Optimal Spatial Mapping of the Clear-Sky Probability Process
4.7 Change-of-Support
  4.7.1 Cloud-Fraction at Resolutions Coarser than 1 km × 1 km
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Plot showing the observed counts for n=20,000 locations (left panel) and the underlying spatial process over the domain $D$ (right panel)</td>
</tr>
<tr>
<td>2.2</td>
<td>Plot showing the total time taken for EM estimation plus MCMC implementation (10,000 iterations, after a burn-in of size 2,000), as a function of sample size; the dark-gray region shows the time for EM estimation, and the light-gray region shows the time for MCMC implementation.</td>
</tr>
<tr>
<td>2.3</td>
<td>Map showing the 100 counties of North Carolina, numbered in alphabetical order of county names. (Source: Cressie and Chan, 1989)</td>
</tr>
<tr>
<td>2.4</td>
<td>Scatter plot showing FTNR on the horizontal axis and FTSIDS on the vertical axis; county 4 is shown as a “+”.</td>
</tr>
<tr>
<td>2.5</td>
<td>Residual diagnostic plots for the Poisson GLM assuming no spatial dependence.</td>
</tr>
<tr>
<td>2.6</td>
<td>Map showing 12 contiguous regions of North Carolina; county 4 is shown as white, with no shading.</td>
</tr>
<tr>
<td>2.7</td>
<td>Empirical semivariogram with a fitted exponential model superimposed (solid line). The empirical semivariogram is calculated for the random deviations, $\sqrt{E(\cdot)U(\cdot)}$, defined in Section 2.4.3.</td>
</tr>
<tr>
<td>2.8</td>
<td>Estimated covariance matrix, $\hat{\mathbf{K}}_{EM}$. The thirteenth row (and column) corresponds to the basis function for county 4.</td>
</tr>
<tr>
<td>2.9</td>
<td>Chloropleth maps showing the trend components, $\log(E(\cdot)) + \mathbf{X}(\cdot)^\top \hat{\mathbf{p}}_{EM}$ (top panel); the mean of the predictive distribution of the random-effects component, $E(\cdot)^{-1/2}E[\mathbf{S}(\cdot)^\top \mathbf{\eta}</td>
</tr>
</tbody>
</table>
2.10 Chloropleth map showing the observed relative risk, SMR (top panel), the predictive mean of the relative risk (middle panel), and the predictive standard deviation of the relative risk (bottom panel), the latter two panels being obtained from the empirical predictive distribution, $[\lambda(\cdot)|Z, \hat{\theta}_{EM}]$. 

2.11 Plots show the mean of the predictive distribution of $\mu_{Z|Y}(\cdot)$ versus the observed number of SIDS ($Z(\cdot)$), on the original scale (left panel), and on the log scale (right panel). County 4 is shown as “+”. 

2.12 Plots show the mean of the predictive distribution of $\lambda$ (relative risk), versus the observed relative risk, SMR. There is considerable shrinkage of SMR, although not for county 4, which is shown as “+”. 

3.1 The top panel shows the identity matrix, and the bottom panel shows the matrix, $\text{ave}(\hat{K}_{EM})K_T^{-1}$, where $\text{ave}(\hat{K}_{EM})$ is the elementwise average of the EM estimates $\{\hat{K}_{EM}[l] : l = 1, \ldots, 1600\}$. The common color bar is shown on the right. 

3.2 Plot showing a histogram of $\{\text{trace}(\hat{K}_{EM}[l]K_T^{-1}) : l = 1, \ldots, 1600\}$. The chi-squared density with degrees of freedom equal to $r = 29$ is overlayed on the histogram. 

3.3 Kernel-density plots showing the distribution of the ratio of the SEHM mean squared prediction error divided by the IEHM mean squared prediction error, for locations with data (solid line) and for locations without data (dashed line). 

3.4 Kernel-density plots comparing the SEHM mean squared prediction errors obtained for locations with data (solid line) and for locations without data (dashed line). 

3.5 Plots showing the Gelman-Rubin-Brooks statistics, for EHM (left panel) and for BHM (right panel), as a function of the number of MCMC samples. The solid line corresponds to the MPSRF for $\eta$; the dashed line corresponds to the maximum of the elementwise PSRFs for $\eta$. Here, the number of observations is $n = 5,000$. 

3.6 Plot showing the computational time for EHM (dark gray) and BHM (light gray). Plotted for EHM is the computational time for EM estimation plus the time it took for the MCMC algorithm to generate 15,000 samples from the empirical predictive distribution, $[\eta, \xi_O|\hat{\theta}_{EM}]$, after discarding 2,000 burn-in samples. Plotted for BHM is the computational time it took for the Gelman-Rubin-Brooks-statistics-calibrated MCMC algorithm to generate 40,000 samples from the posterior distribution, $[\eta, \xi, \theta|Z_O]$, after discarding 2,000 burn-in samples.
3.7 Plot showing the observed data (top-left panel), the true simulated process, $Y(\cdot)$ (top-right panel), the mean of the empirical predictive distribution, $\hat{Y}_{SEHM}(\cdot) \equiv E(Y(\cdot)|Z_O, \hat{\theta}_{EM})$ (bottom-left panel), and the mean of the posterior distribution, $\hat{Y}_{SBHM}(\cdot) \equiv E(Y(\cdot)|Z_O)$ (bottom-right panel).

3.8 Kernel-density plot showing the distribution of the difference, $\hat{Y}_{SEHM}(\cdot) - \hat{Y}_{SBHM}(\cdot)$.

3.9 Kernel-density plots showing the distribution of the ratio, $(\text{var}(Y(s)|Z_O))^{1/2} / (\text{var}(Y(s)|Z_O, \hat{\theta}_{EM}))^{1/2}$, separately for locations $s$ in $D_O$ (locations with data) and for locations $s$ in $D_U$ (locations without data).

3.10 Plot showing the study region corresponding to the AOD dataset. The light-gray regions correspond to land surfaces, and the dark-gray regions correspond to water bodies.

3.11 AOD data in $D$ (left panel) and histogram showing their distribution (right panel).

3.12 Maps to the left show the log(AOD) (top-left panel), the mean (middle-left panel) and standard deviation (bottom-left panel) of the predictive distribution of $Y(\cdot)$, namely $[Y(\cdot)|Z_O, \hat{\theta}_{EM}]$. Maps to the right show the predictive mean of the different components of variability in $Y(\cdot)$, namely, the components due to trend, $X(\cdot)^\top \hat{\beta}_{EM}$ (top-right panel), the random-effects component, $E[S(\cdot)^\top \eta|Z_O, \hat{\theta}_{EM}]$ (middle-right panel), and the fine-scale-variation component, $E[\xi(\cdot)|Z_O, \hat{\theta}_{EM}]$ (bottom-right panel). The middle-left panel which is a map of the mean of the predictive distribution of $Y(\cdot)$, namely $E[Y(\cdot)|Z_O, \hat{\theta}_{EM}]$, is the sum of the three panels shown on the right.

3.13 Boxplots show the variability of the predictive standard deviation of $Y(s_i)$ for values of $m(s_i) = 1, 2, \ldots, 21$.

3.14 Maps show the predictive mean (top-left panel), the pixelwise predictive standard deviation (top-right panel), the pixelwise predictive 2.5 percentile (bottom-left panel), and the pixelwise predictive 97.5 percentile (bottom-right panel) obtained from the empirical predictive-distribution, $[\mu_{Z|Y(\cdot)}|Z_O, \hat{\theta}_{EM}]$. The plots of the predictive mean and the predictive percentiles have the same color scale, where any value greater than 1 has been assigned the highest color-value.

4.1 An example of a granule image obtained by the MODIS instrument on board NASA's Terra satellite (June 29, 2006, 12:45 UTC). The inset shows the location of the granule on a world map. (Source: modis-atmos.gsfc.nasa.gov.)
4.2 A pictorial illustration of a MODIS cloud mask spectral test, which is based on an upper and lower bound. ........................................ 99

4.3 Initial Q-values (top panel) and final Q-values (bottom panel) corresponding to the granule shown in Figure 4.1. .......................... 101

4.4 Geographical location of the data granule on June 29, 2006, 12:45 UTC . . 119

4.5 A two-dimensional bisquare function as a 3-D plot (left) and as an image plot (right). .......................................................... 120

4.6 Centers of the basis function; 'o', '+', and 'x' are use to distinguish the three scales of resolution. .................................................. 121

4.7 Maps showing the components due to trend, \( \mathbf{X}(\cdot)^\top \hat{\mathbf{\beta}}_{EM} \) (top-left panel), the mean of the predictive distribution of the random-effects component, \( E[\mathbf{S}(\cdot)^\top \eta| \mathbf{Q}, \hat{\mathbf{\theta}}_{EM}] \) (top-right panel), the mean of the predictive distribution of the component due to fine scale variation, \( E[\xi(\cdot)| \mathbf{Q}, \hat{\mathbf{\theta}}_{EM}] \) (bottom-left panel), and the mean of the process \( Y(\cdot), E[Y(\cdot)| \mathbf{Q}, \hat{\mathbf{\theta}}_{EM}] \) (bottom-right panel). The bottom-right panel is obtained by adding the three other components. .......................................................... 123

4.8 Maps showing the predictive mean (top-left panel), the pixelwise predictive standard-deviation (top-right panel), the pixelwise 2.5 percentile (bottom-left panel) and the pixelwise 97.5 percentile (bottom-right panel) for the predictive distribution of the clear-sky-probability process. . . . . . . 125
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>69</td>
</tr>
<tr>
<td>3.2</td>
<td>77</td>
</tr>
<tr>
<td>3.3</td>
<td>78</td>
</tr>
</tbody>
</table>

3.1 True parameter values and the sample mean of the EM parameter estimates based on 1600 simulated datasets. Each dataset is of size $n = 20,000$. The empirical root mean squared errors (RMSEs) of the parameter estimates are also reported.

3.2 Gelman-Rubin-Brooks statistics for varying sample sizes ($n$). The number of MCMC samples generated are $L=15,000$ for EHM, and $L=40,000$ for BHM. MPSRF is the multivariate potential scale reduction factor, and max(PSRF) is the maximum of the elementwise potential scale reduction factors (PSRFs).

3.3 Computational time for varying sample sizes ($n$). For EHM, the EM algorithm was used to estimate the parameters, and then an MCMC algorithm was used to generate $L_{EHM} = 15,000$ samples from the empirical predictive distribution, $[\eta_0, \xi_0 | Z_0, \hat{\theta}_{EM}]$. For BHM, an MCMC algorithm was used to generate $L_{BHM} = 40,000$ samples from the posterior distribution, $[\eta, \xi, \theta | Z_0]$. 
CHAPTER 1
INTRODUCTION

Observations are taken in order to inform about the etiology of a scientific phenomenon. The challenge of inferring the underlying scientific process in a coherent manner from the data has been addressed in the statistics literature using hierarchical statistical modeling. In this framework, one specifies a model for the data conditional on an underlying true process and unknown parameters; then a process model is specified for the underlying true process, which is also conditional on unknown parameters. If a parameter model for all the unknown parameters is included, then the hierarchical model is called a Bayesian hierarchical model (BHM). Alternatively, if the parameters are assumed fixed but unknown, and they are estimated from the data, then the hierarchical model is called an empirical hierarchical model (EHM). For a more complete discussion of the BHM and the EHM approaches, see Cressie and Wikle (2011, Chapter 2).

In this dissertation, I consider the EHM approach for spatial generalized linear models (SGLM). Such models are especially appropriate in health studies, spatial econometrics, ecological studies, climate studies, analysis of remote sensing datasets, and so forth, where the data may be counts or binary (discrete), or skewed or positive (continuous). The datasets are spatially referenced, and hence one needs to consider models that involve spatial statistical dependence. There is a huge literature on spatial statistics that builds on the
“first law of geography” (Tobler, 1970), which states that: everything is related to everything else, but near things are more related than distant things. For an overview of the field of spatial statistics, see Cressie (1993).

This dissertation introduces a class of spatial models that is hierarchical, empirical, non-stationary, and could be used for analyzing very-large-to-massive non-Gaussian datasets. Specifically, I consider a hierarchical statistical model consisting of a conditionally specified exponential-family model for the observations (the data model), and a transformed Gaussian field for the underlying (hidden) geostatistical process (the process model). If one also includes a parameter model (i.e., prior distribution), one obtains a BHM, such as the models considered by Omre and Tjelmeland (1997) and Diggle et al. (1998).

In the statistics literature, there are several examples of spatial models within a hierarchical modeling framework (Besag et al., 1991; Diggle et al., 1998; Wikle et al., 1998). Besag et al. (1991) showed how a spatial model for counts in small areas could be decomposed hierarchically, where the logarithm of the (hidden) disease risk incorporates the spatial dependence. They assumed that the counts were (conditionally) Poisson distributed and that the log Poisson means were a Gaussian spatial process, specifically the Gaussian Markov Random Field (MRF) known as the conditional autoregressive (CAR) model. However, a simultaneous autoregressive (SAR) model, or a geostatistical model could also be used. Diggle et al. (1998) employed spatial generalized linear mixed models for spatially dependent non-Gaussian variables observed potentially anywhere in the spatial domain of interest, and they assumed a (hidden) process model that was geostatistical. Their hierarchical model was fully Bayesian and required a Markov Chain Monte Carlo (MCMC) algorithm to obtain the posterior distribution. In a spatio-temporal context, Wikle et al. (1998) developed a fully Bayesian hierarchical-model formulation for modeling a dataset of monthly maximum temperatures.

In contrast to the fully Bayesian approach, Heagerty and Lele (1998) developed a
method for binary data, where they used a composite-likelihood approach (e.g., Lindsay, 1988) to estimate the spatial-hierarchical-model parameters. Zhang (2002) gave a Monte Carlo version of the EM Gradient Algorithm to analyze non-Gaussian data, and Monestiez et al. (2006) developed a method called Poisson kriging for mapping the relative abundance of species.

The class of spatial hierarchical models defined above is not suitable for handling large datasets, which are becoming very common. For example, large datasets include those coming from remote sensing of the Earth by satellites, climate-model output, and small-area samples from national surveys. The existence of large-to-massive datasets have led to an active area of research involving the use of reduced-rank representations of the hidden Gaussian process (e.g., see the review in Wikle, 2010). These reduced-rank methods are based on geostatistical models, where a continuously indexed Gaussian process is used to specify the hidden process.

The intersection of reduced rank modeling and hierarchical modeling has been an active area of research. The case when data appear to be Gaussian has been well explored, both for BHMs (e.g., Banerjee et al., 2008; Lopes et al., 2008; Kang and Cressie, 2011; Katzfuss and Cressie, 2012) and EHMs (e.g., Cressie and Johannesson, 2006, 2008; Katzfuss and Cressie, 2009). When data appear to be non-Gaussian, most hierarchical modeling is based on BHMs (e.g., Wikle et al., 2001; Hooten et al., 2003; Royle and Wikle, 2005; Lopes et al., 2011). In this dissertation, my emphasis is on EHMs for massive, non-Gaussian datasets, based on the dimension-reducing Spatial Random Effects (SRE) model (e.g., Cressie and Johannesson, 2006, 2008). The resulting SGLMs are quite flexible and have computational efficiency that is proportional to the sample size. I consider different types of non-Gaussian datasets and use SGLMs to analyze them.
1.1 Organization of the Dissertation

My PhD dissertation is organized according to a “paper” format. This introductory chapter is followed by three chapters, which have been, or will be, submitted for publication. Each chapter is self-contained and includes an introduction, a conclusion, and an Appendix.

The research presented in this dissertation was carried out with the help of my advisor, Dr. Noel Cressie. Dr. Cressie proposed the problems addressed in each of the chapters and provided general comments on my approach to solve those problems. I was responsible for the technical derivations, the computer code, and writing initial drafts. A brief summary of each chapter is described below.

Chapter 2: Empirical hierarchical modeling for count data using the Spatial Random Effects model

Chapter 2 considers data observed in the form of counts. In this chapter, we consider a hierarchical statistical model made up of a Poisson model for the counts, and an underlying (hidden) geostatistical process for the log of the mean of the Poisson distribution. The hidden process is modeled using the SRE model. We develop maximum likelihood (ML) estimates for the parameters of the geostatistical process using an EM algorithm. With the help of a simulation study, we demonstrate that the total computational time is proportional to the size of the data. We then use the EHM to analyze a spatial dataset of Sudden Infant Death Syndrome (SIDS) counts for the counties of North Carolina.

Chapter 3: Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions

In Chapter 3, we extend the hierarchical-statistical-modeling framework of Chapter 2, which was developed for counts, to include the exponential family of distributions. That is, we developed an SGLM. We investigate the statistical and computational performance of our methodology using a simulation study. We then use the EHM to analyze remote sensing dataset on aerosol optical depth (AOD).

This chapter has been submitted for publication in a peer-reviewed journal as "Sengupta, A., and Cressie, N. Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions."

Chapter 4: Predictive inference for MODIS cloud data: Hierarchical spatial statistical model and massive datasets

In Chapter 4, we use the hierarchical-statistical-modeling framework to analyze the cloud mask data from NASA’s MODIS instrument, which is on board the Terra satellite that was launched in December 1999. Clouds play an important role in climate studies, and hence an accurate quantification of clear-sky probabilities is necessary. To our knowledge, until now there has been no attempt to build a spatial statistical model for these types of massive datasets. In this chapter, we develop a hierarchical spatial statistical model for the MODIS cloud product, which is obtained on a very fine 1 km × 1 km scale. Each of the 2.7 million 1 km × 1 km pixels in our dataset produces a “cloudiness” datum in the interval [0, 1]. The statistical model that we consider in this chapter is hierarchical, spatial, and is based on a transformed SRE model.

In *Proceedings of the 2012 Joint Statistical Meetings*, Alexandria, VA: American Statistical Association.” A submission to a peer-reviewed journal is also being prepared.
CHAPTER 2
EMPIRICAL HIERARCHICAL MODELING FOR COUNT DATA
USING THE SPATIAL RANDOM EFFECTS MODEL

2.1 Introduction

Count data are ubiquitous in spatial econometrics, since many economic datasets are derived from rates (i.e., counts divided by base counts). For example, List (1989) developed models based on counts to determine which county-level attributes are conducive to attracting new foreign manufacturing plants. Cameron and Trivedi (1986) and Wedel et al. (1993) discuss econometric models for count data based on the Poisson distribution. To capture spatial dependency, the commonly used models in the econometrics literature are additive, Gaussian, and include a spatially lagged dependent variable; see Elhorst (2010) for details about these spatial simultaneous autoregressive (SAR) models. Such models do not adapt easily to count data, and the purpose of this chapter is to offer spatial statistical models that do.

The literature on spatial statistical methods for analyzing count data has often been based on Markov random fields (MRFs); for a survey of such approaches, see Besag et al. (1991), Cressie (1993, Chapter 6), and Banerjee et al. (2004, Chapter 3). Besag et al. (1991) showed how a spatial model for counts could be decomposed hierarchically so that a hidden process is used to model the spatial dependence. For example, the counts are assumed (conditionally) Poisson distributed, and the logarithm of the Poisson means follow
a Gaussian spatial process. That process could be a MRF (e.g., a conditional autoregressive (CAR) model), a SAR model, or a geostatistical model. Breslow and Clayton (1993) used a MRF for the random effects in their hierarchical modeling framework, as did Kang et al. (2009). Zhu et al. (2007) developed a maximum-likelihood approach to estimating the parameters in a MRF using a stochastic approximation to the EM algorithm.

Another approach to Bayesian hierarchical modeling for count data was developed by Frühwirth-Schnatter and Wagner (2006), where the mean changes over time according to an underlying latent process. LeSage et al. (2007) adopted this approach to study the spatial impact on how knowledge flows between regions in Europe, as captured through patent citations. For the spatial-effects parameters they used SAR priors.

Diggle et al. (1998) employed spatial generalized linear mixed models (GLMMs) for spatially dependent non-Gaussian variables (e.g., counts) observed over a continuous region, using hidden geostatistical processes. Later, Zhang (2002) developed a Monte Carlo version of the EM Gradient Algorithm to analyze non-Gaussian data, and Monestiez et al. (2006) developed a method called Poisson kriging for mapping the relative abundance of species. Application of the latter article to health data was given by Goovaerts (2005, 2006). These approaches are based on geostatistical models, and they assume stationarity.

Despite the popularity of these approaches, there is often a computational bottleneck when the data size is large. The modeling approach in LeSage et al. (2007) ignores the possible interaction effects between the source and the destination of the patent citations, because of its large dimension. Research presented in the articles by Wikle (2002), Hooten et al. (2003), Royle and Wikle (2005), Rue et al. (2009), Lindgren et al. (2011), Lopes et al. (2011), and this chapter address this issue. Wikle (2002) developed a kernel-based spectral approach for non-Gaussian spatio-temporal processes. Hooten et al. (2003), and Royle and Wikle (2005) extended this method for analyzing binary data and count data, respectively.
Their approach to modeling has been a fully Bayesian one, and is based on Markov Chain Monte Carlo (MCMC).

Rue et al. (2009) proposed a method for carrying out approximate Bayesian inference based on the integrated nested Laplace approximation (INLA). They use the Laplace approximation (see Tierney and Kadane, 1986; Tierney et al., 1989; Kass and Steffey, 1989; Steele, 1996) repeatedly to obtain approximations to the marginal posterior densities of the hidden process values. The underlying latent process is modeled using a GMRF with a sparse precision matrix.

Many of the spatial and spatio-temporal applications for very-large-to-massive datasets center around reduced-rank representations of geostatistical models (e.g., Wikle, 2010), where the hidden process is modeled using a continuous Gaussian field (GF). To bridge the gap, Lindgren et al. (2011) developed an “explicit link” between a continuously indexed GF and a GMRF. They suggested doing the modeling using a GF and all the computations using the GMRF representation. Their approach based on the INLA allows for fast computations, albeit with only a small number of parameters. As pointed in Kang and Cressie (2012), there is a compromise in terms of the flexibility of the spatial models when using the homogeneous MRF to model the spatial dependence over large, heterogeneous regions.

In the context of reduced-rank approaches for generalized linear spatial models, Lopes et al. (2011) introduced a new class of spatio-temporal models for data coming from an exponential family of distributions, where they model the spatial and temporal components using a latent factor-analysis structure; their fully Bayesian model allows for dimension reduction and hence fast computations. They model the spatial variation via the columns of a factor-loading matrix only, and their vector of common factors has a diagonal covariance matrix; priors have to be very tightly constrained for reasons of model identifiability.

In this chapter, we propose a flexible class of spatial models for analyzing count data;
the models are hierarchical, geostatistical, nonstationary, and they are computationally efficient to analyze. We take an empirical hierarchical modeling (EHM) approach, where the unknown parameters are estimated by maximum likelihood (ML) estimation. The model is not fully Bayesian, but Bayes’ Theorem is used to obtain the all-important predictive distribution; for a more complete discussion of the EHM approach, see Cressie and Wikle (2011, Chapter 2).

Our approach to modeling spatial count data is an extension of the Spatial Random Effects (SRE) model of Cressie and Johannesson (2006, 2008), developed for continuous, symmetric data with a continuous spatial index. Dimension reduction is achieved by modeling the underlying spatial process as a linear combination of spatial basis functions on the spatial domain of interest. This is important for spatial best linear unbiased prediction (i.e., kriging), which involves inverting the covariance matrix of the observed random vector. For the SRE model, the matrix inversion is a relatively simple task, the model is well suited to change-of-support, and it avoids any stationarity assumptions for the covariance matrix. Assuming the data are Gaussian, Katzfuss and Cressie (2009) gave an EM algorithm to obtain ML estimates for SRE-model parameters; the Bayesian hierarchical model (BHM) version puts prior distributions on the parameters rather than estimating them (Kang and Cressie, 2011; Katzfuss and Cressie, 2012). Unlike Lopes et al. (2011), the SRE model does not assume a diagonal covariance matrix for the spatial random effects. Instead, it captures spatial dependency using both the pre-specified spatial basis functions, and the covariance matrix of the random effects.

When data come in the form of counts, a spatial analysis using hierarchical statistical models is not straightforward. In the EHM proposed below, we use the EM algorithm (Dempster et al., 1977) to obtain ML estimates of the parameters in the model. However, the expectations in the E-step of the algorithm are not available in closed form, and we use the Laplace approximation to approximate the intractable integrals. Having obtained the
estimates for the unknown parameters, we substitute them into the predictive distribution
and use MCMC to generate samples from it.

The plan of this chapter is as follows. In Section 2.2, we describe our modeling approach for count data, based on a hidden SRE model, and we outline the statistical inference based on generating samples from the predictive distribution. Then, in Section 2.3, we describe the EM algorithm for obtaining ML estimates of the model parameters described in Section 2.2. In Section 2.4, we use our model to analyze a dataset of Sudden Infant Death Syndrome (SIDS) counts for the counties of North Carolina. Discussion and conclusions follow in Section 2.5, and technical derivations are given in the Appendix.

### 2.2 Spatial Random Effects Model for Poisson Data

Our main goal here is to carry out empirical hierarchical modeling and inference in a spatial setting, where the size of the dataset can be very large. In this section, we introduce a hidden SRE model for count data, and we then discuss inference based on the (empirical) predictive distribution.

#### 2.2.1 Model Specification

We consider a hierarchical model for spatial data that come in the form of counts. In this chapter, the counts are obtained from small areas. Let the areas be denoted by \( \{ A_i : i = 1, \ldots, N \} \), located at \( \{ s_i : i = 1, \ldots, N \} \). The spatial domain of interest is \( \bigcup_{i=1}^{\ N} A_i \), and we define our spatial index set to be \( D \equiv \{ s_1, \ldots, s_N \} \).

We model the counts using a Poisson distribution, where the mean depends on an underlying spatial process, denoted by \( Y(\cdot) \), which is unobserved. A datum will be denoted as
$Z(s)$, where $s$ is its location (in $D$). Denoting the observed locations as \{$s_i : i = 1, \ldots, n$\} $\subset D$, we write the data model as:

$$Z(s_i) | Y(s_i) \sim \text{ind. Poisson} \left( \mu_{Z|Y}(s_i) \right), \ i = 1, \ldots, n,$$

(2.1)

where “ind” is an abbreviation for “independent,” and $\mu_{Z|Y}(\cdot)$ is the conditional mean of the Poisson distribution. Following the generalized-linear-modeling approach, we consider the log link, namely,

$$Y(\cdot) \equiv \log(\mu_{Z|Y}(\cdot)).$$

(2.2)

We select the log link because it is the canonical link for the Poisson family of distributions (see McCullagh and Nelder, 1989; Diggle et al., 1998), and because $\mu_{Z|Y}(\cdot) = \exp(Y(\cdot))$ is always non-negative.

We consider a Gaussian spatial process for the underlying (hidden) $Y(\cdot)$, which is indexed by the set $D = \{s_i : i = 1, \ldots, N\}$, where $N \geq n$. Hence, there may be a set of locations where there are no data; without loss of generality, it is given by the set $\{s_i : i = n + 1, \ldots, N\}$. When $n$ is very large, the SAR and CAR models (e.g., Cressie, 1993, Section 6.3) may not be computationally feasible unless their respective spatial-weights matrices are considered sparse. In this chapter, we present a geostatistical model for very large, but finite $N$, where inference is scalable in data size $n$. The process model here is the underlying stochastic process $Y(\cdot)$; we use the reduced rank Spatial Random Effects (SRE) model given by Cressie and Johannesson (2006, 2008), modified here for the count data:

$$Y(s_i) = C(s_i) + X(s_i)^\top \beta + W(s_i) \left( S(s_i)^\top \eta + \xi(s_i) \right), \ i \in \{1, \ldots, N\}.$$

(2.3)

In (2.3), $C(\cdot)$ is an offset term; $W(\cdot)$ is a known weight that accounts for (typically demographic) strata of different rates; $X(\cdot)$ is a vector of known covariates; $\beta$ is the regression parameter associated with the covariates; $S(\cdot) \equiv (S_1(\cdot), \ldots, S_r(\cdot))^\top$ is a vector of $r$ (not necessarily orthogonal) spatial basis functions, where $r$ is fixed; $\eta$ is an $r$-dimensional
vector of spatial random effects; and \( \xi(\cdot) \) is a component that captures extra-Poisson variability. The random effects, \( \eta \), are assumed to have a multivariate Gaussian distribution with mean 0 and a covariance matrix \( K \), where \( K \) is not necessarily a diagonal matrix; that is, \( \eta \sim \text{Gau}(0, K) \). The fine-scale variation component, \( \xi(\cdot) \), is composed of independent \( \text{Gau}(0, \sigma^2_\xi v_\xi(\cdot)) \) random variables, where \( \sigma^2_\xi > 0 \), and \( v_\xi(\cdot) \) is known. The term \( \xi(\cdot) \) captures any extra-Poisson variability coming from overdispersion.

The interest here lies primarily in the hidden spatial process, \( Y(\cdot) \), or equivalently in \( \mu_{Z|Y}(\cdot) = \exp(Y(\cdot)) \), which is unobserved. Our main goals are to estimate unknown parameters and to predict \( Y(s_i) \), for \( s_i \in D \), based on the observed counts \( Z_O = (Z(s_1), \ldots, Z(s_n))^\top \).

In the next section, we draw an analogy between the spatial simultaneous autoregressive (SAR) model and the SRE model proposed in this chapter.

### 2.2.2 The SRE Model and Spatial Simultaneous Autoregression

The spatial simultaneous autoregression (SAR) is the commonly used model in the spatial econometrics literature. In this section, we derive conditions on the covariance matrix of our spatial model on \( Y(\cdot) \), given by (2.3), such that the two models have the same covariance structure.

Rewriting (2.3) in vector notation, we have,

\[
Y = C + X\beta + W(S\eta + \xi),
\]

(2.4)

where \( Y \equiv (Y(s_1), \ldots, Y(s_N))^\top \), \( C \equiv (C(s_1), \ldots, C(s_N))^\top \), \( X \equiv (X(s_1), \ldots, X(s_N))^\top \), \( W \equiv \text{diag}(W(s_1), \ldots, W(s_N)) \), \( S \equiv (S(s_1), \ldots, S(s_N))^\top \), and \( \xi = (\xi(s_1), \ldots, \xi(s_N))^\top \). The corresponding SAR model is given by (e.g., Cressie, 1993, p.406):

\[
Y = C + X\beta + (I_N - B)^{-1}\gamma,
\]

(2.5)

where \( B \) is the spatial-weights matrix for this SAR model, \( I_N \) is the \( N \times N \) identity matrix,
and $\gamma$ is independent and identically distributed (i.i.d.) $\text{Gau}(0, \Lambda)$, where $\Lambda$ is a diagonal matrix.

From (2.4) and (2.5), the means match; to match the variances and the covariances we must have

$$W(SK^T + \sigma_\xi^2 V_{\xi})W = (I_N - B)^{-1}\Lambda(I_N - B^T)^{-1}. \quad (2.6)$$

That is,

$$SK^T + \sigma_\xi^2 V_{\xi} = W^{-1}(I_N - B)^{-1}\Lambda(I_N - B^T)^{-1}W^{-1}. \quad (2.7)$$

Let $S = Q_S R_S$ be the $Q$-$R$ decomposition of $S$, where $Q_S$ is an $n \times r$ orthonormal matrix and $R_S$ is a non-singular $r \times r$ upper-triangular matrix. Then from (2.7), we obtain an expression for $K$, as a function of $\sigma_\xi^2$, as:

$$K(\sigma_\xi^2) = R_S^{-1}Q_S^T \left\{ W^{-1}(I_N - B)^{-1}\Lambda(I_N - B^T)^{-1}W^{-1} - \sigma_\xi^2 V_{\xi} \right\} Q_S(R_S^{-1})^{-1}, \quad (2.8)$$

where $\sigma_\xi^2 > 0$ is such that $\left\{ W^{-1}(I_N - B)^{-1}\Lambda(I_N - B^T)^{-1}W^{-1} - \sigma_\xi^2 V_{\xi} \right\}$ is positive-definite. There can be none, one, or many such $\sigma_\xi^2$. If there exists such a $\sigma_\xi^2 > 0$ such that the aforementioned positive-definiteness condition holds, then the resulting covariance matrix for the equivalent SRE model is:

$$\Sigma_{Y}^{SRE}(\sigma_\xi^2) \equiv W \left[ P(W^{-1}\Sigma_{Y}^{SAR}W^{-1}) + \sigma_\xi^2 \{ V_{\xi} - P(V_{\xi}) \} \right] W,$$

where $P(A) \equiv Q_S A Q_S^T A Q_S$, $\Sigma_{Y}^{SAR} \equiv (I_N - B)^{-1}\Lambda(I_N - B^T)^{-1}$, and recall that $W$ is a known diagonal matrix.

### 2.2.3 Empirical Hierarchical Model

Our main focus here is on prediction of $\mu_{Z|Y}(\cdot) = \exp(Y(\cdot))$; it is seen in Section 2.4, that $\lambda(\cdot) \equiv \mu_{Z|Y}(\cdot)/\exp(C(\cdot))$ is also of interest. After having observed $Z_O$ at locations $\{s_1, \ldots, s_n\}$, respectively, we are interested in inference on the hidden process $Y(s)$ (or
equivalently \( \mu_{Z|Y}(s) \) at any spatial location \( s \in D = \{ s_1, \ldots, s_N \} \). The parameters \( \beta, K \), and \( \sigma^2_\xi \) are unknown and considered to be nuisance parameters for the purpose of prediction. Our hierarchical model becomes empirical when we estimate the parameters (Section 2.3) and substitute them into the predictive distribution of \( Y(\cdot) \).

Recall that \( Z_O = (Z(s_1), \ldots, Z(s_n))^\top \), and write \( Y = (Y_O^\top, Y_U^\top)^\top \), where

\[
Y_O^\top \equiv (Y(s_1), \ldots, Y(s_n)), \quad \text{and} \quad Y_U^\top \equiv (Y(s_{n+1}), \ldots, Y(s_N)).
\]

Similarly, we define \( C \equiv (C_O^\top, C_U^\top)^\top \), \( X \equiv (X_O^\top, X_U^\top)^\top \), \( W \equiv (W_O^\top, W_U^\top)^\top \), \( S \equiv (S_O^\top, S_U^\top)^\top \), and \( \xi \equiv (\xi_O^\top, \xi_U^\top)^\top \). Using the notation \( [U|V] \) to represent the conditional distribution of \( U \) given \( V \), we see that

\[
E \left( Y(s_i) | Z_O, \beta, K, \sigma^2_\xi \right) = C(s_i) + X(s_i)^\top \beta + W(s_i) \left\{ S(s_i)^\top E \left( \eta | Z_O, \beta, K, \sigma^2_\xi \right) \right\}
\]

\[
\text{var} \left( Y(s_i) | Z_O, \beta, K, \sigma^2_\xi \right) = W(s_i) \left\{ S(s_i)^\top \text{var} \left( \eta | Z_O, \beta, K, \sigma^2_\xi \right) S(s_i) + \sigma^2_\xi (s_i) \right\} W(s_i).
\]

Thus, \( \xi_U \) is independent of \( (Z_O, \eta, \xi_O) \), and hence for a site in \( \{ s_i : i = n+1, \ldots, N \} \) that has no observation, we have:

\[
E \left( Y(s_i) | Z_O, \beta, K, \sigma^2_\xi \right) = C(s_i) + X(s_i)^\top \beta + W(s_i) \left\{ S(s_i)^\top E \left( \eta | Z_O, \beta, K, \sigma^2_\xi \right) \right\}
\]

\[
\text{var} \left( Y(s_i) | Z_O, \beta, K, \sigma^2_\xi \right) = W(s_i) \left\{ S(s_i)^\top \text{var} \left( \eta | Z_O, \beta, K, \sigma^2_\xi \right) S(s_i) + \sigma^2_\xi (s_i) \right\} W(s_i).
\]
The goal is to predict \( Y \), given the data. However, the predictive distribution, \( [Y|Z,O,\beta,K,\sigma^2_\xi] \), is not available in closed form. We shall use Markov Chain Monte Carlo (MCMC) (see, e.g., Robert and Casella, 2004) to obtain samples from the predictive distribution, \( [Y|Z,O,\beta,K,\sigma^2_\xi] \). This is equivalent to obtaining samples from the predictive distribution, \( [\eta,\xi,O|Z,O,\beta,K,\sigma^2_\xi] \), and from the predictive (Gaussian) distribution, \( [\xi_U|\sigma^2_\xi] \).

