REDUCED-DIMENSIONAL NON-GAUSSIAN SPATIAL MODELS
AND STATISTICAL METHODS FOR STUDYING THE WEST
ANTARCTIC ICE SHEET

A Dissertation in
Statistics
by
Yawen Guan

© 2017 Yawen Guan

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2017
The dissertation of Yawen Guan was reviewed and approved* by the following:

Murali Haran
Professor of Statistics
Dissertation Advisor
Chair of Committee

Francesca Chiaromonte
Professor of Statistics

Ephraim M. Hanks
Assistant Professor of Statistics

Chris E. Forest
Associate Professor of Climate Dynamics, Department of Meteorology and Atmospheric Science and Department of Geosciences

Aleksandra Slavkovic
Professor of Statistics
Chair of Graduate Program

*Signatures are on file in the Graduate School.
Abstract

My dissertation research focuses on developing efficient statistical methods for high-dimensional non-Gaussian spatial data, and hierarchical models for the West Antarctic Ice Sheet (WAIS). My thesis consists of three projects: (1) A fast projection-based approach to model high-dimensional non-Gaussian data on a continuous spatial domain. This involves developing a novel low-dimensional reparameterization of a spatial generalized linear mixed model (SGLMM). (2) A fast maximum likelihood inferential approach for high-dimensional non-Gaussian spatial data. This involves developing an automated Markov chain Monte Carlo Expectation Maximization algorithms for two reparameterized SGLMMs, one for discrete-domain Gaussian Markov random field models, and the other for continuous-domain Gaussian process spatial models. (3) A rigorous method for inferring important characteristics of the WAIS by combining multiple data sources while respecting physical constraints. Therefore, modeling and computational approaches for high-dimensional non-Gaussian spatial data are the underlying theme for all three projects.
# Table of Contents

| List of Figures                  | vii |
| List of Tables                  | xii |
| Acknowledgments                 | xiii|

## Chapter 1
Introduction

1.1 Models for Large Non-Gaussian Spatial Data ........................................ 3
1.2 Statistics for Studying Ice Thickness .................................................... 8
1.3 Contributions .................................................................................. 11
1.4 Thesis Organization ........................................................................... 12

## Chapter 2
Projection-Based Bayesian Inference for Spatial Generalized Linear Mixed Models

2.1 Introduction ......................................................................................... 14
2.2 Spatial Models ..................................................................................... 16
  2.2.1 Spatial Linear Models ................................................................. 16
  2.2.2 Spatial Generalized Linear Mixed Models ................................. 17
  2.2.3 Model Fitting and Computational Challenges ............................ 18
2.3 Confounding and Restricted Spatial Regression ........................................ 19
2.4 Reducing Dimensions through Projection .............................................. 21
  2.4.1 Random Projection ............................................................................. 23
  2.4.2 Approximation Comparison .............................................................. 25
  2.4.3 Random Projection for Spatial Linear Mixed Models .................... 27
  2.4.4 Random Projection for Spatial GLMMs ............................................. 29
### Chapter 2

- 2.4.5 Random Projection for Areal Data ........................................... 30
- 2.4.6 Rank Selection ........................................................................... 32
- 2.4.7 Computational Gains ................................................................. 33
- 2.5 Simulation Study and Results ..................................................... 36
  - 2.5.1 Linear Case .............................................................................. 36
  - 2.5.2 Binary Case ............................................................................. 37
  - 2.5.3 Poisson Case ........................................................................... 40
  - 2.5.4 Comparison to Predictive Processes ........................................ 42
- 2.6 Applications .................................................................................. 46
  - 2.6.1 Binary Data Application .......................................................... 46
  - 2.6.2 Count Data Application ........................................................... 47
- 2.7 Discussion ...................................................................................... 49

### Chapter 3

**Fast Maximum Likelihood Inference for Projection-Based Spatial Models**

- 3.1 Introduction ................................................................................... 52
- 3.2 Spatial Generalized Linear Mixed Models .................................... 54
  - 3.2.1 Models ..................................................................................... 54
  - 3.2.2 Spatial Confounding ................................................................. 57
- 3.3 A Projection-Based Approach to Dimension Reduction ............... 57
- 3.4 MCMC-EM for Inference ............................................................... 59
  - 3.4.1 E-Step ..................................................................................... 60
  - 3.4.2 M-Step .................................................................................... 62
  - 3.4.3 Stopping Criterion ................................................................. 65
  - 3.4.4 Rank Selection ........................................................................ 66
  - 3.4.5 Spatial Prediction .................................................................... 66
  - 3.4.6 Approximating Eigencomponents .......................................... 67
- 3.5 A Simulation Study ........................................................................ 69
  - 3.5.1 Count Data in Continuous Domain ....................................... 69
  - 3.5.2 Count Data on a Lattice ......................................................... 73
- 3.6 An Application on US Infant Mortality Data ................................ 76
- 3.7 Summary ....................................................................................... 78

### Chapter 4

**Inferring Ice Thickness from a Glacier Dynamics Model and Multiple Surface Datasets**

- 4.1 Introduction ................................................................................... 81
- 4.2 Glacier Dynamics and Data .......................................................... 84
  - 4.2.1 Glacier Dynamics ................................................................. 84
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Approximation performance comparison shows that introducing randomness in $\Phi = K^\alpha \Omega$ improves the Nystöm approximation to the eigenvectors (left column) and eigenvalues (right column). Letting $\Phi$ to be $K^\alpha \Omega$ with small power $\alpha = 1, 2$ also provides better approximation for both $\nu = 0.5$ (first row) and $\nu = 2.5$ (second row).</td>
<td>27</td>
</tr>
<tr>
<td>2.2</td>
<td>Pairwise cross-correlations are close to zero for both projection-based approaches (FRP and RRP). This is true for each pair of random effects (right) and between fixed and random effects (left). In contrast, the cross-correlations for the predictive process (black curves) are much larger. Hence our approaches result in better MCMC mixing. FRP = full model with random projection, RRP = restricted model with random projection and PP = predictive process model.</td>
<td>35</td>
</tr>
<tr>
<td>2.3</td>
<td>Computational time for $10^5$ iterations versus data size for SGLMM and projection-based approach (both FRP and RRP) with rank 50. This illustrates the benefit of the projection-based approach (both FRP and RRP) in terms of computational time. FRP = full model with random projection, and RRP = restricted model with random projection.</td>
<td>35</td>
</tr>
<tr>
<td>2.4</td>
<td>Distribution of posterior mean estimates of $\beta$ among four models and with adjustment. The distributions all center around the true value and are comparable. Random projection models FRP and RRP with rank=50 produce results that are similar to the models they approximate.</td>
<td>38</td>
</tr>
<tr>
<td>2.5</td>
<td>Linear case with $n = 1000$. First row shows the random effects at training locations. Second row shows the random effects on a $20 \times 20$ grid. Left column is the simulated data, middle column shows the estimated and predicted random effects from FRP, and right column shows the results from RRP. Random projections approach works well in recovering the true random effects.</td>
<td>39</td>
</tr>
</tbody>
</table>
2.6 Binary simulation study: distribution of $\beta$ posterior mean estimates for RP models and after adjustment. First row for the confounded case, and second row for the orthogonal case. All distributions center around the true value. For the confounded case (top row), FRP and RRP have similar results; while under the orthogonal case (bottom row), RRP produces much tighter distribution. For both cases, the adjusted inference A-RRP is similar to the FRP. FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.

2.7 First row shows the estimated probability surface at all training locations. Second row shows the predicted probability surface on a 20x20 grid using random projection models. Left column is simulated data, middle column shows the FRP, and right column shows the RRP. Comparing the patterns from the true model (left column) to the ones from projection-based models using rank $m = 50$ (middle and right columns), we see the projection-based models are able to recover the true value quite well. FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.

2.8 Poisson simulation study: distribution of $\beta$ posterior mean estimates for RP models and after adjustment. First row for the confounded case, and second row for the orthogonal case. All distributions center around the true value. For the confounded case (top row), FRP and RRP have similar results; while under the orthogonal case (bottom row), RRP produces much tighter distribution.

2.9 Simulated Poisson example: First row shows the estimated Poisson mean surface (in log scale) at all training locations. Second row shows the predicted Poisson mean surface (in log scale) on a 20x20 grid using projection-based models. Left column is simulated data, middle column shows the FRP, and right column shows the RRP. Comparing the pattern from the true model (left column) to the ones from our models using rank $m = 50$ (middle and right columns), we see the projection models are able to recover the true value quite well. FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.
2.10 Poisson simulation study: compare projection-based models with predictive process on gridded knots. The point estimate distributions for $\beta_1$ are comparable (top left); while the coverage for RRP is much lower than the others, however after adjustment, the coverage for A-RRP is corrected and is comparable to FRP and PP (top right). Both FRP and RRP have better prediction performance than PP (bottom left). The length of the CIs for FRP and PP are comparable, while RRP produce much narrower CI; but after adjustment the CI gets much wider (bottom right). FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.

2.11 Predicted rock profile by restricted model with random projection (right) and full model with random projection (left) using rank $m = 50$.

2.12 Data on abundance of house finch in the US from North American Breeding Bird Survey Dataset (left). Size of the circle is proportional to the bird counts. Predicted bird counts on a grid using the random projection method (right). The large number of observed counts on west coast and lack of observations in the central mid-west region is reflected on the prediction map.

3.1 Trace plots for one random effect for the first (left), second (middle), and the third (right) EM iteration. The number of Monte Carlo samples required reduces from 100,000 to $\sim 20,000$ within 3 EM iterations.

3.2 Parameter estimates with different starting values converge within a few iterations.

3.3 The left panel shows the simulated random effects at the observation (top) and prediction locations (bottom). The right panel shows the estimated (top) and predicted (bottom) random effects using the GLM starting value.

3.4 Monte Carlo sample size is adjusted automatically, with increasing simulation efforts as EM iteration increases (left). We typically obtain a large MC sample when the algorithm stops; vertical line indicates the stopping threshold is reached. Integrated log-likelihood function corresponding to different starting values stabilize as EM iteration increases (right).

3.5 Distributions of the estimates from the MCMC-EM algorithm for spatial counts in continuous domain. It seems that $\beta$ are unbiased and $\theta$ has positive biases.
3.6 Parameter estimates with different starting values converge within a few iterations (top row). Distributions of $\hat{\beta}$ from the MCMC-EM algorithm for count data on a lattice. It seems that $\hat{\beta}$ are unbiased and tightly center around the true value and $\hat{\tau}$ has a negative bias (bottom row).

3.7 Monte Carlo sample size is adjusted automatically, with increasing simulation efforts as EM iteration increases (a). We typically obtain a large MC sample when the algorithm stops; vertical line indicates the stopping threshold is reached. Integrated log-likelihood function corresponding to different starting values stabilize as EM iteration increases (b).

3.8 Parameter updates for infant mortality analysis.

4.1 Depending on different starting locations at the divide, the flow widths vary.

4.2 Thwaites Glacier thickness solutions corresponding to different flow-line widths and rheologic coefficient values. The non-statistical approach provides deterministic ice thickness solutions without uncertainty estimates, and none of these solutions seem to reconstruct the ice thickness well.

4.3 Dynamics model $M$ is based on equations (4.1) and (4.2). It links input processes, model parameters and ice thickness.

4.4 Synthetic ice thickness data with different magnitude of observational errors. Small, medium and large errors have standard deviations of 10, 50 and 100 m, respectively.

4.5 Predicted flow widths using 5, 10 and 25 synthetic ice thickness observations. Flow width predictions using small (a), medium (b) and large (c) magnitude of observational errors. The magnitude of observational errors does not change the shape of the width. Larger number of training locations produces tighter flow width that is more comparable to the synthetic true width.

4.6 Predicted ice thickness (m) from flow model using 5, 10, and 25 synthetic thickness observations. Prediction results are similar under different error schemes, and a larger number of training locations provide more certain thickness prediction.

4.7 Predicted ice thickness from the dynamics model using 5 (a), 10 (c), and 25 (c) training locations. The left panel shows the 95% credible intervals for predicted ice thickness plus observational errors using 5 (b), 10 (d), and 25 (f) training locations.
4.8 Thwaites Glacier flow width estimates using different number of training locations. More training locations recreates the local feature of the width.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Model comparisons for linear case with ( n = 400 )</td>
<td>38</td>
</tr>
<tr>
<td>3.1</td>
<td>Results of our analyses for Poisson data simulated from the SGLMM in a continuous domain</td>
<td>73</td>
</tr>
<tr>
<td>3.2</td>
<td>Coverages for Poisson data simulated from the SGLMM on a continuous domain</td>
<td>73</td>
</tr>
<tr>
<td>3.3</td>
<td>Results of our analyses and computation time (minutes) for Poisson data simulated from the SGLMM on a lattice</td>
<td>75</td>
</tr>
<tr>
<td>3.4</td>
<td>Coverages for Poisson data simulated from the SGLMM on a lattice</td>
<td>76</td>
</tr>
<tr>
<td>3.5</td>
<td>Results of fitting model with rank 50 to the infant mortality data</td>
<td>78</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of Data Sets</td>
<td>91</td>
</tr>
<tr>
<td>4.2</td>
<td>The coverage of 95% CI’s for flow width prediction</td>
<td>101</td>
</tr>
</tbody>
</table>
I sincerely thank my adviser Murali Haran for his guidance and encouragement on my research and professional development. Working with Dr. Murali Haran provided me opportunities to present our work at various workshops and conferences, and the unique opportunities to work with undergraduate summer scholars, including, Haixing Yin, Fanhu Gui, Joseph Kvederis, Ben Roberson, Ruijun Wu and Gwen Wyatt.

I want to thank my collaborators David Pollard and Patrick Applegate for their guidance on ice sheet modeling research. I am grateful to John Hughes, Jim Hodges and Ephraim Hanks for helpful discussions. I also thank Klaus Keller and Kelsey Ruckert for providing wonderful collaboration opportunities. I feel very grateful to have the opportunities to be mentored by and collaborate with many outstanding researchers during my graduate study. The NSF Statistical Methods in the Atmospheric Sciences Network (STATMOS) had organized many great conferences and workshops that I am very grateful to be a part of.