The inference described in this section is not fully Bayesian inference. We do have a hierarchical model, but we do not put prior distributions on the parameters, \( \{\beta,K,\sigma^2_\xi\} \), unlike Diggle et al. (1998). Instead, we estimate the parameters, which is in line with much of the inference in geostatistics for data that are continuous and symmetric (e.g., Cressie, 1993, Chapter 3). In actuality, we substitute in parameter estimates \( \hat{\beta}, \hat{K}, \hat{\sigma}^2_\xi \) to obtain MCMC samples from the empirical predictive distributions, \( [\eta,\xi,O|Z,O,\hat{\beta},\hat{K},\hat{\sigma}^2_\xi] \) and \( [\xi_U|\hat{\sigma}^2_\xi] \). The MCMC algorithm is described in the Appendix, and estimation of the parameters (using the EM algorithm) is presented in Section 2.3.

### 2.2.4 Computational Efficiency for the MCMC Algorithm

In this section, we illustrate the computational efficiencies that are achieved using a hidden SRE model; we shall see that the time it takes to implement the MCMC is of the same order as the sample size \( n \). The illustration is based on a simulated dataset \( Z,O \), which was obtained as follows.

Consider a regular spatial domain, \( D = \{s_1, \ldots, s_N\} \), consisting of \( N = 300 \times 300 = 90,000 \) points on \( \{-149.5, \ldots, -0.5, 0.5, \ldots, 149.5\}^2 \). In this simulation, the hidden process \( Y(\cdot) \) is made up of the three components:

\[
Y(s) = X(s)^\top \beta + S(s)^\top \eta + \xi(s), \ s \in D. \tag{2.12}
\]

That is, here the offset term \( C(\cdot) \) is taken to be zero, and the weights \( W(\cdot) \) are considered to be unity. The large-scale variation, or trend, is assumed to be

\[
X(s)^\top \beta = \beta_0 + \beta_1 \times s_2, \tag{2.13}
\]
where \( s = (s_1, s_2)^\top \) and \( \beta = (\beta_0, \beta_1)^\top \).

We assume the random effects \( \eta \sim \text{Gau}(0, K) \), and \( \xi(\cdot) \) are independent and identically distributed (i.i.d.) \( \text{Gau}(0, \sigma_\xi^2) \), independent of \( \eta \). To specify the SRE model’s covariance matrix \( K \), we started with an exponential covariance function given by
\[
C(u, v) = c_0 \exp \left( -\frac{||u - v||}{a_0} \right),
\]
(2.14)
where \( c_0 \) is the sill and \( a_0 \) is the scale parameter. Here we specified \( c_0 = 1 \) and \( a_0 = 100 \).

Let \( v \equiv (v(s_1), \ldots, v(s_N))^\top \) be a mean-zero spatial Gaussian process defined over \( D \), whose covariance matrix is obtained from the exponential covariance model (2.14); that is, \( v \sim \text{Gau}(0, \Sigma_v) \). We calibrate \( K \) and \( \sigma_\xi^2 \) using the procedure in Kang and Cressie (2011). For the purpose of calibration, we consider only 9,000 regularly spaced locations that cover the entire spatial domain (rather than using all 90,000 locations).

First we calculate \( K^0 \) such that \( ||SK^0S^\top - \Sigma_v|| \) is minimized, where \( || \cdot || \) is the Frobenius norm (e.g., Cressie and Johannesson, 2008). Finally, to control the variability of \( Y \), we choose \( K = gK^0 \), where \( g \) is chosen to preserve the total variation. That is,
\[
\text{trace}(\Sigma_v)/N = 1 = \text{trace}(gSK^0S^\top + \sigma_\xi^2 I_N)/N.
\]
(2.15)

For selecting the large-scale variation parameter \( \beta \), we define the variation of the “signal,” \( V_s \), as:
\[
V_s = \frac{1}{N} \text{trace} \left( SKS^\top + \sigma_\xi^2 I_N \right) + \frac{1}{N} \sum_{i=1}^{N} \left( X(s_i)^\top \beta - \text{average}_{s_i \in D} \left( X(s_i)^\top \beta \right) \right)^2.
\]

The parameter \( \beta \) was selected such that \( V_s \) is approximately 2 (see Aldworth and Cressie, 1999, Section 3.2.4). In (2.13), we specified \( \beta = (2, 0.0125)^\top \), which gives \( V_s = 2.17 \). Additionally, we specified the fine-scale variation proportion (FVP), which is defined as
\[
FVP = \frac{\text{trace}(\sigma_\xi^2 I_N)}{\text{trace}(SKS^\top + \sigma_\xi^2 I_N)},
\]
(2.16)
which from (2.15) is equal to $\sigma_\zeta^2$. In our simulation, $FVP$ was held at 5%; hence $\sigma_\zeta^2 = 0.05$. Using (2.15), we obtained $g = 1.22$.

We simulated $\eta$ and $\xi$ from the Gaussian process defined above and then, using (2.12), we obtained $Y$ over the entire domain $D$. Next, we use the inverse-link function,

$$\mu_{Z|Y}(\cdot) = \exp(Y(\cdot)),$$  

(2.17)

to simulate a realization of the conditionally (conditional on $Y(\cdot)$) independent Poisson random variables, $Z_O$, for only $n$ locations ($n \leq N$); the $n$ locations $\{s_1, \ldots, s_n\}$ were randomly sampled without replacement from the $N = 90,000$ possible locations. The resulting realization of the entire spatial process $Y(\cdot)$ and the $n$ data $Z_O$ are shown in Figure 2.1. In what follows, we vary $n$ and tabulate the computational efficiency as a function of $n$.

Figure 2.1: Plot showing the observed counts for $n=20,000$ locations (left panel) and the underlying spatial process over the domain $D$ (right panel)
Having obtained $Z_0$ for the $n$ locations, we ran the MCMC (here using the true parameter values). For each chosen $n$, we recorded the time it took for our MCMC algorithm to generate 10,000 samples from the target distribution, after allowing a burn-in of 2,000 samples. The simulation was performed on a dual quad core 2.8 GHz 2x Xeon X5560 processor, with 96 Gbytes of memory. The time for the MCMC was recorded as 7.1 mins ($n = 5,000$), 25.4 mins ($n = 20,000$), 43 mins ($n = 35,000$), and 62 mins ($n = 50,000$). This illustrates that the computational complexity is linear in $n$; see Figure 2.2.

2.3 EM Estimation of Parameters

The inference described in Section 2.2.3 assumed that all the parameters were known; in practice, we need to estimate the parameters from the data. In this section, we consider the task of obtaining the ML estimates of the parameters using the EM algorithm.

2.3.1 EM algorithm

The EM algorithm (Dempster et al., 1977) has been employed for estimation of parameters in the presence of missing data; for more details on the EM algorithm, see McLachlan and Krishnan (2008). For the hierarchical model described in Section 2.2.1, the random effects $\eta$ and the fine-scale variation $\xi_0$ are not known and can be treated as missing. Zhang (2002) proposed a Monte Carlo version of the EM Gradient Algorithm, however that will be difficult to adopt in this setting since we do not have a stationary structure for the covariance matrix of $Y$, and we cannot parameterize it in terms of a few elements. The EM algorithm involves iterating between an E (expectation) step and an M (maximization) step, and in our case the E-step is the most problematic. We resolve this problem by using Laplace approximations to evaluate the expectations.

We begin by writing down the “complete data” log likelihood, $L_c$, for the unknown parameters, given the data. The complete data here is made up of the observations $Z_0$ and
the unobserved $\eta$ and $\xi$. Then $L_e$ is simply the logarithm of the joint distribution of $Z, \eta$, and $\xi$.

$$L_e(\beta, K, \sigma_\xi^2 | Z, \eta, \xi) = \log [Z | \beta, \eta, \xi] + \log [\eta | K] + \log [\xi | \sigma_\xi^2]$$

$$= \text{const.} - \sum_{i=1}^n \exp \left\{ C(s_i) + X(s_i)^\top \beta + W(s_i) \left( S(s_i)^\top \eta + \xi(s_i) \right) \right\}$$

$$+ \sum_{i=1}^n Z(s_i) \left\{ C(s_i) + X(s_i)^\top \beta + W(s_i) \left( S(s_i)^\top \eta + \xi(s_i) \right) \right\}$$

$$- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left( \eta \eta^\top K^{-1} \right)$$

$$- \frac{n}{2} \log \sigma_\xi^2 - \frac{1}{2} \text{trace} \left( \xi \xi^\top V^{-1}_{\xi} \right),$$ (2.18)

where recall that $[U|V]$ denotes the density function of $U$ given $V$, $V_{\xi} \equiv \text{diag}(v_\xi(s_1), \ldots, v_\xi(s_n))$, and “const.” denotes a generic constant that does not depend on $\beta$, $K$, or $\sigma_\xi^2$. Assume we are at the $(l+1)$-th iteration of the EM algorithm. At the E-step, we need to take the expectation of the complete data log likelihood with respect to the parameter values at the $l$-th iteration, giving rise to a function we denote as $Q(\cdot, \cdot)$. All the expectations in the E-step are with respect to $(\eta, \xi)$. Let us write $\theta = \{ \beta, K, \sigma_\xi^2 \}$. Then the E-step of the EM algorithm is:

$$Q(\theta, \theta^{[l]}) = E \left( L_e(\theta | Z, \eta, \xi) | Z, \theta^{[l]} \right)$$

$$= \text{const.} - \sum_{i=1}^n \exp \left\{ C(s_i) + X(s_i)^\top \beta \right\} E \left( \exp \left\{ W(s_i) \left( S(s_i)^\top \eta + \xi(s_i) \right) \right\} | Z, \theta^{[l]} \right)$$

$$+ \sum_{i=1}^n Z(s_i) \left\{ C(s_i) + X(s_i)^\top \beta + W(s_i) (S(s_i)^\top \eta + \xi(s_i)) + E(\xi(s_i) | Z, \theta^{[l]}) \right\}$$

$$- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left( E \left( \eta \eta^\top | Z, \theta^{[l]} \right) K^{-1} \right)$$

$$- \frac{n}{2} \log \sigma_\xi^2 - \frac{1}{2} \text{trace} \left( E \left( \xi \xi^\top | Z, \theta^{[l]} \right) V^{-1}_{\xi} \right).$$ (2.19)

The expectations involved in the E-step of the EM algorithm are not available in closed form. One common approach is to implement a stochastic EM (SEM) algorithm (see McLachlan and Krishnan, 2008; Robert and Casella, 2004), where the expectations are evaluated using Monte Carlo integration. In order to implement SEM, we need to simulate
from the posterior distributions, $[\eta|Z_O, K^{[l]}]$ and $[\xi_O|Z_O, \sigma^2_{\xi^{[l]}}]$, within the $(l + 1)$-th iteration of the algorithm; see (2.19). Since these distributions are not available in closed form, an MCMC algorithm would be needed, something we wish to avoid as much as possible. Hence, we derive instead Laplace approximations to the expectations involved in (2.19). The Laplace approximation (LA) is based on a second-order Taylor-series expansion of the logarithm of the integrand around the mode, which leads to a Gaussian density for which the integral can be computed exactly.

To apply the LA, we need to obtain $\left(\hat{\eta}^{[l]}, \hat{\xi}_O^{[l]}\right)$, which maximizes the complete data likelihood. We use a coordinate-wise ascent method, which maximizes alternately with respect to $\eta$, and then $\xi_O$, until convergence. Convergence of the coordinate-wise ascent method typically requires that the objective function (which is $L_c(\theta^{[l]}|Z_O, \eta, \xi_O)$ in our case) be strictly concave and differentiable, with respect to $(\eta, \xi_O)$. Denote $\delta \equiv \left(\eta^T, \xi_O^T\right)^T$. Then it is easy to see that the second-derivative matrix of $L_c(\theta^{[l]}|Z_O, \delta)$, with respect to $\delta$, is negative-definite; hence, the coordinate-wise ascent method will converge to $\hat{\eta}^{[l]}$ and $\hat{\xi}_O^{[l]}$.

Once the maximizer of the complete data likelihood has been obtained, we use the LA to approximate the first two moments of $[\eta, \xi_O, Z_O, \theta^{[l]}]$. We use the first-order approximation that is given in Kass and Steffey (1989). Second-order approximations (e.g., Steele, 1996) are also possible, but they are substantially more complicated.

Using a Taylor-series approximation up to second order, we can approximate the posterior distribution $[\delta|Z_O, \theta^{[l]}]$ with a Gaussian distribution with mean and variance given by, respectively, the posterior mode and the inverse of the negative Hessian of the posterior evaluated at the mode. These normality assumptions can be justified using Laplace’s
method (Tierney and Kadane, 1986), which is shown in the Appendix. As a result, approximately,
\[
E \left( \begin{pmatrix} \eta \\ \xi_O \end{pmatrix} \mid Z_O, \theta^l \right) = \begin{pmatrix} \hat{\eta}^l \\ \hat{\xi}_O^l \end{pmatrix}
\] (2.20)

and, approximately,
\[
\text{var} \left( \begin{pmatrix} \eta \\ \xi_O \end{pmatrix} \mid Z_O, \theta^l \right) = \left\{ \begin{array}{l} -\frac{\partial^2}{\partial \eta \partial \eta} \left( L_c(\theta^l|Z_O, \eta, \xi) \right) \\ -\frac{\partial^2}{\partial \xi_O \partial \eta} \left( L_c(\theta^l|Z_O, \eta, \xi) \right) \\ -\frac{\partial^2}{\partial \xi_O \partial \xi_O} \left( L_c(\theta^l|Z_O, \eta, \xi) \right) \end{array} \right\}^{-1}.
\] (2.21)

To obtain \(\text{var}(\eta|Z_O, \theta^l)\) and \(\text{var}(\xi_O|Z_O, \theta^l)\), we need to invert the matrix of second partial derivatives given above. Recall the Sherman-Morrisson-Woodbury formula (e.g., Henderson and Searle, 1981). Let \(A\) denote an \(r \times r\) matrix and \(D\) denote an \(n \times n\) matrix. Further, let \(U\) be any \(r \times n\) matrix and \(V\) be any \(n \times r\) matrix. The Sherman-Morrisson-Woodbury formula gives us:
\[
(A + UDV)^{-1} = A^{-1} - A^{-1}U(D^{-1} + VA^{-1}U)^{-1}VA^{-1}.
\]

We use this formula in the block-matrix inversion formula given in Duncan (1944) to obtain the following formula:
\[
\begin{pmatrix} A & U \\ V & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - UD^{-1}V)^{-1} & -(A - UD^{-1}V)^{-1}UD^{-1} \\ -D^{-1}V(A - UD^{-1}V)^{-1} & D^{-1} + D^{-1}V(A - UD^{-1}V)^{-1}UD^{-1} \end{pmatrix},
\] (2.22)

where the left-hand side of (2.22) is a general form of the right-hand side (2.21). Now, for generic variables \(u\) and \(v\), define
\[
J(u_0, v_0) = -\frac{\partial^2}{\partial u \partial v} \left( L_c(\theta^l|Z_O, u, v) \right) \bigg|_{u=u_0, v=v_0}.
\]
We look at the different component matrices in the approximation (2.21). The matrix $J(\hat{\xi}_O, \hat{\xi}_O)$ is an $n \times n$ diagonal matrix; its inversion is easy. The matrix $J(\hat{\eta}_O, \hat{\eta}_O)$ is of dimension $r \times r$, where $r << n$. The other two matrices, $J(\hat{\xi}_O, \hat{\eta}_O)$, and $J(\hat{\eta}_O, \hat{\xi}_O)$, have dimension $n \times r$ and $r \times n$ respectively. We use the formula (2.22) to invert (2.21). This leads us to an approximate expression for $\text{var}(\eta|Z_O, \theta^{[l]})$, $\text{var}(\xi_O|Z_O, \theta^{[l]})$, and $\text{cov}(\eta, \xi_O|Z_O, \theta^{[l]})$, namely

$$\text{var}(\eta|Z_O, \theta^{[l]}) = \left( J(\hat{\eta}_O, \hat{\eta}_O) - J(\hat{\eta}_O, \hat{\xi}_O)J(\hat{\xi}_O, \hat{\xi}_O) - 1 J(\hat{\xi}_O, \hat{\eta}_O) \right)^{-1}$$

$$\text{var}(\xi_O|Z_O, \theta^{[l]}) = J(\hat{\xi}_O, \hat{\xi}_O)^{-1} + J(\hat{\xi}_O, \hat{\eta}_O)^{-1} J(\hat{\eta}_O, \hat{\xi}_O)^{-1} J(\hat{\xi}_O, \hat{\eta}_O) \times \left( J(\hat{\eta}_O, \hat{\eta}_O) - J(\hat{\eta}_O, \hat{\xi}_O)J(\hat{\xi}_O, \hat{\xi}_O) - 1 J(\hat{\xi}_O, \hat{\eta}_O) \right)^{-1}$$

$$\times J(\hat{\eta}_O, \hat{\xi}_O)J(\hat{\xi}_O, \hat{\xi}_O)^{-1}$$

$$\text{cov}(\eta, \xi_O|Z_O, \theta^{[l]}) = - \left( J(\hat{\eta}_O, \hat{\eta}_O) - J(\hat{\eta}_O, \hat{\xi}_O)J(\hat{\xi}_O, \hat{\xi}_O) - 1 J(\hat{\xi}_O, \hat{\eta}_O) \right)^{-1}$$

$$\times J(\hat{\eta}_O, \hat{\xi}_O)J(\hat{\xi}_O, \hat{\xi}_O)^{-1}.$$

(2.23)

Note that all we need to invert is the $n \times n$ diagonal matrix, $J(\hat{\xi}_O, \hat{\xi}_O)$, and some fixed-rank $r \times r$ matrices. This makes the computations extremely efficient and allows us to obtain expressions for $E(\eta\eta^T|Z_O, \theta^{[l]})$ and $E(\xi_O\xi_O^T|Z_O, \theta^{[l]})$ in (2.19):

$$E(\eta\eta^T|Z_O, \theta^{[l]}) = \text{var}(\eta|Z_O, \theta^{[l]}) + E(\eta|Z_O, \theta^{[l]})E(\eta|Z_O, \theta^{[l]}\eta^T)$$

$$E(\xi_O\xi_O^T|Z_O, \theta^{[l]}) = \text{var}(\xi_O|Z_O, \theta^{[l]}) + E(\xi_O|Z_O, \theta^{[l]})E(\xi_O|Z_O, \theta^{[l]}\eta^T),$$

(2.24)

where the terms on the right-hand side of (2.24) are approximated (using the LA) from (2.20) and (2.23).

The remaining term in (2.19), for which we need an approximation, is

$$E\left( \exp\left\{ W(s_i) \times \left( S(s_i)^T \eta + \xi(s_i) \right) \right\} |Z_O, \theta^{[l]} \right), i = 1, \ldots, n.$$
From the Appendix, we see that, approximately,

\[
E \left( \exp \left\{ W(s_i) \times \left( S(s_i) \top \eta + \xi(s_i) \right) \right\} \mid Z_O, \theta^{[l]} \right) = \exp \left\{ W(s_i) \times \left( S(s_i) \top \hat{\eta}^{[l]} + \hat{\xi}^{[l]}(s_i) \right) \right\} + \frac{W(s_i)^2}{2} \left( S(s_i) \top \text{var}(\eta|Z_O, \theta^{[l]})S(s_i) + 2S(s_i) \top \text{cov}(\eta, \xi|Z_O, \theta^{[l]} e(s_i) \right) + e(s_i) \top \text{var}(\xi|Z_O, \theta^{[l]} e(s_i) \right),
\]

(2.25)

where \( e(s_i) \) is a vector of length \( n \) that has 1 in the \( i \)-th position and all other entries in \( e(s_i) \) are zero, for \( i = 1, \ldots, n \).

In the M-step, we maximize (2.19) with respect to each of the parameters in \( \theta \): The maximization with respect to \( K \) and \( \sigma^2_\xi \) is obtained by differentiating (2.19) with respect to \( K \) and \( \sigma^2_\xi \), and then by equating the result to zero. The updating equations are obtained in closed form:

\[
\sigma^2_\xi^{[l+1]} = \frac{1}{n} \text{trace} \left( \left( E(\xi|Z_O, \theta^{[l]})E(\xi|Z_O, \theta^{[l]}) \top + \text{var}(\xi|Z_O, \theta^{[l]}) \right) V^{-1}_{\xi|Z_O} \right)
\]

\[
K^{[l+1]} = E(\eta|Z_O, \theta^{[l]})E(\eta|Z_O, \theta^{[l]}) \top + \text{var}(\eta|Z_O, \theta^{[l]}).
\]

(2.26)

However, the maximization with respect to \( \beta \) is not available in closed form; we use a one-step Newton-Raphson update as follows:

\[
\beta^{[l+1]} = \beta^{[l]} - \left\{ \frac{\partial}{\partial \beta} R(\theta) \right\}^{-1} \left. R(\theta^{[l]}) \right|_{\theta = \theta^{[l]}},
\]

(2.27)

where \( R(\theta) \) denotes the score function obtained by differentiating \( Q(\cdot, \cdot) \) given by (2.19) with respect to \( \beta \); \( R(\theta^{[l]}) \) is obtained by evaluating \( R(\theta) \) at \( \theta^{[l]} \).

The iterations from the E-step to the M-step to the E-step, etc., proceed until convergence, that is, until \( \theta^{[l]} \) does not change according to a prespecified tolerance. The resulting EM estimate of \( \theta \) is denoted as \( \hat{\theta}_{EM} \equiv \left\{ \hat{\beta}_{EM}, \hat{K}_{EM}, \hat{\sigma}^2_{\xi,EM} \right\} \).

The EM algorithm does not automatically produce measures of standard errors for the maximum likelihood estimates (MLEs). The asymptotic covariance matrix for the MLEs can be computed in a number of different ways. Notable contributions include those by
Louis (1982) and Oakes (1999). Rather than trying to adapt existing methodology to the hierarchical spatial model proposed in this chapter, we use a parametric bootstrap (e.g., Efron and Tibshirani, 1993, Section 6.5) to compute the standard deviation of the sampling distribution of the MLEs. We proceed by simulating realizations from the fitted empirical hierarchical model in the following sequence: simulate $[Y|\hat{\theta}_{EM}]$, followed by $[Z_{O}|Y]$. Then, based on the simulated $Z_{O}$, we compute the MLEs for $\theta$ using the EM algorithm. We repeat this process 1600 times and compute the mean and the standard deviation of the MLEs over these iterations.

### 2.3.2 Computational Efficiency for EM Estimation

In this section, we assess the computational efficiency of estimation of the parameters $\theta = \{\beta, K, \sigma^2_\xi\}$. Recall the simulation set-up described in Section 2.2.4, where we considered a spatial domain of size $N = 90,000$. Here, we repeat the same simulation experiment as in Section 2.2.4, with the same sample sizes that were used there, where now we use the EM algorithm to estimate the SRE-model parameters for each simulation. The times taken to obtain the EM estimates on a dual quad core 2.8 GHz 2x Xeon X5560 processor with 96 Gbytes of memory, were recorded. The true parameter values used in the simulation were used as the starting values for the EM algorithm, because our interest in this section is only in the computational speed achieved by using the SRE model in the hierarchical spatial statistical model.

As in Section 2.2.4, we worked with four different values for $n$: The time for the EM estimation was recorded as 0.93 mins ($n = 5,000$), 1.95 mins ($n = 20,000$), 5.14 mins ($n = 35,000$), and 7.23 mins ($n = 50,000$). A plot of the computational times for the EM estimation (Section 2.3.1) and the MCMC (Section 2.2.4) is given in Figure 2.2. For each sample size, the vertical bar is divided into two segments, showing the time it took for the
EM estimation separately from the time it took for the MCMC. Clearly, the computational efficiencies are *linear* in the sample size.

![Graph showing computational time vs. sample size](image)

**Figure 2.2**: Plot showing the total time taken for EM estimation plus MCMC implementation (10,000 iterations, after a burn-in of size 2,000), as a function of sample size; the dark-gray region shows the time for EM estimation, and the light-gray region shows the time for MCMC implementation.

### 2.4 Counts from Small Areas: SIDS in North Carolina

In this section, county-level count data of Sudden Infant Death Syndrome (SIDS) is analyzed based on the hierarchical spatial model given in Section 2.2.1. While it is not a large dataset, it illustrates the hierarchical nature of our approach to spatial modeling of counts;
the computational advantages for large datasets are addressed directly through simulation in Sections 2.2.4 and 2.3.2.

2.4.1 Background of the Data and Exploratory Data Analysis

Sudden infant death syndrome (SIDS) is a classification of unexpected, sudden death in apparently healthy babies less than one year of age. However, its etiology is still a mystery. In this section, we analyze a dataset of counts that contains SIDS information for the 100 counties of North Carolina, from 1974 to 1978. Cressie and Chan (1989) used a CAR model of spatial dependency after transforming the SIDS rates. Here we use a hierarchical spatial model to analyze the counts, which avoids having to make Cressie and Chan’s approximate Gaussian assumption on transformed rates. We use the same labels as in Cressie and Read (1985) and Cressie and Chan (1989) to number the counties from 1 to 100, as shown in Figure 2.3. Cressie and Chan (1989) used a heterogeneous CAR model; here we approach the analysis without assuming a CAR model. In the previous analyses of this dataset by Cressie and co-authors, county 4 (Anson county) was detected as an outlier and was set aside when fitting the CAR model. Here, we include county 4 in our analysis, using a separate spatial basis function (an indicator function) to account for its extra variability, something that is easy to do for the SRE model.

We first introduce the following important notations:

- $s_i$ denotes the location of county $i$, for $i = 1, \ldots, n$. In our case, the location is the coordinate of the county seat, and $n = N = 100$ (i.e., no missing data).

- $Z(s_i)$ denotes the observed number of SIDS in county $i$.

- $Z_k(s_i)$ denotes the observed number of SIDS for subgroup $k$ in county $i$, where $k = 1, \ldots, 4$ represents the subgroups, “white male,” “white female,” “non-white male,” and “non-white female,” respectively.
Figure 2.3: Map showing the 100 counties of North Carolina, numbered in alphabetical order of county names. (Source: Cressie and Chan, 1989)

- $B(s_i)$ denotes the total number of live births in county $i$.

- $B_k(s_i)$ denotes the total number of live births for subgroup $k$ in county $i$.

- $E(s_i)$ denotes the expected number of SIDS in county $i$, defined as

$$ E(s_i) \equiv \sum_{k=1}^{4} B_k(s_i) \left( \frac{\sum_{j=1}^{n} Z_k(s_j)}{\sum_{j=1}^{n} B_k(s_j)} \right). $$

This is an example of internal standardization, where the SIDS data is also used to compute the expected number of SIDS for each county.

- $SMR(s_i)$ is the standardized mortality ratio (SMR) in county $i$, defined as

$$ SMR(s_i) \equiv Z(s_i)/E(s_i). $$
The Freeman-Tukey transformation of counts is essentially the square-root transformation, which approximately removes the functional dependence between the mean and the variance of \( \{ \text{SMR}(s_i) \} \). As in Cressie and Read (1985), we shall use the Freeman-Tukey transformed rates to do an exploratory data analysis. For county \( i \), define:

\[
FTNR(s_i) \equiv \sqrt{\frac{1000 [B_3(s_i) + B_4(s_i)]}{B(s_i)}} + \sqrt{\frac{1000 [B_3(s_i) + B_4(s_i) + 1]}{B(s_i)}},
\]

which is the Freeman-Tukey transformed non-white live birth rate (per thousand live births), and

\[
FTSIDS(s_i) \equiv \sqrt{\frac{1000Z(s_i)}{B(s_i)}} + \sqrt{\frac{1000(Z(s_i) + 1)}{B(s_i)}},
\]

which is the Freeman-Tukey transformed SIDS rate (per thousand live births). A scatter plot showing the relationship between \( FTNR(\cdot) \) and \( FTSIDS(\cdot) \) is given in Figure 2.4. The scatter plot shows that the observed SIDS rate depends positively on the non-white live birth rate. This suggests keeping \( FTNR(\cdot) \) as an explanatory variable in our hierarchical spatial analysis.

We continue our exploratory data analysis by fitting a non-spatial Generalized Linear Model (GLM) to \( Z(\cdot) \); for further details on GLMs, see McCullagh and Nelder (1989). In our case, we fitted the model

\[
Z(s_i) \mid \mu_Z(s_i) \sim \text{ind. Poisson}(\mu_Z(s_i)).
\]

For this exploratory data analysis, we consider the log link, and fit a linear model to the log of the mean \( \mu_Z(\cdot) \):

\[
\log \mu_Z(s_i) = \log(E(s_i)) + \beta_0 + \beta_1(FTNR(s_i)), \tag{2.30}
\]

where \( E(\cdot) \), which is defined in (2.28), results in an offset term, and the explanatory variable is \( FTNR(\cdot) \). For fitting the non-spatial GLM defined in equation (2.30), we ignored county 4; it will be included when we fit the hierarchical model that is defined
Figure 2.4: Scatter plot showing FTNR on the horizontal axis and FTSIDS on the vertical axis; county 4 is shown as “+”.

later in Section 2.4.2. The iteratively weighted least squares estimate of $\beta = (\beta_0, \beta_1) \top$ is $\hat{\beta}_{GLM} = (-0.293313, 0.007568) \top$, and the corresponding standard errors obtained are $(\text{var}(\hat{\beta}_{1;GLM}))^{1/2} = 0.154437$ and $(\text{var}(\hat{\beta}_{2;GLM}))^{1/2} = 0.00404$. We also obtain the Pearson residuals corresponding to the GLM fit, which are defined as:

$$r_p(s) \equiv \frac{Z(s) - \exp(\log(E(s)) + X(s) \top \beta)}{\sqrt{\exp(\log(E(s)) + X(s) \top \beta)}}.$$

The diagnostic plots in Figure 2.5 show that the model does capture the large-scale variation. However, the model has no component to capture the small-scale spatial dependency.
2.4.2 Hierarchical Spatial Statistical Model for the SIDS Data

Following the model specifications defined in Section 2.2.1, we specify the data model as:

\[ Z(s_i) | \lambda(s_i) \sim \text{ind. Poisson}(E(s_i)\lambda(s_i)), \]

where \( \lambda(s_i) \equiv \mu_{Z,Y}(s_i)/E(s_i) \) is the underlying relative risk parameter for county \( i \) (e.g., Cressie and Wikle, 2011, p. 194). The process model is given by:

\[
Y(s_i) = \log(E(s_i)\lambda(s_i)) \\
= \log(E(s_i)) + \beta_0 + \beta_1 FTNR(s_i) + E(s_i)^{-1/2} \left\{ S(s_i)^\top \eta + \xi(s_i) \right\},
\]

where, from (2.3), \( C(s_i) \equiv \log(E(s_i)) \) is the offset term, and \( W(s_i) \equiv E(s_i)^{-1/2} \) are the weights (e.g., Cressie and Chan, 1989). Recall that \( E(s_i) \) is the expected number of SIDS.
in county \( i \), obtained from internal standardization. The statistical and epidemiologic literature contains many discussions on the relative merits of internal and external standardization (see Waller and Gotway, 2004, Section 2.3). As is standard in the epidemiology literature, any random variation in \( \{E(s_i)\} \) is assumed small and ignored.

Amongst other things, we wish to predict the relative risks \( \{\lambda(s_i) : i = 1, \ldots, N\} \), for which \( \{SMR(s_i) : i = 1, \ldots, n\} \) given by (2.29) represent noisy and possibly incomplete observations. In (2.32), \( \eta \) is a 13-dimensional vector \((r = 13)\) of random effects: We specify one random effect for each of 12 contiguous regions of North Carolina defined in Cressie and Chan (1989) and shown in Figure 2.6; and the last random effect is for county 4, which was previously treated as an outlier and deleted from the inference. The vector of spatial basis functions \( S(s_i) \) for county \( i \) is defined as,

\[
S(s_i) \equiv (I(i \in 1), \ldots, I(i \in 12), I(i = 4))^\top,
\]

(2.33)

where \( I(i \in j) \) takes a value 1 if county \( i \) lies in region \( j \) \( (j = 1, \ldots, 12) \), and 0 otherwise; \( I(i = 4) \) takes the value 1 if \( i = 4 \), and 0 otherwise. The random-effects vector, \( \eta \), is assumed to be distributed according to a multivariate Gaussian density given by,

\[
\eta \sim \text{Gau}(0, \mathbf{K}).
\]

(2.34)

Notice that we multiply \( S(s_i)^\top \eta \) by the weight \( E(s_i)^{-1/2} \) to account for heterogeneity among the counties.

The final component in (2.32), \( \xi(s_i) \), captures the extra Poisson variability, and it is modeled as,

\[
\xi(s_i) \sim \text{ind. Gau}(0, \sigma_\xi^2).
\]

(2.35)

Likewise, we multiply \( \xi(s_i) \) by \( E(s_i)^{-1/2} \) to account for heterogeneity among the counties.
2.4.3 Starting values for the EM algorithm

In the next section, parameter estimates are obtained using the EM algorithm, but to implement it we need starting values for the parameters. For the SIDS data, there was evidence of sensitivity to the choice of initial values, so we pay particular attention to choosing them well.

We use the non-spatial GLM estimate $\hat{\beta}_{GLM}$, obtained in Section 2.4.1, as our starting value for $\beta$. For $K$ and $\sigma^2_\xi$, we proceed as follows. Define the random variable $U(\cdot)$ as:

$$U(s) \equiv Y(s) - \log(E(s)) - X(s)^T \hat{\beta}, \ s \in D.$$  \hfill (2.36)
Using the linear model for $Y(\cdot)$ specified in (2.32), we obtain the following equivalent definition of $U(\cdot)$:

$$U(s) = E(s)^{-1/2} \left( S(s)^\top \eta + \xi(s) \right).$$

(2.37)

From (2.37), it immediately follows that:

$$\text{var} \left( \sqrt{E(s)} U(s) \right) = S(s)^\top K S(s) + \sigma^2$$

$$\text{cov} \left( \sqrt{E(s)} U(s), \sqrt{E(u)} U(u) \right) = S(s)^\top K S(u).$$

(2.38)

In practice, $Y(\cdot)$ is not observed. Here we will approximate $Y(\cdot)$ as:

$$Y(s) \approx \log \left( Z(s) + 1 \right),$$

where 1 is added to account for zero counts. Consequently, an approximation for $U(\cdot)$ is obtained as:

$$U(s) \approx \log \left( \frac{Z(s) + 1}{E(s)} \right) - X(s)^\top \hat{\beta}_{GLM}.$$

(2.39)

We now proceed by fitting a semivariogram model to the empirical semivariogram obtained for the weighted random deviations, $\sqrt{E(\cdot)} \times U(\cdot)$.

We define the distance between two counties based on the distance between the coordinates of the county seat. First, we obtained the empirical semivariogram estimates (shown as dots in Figure 2.7). We fitted an exponential model to the empirical semivariogram using the weighted-least-squares method, with weights suggested by Cressie (1985).

For fitting the semivariogram model, we weighted the random variables $U(\cdot)$ with $\sqrt{E(\cdot)}$, which takes into account the heterogeneity among the counties (see Cressie and Read, 1989). Figure 2.7 shows the plot of the empirical semivariogram, along with the fitted model. The fitted semivariogram model has a partial-sill, a range, and a nugget effect, and hence we use it to estimate a covariance matrix for $D_E^{1/2} U$, where $U \equiv (U(s_1), \ldots, U(s_n))^\top$, and $D_E \equiv \text{diag} (E(s_1), \ldots, E(s_n))$. Let us write the estimated covariance matrix as:

$$\hat{\text{cov}}(D_E^{1/2} U) = \hat{\Sigma} + \hat{c}_0 I_n,$$

(2.40)
where $\hat{c}_0$ is the estimated nugget variance, and $\hat{\Sigma}$ is the $n \times n$ covariance matrix that is computed using the partial-sill and the range of the fitted exponential semivariogram model. Now, using (2.38) and (2.40), we have approximately:

$$\hat{\Sigma} + \hat{c}_0 I_n = S_O K S_O^\top + \sigma^2_\xi I_n.$$  \hspace{1cm} (2.41)

Hence, $\hat{c}_0$ is used as the starting value for $\sigma^2_\xi$. Upon using the relation,

$$S_O K S_O^\top \approx \hat{\Sigma}$$

and the $Q$-$R$ decomposition of $S_O = Q_S R_S$, we obtain the starting value,

$$K = R_S^{-1} Q_S^\top \hat{\Sigma} Q_S (R_S^\top)^{-1},$$ \hspace{1cm} (2.42)

for $K$ (e.g., Kang and Cressie, 2011).

### 2.4.4 Parameter Estimation and Optimal Spatial Mapping of SIDS

Using starting values specified in Section 2.4.3, we implemented the EM algorithm to obtain parameter estimates for the hierarchical model (2.31)-(2.32) fitted to the SIDS data discussed in Section 2.4.1. The EM algorithm converged in 14 iterations, yielding

$$\hat{\beta}_{EM} = (-0.214945, 0.00452)^\top,$$

and

$$\hat{\sigma}^2_{\xi,EM} = 0.0947.$$
2.8. Notice that the regression coefficient of FTNR is relatively small; its importance has diminished from our initial exploratory data analysis (see Figure 2.4), due to inclusion of $C(s_i) = \log(E(s_i))$ as an offset term in (2.32). Nevertheless, it does explain a small amount of spatial variability, and we keep it in our model.

Next, $\hat{\theta}_{EM} \equiv \{\hat{\beta}_{EM}, \hat{K}_{EM}, \hat{\sigma}^2_{\xi, EM}\}$ is substituted into our MCMC procedure to obtain samples from the (empirical) predictive distribution, $[\eta, \xi_O | Z_O, \hat{\theta}_{EM}]$. We generated 10,000 samples using the MCMC method, after allowing for a burn-in of size 2,000. Once we have this sample, we can obtain the entire predictive distribution, $[Y | Z_O, \hat{\theta}_{EM}]$, and we
Figure 2.8: Estimated covariance matrix, $\hat{K}_{EM}$. The thirteenth row (and column) corresponds to the basis function for county 4.

can compute any desired summary statistic of $Y$. The hierarchical nature of the model allows us to look into the different sources of variability separately. We first obtain choropleth maps separately for the trend component, for the predictive mean of the random effects (the small-scale-variation component) multiplied by the term $E(\cdot)^{-1/2}$, and the predictive mean of the extra-Poisson variation term, also multiplied by $E(\cdot)^{-1/2}$. Recall from (2.32) that these are the three components of variation in our model. We then add them together to obtain the mean of the predictive distribution, $[Y|Z_O, \hat{\theta}_{EM}]$. The choropleth maps produced
are shown in Figure 2.9. We also obtain choropleth maps for the observed SMR, the mean of the predictive distribution of the relative risk, and the standard deviation of the predictive distribution of the relative risk, which are shown in Figure 2.10.

To diagnose the fit of the model, Figure 2.11 shows plots for the observed counts versus the predicted counts, on the original count scale and on the log scale. Further, a plot of the observed relative risk (i.e., the SMR), versus the predictive mean of the relative risk is shown in Figure 2.12. From Figure 2.12, we see the considerable shrinkage of the SMR to the predictive mean of the underlying relative risk. Furthermore, the extra spatial random effect that we used for county 4 accounts for its unusually high risk.

2.5 Discussion and Conclusions

In this chapter, we develop a hierarchical spatial statistical model for counts, based on a hidden SRE model of the spatial covariances of a latent random process. We show a relationship between the SRE model and the SAR model. Importantly, the SRE model allows for dimension reduction, which is advantageous when analyzing a large dataset. From a simulated dataset of counts, we demonstrated that our spatial analysis is scalable in \( n \), the size of the dataset. We also used our model to study SIDS in the counties of North Carolina, from 1974 to 1978, using the flexibility of the model to account for an unusually high SIDS rate in one of the counties. In this example, \( n = N = 100 \) and \( r = 13 \). The SIDS data are counts summed over a period of 4 years. There were only a few zeros, and hence zero-inflated models were not used. For further information on such models in the spatial context, see Agarwal et al. (2002).

We take an empirical hierarchical modeling (EHM) approach and develop the EM algorithm to estimate the unknown parameters in the model. To do inference, we generated
Figure 2.9: Chloropleth maps showing the trend components, $\log(E(\cdot)) + X(\cdot)^T\hat{\beta}_{EM}$ (top panel); the mean of the predictive distribution of the random-effects component, $E(\cdot)^{-1/2}E[S(\cdot)^T\eta|Z_O, \hat{\theta}_{EM}]$ (second panel); and the mean of the predictive distribution of the extra Poisson variability, $E(\cdot)^{-1/2}E[\xi(\cdot)|Z_O, \hat{\theta}_{EM}]$ (third panel). The bottom panel is a chloropleth map of the mean of the predictive distribution of $Y(\cdot)$, namely $E[Y(\cdot)|Z_O, \hat{\theta}_{EM}]$, which is the sum of the top three panels.
Figure 2.10: Chloropleth map showing the observed relative risk, SMR (top panel), the predictive mean of the relative risk (middle panel), and the predictive standard deviation of the relative risk (bottom panel), the latter two panels being obtained from the empirical predictive distribution, $\mathbb{E}[\lambda(Z)|\mathbf{Z}, \hat{\theta}_{EM}]$.

MCMC samples from the (empirical) predictive distribution. Computing for both estimation and prediction is fast. Estimation (here, via the EM algorithm) avoids having to put prior distributions on the parameters of the hierarchical model.
Figure 2.11: Plots show the mean of the predictive distribution of $\mu_{Z|Y}(\cdot)$ versus the observed number of SIDS ($Z(\cdot)$), on the original scale (left panel), and on the log scale (right panel). County 4 is shown as “+”.

Datasets in the form of counts are not uncommon in econometric analysis, where interest often lies in predicting rates (e.g., unemployment rates in small areas, or hotspots of housing foreclosure rates). The hierarchical statistical modeling approach given here provides flexible spatial structure, is relatively straightforward to implement, is computationally scalable in the size of the dataset, and is hierarchical but not fully Bayesian.
Figure 2.12: Plots show the mean of the predictive distribution of $\lambda$ (relative risk), versus the observed relative risk, SMR. There is considerable shrinkage of SMR, although not for county 4, which is shown as “+”.
CHAPTER 3
HIERARCHICAL STATISTICAL MODELING OF BIG SPATIAL DATASETS USING THE EXPONENTIAL FAMILY OF DISTRIBUTIONS

3.1 Introduction

Big spatial datasets are very common in scientific problems, such as those involving remote sensing of the earth by satellites, climate-model output, small-area samples from national surveys, and so forth. In this chapter, our interest lies primarily in datasets that are very large and non-Gaussian in form. We consider a hierarchical statistical model consisting of two levels. At the first level, we have an exponential-family model for the data given a spatial process and parameters (which we call the data model). At the second level, we assume a geostatistical process given parameters (which we call the process model), for some transformation of the mean of the data model.

The exponential family of distributions include commonly used continuous and discrete distributions; for a detailed review, see McCullagh and Nelder (1989, Section 2.2.2). All members of the exponential family have a density or probability mass function that can be written as:

\[ p(z|\gamma) = \exp\left\{ \frac{(z\gamma - b(\gamma))}{\tau^2} - c(z, \tau) \right\}, \]  

(3.1)

where \( \gamma \) is called the canonical parameter or the natural parameter, \( b(\gamma) \) is a function that
depends only on $\gamma$, $c(z, \tau)$ is a function independent of $\gamma$, and $\tau$ is a scaling constant. The representation above is called the canonical form, or the natural form, of the exponential family.

Examples of distributions that belong to the exponential family include the Gaussian distribution, the log Gaussian distribution, the Gamma distribution, the Poisson distribution, the Bernoulli distribution, and the Binomial distribution. Let $Z_1, \ldots, Z_n$ denote data coming from a member of the exponential family such that $\{ [Z_i|\gamma_1, \ldots, \gamma_n] : i = 1, \ldots, n \}$ are mutually independent, and $[Z_i|\gamma_1, \ldots, \gamma_n] \equiv [Z_i|\gamma_i]$, where $[Z_i|\gamma_i]$ has density given by (3.1). Then the expectation of $Z_i$ is given by:

$$E(Z_i|\gamma_i) = b'(\gamma_i), \quad (3.2)$$

where, assuming $b(\cdot)$ is differentiable, the last equality is easily derived. In a generalized linear model (GLM), we write down a regression model for some transformation of the mean $E(Z_i|\gamma_i)$, defined through a link function $g(\cdot)$; that is, we write:

$$g(E(Z_i|\gamma_i)) = X_i^\top \beta, \quad (3.3)$$

where $X_i$ denotes a p-dimensional vector of known covariates, $\beta$ is a p-dimensional vector of regression coefficients, and there are a lot of possible choices for $g(\cdot)$. Consequently, $\gamma_i$ is related to the regression, $X_i^\top \beta$, through (3.2) and (3.3). The maximum likelihood (ML) estimator of $\beta$ can be obtained via iteratively reweighted least squares. For a detailed review of the literature on GLMs, see McCullagh and Nelder (1989) or McCulloch et al. (2001).

When $Z_1, \ldots, Z_n$ are associated with locations in space, the assumption of independence is doubtful. A way to extend the framework above, that takes into account spatial variability, is to replace $\gamma$ in (3.1) with a spatial Gaussian process, $\{ Y(s) : s \in D \}$, where $D$ is the
spatial domain of interest. The covariance between $Y(s)$ and $Y(u)$, for $s, u \in D$, is defined as:

$$C_Y(s, u) \equiv \text{cov}(Y(s), Y(u)).$$

Now consider spatial data $Z(s_1), \ldots, Z(s_n)$ from a GLM where $\{[Z(s_i)|Y(\cdot)] : i = 1, \ldots, n\}$ are mutually independent, and

$$g(E(Z(s_i)|Y(\cdot))) = Y(s_i); \ i = 1, \ldots, n,$$

(3.4)

where $g(\cdot)$ is the link function. The hierarchical modeling framework defined above yields a spatial version of the GLM framework; it was proposed by Diggle et al. (1998), who assumed a Gaussian model for $Y(\cdot)$ and a prior distribution on its parameters. See also Omre and Tjelmeland (1997) for an exposition of the same framework for solving complex problems in petroleum geostatistics.

Lindley and Smith (1972) introduced a Bayesian-linear-model framework, where conditional and prior distributions come from a multivariate Gaussian distribution. In the spatial context, Omre (1987) defined Bayesian kriging for the linear model; for further extensions, see Cressie (1993, Sec. 3.4.4). Besag et al. (1991) showed how a spatial model for counts in small areas could be decomposed hierarchically, where the hidden process $Y(\cdot)$ was used to model the spatial dependence. They assumed that the counts were (conditionally) Poisson distributed, and that the log means were a Gaussian spatial process, specifically a Gaussian Markov Random Field (MRF) known as the conditional autoregressive (CAR) model. However, a simultaneous autoregressive (SAR) model, or a geostatistical model could also be used. Indeed Diggle et al. (1998) employed spatial generalized linear mixed models (GLMMs) for spatially dependent non-Gaussian variables observed potentially anywhere in $D$, and they assumed a hidden geostatistical processes $Y(\cdot)$ with both fixed effects and random effects. Their hierarchical model was fully Bayesian and required a Markov Chain Monte Carlo (MCMC) algorithm to obtain the posterior distribution. In a
spatio-temporal context, Wikle et al. (1998) developed a fully Bayesian hierarchical-model formulation for modeling a dataset of monthly maximum temperatures.

In contrast, Heagerty and Lele (1998) developed a method for binary data where they used a composite-likelihood approach (e.g., Lindsay, 1988) to estimate the spatial hierarchical model parameters. Zhang (2002) gave a Monte Carlo version of the EM Gradient Algorithm to analyze non-Gaussian data, and Monestiez et al. (2006) developed a method called Poisson kriging for mapping the relative abundance of species.

Despite the popularity of the spatial models discussed above, these models might suffer from two major drawbacks: (1) there might be spatial confounding, and (2) there is often a computational bottleneck when the size of the dataset is large. Spatial confounding between the fixed and the random effects was pointed out in articles by Reich et al. (2006), Hodges and Reich (2010), and Paciorek (2010). Reich et al. (2006) and Hodges and Reich (2010) proposed a modeling approach that gets around the problem of spatial confounding by introducing random effects that are orthogonal to the column space of the matrix of covariates. We shall discuss this in more detail in Section 3.2.2.

The computational bottleneck arises due to the general computational cost of $O(n^3)$ to obtain the inverse of an $n \times n$ covariance matrix. It is often referred to as a “big $n$” problem. Many geophysical and environmental datasets are high-dimensional. When the data are Gaussian, reduced-rank-modeling approaches for the hidden Gaussian process $Y(\cdot)$ have been developed to deal with this computational challenge (e.g., Wikle et al., 2001; Cressie and Johannesson, 2006, 2008; Banerjee et al., 2008; Stein, 2008; Lopes et al., 2008). When the data are non-Gaussian, Lopes et al. (2011) take Diggle et al. (1998)’s GLMM approach, but with reduced-rank factor analysis models for $Y(\cdot)$ in place of the intrinsically stationary models used by Diggle et al. (1998). A number of spatial and spatio-temporal applications for very-large-to-massive datasets center around this reduced-rank representation of the hidden continuous Gaussian process (e.g., see the review in Wikle, 2010).
The reduced-rank methods discussed above are based on geostatistical models, where a continuously indexed Gaussian process \( \{ Y(s) : s \in D \} \) is used to specify the hidden process. In the case where \( D \equiv \{ s_1, \ldots, s_N \} \) is a spatial lattice of sites, a geostatistical model for \( Y \equiv (Y(s_1), \ldots, Y(s_N))^\top \) can still be used; such a model captures the spatial dependence through the covariance matrix, \( \Sigma_Y \equiv \text{cov}(Y) \).

A Gaussian MRF that is used to capture the spatial dependence in \( Y \), does so through the (typically sparse) precision matrix \( \Sigma_Y^{-1} \). A detailed discussion of this can be found in Rue and Held (2005, Chapter 5) and Cressie and Wikle (2011, Pages 185-186). Rue and Held (2005, Chapter 5) discuss a way to approximate a geostatistical model with a sparse CAR model, and this relationship has been used by Lindgren et al. (2011) and Simpson et al. (2012) to build hierarchical spatial models with Gaussian-MRF process models that allow fast computations. However, by necessity, they use only a small number of parameters, which could be problematic when modeling spatial dependence over large, continental-scale, heterogeneous regions. In a recent article, Hughes and Haran (2012) consider a Bayesian hierarchical model with a hidden Gaussian MRF and use a dimension-reduction approach to deal with spatial confounding and computational complexity that arise when analyzing a large spatial dataset. They parameterize the precision matrix using an underlying graph, \( G = (V,E) \), where edges represent spatial dependence, and they assume only a small number of parameters.

In this chapter, we assume there are small areas \( \{ A_i : i = 1, \ldots, N \} \) at locations \( D \equiv \{ s_1, \ldots, s_N \} \), respectively. The order of the small areas is immaterial, so we choose to order them such that \( A_1, \ldots, A_n \) have observations \( Z(s_1), \ldots, Z(s_n) \), respectively, associated with them, where \( n \leq N \). Define the observation vector (i.e., data) to be

\[
Z_O = (Z(s_1), \ldots, Z(s_n))^\top; \quad 1 \leq n \leq N.
\]

We propose a flexible class of spatial models for analyzing these (potentially) non-Gaussian lattice data. The models are hierarchical, where the data model comes from the exponential
family of distributions, and the process model is geostatistical and nonstationary (Section 3.2). These models are computationally efficient to implement, and we take an empirical hierarchical modeling (EHM) approach where any unknown parameters are estimated by ML estimation. Hence, the model is not fully Bayesian, but Bayes’ Theorem is used to obtain the all-important predictive distribution; for the special case where data are spatial counts, we have demonstrated its feasibility (Chapter 2; Sengupta and Cressie, 2013). For a more complete discussion of the EHM approach, see Cressie and Wikle (2011, Chapter 2).

Our spatial statistical analysis of the lattice data \( Z_0 \) is a combination of the GLMM framework of Diggle et al. (1998), the use of the Spatial Random Effects (SRE) model of Cressie and Johannesson (2006, 2008), developed for Gaussian data with a continuous spatial index, and a fast EM algorithm for estimating any unknown parameters. The SRE model is a geostatistical model that achieves dimension reduction by modeling the underlying spatial process as a linear combination of specified spatial basis functions on a spatially continuous domain; in what is to follow, we use it on a discrete spatial lattice. The dimension reduction is important for spatial best linear unbiased prediction (i.e., kriging), since it involves inverting the \( n \times n \) covariance matrix of \( Z_0 \). Using the SRE model, the matrix inversion is a relatively simple task, the model is well suited to change-of-support, and it avoids any stationarity assumptions for the covariance matrix. Unlike the model used in Lopes et al. (2011), the SRE model does not assume a diagonal covariance matrix for the spatial random effects. Instead, it captures spatial-statistical dependence using both the modeler-specified spatial basis functions and correlated random effects. Assuming the data are Gaussian, Katzfuss and Cressie (2009) gave an EM algorithm to obtain ML estimates for SRE-model parameters; and there is also a Bayesian-hierarchical-model (BHM) version that puts prior distributions on the parameters rather than estimating them (Kang and Cressie, 2011).
When the data are non-Gaussian, estimation of the parameters in a hierarchical statistical model is not as straightforward. In the EHM proposed in Section 3.2, we use the EM algorithm (Dempster et al., 1977) to obtain ML estimates of the parameters in the model. Since the expectations in the E-step of the algorithm are not available in closed form, we use a Laplace approximation to approximate the intractable integrals. Having obtained the estimates for the unknown parameters, we substitute them into the predictive distribution and use an MCMC algorithm to generate samples from it. Thus, our use of EHM for non-Gaussian data, with parameter estimates substituted into optimal predictors, is the direct analogue of kriging (used ubiquitously in geostatistical and environmental applications). We handle big spatial datasets by embedding the SRE model into our EHM.

The plan for this chapter is as follows. In Section 3.2, we describe a hierarchical model for non-Gaussian spatial data, whose data model comes from the exponential family and whose process model is based on a hidden SRE model. We also address the issue of spatial confounding in Section 3.2. In Section 3.3, we outline the statistical inference based on generating MCMC samples from the predictive distribution. Then, in Section 3.4, we describe the EM algorithm for obtaining ML estimates of the model parameters described in Section 3.2. In Section 3.5, we carry out a simulation experiment to assess the performance of our EHM approach. In Section 3.6, we use our EHM approach to analyze a large, spatial, remote sensing dataset of aerosol optical depth (AOD) from the MISR instrument on the Terra satellite. Discussion and conclusions follow in Section 3.7, and technical derivations are given in the Appendix.
3.2 Hierarchical Statistical Model

In this section, we give details of the hierarchical statistical model that we use to model non-Gaussian data. Specifically, the data model comes from the exponential family of distributions, and the process model is a (transformed) Gaussian spatial process. We consider lattice data obtained from among small areas \( \{A_i : i = 1, \ldots, N\} \), located at \( \{s_i : i = 1, \ldots, N\} \), respectively, although some locations have missing data. Thus, the spatial domain is the discrete spatial lattice \( D \equiv \{s_1, \ldots, s_N\} \). Without loss of generality, the locations where there are observations are denoted as \( \{s_1, \ldots, s_n\} \subset D \), where \( 1 \leq n \leq N \). Hence, the set of unobserved locations are \( \{s_i : i = n + 1, \ldots, N\} \), if \( n < N \).

3.2.1 Components of the Hierarchical Statistical Model

1. Conditional distribution of the data given the process (data model)

Recall \( Z_O = (Z(s_1), \ldots, Z(s_n))^\top \) denotes the vector of observations, and \( Y(s) \) denotes the hidden process at location \( s \in D \). Further, define the random process \( Y(\cdot) \equiv \{Y(s) : s \in D\} \). Then assume that \( [Z(s_i)|Y(\cdot)] = [Z(s_i)|Y(s_i)] \), and furthermore that it is a member of the exponential family (e.g., McCullagh and Nelder, 1989, Chapter 2). Conditional independence of the data given the process yields,

\[
[Z_O|Y(\cdot)] = \prod_{i=1}^n [Z(s_i)|Y(s_i)],
\]

where

\[
Z(s_i)|Y(s_i) \sim \text{ind. exponential family}\left(\mu_{Z|Y}(s_i), V(\mu_{Z|Y}(s_i))\right), \; i = 1, \ldots, n; (3.5)
\]

the conditional mean, \( \mu_{Z|Y}(s_i) \equiv E(Z(s_i)|Y(s_i)) \), depends on \( Y(s_i) \); and the variance
of the conditional distribution, $[Z(s_i) | Y(s_i)]$, is expressed as a function of the conditional mean through $V(\mu_{Z|Y}(s_i))$. The function $V(\cdot)$ denotes the mean-variance relationship for the exponential family. The distribution in (3.5) can be written as:

$$f_{Z|Y}(z(s_i) | Y(s_i)) = \exp \left\{ \frac{(z(s_i)\gamma(s_i) - b(\gamma(s_i)))}{\tau^2} - c(z(s_i), \tau) \right\},$$

(3.6)

where for convenience we have written the distribution in its canonical form. The quantities $\gamma(s_i)$ and $b(\gamma(s_i))$ depend on $Y(s_i)$ in a way determined by which member of the exponential family in (3.5) is chosen.

2. Link function

We proceed by modeling a transformation, $g(\cdot)$, of the mean $\mu_{Z|Y}(\cdot)$ as a sum of the two components:

$$g(\mu_{Z|Y}(s)) = t(s) + v(s); \ s \in D,$$

(3.7)

where $g(\mu_{Z|Y}(s))$ is the link function evaluated at the (conditional) mean, $t(s)$ is deterministic large-scale spatial variation (or the trend term), and $v(s)$ denotes random, mean-zero, small-scale spatial variation. If $g(\mu_{Z|Y}(\cdot)) \equiv \gamma(\cdot)$ in (3.6), then $g(\cdot)$ is the canonical link function, which plays an important role in the GLM (McCullagh and Nelder, 1989, Section 2.2.3). Examples of canonical links include the logit link for the Binomial distribution, the log link for the Poisson distribution, and the inverse link for the Gamma distribution. However, the canonical link is not the only choice. Some popular non-canonical links include the probit link for the Binomial distribution and the log link for the Gamma distribution.

3. Process model

The process $Y(\cdot)$ is defined as:

$$Y(\cdot) \equiv g(\mu_{Z|Y}(\cdot)).$$

(3.8)
Thus, $Y(\cdot)$ is related to the mean of the observed process through the link function. If we work with the canonical link, we have the special case $Y(\cdot) \equiv \gamma(\cdot)$.

From (3.7),

$$Y(\cdot) = t(\cdot) + v(\cdot), \quad (3.9)$$

where recall that $t(\cdot)$ is the deterministic spatial trend and $v(\cdot)$ is a random mean-zero spatial process.

4. Spatial trend

The trend, or large-scale spatial variation, is modeled as a linear combination of known covariates, $X(s) \equiv (X_1(s), \ldots, X_p(s))^\top$:

$$t(s) = C(s) + X(s)^\top \beta, \quad (3.10)$$

where $C(s)$ is a known offset term, and $\beta$ is a $p$-dimensional vector of unknown regression coefficients that need to be estimated. Recall that $Y = (Y(s_1), \ldots, Y(s_N))^\top$, and hence (3.9) becomes,

$$Y = C + X\beta + v, \quad (3.11)$$

where $X \equiv (X_O^\top, X_U^\top)^\top$, $X_O \equiv (X(s_1), \ldots, X(s_n))^\top$, $X_U \equiv (X(s_{n+1}), \ldots X(s_N))^\top$, $v \equiv (v_O^\top, v_U^\top)^\top$, $v_O \equiv (v(s_1), \ldots, v(s_n))^\top$, $v_U \equiv (v(s_{n+1}), \ldots, v(s_N))^\top$, and $C \equiv (C(s_1), \ldots, C(s_N))^\top$.

5. Spatial Random Effects (SRE) model for $v(\cdot)$

We use a geostatistical model for $v(\cdot)$, in contrast to the MRF used by Besag et al. (1991). The possibility of big data, $Z_O$, motivates us to propose the Spatial Random Effects (SRE) model:

$$v(\cdot) = S(\cdot)^\top \eta + \xi(\cdot), \quad (3.12)$$
where \( S(\cdot) \) is an \( r \)-dimensional vector of known spatial basis functions; \( \eta \) is a vector of random effects that is assumed to have a \( \text{Gau}(0, K) \) distribution; and \( \xi(\cdot) \) is a fine-scale-variation component that is assumed to be spatially independent with a \( \text{Gau}(0, v_\xi(\cdot)\sigma^2_\xi) \) distribution and \( v_\xi(\cdot) \) known. Other possible approaches to spatial prediction where datasets are very-large-to-massive are discussed in Section 3.1.

Recall that \(|D| = N \geq n\), where \( n \) may be very large; however, the random-effects vector \( \eta \) is only of dimension \( r \) (\( r \ll n \)). We do not assume any particular structure for the \( r \times r \) covariance matrix \( K \), nor do we necessarily try to parameterize it using just a few parameters. The spatial dependence in \( Y \) is captured using both \( K \) and the spatial basis functions \( S(\cdot) \). Dimension reduction is achieved by modeling the underlying \( N \)-dimensional spatial process as a linear combination of \( r \) fixed spatial basis functions over the entire spatial domain of interest. In Section 3.5, we show that this leads to substantial computational gain, which is especially significant when dealing with very large datasets. As well as computational speed-ups, the hierarchical model given by (3.6), (3.11), and (3.12) avoids making second-order stationarity assumptions, and it is well suited to change-of-support.

### 3.2.2 Spatial Confounding of Fixed and Random Effects

Our interest in this chapter lies primarily in inference on the hidden spatial process \( Y(\cdot) \) or, equivalently, in inference on \( \mu_{Z|Y}(\cdot) = g^{-1}(Y(\cdot)) \). That is, we wish to predict \( Y(\cdot) \) over the entire spatial domain \( D \), based on the data \( Z_O = (Z(s_1), \ldots, Z(s_n))^\top \). We first discuss confounding for the case where there is no dimension reduction, namely for a full-rank spatial generalized linear mixed model (SGLMM). The process model for a full-rank SGLMM is given by:

\[
g(\mu_{Z|Y}(\cdot)) = X(\cdot)^\top \beta + v(\cdot),
\]

(3.13)
where recall that $X(·)$ is a $p$-dimensional vector of known covariates, $\beta$ is a $p$-dimensional vector of regression coefficients, and $\nu(·)$ is the random effect. Define $g_O \equiv (g(\mu_{Z|Y}(s_1)), \ldots, g(\mu_{Z|Y}(s_n)))^\top$, and rewrite (3.13) in vector notation as,

$$g_O = X_O \beta + \nu_O = X_O \beta + I_n \nu_O,$$

(3.14)

where $X_O \equiv (X(s_1), \ldots, X(s_n))^\top$, and $\nu_O \equiv (\nu(s_1), \ldots, \nu(s_n))^\top$. The last equality emphasizes the matrix coefficients of the fixed and random effects. Reich et al. (2006) and Hodges and Reich (2010) used a reparameterization of (3.14) to show that such a SGLMM exhibits spatial confounding for fully Bayesian inference. Specifically, posterior inference for $\beta$ tends to be biased, and its posterior variance is inflated. This happens because a subspace of the column space of $I_n$ coincides with the column space of $X_O$ (Paciorek, 2010), as follows. Write

$$P \equiv X_O (X_O^\top X_O)^{-1} X_O^\top,$$

(3.15)

where $P$ is the projection matrix onto the span of $X_O$, and recall that $\text{rank}(P) = p$. Let $P^\perp \equiv I_n - P$ be the orthogonal complement of $P$, where $\text{rank}(P^\perp) = n - p$. Let $B$ and $B^\perp$ be the orthogonal bases of $\text{span}(X_O)$ and $\text{span}(X_O)^\perp$, respectively. Consequently, we can write (3.14) as

$$g_O = X_O \beta + B \delta_1 + B^\perp \delta_2,$$

(3.16)

where $\delta_1$ and $\delta_2$ are respectively the $n$-dimensional and $(n-p)$-dimensional spatial-random-effects vectors under this new parameterization. Notice that $B$ is collinear with $X_O$, which is the source of the spatial confounding. Reich et al. (2006) suggested deleting the component $B \delta_1$ from the model, by setting $\delta_1 = 0$, but Hughes and Haran (2012) pointed out that for a Gaussian MRF, this can result in negative spatial dependence. They proposed a model that alleviates spatial confounding, reduces the dimension of the random effects, and only allows for positive spatial dependence among the random effects.
Our approach to modeling is also based on reducing the dimension of the random effects. We use spatial basis functions to achieve dimension reduction but allow general dependence between the random effects. Recall the SRE model (3.12), which gives

\[ \mathbf{v}_O = \mathbf{S}_O \mathbf{\eta} + \xi_0, \]  

(3.17)

where \( \mathbf{S}_O \equiv (\mathbf{S}(s_1), \ldots, \mathbf{S}(s_n))^\top \) is typically sparse, and \( \xi_0 \equiv (\xi(s_1), \ldots, \xi(s_n))^\top \). The basis functions are introduced to capture the small-scale spatial variation in the model, and their optimal choice is an area of ongoing research (e.g., Bradley et al., 2011). As long as \( \mathbf{X}_O \) is not perfectly collinear with \( \mathbf{S}_O \), the large-scale variability that is captured by the fixed-effects component will not be fully explained by the random effects. In this chapter, we take an empirical-Bayesian approach, where we use the EM algorithm to estimate the unknown parameters (Section 3.4), and then we substitute in the estimates to obtain MCMC samples from the empirical predictive distribution (Section 3.3). That is, the EM estimate of \( \mathbf{\beta} \) (and \( \mathbf{K} \) and \( \sigma_\xi^2 \)) is held fixed in the MCMC, which is consistent with the treatment of large-scale variation in kriging when, in practice, the spatial trend (and the variogram) is unknown and has to be estimated (e.g., Cressie, 1993, Section 3.5). When \( \mathbf{\beta} \) is held fixed in the MCMC, (empirical) Bayesian inference on the random-effects term is no longer confounded. Consequently, an EHM approach mitigates spatial confounding in the SGLMM used in the process model.

### 3.3 Empirical-Bayesian Inference

Our main focus in this chapter is on prediction of \( Y(\cdot) \) or of \( \mu_{Z|Y}(\cdot) \). That is, after having observed \( \mathbf{Z}_O \) at locations \( \{s_1, \ldots, s_n\} \), we wish to make inference on \( \mathbf{Y} = (Y(s_1), \ldots, Y(s_N))^\top \) or some function of \( \mathbf{Y} \). The parameters \( \mathbf{\theta} \equiv \{\mathbf{\beta}, \mathbf{K}, \sigma_\xi^2\} \) are also of interest, but instead of putting a prior distribution on them, we estimate them using an EM algorithm (Section 3.4). Our hierarchical model becomes an empirical hierarchical model when we substitute
the estimated parameters \( \hat{\theta} \) in place of \( \theta \), into the predictive distribution, \([Y|Z_O, \theta]\). With a slight abuse of notation, we write this as \([Y|Z_O, \hat{\theta}]\) and refer to it as the empirical predictive distribution.

Recall that \( Z_O = (Z(s_1), \ldots, Z(s_n))^\top \), and write \( Y \equiv (Y_{O_1}^T, Y_{U_1}^T)^T \), where

\[
Y_O \equiv (Y(s_1), \ldots, Y(s_n))^\top, \quad \text{and} \quad Y_U \equiv (Y(s_{n+1}), \ldots, Y(s_N))^\top.
\]

Similarly, \( X \equiv (X_O^T, X_U^T)^T, S \equiv (S_O^T, S_U^T)^T, \) and \( \xi \equiv (\xi_O^T, \xi_U^T)^T \). Now,

\[
[Z_O|\eta, \xi_O, \theta] = [\xi_O|\xi_U, Z_O, \eta, \theta]
\]

\[
= \frac{[Z_O|\eta, \xi_O, \theta][\eta|\xi_U][\xi_O]\var[\xi_U]}{\int[Z_O|\eta, \xi_O][\eta|\xi_U][\xi_O]\var[\xi_U]d\xi_U}
\]

\[
= [\xi_U]\var[\xi_U].
\]

(3.18)

Thus, given \( \theta, \xi_U \) is conditionally independent of \( (Z_O, \eta, \xi_O) \), and hence for an unobserved site in \( \{s_i : i = n + 1, \ldots, N\} \), we have:

\[
E\left(Y(s_i)|Z_O, \beta, K, \sigma^2_\xi\right) = C(s_i) + X(s_i)\beta + S(s_i)\beta + E\left(\eta|Z_O, \beta, K, \sigma^2_\xi\right)
\]

\[
\var\left(Y(s_i)|Z_O, \beta, K, \sigma^2_\xi\right) = S(s_i)^T\var(\eta|Z_O, \beta, K, \sigma^2_\xi)S(s_i) + \sigma^2_\xi\var(\xi(s_i)).
\]

(3.19)

For a site \( s_i \in \{s_1, \ldots, s_n\} \), where an observation is available, we have

\[
E\left(Y(s_i)|Z_O, \beta, K, \sigma^2_\xi\right) = C(s_i) + X(s_i)\beta + S(s_i)\beta + E\left(\eta|Z_O, \beta, K, \sigma^2_\xi\right) + E\left(\xi(s_i)|Z_O, \beta, K, \sigma^2_\xi\right)
\]

\[
\var\left(Y(s_i)|Z_O, \beta, K, \sigma^2_\xi\right) = S(s_i)^T\var(\eta|Z_O, \beta, K, \sigma^2_\xi)S(s_i) + \var(\xi(s_i)|Z_O, \beta, K, \sigma^2_\xi)
\]

\[\quad + 2S(s_i)^T\cov(\eta, \xi(s_i)|Z_O, \beta, K, \sigma^2_\xi). \]

(3.20)

The goal here is to predict \( Y \) (or some function of \( Y \)), given the data. However, the predictive distribution, \([Y|Z_O, \theta]\), is not available in closed form, nor is \( \theta \) known. We shall use a combination of EM estimation of \( \theta \) to yield \( \hat{\theta}_{EM} \), and we shall use an MCMC algorithm (see, e.g., Robert and Casella, 2004) to yield samples from the predictive distribution, \([Y|Z_O, \theta]\), where \( \hat{\theta}_{EM} \) is substituted in for \( \theta \). In actuality, this is achieved by obtaining
samples from the predictive distribution, \([\eta, \xi_o|Z,o, \theta]\), and the distribution \([\xi_o|\sigma^2_{\xi}]\), where \(\theta = \hat{\theta}_{EM}\) and \(\sigma^2_{\xi} = \hat{\sigma}^2_{\xi,EM}\) are respectively substituted in. The EM algorithm to obtain \(\hat{\theta}_{EM}\) is presented in the next section, where it is seen that the E-step requires a Laplace approximation. The MCMC algorithm to obtain the predictive distribution is described in the Appendix.