I am grateful to Professor Sanjay Srinivasan at the Department of Energy and Mineral Engineering at Penn State University for providing the seismic data set. I also want to thank Knut Christianson and Nick Holschuh for the helpful discussions on the Thwaites Glacier data sets. This work was partially supported by the National Science Foundation through (1) NSF-DMS-1418090 and (2) Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO1240507. I would also like to acknowledge the high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR’s Computational and Information Systems Laboratory, sponsored by the National Science Foundation.
Spatial models for non-Gaussian data are used extensively in many disciplines, for example, when interpolating ice sheet thickness in West Antarctica, modeling plant or animal species counts in ecology (Oliveira 2013; Wikle 2002), forest presence-absence data in environmental studies (Finley et al. 2008), and disease incidence in epidemiology (Waller and Carlin 2010). These spatial models are used for interpolation or for carrying out regression while adjusting for spatial dependence. Spatial generalized linear mixed models (SGLMMs) proposed by Diggle et al. (1998); Besag et al. (1991) have become very popular for modeling spatial data, particularly non-Gaussian spatial data. While they are very flexible, SGLMMs pose some challenges in computation when the data set is large. There are also issues with regression parameter interpretation due to spatial confounding between fixed and random effects in these models. These challenges motivate the development of a computationally efficient projection-based SGLMM in this thesis. Computational algorithms, specifically Markov chain Monte Carlo (MCMC) for Bayesian inference, and Markov chain Monte Carlo expectation-maximization (MCMC-EM) for maximum likelihood inference, are important components of my dissertation research.
An important motivating problem for my dissertation is research on studying the West Antarctic Ice Sheet (WAIS). The present and future of this ice sheet are important to study because its melting is directly linked to sea level rise. In the problem I study in my dissertation, disparate sources of information are available regarding the Thwaites glacier in West Antarctica. These data include physical models and various data sets measuring different properties of the ice sheet. A statistical challenge is to develop a computationally efficient method that can integrate these multiple sources of information in a single framework to provide parameter estimates along with its uncertainty, and probabilistic predictions for the unobserved or partially observed variables of interest. These motivate a Bayesian approach that combines a glacier dynamics model with multiple data sets for reconstructing the ice profile at the Thwaites Glacier.
1.1 Models for Large Non-Gaussian Spatial Data

Spatial data are data that have locations associated with them. Here I will mainly deal with point-level data irregularly-positioned in a continuous domain, and areal data located on a lattice. When spatial data are high-dimensional, computation for models for such data typically becomes very challenging. In cases where spatial regression is of interest, confounding between spatially dependent random effects and predictors can make parameter interpretation problematic. For observations that are modeled well by the Gaussian assumption, there is a large body of research to deal with such challenges. I will discuss below some of these current approaches and provide some insights on how they may or may not extend easily to non-Gaussian observations. I will also introduce some useful methods for large non-Gaussian data and draw connections wherever possible to the proposed methods in this thesis.

Spatially observed response variables often exhibit correlations after controlling for available covariates. In a linear regression setting, this dependence is often captured by random effects modeled using a Gaussian process (Cressie, 1992). These models are hence called spatial linear mixed models (SLMMs). This model was extended to non-Gaussian observations in two seminal papers, Diggle et al. (1998), with a focus on point-level data, and, Besag et al. (1991), with a focus on areal data (e.g. county-level data). This extension is analogous to the transition from a linear model to a generalized linear model and hence these models may be referred to as spatial generalized linear mixed models (SGLMMs). Both Diggle et al. (1998) and Besag et al. (1991) develop Bayesian inferential approaches. The Bayesian framework is flexible and can easily integrate multiple sources of information (e.g., physical models, multiple data sets, and complex sources of error) into a single framework. In fact, the Bayesian approach has become a very
popular inferential method for spatial data in statistics as well as various applied
disciplines. For instance, based on current Google scholar citation counts, Diggle
et al. (1998) has over 1300 citations and Besag et al. (1991) has over 2600 citations.
However, SGLMMs pose three major challenges: (1) computational issues due to
high-dimensional random effects that are typically heavily cross-correlated – these
often result in slow-mixing Markov chain Monte Carlo algorithms (MCMC); (2)
expensive matrix calculations involving large matrices; and (3) difficult parameter
interpretation due to confounding between fixed and random effects (Reich et al.,
2006; Hughes and Haran, 2013; Hanks et al., 2015). This problem is similar to the
multicollinearity problems with standard regression models.

My contribution with this dissertation is to provide a novel projection-based
approach for fast computation for both the discrete and the continuous domains.
I first develop methods for this in the Bayesian paradigm. Maximum likelihood
inference using Monte Carlo MLE (Geyer, 1991) and Monte Carlo EM (Lange,
1995) has been studied for these models, for instance, Christensen (2004) and
Zhang (2002), respectively; however it is computationally challenging for large
non-Gaussian data due to the high dimensional integrations required when com-
puting the likelihood function. In this dissertation I develop a maximum likelihood
procedure that works well for high-dimensional problems, and provide a fast and
automated algorithm for researchers who want to fit these models quickly and rou-
tinely. This has the benefit of permitting maximum likelihood inference as well as
potentially reducing computational costs when compared to Bayesian approaches.

Existing fast computational methods for high-dimensional spatial models largely
focus on the linear (Gaussian) case. These includes fixed rank approaches (cf. Hig-
don, 1998; Cressie and Johannesson, 2008; Nychka et al., 2015; Banerjee et al.,
2012), where the main idea is to approximate the infinite-dimensional Gaussian
process with a fixed number of basis functions and their weights. The choice of basis functions is flexible, for example, they can be kernel functions (Higdon 1998), different types of wavelets or local bisquare functions (Cressie and Johannesson 2008), radial basis functions with multiresolution (Nychka et al. 2015), or approximated eigenfunctions of the Gaussian process (Banerjee et al. 2012); The choice of model for weights is also flexible, for example, a mean-zero multivariate normal with unknown covariance structure to be estimated (Cressie and Johannesson 2008), independent normal variables (Higdon 1998, Banerjee et al. 2012) or a Gaussian Markov random field (Nychka et al. 2015). Some approaches such as covariance tapering (cf. Furrer et al. 2006; Kaufman et al. 2008; Shaby and Ruppert 2012) do not necessarily reduce the dimension of the random effects, but rather utilize sparse matrix operations to speed up the likelihood evaluations by introducing zeros in the covariance matrix. Another promising area is likelihood approximation (cf. Vecchia 1988; Stein et al. 2004; Caragea and Smith 2006; Bevilacqua et al. 2012; Datta et al. 2016), where the main idea is to approximate the likelihood function in order to reduce computational costs while still preserving some desirable properties of the resulting estimators. Approximate likelihoods can be constructed in a number of different ways, for instance by partitioning the data into clusters (Vecchia 1988; Stein et al. 2004), blocks (Caragea and Smith 2006) or neighborhood (Datta et al. 2016) and assuming conditionally independence, or by a weighted composite likelihood approach (Bevilacqua et al. 2012) to reduce computations of the likelihood function. Hybrid approaches are also possible, for instance, Lindgren et al. (2011); Simpson et al. (2012) propose using a Markov random field to approximate the latent Gaussian process to utilize the sparse structure of the precision matrix, and then combine this model with Laplace approximation for fast inference. Furthermore, numerous new methods
have been/are being developed for the linear spatial model constantly. However, extending these approaches to non-Gaussian spatial models is non-trivial and it is not clear whether and how these methods may be adapted for addressing spatial confounding.

Some approaches have been developed for modeling large non-Gaussian spatial data. The predictive process approach (Banerjee et al., 2008) has been studied in the SGLMM context, and it is an important contribution to the literature. The main idea of this approach is to reduce the dimension of the random effects by first selecting a small number of locations (knots), then approximate the original spatial process with a reduced-rank process conditional on the process at these knots. This has been used widely but it requires users to provide reference knots, which can be challenging to specify (Finley et al., 2009). Also, it underestimates the variance of the random effects, a problem that may be corrected by adding an independent error component (Finley et al., 2009). The problem with this correction is that, while it does not change the computation in the linear case, it makes computations very challenging in the SGLMM case as the correction results in a much higher-dimensional process. The INLA approach (Rue et al., 2009; Eidsvik et al., 2009, 2012) provides a fast approximate numerical method for carrying out inference for latent Gaussian random field models. INLA is very fast and hence potentially very useful but it is often not flexible enough. In particular, when the mean function in the model is complicated or is represented by a physical model (as is often the case in atmospheric and earth sciences), it is unclear how to use an INLA approach. In addition, there may be issues with the accuracy of INLA-based inference. The methods I proposed are not incompatible with INLA; I will discuss how the two methods can be combined. In addition to these issues, both INLA and the predictive process do not provide an obvious solution to the confounding
issue. Another approach, but one that only applies to the lattice model case, is one that uses the neighborhood structure, essentially the graph that describes the underlying dependence model, to construct a reduced-rank SGLMM (Hughes and Haran, 2013). This approach also alleviates confounding in a computationally efficient manner. We find later that it is at least as efficient as the methods I develop in this thesis. However, its major limitation is the fact that the methodology does not extend to the continuous-domain setting.

In this thesis, I describe a novel method that utilizes the principal components of the covariance matrix to achieve fast computation for fitting traditional SGLMMs. This projection-based approach is similar to the fixed-rank approach except that it allows the basis functions to adapt to the data; it uses the approximated eigenfunctions as a basis. This therefore has the added benefit of resulting in nearly independent random effects, which can help when constructing efficient MCMC algorithms for inference. This method simultaneously alleviates spatial confounding using a similar recipe as Hughes and Haran (2013); Reich et al. (2006); Hodges and Reich (2010), where the random effects are restricted to be orthogonal to the fixed effects. In Chapter 2, I describe, in detail, the projection-based method for both the continuous and discrete spatial domains under the Bayesian inferential framework. Also in this chapter, I compare the projection-based method to the predictive process and the reduced-rank approach proposed by Hughes and Haran (2013). In Chapter 3, I propose a fast maximum likelihood inference approach for SGLMMs based on two projection-based models, that is, the discrete domain spatial model for lattice data (Hughes and Haran, 2013) and the continuous domain case proposed in Chapter 2. In both chapters, I show, via simulations and real data applications, that my new projection-based approach works well both in terms of inference and prediction.
1.2 Statistics for Studying Ice Thickness

The West Antarctic ice sheet (WAIS) is currently the largest marine ice sheet with much of the area lying below sea level; therefore, it is unstable and is capable of rapid change. It is drained by fast-flowing glaciers, including the Pine Island Glacier and Thwaites Glacier ice streams, which are major contributors to ice loss, and consequently to sea level rise (Davies, 2014). Hence, understanding the stability and dynamics of glaciers is critical for predicting the future of the ice sheet. Disparate sources of information, including (1) physical models that describe glacier behavior and (2) multiple data sets acquired via different instruments are available, and are important for studying the glaciers in their own right. However, a statistical approach that integrates these multiple sources of information can be helpful to quantify the imperfection of each of the sources, such as the mismatch between the mathematical model and the true state of the glaciers (model discrepancy), and the observational errors inherited from the data sets. The idea of combining physical and statistical modeling is not new, and has been a promising yet challenging area of research (Berliner, 2003).

Glacier dynamics are driven by the interplay between the topography, temperature and basal conditions beneath the ice. The interaction among these major processes are often modeled by a numerical model consists of differential equations (cf. Warner and Budd, 2000; Van der Veen, 2013). Depending on the resolutions and complexities of physical processes, there are simple one-dimensional flowline models that can be solved numerically within a second of computational time, to complex three-dimensional model that can take hours. Here, we employ a one-dimensional flowline model, which describes the ice flow along the fastest flow path (centerline). We recognize that the flowline model does not account for the horizontal characteristics of an ice stream. Therefore, we correct this via statistical
modeling of the flow width. The advantage of working with a simple physical model is mainly its fast computation so that a fully Bayesian inversion method can be applied. The disadvantage is the simplified physics incorporated in the model. In our problem, we restrict ourselves to grounded ice stream where the simple flowline model, combined with the flow width adjustment, describes the major mechanism of the underlying process. The flowline model relies on two fundamental laws: the conservation law of mass, and the Glen’s flow law describing how ice deforms due to stress (Huybrechts 2007). It provides a deterministic relationship between the processes that can be measured relatively easy, such as surface elevation, surface ice velocity and accumulation rate, and the processes for which we have little or no information, such as ice thickness and the flow width. Therefore, we can use this relationship to deduce quantities that are scientifically of interest.

Available data inputs for the flowline model are largely surface observations, including surface elevation using Airborne Topographic Mapper (ATM) instrumentation, surface velocity of ice by speckle tracking, and ice accumulation (snowfall) rates from a calibrated regional climate model. For the Thwaites Glacier, we also have dense ice thickness observations, which are typically sparse for most glaciers; they are derived by subtracting bed elevation from surface elevation. Bed elevation is obtained by aircraft equipped with measuring instrument flying near the centerline of the glacier. The scientific goals here are to estimate or interpolate bedrock topography along the centerline given fairly accurate surface data sets and sparsely observed ice thickness, and to estimate the ice deformation coefficient, which describes the temperature and fabric of the ice.

A sound statistical modeling approach should combine the underlying physics and multiple relevant data sets to reconstruct the ice sheet. Because pure statistical interpolation may violate the process generating mechanism, it will typically
result in physically implausible solutions, and the resulting interpolations would not contribute to our understanding of the glacier dynamics. On the other hand, a purely physics-based approach often neglects errors and uncertainties in observations and physical model (Kennedy and O’Hagan 2001), resulting in unreliable estimates. Some major challenges involved in developing such statistical methods are (1) dealing with high-dimensional observations that have different spatial resolutions and are subject to observational errors, (2) modeling the completely unobserved quantities, including the ice deformation coefficient (a scalar) and the spatially varying flow width (a process) and (3) providing solutions in a computationally efficient manner.

There is a large amount of research that combines noisy observations and deterministic models. However, this “data fusion” problem remains an active research area. Some approaches utilize physical model outputs without explicitly solving the model (cf. Reich and Fuentes 2007; Wikle et al. 2001). Some solve the physical model explicitly with a focus on inferring model parameters, which are typically scalar (cf. Li et al. 2005; Ramsay et al. 2007; Xun et al. 2013). However, these approaches do not easily adapt to our problem, because the unknown flow width in our glacier dynamics model varies along the flowline. Therefore, the flow width should be modeled as a process that varies along the flowline instead of oversimplifying the flow width by treating it as a scalar. Moreover, the multiple data inputs to the physical model have different spatial resolutions and are subject to observational errors.

Some closely related recent works, which integrate multiple data sets and use a glacier dynamics model, are proposed by Berliner et al. (2008a,b). They illustrated the usefulness of physical-statistical modeling in estimating unknown parameters for understanding glacier properties. The problem considered in this thesis differs
from their work. Their focuses are on parameter estimation and studying how the
assumed smoothness of the observational processes affects parameter estimation.
In contrast, our focus is on using sparse thickness data combined with other high
quality surface observations and physical models to reconstruct the ice profile.

1.3 Contributions

The projection-based model developed in Chapter 2 provides an alternative to
existing methods for dealing with the computational and confounding issues in
high-dimensional SGLMMs. The reparameterization approach we develop is flexi-
ble and easy to implement in practice for latent-Gaussian process models. I have
conducted extensive simulation studies for linear (Gaussian), count and binary data
to inform inference and practical implementation of the projection-based model.
This work, co-authored by Murali Haran, has been revised and resubmitted to
Journal of Computational and Graphical Statistics.