### 3.4 EM Estimation of Parameters

In this section, we obtain the ML estimates of the parameters using the EM algorithm. The EM algorithm (Dempster et al., 1977) has been employed for estimation of parameters in the presence of missing data; for more details, see McLachlan and Krishnan (2008). For the hierarchical model described in Section 3.2, the random effects, \(\eta\), and the fine-scale variation, \(\xi_o\), are not known and can be treated as “data” that complete the likelihood. The EM algorithm involves iterating between an E (expectation)-step and an M (maximization)-step, and in our case the E-step is the most problematic. We resolve this problem by using Laplace approximations to evaluate the expectations required in the E-step.

Recall that

\[ g(\mu_z|\cdot) = Y(\cdot), \]

where \(g(\cdot)\) is the link function. We now rewrite \(\gamma(\cdot)\) and \(b(\gamma(\cdot))\) in (3.6) as functions of \(Y(\cdot)\). Define:

\[ \gamma(\cdot) \equiv h_1(Y(\cdot)) \]

\[ b(\gamma(\cdot)) \equiv h_2(Y(\cdot)). \]

(3.21)

Then, under this re-parameterization, the conditional density of \([Z(s)|Y(s)]\), for \(s \in \{s_1, \ldots, s_n\}\), is given by:

\[ f_{Z|Y}(z(s)) = \exp \left\{ (z(s)h_1(Y(s)) - h_2(Y(s))) / \tau^2 - c(z(s), \tau) \right\}. \]

(3.22)
Note that if the canonical link is considered, we have $\gamma(\cdot) = Y(\cdot)$, and hence

\[ h_1(Y(\cdot)) = Y(\cdot) \]
\[ h_2(Y(\cdot)) = b(Y(\cdot)). \]

The “complete data” log likelihood, $L_c$, for the unknown parameters is made up of the observations $Z_O$ and the unobserved $\eta$ and $\xi_O$. Then $L_c$ is simply the logarithm of the joint distribution of $Z_O$, $\eta$, and $\xi_O$, given the parameters $\theta = \{\beta, K, \sigma^2_\xi\}$. That is,

\[
L_c(\theta|Z_O, \eta, \xi_O) = \log [Z_O|\beta, \eta, \xi_O] + \log [\eta|K] + \log [\xi_O|\sigma^2_\xi]
= \text{const.} + \left\{ \sum_{i=1}^n Z(s_i)h_1(C(s_i) + X(s_i)^T\beta + S(s_i)^T\eta + \xi(s_i)) 
- \sum_{i=1}^n h_2(C(s_i) + X(s_i)^T\beta + S(s_i)^T\eta + \xi(s_i)) \right\} / \tau^2 
- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} (\eta\eta^T K^{-1}) 
- \frac{n}{2} \log \sigma^2_\xi - \frac{1}{2\sigma^2_\xi} \text{trace} (\xi_O\xi_O^T V^{-1}_{\xi_O}), \tag{3.24}
\]

where $[A|B]$ denotes the density function of $A$ given $B$, $V_{\xi,O} \equiv \text{diag}(v_\xi(s_1), \ldots, v_\xi(s_n))$, and “const.” denotes a generic constant that does not depend on $\theta$. The EM algorithm is based on $L_c$ and an iteration procedure that we now describe. Assume we have completed the $l$-th iteration of the EM algorithm; that is, we have an estimate $\theta^{[l]}$ of $\theta$. 

58
3.4.1 The E-step

At the \((l+1)\)-th iteration, the E-step is:

\[
Q(\theta, \theta^{[l]}) \equiv E\left( L_c(\theta|Z_O, \eta, \xi_O) | \theta^{[l]} \right)
\]

\[
= \text{const.} + \left\{ \sum_{i=1}^{n} Z(s_i) E\left( h_1(C(s_i) + X(s_i)^{\top} \beta + S(s_i)^{\top} \eta + \xi(s_i)) | Z_O, \theta^{[l]} \right) \\
- \sum_{i=1}^{n} E\left( h_2(C(s_i) + X(s_i)^{\top} \beta + S(s_i)^{\top} \eta + \xi(s_i)) | Z_O, \theta^{[l]} \right) \right\} / \tau^2 \\
- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace}\left( E\left( \eta \eta^{\top} | Z_O, \theta^{[l]} \right) K^{-1} \right) \\
- \frac{n}{2} \log \sigma^2_\xi - \frac{1}{2 \sigma^2_\xi} \text{trace}\left( E\left( \xi_O \xi_O^{\top} | Z_O, \theta^{[l]} \right) V^{-1}_{\xi,O} \right). \tag{3.25}
\]

The expectations involved in the E-step of the EM algorithm are with respect to the unobserved variables \(\eta\) and \(\xi_O\), and they are not available in closed form.

When the integrals in the E-step are problematic, one approach may be to implement a stochastic EM (SEM) algorithm (e.g., see McLachlan and Krishnan, 2008; Robert and Casella, 2004), where the expectations are evaluated using Monte Carlo integration. When datasets are large, this computation can be very slow, and hence the EM algorithm can be very slow to converge. In our approach, we derive Laplace approximations (LA) to approximate the expectations involved in (3.25), which are based on second-order Taylor-series expansions of the logarithm of the integrands around their respective modes.

To apply the LA, we need to obtain the mode, \(\widehat{\eta}^{[l]}, \widehat{\xi}_O^{[l]}\), of \(L_c\) considered as a function of \(\eta\) and \(\xi_O\). Sengupta and Cressie (2013) use a coordinate-wise ascent method for the Poisson GLM and canonical log link, which maximizes alternately with respect to \(\eta\), and then with respect to \(\xi_O\), until convergence. We do the same here for the general hierarchical model described in Section 3.2.

Following Sengupta and Cressie (2013), we use a second-order Taylor-series approximation to approximate the posterior distribution of \([\eta, \xi_O|Z_O, \theta^{[l]}]\) with a Gaussian distribution with mean and variance given by the posterior mode and the inverse of the negative
Hessian of the posterior evaluated at the mode; see the justification given in Kass and Steffey (1989). Details of our approximations can be found in the Appendix, where it is seen that the posterior distribution, $[\eta, \xi_0 | Z_0, \theta^{[l]}]$, is approximately a multivariate Gaussian density, with mean and variance given by

$$E \left( \begin{pmatrix} \eta \\ \xi_0 \end{pmatrix} \bigg| Z_0, \theta^{[l]} \right) = \begin{pmatrix} \hat{\eta}^{[l]} \\ \hat{\xi}_0^{[l]} \end{pmatrix},$$

(3.26)

and

$$\text{var} \left( \begin{pmatrix} \eta \\ \xi_0 \end{pmatrix} \bigg| Z_0, \theta^{[l]} \right) = \left\{ \begin{array}{ccc} -\frac{\partial^2}{\partial \eta \partial \eta} (L_c(\theta^{[l]} | Z_0, \eta, \xi_0)) & -\frac{\partial^2}{\partial \eta \partial \xi_0} (L_c(\theta^{[l]} | Z_0, \eta, \xi_0)) \\ -\frac{\partial^2}{\partial \xi_0 \partial \eta} (L_c(\theta^{[l]} | Z_0, \eta, \xi_0)) & -\frac{\partial^2}{\partial \xi_0 \partial \xi_0} (L_c(\theta^{[l]} | Z_0, \eta, \xi_0)) \end{array} \right\}^{-1} \mid _{\eta = \hat{\eta}^{[l]}, \xi_0 = \hat{\xi}_0^{[l]}}.$$  

(3.27)

respectively. To obtain $\text{var}(\eta | Z_0, \theta^{[l]})$ and $\text{var}(\xi_0 | Z_0, \theta^{[l]})$, we need to invert the matrix of partial derivatives shown just above. Let $A$ denote an $r \times r$ matrix and $B$ denote an $n \times n$ matrix. Further, let $U$ be any $r \times n$ matrix and $V$ be any $n \times r$ matrix. Then, a block-matrix-inversion formula (e.g., Duncan, 1944) is given by:

$$\begin{pmatrix} A & U \\ V & B \end{pmatrix}^{-1} = \begin{pmatrix} (A - UB^{-1}V)^{-1} & -(A - UB^{-1}V)^{-1}UB^{-1} \\ -(B - VA^{-1}U)^{-1}VA^{-1} & (B - VA^{-1}U)^{-1} \end{pmatrix}. \tag{3.28}$$

Now recall the Sherman-Morrison-Woodbury formula (e.g., Henderson and Searle, 1981):

$$(B - VA^{-1}U)^{-1} = B^{-1} + B^{-1}V(A - UB^{-1}V)^{-1}UB^{-1}.$$
We use this formula in the block-matrix-inversion formula (3.28) to obtain the following equivalent block-matrix-inversion formula, which we use to obtain the inverse in (3.27):

\[
\begin{pmatrix}
A & U \\
V & B
\end{pmatrix}^{-1}
= \begin{pmatrix}
(A - UB^{-1}V)^{-1} & -(A - UB^{-1}V)^{-1} UB^{-1} \\
-B^{-1}V(A - UB^{-1}V)^{-1} & B^{-1} + B^{-1}V(A - UB^{-1}V)^{-1} UB^{-1}
\end{pmatrix},
\]

(3.29)

where the lower off-diagonal block is obtained using the Sherman-Morrison-Woodbury formula as follows:

\[
(B - VA^{-1}U)^{-1} VA^{-1} = \{B^{-1}V(A - UB^{-1}V)^{-1} UB^{-1} + B^{-1}\} VA^{-1}
= B^{-1}V(A - UB^{-1}V)^{-1} \{UB^{-1}VA^{-1} + (A - UB^{-1}V)A^{-1}\}
= B^{-1}V(A - UB^{-1}V)^{-1}.
\]

(3.30)

Now, for generic variables \(u\) and \(v\), define

\[
J(u_0, v_0) = -\frac{\partial^2}{\partial u \partial v} \left. L_c(\theta^{[l]} | Z_O, u, v) \right|_{u=u_0, v=v_0}.
\]

We consider the different component matrices in the \((r + n) \times (r + n)\) matrix of partial derivatives given in (3.27). The matrix \(J(\hat{\xi}_O^{[l]}, \hat{\eta}_O^{[l]})\) is an \(n \times n\) diagonal matrix; its inversion is easy. The matrix \(J(\hat{\eta}_O^{[l]}, \hat{\xi}_O^{[l]})\) is of dimension \(r \times r\), where \(r << n\). The other two matrices, \(J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})\) and \(J(\hat{\eta}_O^{[l]}, \hat{\xi}_O^{[l]})\), have dimension \(n \times r\) and \(r \times n\), respectively. We can then use
formula (3.29) to invert the matrix in (3.27), which gives, approximately,

$$\text{var}(\eta|Z_O, \theta^{[l]}) = \left( J(\hat{\eta}^{[l]}, \hat{\eta}^{[l]}) - J(\hat{\eta}^{[l]}, \hat{\xi}_O^{[l]}) J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\hat{\xi}_O^{[l]}, \hat{\eta}^{[l]}) \right)^{-1}$$

$$\text{var}(\xi_O|Z_O, \theta^{[l]}) = J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} + J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\hat{\xi}_O^{[l]}, \hat{\eta}^{[l]})$$

$$\times \left( J(\hat{\eta}^{[l]}, \hat{\eta}^{[l]}) - J(\hat{\eta}^{[l]}, \hat{\xi}_O^{[l]}) J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\hat{\xi}_O^{[l]}, \hat{\eta}^{[l]}) \right)^{-1}$$

$$\times J(\hat{\eta}^{[l]}, \hat{\xi}_O^{[l]}) J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} \times J(\hat{\eta}^{[l]}, \hat{\xi}_O^{[l]}) J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1}.$$  \hspace{1cm} (3.31)

In the formulas given just above, all we need to invert is the \(n \times n\) diagonal matrix, \(J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})\), and some fixed-rank \(r \times r\) matrices. This makes the computations extremely efficient and allows us to obtain the expressions for \(E(\eta\eta^\top|Z_O, \theta^{[l]})\) and \(E(\xi_O\xi_O^\top|Z_O, \theta^{[l]})\) in (3.25) as follows:

$$E(\eta\eta^\top|Z_O, \theta^{[l]}) = \text{var}(\eta|Z_O, \theta^{[l]}) + E(\eta|Z_O, \theta^{[l]}) E(\eta|Z_O, \theta^{[l]})^\top$$

$$E(\xi_O\xi_O^\top|Z_O, \theta^{[l]}) = \text{var}(\xi_O|Z_O, \theta^{[l]}) + E(\xi_O|Z_O, \theta^{[l]}) E(\xi_O|Z_O, \theta^{[l]})^\top,$$  \hspace{1cm} (3.32)

where the terms on the right-hand side of (3.32) are evaluated approximately using (3.26) and (3.31).

The remaining terms in (3.25), for which we need an approximation, are

$$E \left( h_k(C(s) + X(s)^\top \beta + S(s)^\top \eta + \xi(s))|Z_O, \theta^{[l]} \right), s \in \{s_1, \ldots, s_n\}, k = 1, 2.$$

For the particular case of count data and the canonical link considered in Sengupta and Cressie (2013), analytical expressions were obtained based on the Gaussian approximation for \([\eta, \xi_O|Z_O, \theta^{[l]}]\) discussed above. In the general case considered here, a second-order
Taylor-series expansion is needed to evaluate the required expectations. From the Appendix, we see that, approximately,

\[
E\left(h_k(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i))|Z_O, \theta^{[l]}\right)
= h_k(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \eta^{[l]} + \hat{\xi}^{[l]}(s_i))
+ \frac{1}{2}h''_k(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \eta^{[l]} + \hat{\xi}^{[l]}(s_i)) \times \left(S(s_i)^\top \text{var}(\eta|Z_O, \theta^{[l]}) S(s_i) + 2S(s_i)^\top \text{cov}(\eta, \xi_O|Z_O, \theta^{[l]}) e(s_i) + e(s_i)^\top \text{var}(\xi_O|Z_O, \theta^{[l]}) e(s_i)\right),
\]

(3.33)

where \(k = 1, 2\), and \(e(s_i)\) is a vector of length \(n\) whose \(i\)-th element is 1 and all other entries are 0, for \(i = 1, \ldots, n\).

### 3.4.2 The M-step

Following the E-step, we perform the M-step, which involves maximizing (3.25) with respect to each of the parameters in \(\theta\). The maximization with respect to \(K\) and \(\sigma^2_\xi\) is obtained by differentiating (3.25) with respect to \(K\) and \(\sigma^2_\xi\), equating to zero, and solving the resulting equations. The solutions at the \((l + 1)\)-th iteration are:

\[
\sigma^{2[l+1]}_\xi = \frac{1}{n} \text{trace} \left( \left( E(\xi_O|Z_O, \theta^{[l]}) E(\xi_O|Z_O, \theta^{[l]})^\top + \text{var}(\xi_O|Z_O, \theta^{[l]}) \right) \text{var}^{-1}(\xi_O) \right)
\]

\[
K^{[l+1]} = E(\eta|Z_O, \theta^{[l]}) E(\eta|Z_O, \theta^{[l]})^\top + \text{var}(\eta|Z_O, \theta^{[l]})
\]

(3.34)

However, the maximization of (3.25) with respect to \(\beta\) is not available in closed form; we use a Newton-Raphson update at each M-step as follows:

\[
\beta^{[l+1]} = \beta^{[l]} - \left[ \frac{\partial}{\partial \beta} \mathbf{R}(\theta) \right]^{-1}_{\theta = \theta^{[l]}} \mathbf{R}(\theta^{[l]}).
\]

(3.35)

In (3.35), \(\mathbf{R}(\theta)\) denotes the score function obtained by taking the partial derivative of \(Q(\theta, \theta^{[l]})\), given by (3.25), with respect to \(\beta\), and \(\mathbf{R}(\theta^{[l]})\) is obtained by evaluating \(\mathbf{R}(\theta)\) at \(\theta^{[l]}\). The score function and the derivative required in (3.35) are evaluated in the Appendix.
3.4.3 Starting Values for the EM Algorithm

In order to implement the EM algorithm, we need to specify some starting values for the parameters. Although in the simulation study described in Section 3.5, we use the true parameter values as our starting values, for real data applications we do not have that luxury. In this section, we give a recommendation for initializing the EM algorithm. We shall use this method to obtain the starting values for the EM algorithm when analyzing the large remote sensing dataset in Section 3.6.

One may proceed by using the classical fixed-effects GLM estimate, $\hat{\beta}_{GLM}$, as the starting value for $\beta$; here, $\hat{\beta}_{GLM}$ is obtained using the iterated reweighted least squares algorithm (see McCulloch et al., 2001, Chapter 5).

Recall that the spatial trend is: $t(s_i) = C(s_i) + X(s_i)^\top \hat{\beta}$; consider the detrended process,

$$U(s_i) \equiv Y(s_i) - t(s_i),$$

which has mean zero and

$$\text{var}(U(s_i)) = S(s_i)^\top KS(s_i) + \sigma_\xi^2 V_{\xi}(s_i).$$

Writing $U_O \equiv (U(s_1), \ldots, U(s_n))^\top$, we obtain:

$$\text{cov}(U_O) \equiv \Sigma_{U,O} = S_O KS_O^\top + \sigma_\xi^2 V_{\xi,O},$$

where recall that $V_{\xi,O}$ is a known diagonal matrix.

To obtain method-of-moments estimates of $K$ and $\sigma_\xi^2$ that can be used as starting values, we replace $Y(s_i)$ with $g(Z(s_i) + c)$, where $c$ is some user-specified constant that is added to the data to ensure that the transformation is defined everywhere within the range of the data. For example, for Poisson data and the canonical log link, $\log(Z(s_i) + 0.5)$ avoids a singularity when $Z(s_i) = 0$.

Consequently, an approximation for $U(\cdot)$ is obtained as:

$$\hat{U}(s_i) \equiv g(Z(s_i) + c) - C(s_i) - X(s_i)^\top \hat{\beta}_{GLM}, \quad i = 1, \ldots, n.$$
Define $s_U^2 \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{U}(s_i)^2$, and choose

$$\hat{\Sigma}_{U;O} = s_U^2 I_n,$$  \hspace{1cm} (3.40)

simply to capture the total variation through the trace operator. We apportion approximately 90% of this to the smooth small-scale variation and 10% to the fine-scale variation (e.g., Katzfuss and Cressie, 2011). That is, we select our starting values for $K$ and $\sigma_\xi^2$ to satisfy

$$S_O K^{[0]} S_O^T \approx 0.9 \times \hat{\Sigma}_{U;O},$$

$$\sigma_\xi^{2[0]} = 0.1 \times \text{trace}(\hat{\Sigma}_{U;O}) / \text{trace}(V_{\xi;O}),$$ \hspace{1cm} (3.41)

as follows. Using (3.41), and the $Q$-$R$ decomposition, $S_O = Q_S R_S$, we obtain the starting value for $K$ as

$$K^{[0]} = R_S^{-1} Q_S^T \left( 0.9 \times \hat{\Sigma}_{U;O} \right) Q_S (R_S^T)^{-1}.$$ \hspace{1cm} (3.42)

Note that this approximate 90-10 apportionment of the total variability could be done differently, depending on the data’s smooth-scale variation relative to their fine-scale variation.

### 3.4.4 Properties of the Resulting EM Algorithm

Suppose the algorithm is initialized with parameter values $\theta^{[0]} \in \Theta$, where $\Theta$ is the parameter space. Then it can be seen from (3.34) that $\theta^{[l]} \in \Theta$, $l = 1, 2, \ldots$, which is a desirable property. For example, this means that if the starting value for $K$ is a covariance matrix, then all future EM updates will also be symmetric and at least non-negative definite. Likewise, if we choose $\sigma_\xi^{2[0]} > 0$, then it is guaranteed that the EM estimate satisfies $\hat{\sigma}_\xi^{2,EM} \geq 0$.

The most appealing feature of the resulting EM algorithm is computational. The E-step requires one optimization to obtain the posterior mode. Then the SRE-model assumption and the Sherman-Morrison-Woodbury formula make the LA computations extremely efficient. The computational complexity of the EM algorithm is linear in the sample size $n$ (see Section 3.5.4). This is a highly desirable property when dealing with big data. In Section
3.5, the computational performance of this algorithm and the variability of the estimates are assessed through simulation.

3.5 A Simulation Study

In this section, we investigate statistical properties of our EHM approach using a simulation experiment, where we simulate Poisson data over a regular spatial domain using the hierarchical model set-up as described in Section 3.2. Further, we demonstrate the computational gain that is achieved by using an EHM approach as opposed to a BHM approach.

3.5.1 Simulation Set-Up

We generate count data from a Poisson distribution whose mean is obtained by exponentiating an underlying spatial Gaussian process $Y(\cdot)$. We consider a regular spatial domain, $D = \{s_1, \ldots, s_N\}$, consisting of $N = 300 \times 300 = 90,000$ points on $\{-149.5, \ldots, -0.5, 0.5, \ldots, 149.5\}^2$. In this simulation, the hidden process $Y(\cdot)$ is made up of three additive components:

$$
Y(s) = X(s)^\top \beta + S(s)^\top \eta + \xi(s); \ s \in D,
$$

(3.43)

where the fine-scale heterogeneity term $v_\xi(\cdot) = 1$, and the offset term $C(\cdot) = 0$. The large-scale variation, or trend, is assumed to be,

$$
X(s)^\top \beta = \beta_0 + \beta_1 \times s_2,
$$

(3.44)

where $s = (s_1, s_2)^\top$ and $\beta = (\beta_0, \beta_1)^\top$.

Recall that the random-effects vector $\eta \sim \text{Gau}(0, K)$, and here $\xi(\cdot)$ is a process of independent and identically distributed (i.i.d.) $\text{Gau}(0, \sigma_\xi^2)$ random variables, independent
of \( \eta \). To specify the SRE model’s covariance matrix \( K \), we started with an exponential covariance function given by

\[
C(u, v) = c_0 \exp \left( -\frac{||u - v||}{a_0} \right),
\]

(3.45)

where \( c_0 \) is the sill and \( a_0 \) is the scale parameter. Here we specified \( c_0 = 1 \) (without loss of generality) and \( a_0 = 100 \) (to capture moderate-to-strong spatial dependence). Let \( \nu \equiv (\nu(s_1), \ldots, \nu(s_N))^\top \) be a mean-zero spatial Gaussian process defined over \( D \), whose covariance matrix is obtained from the exponential covariance model (3.45); that is, \( \nu \sim \text{Gau}(0, \Sigma) \). We calibrate \( K \) and \( \sigma^2_\xi \) using the procedure given in Kang and Cressie (2011).

For just the calibration, we considered only 9,000 regularly spaced locations (sampling every tenth location from the list of all 90,000 locations) that covers the entire spatial domain, rather than using all 90,000 locations.

First we calculated \( K^0 \) such that \( ||SKS^\top - \Sigma_\nu|| \) was minimized, where \( || \cdot || \) is the Frobenius norm (e.g., Cressie and Johannesson, 2008). Finally, to control the variability of \( Y \), we chose \( K = gK^0 \), where \( g \) was chosen to preserve the total variation. That is,

\[
\text{trace}(\Sigma_\nu)/N = 1 = \text{trace}(gSKS^\top + \sigma^2_\xi I_N)/N.
\]

(3.46)

For selecting the large-scale-variation parameter \( \beta \), we defined the variation of the “signal,” \( V_s \), as:

\[
V_s \equiv \frac{1}{N} \text{trace} \left( SKS^\top + \sigma^2_\xi I_N \right) + \frac{1}{N} \sum_{i=1}^{N} \left( \left( X(s_i)^\top \beta - \text{ave}_{s_i \in D} (X(s_i)^\top \beta) \right)^2 \right).
\]

The parameter \( \beta \) was selected such that \( V_s \) was approximately 2 (see Aldworth and Cressie, 1999, Section 3.2.4). In (3.44), we specified \( \beta = (2, 0.0125)^\top \), which gave \( V_s = 2.17 \). Additionally, we specified the fine-scale-variation proportion (FVP),

\[
FVP \equiv \frac{\text{trace} \left( \sigma^2_\xi I_N \right)}{\text{trace} \left( SKS^\top + \sigma^2_\xi I_N \right)},
\]

(3.47)
which from (3.46) is equal to $\sigma^2_\xi$. In our simulation, $FVP$ was held at 5%; hence, $\sigma^2_\xi = 0.05$. Using (3.46), we obtained $g = 1.22$.

We simulated $\eta$ and $\xi$ from the Gaussian process defined above and then, using (3.43), we obtained $Y$ over the entire domain $D$. Next, we used the inverse of the log link function,

$$
\mu_{Z|Y}(\cdot) = \exp(Y(\cdot)),
$$

(3.48)
to simulate a realization of the conditionally (conditional on $Y(\cdot)$) independent Poisson random variables, $Z_O$, for only $n$ locations ($n \leq N$); the $n$ locations $\{s_1, \ldots, s_n\}$ were randomly sampled without replacement from the $N = 90,000$ possible locations.

We will use this set-up to investigate the performance of the EM-based parameter estimates (Section 3.5.2), to compare the predictive performance of our EHM approach to that of an independent hierarchical GLM (Section 3.5.3), to compare the computational efficiency of our EHM approach to that of a competing Bayesian hierarchical modeling (BHM) approach (Section 3.5.4), and finally to do a sensitivity study of the EHM and the BHM approaches (Section 3.5.5). In Sections 3.5.2, 3.5.3, and 3.5.5, we hold $n$ fixed at 20,000. In Section 3.5.4, we vary $n$ and tabulate the computational efficiency as a function of $n$. We use the true parameter values as starting values for the EM algorithm, and for specifying hyperparameters for the BHM approach.

### 3.5.2 Assessment of the EM Estimates

In this section, we assess the performance of the EM estimates. Holding $n$ fixed at 20,000, we simulated 1600 vectors $Z_O^{[1]}, \ldots, Z_O^{[1600]}$ as specified in Section 3.5.1. For each of the simulated datasets, $Z_O^{[l]}$, where $l = 1, \ldots, 1600$, we used the EM algorithm described in Section 3.4 to estimate the unknown parameters.

We calculated the average and the empirical root mean squared error (RMSE) for the parameters $\beta = (\beta_0, \beta_1)^\top$ and $\sigma^2_\xi$; the results are summarized in Table 3.1, and they show very good agreement with the true values.
Table 3.1: True parameter values and the sample mean of the EM parameter estimates based on 1600 simulated datasets. Each dataset is of size \( n = 20,000 \). The empirical root mean squared errors (RMSEs) of the parameter estimates are also reported.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>Sample mean based on the 1600 simulated datasets</th>
<th>RMSEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 )</td>
<td>2.0</td>
<td>1.922</td>
<td>0.0954</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>0.0125</td>
<td>0.01262</td>
<td>0.0002</td>
</tr>
<tr>
<td>( \sigma^2_\varepsilon )</td>
<td>0.05</td>
<td>0.0507</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Now we consider the EM estimate of \( \mathbf{K} \). The elementwise mean of the EM estimates, \( \{ \hat{\mathbf{K}}_{EM}^{[l]} : l = 1, \ldots, 1600 \} \), is computed as:

\[
\text{ave}(\hat{\mathbf{K}}_{EM}) \equiv \frac{1}{1600} \sum_{l=1}^{1600} \hat{\mathbf{K}}_{EM}^{[l]}.
\] (3.49)

Figure 3.1 shows an image plot of the matrix \( \mathbf{H} \equiv \{ \text{ave}(\hat{\mathbf{K}}_{EM})\mathbf{K}^{-1}_T \} \), where \( \mathbf{K}_T \) is the true covariance matrix for \( \mathbf{\eta} \). We compare the matrix \( \mathbf{H} \) to the identity matrix. This gives a visual representation of how close the elementwise mean of the EM estimates of \( \mathbf{K} \) is to the true value \( \mathbf{K}_T \).

We also computed \( \text{trace}(\hat{\mathbf{K}}_{EM}^{[l]}\mathbf{K}_T^{-1}) \), for \( l = 1, \ldots, 1600 \). Now, had we observed \( \mathbf{\eta}^{[l]} \), the ML estimate of \( \mathbf{K} \) would be given by:

\[
\hat{\mathbf{K}}_{ML;\mathbf{\eta}}^{[l]} = \mathbf{\eta}^{[l]\top}\mathbf{\eta}^{[l]},
\] (3.50)

for which

\[
\text{trace}(\hat{\mathbf{K}}_{ML;\mathbf{\eta}}\mathbf{K}^{-1}_T) = \text{trace}(\mathbf{\eta}^{[l]\top}\mathbf{\eta}^{[l]\top}\mathbf{K}^{-1}_T) = \mathbf{\eta}^{[l]\top}\mathbf{K}^{-1}_T\mathbf{\eta}^{[l]} \sim \chi^2_r.
\] (3.51)
Figure 3.1: The top panel shows the identity matrix, and the bottom panel shows the matrix, $\text{ave}(\hat{K}_{EM})K^{-1}$, where $\text{ave}(\hat{K}_{EM})$ is the elementwise average of the EM estimates $\{\hat{K}_{EM}^{[l]} : l = 1, \ldots, 1600\}$. The common color bar is shown on the right.

So, we might expect the distribution of $\text{trace}(\hat{K}_{EM}^{[l]}K^{-1}_T)$ to look similar to a $\chi^2_r$ distribution. Recall that $r = 29$ in our case. Figure 3.2 shows a histogram of $\{\text{trace}(\hat{K}_{EM}^{[l]}K^{-1}_T) : l = 1, \ldots, 70\}$. 
upon which a $\chi^2_{29}$ density is superimposed. The sample mean and the sample variance of \( \text{trace}(\hat{K}_{EM}^{[l]} K_T^{-1}) \) are 29.4194 and 59.821, respectively, which we compare to $E(\chi^2_{29}) = 29$ and $\text{var}(\chi^2_{29}) = 58$.

Figure 3.2: Plot showing a histogram of \( \{\text{trace}(\hat{K}_{EM}^{[l]} K_T^{-1})\} : l = 1, \ldots, 1600 \). The chi-squared density with degrees of freedom equal to $r = 29$ is overlayed on the histogram.

Overall, the EM algorithm seems to perform well, despite the approximations involved...
in the E-step of the EM algorithm. Next, we shall investigate the predictive properties of our EHM approach.

### 3.5.3 Predictive Properties

In this section, we assess the predictive properties for the EHM approach described in Sections 3.2-3.4. For \( n = 20,000 \) and each of the simulated datasets \( \{Z_{O}^{[l]} : l = 1, \ldots, 100\} \), we implemented the EM algorithm to obtain \( \hat{\theta}_{EM}^{[l]} \equiv (\hat{\beta}_{EM}^{[l]}, \hat{K}_{EM}^{[l]}, \hat{\sigma}_{\xi,EM}^{2[l]}) \). Then, using the MCMC algorithm described in Section 3.3, we obtained samples from the empirical predictive distribution, \( [\eta, \xi_{O}|Z_{O}^{[l]}, \hat{\theta}_{EM}^{[l]}] \). For each of the 100 simulated datasets, we generated 25,000 MCMC samples, after discarding a burn-in sample of size 2,000. Recall that our EHM approach yields the predictor of \( Y(\cdot) \) based on \( Z_{O}^{[l]} \), as the mean of the resulting MCMC samples from \( [Y(\cdot)|Z_{O}^{[l]}, \hat{\theta}_{EM}^{[l]}] \). Here we compare this to one derived from a spatially independent GLM, namely

\[
Y(\cdot) = X(\cdot)^{\top} \beta + \xi(\cdot),
\]

where \( \xi(\cdot) \sim \text{i.i.d. Gau}(0, \sigma_{\xi}^{2}) \). To estimate the parameters of the resulting EHM, we used a variant of the EM algorithm described in Section 3.4 with \( \eta = 0 \), that is, no spatial random-effects component. The MCMC algorithm from which the empirical predictive distribution is obtained is, likewise, a special case of that given in Section 3.3, with \( \eta = 0 \).

In what follows, we denote the 20,000 locations with data as \( D_{O} \) and the complementary set of 70,000 locations without data as \( D_{U} \). Recall that \( D_{O} \) is obtained by random sampling from \( D \) without replacement; for the 100 datasets, the set of locations \( D_{O} \) (and hence \( D_{U} \)) are held fixed.

Using obvious notation where “S” denotes “spatial” and “I” denotes “independent,” define \( \hat{Y}_{SEHM}(\cdot) \) and \( \hat{Y}_{IEHM}(\cdot) \) to be the means of the predictive distributions, \( [Y(\cdot)|Z_{O}^{[l]}, \hat{\theta}_{SEM}^{[l]}] \), and \( [Y(\cdot)|Z_{O}^{[l]}, \hat{\theta}_{IEM}^{[l]}] \), respectively. Importantly, \( Z_{O}^{[1]}, \ldots, Z_{O}^{[100]} \) are simulated according to the set-up given in Section 3.5.1.
Consider the ratio of the mean squared prediction errors,
\[
e(s) \equiv \frac{1}{100} \sum_{l=1}^{100} (\hat{Y}_{SEHM}^{[l]}(s) - Y^{[l]}(s))^2 \div \frac{1}{100} \sum_{l=1}^{100} (\hat{Y}_{IEHM}^{[l]}(s) - Y^{[l]}(s))^2; \quad s \in D,
\]
where \(Y^{[l]}(\cdot)\) is the true process (Section 3.5.1). From (3.53), we made kernel-density plots showing the distribution of \(e(\cdot)\) for locations in \(D_O\) and for those in \(D_U\), separately. These plots are shown in Figure 3.3, from which we see that SEHM has higher relative efficiency for locations in \(D_U\) than for those in \(D_O\). Clearly, for locations without data (i.e., \(D_U\)), SEHM borrows strength efficiently from nearby observations, and hence it performs much better than IEHM in terms of smaller mean squared prediction error.

Figure 3.3: Kernel-density plots showing the distribution of the ratio of the SEHM mean squared prediction error divided by the IEHM mean squared prediction error, for locations with data (solid line) and for locations without data (dashed line).
What about absolute performance of our EHM approach for the locations with and without data? We made kernel-density plots that compare the distribution of mean squared prediction errors,

$$\frac{1}{100} \sum_{l=1}^{100} (\hat{Y}_{SEHM}^{[l]}(s) - Y^{[l]}(s))^2,$$

for locations $s$ in $D_O$ to those in $D_U$ (see Figure 3.4). Generally, Figure 3.4 shows that mean squared prediction errors are smaller in $D_O$ than in $D_U$. Since a datum $Z(s)$ at location $s$ is very informative about the hidden value $Y(s)$ at $s$, this is to be expected.

Figure 3.4: Kernel-density plots comparing the SEHM mean squared prediction errors obtained for locations with data (solid line) and for locations without data (dashed line)
3.5.4 Computational Time: EHM versus BHM

In this section, we illustrate the computational gain achieved by using an EHM approach as opposed to using a comparable BHM approach. In what follows, whenever we say EHM (BHM), we mean a spatial EHM (BHM).

Recall that part of our EHM approach involves estimating the unknown parameters using an EM algorithm, followed by an MCMC algorithm that generates samples from the empirical predictive distribution, \([\eta, \xi_0|Z_0, \hat{\theta}_{EM}]\), where \(\hat{\theta}_{EM} \equiv (\hat{\beta}_{EM}, \hat{K}_{EM}, \hat{\sigma}^2_{\xi,EM})\). In a BHM approach, priors are put on \(\beta, K, \) and \(\sigma^2_{\xi}\), and an MCMC algorithm is used to generate samples from the posterior distribution, \([\eta, \xi, \theta|Z_0]\). Priors are assigned following Kang and Cressie (2011), the details of which are given in the Appendix.

Generally, the MCMC algorithm mixes more slowly for the BHM than for the EHM. Hence, we need to calibrate the MCMC sample sizes properly before we can compare the computational times. Suppose the number of MCMC samples from the empirical predictive distribution, \([\eta, \xi_0|Z_0, \hat{\theta}_{EM}]\), is \(L_{EHM}\), and suppose that \(L_{BHM}\) is the number of MCMC samples obtained from the posterior distribution, \([\eta, \xi, \theta|Z_0]\).

To calibrate the MCMC sample sizes, there are different diagnostic measures that could be used (e.g., Robert and Casella, 2004, Chapter 12). In this chapter, we shall use the diagnostics proposed by Gelman and Rubin (1992) and Brooks and Gelman (1998). The Gelman-Rubin statistic, or potential scale reduction factor (PSRF), is based on the idea of generating several MCMC chains, each of length \(L\), and then comparing the variability based on these individual chains to that based on the combined chain. If PSRF is close to 1, we can conclude that each set of \(L\) simulated values is close to the target distribution; if PSRF is large, \(L\) may be too small. Brooks and Gelman (1998) proposed the multivariate potential scale reduction factor (MPSRF), which is a multivariate extension of the PSRF, that can be used for assessing convergence of several parameters simultaneously.

For fixed data size \(n\), we generated five MCMC chains, each of length \(L\). Then we
found the values of $L_{EHM}$ and $L_{BHM}$ that have comparable MSPRFs close to 1. We started with $n = 5,000$ and found that for the elements of $\xi$, mixing was achieved quickly for both EHM and BHM. However, mixing for $\eta$ is comparatively slow for EHM and even slower for BHM, so we calibrated the MCMC sample sizes based on the convergence diagnostics for $\eta$. Figure 3.5 shows plots of the MPSRF and the maximum of elementwise PSRFs as functions of $L$. From Figure 3.5, we selected $L_{EHM} = 15,000$, and $L_{BHM} = 40,000$, which resulted in MPSRFs of 1.08 for EHM and 1.07 for BHM.

![Figure 3.5](image.png)

Figure 3.5: Plots showing the Gelman-Rubin-Brooks statistics, for EHM (left panel) and for BHM (right panel), as a function of the number of MCMC samples. The solid line corresponds to the MPSRF for $\eta$; the dashed line corresponds to the maximum of the elementwise PSRFs for $\eta$. Here, the number of observations is $n = 5,000$.

Next we investigated how the MPSRF and the PSRFs changed as $n$ changed. By
holding $L_{EHM} = 15,000$ and $L_{BHM} = 40,000$, and varying $n$, Table 3.2 shows that the Gelman-Rubin and Gelman-Brooks statistics are robust to change in the sample size, $n$. Consequently, we compare the computational times for EHM and BHM, for all $n$, using $L_{EHM} = 15,000$ and $L_{BHM} = 40,000$.

Table 3.2: Gelman-Rubin-Brooks statistics for varying sample sizes ($n$). The number of MCMC samples generated are $L=15,000$ for EHM, and $L=40,000$ for BHM. MPSRF is the multivariate potential scale reduction factor, and max(PSRF) is the maximum of the elementwise potential scale reduction factors (PSRFs).

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>EHM (L=15,000)</th>
<th>BHM (L=40,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\eta$</td>
<td>$\xi_O$</td>
</tr>
<tr>
<td></td>
<td>MPSRF</td>
<td>max(PSRF)</td>
</tr>
<tr>
<td>5,000</td>
<td>1.08</td>
<td>1.028</td>
</tr>
<tr>
<td>10,000</td>
<td>1.07</td>
<td>1.028</td>
</tr>
<tr>
<td>15,000</td>
<td>1.09</td>
<td>1.027</td>
</tr>
<tr>
<td>20,000</td>
<td>1.07</td>
<td>1.027</td>
</tr>
</tbody>
</table>

The simulation experiment was performed on a dual quad core 2.8 GHz 2x Xeon X5560 processor, with 96 Gbytes of memory. The computational times for the EHM and BHM are given in Table 3.3, and a visual illustration is given in Figure 3.6. From Table 3.3 and Figure 3.6, we see that EHM is on the order of 7-10 times faster than BHM. Nevertheless, in both cases, the computational time increases approximately linearly in $n$, which is due to the dimension reduction afforded by the SRE model given by (3.12).
Table 3.3: Computational time for varying sample sizes \((n)\). For EHM, the EM algorithm was used to estimate the parameters, and then an MCMC algorithm was used to generate \(L_{EHM} = 15,000\) samples from the empirical predictive distribution, \(\eta, \xi | Z_O, \hat{\theta}_{EM}\). For BHM, an MCMC algorithm was used to generate \(L_{BHM} = 40,000\) samples from the posterior distribution, \(\eta, \xi, \theta | Z_O\).

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>EHM (L=15,000)</th>
<th>BHM (L=40,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM Estimation</td>
<td>MCMC Implementation</td>
</tr>
<tr>
<td>5,000</td>
<td>0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>20,000</td>
<td>0.02</td>
<td>0.62</td>
</tr>
<tr>
<td>35,000</td>
<td>0.02</td>
<td>1.01</td>
</tr>
<tr>
<td>50,000</td>
<td>0.04</td>
<td>1.45</td>
</tr>
</tbody>
</table>

3.5.5 Sensitivity Study Comparing EHM to BHM

In this section, we describe a sensitivity study to demonstrate the precision and accuracy of the EHM predictions, when compared to BHM predictions (e.g., Kang et al., 2009).

Using the methods described in Section 3.5.1, we simulated \(Z_O\). with \(n = 20,000\). From those simulated data, we obtained samples from the empirical predictive distribution \(Y(\cdot) | Z_O, \hat{\theta}_{EM}\), which is our EHM approach, and from the posterior distribution \(Y(\cdot) | Z_O\), which is the BHM approach. First, we did a visual assessment of the predictions, \(\hat{Y}_{SEHM}(\cdot) \equiv E(Y(\cdot) | Z_O, \hat{\theta}_{EM})\) and \(\hat{Y}_{SBHM}(\cdot) \equiv E(Y(\cdot) | Z_O)\), which are shown in Figure 3.7, along with the data, \(\{Z(s_i), i = 1, \ldots, n = 20,000\}\), and the true underlying process, \(Y(\cdot)\).
Figure 3.6: Plot showing the computational time for EHM (dark gray) and BHM (light gray). Plotted for EHM is the computational time for EM estimation plus the time it took for the MCMC algorithm to generate 15,000 samples from the empirical predictive distribution, \( \eta, \xi|\hat{\theta}_{EM} \), after discarding 2,000 burn-in samples. Plotted for BHM is the computational time it took for the Gelman-Rubin-Brooks-statistics-calibrated MCMC algorithm to generate 40,000 samples from the posterior distribution, \( \eta, \xi, \theta|Z_O \), after discarding 2,000 burn-in samples.

Figure 3.7 gives the visual impression that there is no difference in the predictions obtained using EHM and BHM, which is confirmed with a kernel-density plot showing the distribution of the difference, \( \hat{Y}_{SEHM}(\cdot) - \hat{Y}_{SBHM}(\cdot) \); see Figure 3.8.

Next we computed the ratio,

\[
    r(\cdot) = \frac{(\text{var}(Y(\cdot)|Z_O))^{1/2}}{(\text{var}(Y(\cdot)|Z_O, \hat{\theta}_{EM})^{1/2}}. \tag{3.54}
\]

The distribution of the ratio of the standard deviations is shown in Figure 3.9, separately for locations in \( D_O \) (where data are observed) and \( D_U \) (where data are not observed). From
Figure 3.7: Plot showing the observed data (top-left panel), the true simulated process, \( Y(\cdot) \) (top-right panel), the mean of the empirical predictive distribution, \( \tilde{Y}_{SEHM}(\cdot) \equiv E(Y(\cdot) | Z_O, \hat{\theta}_{EM}) \) (bottom-left panel), and the mean of the posterior distribution, \( \tilde{Y}_{SBHM}(\cdot) \equiv E(Y(\cdot) | Z_O) \) (bottom-right panel).

Figure 3.9, we see that the ratio is mostly larger than 1; it is always larger than 1 in \( D_U \), and it is larger than 1 for 87.5\% of locations in \( D_O \). Thus, our EHM approach tends to yield credible intervals for \( Y(\cdot) \) that are narrower than those obtained from a BHM approach. From this experiment, we see that for \( s \in D_O \), EHM-based credible intervals tend to be
narrower by a factor of 0.8, while for \( s \in D_U \), the factor is 0.75. These results are consistent with other spatial studies (e.g., Kang et al., 2009).

### 3.6 Analysis of Aerosol Optical Depth from the MISR Instrument

In this section, we use the methodology presented in the previous sections to analyze a large, spatial, remotely sensed dataset on aerosol optical depth (AOD) retrieved by the
Figure 3.9: Kernel-density plots showing the distribution of the ratio, \( \frac{\text{var}(Y(s)|Z_O)^{1/2}}{\text{var}(Y(s)|Z_O, \hat{\theta}_{EM})^{1/2}} \), separately for locations \( s \) in \( D_O \) (locations with data) and for locations \( s \) in \( D_U \) (locations without data).

Multi-angle Imaging SpectroRadiometer (MISR) instrument on NASA’s Terra satellite. An analysis of this dataset was done by Shi and Cressie (2007); they used a log transformation of the data and then analyzed log(AOD) using a Gaussian model, however they did not give results back on the original AOD scale. The key feature of our current analysis is to model AOD directly, using a hierarchical spatial statistical model with a Gamma data model. The methodology we have developed in the previous sections allows us to obtain optimal spatial predictions, posterior standard errors, and 95% prediction intervals on the original AOD scale.
3.6.1 Background to the Dataset

The Terra satellite was launched by NASA on December 18, 1999, as part of the Earth Observing System (EOS). The MISR instrument is one of the key instruments on board that collects global aerosol information, and it covers the entire globe in 16 days. Level-2 AOD data are collected at a 17.6 km × 17.6 km spatial resolution; they can then be converted to level-3 AOD data at a lower spatial resolution (of 0.5° × 0.5°) by averaging all the level-2 observations that fall within the level-3 pixels. (Here, and in what follows, when we say level-3 pixel, we mean a pixel at the spatial resolution of 0.5° × 0.5°.) Due to orbit geometry, clouds, or non-retrievals, data can be missing in many regions. We use our model to predict the true AOD at level-3 pixels, both where there are data and where there are no data.

We analyze here a spatial dataset of lattice data consisting of level-3 AOD values observed between August 2-9, 2001, within a study region \( D \) bounded by longitudes \(-125°\) and \(+3°\) and latitudes \(-20°\) and \(+44°\). This is the same dataset that was analyzed in Shi and Cressie (2007), and was part of a spatio-temporal dataset in Kang et al. (2010), although exclusively on the log(AOD) scale. The region covers North and South America, the western part of the Sahara desert in Africa, the Iberian Peninsula in Europe, and parts of the Atlantic and Pacific Oceans (see Figure 3.10). There are \( N \equiv 128 \times 256 = 32,768 \) level-3 pixels in \( D \). The \( n = 21,759 \) data in \( D_O \) are shown in the left panel of Figure 3.11, where white pixels define the no-data locations (i.e., \( D_U \)); a histogram for the data is shown on the right panel of Figure 3.11.

3.6.2 Hierarchical Spatial Statistical Modeling of AOD

In this section, we do some initial data analysis of the AOD dataset by fitting a weighted generalized linear model that does not contain spatial dependence (McCullagh and Nelder, 1989), followed by a full spatial analysis of the dataset. Recall from Section 3.6.1 that
Figure 3.10: Plot showing the study region corresponding to the AOD dataset. The light-gray regions correspond to land surfaces, and the dark-gray regions correspond to water bodies.

Figure 3.11: AOD data in $D$ (left panel) and histogram showing their distribution (right panel).

$Z(s_i)$ is the average AOD obtained by averaging all the level-2 observations that fall within the level-3 pixel located at $s_i$. Let $m(s_i)$ denote the number of level-2 observations that are
averaged to obtain $Z(s_i)$, for $i = 1, \ldots, n$. We denote the level-2 observations within the level-3 pixel located at $s_i$ as $Z_j(s_i)$, $j = 1, \ldots, m(s_i)$, so that $Z(s_i) \equiv \sum_{j=1}^{m(s_i)} Z_j(s_i)/m(s_i)$.

Conditional on an underlying spatial process $Y(\cdot)$, we assume independent Gamma distributions for the level-2 observations. That is, conditional on $Y(\cdot)$, $Z_j(s)$ and $Z_k(u)$ are independent, except when $s = u$ and $j = k$. We further assume local homogeneity within a level-3 pixel; that is,

$$Z_j(s_i) | Y(s_i) \sim \text{i.i.d Gamma}(\nu, \mu_{Z|Y}(s_i) / \nu); \ j = 1, \ldots, m(s_i),$$

where $\mu_{Z|Y}(s_i) \equiv E(Z_1(s_i) | Y(\cdot)) = E(Z_1(s_i) | Y(s_i))$ is the mean of the conditional distribution $[Z_j(s_i) | Y(s_i)]; \nu > 0$ is the shape parameter of the Gamma distribution; and, consequently, $\mu_{Z|Y}(s_i) / \nu (\nu > 0)$ is its scale parameter for the level-3 pixel at $s_i$. That is, the density function for $Z_j(s_i) | Y(s_i)$, under this parameterization, is

$$f_{Z|Y}(z_j(s_i) | Y(s_i)) = \frac{(z_j(s_i) \nu)^\nu \exp(-z_j(s_i) \nu / \mu_{Z|Y}(s_i))}{z_j(s_i) \Gamma(\nu) \mu_{Z|Y}(s_i)^\nu}; \ z_j(s_i) \geq 0.$$  

(3.56)

From (3.55), and (3.56), we obtain the conditional distribution of the level-3 datum at $s_i$ as,

$$Z(s_i) | Y(s_i) \sim \text{Gamma}(m(s_i) \nu, \mu_{Z|Y}(s_i) / (m(s_i) \nu)); \ i = 1, \ldots, n,$$

where the distributions are assumed independent. Thus, we see that the between-pixel heterogeneity shows up in the scale and the shape parameters, although $E(Z(s_i) | Y(s_i))$ is $\mu_{Z|Y}(s_i)$ and does not depend on $m(s_i)$. This yields the loglikelihood,

$$L(\beta, \nu) = \sum_{i=1}^{n} \left\{ (m(s_i) \nu - 1) \log(Z(s_i)) + m(s_i) \nu \log(m(s_i) \nu) - \frac{Z(s_i)m(s_i) \nu}{\exp(X(s_i)^T \beta)} 

- \log \Gamma(m(s_i) \nu) - m(s_i) \nu (X(s_i)^T \beta) \right\}. $$

(3.58)

The canonical link for the Gamma distribution is the reciprocal link, namely, $\gamma(s) = (\mu_{Z|Y}(s))^{-1}$, which leads to constraints on the conditional mean that are not easy to model.
Guided by previous analyses of AOD where log data were analyzed, we use a log link. That is,

\[ \log(\mu_{Z_i|Y_i}) = X(s_i)^\top \beta; \ i = 1, \ldots, N, \quad (3.59) \]

where \( X(s_i) \) is a \( p \)-dimensional vector of known covariates, and there is no offset term \( C(\cdot) \) in this model. After some initial exploratory data analysis considering the covariates used in Kang et al. (2010), we selected the covariates in (3.59) to be the indicator functions for each of the Americas, Africa (the Sahara desert), the south-western tip of Europe (Iberian Peninsular), and oceans; and we also included latitude as a covariate.

From the weighted GLM (WGLM) given by (3.56) and (3.59), we obtained the ML estimate, \( \hat{\beta}_{WGLM} \), of \( \beta \), which does not depend on \( \nu \). Note that the estimate \( \hat{\beta}_{WGLM} \) is different than what one would obtain using a standard R or Matlab package, since they do not consider the different \( \{m(s_i) : i = 1, \ldots, n\} \) that appear in the loglikelihood given by (3.58). The maximum likelihood estimate of \( \nu \) is obtained by maximizing \( L(\hat{\beta}_{WGLM}, \nu) \) with respect to \( \nu \) and results in \( \hat{\nu} = 0.3637 \). These ML estimates are used in the hierarchical statistical analysis that follows.

As an aside, if we transform the data as, \( \tilde{Z}(s_i) \equiv m(s_i)Z(s_i); \ i = 1, \ldots, n \), then the distribution of \( \tilde{Z}(s_i) \) is Gamma\((m(s_i)\nu, \mu_{Z|Y_i}(s_i))\), where \( \mu_{Z|Y_i}(s_i) \equiv m(s_i)\mu_{Z|Y_i}(s_i) \). Hence, the log link is:

\[ \log(\mu_{Z|Y_i}(s_i)) = \log(m(s_i)\mu_{Z|Y_i}(s_i)) = \log(m(s_i)) + X(s_i)^\top \beta, \quad (3.60) \]

where there is now an offset term \( C(s_i) = \log(m(s_i)) \). Since the information content of \( \{\tilde{Z}(s_i)\} \) and \( \{Z(s_i)\} \) are the same, the ML estimates of \( \beta \) and \( \nu \) are unchanged.

Our hierarchical spatial statistical model consists of a data model and a process model; recall that unknown parameters are estimated. The data model is given by (3.57), where \( \nu = 0.3637 \), obtained above. We assume the log link,

\[ Y(\cdot) = \log(\mu_{Z|Y}(\cdot)), \quad (3.61) \]
and the process model is:

\[ Y(s_i) = X(s_i)\top \beta + S(s_i)\top \eta + \xi(s_i); \quad i = 1, \ldots, N, \] (3.62)

where recall that \( N = 128 \times 256 = 32,768 \) level-3 pixels, and \( X(\cdot) \) is a 5-dimensional vector made up of the same covariates used in the initial data analysis. In (3.62), the \( r \)-dimensional vector of random effects, \( \eta \), is assumed to have a \( \text{Gau}(0, K) \) distribution, where the covariance matrix \( K \) is fixed but unknown and will be estimated. We use mutiresolutional W-wavelet basis functions for \( S(\cdot) \); see Kang et al. (2010) and Kang and Cressie (2011). That is, we choose all 32 W-wavelets from the first resolution, and 62 W-wavelets from the second resolution, resulting in \( r = 32 + 62 = 94 \). The \( N \times r \) matrix \( S \) of basis functions is further rescaled by dividing each column of \( S \) by the standard deviation of the elements of the corresponding column. Finally, the component \( \xi(\cdot) \) denotes the fine-scale-variation parameter, and we model it using a \( \text{Gau}(0, \sigma^2_\xi) \) distribution.

### 3.6.3 Parameter Estimation and Optimal Spatial Mapping of AOD

We use the EM algorithm (Section 3.4) to estimate the parameters \( \theta = \{ \beta, K, \sigma^2_\xi \} \). To implement the EM algorithm, we obtain the starting values using the methods discussed in Section 3.4.3, with \( \hat{\beta}_{WGLM} \) used as the starting value for \( \beta \). The EM estimates, \( \hat{\theta}_{EM} \equiv \{ \hat{\beta}_{EM}, \hat{K}_{EM}, \hat{\sigma}^2_{\xi,EM} \} \), are then substituted into an MCMC algorithm (Appendix B.3) to obtain samples from the (empirical) predictive distribution, \( [\eta, \xi_O|Z_O, \hat{\theta}_{EM}] \). We generated 20,000 MCMC samples, after discarding 2,000 samples as burn-in. These MCMC samples, together with MCMC samples from \( [\xi_U|\hat{\sigma}^2_{\xi,EM}] \), give us the entire empirical predictive distribution, \( [Y|Z_O, \hat{\theta}_{EM}] \), or any desired transformation or summary of it. For example, we can obtain \( [\mu_Z|Y|Z_O, \hat{\theta}_{EM}] \), where \( \mu_Z|Y \equiv (\mu_{Z|Y}(s_1), \ldots, \mu_{Z|Y}(s_N))\top \) and \( \mu_{Z|Y}(\cdot) = \exp(Y(\cdot)) \), whose moments and quantiles are immediately computable.

Using the MCMC samples, we first computed the predictive mean and the predictive standard deviation of the process \( Y(\cdot) \); see the left panels of Figure 3.12. These panels are
comparable to the optimal predictions in Shi and Cressie (2007), Kang et al. (2010), and Kang and Cressie (2011), which are on the log scale. The predictive mean of \( Y(\cdot) \) shows that high aerosol particles are emitted from the Sahara desert and make their way across the Atlantic Ocean to North America via mid-latitude trade winds. The map of predictive standard deviations reflects the satellite tracks and regions of missing data, as it should.

The additive nature of the model for \( Y(\cdot) \) allows us to map and interpret different sources of variability separately. Specifically, the right panels of Figure 3.12 show image plots for the trend component \( \mathbf{X}(\cdot) \top \hat{\mathbf{\beta}}_{EM} \), for the predictive mean of the small-scale variation component \( \mathbf{S}(\cdot) \top \eta \), and for the predictive mean of the fine-scale-variation component \( \xi(\cdot) \).

Adding them together, we obtain the predictive mean of \( Y(\cdot) \) shown in the middle-left panel of Figure 3.12.

Recall that the datum \( Z(s_i) \) was obtained by averaging \( m(s_i) \) level-2 observations observed in the level-3 pixel located at \( s_i; i = 1, \ldots, n \). We incorporated that heterogeneity in our hierarchical model through (3.57), and to assess its impact we made side-by-side boxplots showing how the predictive standard deviation of \( Y(\cdot) \) varies for different values of \( m(s_i) \); see Figure 3.13. As expected, the predictive standard deviation of \( Y(s_i) \) decreases as \( m(s_i) \) increases, reflecting the importance of the data model in this spatial statistical analysis.

Our goal in this analysis is to make inference on the original AOD scale. Here we obtained maps of the mean, the standard deviation, the 2.5 percentile, and the 97.5 percentile of each of the \( N \) elements of \( \mathbf{\mu}_{Z|Y} \) in the (empirical) predictive distribution \( [\mathbf{\mu}_{Z|Y}|Z_O, \hat{\mathbf{\theta}}_{EM}] \); see Figure 3.14. Notice that the map of the predictive standard deviation shows a mean-variance relationship, which is the consequence of the Lognormal process model for \( \mu_{Z|Y}(\cdot) \).

The maps showing the 2.5 percentile and the 97.5 percentile give the upper bound and lower bound, respectively, of pixelwise 95% credible intervals. All panels in Figure 3.14 show maps on the original AOD scale, where they are most interpretable scientifically.
3.7 Discussion and Conclusions

In this chapter, we have developed a hierarchical spatial statistical model where the data model belongs to the exponential family of distributions. The process model is spatially dependent and is based on a hidden SRE model for the underlying latent random process. This allows for nonstationarity and dimension reduction, which is advantageous when analyzing big, spatially heterogeneous datasets. The model parameters are assumed fixed but unknown and are estimated.

The model proposed in this chapter is spatial-only. However, it could be extended to a hierarchical spatio-temporal model in an obvious way. We could use the same data model and a process model where the reduced-dimensional basis function coefficients evolve over time (e.g., Wikle et al., 2001; Cressie et al., 2010). There remain the problems of estimation of spatio-temporal-model parameters and optimal filtering, smoothing, and forecasting from the empirical predictive distribution.

Because of our empirical hierarchical modeling (EHM) approach, we are able to avoid spatial confounding between fixed-effects and random-effects terms in the process model. We have developed an EM algorithm to estimate the unknown parameters; since the expectations required in the E-step of the EM algorithm are not available in closed form, we developed a Laplace approximation for them.

Based on a simulation experiment, we assessed the performance of EM estimation of the parameters, and then we investigated the predictive properties of our EHM approach. We further used the simulation set-up to compare the performance of our EHM approach to that of a comparable BHM approach, both in terms of computational efficiency (EHM is 7-10 times faster) and in terms of width of credible intervals (EHM is 75-80% more liberal).

Finally, we used our methodology to analyze a big, spatially heterogeneous dataset on AOD. Based on a Gamma data model and a Lognormal process model, and after properly
accounting for sources of heterogeneity, we obtained a map of optimal spatial predictions of AOD on the original scale, along with maps quantifying the uncertainty of that prediction.

In conclusion, we have presented an empirical hierarchical modeling (EHM) approach that captures non-linear, non-Gaussian, spatial variability, and is well suited to the analysis of big data.
Figure 3.12: Maps to the left show the log(AOD) (top-left panel), the mean (middle-left panel) and standard deviation (bottom-left panel) of the predictive distribution of $Y(\cdot)$, namely $\mathbb{E}[Y(\cdot)|Z_O, \hat{\theta}_{EM}]$. Maps to the right show the predictive mean of the different components of variability in $Y(\cdot)$, namely, the components due to trend, $X(\cdot)^T \hat{\beta}_{EM}$ (top-right panel), the random-effects component, $E[S(\cdot)^T \eta|Z_O, \hat{\theta}_{EM}]$ (middle-right panel), and the fine-scale-variation component, $E[\xi(\cdot)|Z_O, \hat{\theta}_{EM}]$ (bottom-right panel). The middle-left panel which is a map of the mean of the predictive distribution of $Y(\cdot)$, namely $\mathbb{E}[Y(\cdot)|Z_O, \hat{\theta}_{EM}]$, is the sum of the three panels shown on the right.
Figure 3.13: Boxplots show the variability of the predictive standard deviation of $Y(s_i)$ for values of $m(s_i) = 1, 2, \ldots, 21$. 
Figure 3.14: Maps show the predictive mean (top-left panel), the pixelwise predictive standard deviation (top-right panel), the pixelwise predictive 2.5 percentile (bottom-left panel), and the pixelwise predictive 97.5 percentile (bottom-right panel) obtained from the empirical predictive-distribution, $[\mu_{Z|Y}(\cdot)|Z_0, \hat{\theta}_EM]$. The plots of the predictive mean and the predictive percentiles have the same color scale, where any value greater than 1 has been assigned the highest color-value.
CHAPTER 4

PREDICTIVE INFERENCE FOR MODIS CLOUD DATA:
HIERARCHICAL SPATIAL STATISTICAL MODEL AND
MASSIVE DATASETS

4.1 Introduction

Clouds are generally characterized by higher reflectances and lower temperatures than the Earth’s surface (Ackerman et al., 2010). They play an important role in climate research and must be accurately described in order to properly assess climatic processes and climate change. The accuracy of remote sensing retrievals can be affected by cloud contamination of the atmospheric column. Retrievals of certain atmospheric quantities require clear skies (e.g., land surface properties or atmospheric aerosols etc.). Retrieval of cloud properties are done for pixels that are cloudy. So it is important to know if a pixel is cloudy or if it is clear. The Moderate Resolution Imaging Spectroradiometer (MODIS) offers the opportunity for multispectral approaches to cloud detection.

Our interest is in the MODIS instrument that is on board the Terra satellite, which was launched by NASA in December 1999. The Level-2 MODIS cloud mask product (Platnick et al., 2003) is produced for pixel arrays at a spatial resolution of 1 km×1 km. Each MODIS cloud mask product file covers data collected over a five-minute time interval, which is called a granule. Each granule contains data on approximately $2.75 \times 10^6$ pixels of 1 km×1 km resolution. In this chapter, a granule of Terra MODIS data will be used
to illustrate our statistical-modeling approach. The granule corresponds to June 29, 2006, 12:45 UTC. A true-color composite image of the granule is shown in Figure 4.1.

Figure 4.1: An example of a granule image obtained by the MODIS instrument on board NASA's Terra satellite (June 29, 2006, 12:45 UTC). The inset shows the location of the granule on a world map. (Source: modis-atmos.gsfc.nasa.gov)

The MODIS instrument collects data on spectral radiances that are then processed using the MODIS cloud detection algorithm (e.g., Platnick et al., 2003; Ackerman et al., 1998, 2010) to produce a Level-2 cloud-mask classification (MOD06 product). The MODIS cloud-detection algorithm is based on a number of spectral tests; different tests can have different results for a particular pixel. The results from all tests are then combined to determine an overall “confidence” value, $Q(s)$, for a pixel located at $s$ to be clear (i.e., cloud
free). If $Q(s) = 1$, it signifies high confidence for the pixel to be clear, and if $Q(s) = 0$, it signifies high confidence for the pixel to be cloudy. Then, “clear-sky restoral” tests are performed that check for unambiguous clear-sky signals. We denote MODIS cloud mask product as $Q(\cdot)$, and we review the algorithm that results in $Q(\cdot)$ in Section 4.2.

The MODIS cloud-mask algorithm does not consider any potential spatial-statistical dependence in $Q(\cdot)$. Further, there is no quantification of the uncertainty associated with estimates of clear-sky/cloudy probabilities. In this chapter, we propose a hierarchical spatial statistical model for analyzing MODIS cloud data. Our goal is to produce optimal spatial-prediction maps for the underlying clear-sky/cloudy process, along with measures of prediction uncertainties.

In the rest of the chapter, we concentrate on the particular granule discussed above (see Figure 4.1). Our data are the MODIS cloud mask product, $Q(\cdot)$, which is available at 1 km×1 km pixels. Henceforth, each of these pixels will be called a “basic areal unit” (BAU). The number of BAUs in the granule shown in Figure 4.1 is $N = 2,748,620$. In general, we assume that we have data for $n$ BAUs, where $n \leq N$. For the particular granule that we consider in this chapter, we have $n = N$ (i.e., there are no BAUs without data). A full-rank spatial-statistical modeling approach for the granule would require specifying an $N \times N$ covariance matrix for the underlying spatial (transformed) clear-sky-probability process. To produce optimal spatial-statistical predictions, we would need to invert the $N \times N$ covariance matrix, something that is not computationally feasible for $N$ larger than several thousand.

The computational bottleneck that arises due to the computational cost of inverting the $N \times N$ covariance matrix referred to above, is often referred to as a “big $N$” problem. When the data appear to be Gaussian, reduced-rank-modeling approaches have been developed to deal with this computational challenge (e.g., Wikle and Cressie, 1999; Wikle et al., 2001; Cressie and Johannesson, 2006, 2008; Banerjee et al., 2008; Stein, 2008; Lopes et al.,
For data appearing to come from the exponential family of distributions, Lopes et al. (2011) took the hierarchical generalized linear mixed modeling framework proposed by Diggle et al. (1998), and they introduced a new class of spatio-temporal models using a latent factor-analysis structure; their fully Bayesian model allows for dimension reduction and hence fast computations. A number of spatial and spatio-temporal applications for very-large-to-massive datasets center around these reduced-rank representations of a hidden continuous Gaussian process (e.g., see the review in Wikle, 2010).

To solve the “big N” problem that arises in our application, we shall use the reduced-rank modeling approach developed by Cressie and Johannesson (2006, 2008), although our data are bimodal and constrained to $[0,1]$. Our modeling approach is a combination of the GLMM framework of Diggle et al. (1998) and use of the Spatial Random Effects (SRE) model of Cressie and Johannesson (2006, 2008), although they developed it for Gaussian data with a continuous spatial index. We take an empirical hierarchical modeling (EHM) approach and, unlike a Bayesian hierarchical modeling (BHM) approach, we treat the model’s parameters as fixed but unknown. We estimate these parameters using an EM algorithm (e.g., Dempster et al., 1977). With this approach, computation of optimal spatial predictions are feasible, and no prior specification of parameters is needed. For a more complete discussion of the EHM and BHM approaches, see Cressie and Wikle (2011, Chapter 2).

model was developed for big, spatial, discrete and continuous data, that includes the SRE model as a component of the process model.

With regard to applications, the SRE model and the methodologies associated with it have been successful in analyzing massive remote sensing datasets (e.g., Cressie and Johannesson, 2006, 2008; Shi and Cressie, 2007; Kang et al., 2010; Katzfuss and Cressie, 2011). The models were Gaussian and additive. In Chapter 3, the SRE model was used in a hierarchical framework to analyze highly skewed, non-negative, remotely sensed Aerosol Optical Depth data, where the models were non-Gaussian and non-additive.

The plan for the rest of the chapter is as follows: In Section 4.2 we describe the MODIS cloud mask product. Details of the hierarchical spatial statistical model will be presented in Section 4.3. In Section 4.4, we outline the statistical inference based on generating MCMC samples from the predictive distribution. In Section 4.5 we describe the EM algorithm for obtaining the ML estimates of the model parameters described in Section 4.3. Then, in Section 4.6, we analyze the granule of MODIS cloud data shown in Figure 4.1 using the methodologies developed in this chapter. In Section 4.7 we discuss the change-of-support relationship. Discussion and conclusions follow in Section 4.8, and the technical derivations are given in the Appendix.

4.2 MODIS Cloud-Mask Product

The MODIS cloud-mask algorithm (e.g., Ackerman et al., 1998, 2010) identifies different conceptual domains according to surface type and solar illumination. Once a pixel is assigned to a domain, a battery of spectral tests is applied, where each test attempts to detect the presence of cloud in the pixel, by returning a confidence level for the pixel to be clear, ranging from 1 (high-confidence clear), to 0 (low-confidence clear, that is, high-confidence cloudy). Individual spectral tests are based on an upper and lower bound (see below).

Tests capable of detecting similar conditions are grouped together. Denote the total
number of groups by \( N_G \), and assume that there are \( m_i \) spectral tests within the \( i \)-th group, \( i = 1, \ldots, N_G \). For the \( j \)-th test within the \( i \)-th group, if the observed light radiance falls below (above) the lower (respectively, upper) bound, then the clear-sky confidence level, \( F_{ij} \), is 0 (respectively, 1). A pictorial illustration of such a spectral test in the MODIS cloud-mask algorithm is given in Figure 4.2: If the observed radiance of the reflected light falls in the “high-confidence cloudy” region (i.e., below the lower bound), then \( F_{ij} \) is assigned a value 0 (i.e., cloudy), and if the observed light radiance falls in the “high-confidence clear” region (i.e., above the upper bound), then \( F_{ij} \) is assigned a value 1 (i.e., clear). When the observed value falls in the “intermediate” region (i.e., between the lower and upper bounds), \( F_{ij} \) is assigned a value between 0 and 1 using linear interpolation; see Figure 4.2.