In Chapter 3, I provide a fast maximum likelihood inference approach for high-
dimensional non-Gaussian spatial data in the SGLMMs context. Markov chain
Monte Carlo expectation-maximization (MCMC-EM) algorithms are commonly
used for maximum likelihood inference for SGLMMs, but they have so far been
infeasible for high-dimensional data. The method developed here therefore con-
tribute to the statistical methodology in this area by developing maximum like-
lihood inference for SGLMMs with reduced-dimensional reparameterizations. A
significant portion of the work here is computational, specifically on developing
MCMC-EM algorithms to fit these models. The manuscript based on this work,
co-authored by Murali Haran, is targeted for submission to a statistics methods
journal.

In Chapter 4, I develop a hierarchical Bayesian model that integrates multiple
ice sheet surface data sets with a glacier dynamics model. This approach allows us to (i) infer important parameters describing the glacier dynamics, (ii) learn about ice sheet thickness, and (iii) account for errors in the observations and the model. This approach requires only sparsely observed thickness data that is relatively expensive to obtain, and high-quality surface data sets that are relatively easy to obtain from satellite imagery. This methodology may potentially be extended to other glaciers on the WAIS. This work co-authored by Murali Haran and David Pollard has been tentatively accepted for publication in *Environmetrics*.

### 1.4 Thesis Organization

The outline of the remainder of this thesis is as follows. In Chapter 2, I provide some brief background on spatial models, and computational and inferential challenges. I then propose a novel projection-based approach for latent Gaussian process models and discuss results from applications to simulated and real data.

In Chapter 3, I propose a Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm for fast maximum likelihood inference for projection-based methods in spatial generalized linear mixed models. In Chapter 4, I first outline the scientific problem and statistical challenges in the context of inferring various properties of an ice sheet. Then I propose a Bayesian hierarchical approach for modeling ice glaciers on WAIS. Finally, in Chapter 5, I discuss some potential directions of future work.
In this chapter, I propose a novel projection-based method for modeling large non-Gaussian data under the Bayesian framework, which has been the default inferential framework for these models for the last two decades. This method reduces the dimension of random effects based on principle component analysis, and allows a fast approximation of the leading eigencomponents relying on the random projections algorithm. I show how we can build upon this projection-based approach to address the computational and inferential challenges of fitting spatial generalized linear mixed models for high-dimensional data on both the continuous and lattice domains. I study the inference and prediction performance of the proposed method via extensive simulation studies and real data applications.
2.1 Introduction

Gaussian and non-Gaussian spatial data arise in a number of disciplines, for example, species counts in ecology, tree presence-absence data, and disease incidence data. Models for such data are important for scientific applications, for instance when fitting spatial regression models or when interpolating observations across continuous spatial domains. Spatial generalized linear mixed models (SGLMMs) are popular and flexible models for spatial data, both for continuous domain or “point-referenced” data (Diggle et al., 1998), where the spatial dependence is captured by random effects modeled using a Gaussian process, as well as for lattice or areal data (cf. Besag et al., 1991; Rue and Held, 2005) where dependence is captured via random effects modeled with Gaussian Markov random fields. SGLMMs have become very popular in a wide range of disciplines. In practice, however, SGLMMs pose some computational and inferential challenges: (i) computational issues due to high-dimensional random effects that are typically strongly cross-correlated – these often result in slow mixing Markov chain Monte Carlo (MCMC) algorithms; (ii) computations involving large matrices; and (iii) spatial confounding between fixed and random effects, which can lead to variance-inflated estimation of regression coefficients (Reich et al., 2006; Hughes and Haran, 2013; Hanks et al., 2015). In this manuscript we provide an approach for reducing the dimensions of the spatial random effects in SGLMM models. Our approach simultaneously addresses both computational issues as well as the confounding issue.

There is a large literature on fast computational methods for spatial models (cf. Cressie and Johannesson, 2008; Banerjee et al., 2008; Higdon, 1998; Shaby and Ruppert, 2012; Datta et al., 2016, among many others). These methods have been very useful in practice, but they largely focus on linear (Gaussian) spatial models and do not consider the spatial confounding issue. The predictive process approach
(Banerjee et al., 2008) has been an important contribution to the literature, and has also been studied in the SGLMM context. However, the predictive process approach requires that users provide reference knots, which can be challenging to specify; our method is more automated. We also find that in some cases we obtain similar performance to the predictive process at far lower computational cost. Crucially, our approach is also able to easily address the spatial confounding issue.

INLA (Rue et al., 2009) provides a sophisticated numerical approximation approach for SGLMMs. As we later discuss, INLA may be used in combination with the projection-based reparameterization approach we develop in this manuscript. This is useful for addressing confounding while also reducing computational costs.

Restricted spatial regression models for areal and point-referenced spatial data (Reich et al., 2006; Hanks et al., 2015) address the confounding issue. However, these models are computationally intensive for large data sets. For areal data, Hughes and Haran (2013) alleviate confounding in a computationally efficient manner by proposing a reparameterization that utilizes the underlying graph to reduce the dimension of random effects. To our knowledge, no existing approach alleviates spatial confounding and is computationally efficient for point-referenced non-Gaussian data. In this manuscript we describe a novel method that utilizes the principal components of covariance matrices to achieve fast computation for fitting traditional SGLMMs as well as restricted spatial regression.

Our method relies on the random projections algorithm (Banerjee et al., 2012; Sarlos, 2006; Halko et al., 2011), which allows a fast approximation of the leading eigencomponents. We show how we can build upon this projection-based approach to address the computational and inferential challenges of SGLMMs. The outline of the remainder of the chapter is as follows. In Section 2.2, we introduce spatial linear mixed models and explain how a generalized linear model formulation of
these models is appropriate for non-Gaussian observations. We also examine the computational challenges and some current approaches. In Section 2.3, we explain spatial confounding, how it affects interpretation of regression parameters, and describe how to alleviate confounding via orthogonalization. In Section 2.4, we describe our projection-based approach for both the continuous domain and lattice case. We study the inference and prediction performance of the proposed method via a simulation study in Section 2.5, and study our method in the context of applications in Section 2.6. We conclude with a discussion of our work in Section 2.7.

2.2 Spatial Models

2.2.1 Spatial Linear Models

Let \( Y(s) \) denote an observation, and \( x(s) \) a \( p \)-dimensional vector of covariates at location \( s \) in a continuous spatial domain \( D \subseteq \mathcal{R}^d \), where \( d \) is typically 2 or 3. Given data locations \( S = \{s_1, \ldots, s_n\} \), the observations \( Y = [Y(s_1), \ldots, Y(s_n)]^T \) may show residual spatial structure after controlling for \( X = [x(s_1), \ldots, x(s_n)]^T \). This can be taken into account by including spatially dependent random effects \( W(s) \) to model the residual dependence,

\[
Y(s) = x(s)^T \beta + W(s) + \epsilon(s), \tag{2.1}
\]

where \( \beta \) are regression parameters. \( \{\epsilon(s) : s \in D\} \) is a small-scale (nugget) spatial effect/measurement error process, modeled as an uncorrelated Gaussian process with mean 0 and variance \( \tau^2 \). For the continuous spatial domain, the random effects \( \{W(s) : s \in D\} \) are typically modeled by a zero-mean stationary Gaussian
process with a positive definite covariance function $C(\cdot)$. Hence, for a finite set of locations, $\mathbf{W} = [W(s_1), \ldots, W(s_n)]^T$ follows a multivariate normal distribution $\text{MVN}(0, \Sigma)$, with $\Sigma_{ij} = \text{cov}(W(s_i), W(s_j)) = C(||s_i, s_j||)$.

A commonly used class of covariance functions, assuming stationarity and isotropy, is the Matérn class (Stein, 1999),

$$C(s_i, s_j) = C(h) = \sigma^2 \rho(h; \phi, \nu) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}h}{\phi}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu}h}{\phi}\right),$$

where $h = ||s_i - s_j||$ denotes the Euclidean distance between pairs of locations, $\sigma^2$ is a variance parameter, and $\rho$ is a positive definite correlation function parameterized by $\phi$, the spatial range parameter, and $\nu$, the smoothness parameter. $\Gamma(\cdot)$ is the gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind.

### 2.2.2 Spatial Generalized Linear Mixed Models

A popular way to model spatial non-Gaussian data is by using spatial generalized linear mixed models (SGLMMs) (cf. Diggle et al., 1998; Haran, 2011). Let \( \{Z(s) : s \in \mathcal{D}\} \) denote a non-Gaussian spatial field, and $g(\cdot)$ a known link function. Then, the conditional mean, $E[Z(s) \mid \beta, W(s)]$ may be modeled as

$$\eta(s) \equiv g\{E[Z(s) \mid \beta, W(s)]\} = \mathbf{x}(s)^T \beta + W(s), \quad s \in \mathcal{D}. \tag{2.2}$$

Conditional on $W(s)$, $\mathbf{Z} = [Z(s_1), \ldots, Z(s_n)]^T$ are mutually independent, following a classical generalized linear model (cf. Diggle et al., 1998). We provide two commonly used examples of SGLMMs for spatial binary and count data to illustrate our projection-based approach, the Poisson with log link and binary with logit link respectively. The projection-based approach presented in this chapter generalizes to other link functions and observation models, as well as to cases where an
additional nugget term is added to the model (2.2) (cf. Berrett and Calder 2016). Details for the nugget model are provided in Appendix A.

2.2.3 Model Fitting and Computational Challenges

The hierarchical structure of spatial models makes it convenient to use a Bayesian inferential approach. Often in practice, we fix the value of $\nu$ and assign prior, $p(\theta, \beta)$, to parameters $\theta$ and $\beta$ where $\theta = (\phi, \sigma^2, \tau^2)^T$, then use Markov Chain Monte Carlo (MCMC) to sample from the posterior $\pi(\beta, \theta, W | Z)$. Fitting SGLMMs generally requires the evaluation of an $n$-dimension multivariate normal likelihood for every MCMC iteration, with matrix operations of order $n^3$ floating point operations (flops). There are often strong correlations between the fixed and random effects (Hodges and Reich 2010), and strong cross-correlations among the spatially dependent random effects. It is well known that this dependence typically results in poor mixing in standard MCMC algorithms (cf. Christensen et al. 2006, Haran et al. 2003, Rue and Held 2005). Furthermore, when the data locations are near each other, the covariance matrix may be near singular, resulting in numerical instabilities (Banerjee et al. 2012). These issues motivate the development of our reduced-dimensional approach to inference for SGLMMs.

Considerable work has been done to address the above issues in the linear case, where model inference and prediction are based on the marginal distribution $Y | \beta, \phi, \sigma^2, \tau^2 \sim \text{MVN}(X\beta, \Sigma + \tau^2 I)$. Several methods rely on low rank approximations or multi-resolution approaches to reduce computations involving the $n \times n$ covariance matrix $\Sigma$ (cf. Banerjee et al. 2012, Sang and Huang 2012, Nychka et al. 2015, Cressie and Johannesson 2008). However, these methods do not readily extend to SGLMMs because the marginal distribution for $Z | \beta, \phi, \sigma^2$ is not available. A notable exception is the predictive process ap-
proach (Banerjee et al., 2008), where the extension to SGLMMs has been well studied. This approach replaces random effect $W$ by $W^*$, the realization of $W(s)$ at $m \ll n$ reference knots $S^* = \{s^*_1, \ldots, s^*_m\}$; $W \approx C(s, s^*)C^{-1}W^*$, where $C^* = C(s^*, s^*)$ denotes the $m \times m$ covariance matrix $\text{var}(W^*)$. Correspondingly, $\Sigma$ is approximated by a low rank matrix $c^T C^* c$, where $c^T = C(s, s^*)$ denotes the covariance matrix $\text{cov}(W, W^*)$. This method can be applied to both the linear and the generalized case. However, Finley et al. (2009) points out that the predictive process underestimates the variance of $W$ and proposed a modified predictive process by defining $W_{\text{mod}}(s) = C(s, s^*)C^{-1}W^* + \tilde{\epsilon}(s)$, where $\tilde{\epsilon}(s) \sim N(0, C(s, s) - c^T(s, s^*)C^{-1}c(s, s^*))$. For the linear case, this adjustment adds little extra computation. However, for the SGLMM case, the modified predictive process puts us back to working with a high-dimensional random effect $W_{\text{mod}}$. Furthermore, determining the number and placement of reference knots is a non-trivial challenge (see Finley et al., 2009 for some potential strategies).

Another challenge with SGLMMs arises from the strong-correlations among random effects, which often results in poor Markov chain mixing. Reparameterization techniques (Christensen et al., 2006) can help with mixing; however, for high-dimensional spatial data, the reparameterizing step is computationally expensive and may not result in fast mixing.

## 2.3 Confounding and Restricted Spatial Regression

Spatial confounding occurs when the spatially observed covariates are collinear with the spatial random effects. This is a common problem for both point-referenced and areal data (cf. Hanks et al., 2015; Reich et al., 2006). Here we
demonstrate the confounding problem in a continuous domain. Let \( \eta = [\eta(s_1), \ldots, \eta(s_n)]^T \) denote the transformed site-specific conditional means, where \( \eta(s_i) = g\{E[Z(s_i) \mid \beta, W(s_i)]\} \). The SGLMM is then

\[
\eta = X\beta + W, \quad W \sim MVN(0, \sigma^2 R(\phi)),
\]  

(2.3)

where the covariance \( \Sigma \) is \( \sigma^2 R(\phi) \), with \( R(\phi) \) a positive definite correlation matrix, \( R_{ij}(\phi) = \rho(||s_i - s_j||; \phi) \). \( X \) are spatially observed covariates that may explain the random field of interest. \( W \) is used as a smoothing device. When both \( X \) and \( W \) are spatially smooth, they are often collinear (cf. Hanks et al., 2015). This confounding problem may lead to variance inflation of the fixed effects (Hodges and Reich, 2010).

Let \( P_{[X]} \) and \( P_{[X]}^\perp \) denote orthogonal projections onto the space spanned by \( X \) and its complement, respectively. Model (2.3) can be equivalently written as

\[
\eta = X\beta + P_{[X]}W + P_{[X]}^\perp W = X[\beta + (X^TX)^{-1}X^TW] + P_{[X]}^\perp W = X\tilde{\beta} + P_{[X]}^\perp W.
\]

(2.4)

In some cases, it may be reasonable to fit model (2.4) to address the confounding issue by restricting the random effects to be orthogonal to the fixed effects in \( X \) (Reich et al., 2006; Hughes and Haran, 2013; Hanks et al., 2015). We refer to this as restricted spatial regression (RSR) in the remaining sections. After fitting the RSR via MCMC, we can obtain valid inference for \( \beta \) using an \textit{a posteriori} adjustment based on the MCMC samples (Hanks et al., 2015). Let \( k \) indicate the \( k^{th} \) MCMC sample, then

\[
\beta^{(k)} = \tilde{\beta}^{(k)} - (X^TX)^{-1}X^TW^{(k)}.
\]

(2.5)
Fitting RSR is just as computationally expensive as regular SGLMMs because the dimension of random effects $P_{[X]} W$ remains large; their strong correlations lead to slow MCMC mixing.