Figure 4.2: A pictorial illustration of a MODIS cloud mask spectral test, which is based on an upper and lower bound.
For a given pixel, a minimum confidence level is determined for the \( i \)-th group as:

\[
G_i = \min \{ F_{ij} : j = 1, \ldots, m_i \}, \text{ for } i = 1, \ldots, N_G.
\]  

(4.1)

The overall clear-sky confidence value, \( Q \), for that pixel, is then defined as:

\[
Q \equiv \left( \prod_{i=1}^{N_G} G_i \right)^{1/N_G}.
\]  

(4.2)

This approach is clear-sky conservative in the sense that if one of the tests concludes that the pixel is cloudy (i.e., if one \( F_{ij} = 0 \)), then the overall clear-sky confidence value is 0.

The Q-values obtained above (called the “initial” Q-values) are then subject to “clear-sky restoral tests” (e.g., Ackerman et al., 2010; Heidinger, 2010). These tests check for unambiguous clear-sky signals. For example, spectral tests might indicate that a pixel located at \( s \) is cloudy (i.e., \( Q(s) = 0 \)); but, if all its neighboring pixels are clear, then the pixel is restored as “probably clear” by setting \( Q(s) = 0.96 \). Here “cloudy,” “probably cloudy,” “clear,” and “probably clear” are the possible classifications for a pixel, and they are based on thresholding the Q-values (e.g., Platnick et al., 2003). There are other clear-sky restoral tests for different land surfaces, coastal waters, and sun glint. Final Q-values are obtained after applying the clear-sky restoral tests; see Figure 4.3 for the difference between initial and final Q-values obtained for the granule shown in Figure 4.1. Noticeable in Figure 4.1 is a strip of sun glint reflecting off the ocean, which appears in the top panel of Figure 4.3 (initial Q-values) but not in the bottom panel (final Q-values). Thus, restoral tests are important, since there are parts of the world and situations where the cloud-mask algorithm tends to over-predict clouds (e.g., regions with sun glint).

In this chapter, we analyze the spatial dataset of final Q-values (denoted by \( Q(\cdot) \)), which we refer to as the MODIS cloud data. In the next section, we develop a hierarchical spatial statistical modeling framework that will be used for predicting the underlying clear-sky-probability process, given the data. Our approach also allows us to quantify the uncertainty associated with our predictions. These models allow for spatial change-of-support, which
Figure 4.3: Initial Q-values (top panel) and final Q-values (bottom panel) corresponding to the granule shown in Figure 4.1.

enables us to predict cloud-fraction at any desired resolution coarser than 1 km × 1 km (see Section 4.7).
4.3 Hierarchical Model for the Cloud Data

In this section, we propose an empirical hierarchical model for final Q-values obtained from the MODIS cloud mask product. We index the set of BAUs with data as $D_O \equiv \{s_1, \ldots, s_n\}$, and the complimentary set of BAUs without data as $D_U \equiv \{s_{n+1}, \ldots, s_N\}$. Hence, our data are $\{Q(s_i) : i = 1, \ldots n\}$ (see Section 4.2), where recall that for the granule shown in Figure 4.1, we have $n = N = 2,748,620$. We introduce a hidden variable $W(s_i)$, that denotes the state of a pixel, namely 0 or 1 (cloudy or clear), located at $s_i; i = 1, \ldots, N$. Then we assume a hidden spatial process $Y(\cdot)$ that controls the probability of $W(\cdot)$ being 1, where both $W(\cdot)$ and $Y(\cdot)$ are defined over the entire spatial domain, $D \equiv D_O \cup D_U$.

Our hierarchical spatial statistical model consists of a data model and a two-stage process model. Next we give details of the data model and the process models.

4.3.1 Data Model

We model the pixel-level conditional probabilities, $\{[Q(s_i)|W(s_i), \text{parameters}] : i = 1, \ldots, n\}$ using a “0-1 inflated” Beta distribution. Conditional on $W(s_i) = 0$, $Q(s_i)$ will be modeled using a zero-inflated Beta distribution; and, conditional on $W(s_i) = 1$, $Q(s_i)$ will be modeled using a one-inflated Beta distribution. The 0-1 inflation deals with those $\{Q(s_i)\}$ that are exactly 0 or 1. Then, our data model is as follows: For $i = 1, \ldots, n$, independently,

$$[Q(s_i)|W(s_i) = 0, P_0, \alpha_0] = \left\{P_0 I(Q(s_i) = 0) + (1 - P_0) f_1, \alpha_0(Q(s_i))\right\};$$

(4.3)

and, for $i = 1, \ldots, n$, independently,

$$[Q(s_i)|W(s_i) = 1, P_1, \alpha_1] = \left\{P_1 I(Q(s_i) = 1) + (1 - P_1) f_1, \alpha_1(Q(s_i))\right\}. \quad (4.4)$$

In (4.3) and (4.4),

$$f_{a,b}(Q(s_i)) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} Q(s_i)^{a-1}(1-Q(s_i))^{b-1}I(0 < Q(s_i) < 1), \quad (4.5)$$

102
which is the density of a Beta\((a, b)\) random variable, where \(a > 0\) and \(b > 0\). The parameters \(P_0, \alpha_0, P_1,\) and \(\alpha_1\) in the data model are unknown and need to be estimated.

Note that, while specifying the data model we have fixed one of the parameters of the Beta distribution to reduce the number of parameters. Keeping one parameter for the Beta distribution is enough to capture the mean of the the conditional distribution given in the data models; the portion of the total variability that is left unaccounted for, will be captured using the process models. For the particular granule that we consider in this chapter (see Figure 4.1), after doing some initial exploratory analysis, we selected the parameterization given in (4.3) and (4.4). When analyzing a different granule, depending on the dataset, one might use \(f_{\alpha_0,1}(\cdot)\) and/or \(f_{\alpha_1,1}(\cdot)\) in (4.3) and/or (4.4), respectively.

### 4.3.2 Process Model

Next, we specify the two-stage process model. Recall that the process variables \(W(\cdot)\) and \(Y(\cdot)\) are defined over the entire spatial domain, \(D\).

“Process model 1” represents the distribution of \(\{W(s_i) : i = 1, \ldots, N\}\), conditional on the hidden spatial process \(Y(\cdot)\). We assume a set of independent Bernoulli random variables for process model 1: For \(i = 1, \ldots, N\), independently,

\[
W(s_i) | Y(\cdot) \sim \text{Bernoulli} \left( \frac{\exp(Y(s_i))}{1 + \exp(Y(s_i))} \right),
\]

where recall that \(W(s_i) = 1\) (respectively, 0) means that the pixel located at \(s_i\) is clear (respectively, cloudy). Then \(Y(\cdot)\) is the logit transform of the clear-sky-probability process, \(p(\cdot)\), and conversely,

\[
p(\cdot) = \left( \frac{\exp(Y(\cdot))}{1 + \exp(Y(\cdot))} \right).
\]

At the second stage of the process model (“process model 2”), we use the reduced-rank
Spatial Random Effects (SRE) model (e.g., Cressie and Johannesson, 2006, 2008) to define the smooth spatial dependence in \( Y(\cdot) \). Process model 2 is:

\[
Y(s_i) = X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i); \quad i = 1, \ldots, N,
\]

(4.8)

where \( X(s_i) \) is a vector of known covariates; \( \beta \) denotes the set of unknown regression coefficients; \( S(\cdot) \equiv (S_1(\cdot), \ldots, S_r(\cdot))^\top \) is a vector of \( r \) (not necessarily orthogonal) spatial basis functions, where \( r << N \) is fixed; \( \eta \) is an \( r \)-dimensional vector of spatial random effects assumed to have a \( \text{Gau}(0, K) \) distribution, where the covariance matrix \( K \) is unknown and needs to be estimated; \( \xi(\cdot) \) is a fine-scale-variation process modeled as independent \( \text{Gau}(0, \sigma_\xi^2) \) random variables, where \( \sigma_\xi^2 \) is unknown and needs to be estimated.

### 4.4 Empirical Bayesian Inference

As mentioned before, our main focus in this paper is on prediction of the clear-sky probability process, 

\[
p(\cdot) = \frac{\exp(Y(\cdot))}{1 + \exp(Y(\cdot))}.
\]

Consequently, the probability of cloudiness, \( q(\cdot) \), can be obtained as:

\[
q(\cdot) = 1 - p(\cdot).
\]

(4.9)

After having obtained data \( Q(s_1), \ldots, Q(s_n) \), we are interested in inference on the hidden process \( Y(\cdot) \) (or equivalently \( p(\cdot) \)) at spatial locations \( \{s_i : i = 1, \ldots, N\} \). The parameters \( \theta \equiv \{P_0, P_1, \alpha_0, \alpha_1, \beta, K, \sigma_\xi^2\} \) are unknown and considered to be nuisance parameters for the purpose of prediction. Our hierarchical model becomes empirical when we estimate the parameters (Section 4.5) and substitute them into the predictive distribution of \( Y(\cdot) \).
Let us define, \( Q_O \equiv (Q(s_1), \ldots, Q(s_n))^\top \), \( W \equiv (W(s_1), \ldots, W(s_N))^\top \), \( Y \equiv (Y(s_1), \ldots, Y(s_N))^\top \), and \( \xi \equiv (\xi(s_1), \ldots, \xi(s_N))^\top \). The goal is to predict the process variables, \((W, Y)\), given the data, and the parameters. The predictive distribution is given by:

\[
[W, Y|Q_O, \theta] \propto [Q_O|W, Y, \theta][W|Y, \theta][Y|\theta]. \tag{4.10}
\]

However, the predictive distribution is not available in closed form, nor is \( \theta \) known. We shall use a combination of EM estimation of \( \theta \) to yield \( \hat{\theta}_{EM} \), and shall use a Markov Chain Monte Carlo (MCMC) algorithm (see, e.g., Robert and Casella, 2004) to yield samples from the predictive distribution, \([W, Y|Q_O, \theta]\), where \( \hat{\theta}_{EM} \) is substituted in for \( \theta \). In actuality, this is achieved by obtaining samples from the predictive distribution, \([W, \eta, \xi|Q_O, \hat{\theta}_{EM}]\).

The inference described in this this chapter is based on a hierarchical statistical model, but we do not put prior distributions on the parameters \( \theta \). Following the classical geostatistical paradigm for continuous symmetric variables (e.g., Cressie, 1993, Chapter 3), we estimate the parameters \( \theta \) using the EM algorithm (see Section 4.5). We then substitute the parameter estimates into the MCMC algorithm to obtain samples from the empirical predictive distribution, \([W, \eta, \xi|Q_O, \hat{\theta}_{EM}]\), where \( \hat{\theta}_{EM} \) is the EM estimate of \( \theta \). The MCMC algorithm to obtain the predictive distribution is described in the Appendix.

### 4.5 EM Estimation of Parameters

The EM algorithm (Dempster et al., 1977) has been employed for estimation of parameters in the presence of missing data; for more details on the EM algorithm, see McLachlan and Krishnan (2008). For the hierarchical model described in Section 4.3, and for locations in \( D_O \), the process \( W(\cdot) \), the random effects \( \eta \) and the fine-scale-variation term \( \xi(\cdot) \) are not observed and can be treated as missing data. The EM algorithm involves iterating between an E (expectation) step and an M (maximization) step, and in our case the E-step is the
most problematic. We resolve this problem by using Laplace approximations to evaluate the expectations.

The “complete data” log likelihood, $L_c$, for the unknown parameters is made up of the observations $Q_0$ and the “missing data” $W_0$, $\eta$, and $\xi_0$, where $W_0 \equiv (W(s_1), \ldots, W(s_n))^\top$, and $\xi_0 \equiv (\xi(s_1), \ldots, \xi(s_n))^\top$. Then $L_c$ is simply the logarithm of the joint distribution of $Q_0$, $W_0$, $\eta$, and $\xi_0$, given the parameters $\theta = \{P_0, \alpha_0, P_1, \alpha_1, \beta, K, \sigma_\xi^2\}$. That is,

\[
L_c(\theta | Q_0, \eta, \xi_0) = \log [Q_0 | W_0, \beta, \eta, \xi_0] + \log [W_0 | \beta, \eta, \xi_0] + \log [\eta | K] + \log [\xi_0 | \sigma_\xi^2] \\
= \text{const.} + \sum_{i:Q(s_i)=1} W(s_i) \log P_1 + \sum_{i:Q(s_i)=0} (1 - W(s_i)) \log P_0 \\
+ \sum_{i:0<Q(s_i)<1} W(s_i) \log \left\{ (1 - P_1) f_1,\alpha_1 (Q(s_i)) \right\} \\
+ \sum_{i:0<Q(s_i)<1} (1 - W(s_i)) \log \left\{ (1 - P_0) f_1,\alpha_0 (Q(s_i)) \right\} \\
+ \sum_{i=1}^n W(s_i)(X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i)) \\
- \sum_{i=1}^n \log \left( 1 + \exp (X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i)) \right) \\
- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left( \eta \eta^\top K^{-1} \right) \\
- \frac{n}{2} \log \sigma_\xi^2 - \frac{1}{2\sigma_\xi^2} \text{trace} \left( \xi_0 \xi_0^\top \right),
\]

where, $f_{a,b}(Q(\cdot))$ is given by (4.5), $[U|V]$ denotes the density function of $U$ given $V$, and “const.” denotes a generic constant that does not depend on the parameters. Assume that we have completed the $l$-th iteration of the EM algorithm.
4.5.1 E-step

At the \( l \)-th iteration of the EM algorithm, the E-step involves taking expectation of the complete data loglikelihood with respect to (w.r.t.) the parameter values at the \( l \)-th iteration. We obtain the expectation iteratively, first w.r.t. \([W|\eta, \xi, Q, \theta^{[l]}]\) and then w.r.t. \([\eta, \xi|Q, \theta^{[l]}]\).

We denote the expectation w.r.t. \([W|\eta, \xi, Q, \theta^{[l]}]\) as \(Q_1(\theta^{[l]}, \eta, \xi)\). We obtain:

\[
Q_1(\theta^{[l]}, \eta, \xi) = \text{const.} + \sum_{i: Q(s_i) = 1} E \left( W(s_i)|\eta, \xi, Q, \theta^{[l]} \right) \log P_1 \\
+ \sum_{i: Q(s_i) = 0} \left( 1 - E \left( W(s_i)|\eta, \xi, Q, \theta^{[l]} \right) \right) \log P_0 \\
+ \sum_{i: 0 < Q(s_i) < 1} \left( 1 - E \left( W(s_i)|\eta, \xi, Q, \theta^{[l]} \right) \right) \log \left\{ (1 - P_1)f_{1, a_1}(Q(s_i)) \right\} \\
+ \sum_{i: 0 < Q(s_i) < 1} \left( 1 - E \left( W(s_i)|\eta, \xi, Q, \theta^{[l]} \right) \right) \log \left\{ (1 - P_0)f_{1, a_0}(Q(s_i)) \right\} \\
+ \sum_{i = 1}^n E \left( W(s_i)|\eta, \xi, Q, \theta^{[l]} \right) \left( X(s_i) + S(s_i) \eta + \xi(s_i) \right) \\
- \sum_{i = 1}^n \log \left( 1 + \exp(X(s_i)^T \beta + S(s_i)^T \eta + \xi(s_i)) \right) \\
- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left( \eta \eta^T K^{-1} \right) \\
- \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \text{trace} \left( \xi \xi^T \right). \tag{4.12}
\]

It is easy to see that the conditional distribution, \([W|\eta, \xi, Q, \theta^{[l]}]\), is a binary variable with the probability that \(W(s_i) = 1\) given by:

\[
p^{[l]}(s_i; \eta, \xi)/g^{[l]}_1(Q(s_i)) \\
p^{[l]}(s_i; \eta, \xi) / g^{[l]}_1(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi)) / g^{[l]}_0(Q(s_i)), \tag{4.13}
\]

where \(p^{[l]}(s_i; \eta, \xi)\) is obtained as

\[
p^{[l]}(s_i; \eta, \xi) = \frac{\exp(X(s_i)^T \beta^{[l]} + S(s_i)^T \eta + \xi(s_i))}{1 + \exp(X(s_i)^T \beta^{[l]} + S(s_i)^T \eta + \xi(s_i))}, \tag{4.14}
\]

107
and we define,

\[ g_1^{|l|}(Q(s_i)) \equiv [Q(s_i)|W(s_i) = 1, P_1^{|l|}, \alpha_1^{|l|}] = \left\{ P_1^{|l|} I(Q(s_i) = 1) + (1 - P_1^{|l|}) f_{1,\alpha_1^{|l|}}(Q(s_i)) \right\}, \]

and

\[ g_0^{|l|}(Q(s_i)) \equiv [Q(s_i)|W(s_i) = 0, P_0^{|l|}, \alpha_0^{|l|}] = \left\{ P_0^{|l|} I(Q(s_i) = 0) + (1 - P_0^{|l|}) f_{1,\alpha_0^{|l|}}(Q(s_i)) \right\}. \]

Note that, \( g_1^{|l|}(Q(s_i) = 0) = 0 \), and \( g_0^{|l|}(Q(s_i) = 1) = 0 \).

The resultant expression for \( Q_1(\theta^{|l|}; \eta, \xi_O) \) is:

\[
Q_1(\theta^{|l|}; \eta, \xi_O) = \text{const.} + \sum_{i:Q(s_i) = 1} \left( \frac{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i))}{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i)) + (1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))} \right) \log P_1 \\
+ \sum_{i:Q(s_i) = 0} \left( \frac{(1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))}{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i)) + (1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))} \right) \log P_0 \\
+ \sum_{i:0 < Q(s_i) < 1} \left( \frac{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i))}{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i)) + (1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))} \right) \times \log \left\{ (1 - P_1) f_{1,\alpha_1}(Q(s_i)) \right\} \\
+ \sum_{i:0 < Q(s_i) < 1} \left( \frac{(1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))}{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i)) + (1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))} \right) \times \log \left\{ (1 - P_0) f_{1,\alpha_0}(Q(s_i)) \right\} \\
+ \sum_{i=1}^n \left\{ \left( \frac{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i))}{p^{|l|}(s_i; \eta, \xi_O) g_1^{|l|}(Q(s_i)) + (1 - p^{|l|}(s_i; \eta, \xi_O)) g_0^{|l|}(Q(s_i))} \right) \times \left( X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i) \right) \\
- \log \left( 1 + \exp(X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i)) \right) \right\} \\
- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left( \eta \eta^\top K^{-1} \right) - \frac{n}{2} \log \sigma_x^2 - \frac{1}{2 \sigma_x^2} \text{trace} \left( \xi_O \xi_O^\top \right), \tag{4.15} \]

where from (4.14), \( p^{|l|}(s_i; \eta, \xi_O) \) depends on \( (\eta, \xi_O) \).
Next we obtain the expectation of $Q_1(\theta, \theta^{[l]}; \eta, \xi_{O})$ w.r.t. $[\eta, \xi_{O}|Q_{O}, \theta^{[l]}]$. The resultant expression will be denoted as $Q_2(\theta, \theta^{[l]})$, which is obtained as

$$Q_2(\theta, \theta^{[l]}) = \text{const.}$$

$$+ \sum_{\ell:Q(s_i) = 1} \mathbb{E} \left( \frac{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))} \bigg| Q_{O}, \theta^{[l]} \right) \log P_1$$

$$+ \sum_{\ell:Q(s_i) = 0} \mathbb{E} \left( \frac{(1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))} \bigg| Q_{O}, \theta^{[l]} \right) \log P_0$$

$$+ \sum_{\ell:0 < Q(s_i) < 1} \mathbb{E} \left( \frac{(1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))} \bigg| Q_{O}, \theta^{[l]} \right)$$

$$\times \log \left\{ (1 - P_1) f_{1, \ell_1}(Q(s_i)) \right\}$$

$$+ \sum_{\ell:0 < Q(s_i) < 1} \mathbb{E} \left( \frac{(1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))} \bigg| Q_{O}, \theta^{[l]} \right)$$

$$\times \log \left\{ (1 - P_0) f_{1, \ell_0}(Q(s_i)) \right\}$$

$$+ \sum_{i=1}^{n} \left\{ \mathbb{E} \left( \frac{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_{O}) g^{[l]}_1(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_{O})) g^{[l]}_0(Q(s_i))} \right) \right. $$

$$\times \left( X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i) \right) \bigg| Q_{O}, \theta^{[l]} \bigg)$$

$$- \mathbb{E} \left( \log \left( 1 + \exp \left( X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta + \xi(s_i) \right) \right) \bigg| Q_{O}, \theta^{[l]} \right) \right\}$$

$$- \frac{1}{2} \log |K| - \frac{1}{2} \text{trace} \left( E \left( \eta \eta^\top | Q_{O}, \theta^{[l]} \right) K^{-1} \right) - \frac{n}{2} \log \sigma^2_\eta - \frac{1}{2 \sigma^2_\xi} \text{trace} \left( E(\xi_{O} \xi_{O}^\top | Q_{O}, \theta^{[l]} \right) \right).$$

(4.16)

The expectations involved in (4.16) are w.r.t. $[\eta, \xi_{O}|Q_{O}, \theta^{[l]}]$, and they are not available in closed form.

When the integrals in the E-step are problematic, one common approach is to implement a stochastic EM (SEM) algorithm (see McLachlan and Krishnan, 2008; Robert and Casella, 2004), where the expectations are evaluated using Monte Carlo integration. When datasets are massive, this computation will be extremely slow, and hence the EM algorithm will be computationally very intensive. Following the ideas in Chapter 2 and Chapter 3,
we derive the Laplace approximations (LA) to approximate the expectations involved in (4.16), which is based on a second-order Taylor-series expansion of the logarithm of the integrands around their respective modes.

To apply the LA, we need to obtain the \( \hat{\eta} \) that maximizes \( Q_1(\theta^{[l]}, \theta^{[l]}, \eta, \xi_O) \). Following the approach in Chapter 2 and Chapter 3, we will use a coordinate-wise ascent method, which maximizes alternately with respect to \( \eta \) and then \( \xi_O \), until convergence. We proceed as follows:

1. Start with \( \eta^0 \) and \( \xi_O^0 \); \( t = 0 \).

2. Evaluate \( p^{[l]}(s_i; \eta, \xi_O) \), \( i = 1, \ldots, n \), at \( (\eta, \xi_O) = (\eta', \xi'_O) \), and substitute it in (4.15).

3. Maximize the resultant expression alternately w.r.t. \( \eta \) and \( \xi_O \). Call the updates \( \eta^{t+1} \) and \( \xi_O^{t+1} \).

4. If the updates and the current values of \( \eta \) and \( \xi_O \) differ by less than a prespecified tolerance level, then quit; we have \( \eta^{[l]} = \eta' \) and \( \xi_O^{[l]} = \xi'_O \). Otherwise, \( t = t + 1 \), and go to step 2.

Once the maximizer of \( Q_1(\theta^{[l]}, \theta^{[l]}, \eta, \xi_O) \) is obtained, we follow the methodology developed in Chapter 3. That is, we use a second-order Taylor-series expansion to approximate the posterior distribution, \([\eta, \xi_O|Q_O, \theta^{[l]}]\), with a Gaussian distribution whose mean and variance are given by the posterior mode and the inverse of the negative Hessian of the posterior, evaluated at the mode; see the justification in Kass and Steffey (1989). Details of our approximation can be found in the Appendix, where it is seen that the posterior distribution, \([\eta, \xi_O|Q_O, \theta^{[l]}]\), is approximately a multivariate Gaussian density, with mean and variance given by:

\[
E\left(\begin{pmatrix} \eta \\ \xi_O \end{pmatrix} | Q_O, \theta^{[l]}\right) = \begin{pmatrix} \hat{\eta}^{[l]} \\ \hat{\xi}_O^{[l]} \end{pmatrix},
\]

(4.17)
and

\[
\begin{align*}
\text{var} \left( \begin{pmatrix} \eta \\ \xi_O \end{pmatrix} \mid Q_O, \theta[i] \right) &= \left\{ \begin{pmatrix} -\frac{\partial^2}{\partial \eta \partial \eta} (Q_1(\theta[i], \theta[i]; \eta, \xi_O)) & -\frac{\partial^2}{\partial \eta \partial \xi_O} (Q_1(\theta[i], \theta[i]; \eta, \xi_O)) \\ -\frac{\partial^2}{\partial \xi_O \partial \eta} (Q_1(\theta[i], \theta[i]; \eta, \xi_O)) & -\frac{\partial^2}{\partial \xi_O \partial \xi_O} (Q_1(\theta[i], \theta[i]; \eta, \xi_O)) \end{pmatrix} \right\}^{-1}
\end{align*}
\]

respectively. To obtain \( \text{var}(\eta|Q_O, \theta[i]) \) and \( \text{var}(\xi_O|Q_O, \theta[i]) \), we need to invert the matrix of second-order partial derivatives shown just above. Let \( A \) denote an \( r \times r \) matrix and \( B \) denote an \( n \times n \) matrix. Further, let \( U \) be any \( r \times n \) matrix and \( V \) be any \( n \times r \) matrix. Then, a block-matrix-inversion formula (e.g., Duncan, 1944) is given by:

\[
\begin{pmatrix} A & U \\ V & B \end{pmatrix}^{-1} = \begin{pmatrix} (A - UB^{-1}V)^{-1} & -(A - UB^{-1}V)^{-1}UB^{-1} \\ -(B - VA^{-1}U)^{-1}VA^{-1} & (B - VA^{-1}U)^{-1} \end{pmatrix}. \tag{4.19}
\]

Now recall the Sherman-Morrison-Woodbury formula (e.g., Henderson and Searle, 1981):

\[
(B - VA^{-1}U)^{-1} = B^{-1} + B^{-1}V(A - UB^{-1}V)^{-1}UB^{-1}.
\]

We use this formula in the block-matrix-inversion formula (4.19) to obtain the following equivalent block-matrix-inversion formula, which we use to obtain the inverse in (4.18) (see Chapter 3):

\[
\begin{pmatrix} A & U \\ V & B \end{pmatrix}^{-1} = \begin{pmatrix} (A - UB^{-1}V)^{-1} & -(A - UB^{-1}V)^{-1}UB^{-1} \\ -(B - VA^{-1}U)^{-1}VA^{-1} & (B - VA^{-1}U)^{-1} \end{pmatrix}. \tag{4.20}
\]

Now, for generic variables \( u \) and \( v \), define

\[
J(u_0, v_0) = -\left. \frac{\partial^2}{\partial u \partial v} \left( Q_1(\theta[i], \theta[i]; u, v) \right) \right|_{u=u_0, v=v_0}.
\]
We consider the different component matrices in the \((r + n) \times (r + n)\) matrix of partial derivatives given in (4.18). The matrix \(J(\hat{\xi}_O^{[l]}, \hat{\eta}_O^{[l]})\) is an \(n \times n\) diagonal matrix; its inversion is easy. The matrix \(J(\hat{\eta}_O^{[l]}, \hat{\eta}_O^{[l]})\) is of dimension \(r \times r\), where \(r < n\). The other two matrices, \(J(\hat{\xi}_O^{[l]}, \hat{\eta}_O^{[l]})\) and \(J(\hat{\eta}_O^{[l]}, \hat{\xi}_O^{[l]})\), have dimension \(n \times r\) and \(r \times n\), respectively. We use formula (4.20) to invert the matrix in (4.18), which gives, approximately,

\[
\begin{align*}
\text{var}(\eta|Q_O, \theta^{[l]}) &= \left( J(\eta^{[l]}, \hat{\eta}_O^{[l]}) - J(\hat{\eta}_O^{[l]}, \xi_O^{[l]}) J(\xi_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\xi_O^{[l]}, \hat{\eta}_O^{[l]}) \right)^{-1} \\
\text{var}(\xi_O|Q_O, \theta^{[l]}) &= J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} + J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]}) \\
&\quad \times \left( J(\eta^{[l]}, \hat{\eta}_O^{[l]}) - J(\hat{\eta}_O^{[l]}, \xi_O^{[l]}) J(\xi_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\xi_O^{[l]}, \hat{\eta}_O^{[l]}) \right)^{-1} \\
&\quad \times J(\hat{\eta}_O^{[l]}, \xi_O^{[l]}) J(\xi_O^{[l]}, \xi_O^{[l]})^{-1} \\
\text{cov}(\eta, \xi_O|Q_O, \theta^{[l]}) &= - \left( J(\eta^{[l]}, \hat{\eta}_O^{[l]}) - J(\hat{\eta}_O^{[l]}, \xi_O^{[l]}) J(\xi_O^{[l]}, \hat{\xi}_O^{[l]})^{-1} J(\xi_O^{[l]}, \hat{\eta}_O^{[l]}) \right)^{-1} \\
&\quad \times J(\hat{\eta}_O^{[l]}, \xi_O^{[l]}) J(\xi_O^{[l]}, \xi_O^{[l]})^{-1}.
\end{align*}
\]

In the formulas given just above, all we need to invert is the \(n \times n\) diagonal matrix, \(J(\hat{\xi}_O^{[l]}, \hat{\xi}_O^{[l]})\), and some fixed-rank \(r \times r\) matrices. This makes the computations extremely efficient and allows us to obtain the expressions for \(E(\eta\eta^\top|Q_O, \theta^{[l]})\) and \(E(\xi_O\xi_O^\top|Q_O, \theta^{[l]})\) in (4.16) as follows:

\[
\begin{align*}
E(\eta\eta^\top|Q_O, \theta^{[l]}) &= \text{var}(\eta|Q_O, \theta^{[l]}) + E(\eta|Q_O, \theta^{[l]}) E(\eta|Q_O, \theta^{[l]})^\top \\
E(\xi_O\xi_O^\top|Q_O, \theta^{[l]}) &= \text{var}(\xi_O|Q_O, \theta^{[l]}) + E(\xi_O|Q_O, \theta^{[l]}) E(\xi_O|Q_O, \theta^{[l]})^\top,
\end{align*}
\]

where the terms on the right-hand side of (4.22) are evaluated approximately using (4.17) and (4.21).

The remaining terms in (4.16), for which we need an approximation, are

\[
E \left( \log \left( 1 + \exp \left\{ X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta + \tilde{\xi}(s_i) \right\} \right| Q_O, \theta^{[l]} \right),
\]

112
Using (4.23) repeatedly, we obtain the two required expectations (see Appendix) as,

\[ E \left( \frac{p^{[l]}(s_i; \eta, \xi_o) g_1^{[l]}(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_o) g_1^{[l]}(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_o)) g_0^{[l]}(Q(s_i))} \times \left( X(s_i)^{\top} \beta + S(s_i)^{\top} \eta + \xi(s_i) \right) | Q_o, \theta^{[l]} \right), \]

where \( i = 1, \ldots, n \). From the Appendix, we see that the expression for the expectation of a generic function \( b(q(s_i)^{\top} \delta) \), where \( \delta \equiv (\eta^\top, \xi_o^\top) \), conditional on \( Q_o \) and \( \theta^{[l]} \) can be written approximately as:

\[
E \left( b(q(s_i)^{\top} \delta) | Q_o, \theta^{[l]} \right) \\
\qquad \approx b(q(s_i)^{\top} \delta^{[l]}) + \frac{1}{2} \text{trace} \left( \text{var}(\delta | Q_o, \theta^{[l]}) \left( b''(q(s_i)^{\top} \delta^{[l]}) \times q(s_i) q(s_i)^{\top} \right) \right) \\
\qquad = b(q(s_i)^{\top} \delta^{[l]}) + \frac{1}{2} b''(q(s_i)^{\top} \delta^{[l]}) \times q(s_i)^{\top} \text{var}(\delta | Q_o, \theta^{[l]}) q(s_i). \tag{4.23}
\]

Using (4.23) repeatedly, we obtain the two required expectations (see Appendix) as,

\[
E \left( \log \left( 1 + \exp \left\{ X(s_i)^{\top} \beta^{[l]} + S(s_i)^{\top} \eta + \xi(s_i) \right\} | Q_o, \theta^{[l]} \right) \right) \\
\qquad \approx \log \left( 1 + \exp \left\{ X(s_i)^{\top} \beta^{[l]} + S(s_i)^{\top} \eta^{[l]} + \xi^{[l]}(s_i) \right\} \right) \\
\qquad \quad + \frac{1}{2} \exp \left\{ X(s_i)^{\top} \beta^{[l]} + S(s_i)^{\top} \eta^{[l]} + \xi^{[l]}(s_i) \right\} \left( 1 + \exp \left\{ X(s_i)^{\top} \beta^{[l]} + S(s_i)^{\top} \eta^{[l]} + \xi^{[l]}(s_i) \right\} \right)^{-1} \\
\qquad \times \left( S(s_i)^{\top} \text{var}(\eta | Q_o, \theta^{[l]}) S(s_i) + 2 S(s_i)^{\top} \text{cov}(\eta, \xi_o | Q_o, \theta^{[l]}) e(s_i) + e(s_i)^{\top} \text{var}(\xi_o | Q_o, \theta^{[l]}) e(s_i) \right), \tag{4.24}
\]

113
and, 

\[ E \left\{ \left( \frac{p_i(s; \eta, \xi)}{p_i(s; \eta, \xi)} g_1(Q(s)) \right) \times \left( X(s)^T \beta + S(s)^T \eta + \xi \right) \mid O, \theta \right\} \]

\[ \approx \left\{ \left( \frac{p_i(s; \eta, \xi)}{p_i(s; \eta, \xi)} g_1(Q(s)) \right) \times \left( X(s)^T \beta + S(s)^T \eta + \xi \right) \mid O, \theta \right\} \]

\[ + \frac{1}{2} \left( \frac{g_1(Q(s)) - g_1(Q(s))}{g_1(Q(s)) + g_1(Q(s))} \right) \times \left( X(s)^T \beta + S(s)^T \eta + \xi \right) \]

\[ \times \left( X(s) \right) \cdot \text{var}(\eta | O, \theta) S(s) + 2S(s) \cdot \text{cov}(\eta, \xi | O, \theta) e(s) + e(s) \cdot \text{var}(\xi | O, \theta) e(s) \right) \]

+ terms independent of \( \theta \),

(4.25)

where 

\[ p_i(s; \eta, \xi) = \frac{\exp(X(s)^T \beta + S(s)^T \eta + \xi)}{1 + \exp(X(s)^T \beta + S(s)^T \eta + \xi)} \]

and \( e(s_i) \) is a vector of length \( n \) that has 1 at the \( i \)-th position, and all other entries in \( e(s_i) \) are zero; \( i = 1, \ldots, n \).

### 4.5.2 M-step

Following the E-step, we perform the M-step, where we maximize (4.16) with respect to each of the parameters in \( \theta \): The maximization w.r.t. \( P_0, P_1, K \) and \( \sigma_\xi^2 \) is obtained by differentiating (4.16) w.r.t. \( P_0, P_1, K \), and \( \sigma_\xi^2 \), equating the result to zero, and then solving
the resulting equations. The updating equations are obtained in closed form:

\[
\sigma^2_{\xi}^{[l+1]} = \frac{1}{n} \text{trace} \left( E(\xi_{O} | Q_{O}, \theta^{[l]}) E(\xi_{O} | Q_{O}, \theta^{[l]})^\top + \text{var} \left( \xi_{O} | Q_{O}, \theta^{[l]} \right) \right)
\]

\[
K^{[l+1]} = E(\eta | Q_{O}, \theta^{[l]}) E(\eta | Q_{O}, \theta^{[l]})^\top + \text{var} \left( \eta | Q_{O}, \theta^{[l]} \right)
\]

\[
P_0^{[l+1]} = \frac{\sum_{i:Q(s_i)=0} \left( 1 - E(W(s_i) | Q_{O}, \theta^{[l]}) \right)}{\sum_{i:Q(s_i)<1} \left( 1 - E(W(s_i) | Q_{O}, \theta^{[l]}) \right)}
\]

\[
P_1^{[l+1]} = \frac{\sum_{i:Q(s_i)=1} E(W(s_i) | Q_{O}, \theta^{[l]})}{\sum_{i:Q(s_i)>0} E(W(s_i) | Q_{O}, \theta^{[l]})},
\]

(4.26)

where \( E(W(s_i) | Q_{O}, \theta^{[l]}), \) is obtained as (see Section 4.5.1):

\[
E(W(s_i) | Q_{O}, \theta^{[l]}) \
\approx \left( \frac{p^{[l]}(s_i, \eta^{[l]}, \xi^{[l]}_{O}) g^{[l]}_{\theta}(Q(s_i))}{p^{[l]}(s_i, \eta^{[l]}, \xi^{[l]}_{O}) g^{[l]}_{\theta}(Q(s_i)) + (1 - p^{[l]}(s_i, \eta^{[l]}, \xi^{[l]}_{O})) g^{[l]}_{0}(Q(s_i))} \right) \\
\quad + \frac{1}{2} \left( \frac{g^{[l]}_{0}(Q(s_i)) - g^{[l]}_{1}(Q(s_i)) \exp(X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta^{[l]} + \xi^{[l]}_{O}(s_i))}{g^{[l]}_{0}(Q(s_i)) + g^{[l]}_{1}(Q(s_i)) \exp(X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta^{[l]} + \xi^{[l]}_{O}(s_i))} \right) ^2 \\
\quad \times \left( g^{[l]}_{0}(Q(s_i)) g^{[l]}_{1}(Q(s_i)) \exp \left( X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta^{[l]} + \xi^{[l]}_{O}(s_i) \right) \right) \\
\quad \times \left( S(s_i)^\top \text{var}(\eta | Q_{O}, \theta^{[l]}) S(s_i) + 2 S(s_i)^\top \text{cov}(\eta, \xi^{[l]}_{O} | Q_{O}, \theta^{[l]}) e(s_i) + e(s_i)^\top \text{var}(\xi^{[l]}_{O} | Q_{O}, \theta^{[l]}) e(s_i) \right)
\]

(4.27)

The maximizations with respect to \( \alpha_0, \alpha_1, \) and \( \beta \) are not available in closed form. We use a one-step Newton-Raphson update for these parameters, within each iteration of the EM algorithm. For \( \beta \), we obtain the update as follows:

\[
\beta^{[l+1]} = \beta^{[l]} - \left[ \frac{\partial}{\partial \beta} R(\theta) \right]^{-1}_{\theta = \theta^{[l]}} R(\theta^{[l]}).
\]

(4.28)

In (4.28) above, \( R(\theta) \) denotes the score function obtained by differentiating \( Q_2(\theta, \theta^{[l]}) \), given by (4.16), w.r.t. \( \beta \), and \( R(\theta^{[l]}) \) is obtained by evaluating \( R(\theta) \) at \( \theta^{[l]} \). A one-step Newton-Raphson update for \( \alpha_0 \) and \( \alpha_1 \) is obtained in a similar manner.
4.5.3 Starting Values for the EM Algorithm

In order to implement the EM algorithm, we need to specify some starting value for the parameters; it is an important part of the algorithm to select them appropriately. In this section, we specify how we select starting values for the EM algorithm discussed above.

Following the NASA cloud-mask classification scheme, we make use of the classification categories that are released as the level-2 cloud-mask product. A pixel is classified into one of the four possible categories by thresholding the final Q-values, \( Q(\cdot) \) (e.g., Platnick et al., 2003). The four categories included in the cloud-mask output are: 1) confident clear (\( Q > 0.99 \)), 2) probably clear (\( 0.95 < Q \leq 0.99 \)), 3) probably cloudy (\( 0.66 < Q \leq 0.95 \)), and 4) cloudy (\( Q \leq 0.66 \)). If a pixel located at \( s_i \) is classified as clear or probably clear (cloudy or probably cloudy), we define \( W_i^{[0]}(s_i) = 1 \) (respectively, \( W_i^{[0]}(s_i) = 0 \)), \( i = 1, \ldots, n \).

The starting values for \( P_0 \) and \( P_1 \) are now obtained as,

\[
P_0^{[0]} = \frac{\sum_{i:Q(s_i)=0} (1 - W_i^{[0]}(s_i))}{\sum_{i:Q(s_i)<1} (1 - W_i^{[0]}(s_i))} \\
P_1^{[0]} = \frac{\sum_{i:Q(s_i)=1} W_i^{[0]}(s_i)}{\sum_{i:Q(s_i)>0} W_i^{[0]}(s_i)}
\]  

(4.29)

To obtain the starting values for \( \alpha_0 \) and \( \alpha_1 \), we maximize the likelihoods,

\[
L_0(\alpha_0) = \text{const.} + \sum_{i:Q(s_i)=0} (1 - W_i^{[0]}(s_i)) \log P_0^{[0]} \\
+ \sum_{i:0<Q(s_i)<1} (1 - W_i^{[0]}(s_i)) \log \left\{ (1 - P_0^{[0]}) f_{1,\alpha_0}(Q(s_i)) \right\}, 
\]

(4.30)

and

\[
L_1(\alpha_1) = \text{const.} + \sum_{i:Q(s_i)=1} W_i^{[0]}(s_i) \log P_1^{[0]} \\
+ \sum_{i:0<Q(s_i)<1} W_i^{[0]}(s_i) \log \left\{ (1 - P_1^{[0]}) f_{1,\alpha_1}(Q(s_i)) \right\},
\]

(4.31)
respectively.

Next we fit a non-spatial fixed-effects generalized linear model (GLM) to \( \{ W^{[0]}(s_i) : i = 1, \ldots, n \} \), using the iterated reweighted least squares algorithm (e.g., McCullagh and Nelder, 1989, Ch. 2). The resulting GLM estimate, \( \hat{\beta}_{GLM} \), will be used as the starting value for \( \beta \), which we will denote as \( \beta^{[0]} \).

Next consider the detrended process,

\[
U(s_i) \equiv Y(s_i) - X(s_i) \top \beta^{[0]},
\]

which has mean zero and

\[
\text{var}(U(s_i)) = S(s_i) \top KS(s_i) + \sigma_s^2.
\]

Writing \( U_O \equiv (U(s_1), \ldots, U(s_n)) \top \), we obtain:

\[
\text{cov}(U_O) \equiv \Sigma_{U;O} = S_OK_S \top + \sigma_s^2 I_n.
\]

To obtain method-of-moments estimates of \( K \) and \( \sigma_s^2 \) that can be used as starting values, we replace \( Y(s_i) \) with \( \log \left( \frac{Z(s_i) + c}{1 - Z(s_i) + c} \right) \), where \( c \) is some user-specified constant that is added to the data to ensure that the transformation is defined everywhere within the range of the data (see Chapter 3).

Consequently, an approximation for \( U(\cdot) \) is obtained as:

\[
\hat{U}(s_i) \equiv \log \left( \frac{Z(s_i) + c}{1 - Z(s_i) + c} \right) - X(s_i) \top \hat{\beta}^{[0]}, \quad i = 1, \ldots, n.
\]

Define \( s_U^2 \equiv \frac{1}{n} \sum_{i=1}^n \hat{U}(s_i)^2 \), and choose

\[
\hat{\Sigma}_{U;O} = s_U^2 I_n,
\]

simply to capture the total variation through the trace operator. We apportion approximately 90\% of this to the smooth small-scale variation and 10\% to the fine-scale variation (e.g., Katzfuss and Cressie, 2011). That is, we select our starting values for \( K \) and \( \sigma_s^2 \) to satisfy

\[
S_OK^{[0]}S_O \top \approx 0.9 \times \hat{\Sigma}_{U;O}
\]

\[
\sigma_s^2 = 0.1 \times \text{trace}(\hat{\Sigma}_{U;O}) / n,
\]

(4.37)
as follows. Using (4.37), and the \( Q-R \) decomposition, \( S_O = Q_S R_S \), we obtain the starting value for \( K \) as

\[
K^{[0]} = R_S^{-1} Q_S^\top \left( 0.9 \times \hat{\Sigma}_{U,O} \right) Q_S (R_S^\top)^{-1}.
\]  

(4.38)

Note that this approximate 90-10 apportionment of the total variability could be done differently, depending on the data’s smooth-scale variation relative to their fine-scale variation.

### 4.6 Spatial Statistical Analysis of MODIS Cloud Data

In this section, we carry out a spatial statistical analysis of the granule of MODIS cloud data shown in Figure 4.1, using the hierarchical spatial statistical model discussed in Section 4.3, and use it to predict the underlying clear-sky-probabilities for the entire study region, \( D = D_O \cup D_U \), along with estimates of prediction uncertainties.

Our data are the MODIS cloud-mask data (the final Q-values) shown in Figure 4.3. A map showing the geographical location of the granule is given in Figure 4.4.

Recall that the granule of MODIS data shown in Figure 4.1 was collected by the MODIS instrument on the Terra satellite and corresponds to June 29, 2006, 12:45 UTC. The number of 1 km \( \times \) 1 km pixels within the study region is \( N = 2,748,620 \), and we have data for all the locations; hence \( n = 2,748,620 \). The study region spans a region from longitudes -45 degrees to -10 degrees, and from latitudes 20 degrees to 45 degrees.

#### 4.6.1 Spatial Basis Functions

In this section, we discuss the choice of spatial basis function. The selection of the basis functions, its type and number, is a current area of research (e.g., Bradley et al., 2011). For the purpose of this analysis, we selected as basis functions the bisquare functions (e.g., Cressie and Johannesson, 2006, 2008). The generic form of a bisquare function is,

\[
b(s) = \left\{ 1 - \left( \frac{||s - c||}{w} \right) \right\}^2 I(||s - c|| < w),
\]  

(4.39)
where $c$ is the center of the basis function, and $I(A)$ is an indicator function equal to 1 if $A$ is true, and equal to 0 otherwise. Centers $\{c_i\}$ in $D$ are usually chosen according to a multi-resolution scheme (e.g., a quad-tree). Then the “aperture” $w$ is given by (Cressie and Johannesson, 2008),

$$w = 1.5 \times \text{shortest great arc distance between like-resolution center points}$$

A pictorial illustration of the bisquare basis function is given in Figure 4.5. Other choices for basis functions are also possible (e.g., EOFs in Wikle and Cressie, 1999; W-wavelets in Shi and Cressie, 2007).
Figure 4.5: A two-dimensional bisquare function as a 3-D plot (left) and as an image plot (right).

We employ several resolutions of the basis function to capture the different scales of spatial variability; here we use three resolutions to obtain \( \{b_i(s) : i = 1, \ldots, (r_1 + r_2 + r_3)\} \), where \( r_1 = 12, r_2 = 34, \) and \( r_3 = 102 \), are the number of basis functions at the three resolutions. The centers of the bisquare basis functions were selected using a quad-tree structure (e.g., Cressie and Kang, 2010), ensuring that the centers for the different resolutions do not match. The number of basis functions were determined to allow full coverage of the spatial domain. We also included centers of the bisquare function outside the study region to account for the boundary effects (e.g., Cressie and Kang, 2010). We further standardized the bisquare function \( b_i(\cdot) \) to obtain the \( i \)-th basis function,

\[
S_i(s) \equiv \frac{b_i(s) - \text{ave}_{s \in D} (b(s))}{\left\{ \text{var}_{s \in D} (b(s)) \right\}^{1/2}}; \quad i = 1, \ldots, (r_1 + r_2 + r_3),
\]
where \( \text{ave}_{s \in D}(\cdot) \) and \( \text{var}_{s \in D}(\cdot) \) are spatial moments taken over the domain of interest. The locations of the basis function centers for all three resolutions are shown in Figure 4.6.

Figure 4.6: Centers of the basis function; ’o’, ’+’, and ’x’ are use to distinguish the three scales of resolution.
4.6.2 Optimal Spatial Mapping of the Clear-Sky Probability Process

We now discuss the results obtained by fitting the hierarchical model to the MODIS cloud-mask product shown in Figure 4.3, and we produce optimal spatial maps showing the predictions for the underlying clear-sky probability process, along with maps showing the prediction uncertainties.

First, let us consider the covariates \( X(\cdot) \) in (4.8). We include the vector 1, and latitude as a covariate. Further, instead of using the coarsest-resolution \( S_1(\cdot), \ldots, S_r(\cdot) \) as spatial basis function, we use them as covariates in \( X(\cdot) \) (e.g., Shi and Cressie, 2007).

The second term of (4.8) involves an \( r \)-dimensional vector, \( S(\cdot) \), of spatial basis functions, which in our case is made up of the bisquare functions at the second and the third resolutions (see Figure 4.6). Now, there are regions in the study region that are affected by sun glint (see Figure 4.1), which the MODIS cloud algorithm attempts to account for by implementing clear-sky restoral tests. Nevertheless, the presence or absence of sun glint is a source of variability that exists for the granule we consider. Hence, we include the sun-glint indicator flag (which takes a value 1 if a pixel is affected by sun glint, and is 0 otherwise) as a column in \( S(\cdot) \). That is, \( r = 1 + r_2 + r_3 = 1 + 34 + 102 = 137 \).

With the above model specifications, and starting values obtained using the methodology discussed in Section 4.5.3, we implemented the EM algorithm to obtain the parameter estimates for the hierarchical model defined in Section 4.3. After having obtained the EM estimates, \( \hat{\theta}_{EM} = \left\{ \hat{P}_{0;EM}, \hat{P}_{1;EM}, \hat{\alpha}_{0;EM}, \hat{\alpha}_{1;EM}, \hat{\beta}_{EM}, \hat{K}_{EM}, \hat{\sigma}^2_{\xi;EM} \right\} \), we substitute these estimates into the MCMC algorithm (see Appendix) to obtain samples from the (empirical) predictive distribution, \( [W, \eta, \xi|Q_O, \hat{\theta}_{EM}] \). We generated 10,000 MCMC samples, after discarding 1,000 samples as burn-in. Because of storage issues involved with storing the \( N \)-dimensional vector \( \xi \), we saved every fifth MCMC sample generated. The EM algorithm converged after 14 iterations, and the computational time for the EM algorithm was 27.76 minutes. The computational time for the MCMC was 12.73 hours. All the computations
were performed on a dual quad core 2.8 GHz 2x Xeon X5560 processor, with 96 Gbytes of memory.

The hierarchical nature of the model allows us to look into the different sources of variability, separately at the logit scale (i.e., the different components of $Y(\cdot)$). The predictive mean of these different components of variability are shown in Figure 4.7.

Figure 4.7: Maps showing the components due to trend, $X(\cdot)\hat{\beta}_{EM}$ (top-left panel), the mean of the predictive distribution of the random-effects component, $E[S(\cdot)^T\eta|Q,\hat{\theta}_{EM}]$ (top-right panel), the mean of the predictive distribution of the component due to fine scale variation, $E[\xi(\cdot)|Q,\hat{\theta}_{EM}]$ (bottom-left panel), and the mean of the process $Y(\cdot)$, $E[Y(\cdot)|Q_O,\hat{\theta}_{EM}]$ (bottom-right panel). The bottom-right panel is obtained by adding the three other components.
Now, using the MCMC samples referred to above, we computed the predictive mean and the predictive standard deviation of the clear-sky-probability process, $p(s_i)$, given by (4.7). We also obtained the pixelwise 2.5 and 97.5 percentiles of each of the $N$ elements of $\mathbf{p} \equiv (p(s_1), \ldots, p(s_N))^\top$ from the predictive distribution, $[\mathbf{p}|\mathbf{Q}_O, \hat{\theta}_{EM}]$, computed from $[\mathbf{W}, \eta, \xi|\mathbf{Q}_O, \hat{\theta}_{EM}]$. Figure 4.8 shows maps of the pixelwise predictive mean, the pixelwise predictive standard deviation, and the pixelwise 2.5 and 97.5 percentiles, respectively; the latter two quantities are the end-points of a pixelwise 95% prediction interval.

4.7 Change-of-Support

In this section, we discuss spatial change-of-support and outline the role it plays in predicting cloud-fraction (defined later) at resolutions coarser than 1 km × 1 km (i.e., upscaling). We finish the section with a discussion of how one could adapt our modeling framework to predict clear-sky probabilities (or conversely, cloudiness probabilities) at resolutions finer than 1 km × 1 km (i.e., downscaling).

4.7.1 Cloud-Fraction at Resolutions Coarser than 1 km × 1 km

In applications such as weather forecasting and climate projections, researchers are interested in geophysical questions that involve cloud-fraction at resolutions coarser than 1 km × 1 km. In what follows, we define cloud-fraction and state how it is currently being estimated. Then we discuss its estimation and statistical inference based on our hierarchical spatial statistical model.

Cloud-fraction is defined as follows: Let $C(s) = 1$ if there is cloud present (and 0 otherwise) at $s \in D$. Consider an area $B \subset D$. Then the cloud-fraction for $B$ is defined formally as $\int_B C(s) ds / |B|$, where $|B|$ denotes the area of $B$. Usually $B$ is at a coarser resolution than the
Figure 4.8: Maps showing the predictive mean (top-left panel), the pixelwise predictive standard-deviation (top-right panel), the pixelwise 2.5 percentile (bottom-left panel) and the pixelwise 97.5 percentile (bottom-right panel) for the predictive distribution of the clear-sky-probability process.

BAUs, in which case the cloud-fraction for $B$ can be numerically approximated as follows. Define

$$CF(B) \equiv \frac{\sum_{s_i \in B} I\{W(s_i) = 0\}}{\sum_{s_i \in B} 1},$$

(4.41)

where recall that $W(s_i) = 0$ means that the BAU located at $s_i$ is cloudy. This is the definition used by Zhao and Di Girolamo (2006).
The cloud-fraction is important for a variety of different applications (e.g., see the introduction in Di Girolamo and Davies, 1997). Currently, the cloud-fraction, $CF(B)$, is estimated as follows: Classify each 1 km×1 km pixel as cloudy or clear by thresholding the final Q-value. If $Q(s_i) > 0.95$, then the pixel $s_i$ is classified as clear ($W(s_i) = 1$), and if $Q(s_i) \leq 0.95$, then the pixel $s_i$ is classified as cloudy ($W(s_i) = 0$). An estimate of cloud-fraction is then given by,

$$
\hat{CF}(B) = \frac{\sum_{s_i \in B} I \{ Q(s_i) \leq 0.95 \}}{\sum_{s_i \in B} 1}.
$$

(4.42)

Notice that the estimate (4.42) is affected by the threshold value used for cloud-mask classification.

The cloud-fraction, $CF(B)$, is often used to determine if remote sensing retrievals of other geophysical quantities will be computed for an instrument whose footprint is $B$. For example, for atmospheric aerosols, no retrieval would be reported if $CF(B)$ exceeded a pre-specified tolerance value, $k$. Therefore, the associated geophysical quantity of interest is,

$$
\delta(B) \equiv I \{ CF(B) > k \}.
$$

(4.43)

Now,

$$
E(\delta(B)) = Pr \{ CF(B) > k \},
$$

(4.44)

$$
\text{var}(\delta(B)) = Pr \{ CF(B) > k \} (1 - Pr \{ CF(B) > k \}),
$$

(4.45)

and

$$
\text{cov}(\delta(B_1), \delta(B_2)) = Pr \{ CF(B_1) > k, CF(B_2) > k \}
- Pr \{ CF(B_1) > k \} Pr \{ CF(B_2) > k \}.
$$

(4.46)

Therefore, to make inference on cloud-fraction, we need the predictive distribution, $[\delta(\cdot)|Q_O, \hat{\theta}_{EM}]$. 

126
Using the hierarchical modeling framework defined in Section 4.3 and MCMC samples from the predictive distribution, \([W, \eta, \xi|Q_O, \hat{\theta}_{EM}]\), we can compute the predictive distribution of any desired summary statistic of \(\{W(s_i) : i = 1, \ldots, N\}\), including \(\delta(\cdot)\), to answer related geophysical questions of interest.

### 4.7.2 Downscaling to Obtain Clear-Sky Probability at Finer Resolutions

In some applications where researchers want to study local weather phenomena, interest lies in predicting the clear-sky probabilities (or equivalently, the probabilities of cloudiness) at resolutions finer than the 1 km × 1 km pixels. The method of obtaining high-resolution information from relatively coarse-resolution data is called downscaling.

When interest lies in downscaling, one can proceed by defining the *process models* for \(W(\cdot)\) and \(Y(\cdot)\) at the finest resolution at which we are interested in doing inference. Then, the *data model* will be specified at the (coarser) resolutions at which we have data, conditional on the process variables at the finest resolution. We can write the distribution, \([process 1, process 2|data, parameters]\), as

\[
[process 1, process 2|data, parameters] \\
\propto [data|process 1, process 2, parameters][process 1|process 2, parameters] \\
[process 2|parameters], 
\]  

(4.47)

where the data are at a coarser resolution than the processes. Hence, to carry out inference on \(W(\cdot)\) (i.e., process 1), we simulate from the predictive distribution,

\([process 1, process 2|data, parameters]\).

Consequently, \([W(s)|data, parameters]\) is the predictive distribution of \(W(s)\), and the Bayes’ estimate of cloudiness is \(Pr(W(s) = 0|data, parameters)\), where \(s\) is a location at the finest resolution.
4.8 Discussion and Conclusions

In this chapter, we developed a hierarchical spatial statistical model for analyzing a remote-sensing dataset on clouds from NASA’s MODIS instrument. The dataset is massive in size ($N = n = 2,748,620$). However, use of the reduced-rank SRE model to capture the spatial covariance of the latent process $Y(\cdot)$ allows for fast computations.

We took an empirical hierarchical modeling (EHM) approach, where the unknown model parameters were estimated using an EM algorithm. Alternatively, one could take a Bayesian hierarchical modeling (BHM) approach, where a prior distribution is put on the parameters. Kang and Cressie (2011) developed the “Givens angle prior” for $K$, which could be adapted for analyzing the cloud data using a BHM approach in very much the same way as was done in Chapter 3. While the prediction intervals computed using the EHM approach tend to be too liberal when compared to those using a BHM approach, EHM is an order of magnitude faster (e.g., Chapter 3).

Within the hierarchical modeling framework that we developed in this chapter, we used the SRE model to define the underlying Gaussian field. Other than computational speed-ups, these models do not rely on specifying a spatial-weights matrix, and no assumptions of homogeneity, stationarity, or isotropy are needed. The hierarchical statistical modeling framework considered here enables us to do inference on cloud-fraction at resolutions coarser than 1 km $\times$ 1 km pixels and, in principle, at finer resolutions as well.

To our knowledge this is a first attempt to develop a hierarchical spatial statistical model for a cloud dataset at such a fine resolution. The spatial model developed here could be extended to a spatio-temporal setting that might be used to improve a climate model’s subgrid-scale physical parameterization. The spatial statistical model developed here could also be used to develop a data-fusion methodology to incorporate cloud data (e.g., to fuse water-vapor data from NASA’s AIRS instrument with cloud data from MODIS).
A.1 Approximations Involved in the EM Algorithm

Let \( \delta \equiv (\eta^\top, \xi^\top)^\top \) be an \( m \)-dimensional vector, where \( m = r + n \). Here we derive the Laplace approximation to the density \([\delta|Z_O, \theta^{[l]}]\). Let \( \hat{\delta}^{[l]} \) maximize the complete data log likelihood, \( L_c(\theta^{[l]}|Z_O, \delta) \). Now, the density for the distribution of \([\delta|Z_O, \theta^{[l]}]\) is given by:

\[
p(\delta|Z_O, \theta^{[l]}) \propto \exp \left( L_c(\theta^{[l]}|Z_O, \delta) \right).
\]

(A.1)

A Taylor-series approximation of \( L_c(\theta^{[l]}|Z_O, \delta) \) around \( \hat{\delta}^{[l]} \) yields:

\[
L_c(\theta^{[l]}|Z_O, \delta) = L_c(\theta^{[l]}|Z_O, \hat{\delta}^{[l]}) + \frac{1}{2} (\delta - \hat{\delta}^{[l]})^\top \left[ \frac{\partial^2}{\partial \delta \partial \delta} L_c(\theta^{[l]}|Z_O, \delta) \right]_{\delta = \hat{\delta}^{[l]}} (\delta - \hat{\delta}^{[l]}) + \text{higher-order terms}
\]

(A.2)

where \( Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O) \equiv - \left[ \frac{\partial^2}{\partial \delta \partial \delta} L_c(\theta^{[l]}|Z_O, \delta) \right]_{\delta = \hat{\delta}^{[l]}} \). In (A.2) above, notice that the first-order linear term is zero since the first-order derivative of \( L_c(\theta^{[l]}|Z_O, \delta) \), with respect to \( \delta \), evaluated at \( \delta = \hat{\delta}^{[l]} \), is zero (recall that \( \hat{\delta}^{[l]} \) maximizes \( L_c(\theta^{[l]}|Z_O, \delta) \)). Therefore, for the density of \( \delta|Z_O, \theta^{[l]} \), we have approximately,

\[
p(\delta|Z_O, \theta^{[l]}) \propto \exp \left( L_c(\theta^{[l]}|Z_O, \hat{\delta}^{[l]}) \right) \\
\times \exp \left( -\frac{1}{2} (\delta - \hat{\delta}^{[l]})^\top Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)(\delta - \hat{\delta}^{[l]}) \right).
\]

(A.3)
Thus, $p(\delta | Z_O, \theta [l])$ is approximately proportional to a Gaussian density. Evaluating
the proportionality constant on the right-hand side of (A.3) yields:

$$\int p(\delta | Z_O, \theta [l]) d\delta = \exp \left( L_c(\theta [l] | Z_O, \hat{\delta} [l]) \right) (2\pi)^{m/2} |Q_{LA}(\delta [l], \theta [l] | Z_O)|^{-1/2}. \quad (A.4)$$

Therefore, the density is approximately $\text{Gau}(\hat{\delta} [l], Q_{LA}(\delta [l], \theta [l] | Z_O)^{-1})$. So, approximately,

$$E(\delta | Z_O, \theta [l]) = \hat{\delta} [l] \quad \text{var}(\delta | Z_O, \theta [l]) = Q_{LA}(\delta [l], \theta [l] | Z_O)^{-1}. \quad (A.5)$$

Consider now $E \left( \exp \left\{ W(s_i) \times (S(s_i)^T \eta + \xi(s_i)) \right\} | Z_O, \theta [l] \right) \equiv E \left( \exp \left\{ W(s_i) \times (q(s_i)^T \delta) \right\} | Z_O, \theta [l] \right)$, for $i = 1, \ldots, n$. Then approximately,

$$E \left( \exp \left\{ W(s_i) \times (q(s_i)^T \delta) \right\} | Z_O, \theta [l] \right) = \exp \left\{ W(s_i) \times (q(s_i)^T \hat{\delta} [l]) \right\} + \frac{W(s_i)^2}{2} q(s_i)^T Q_{LA}(\delta [l], \theta [l] | Z_O)^{-1} q(s_i). \quad (A.6)$$

Recall that $\delta \equiv (\eta^T, \xi_{O}^T)^T$. Therefore, from (A.5) and (A.6), we obtain the approximations to the expectations, involved in the E-step of the EM algorithm, that were used in (2.20), (2.21), and (2.25).

### A.2 MCMC Algorithm

Here we describe the MCMC procedure that is used to obtain samples from the predictive distribution, $[\eta, \xi_O | Z_O, \beta, K, \sigma_{\xi}^2]$. We implement the MCMC procedure with a Gibbs sampler, incorporating Metropolis-Hastings steps where necessary. The full conditional distributions, as well as details of the Metropolis-Hastings steps, are described as follows.

The joint distribution, $[Z_O, \eta, \xi_O | \beta, K, \sigma_{\xi}^2]$, can be written as:

$$[Z_O, \eta, \xi_O | \beta, K, \sigma_{\xi}^2] = [Z_O | \eta, \xi_O, \beta] \times [\eta | K] \times [\xi_O | \sigma_{\xi}^2]. \quad (A.7)$$
Let “[A|B,·]” denote the full conditional distribution of the unknown A given B and all other unknowns (and the data). The Gibbs sampler uses the following steps to generate samples from the predictive distribution, \([\eta, \xi_0|Z_0, \beta, K, \sigma^2_\xi]\).

1. At \(t = 0\), we select starting values \(\eta^{[0]}\) and \(\xi_0^{[0]}\).

2. \(t=t+1\); simulate successively from the full conditionals, \([\eta^{[t+1]}|\xi_0^{[t]}, \cdot]\) and \([\xi_0^{[t+1]}|\eta^{[t+1]}, \cdot]\).

3. Repeat step 2 to generate as many samples as needed.

4. Discard an initial number of samples as “burn-in.”

The full conditionals are not available in closed form, so we use a Metropolis step within the Gibbs sampler. We describe a generic version of this below and then we give the particular case of sampling from the full conditionals, \([\eta^{[t+1]}|\xi_0^{[t]}, \cdot]\) and \([\xi_0^{[t+1]}|\eta^{[t+1]}, \cdot]\).

Suppose \(a\) is the random variable (or a block of random variables) that we are updating, and \(a_0\) is the most recently sampled value. We follow the steps below to obtain a new sample of \(a\):

1. Draw a trial value \(a_1\) from a proposal density, \(g(a)\).

2. Generate \(U_1\) uniformly on \((0, 1)\).

3. Compute the joint density of \(a\) and all other unknowns, \(l(a_0, rest)\) and \(l(a_1, rest)\) where “rest” denotes all the other unknowns fixed at their most recently sampled value.

4. If \(U_1 < \min\left\{\frac{l(a_1, rest)g(a_0)}{l(a_0, rest)g(a_1)}\right\}\), accept the new \(a_1\) and keep it for the most current iteration; otherwise, the value \(a_0\) is retained.

When sampling from \([\eta^{[t+1]}|\xi_0^{[t]}, \cdot]\), we update \(\eta\) as a block. For the proposal distribution, we use the formulation given in Chib and Greenberg (1995, 1996) and Chib and Winkelmann (2001). They suggest matching the proposal density to the target density at the mode.
by using a multivariate normal, or a mutivariate t-distribution as the proposal, for which the location is given by the mode of the target distribution, and the dispersion is given by the negative inverse of the Hessian matrix, evaluated at the mode. Let $\eta_{\text{mode}}$ denote the posterior mode of $[\eta|\xi_O, \cdot]$ and $\Sigma_\eta$ denote the negative inverse of the Hessian matrix, evaluated at the mode. The mode can be computed using a Newton-Raphson method (see Robert and Casella, 2004, Section 1.4). The proposal for $\eta$ is then taken as $\text{Gau}(\eta_{\text{mode}}, c_\eta \Sigma_\eta)$, where $c_\eta$ is a tuning parameter that is adjusted to control the acceptance rate. In the real-data application presented in Section 2.4, we held $c_\eta$ at 1.9 and achieved an acceptance rate of 24% for $\eta$, which is close to optimal (see Roberts et al., 1997).

For sampling $[\xi^{[t+1]}_O | \eta^{[t+1]}, \cdot]$, we update $\xi_O$ elementwise. Here we also compute the mode of $[\xi_O | \eta, \cdot]$, which we denote by $\xi_{\text{mode}}$, and the inverse of the negative Hessian matrix, evaluated at the mode, which we denote by $\Sigma_\xi$. For updating the $i$-th element of $\xi_O$ we use the proposal: $\text{Gau}\left((\xi_{\text{mode}})_i, c_\xi (\Sigma_\xi)_{i,i}\right)$, where $(\xi_{\text{mode}})_i$ denotes the $i$-th element of $\xi_{\text{mode}}$, and $(\Sigma_\xi)_{i,i}$ denotes the $(i,i)$-th element of the matrix $\Sigma_\xi$. In the real-data application presented in Section 2.4, we fixed $c_\xi = 10$, and achieved an acceptance rate between 37% and 40% for the different elements of $\xi_O$. 

132
Appendix B

APPENDIX FOR CHAPTER 3

B.1 Approximations Involved in the EM Algorithm

Let $\delta \equiv (\eta^T, \xi^T)^T$ be an $m$ ($m = r + n$)-dimensional vector. Here we derive the Laplace approximation to the density $[\delta | Z_O, \theta^{[l]}]$. Let $\hat{\delta}^{[l]}$ maximize the complete data log likelihood, $L_c(\theta^{[l]}|Z_O, \delta)$. Now, the density for the distribution of $[\delta | Z_O, \theta^{[l]}]$ is given by:

$$p(\delta|Z_O, \theta^{[l]}) \propto \exp \left( L_c(\theta^{[l]}|Z_O, \delta) \right).$$

(B.1)

A second-order Taylor-series approximation of $L_c(\theta^{[l]}|Z_O, \delta)$ around $\hat{\delta}^{[l]}$ yields:

$$L_c(\theta^{[l]}|Z_O, \delta) = L_c(\theta^{[l]}|Z_O, \hat{\delta}^{[l]}) + \frac{1}{2}(\delta - \hat{\delta}^{[l]})^T \left[ \frac{\partial^2}{\partial \delta^T \partial \delta} L_c(\theta^{[l]}|Z_O, \delta) \right]_{\delta = \hat{\delta}^{[l]}} (\delta - \hat{\delta}^{[l]})$$

+ higher-order terms

$$\approx L_c(\theta^{[l]}|Z_O, \hat{\delta}^{[l]}) - \frac{1}{2}(\delta - \hat{\delta}^{[l]})^T Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)(\delta - \hat{\delta}^{[l]}),$$

(B.2)

where $Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O) \equiv - \left[ \frac{\partial^2}{\partial \delta^T \partial \delta} L_c(\theta^{[l]}|Z_O, \delta) \right]_{\delta = \hat{\delta}^{[l]}}$. In (B.2) above, notice that the first-order linear term is zero since the first-order derivative of $L_c(\theta^{[l]}|Z_O, \delta)$ with respect to $\delta$, evaluated at $\delta = \hat{\delta}^{[l]}$, is zero (recall that $\hat{\delta}^{[l]}$ maximizes $L_c(\theta^{[l]}|Z_O, \delta)$). Therefore, for the density of $[\delta | Z_O, \theta^{[l]}]$, we have approximately,

$$p(\delta|Z_O, \theta^{[l]}) \propto \exp \left( L_c(\theta^{[l]}|Z_O, \hat{\delta}^{[l]}) \right)$$

$$\times \exp \left( -\frac{1}{2}(\delta - \hat{\delta}^{[l]})^T Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)(\delta - \hat{\delta}^{[l]}) \right).$$

(B.3)
Thus, \( p(\delta|Z_O, \theta^{[l]}) \) is approximately proportional to a Gaussian density. Evaluating the proportionality constant on the right-hand side of (B.3) yields the approximation:

\[
\int p(\delta|Z_O, \theta^{[l]}) d\delta = \exp \left( L_c(\theta^{[l]}|Z_O, \hat{\delta}^{[l]}) \right) (2\pi)^{m/2} |Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)|^{-1/2}, \tag{B.4}
\]

and hence the first two moments are approximately,

\[
E(\delta|Z_O, \theta^{[l]}) = \hat{\delta}^{[l]}, \tag{B.5}
\]
\[
\text{var}(\delta|Z_O, \theta^{[l]}) = Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)^{-1}.
\]

Next, for \( k = 1, 2 \), we derive the expectation:

\[
E \left( h_k \left( C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i) \right) \big| Z_O, \theta^{[l]} \right) \\
\equiv E \left( h_k \left( C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \delta \right) \big| Z_O, \theta^{[l]} \right). \tag{B.6}
\]

Using a second-order Taylor-series expansion of \( h_k(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \delta) \) around \( \hat{\delta}^{[l]} \), we obtain:

\[
h_k \left( C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \delta \right) \\
= h_k \left( C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}^{[l]} \right) \\
+ (\delta - \hat{\delta}^{[l]})^\top \left( h_k'(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}^{[l]}) \times q(s_i) \right) \\
+ \frac{1}{2} (\delta - \hat{\delta}^{[l]})^\top \left( h_k'' \left( C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}^{[l]} \right) \times q(s_i)q(s_i)^\top \right) (\delta - \hat{\delta}^{[l]}) \\
+ \text{higher-order terms}, \tag{B.7}
\]

where the vector \( h_k'(x_0) \equiv \frac{d}{dx} h_k(x)|_{x=x_0} \), and the matrix \( h_k''(x_0) \equiv \frac{d^2}{dx^2} h_k(x)|_{x=x_0} \).
Taking expectations, we obtain:

\[
E\left(h_k(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \delta) | Z_O, \theta[l]\right)
\approx h_k\left(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right)

+ E\left((\delta - \hat{\delta}[l]) | Z_O, \theta[l]\right)^\top \left(h_k'\left(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right) \times q(s_i)\right)

+ \frac{1}{2} \text{tr}\left(E\left((\delta - \hat{\delta}[l])(\delta - \hat{\delta}[l])^\top | Z_O, \theta[l]\right)\right)

\times \left(h_k''\left(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right) \times q(s_i)q(s_i)^\top\right) \right). \tag{B.8}
\]

The second term in (B.8) is zero, since \(\hat{\delta}[l]\) is the expectation of the Gaussian density that approximates the posterior density, \(\delta | Z_O, \theta[l]\); see (B.5). Consequently, we obtain:

\[
E(h_k(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \delta) | Z_O, \theta[l])
\approx h_k\left(X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right)

+ \frac{1}{2} \text{tr}\left(Q_{LA}(\delta[l], \theta[l] | Z_O)^{-1}\left(h_k''\left(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right) \times q(s_i)q(s_i)^\top\right)\right)

= h_k\left(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right)

+ \frac{1}{2} h_k''\left(C(s_i) + X(s_i)^\top \beta + q(s_i)^\top \hat{\delta}[l]\right) \times q(s_i)^\top Q_{LA}(\delta[l], \theta[l] | Z_O)^{-1} q(s_i). \tag{B.9}
\]

Recall that \(\delta \equiv (\eta^\top, \xi^\top)^\top\). Therefore, from (B.5) and (B.9), we obtain the approximations to the expectations involved in the E-step of the EM algorithm, that are used in (3.26), (3.27), and (3.33).