### 2.4 Reducing Dimensions through Projection

Instead of working with the original size of the random effects $W$ in the model, we consider a reduced dimensional approximation. We want to reduce the dimension of random effects from $n$ to $m$ so that: (i) for a fixed $m$, the approximation to the original process comes close to minimizing the variance of the truncation error (details below), (ii) the $m$ random effects are nearly uncorrelated, and (iii) we reduce the number of random effects as far as possible in order to reduce the dimensionality of the posterior distribution.

Let $\delta$ denote a vector of the reduced-dimensional reparameterized random effects. The main idea of our approach is to obtain $\delta$ by projecting $W$ to its first-$m$ principal direction $V_m = [v_1, \ldots, v_m]$, and scaling by its eigenvalues $\Lambda_m = \text{diag}(\lambda_1, \ldots, \lambda_m)$. Conditional on $\phi$, $W$ is multivariate normal. Hence, $\delta = (V_m \Lambda_m^{-1/2})^T W$ is conditionally independent given $\phi$, $\delta | \phi \sim N(0, I)$. This reparameterization utilizing principal components minimizes the variance of the truncation error (details are provided below), decorrelates the random effects and reduces its dimension to $m$. However, exact eigendecomposition is computationally infeasible for high-dimensional observations, so we approximate the eigencomponents, $U_m \approx V_m$ and $D_m \approx \Lambda_m$, using a recently developed stochastic matrix approximation ([Halko et al. 2011](#)). Because the eigencomponents are approximated reasonably well by the random projections algorithm as illustrated in Section 2.4.2, our reparameterization, $\delta = (U_m D_m^{-1/2})^T W$, is therefore close to the one based on the exact eigendecomposition. In the remainder of this section we
describe the motivation and properties of the reparameterization approach. We describe random projections for fast approximations of the eigencomponents and illustrate its approximation performance. We then explain how to fit the reparameterized SGLMMs with random projection. We conclude this section by showing how our approach is also applicable to areal data, and compare it to the method in [Hughes and Haran 2013].

Our reparameterized random effects model achieves (i)-(iii) above. Consider the spatial process \{W(s) : s ∈ D\} in (2.1) defined on a compact subset \(D\) of \(\mathcal{R}^d\). Let \(\{ψ_i(s) : i = 1, \ldots, ∞\}\) and \(\{λ_i : i = 1, \ldots, ∞\}\) be orthonormal eigenfunctions and eigenvalues, respectively, of the covariance function \(C(·)\) of \(W(s)\). By Mercer’s theorem, they satisfy \(C(s, s') = ∑_i=1^∞ λ_i ψ_i(s)ψ_i(s')\) [Adler 1990]. By the Karhunen-Loève (K-L) expansion, we can write \(W(s) = ∑_i=1^∞ ξ_i \sqrt{λ_i} ψ_i(s)\), where \(\{ξ_i : i = 1, \ldots, ∞\}\) are orthonormal Gaussian [Adler 1990]. Assuming the eigenvalues are in descending order, \(λ_1 ≥ λ_2 ≥ \ldots\), the truncated K-L expansion for \(W(s)\), \(\tilde{W}(s) = ∑_i=1^m ξ_i \sqrt{λ_i} ψ_i(s)\), minimizes the mean square error, \(||W − \tilde{W}||\), among all basis sets of order \(m\) [Banerjee et al. 2012; Cressie and Wikle 2015]. A discrete analogue for the truncated expansion of the process realization \(W\) is similar to the above. \(W\) has expansion \(W = ∑_i=1^n ξ_i \sqrt{λ_i} v_i\), while its rank-\(m\) approximation is \(\tilde{W} = ∑_i=1^m ξ_i \sqrt{λ_i} v_i\), where \(\{λ_i, v_i\}\) are eigen-pairs of \(Σ\). Let \(V_n = [v_1, \ldots, v_n]\) denote an \(n×n\) matrix of eigenvectors, and \(Λ_n = \text{diag}(λ_1, \ldots, λ_n)\) an \(n×n\) diagonal matrix of eigenvalues. We define \(V_m = [v_1, \ldots, v_m]\) and \(Λ_m = \text{diag}(λ_1, \ldots, λ_m)\) similarly. Then, the truncated expansion has variance \(\text{var}(\tilde{W}) = \tilde{Σ} = V_m Λ_m V_m^T\), and it minimizes the variance of the truncation error \(||Σ − \tilde{Σ}|| = ∑_{i=m+1}^n λ_i\) [Banerjee et al. 2012].

Our reparameterization is a principal component analysis (PCA) based approach with the advantage that a reasonably small rank \(m\) captures most of the
spatial variation. For instance, we later demonstrate in a simulated example of data size \( n=1,000 \), rank \( m=50 \) is sufficient to achieve reasonable performance. We discuss heuristics for choosing an appropriate value for \( m \) in Section 2.4.6.

2.4.1 Random Projection

Random projection is an approach that facilitates fast approximations of matrix operations (see Halko et al., 2011, and references therein). Here we use it to approximate the principal components of covariance matrices. Before introducing the random matrix approach, we first describe a deterministic approach to approximate eigendecomposition. Various algorithms that approximate eigencomponents using a submatrix of the original matrix are compared in Homrighausen and McDonald (2016). In our implementation, we used the Nyström method (Williams and Seeger, 2001; Drineas and Mahoney, 2005). Let \( K \) denote an \( n \times n \) positive definite matrix to be decomposed; we can further denote its partition as \( K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \), where \( K_{11} \) is \( k \times k \) dimensions. The central idea of Nyström’s method is to compute exact eigendecomposition on the lower-dimensional submatrix \( K_{11} \), then use the resulting lower-dimensional eigencomponents to approximate eigencomponents of \( K \). Let \( V_c(A) \) and \( \Lambda_c(A) \) be matrices of the first \( c \) eigenvectors and eigenvalues, respectively, of a positive definite matrix \( A \); therefore, both have \( c \) columns. Then \( \Lambda_k(K) \) is approximated by \( \tilde{\Lambda}_k(K) = \frac{n}{k} \Lambda_k(K_{11}) \), a \( k \times k \) diagonal matrix whose elements are the approximated eigenvalues and \( V_k(K) \) is approximated by scaling \( V_k(K_{11}) \) up to high dimensions via

\[
\tilde{V}_k(K) = \sqrt{\frac{k}{n}} \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} V_k(K_{11}) \Lambda_k(K_{11})^{-1} = \sqrt{\frac{k}{n}} \begin{bmatrix} V_k(K_{11}) \\ K_{21}V_k(K_{11}) \Lambda_k(K_{11})^{-1} \end{bmatrix}.
\]
From the Nyström method, we also obtain an approximation to $K$ by
$$
\tilde{K} = \tilde{V}_k(K)\tilde{\Lambda}_k(K)\tilde{V}_k(K)^T = \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} K_{11}^{-1} [K_{11}, K_{12}].
$$
The error in the approximation is $||K_{22} - K_{21}K_{11}^{-1}K_{12}||$, which reflects the information lost from truncating $K$ (Belabbas and Wolfe, 2009). However, the approximated eigenvectors above are not guaranteed to be orthogonal, hence we adopt a slight variant of the form (similar to Algorithm 5.5 in Halko et al., 2011). Let $C$ denote
$$
\begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix} V_k(K_{11})\Lambda_k(K_{11})^{-1/2};
$$
then, its singular value decomposition is $\text{SVD}(C) = U(C)S(C)V(C)^T$, where $S(C)$ is a $k \times k$ diagonal matrix with elements equal to the non-zero singular values, $U(C)$ is an $n \times k$ matrix of the left singular vectors, and $V(C)$ is a $k \times k$ matrix of the right singular vectors of $C$. (From here on, we suppress the dependencies of $U$ and $S$ on $C$.) Then $\tilde{K}$ can also be expressed as $CC^T = U S^2 U^T$, satisfying $\tilde{K}U = US^2$.

Therefore, $U$ and $S^2$ are the first $k$ eigencomponents of $\tilde{K}$, and they are used as our approximation to the eigencomponents of $K$, respectively. This Nyström’s approximation is summarized as step 2 in Algorithm 1.

Nyström’s method obtains the column space of $K$ from subsampling its columns
$$
K\Phi = \begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix},
$$
where $\Phi$ is an $n \times k$ matrix by permuting the rows of $[I_{k \times k}, 0_{k \times (n-k)}]^T$; once $\Phi$ is fixed, the approximation is deterministic. Alternatives to approximate the column space of $K$ involving randomness have been proposed, such as weighted random subsampling of the rows and columns of $K$ by Frieze et al. (2004), subsampled randomized Hadamard transform by Tropp (2011) or a random projection matrix with iid elements (Halko et al., 2011; Bingham and Mannila, 2001; Banerjee et al., 2012). Here we adopt the latter method with iid Gaussian random variables.

Rather than truncating $K$, we take $\Phi$ to be $K^\alpha \Omega$, where $\Omega$ is a $n \times k$ random matrix with $\Omega_{ij} \sim N(0, 1/\sqrt{k})$, and $\alpha = 0, 1$, or 2 takes a small non-negative integer value for improving approximation (see a comparison for $\alpha$ in Section 2.4.2). Then
\( K \Phi \) is randomly weighted linear combination of columns of \( K \), by construction, it approximates the column space of \( K \). The random matrix \( \Omega \) is a low-dimensional embedding: \( R^{n \times n} \rightarrow R^{n \times k} \), that satisfies Johnson-Lindenstrauss’s transformation; it has a low distortion such that \( \| \Omega^T v \| - \| v \| \) is small for all \( v \in V \subset R^n \) with high probability (for details on the embedding, we refer the readers to Dasgupta and Gupta, 2003). Taking small powers of \( K \) in the projection matrix \( \Phi = K^\alpha \Omega \), enhances our approximation performance but this involves a tradeoff in terms of computational speed; in our implementation, we see substantial improvement by taking \( \alpha = 1 \) or 2. We let \( k = m + l \), where \( m \) denotes the target rank, and \( l \) is an oversampling factor typically set to 5 or 10 to reduce approximation error (Halko et al., 2011). In our implementation, we noticed that taking small \( l \) is not enough to give a good approximation of the eigenvectors corresponding to the smaller eigenvalues; therefore, we take \( l = m \). The random projection approach to approximate the column space of \( K \) is summarized as step 1 in Algorithm 1.

In the context of Gaussian process regression, Banerjee et al. (2012) used a similar random projection algorithm to approximate the covariance matrix \( \Sigma \). Here we will directly approximate the eigencomponents of the correlation matrix \( R(\phi) \), because multiplying \( \sigma^2 \) does not affect the approximation. In our MCMC implementation, we obtain \( U_m \) and \( D_m \), the approximated leading \( m \) eigencomponents of \( R(\phi) \), using Algorithm 1 for every \( \phi \) value. Then we can obtain reparameterized random effects \( \delta | \phi = (U_mD_m^{-1/2})^TW \).

### 2.4.2 Approximation Comparison

To illustrate the performance of introducing randomness in approximating eigencomponents using Nystöm’s method as described in Section 2.4.1, we perform a numerical experiment to compare the approximation performance of the leading \( m \)
Algorithm 1 Random projection algorithm:
Given a positive semi-definite matrix $K$, this algorithm approximates the leading $m$ eigencomponents of $K$ by utilizing Nyström’s method.

1. Low dimensional projection from $R^{n \times n}$ to $R^{n \times k}$, $m < k << n$:
   Form $\Phi = K\Omega$ where $\Omega \in R^{n \times k}$ with $\Omega_{ij} = N(0, 1/\sqrt{k})$

2. Nyström’s method to approximate eigendecomposition:
   Form $K_{11} = \Phi^T K \Phi$
   SVD for $K_{11}$: $V_k(K_{11})\Lambda_k(K_{11})V_k(K_{11})^T$
   Form Nyström extension $C = [K\Phi][V_k(K_{11})\Lambda_k(K_{11})^{-1/2}]$
   SVD for $C$: $USV^T$

3. Take the first $m$ columns of $U$, and the first $m$ diagonal elements of $S^2$ as our approximation to the leading $m$ eigencomponents of $K$

eigenvectors and eigenvalues of $K$. Direct comparison of eigenvectors are difficult, because they are only uniquely defined up to a sign change. Let $\triangle(U, V)$ denote the distance between two subspaces $U$ and $V$. Here we follow Homrighausen and McDonald (2016) and define

$$\triangle(U, V) = \| \Pi_U - \Pi_V \|_F,$$

where $\Pi_U$ and $\Pi_V$ are the orthogonal projection associated with $U$ and $V$, respectively. The smaller the distance between the subspaces generated by the approximated eigenvectors $U_m$ and the true eigenvectors $V_m$, the better the approximation.

To measure the approximation performance of the eigenvalues, for any vector containing the estimated leading eigenvalues in descending order, we compare it to the true leading eigenvalues $\lambda$ using $\| \hat{\lambda} - \lambda \|_2$. We simulate 1000 random locations in unit domain; based on these data points, we compute the correlation matrices $K$ using the Matérn covariance function with $\nu = 0.5$ or 2.5, and $\phi = 0.3$. Figure 2.1 shows the approximation results when the projection matrix $\Phi$ is $[I_{k \times k}, 0_{k \times (n-k)}]^T$ with rows permuted, or $K^\alpha \Omega$ with $\alpha = 0, 1, 2$. We see that when random projec-
tions are used the approximation is improved. In addition, there are advantages in taking $\Phi$ to be $K^\alpha \Omega$, where in practice $\alpha = 1$ appears to be a good choice.

Figure 2.1: Approximation performance comparison shows that introducing randomness in $\Phi = K^\alpha \Omega$ improves the Nystöm approximation to the eigenvectors (left column) and eigenvalues (right column). Letting $\Phi$ to be $K^\alpha \Omega$ with small power $\alpha = 1, 2$ also provides better approximation for both $\nu = 0.5$ (first row) and $\nu = 2.5$ (second row).

2.4.3 Random Projection for Spatial Linear Mixed Models

Here we illustrate the random projection approach for a spatial linear mixed model (SLMM) with an emphasis on dealing with confounding. Banerjee et al. (2012) proposes using random projection for efficient Gaussian process regression. In this subsection we extend their approach so it applies to both SLMMs and the linear
restricted spatial regression model. This description also serves as an introduction to our more general approach to SGLMMs.

For the linear case, model fitting is based on the marginal distribution of \( Y | \beta, \phi, \sigma^2, \tau^2 \). The main computational challenge is therefore due to the expense in calculating inverses and determinants for large covariance matrices. Random projection may be used to approximate the correlation matrix using its principal components. To fit the full model with random projection (FRP), we apply Algorithm 1 to approximate \( R(\phi) \approx \tilde{R}(\phi) = U_m D_m U_m^T \). We rewrite the model as follows,

\[
Y = X\beta + U_m D_m^{1/2} \delta + \epsilon, \quad \epsilon \sim N(0, \tau^2 I), \quad \delta \sim N(0, \sigma^2 I).
\]

Marginally: \( Y | \beta, \phi, \sigma^2, \tau^2 \sim N \left( X\beta, \sigma^2 \tilde{R}(\phi) + \tau^2 I \right) \). (2.6)

Analogously, our RSR model with random projection (RRP) is

\[
Y = X\beta + P_{[X]}^{-1} U_m D_m^{1/2} \delta + \epsilon, \quad \epsilon \sim N(0, \tau^2 I), \quad \delta \sim N(0, \sigma^2 I).
\]

Marginally: \( Y | \beta, \phi, \sigma^2, \tau^2 \sim N \left( X\beta, \sigma^2 P_{[X]}^{-1} \tilde{R}(\phi) P_{[X]} + \tau^2 I \right) \). (2.7)

Hereafter, \( \tilde{R}(\phi) \) will be referred to as \( \tilde{R} \) to suppress its dependency on the unknown parameter \( \phi \). Fitting the FRP model (2.6) involves evaluating the inverse and determinant of \( \sigma^2 \tilde{R} + \tau^2 I = \sigma^2 U_m D_m U_m^T + \tau^2 I \). Then, by the Sherman-Morrison-Woodbury identity (Harville 1997), we have \( (\sigma^2 U_m D_m U_m^T + \tau^2 I)^{-1} = \tau^{-2} I - \tau^{-2} U_m (\sigma^2 D_m^{-1} + \tau^{-2} U_m^T U_m)^{-1} U_m^T \tau^{-2} \). The matrix inversion of \( \sigma^2 D_m^{-1} + \tau^{-2} U_m^T U_m \) can be further reduced to inverting an \( m \times m \) diagonal matrix \( \sigma^2 D_m^{-1} + \tau^{-2} I \) with cost of \( m \) flops, since \( U_m \) has orthonormal columns. The determinant calculation can also be simplified. By the determinant lemma (Harville 1997), \( |\sigma^2 \tilde{R} + \tau^2 I| = |\sigma^2 D_m^{-1} + \tau^{-2} I| \times |\tau^2 I| \times |D_m| = \prod (\sigma^2 D_{m,ii}^{-1} + \tau^{-2}) \times \tau^{2n} \times \prod D_{m,ii} \). Similarly, fitting the RRP model (2.7) in-
volves calculating the inverse and determinant of $\sigma^2 P_{[X]} P_{[X]}^\perp \tilde{R} P_{[X]}^\perp + \tau^2 I$, for which the dominant cost is tied to the $m \times m$ matrix $\sigma^2 D_m^{-1} + \tau^{-2} U_m^T P_{[X]} P_{[X]}^\perp U_m$.

For the linear case we can simply approximate the correlation matrix of the random effect $R$ with $\tilde{R}$ without explicitly reparameterizing the random effects. However, for SGLMMs, we do not have closed-form marginal distribution. It is therefore necessary to obtain the reduced random effects $\delta$ and carry out inference based on $\pi(\theta, \beta, \delta \mid Z)$.

### 2.4.4 Random Projection for Spatial GLMMs

Here we describe how to reparameterize and reduce the dimension of our models such that the resulting model is easier to fit and preserves the desirable properties of the original model. We do this for both cases, first where confounding may not be an issue and the second where we want to address confounding.

We apply Algorithm 1 to $R$ to obtain $U_m$ and $D_m$. If confounding is not an issue, we replace random effect $W$ with $U_m D_m^{1/2} \delta$. The SGLMM (2.3) may be rewritten as

$$g \{E(Z_i \mid \beta, U_m, D_m, \delta)\} = X_i \beta + (U_m D_m^{1/2})_i \delta,$$

$$\delta \mid \theta \text{ approx } \sim N(0, \sigma^2 I). \quad (2.8)$$

We refer to this as the full model with random projection (FRP). Essentially, the spatial dependence in $W$ is transformed into a reduced-dimension spatially independent variable $\delta$ and synthetic spatial variable $U_m D_m^{1/2}$. This combines the idea of spatial filtering (Getis and Griffith, 2002) and PCA, thereby reducing the dimension of the posterior distribution. By also reducing the correlations among the parameters, our approach improves the mixing of MCMC algorithms for sampling from the posterior. Once priors $p(\beta, \theta)$ are specified, we can sample from the full conditionals using Metropolis-Hastings random-walk updates.
To address the confounding problem, we follow the RSR approach to restrict the random effects to be orthogonal to the fixed effects (Hodges and Reich [2010]). We can project our reduced random effects $U_mD_m^{1/2}\delta$ to the orthogonal span of $X$. The restricted model with random projection (RRP) can be summarized as follows:

$$g\{E(Z_i \mid \beta, U_m, D_m, \delta)\} = X_i\beta + (P_{[X]}^\perp U_mD_m^{1/2})_i\delta,$$

$$\delta \mid \theta \approx \text{N}(0, \sigma^2 I).$$

(2.9)

Fitting RRP is similar to FRP, except that in the data likelihood $\prod_{i=1}^n f(Z_i \mid \beta, U_m, D_m, \delta)$, $U_m$ is replaced by $P_{[X]}^\perp U_m$.

It is tempting to first replace $W$ by $P_{[X]}^\perp W \sim \text{MVN}(0, \sigma^2 P_{[X]}^\perp R_{[X]}^\perp P_{[X]}^\perp)$, then reduce the dimension of random effects by approximating the matrix $P_{[X]}^\perp R_{[X]}^\perp P_{[X]}^\perp$. However the order of approximation and projection affects the inference of the covariance parameter $\phi$. Although the projected eigenvectors of $R$, i.e. $P_{[X]}^\perp U_m$, is the same as the eigenvectors of $P_{[X]}^\perp R_{[X]}^\perp P_{[X]}^\perp$, the ordering of the eigencomponents may change depending on the direction of projection. To better approximate the original random effects $W$, rather than $P_{[X]}^\perp W$, we therefore first approximate the correlation matrix and then perform the requisite orthogonal projection.

### 2.4.5 Random Projection for Areal Data

Our approach reduces the dimension by decomposing its correlation matrix. Hence, it can be easily applied to Gaussian Markov random field models as well. Here we develop FRP and RRP for non-Gaussian areal data. Note that RRP model for areal data is similar to the approach proposed by Hughes and Haran (2013). Both methods adjust for confounding and reduce the dimension of random effects. The only difference is that we decompose the covariance matrix, whereas their decomposition is performed on the Moran operator (details are provided later in
Consider spatial data located on a discrete domain, for instance mortality rates by county across the U.S. If we describe the data locations via nodes on an undirected graph with edges only between nodes that are considered neighbors, we can model the spatial dependence via a Gaussian Markov random field model. To model the dependence of $W = (W_1, ..., W_n)^T$, where the index indicates block, we define the neighboring structure among blocks through an $n \times n$ adjacency matrix $A$ with $\text{diag}(A) = 0$ and $A_{ij} = 1$ if the $i^{th}$ and $j^{th}$ locations are connected, or $A_{ij} = 0$ if they are not connected (Besag et al., 1991). A common model for $W$ is an intrinsic conditionally auto-regressive (ICAR) or Gaussian Markov Random Field prior:

$$p(W | \tau) \propto \tau^{\text{rank}(Q)/2} \exp\left(-\frac{\tau}{2} W^T Q W\right),$$

where $\tau$ is a smoothing parameter that controls the smoothness of the spatial field, and $Q = \text{diag}(A1) - A$ is the precision matrix ($1$ is a $n$-dimensional vector of 1s).

For the discrete domain, Hughes and Haran (2013) use the eigenvectors of the Moran operator, $M = P_{[X]}^\perp AP_{[X]}^\perp$, to reduce the dimension of the random effects. Their method alleviates confounding while preserving spatial dependence structure implied by the underlying graph. The eigenvectors can be interpreted as spatial patterns corresponding to different degrees of spatial dependency.

We can also fit both FRP and RRP to areal data and achieve similar dimension reduction and computational gains. First we obtain the covariance matrix by taking the Moore-Penrose generalized inverse (Penrose, 1955) of the precision matrix $Q$. Then apply random projection on the covariance matrix $Q^{-1}$, and proceed with either FRP or RRP as described in the preceding section. The eigenvectors corresponding to large eigenvalues represent large-scale spatial variation. The advantage of this PCA approach is that a relative small number of PC’s is enough
to capture most spatial variation. Computationally, our method is not as efficient as Hughes and Haran (2013) because of the extra cost of inverting the precision matrix. However, for the ICAR model, the precision matrix is fixed and defined beforehand, so the inversion and random projection only need to be performed once. The model estimates from RRP are comparable to those obtained by Hughes and Haran (2013).

2.4.6 Rank Selection

Here we provide a general guideline for selecting the rank for projection-based models. From a Bayesian perspective, the rank can be determined by model comparison criteria such as DIC (Spiegelhalter et al., 2002). We can fit several models using different number of ranks and select the one with the smallest DIC. However, to reduce computational time, we recommend the following procedure to select the appropriate rank before fitting either FRP or RRP model.

Since the projection-based models combine the idea of spatial filtering and PCA dimension reduction, we can fit non-spatial generalized linear models with predictors $X$ and synthetic spatial variables $U_m D_m^{1/2}$ for $m = 1, 2, \ldots$, and then select the initial rank based on variable selection criterion such as BIC. The synthetic spatial variables $U_m D_m^{1/2}$ can be obtained by performing eigendecomposition on the correlation matrix $R(0)$ for an appropriate range value $\phi(0)$, for example, $\phi(0)$ is half of the maximum distance among observations. If performing a full eigendecomposition is computationally infeasible, one may perform random projection using Algorithm 1 to approximate the leading eigencomponents. In our Poisson simulation study, the overall smallest BIC corresponds to $m = 75$ for $\nu = 0.5$, $m = 50$ for $\nu = 1.5$, $m = 40$ for $\nu = 2.5$, and $m = 30$ for $\nu = \infty$. To be careful, we then advocate seeing if increasing the rank leads to a marked improvement
in the model, for instance by again using criteria like DIC. If it does, it may be useful to try increasing the rank again; if there is not much change, we can stop and simply use the current rank. For each of the above smoothness values, we also conduct simulation studies using several ranks to study the effectiveness of the rank selection method and the performance of our projection-based approach. Our simple heuristic appears to work well in our simulation studies, the DIC values from model fits agree with BIC selection and the prediction performance increases by only a small margin above the selected rank. We find that it is not necessary to repeatedly fit many models with increasing rank. If researchers want a more rigorous comparison of models with different ranks and are willing to implement Bayes factor calculations, comparing Bayes factors would be a useful alternative to what we have proposed.

2.4.7 Computational Gains

The advantages of our reparameterization schemes are shorter computational time per iteration and less MCMC iterations to achieve convergence. These result from: (1) reducing large matrix operations, (2) reducing the number of random effects, and (3) improving MCMC algorithm mixing. Although the main computational cost of our approach is of order $O(n^2m)$ from applying Algorithm 1, it is dominated by matrix multiplications that can be easily parallelized by multi-core processors. Leveraging parallel computing for matrix multiplication, the remaining dominant cost of fitting our model is of order $O(nm^2)$ due to the singular value decomposition of $n \times m$ matrices. To illustrate the computational gain of the projection-based models, we fit both the SGLMM and the projection-based models to simulated Poisson data. We fit the SGLMM using one-variable-at-a-time Metropolis-Hastings random-walk updates. To fit the projection-based models, we update random
effects $\delta$ in a block using spherical normal proposal; simple updating scheme for $\delta$ is sufficient because it has a smaller dimension and are decorrelated. In our implementation, Algorithm [1] is coded in C++ using Intel’ Math Kernel Library BLAS and LAPACK routines for matrix operations; the MCMC is written in programming language R [R Core Team 2014]. All the code was run on National Center for Atmospheric Research’s Yellowstone supercomputer [Computational and Information Systems Laboratory, 2016].

To see the improvement in MCMC mixing, we compute the effective sample size (ESS) using the R coda package [Plummer et al. 2006]; it provides the number of independent samples roughly comparable to the number of dependent samples produced by the MCMC algorithm, therefore a larger ESS implies better Markov chain mixing. Based on our results, the projection-based models have better mixing, for example, the univariate ESSs of the RRP are, on average, 12 times larger than the ones from the SGLMM and three times the ones from the predictive process (using the R package by Finley et al. 2013) for the same number of MCMC iterations. The mixing improvement of our projection-based models is implied from the a posteriori correlations (Figure 2.2); our projection-based models (both FRP and RRP) produce weakly correlated random effects compared to the predictive process. The improvement in computational time is illustrated in Figure 2.3. The time required increases dramatically for SGLMM as the data size increases, however we can still fit the random projection model in a reasonable amount of time. We also compute ESS per second to compare MCMC efficiency; our RRP model more than 120 times more efficient than the SGLMM.
Figure 2.2: Pairwise cross-correlations are close to zero for both projection-based approaches (FRP and RRP). This is true for each pair of random effects (right) and between fixed and random effects (left). In contrast, the cross-correlations for the predictive process (black curves) are much larger. Hence our approaches result in better MCMC mixing. FRP = full model with random projection, RRP = restricted model with random projection and PP = predictive process model.

Figure 2.3: Computational time for $10^5$ iterations versus data size for SGLMM and projection-based approach (both FRP and RRP) with rank 50. This illustrates the benefit of the projection-based approach (both FRP and RRP) in terms of computational time. FRP = full model with random projection, and RRP = restricted model with random projection.
2.5 Simulation Study and Results

In this section, we apply our approaches to simulated linear, binary and Poisson data. For each case, we simulate 100 data sets where the locations are in the unit domain \([0,1]^2\). We fit both FRP and RRP models to simulated data with size of \(n = 1000\) at random locations, then make predictions on a \(20 \times 20\) grid. We adjust the regression parameters of RRP using equation (2.5) (denote this adjusted inference A-RRP). Throughout the simulation study, we let \(X\) be the xy-coordinate of the observations and \(\beta = (1,1)^T\). We simulate \(W\) from the Matérn covariance function with \(\theta = (\nu,\sigma^2,\phi)^T = (2.5,1,0.2)^T\), which has the form as below (Rasmussen and Williams, 2005, Section 4.2):

\[
C(h) = \sigma^2 \left( 1 + \frac{\sqrt{5}|h|}{\phi} + \frac{5|h|^2}{3\phi^2} \right) \exp \left( -\frac{\sqrt{5}|h|}{\phi} \right)
\]

We use a vague multivariate normal prior \(N(0,100I)\) for regression coefficients \(\beta\), inverse gamma prior \(IVG(2,2)\) for \(\sigma^2\) and uniform prior \(U(0.01,1.5)\) for \(\phi\). We have experimented with different choice of prior; the inference performances are similar. To evaluate our approaches, we compare inference performance with a focus on the posterior mean estimates and 95% equal-tail credible intervals of \(\beta\), and we compare prediction performance based on mean square error.