### B.2 Evaluations for the One-Step Newton-Raphson Update for \(\beta\)

In this part of the Appendix, we evaluate the expressions involved in the one-step Newton-Raphson update for \(\beta\), which was discussed at the end of Section 3.4.2. Specifically, we will evaluate the score function \(R(\theta)\) and its derivative with respect to \(\beta\), assuming as many derivatives for \(h_1(\cdot)\) and \(h_2(\cdot)\) as necessary.
The expression for $Q(\cdot, \cdot)$ given by (3.25), after substituting in the approximations to the required expectations, becomes:

$$
Q(\theta, \theta^{[\ell]}) = \text{const.} + \left\{ \sum_{i=1}^{n} Z(s_i) \left\{ h_1(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \right. \\
+ \frac{1}{2} h_1''(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \times q(s_i)^\top Q_{LA}(\delta^{[\ell]}, \theta^{[\ell]}|Z_O)^{-1} q(s_i) \right\} \\
- \sum_{i=1}^{n} \left\{ h_2(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \right. \\
+ \frac{1}{2} h_2''(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \times q(s_i)^\top Q_{LA}(\delta^{[\ell]}, \theta^{[\ell]}|Z_O)^{-1} q(s_i) \right\} / \tau^2 \\
- \frac{1}{2} \log |\mathbf{K}| - \frac{1}{2} \text{trace} \left( \hat{E} \left( \eta \eta^\top |Z_O, \theta^{[\ell]} \right) \mathbf{K}^{-1} \right) \\
- \frac{n}{2} \log \sigma_{\xi}^2 - \frac{1}{2\sigma_{\xi}^2} \text{trace} \left( \hat{E} \left( \xi \xi^\top |Z_O, \theta^{[\ell]} \right) \mathbf{V}_{\xi,\xi}^{-1} \right),
$$

(B.10)

where $q(s)$ and $Q_{LA}(\delta^{[\ell]}, \theta^{[\ell]}|Z_O)$ are defined in Appendix A; the approximations, $\hat{E} \left( \eta \eta^\top |Z_O, \theta^{[\ell]} \right)$ and $\hat{E} \left( \xi \xi^\top |Z_O, \theta^{[\ell]} \right)$, to the respective expectations, are given by (3.32) (which follows from Appendix B.1).

Now, to obtain the score function, $\mathbf{R} (\theta)$, we differentiate (B.10) with respect to $\beta$, resulting in:

$$
\mathbf{R}(\theta) = \left\{ \sum_{i=1}^{n} Z(s_i) \left\{ h_1'(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \right. \\
+ \frac{1}{2} h_1''(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \times q(s_i)^\top Q_{LA}(\delta^{[\ell]}, \theta^{[\ell]}|Z_O)^{-1} q(s_i) \right\} X(s_i) \\
- \sum_{i=1}^{n} \left\{ h_2'(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \right. \\
+ \frac{1}{2} h_2''(C(s_i) + X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^{[\ell]} + \xi^{[\ell]}(s_i)) \times q(s_i)^\top Q_{LA}(\delta^{[\ell]}, \theta^{[\ell]}|Z_O)^{-1} q(s_i) \right\} / \tau^2 \\
$$

(B.11)

The Newton-Raphson update (3.35) also requires the partial derivative of $\mathbf{R}(\theta)$ with
samples from the predictive distribution, \( \mathbf{R}(\theta) \) incorporates Metropolis-Hastings steps where necessary. The full conditional distributions, as well as details of the Metropolis Hastings steps, are described in the following paragraph.

The joint distribution, \( [\mathbf{Z}_O, \eta, \xi_O|\theta] \), can be written as:

\[
[Z_O, \eta, \xi_O|\theta] \equiv [Z_O|\eta, \xi_O, \beta] \times [\eta|K] \times [\xi_O|\sigma^2_\xi].
\] (B.13)

Let “\([A|B, \cdot]\)” denote the full conditional distribution of the unknown \( A \) given \( B \) and all other unknowns (and the data). The Gibbs sampler uses the following steps to generate samples from the predictive distribution, \( [\eta, \xi_O|Z_O, \theta] \).

1. At \( t = 0 \), we select starting values \( \eta^{[0]} \) and \( \xi_O^{[0]} \).

2. \( t = t + 1 \); simulate successively from the full conditionals, \( [\eta^{[r+1]}|\xi_O^{[r]}, \cdot] \) and \( [\xi_O^{[r+1]}|\eta^{[r+1]}, \cdot] \).
3. Repeat step 2 to generate as many samples as needed.

4. Discard an initial number of samples as “burn-in.”

The full conditionals are not available in closed form, so we use a Metropolis-Hastings step within the Gibbs sampler. A generic version of the algorithm that we have used to draw samples from the full conditionals, \([\eta^{[t+1]}|\xi^{[t]},\cdot]\) and \([\xi^{[t+1]}|\eta^{[t+1]},\cdot]\) (at the \((t + 1)\)-th stage), is discussed below. Suppose \(a\) is the random variable (or a block of random variables) that we are updating, and \(a_0\) is the most recently sampled value. We follow the steps below to obtain a new sample of \(a\):

1. Draw a trial value \(a_1\) from a proposal density, \(\text{Gau}(a_0, \Sigma_a)\).

2. Generate \(U_1\) uniformly on \((0, 1)\).

3. Compute the joint density of \(a\) and all other unknowns, \(l(a_0, \text{rest})\) and \(l(a_1, \text{rest})\) where “rest” denotes all the other unknowns fixed at their most recently sampled value.

4. If \(U_1 < \min\left\{\frac{l(a_1, \text{rest})}{l(a_0, \text{rest})}, 1\right\}\), accept the trial value \(a_1\) and keep it for the most current iteration; otherwise, the value \(a_0\) is retained.

When sampling from \([\eta^{[t+1]}|\xi^{[t]},\cdot]\), we update \(\eta\) as a block. To sample from \([\xi^{[t+1]}|\eta^{[t+1]},\cdot]\), we update \(\xi\) elementwise.

### B.4 BHM: Prior Specifications and the MCMC Algorithm

In this part of the Appendix, we present the prior distributions (or the parameter model) of BHM and fully Bayesian inference using the MCMC algorithm.

Following Kang and Cressie (2011), the prior distribution of \(\theta = (\beta, K, \sigma^2_\xi)\) is assumed to be made up of mutually independent components:

\[
[\beta, K, \sigma^2_\xi] = [\beta] \cdot [K] \cdot [\sigma^2_\xi].
\]  
(B.14)
Next we assume that the \( p \)-dimensional fixed-effects parameters, \( \mathbf{\beta} \), have a Gaussian prior distribution,

\[
\mathbf{\beta} \sim \text{Gau}(\mathbf{\mu}_\beta, \Sigma_\beta),
\]

(B.15)

where \( \mathbf{\mu}_\beta \) and \( \Sigma_\beta \equiv \text{diag}(\sigma_{\beta,1}^2, \ldots, \sigma_{\beta,p}^2) \) are known hyperparameters. For fine-scale-variance parameter \( \sigma_\xi^2 \), we assume that \( \sigma_\xi \sim \text{Uniform}(0, \kappa_\xi) \), where \( \kappa_\xi \) is a known hyperparameter.

Finally, the prior distribution on \( \mathbf{K} \) is based on the spectral decomposition,

\[
\mathbf{K} = \mathbf{P} \Lambda \mathbf{P}^\top,
\]

(B.16)

where \( \Lambda \equiv \text{diag}(\lambda_1, \ldots, \lambda_r) \), \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0 \), and \( \mathbf{P} \) is an orthogonal matrix that can be parametrized in terms of the \( r(r-1)/2 \) Givens angles,

\[
\mathbf{\theta}_G \equiv \{ \theta_{ij} : i = 1, \ldots, r-1, j = i+1, \ldots, r \}.
\]

In terms of these Givens angles, we can write \( \mathbf{P} \) as (e.g., Kang and Cressie, 2011):

\[
\mathbf{P} = (\mathbf{G}_{12}\mathbf{G}_{13} \ldots \mathbf{G}_{1r})(\mathbf{G}_{23} \ldots \mathbf{G}_{2r}) \ldots \mathbf{G}_{(r-1)r},
\]

where \( \mathbf{G}_{ij} \) is the Givens rotation matrix corresponding to the Givens angle \( \theta_{ij} \), which is obtained by modifying the \( r \times r \) identity matrix as follows: The \( i \)-th and the \( j \)-th diagonal elements of 1 are both replaced by \( \cos(\theta_{ij}) \), and the \( (i, j) \)-th and \( (j, i) \)-th elements of 0 are replaced by \( -\sin(\theta_{ij}) \) and \( \sin(\theta_{ij}) \), respectively.

We assign priors to the eigenvalues \( \{\lambda_i : i = 1, \ldots, r\} \) and the Givens angles \( \mathbf{\theta}_G \), using models discussed in Kang and Cressie (2011). That is,

\[
[\lambda_1, \ldots, \lambda_r] = [\lambda_{1,1}, \ldots, \lambda_{1,q_1}] \cdots [\lambda_{K,1}, \ldots, \lambda_{K,q_K} | \lambda_{K-1,q_{K-1}}],
\]

(B.17)

where \( \lambda_{k,1}, \ldots, \lambda_{k,q_k} \) are the eigenvalues corresponding to the \( q_k \) basis functions from the \( k \)-th resolution, \( k = 1, \ldots, K \), and \( \sum_{k=1}^{K} q_k = r \). Finally, \( \lambda_{k,1}, \ldots, \lambda_{k,q_k} \) are assumed to be distributed as order statistics corresponding to i.i.d. truncated Lognormal random variables with known hyperparameters, mean \( \mu_k \) and variance \( \sigma_k^2 \), for \( k = 1, \ldots, K \), where the Lognormal distribution is restricted to \( (0, \lambda_{k-1,q_{k-1}}) \).
We define the prior on $\theta_{ij}$ through a prior on the logit transformation of $\theta_{ij}$, namely

$$h(\theta_{ij}) \equiv \log \left( \frac{\pi/2 + \theta_{ij}}{\pi/2 - \theta_{ij}} \right).$$  \tag{B.18}$$

Then we assign independent priors on $h(\theta_{ij})$ as

$$h(\theta_{ij}) \sim \text{Gau}(c_k, \tau^2_k),$$  \tag{B.19}$$

if $i, j$ both belong to the same resolution $k$, where $k = 1, \ldots, K$; otherwise,

$$h(\theta_{ij}) \sim \text{Gau}(0, \tau^2_0),$$  \tag{B.20}$$

if $i, j$ belong to different resolutions. The hyperparameters $\{c_k\}, \{\tau^2_k\}$, and $\tau^2_0$ are assumed known.

We also specify the hyperparameters following the recommendations in Kang and Cressie (2011). In the simulation study described in this article, the true parameter values, $\theta_T$, were used to specify the hyperparameters. We selected $\mu_\beta = \beta_T$, and the elements of the covariance matrix $\Sigma_\beta$ were specified as three times the square of the standard-errors obtained by fitting a classical fixed-effects Poisson GLM (e.g., McCullagh and Nelder, 1989, Chapter 6) to the data, with the same covariates that were used for the simulation.

Next we chose $\kappa_\xi = 10\sigma_{\xi:T}$.

Finally, to specify the hyperparameters in the prior on $K$, we first obtained:

$$K_T = P_T \Lambda_T P_T^T,$$

where $\Lambda_T \equiv (\lambda_{1:T}, \ldots, \lambda_{r:T})$. We also computed the Givens angles for $K_T$, namely,

$$\{\theta_{ij:T} : i = 1, \ldots, r-1; j = i+1, \ldots, r\}.$$

For $k = 1, \ldots, K$, we specified:

$$\mu_k = \frac{q_k}{\sum_{i=1}^{q_k} \log(\lambda_{k:i:T}) / q_k}$$

$$\sigma^2_k = \frac{q_k}{\sum_{i=1}^{q_k} (\log(\lambda_{k:i:T} - \mu_k)^2 / (q_k - 1)).$$  \tag{B.21}$$
Similarly, we specified \( \{ c_k \} \), \( \{ \tau_k^2 \} \), and \( \tau_0^2 \) as:

\[
    c_k = \sum_{(i,j) \in N_k} h(\theta_{ij:T}) / |N_k|,
\]

\[
    \tau_k^2 = \sum_{(i,j) \in N_k} (h(\theta_{ij:T}) - c_k)^2 / (|N_k| - 1),
\]

\[
    \tau_0^2 = \sum_{(i,j) \in N_0} h(\theta_{ij:T})^2 / |N_0|,
\]

where \( h(\cdot) \) is given by (B.18), \( N_k \equiv \{(i, j) : \text{the } i \text{-th and the } j \text{-th basis functions are both of the } k \text{-th resolution}\}, k = 1, \ldots, K \), and \( N_0 \equiv \{(i, j) : \text{the } i \text{-th and the } j \text{-th basis functions are of different resolutions}\} \).

Finally, we implemented the MCMC procedure with a Gibbs sampler to generate samples from the posterior distribution, \( \eta, \xi_O, \xi_U, \theta | Z_O \). The full conditionals of \( \sigma^2_\xi \) and \( \xi_U \) can be derived in closed form. The full conditional of \( \xi_U \) is:

\[
    [\xi_U | Z_O, \eta, \xi_O, \theta] = [\xi_U | \theta].
\]

The full conditional of \( \sigma^2_\xi \) is a truncated Inverse-Gamma distribution, namely, \( \text{IG}((N - 1)/2, \xi^\top \xi/2) \cdot I(0 < \sigma_\xi < k) \) (see Kang and Cressie, 2011), where recall that \( \xi = (\xi_O^\top, \xi_U^\top)^\top \). The other full conditionals are not available in closed form, so we incorporated a Metropolis-Hastings step, with random walk proposals, to simulate from them. Details of the Metropolis-Hastings algorithm is given in Appendix B.3. We updated \( \beta \) and \( \eta \) in blocks, and \( \xi_O \) elementwise. When sampling the eigenvalues, we updated in blocks according to resolution. If the total ordering of the eigenvalues was broken, we rejected the sample and a new sample was drawn until the ordering of the eigenvalues was preserved (Kang and Cressie, 2011). When sampling the Givens angles, we updated the Givens angles corresponding to the same resolution, \( \{ \theta_{ij} : (i, j) \in N_k \} \), as a block, for \( k = 1, \ldots, K \), and the Givens angles \( \{ \theta_{ij} : (i, j) \in N_0 \} \) were updated as a block.
Appendix C
APPENDIX FOR CHAPTER 4

C.1 Approximations Involved in the EM Algorithm

Let \( \delta \equiv (\eta^\top, \xi^\top)^\top \) be an \( m \) \((m = r + n)\) dimensional vector. Here we derive the Laplace approximation to the density \([\delta|Q_O, \theta^{[l]}]\). Let \( \delta^{[l]} \) maximize \( Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \) (see (4.15)). Recall that \( Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \) was obtained by integrating out the clear/cloudy latent variable \( W_O \) from the complete data log-likelihood. Now, the density for the distribution of \([\delta|Q_O, \theta^{[l]}]\) is given by:

\[
p(\delta|Q_O, \theta^{[l]}) \propto \exp \left( Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \right). \tag{C.1}
\]

A Taylor-series approximation of \( Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \) around \( \hat{\delta}^{[l]} \) yields:

\[
Q_1(\theta^{[l]}, \theta^{[l]}, \delta) = Q_1(\theta^{[l]}, \theta^{[l]}, \hat{\delta}^{[l]}) + \frac{1}{2} (\delta - \hat{\delta}^{[l]})^\top \left( \frac{\partial^2}{\partial \delta^\top \partial \delta} Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \right)_{\delta = \hat{\delta}^{[l]}} (\delta - \hat{\delta}^{[l]}) + \text{higher-order terms} \approx Q_1(\theta^{[l]}, \theta^{[l]}, \hat{\delta}^{[l]}) - \frac{1}{2} (\delta - \hat{\delta}^{[l]})^\top P_{LA}(\delta^{[l]}, \theta^{[l]}|Q_O)(\delta - \hat{\delta}^{[l]}), \tag{C.2}
\]

where \( P_{LA}(\delta^{[l]}, \theta^{[l]}|Q_O) \equiv - \left( \frac{\partial^2}{\partial \delta^\top \partial \delta} Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \right)_{\delta = \hat{\delta}^{[l]}} \). In (C.2) above, notice that the first-order linear term is zero since the first-order derivative of \( Q_1(\theta^{[l]}, \theta^{[l]}, \delta) \), with respect
to \( \delta \), evaluated at \( \delta = \hat{\delta}^{[l]} \), is zero (recall that \( \hat{\delta}^{[l]} \) maximizes \( Q_1(\theta^{[l]}, \theta^{[l]}; \delta) \)). Therefore, for the density of \( [\delta|Q_O, \theta^{[l]}] \), we have approximately,

\[
p(\delta|Q_O, \theta^{[l]}) \propto \exp \left( Q_1(\theta^{[l]}, \theta^{[l]}; \hat{\delta}^{[l]}) \right) \\
\times \exp \left( -\frac{1}{2} (\delta - \hat{\delta}^{[l]})^\top P_LA(\delta^{[l]}, \theta^{[l]}|Q_O)(\delta - \hat{\delta}^{[l]}) \right). \tag{C.3}
\]

Thus, \( p(\delta|Q_O, \theta^{[l]}) \) is approximately proportional to a Gaussian density. Evaluating the proportionality constant on the right-hand side of (C.3) yields:

\[
\int p(\delta|Q_O, \theta^{[l]}) d\delta = \exp \left( Q_1(\theta^{[l]}, \theta^{[l]}; \hat{\delta}^{[l]}) \right) (2\pi)^{m/2} |P_LA(\delta^{[l]}, \theta^{[l]}|Q_O)|^{-1/2}, \tag{C.4}
\]

which means that the density is approximately \( \text{Gau}(\delta^{[l]}, P_LA(\delta^{[l]}, \theta^{[l]}|Q_O)^{-1}) \). That is, approximately,

\[
E(\delta|Q_O, \theta^{[l]}) = \hat{\delta}^{[l]} \\
\text{var}(\delta|Q_O, \theta^{[l]}) = P_LA(\delta^{[l]}, \theta^{[l]}|Q_O)^{-1}. \tag{C.5}
\]

The remaining expectations required in the E-step are:

\[
E \left( \log \left( 1 + \exp \left\{ X(s_i) \tau^T + S(s_i) \tau^T + \xi(s_i) \right\} |Q_O, \theta^{[l]} \right) \right),
\]

and

\[
E \left\{ \left( \frac{p^{[l]}(s_i; \eta, \xi_O) g_1^{[l]}(Q(s_i))}{p^{[l]}(s_i; \eta, \xi_O) g_1^{[l]}(Q(s_i)) + (1 - p^{[l]}(s_i; \eta, \xi_O)) g_0^{[l]}(Q(s_i))} \right) \right. \\
\left. \times \left( X(s_i) \tau^T + S(s_i) \tau^T + \xi(s_i) \right) |Q_O, \theta^{[l]} \right\} \right.
\]

\[
= E \left\{ \left( \frac{\exp \left\{ X(s_i) \tau^T + S(s_i) \tau^T + \xi(s_i) \right\} g_1^{[l]}(Q(s_i))}{\exp \left\{ X(s_i) \tau^T + S(s_i) \tau^T + \xi(s_i) \right\} g_1^{[l]}(Q(s_i)) + g_0^{[l]}(Q(s_i))} \right) \right. \\
\left. \times \left( X(s_i) \tau^T + S(s_i) \tau^T + \xi(s_i) \right) |Q_O, \theta^{[l]} \right\} \right.
\]

where \( i = 1, \ldots, n \). First we will derive a general expression for

\[
E \left\{ b \left( X(s_i) \tau^T + S(s_i) \tau^T + \xi(s_i) \right) |Q_O, \theta^{[l]} \right\} \\
\equiv E \left( b \left( X(s_i) \tau^T + q(s_i) \tau^T \delta \right) |Q_O, \theta^{[l]} \right).
\]

143
where $b(\cdot)$ is a generic function. Using a Taylor-series expansion of $b(X(s_i)\top \beta[l] + q(s_i)\top \delta)$ around $\hat{\delta}^{[l]}$, we obtain:

$$b(X(s_i)\top \beta[l] + q(s_i)\top \delta)$$

$$= b(X(s_i)\top \beta[l] + q(s_i)\top \hat{\delta}^{[l]})$$

$$+ (\delta - \hat{\delta}^{[l]})\top \left( b'(X(s_i)\top \beta[l] + q(s_i)\top \hat{\delta}^{[l]}) \times q(s_i) \right)$$

$$+ \frac{1}{2} (\delta - \hat{\delta}^{[l]})\top \left( b''(X(s_i)\top \beta[l] + q(s_i)\top \hat{\delta}^{[l]}) \times q(s_i)q(s_i)\top \right) (\delta - \hat{\delta}^{[l]})$$

$$+ \text{higher-order terms} \quad (C.6)$$

where $b'(x_0) \equiv \frac{d}{dx} b(x) \big|_{x=x_0}$, and $b''(x_0) \equiv \frac{d^2}{dx} b(x) \big|_{x=x_0}$.

Taking expectations, we obtain:

$$E \left\{ b(X(s_i)\top \beta[l] + q(s_i)\top \delta) \mid Q_O, \theta[l] \right\}$$

$$\approx b(X(s_i)\top \beta[l] + q(s_i)\top \hat{\delta}^{[l]})$$

$$+ E \left( (\delta - \hat{\delta}^{[l]}) \mid Z, \theta[l] \right)\top \left( b'(X(s_i)\top \beta[l] + q(s_i)\top \hat{\delta}^{[l]}) \times q(s_i) \right)$$

$$+ \frac{1}{2} \text{trace} \left( E \left( (\delta - \hat{\delta}^{[l]}) (\delta - \hat{\delta}^{[l]})\top \mid Z, \theta[l] \right) \right) \times \left( b''(X(s_i)\top \beta[l] + q(s_i)\top \hat{\delta}^{[l]}) \times q(s_i)q(s_i)\top \right) \). \quad (C.7)$$
The second term on the right-hand side of (C.7) is zero, since $\hat{\delta}^{[l]}$ is the expectation of the Gaussian density that approximates the posterior density $[\delta|Q_O, \theta^{[l]}]$ (see (C.5)). Consequently, we obtain:

$$E \left( b (X(s_i)^\top \beta^{[l]} + q(s_i)^\top \delta | Z, \theta^{[l]}) \right)$$

$$\approx b(X(s_i)^\top \beta^{[l]} + q(s_i)^\top \delta^{[l]}$$

$$+ \frac{1}{2} \text{trace} \left( Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)^{-1} \left( b''(X(s_i)^\top \beta^{[l]} + q(s_i)^\top \delta^{[l]} \times q(s_i)q(s_i)^\top) \right) \right)$$

$$= b(X(s_i)^\top \beta^{[l]} + q(s_i)^\top \delta^{[l]}$$

$$+ \frac{1}{2} b''(X(s_i)^\top \beta^{[l]} + q(s_i)^\top \delta^{[l]} \times q(s_i)^\top \delta^{[l]} - q(s_i)^\top \delta^{[l]} \times q(s_i)^\top)q(s_i)^\top) - q(s_i)$$

$$= \frac{1}{2} Q_{LA}(\delta^{[l]}, \theta^{[l]}|Z_O)^{-1} - q(s_i).$$

(C.8)

Now, to obtain the remaining expectations in the E-step, we will repeatedly use the formula derived above in (C.8). Let us consider a function $h(x)$ defined as:

$$h(x) = \log(1 + \exp(x)).$$

Taking the first derivative, we obtain:

$$h'(x) = \frac{\exp(x)}{1 + \exp(x)},$$

(C.9)

and taking the second derivative, we obtain:

$$h''(x) = \frac{\exp(x)}{(1 + \exp(x))^2}.$$  

(C.10)

Using (C.10) in (C.8), we obtain the required expectation as:

$$E \left\{ \log \left( 1 + \exp \left\{ X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta + \xi(s_i) \right\} \right) \right\}$$

$$\approx \log \left( 1 + \exp \left\{ X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta^{[l]} + \hat{\xi}^{[l]}(s_i) \right\} \right)$$

$$+ \frac{1}{2} \text{exp} \left\{ X(s_i)^\top \beta^{[l]} + S(s_i)^\top \eta^{[l]} + \hat{\xi}^{[l]}(s_i) \right\}$$

$$\times \left( S(s_i)^\top \text{var}(\eta|Q_O, \theta^{[l]})S(s_i) + 2S(s_i)^\top \text{cov}(\eta, \xi_O|Q_O, \theta^{[l]})e(s_i)$$

$$+ e(s_i)^\top \text{var}(\xi_O|Q_O, \theta^{[l]})e(s_i) \right\}. $$

(C.11)
Now we consider a function $h(x)$ defined as:

$$h(x) = \frac{c_1 \exp(x)}{c_0 + c_1 \exp(x)}.$$  

Taking the first-derivative, we obtain:

$$h'(x) = \frac{c_0 c_1 \exp(x)}{(c_0 + c_1 \exp(x))^2}.$$  (C.12)

and taking the second-derivative, we obtain:

$$h''(x) = \frac{c_0 c_1 \exp(x)(c_0 - c_1 \exp(x))}{(c_0 + c_1 \exp(x))^3}.$$  (C.13)

Using (C.13) in (C.8), we obtain the required expectation as:

$$E\left\{ \frac{p[l](s_i; \eta, \xi_O)g[1](Q(s_i))}{p[l](s_i; \eta, \xi_O)g[1](Q(s_i)) + (1 - p[l](s_i; \eta, \xi_O))g[0](Q(s_i))} \right.$$  

$$\times \left\{ \left( X(s_i)^\top \beta + S(s_i)^\top \eta + \xi(s_i) \right) | Q_O, \theta[l] \right\}$$  

$$\approx \left\{ \frac{p[l](s_i; \hat{\eta}^l, \hat{\xi}_O)g[1](Q(s_i))}{p[l](s_i; \hat{\eta}^l, \hat{\xi}_O)g[1](Q(s_i)) + (1 - p[l](s_i; \hat{\eta}^l, \hat{\xi}_O))g[0](Q(s_i))} \right.$$

$$\times \left\{ \left( X(s_i)^\top \beta + S(s_i)^\top \hat{\eta}^l + \hat{\xi}(s_i) \right) | Q_O, \theta[l] \right\}$$

$$+ \frac{1}{2} \left( \frac{g[0](Q(s_i)) - g[1](Q(s_i)) \exp(X(s_i)^\top \beta[l] + S(s_i)^\top \hat{\eta}^l + \hat{\xi}(s_i))}{g[0](Q(s_i)) + g[1](Q(s_i)) \exp(X(s_i)^\top \beta[l] + S(s_i)^\top \hat{\eta}^l + \hat{\xi}(s_i))} \right)^3$$

$$\times \left( g[0](Q(s_i))g[1](Q(s_i)) \exp(X(s_i)^\top \beta[l] + S(s_i)^\top \hat{\eta}^l + \hat{\xi}(s_i)) \right) \times (X(s_i)^\top \beta)$$

$$\times \left( S(s_i)^\top \text{var}(\eta|Q_O, \theta[l])S(s_i) + 2S(s_i)^\top \text{cov}(\eta, \xi_O|Q_O, \theta[l])e(s_i) \right)$$

$$+ e(s_i)^\top \text{var}(\xi_O|Q_O, \theta[l])e(s_i)$$

$$+ \text{terms independent of } \theta,$$  (C.14)

where $p[l](s_i; \hat{\eta}^l, \hat{\xi}_O)$ is defined as:

$$p[l](s_i; \hat{\eta}^l, \hat{\xi}_O) \equiv \frac{\exp(X(s_i)^\top \beta[l] + S(s_i)^\top \hat{\eta}^l + \hat{\xi}(s_i))}{1 + \exp(X(s_i)^\top \beta[l] + S(s_i)^\top \hat{\eta}^l + \hat{\xi}(s_i))},$$

and $e(s_i)$ is a vector of length $n$ that has 1 at the $i$-th position and all other entries in $e(s_i)$ are zero, $i = 1, \ldots, n$. The terms independent of $\theta$ do not play any role in the M-step, and hence they are not computed.
C.2 MCMC Algorithm

The joint distribution, \( [Q_O, W, \eta, \xi|\beta, K, \sigma^2_\xi] \), can be written as:

\[
[Q_O, W, \eta, \xi|\beta, K, \sigma^2_\xi] \equiv [Q_O|W, \eta, \xi, \beta] \times [W|\eta, \xi, \beta] \times [\eta|K] \times [\xi|\sigma^2_\xi]. \quad \text{(C.15)}
\]

Let “\([A|B, \cdot]\)” denote the full conditional distribution of the unknown \( A \) given \( B \) and all other unknowns (and the data). The Gibbs sampler uses the following steps to generate samples from the predictive distribution, \([W, \eta, \xi|Q_O, \beta, K, \sigma^2_\xi]\).

1. At \( t = 0 \), select starting values \( W[0], \eta[0] \) and \( \xi[0] \).
2. \( t = t+1 \); simulate successively from the full conditionals, \([W[t+1]|\eta[t], \xi[t], \cdot]\), \([\eta[t+1]|W[t+1], \xi[t], \cdot]\) and \([\xi[t+1]|W[t+1], \eta[t+1], \cdot]\).
3. Repeat step 2 to generate as many samples as needed.
4. Discard an initial number of samples as “burn-in.”

The full conditional, \([W[t+1]|\eta[t], \xi[t], \cdot]\), is available in closed form, and it is straightforward to simulate from this distribution. The full conditionals, \([\eta[t+1]|W[t+1], \xi[t], \cdot]\) and \([\xi[t+1]|W[t+1], \eta[t+1], \cdot]\), are not available in closed form, so we use a Metropolis step within the Gibbs sampler. A generic version of this Metropolis algorithm is now presented.

Suppose \( a \) is the random variable (or a block of random variables) that we are updating, and \( a_0 \) is the most recently sampled value. We follow the steps below to obtain a new sample of \( a \):

1. Draw a trial value \( a_1 \) from a proposal density, \( g(a) \).
2. Generate \( U_1 \) uniformly on \((0, 1)\).
3. Compute the joint density of \( a \) and all other unknowns, \( l(a_0, \text{rest}) \) and \( l(a_1, \text{rest}) \) where “rest” denotes all the other unknowns fixed at their most recently sampled value.
4. If $U_1 < \min \left\{ \frac{f(a_1, \text{rest})g(a_0)}{f(a_0, \text{rest})g(a_1)} \right\}$, accept the new $a_1$ and keep it for the most current iteration; otherwise, the value $a_0$ is retained.

When sampling from $[\eta^{[r+1]}|\xi^{[r]}, \cdot]$, we update $\eta$ as a block. For the proposal, we use a mixture distribution (see Tierney, 1994). We use a mixture of the random walk proposal (see Section 7.5 Robert and Casella, 2004) and a tailored multivariate normal proposal (see Chib and Greenberg, 1995, 1996). Let $\eta_{\text{mode}}$ denote the posterior mode of $[\eta|\xi, \cdot]$ and $\Sigma_\eta$ denote the negative inverse of the Hessian matrix, evaluated at the mode. The mode can be computed using a Newton-Raphson method (see Robert and Casella, 2004, Section 1.4). The tailored proposal for $\eta$ was then taken as $\text{Gau}(\eta_{\text{mode}}, c_{1,\eta} \Sigma_\eta)$, where $c_{1,\eta}$ is a tuning parameter that is adjusted to control the acceptance rate. The random walk proposal was taken to be $\text{Gau}(\eta^{[r]}, c_{2,\eta} \Sigma_\eta)$, where again $c_{2,\eta}$ is a tuning parameter. The mixing proportion for the two distributions can also be tuned to achieve the desired control rate. We worked with a mixing proportion of 0.5 and selected $c_{1,\eta} = 1$ and $c_{2,\eta} = 2.3$. With these values for the tuning parameters, we achieved an acceptance rate of 22.1% for $\eta$, which is close to optimal (see Roberts et al., 1997).

For sampling $[\xi^{[r+1]}|\eta^{[r+1]}, \cdot]$, we update $\xi$ elementwise. Here we work with the tailored proposal. We compute the mode of $[\xi|\eta, \cdot]$ which we denote by $\xi_{\text{mode}}$, and the inverse of the negative Hessian matrix, evaluated at the mode, which we denote by $\Sigma_\xi$. For updating the $i$-th element of $\xi$, we use the proposal: $\text{Gau}\left( (\xi_{\text{mode}})_i, c_\xi (\Sigma_\xi)_{i,i} \right)$, where $(\xi_{\text{mode}})_i$ denotes the $i$-th element of $\xi_{\text{mode}}$, and $(\Sigma_\xi)_{i,i}$ denotes the $(i,i)$-th element of the matrix $\Sigma_\xi$. We fixed $c_\xi = 8$, and achieved an acceptance rate between 40.53% and 46.13% for the different elements of $\xi$. 

148


Duncan, W. J. (1944). “Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices).” *The London,


155