2.5.1 Linear Case

The random projection models are first assessed under the linear case. Let \(x_1, x_2\) denote the xy-coordinates. We simulate data from

\[
Y = x_1 + x_2 + W + \epsilon, \quad \epsilon \sim N(0,\tau^2I), \quad W \sim MVN(0,\sigma^2R(\phi)),
\]
where the noise $\epsilon$ has variance $\tau^2 = 0.1$.

We fit both FRP and RRP models using rank $m = 50$ based on the marginal distribution of $Y \mid \beta, \phi, \sigma^2, \tau^2$ as described in (2.6) and (2.7), and we use $IVG(2, 1)$ prior for $\tau^2$. For the linear case, fitting the full SLMM and RSR model is fast for data of size $n = 400$, so we fit both of our projection-based approaches, the spatial linear mixed model and restricted spatial regression model for overall comparisons. Our results from 100 simulated data sets show that inference and prediction provided by the random projection models are similar to the original models they approximate (Table 2.1). As noted by Hanks et al. (2015), when the data are simulated from the full SLMM, we see a low $\beta$ coverage for the RSR model; therefore, its approximated version RRP also has a low coverage. However, this problem is resolved after a simple adjustment (A-RRP) as recommended by Hanks et al. (2015). Also, the distribution of $\beta$ estimates all center around the true value and are comparable among all four models (Figure 2.4).

We also conduct a simulation study for larger data size $n = 1000$, and we fit both FRP and RRP with rank $m = 50$. Our results show that the distributions of $\beta$ estimates for both FRP and RRP are centered around the true value, and the distributions are comparable. Coverage of 95% credible intervals for FRP and A-RRP are comparable to the nominal rate. For prediction performance, the mean square error is similar for both models and the predicted observations at testing locations recover the spatial patterns well (Figure 2.5).

### 2.5.2 Binary Case

The main goal of our approximation method is to fit spatial generalized linear mixed models for large data sets. Here we examine our model performance under the binary case generated with a logit link function $\text{logit}(p) = \log \{p/(1 - p)\}$. 
Table 2.1: Model comparisons for linear case with \( n = 400 \).

<table>
<thead>
<tr>
<th></th>
<th>SLMM</th>
<th>FRP</th>
<th>RSR</th>
<th>RRP</th>
<th>A-RRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 ) (coverage)</td>
<td>1.01 (0.99)</td>
<td>0.98 (0.97)</td>
<td>1.00 (0.07)</td>
<td>1.00 (0.07)</td>
<td>1.00 (0.97)</td>
</tr>
<tr>
<td>( \beta_1 ) mse</td>
<td>0.39</td>
<td>0.46</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>( \beta_2 ) (coverage)</td>
<td>1.02 (0.95)</td>
<td>1.01 (0.95)</td>
<td>1.02 (0.03)</td>
<td>1.02 (0.03)</td>
<td>1.02 (0.94)</td>
</tr>
<tr>
<td>( \beta_2 ) mse</td>
<td>0.60</td>
<td>0.59</td>
<td>1.06</td>
<td>1.06</td>
<td>1.06</td>
</tr>
<tr>
<td>( \phi )</td>
<td>0.21</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
<td>NA</td>
</tr>
<tr>
<td>( \phi ) mse</td>
<td>0.62</td>
<td>0.61</td>
<td>0.62</td>
<td>0.63</td>
<td>NA</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>1.25</td>
<td>1.34</td>
<td>1.24</td>
<td>1.20</td>
<td>NA</td>
</tr>
<tr>
<td>( \sigma^2 ) mse</td>
<td>1.32</td>
<td>1.54</td>
<td>1.26</td>
<td>1.18</td>
<td>NA</td>
</tr>
<tr>
<td>pmse</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>NA</td>
</tr>
</tbody>
</table>

Figure 2.4: Distribution of posterior mean estimates of \( \beta \) among four models and with adjustment. The distributions all center around the true value and are comparable. Random projection models FRP and RRP with rank=50 produce results that are similar to the models they approximate.

We compare two simulation schemes: the confounded case \( \eta = x_1 + x_2 + W \), and the orthogonal case \( \eta = x_1 + x_2 + P_{[X]} W \). For both cases we use the same parameter values as the linear case and simulate \( W \) from \( N(0, \sigma^2 R(\phi)) \). We consider two simulation schemes because in practice we do not know whether there are spatial latent variables that may be collinear with our covariates. A careful approach, therefore, involves fitting both FRP (2.8) and RRP (2.9) models under both schemes to get a fair assessment of the FRP and RRP approaches. Because it is hard to fit full SGLMMs and the RSR models for moderate data size, we...
compare only FRP and RRP models for 100 simulated datasets of size $n = 1000$. Although we do not have a comparison with the original model fit, we can look at how well the true parameters are recovered and compare the prediction mean square error to judge the projection-based models.

Our simulation results show that under the confounded case, $\beta$ estimates for both FRP and RRP have similar distributions (Figure 2.6). However the coverage of RRP, about 41%, is much lower than the 95% nominal rate. This is because the credible intervals obtained under the RRP are similar to the ones for RSR models, which are likely to be inappropriately narrow (Hanks et al. 2015); the mean length of the credible intervals (with 95% intervals) under RRP model is 0.84(0.70, 1.09) compared to 4.15(2.61, 7.01) under the FRP model. However this problem is resolved after the adjustment; the coverage of A-RRP is comparable to the nominal rate and its interval length is 4.17(2.94,7.08), similar to the one
from the FRP model. Under the orthogonal case, in contrast, RRP performs much better than FRP. The point estimates from RRP are distributed tightly around the true values (Figure 2.6). Its credible interval has better coverage than the FRP; they are 94.95% and 100% respectively. Moreover, RRP has much narrower credible intervals, it is 0.81(0.733, 0.95) compared to 4.11(2.75, 6.66) under the FRP. And again the adjusted inference A-RRP is similar to that of FRP. Figure 2.7 shows the estimated probability surface for the binary field at the training locations and the predicted probability surface at the testing locations under the confounded simulation scheme. We see that our projection-based approaches work well in recovering the true spatial pattern (results for the orthogonal simulation scheme are similar, hence not shown). Although the predictive surface seems somewhat smoother than the true surface, this could be because binary outcomes do not provide enough information for the latent variable.

2.5.3 Poisson Case

We also examine our model performance under the Poisson case. The results are similar to the Binary case. We simulate 100 Poisson data sets with a natural logarithm link function using the same parameter values as the linear case for data sizes of \( n = 1000 \); again, both simulation schemes are considered. Under the confounded simulation scheme, FRP and RRP have similar distributions for point estimates (Figure 2.8); however, the credible interval (CI) of RRP is inappropriately narrow with length 0.246 (0.166, 0.367) and a coverage of 14% compared to the FRP, the CI of which has length 3.123 (1.664, 5.743) and a coverage of 91%. Under the orthogonal simulation scheme, RRP performs much better than FRP; its point estimates are closely centered around the true value (2.8), its CI is 0.225 (0.176, 0.299), much narrower compared to 2.985 (1.666, 4.995) of the FRP, and
both RRP and FRP have coverages that are comparable to the nominal rate.

For both cases, the adjusted inference A-RRP is similar to the FRP, hence we can fit only the RRP model for its computational benefits and recover the results for fitting FRP. Figure 2.9 shows the estimated expectation of the Poisson process (log scale) at the training locations and the predicted expectation (log scale) at...
Figure 2.7: First row shows the estimated probability surface at all training locations. Second row shows the predicted probability surface on a 20x20 grid using random projection models. Left column is simulated data, middle column shows the FRP, and right column shows the RRP. Comparing the patterns from the true model (left column) to the ones from projection-based models using rank $m = 50$ (middle and right columns), we see the projection-based models are able to recover the true value quite well. FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.

the testing locations under the confounded simulation scheme. We see that the projection-based models work well in recovering the true (results for the orthogonal simulation scheme are similar, hence not shown).

2.5.4 Comparison to Predictive Processes

The predictive process approach \cite{Banerjee} has been very influential among reduced rank approaches. In the context of spatial generalized linear mixed models for non-Gaussian data, we believe our approach offers some benefits over the predictive process approach: (i) We avoid having to choose the number and
Figure 2.8: Poisson simulation study: distribution of $\beta$ posterior mean estimates for RP models and after adjustment. First row for the confounded case, and second row for the orthogonal case. All distributions center around the true value. For the confounded case (top row), FRP and RRP have similar results; while under the orthogonal case (bottom row), RRP produces much tighter distribution.

locations of knots. Instead, our approach requires specifying the rank for which we have a heuristic; no further user specifications are required. (ii) We provide an approach to easily alleviate spatial confounding. (iii) The reparameterization in our projection-based approach results in decorrelated parameters in the posterior, thereby allowing for a faster mixing MCMC algorithm.

To compare the performance of our projection-based approaches with the predictive process, we simulate 100 Poisson data sets from the traditional SGLMM.
Figure 2.9: Simulated Poisson example: First row shows the estimated Poisson mean surface (in log scale) at all training locations. Second row shows the predicted Poisson mean surface (in log scale) on a 20x20 grid using projection-based models. Left column is simulated data, middle column shows the FRP, and right column shows the RRP. Comparing the pattern from the true model (left column) to the ones from our models using rank $m = 50$ (middle and right columns), we see the projection models are able to recover the true value quite well. FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.

We fit both FRP and RRP with rank $m = 50$ to the datasets, and compare their results with the predictive process with reference points on a $7 \times 7$ grid. In this simulation study, our projection-based approaches provide comparable inference and smaller mean prediction square error (MPSE) (Figure 2.10).
Figure 2.10: Poisson simulation study: compare projection-based models with predictive process on gridded knots. The point estimate distributions for $\beta_1$ are comparable (top left); while the coverage for RRP is much lower than the others, however after adjustment, the coverage for A-RRP is corrected and is comparable to FRP and PP (top right). Both FRP and RRP have better prediction performance than PP (bottom left). The length of the CIs for FRP and PP are comparable, while RRP produce much narrower CI; but after adjustment the CI gets much wider (bottom right). FRP = full model with random projection, RRP = restricted model with random projection, A-RRP = adjusted inference for RRP.
2.6 Applications

2.6.1 Binary Data Application

We apply our approach to classify rock types using a reference synthetic seismic data set. Fluvsim is a computer program that produces realistic geological structures using a sequential scheme; it is used for modeling complex fluvial reservoirs (Deutsch and Wang, 1996). The high-resolution 100x120x10 3-dimensional grid data set is simulated from the program fluvsim conditioning on well observations. A similar reference data set obtained from fluvsim has been used to test the classification method in John et al. (2008). Here we illustrate our projection-based approach on one layer of the rock profile. In the data set, there are five rock types: crevasse, levies, border, channel and mud stone. We combine crevasse and mud stone as one group and treat the rest as the other group for binary classification. Along with the rock type data, we also have acoustic impedance data that is associated with the rock properties; it is desirable to identify rock types from seismic-amplitude data using statistical methods (John et al., 2008).

We fit both FRP and RRP models at 2000 randomly-selected locations and predict the rock profile on a 24x30 grid. Prior to fitting the projection-based models, the rank is selected by fitting non-spatial logistic regression models with synthetic spatial variables as described in Section 2.4.6. The BIC values from the resulting models suggest that rank $m = 50$ is sufficient. In order to help diagnose convergence, we ran multiple chains starting at dispersed initial values and compared the resulting marginal distributions while also ensuring that the MCMC standard errors for the expected value of each parameter of the distribution was below a threshold of 0.02 (cf. Flegal et al., 2008).

Estimated coefficients corresponding to x-coordinates, y-coordinates and impedance
covariates differ between FRP and RRP models, which are \((-0.717, 0.448, -0.233)^T\) and \((-1.587, 2.435, -0.396)^T\), respectively. Although interpretations for the estimates are slightly different the predictions, which are of primary interest, are identical between the two models (see Figure 2.11). We also assess the predicted rock profile using higher ranks; however, the results are similar to using rank 50, hence they are not shown here. The time to fit either the FRP or RRP model is about 10 hours, whereas for the full model, it would have taken about three weeks to run the same number of MCMC iterations. In general, fitting SGLMM to binary observations is harder due to the poor Markov chain mixing; therefore comparison with the full model is prohibitively expensive.

![Figure 2.11: Predicted rock profile by restricted model with random projection (right) and full model with random projection (left) using rank \(m = 50\).](image)

### 2.6.2 Count Data Application

Here we illustrate the usefulness of the projection-based models in the context of an environmental study. We consider the relative abundance of house finch (*Carpodacus mexicanus*), a bird species that is native to western North America ([Elliott and Arbib Jr, 1953](Elliott, 1953)). Figure 2.12 shows the number of bird counts obtained
in 1999 by the North American Breeding Bird Survey with the size of the circle is proportional to the number of counts. The bird surveys are obtained along more than 3,000 routes across the continental US. There are 50 stops per route, spaced roughly 0.5 miles apart. The observer make a three-minute point count at each stop. The bird count is then the total number of birds heard or seen for all 50 stops (Pardieck et al., 2016).

The data set being analyzed has 1257 highly irregular sampling locations. Here we fit the FRP model to approximate the SGLMM with only the intercept term for spatial interpolation. The time to fit FRP is about 7 hours, while the full model would take almost 2 days for the same number of MCMC iterations. Figure 2.12 shows the abundance map predicted by FRP on a high resolution of 40 x 100 grid. Not surprising, the abundance map is smooth. This reflects that the bird counts are very small in the center and most of the east coast of the US. Our map is also consistent with the observation that large counts are centered near New York area and the West Coast.

Figure 2.12: Data on abundance of house finch in the US from North American Breeding Bird Survey Dataset (left). Size of the circle is proportional to the bird counts. Predicted bird counts on a grid using the random projection method (right). The large number of observed counts on west coast and lack of observations in the central mid-west region is reflected on the prediction map.
2.7 Discussion

In this chapter, we have proposed projection-based models for fast approximation to SGLMMs and RSR models. Our simulation study shows that our low rank models have good inference and prediction performance. The advantages of our approach include: (1) a reduction in the number of random effects, which lowers the dimensionality of the posterior distribution and decreases the computational cost of likelihood evaluations at each iteration of the MCMC algorithm; (2) reparameterized and therefore approximately independent random effects, resulting in faster mixing MCMC algorithms; and (3) the ability to adjust for spatial confounding.

Our simulation study shows both the restricted and unrestricted models provide similar results in prediction. RRP provides superior inference when the true model does not have confounding (and hence the spurious confounding effect needs to be removed); it is also computationally more efficient due to its faster mixing. Therefore, we recommend that in general users fit RRP models. If there is concern that the true model may actually exhibit confounding, we recommend adjusting the fixed effects \textit{a posteriori} to recover the inference from FRP as recommended in \cite{Hanks2015}. As we demonstrate here, this is easy to do in practice.

The current methods rely on parallelization to handle large matrix computations; we have successfully carried these out for \( n \) of around 10,000. If we combine parallelization with a discretization of possible values of \( \phi \) (to allow for pre-computing the eigendecomposition of the covariance matrix), this approach will likely scale to tens of thousands of data points. The INLA approach \cite{Rue2009} provides a fast approximate numerical method for carrying out inference for latent Gaussian random field models. An interesting avenue for future research is combining our reduced-dimensional reparameterization with INLA.

There have been a number of recent proposals for dimension reduction and com-
putationally efficient approaches for spatial models. These include the fixed rank approximation by Cressie and Johannesson (2008), predictive process by Banerjee et al. (2008) and random projection approach for the linear case Banerjee et al. (2012). Our approach can be thought of as a fixed rank approach, but we use the approximated principal eigenfunctions as our basis. The advantage is that we have independent basis coefficients and our approximation minimizes the variance of the truncation error as described in Section 2.4. Our approach is also related to the predictive process in that we effectively subsample random effects (see discussion in Banerjee et al. 2012). Developing extension of this methodology to spatial-temporal and multivariate spatial processes may provide fruitful avenues for future research.

Although this method is much faster than the traditional method, it still takes several hours to fit. This is because for every proposed \( \phi \) value, we approximate the eigencomponents of the correlation matrix \( R(\phi) \). Although the approximation is fast and parallelized, repetitive approximation for every MCMC iteration slows down the method. Also, the eigenvectors are unique up to a sign change, therefore, when approximating the eigenvectors for a new proposed range value \( \phi^* \), we choose the sign of the eigenvector so that \( \langle u_i, u_i^* \rangle > 0 \), for \( i = 1, \ldots, m \); here, \( u_i \) and \( u_i^* \) denote the \( i^{th} \) eigenvectors corresponding to \( R(\phi) \) and \( R(\phi^*) \). The above two caveats motivate the development of a Markov chain Monte Carlo expectation maximization algorithm in Chapter 3.
Chapter 3

Fast Maximum Likelihood Inference
for Projection-Based Spatial Models

The projection-based spatial modeling approach developed in the previous chapter appears to be a promising method for high-dimensional non-Gaussian data. In particular, the Bayesian methods that we develop are much faster than fitting Bayesian approaches for traditional spatial generalized linear mixed models. In this chapter, I consider the challenge of developing a maximum likelihood approach for fitting SGLMMs; maximum likelihood has not been widely used in this context largely due to computational challenges. I develop a fast projection-based SGLMM fitting procedure using the Markov chain Monte Carlo expectation-maximization algorithm; this provides a maximum likelihood alternative to the Bayesian approach. The algorithm proposed here is highly automated and computationally efficient for fitting high-dimensional non-Gaussian data in both continuous and discrete spatial domains.
3.1 Introduction

Non-Gaussian spatial data arise in a number of disciplines, for instance when modeling disease incidence in epidemiology (see, for example, [Diggle et al., 1998, Hughes and Haran, 2013]) or modeling weed counts and plant disease in agriculture ([Christensen and Waagepetersen, 2002, Zhang, 2002]). Spatial generalized linear mixed models (SGLMMs) are a convenient and flexible model for such data. Following two seminal papers, [Diggle et al., 1998] and [Besag et al., 1991], SGLMMs have been very popular, not only in mainstream statistics but also in many other disciplines. These models are useful both for data observed on a continuous spatial domain, such as at irregularly-positioned sampling locations ([Diggle et al., 1998]) and data observed on a discrete spatial domain such as county-level data ([Besag et al., 1991]). In this article, we propose a fast maximum likelihood inference method for both the continuous and lattice cases for large data sets.

Inference for SGLMMs is commonly carried out under the Bayesian paradigm (see [Banerjee et al., 2014, Haran, 2011]). However, constructing efficient Markov Chain Monte Carlo (MCMC) samplers for fitting such models to large data sets is often challenging. There are two major computational challenges: (1) computational issues due to high-dimensional random effects that are typically heavily cross-correlated – these often result in slow mixing Markov chain Monte Carlo algorithms; (2) expensive calculations involving large matrices. An additional issue is spatial confounding between fixed and random effects – this can result in slow mixing and problems with parameter interpretation ([Guan and Haran, 2017, Reich et al., 2006, Hughes and Haran, 2013]). Under a Bayesian framework, the high-dimensional computational challenges for SGLMMs have been addressed via the predictive process approach ([Banerjee et al., 2008], the MCMC mixing issues have been addressed by various reparameterizations (cf. [Christensen et al., 2006, Haran](#))
et al., 2003; Rue and Held, 2005), and the confounding issues have been addressed in Reich et al. (2006). Rue et al. (2009) provides a fast inferential approach based on nested Laplace approximations, and Lindgren et al. (2011) suggest how this approximation may be adapted to continuous domain SGLMMs. Recently, via projection-based methods, Hughes and Haran (2013) and Guan and Haran (2017) have addressed both the above computational issues as well as confounding issues, also using a Bayesian approach.

Our contribution in this manuscript is to provide an efficient projection-based approach that addresses all these issues in a maximum likelihood framework. We develop an efficient Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm for estimation and show that it works well in practice. Our approach provides the ability to fit SGLMMs routinely by (i) having an automated algorithm for estimation, (ii) reducing the computational cost of the estimation algorithm, (iii) addressing spatial confounding issues, and (iv) sidestepping the need to provide hyperpriors for parameters about which there is often little available information. We believe (as applied statisticians ourselves!) that the above characteristics are useful to researchers who use SGLMMs. For problems that involve fitting an SGLMM to a spatial data set in more complicated settings where an additional hierarchy in the modeling framework becomes necessary, for instance where multiple data sets need to be integrated, we would likely revert to a Bayesian approach.

There is a body of work on maximum likelihood inference for SGLMMs, for instance, Monte Carlo maximum likelihood method by Christensen (2004) and Monte Carlo EM by Zhang (2002). Both work provide ways to construct efficient approximations for inference when the data sets are small, but they do not extend easily to large spatial data sets; because, they both require likelihood evaluations
and sampling the latent variables, which are difficult to perform when the data sets are large. More recently Bonat and Ribeiro (2016) develops a fast approximate likelihood-based approach for inference. This substitutes a Laplace approximation for a Monte Carlo step within the E-M algorithm. However, it is unclear how well this approach will work for high-dimensional problems, where Gaussian approximation to the full conditional distribution of the latent variable has high dimensions, and it is not obvious how to adapt their methods to address spatial confounding issues.

The outline of the remainder of the chapter is as follows. In Section 3.2 we describe SGLMMs and spatial confounding. We introduce in Section 3.3 the projection-based SGLMMs and in Section 3.4 the MCMC-EM algorithm for maximum likelihood inference. We study our method via a simulation study in Section 3.5 and apply it to the US infant mortality data set in Section 3.6. We conclude with a discussion and potential areas for future work in Section 3.7.

### 3.2 Spatial Generalized Linear Mixed Models

#### 3.2.1 Models

SGLMMs provide a framework for analyzing spatially dependent non-Gaussian observations. Let $Z(s)$ denote the response variable, $x(s) = (x_1(s), ..., x_p(s))^T$ denote the explanatory variables, and $W(s)$ represent a spatial random field, where $s \in \mathbb{R}^2$ indicates a spatial location. Because data are obtained at a finite collection of locations $\mathcal{S} = \{s_1, \ldots, s_n\}$, we write $Z_i = Z(s_i)$, and let $Z = (Z_1, ..., Z_n)^T$, $X = (x_1, ..., x_n)$ and $W = (W_1, \ldots, W_n)^T$ be the corresponding finite counterparts. SGLMMs may be defined as follows.

(a) Model for spatial random effects. This can change depending on whether the
data are lattice or continuous-domain.

(i) For continuous domain, \( W(s), s \in \mathcal{D} \subset \mathbb{R}^2 \) is a zero-mean stationary Gaussian random field with \( \text{cov}(W(s + h), W(s)) = C(h) \) for all \( s \), where the covariance function \( C(\cdot) \) depends on a vector of parameters \( \theta \). A frequently used covariance function, assuming stationarity and isotropy, is the Matérn class (Stein, 1999). \( W \) follows a multivariate normal distribution,

\[
f(W|\theta) \propto |\Sigma_\theta|^{-1/2} \exp\left(-\frac{1}{2} W^T \Sigma_\theta^{-1} W\right)
\]

(ii) For discrete spatial domain, \( W \) is often modeled by a zero-mean Markov random field. Here the index indicates a node on a lattice, typically denoting a geographic block. The neighboring structure among blocks is defined through an adjacency matrix \( A \). It is an \( n \) by \( n \) matrix where \( \text{diag}(A) = 0 \) and \( A_{ij} = 1 \), if the \( i^{th} \) and \( j^{th} \) locations are connected (Besag et al., 1991). A popular model for \( W \) is the intrinsic conditionally auto-regressive (InCAR) or Gaussian Markov random field,

\[
f(W|\tau) \propto \tau^{\text{rank}(Q)/2} \exp\left(-\frac{\tau}{2} W^T Q W\right),
\]

where \( \tau \) is a parameter that controls the smoothness of the spatial field, and \( Q = \text{diag}(A1) - A \) is the precision matrix and \( 1 \) is an \( n \)-dimensional vector of ones.

(b) Conditional on random effects \( W \) and regression parameters \( \beta \), observations \( Z \) are independently distributed with distribution function \( \prod_{i=1}^n f_{Z_i|W_i}(Z_i|W_i, \beta) \). Each observation has a site-specific conditional mean \( \mu_i = E[Z_i|W_i, \beta] \).

(c) For some link function \( g, g(\mu_i) = x_i^T \beta + W_i \).

In the remaining sections, we use \( \theta \) to denote parameters of the spatial random
fields for both continuous and discrete cases. The observed-data likelihood is hence

\[ L(\beta, \theta; Z) = \int_{\mathbb{R}^n} \left\{ \prod_{i=1}^{n} f_{Z_i|W_i}(Z_i|W_i, \beta) \right\} f_W(W|\theta) dW. \] (3.1)

The likelihood function involves a high-dimensional integration and is typically not available in closed-form. Therefore, direct maximization of (3.1) is infeasible. Monte Carlo maximum likelihood (Geyer and Thompson 1992) and Monte Carlo versions of EM algorithms (cf., Wei and Tanner 1990, McCulloch 1994) were proposed for SGLMMs (Christensen 2004; Zhang 2002). These Monte Carlo methods require simulations from the conditional distribution of random effects given the data, \( f_W(W|\theta, Z) \), for both inference and prediction. These methods work quite well for data sets that are relatively small, say in the hundreds. When confronted with thousands of data points or more, these methods become computationally challenging. This is largely because, like in the Bayesian approach, the number of random effects grows with the size of the data. This results in a high-dimensional integration problem at each step of the EM algorithm. This, in turn, leads to an unstable MCMC-EM algorithm. Furthermore, it becomes difficult to construct a fast mixing MCMC algorithm at each expectation step because the random effects are usually highly cross-correlated. In addition to addressing these challenges via our projection-based approach, we provide some guidance on how to tune the algorithm, including, for instance, how to determine appropriate Monte Carlo sample sizes for each step of the algorithm. We present projection-based models in Section 3.3 and an automated MC-EM algorithm in Section 3.4 for fast maximum likelihood inference for SGLMMs.
3.2.2 Spatial Confounding

Let $P_{[X]} = X(X^T X)^{-1}X^T$ and $P_{[X]}^\perp = I - P_{[X]}$ denote the orthogonal projections onto the span of $X$ and its complement, respectively. The confounding problem therefore arises in much the same way as in multicolinearity problems with standard regression models. The only difference here is that the confounding arises because of the spatial random effects. The linear model for cite-specific conditional means, $\mu = (\mu_1, \ldots, \mu_n)^T$, is $g(\mu) = X\beta + W = X\beta + P_{[X]}W + P_{[X]}^\perp W$. Since $P_{[X]}W$ is confounded with $X$, Hodges and Reich (2010) suggests that it be removed from the model to alleviate spatial confounding, however, Hanks et al. (2015) argues that when $P_{[X]}W$ is “removed” from the model, its effect is combined with $\beta$ and an a posteriori adjustment should be performed to obtain valid inference about $\beta$. This way of restricting random effects to be orthogonal to fixed effects is also called restricted spatial regression model. Methods for addressing these problems have been developed and studied for both continuous and discrete domain data (cf. Reich et al., 2006; Hanks et al., 2015; Guan and Haran, 2017; Hughes and Haran, 2013).

3.3 A Projection-Based Approach to Dimension Reduction

To address the computational and confounding issues described in the last section, we consider two projection-based models for the continuous and discrete spatial domains (Guan and Haran, 2017; Hughes and Haran, 2013). Both models leverage efficient reparameterizations to (1) reduce the dimension of the random effects and (2) alleviate spatial confounding. They share a common form, $P_{[X]}^\perp W \approx M\delta$, where $\delta$ is an $m$–dimensional vector with nearly independent elements and $M$ is
an $n \times m$ projection matrix that preserves the spatial information of $W$. Details on the reparameterization are provided below:

(i) For the continuous case, the Gaussian random field $W(s)$ is determined by its parameterized covariance function $C(\cdot)$. For example, the exponential covariance function $C(h) = \sigma^2 \exp(-h/\phi)$ has parameters $\theta = (\sigma^2, \phi)^T$. The resulting covariance for $W$ is $\Sigma_\theta = \sigma^2 R_\phi$, where $R_\phi$ denotes the correlation matrix with $R_{ij} = \exp(-||s_i - s_j||/\phi)$. Guan and Haran (2017) propose to reparameterize $W$ using the first $m(< n)$ principal component of $R_\phi$ and then project the reduced-dimensional random effects to the orthogonal span of $X$. Let $U_\phi = [u_1, ..., u_m]$ denote the first $m$ eigenvectors and $D_\phi = \text{diag}(\lambda_1, \ldots, \lambda_m)$ a diagonal matrix containing eigenvalues of $R_\phi$, and let $M_\phi = P_{[X]}^\perp U_\phi D_\phi^{1/2}$. Then, the reparameterized random effects $\tilde{W} = U_\phi D_\phi^{1/2} \delta$, resulting in independent random effects $\delta | \sigma^2, \phi \sim N(0, \sigma^2 I)$, and $P_{[X]}^\perp \tilde{W} = M_\phi \delta$ is then restricted to be orthogonal to fixed effects. The resulting hierarchical model is

$$g \{ E[Z|\beta, M_\phi, \delta]) \} = X\beta + M_\phi \delta,$$

$$\delta | \sigma^2, \phi \sim N(0, \sigma^2 I).$$

If exact eigendecomposition computation is infeasible, say when there are several thousands of data points, then it can be approximated efficiently by a probabilistic version of Nyström’s method. We present an outline of the approximation algorithm in Section 3.4.6; details are provided in Guan and Haran (2017); Banerjee et al. (2012).

(ii) For the discrete case, we can reduce the dimension of the model by using the first $m$ principal components $M$ of the Moran operator $P^\perp AP^\perp$ (Hughes and
as follows,

\[
g \{E[Z|\beta, \delta]\} = X\beta + M\delta,
\]

\[
p(\delta|\tau) \propto \tau^{q/2} \exp \left(-\frac{\tau}{2} \delta^T Q_\delta \delta\right)
\]

where \(Q_\delta = M^T Q M\).

### 3.4 MCMC-EM for Inference

We propose an automated Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm for maximizing the observed-data likelihood,

\[
L(\beta, \theta; Z) = \int \mathbb{R}^m \left\{ \prod_{i=1}^n f_{Z_i|M\delta}(Z_i|M\delta, \beta) \right\} f_\delta(\delta|\theta) \, d\delta.
\] (3.2)

Compared to standard SGLMMs (3.1), the integral has smaller dimensions. For instance, in our simulation study \(m = 50\) is sufficient for a data size of 1000 based on the rank selection guidelines provided in Section 3.4.4; moreover, \(\delta\) is less correlated than the original random effects which makes it easier to construct an efficient MCMC algorithm at each iteration of the MCMC-EM algorithm.

For ease of representation, we write \(\psi = (\beta, \theta)\), and let \(f_{Z,\delta}(Z, \delta; \psi)\) denote the integrand in (3.2). In an EM algorithm, random effects \(\delta\) are treated as missing data; the complete-data log likelihood, \(\ln f_{Z,\delta}(Z, \delta; \psi)\), is in closed-form and therefore easier to work with. We outline the MCMC-EM algorithm here and provide details in subsequent sections. Let \(\hat{\psi}^{(t)}\) be the current estimate of \(\hat{\psi}\). Then for \(t = 1, 2, 3, \ldots\), the MCMC-EM algorithm iterates between the following two steps,

E-step:
(a) **Simulation step**: obtain an MCMC sample \( \delta^{(t,1)}, \ldots, \delta^{(t,k_t)} \), with a sample size of \( k_t \), from \( f_{\delta|Z}(\delta|Z, \psi^{(t)}) \) under the current estimates \( \psi^{(t)} \).

(b) **Monte Carlo integration step**: approximate the conditional expectation using the sample from (a),

\[
Q(\psi, \psi^{(t)}) = E[\ln f_{Z,\delta}(Z, \delta; \psi)|Z, \psi^{(t)}] \\
\approx \frac{1}{k_t} \sum_{k=1}^{k_t} \ln f_{Z,\delta}(Z, \delta^{(t,k)}; \psi) \\
= \hat{Q}(\psi, \psi^{(t)}).
\]

**M-step:**

Find \( \psi^{(t+1)} \) to satisfy \( Q(\psi^{(t+1)}, \psi^{(t)}) \geq Q(\psi^{(t)}, \psi^{(t)}) \).

A reasonable stopping criteria is to run the algorithm until the increase of the Q-function is, with a high probability, smaller than a predetermined threshold \( \epsilon \).

### 3.4.1 E-Step

Monte Carlo samples from the conditional distribution can be easily obtained using an MCMC algorithm (see Robert and Casella 1999 for an introduction). The projection-based models have a reduced-dimensional and de-correlated random effects; this is advantageous in constructing MCMC over the traditional SGLMMs. Here we use the Metropolis-Hastings algorithm with a multivariate normal proposal function. Given \( \psi \), the conditional density is \( f_{\delta|Z}(\delta|Z, \psi) \propto f(Z|M\delta, \beta)f(\delta|\theta) \); we propose \( \delta' \) from \( \text{MVN}(\delta, \cdot) \), and accept it with probability

\[
\min \left\{ 1, \frac{f(Z|M\delta', \beta)f(\delta'|\theta)}{f(Z|M\delta, \beta)f(\delta|\theta)} \right\}.
\]

We utilize several strategies for constructing an efficient MCMC algorithm, including: (1) We use adaptive MCMC (Roberts and Rosenthal 2009) to avoid tedious manual tuning and to maintain desirable acceptance rate; for the \((t + 1)^{th}\)
EM iteration, we adjust the variance of the proposal function using 

\[ 0.95 \times \frac{2.38^2}{q} \times \Sigma_t + 0.05 \times (0.1)^2 / q \times I_q, \]

where \( \Sigma_t \) is the sample covariance of the target distribution based on the current \( k_t \) sample. (2) We initiate the MCMC using the last iteration of MCMC from the previous EM update, \( \delta^{(t+1,1)} = \delta^{(t,k_t)} \), to obtain a good starting value. (3) We automatically adjust the Monte Carlo sample size for each EM iteration using the ascent-based approach proposed by Caffo et al. (2005) in order to recover EM’s ascent property and allocate computing resources efficiently. A sketch of our implementation is provided below.

The Monte Carlo sample size \( k_t \) at the \( t^{th} \) EM iteration is chosen automatically to increase the \( Q \)-function with a high probability. Let \( \Delta Q(\psi^{(t,k_t)}, \psi^{(t-1)}) \equiv Q(\psi^{(t,k_t)}, \psi^{(t-1)}) - Q(\psi^{(t-1)}, \psi^{(t-1)}) \) be the change in the \( Q \)-function. Its approximation \( \hat{\Delta} Q(\psi^{(t,k_t)}, \psi^{(t-1)}) \) or simply \( \hat{\Delta} Q \) computed in the Monte Carlo integration step, when suitably normalized, has a limiting normal distribution centered at \( \Delta Q \) and a variance \( \sigma_{\Delta Q}^2 \). Let \( z_\alpha \) be the \( (1 - \alpha)^{th} \) percentile of a standard normal random variable \( z \). We compute the asymptotic lower bound, \( \hat{\Delta} Q - z_\alpha \text{ASE} \), where ASE denotes the asymptotic standard error estimated using batch means (Flegal et al., 2008). If the asymptotic lower bound is negative, then the increase in the \( Q \)-function is indistinguishable due to a large Monte Carlo error, indicating that a larger sample size is required. Using this as a guideline, we increase sample size from \( k_t \) to \( k_t + k_t/2 \) until the asymptotic lower bound is positive. The required Monte Carlo sample sizes are typically small in the early EM iterations, and gradually increase as the parameter estimates get near the optimal region (Figure 3.4). To insure that a large enough Monte Carlo sample is obtained at the first EM iteration to explore the parameter space and to estimate the correlation structure of the target distribution, we run the MCMC until the multivariate effective sample size (Gong and Flegal 2015) is at least 10 times the dimension of the target
distribution.

Figure 3.1 shows the trace plots of one random effect for three consecutive EM iterations in a simulated example (we shall return to the simulation details in Section 3.5), we see that MCMC mixing improves substantially and the required MCMC sample size are adjusted automatically.

3.4.2 M-Step

We now present an EM gradient algorithm for updating the parameter estimates $\psi^{(t+1)}$ in the M-step. In the EM gradient algorithm, a one-step Newton-Raphson replaces the maximization which can reduce computing time [Lange 1995, Zhang 2002]. Here, we introduce the algorithm and provide details on fitting the projection-based models. To maximize $Q(\psi, \psi^{(t)})$, we find its first and second derivative, $Q'$ and $Q''$, with respect to $\psi$, then update the parameters by letting $\psi^{(t+1)} = \psi^{(t)} - Q''(\psi^{(t)})^{-1}Q'(\psi^{(t)})$. When the derivatives, $\partial/\partial \psi \ln f_{Z,\delta}(Z, \delta; \psi)$ and $\partial^2/\partial \psi \partial \psi^T \ln f_{Z,\delta}(Z, \delta; \psi)$, have closed-form, their respective conditional expectations,

$$Q' = E \left[ \partial/\partial \psi \ln f_{Z,\delta}(Z, \delta; \psi) | Z \right]$$

and

$$Q'' = E \left[ \partial^2/\partial \psi \partial \psi^T \ln f_{Z,\delta}(Z, \delta; \psi) | Z \right] ,$$

(3.3)
are approximated using Monte Carlo samples from the simulation step. For the projection-based models, $Q''$ is block-diagonal; therefore, $\beta$ and $\theta (\psi = (\beta, \theta))$ can be updated separately.

Parameter estimation for regression parameters $\beta$ is the same for both continuous and discrete cases. If the conditional distribution of the response variable is from the exponential family, for instance, the binomial or Poisson model, and the link function is canonical, then the derivatives in (3.3) with respect to $\beta$ are simply

$$
\frac{\partial \ln f(Z|M\delta, \beta)}{\partial \beta} = X^T (Z - E[Z|M\delta, \beta]),
$$

and

$$
\frac{\partial^2 \ln f(Z|M\delta, \beta)}{\partial \beta \partial \beta^T} = -X^T V(Z|M\delta, \beta) X,
$$

where $V(Z|M\delta, \beta)$ is a diagonal matrix with elements equal to the conditional variance of $Z$. These are approximated by the averages $1/k_t \sum_k X^T (Z - E[Z|M\delta^{(t,k)}, \beta^{(t)}])$ and $1/k_t \sum_k X^T V(Z|M\delta^{(t,k)}, \beta^{(t)}) X$, using MCMC samples $\delta^{(t,i)}, i = 1, \ldots, k_t$ from $f_{\delta|Z}(\delta|Z, \psi^{(t)})$.

Parameter estimation of $\theta$ for the continuous and discrete cases are discussed separately. This is because, under the continuous case, the derivatives with respect to $\phi$ are not in closed-form.

(i) In the continuous case $\theta = (\sigma^2, \phi)$. For a given $\phi$, $\sigma^2$ is estimated using equation (3.3), approximated as follows,

$$
E \left[ \frac{\partial \ln f(\delta|\theta)}{\partial \sigma^2} | Z, \psi^{(t)} \right] \approx -\frac{q}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \frac{1}{k_t} \sum_k \delta^{(t,k)} T \delta^{(t,k)},
$$

$$
E \left[ \frac{\partial^2 \ln f(\delta|\theta)}{\partial (\sigma^2)^2} | Z, \psi^{(t)} \right] \approx \frac{q}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \frac{1}{k_t} \sum_k \delta^{(t,k)} T \delta^{(t,k)}.
$$

However, analytical derivatives in (3.3) with respect to $\phi$ are not available, as the projection matrix $M = M_\phi$ is related to $\phi$ in a complicated fashion.
Therefore, we estimate $\phi$ via a numerical routine. At the $t^{th}$ iteration we first obtain estimates $(\beta(\phi(t+1)), \sigma(\phi)^2(t+1))$ conditioning on $\psi(t)$; they are plugged into $\hat{Q}(\psi, \psi(t))$ to obtain $\hat{Q}(\phi)$. We then perform a numerical search on the neighboring values of $\phi(t)$ to find $\phi(t+1)$ that satisfies $\hat{Q}(\phi(t+1)) > \hat{Q}(\phi(t))$. This is reduced to computing the difference in

$$
\hat{Q}(\phi^*) - \hat{Q}(\phi(t)) = 
- \frac{1}{2} \left( \sum_{i=1}^{m} \ln(d_{\phi^*,i}) - \sum_{i=1}^{m} \ln(d_{\phi,i}) \right) - \frac{1}{2\sigma^2(t+1)} \times
\frac{1}{k_t} \sum_{k=1}^{k_t} (M\delta^{(t,k)})^T (U_{\phi^*}D_{\phi^*}^{-1}U_{\phi^*}^T - U_{\phi}D_{\phi}^{-1}U_{\phi}^T) (M\delta^{(t,k)}),
$$

where $\phi^*$ is a neighboring value of $\phi(t)$. The above comparison is performed for several neighboring values, and the one with the largest difference is set to $\phi(t+1)$. The major computation involved is computing the eigencomponents of $R_{\phi^*}$; performing eigen decompositions several iterations for data size up to a couple of thousands is relatively fast, and we can parallelize this comparison for multiple $\phi$ at the same time using a multicore machine. If the data size is much larger than a few thousands, we can approximate the eigencomponents using Algorithm 2; details are provided in Section 3.4.6.

(ii) In the discrete case $\theta = \tau$. The derivatives with respect to the smoothing parameter $\tau$ are

$$
\frac{\partial \ln f(\delta | \tau)}{\partial \tau} = \frac{q}{2\tau} - \frac{1}{2} \delta^T Q_\delta \delta \approx \frac{q}{2\tau} - \frac{1}{2k_t} \sum_{k} \delta^{(t,k)}^T Q_\delta \delta^{(t,k)},
$$

and

$$
\frac{\partial^2 \ln f(\delta | \tau)}{\partial \tau^2} = - \frac{q}{2\tau^2},
$$

they are approximated similarly.
The asymptotic standard errors for the maximum likelihood estimates of the parameters, except the range $\phi$, are approximated using the observed information matrix. The observed information matrix, defined as

$$J(\hat{\psi}) = -\partial^2 / \partial \psi \partial \psi^T \ln L(\hat{\psi}; Z),$$

is readily obtainable from the last iteration of the maximization step. Parametric bootstrap (Efron and Tibshirani, 1994) is another useful approach often used for obtaining standard errors of the estimates. For the parametric bootstrap, we first fit the projection-based model to the data using the MCMC-EM algorithm. Then, we simulate data sets from the fitted traditional SGLMM. For each simulated data set, we again fit the projection-based model using the MCMC-EM algorithm. We use the bootstrap to study the variability in our parameter estimates. We noticed that when confounding exists, the asymptotic standard errors approximated from inverting the observed information matrix tend to be much smaller than the ones obtained from bootstrap. This results in low coverages for the 95% confidence intervals for the regression parameters. Hence, in such cases, we advocate using a bootstrap approximation rather than the observed information matrix.

### 3.4.3 Stopping Criterion

We use a stopping rule similar to the framework of determining Monte Carlo sample sizes based on the ascent-based approach (Caffo et al., 2003). We stop the algorithm when $\triangle \hat{Q}$ is less than $\epsilon$ with a high probability, that is, when the asymptotic upper bound $\triangle \hat{Q} + z_\gamma \text{ASE} < \epsilon$. The ASE is approximated using batch means. This indicates that the integrated log-likelihood evaluated at the current estimates $\hat{Q}(\psi(t), \psi^{(t-1)})$ stabilizes.

Other standard stopping rules (McLachlan and Krishnan, 2007), for example,
based on small absolute change in parameter estimates could be used as an alternative or combined stopping criterion; we see, in our simulation study, these make little difference.

### 3.4.4 Rank Selection

Here we introduce two ways of selecting the rank $m$. The projection-based model is based on the spatial filtering (Griffith, 2013) and principal component analysis. The rank of the projection-based model can be selected by determining the proportion of variation explained by assessing the eigenvalues of the correlation matrix. We first compute $R_{\phi(0)}$ using the initial range value $\phi(0)$ and perform eigendecomposition. Then, the rank can be determined based on the desired proportion of variation, for example 85%, explained by $\sum_{i=1}^{m} \lambda_i / \sum_{i} \lambda_i$. An alternative approach for selecting the rank is to fit a non-spatial generalized linear model with predictors $X$ and synthetic spatial variables $U_mD_m^{1/2}$ for $m = 1, 2, \ldots$, and then select the rank based on variable selection criterion such as BIC.

Both of the above serve as general guidelines for selecting the initial rank, based on this, we can fit several models with different ranks and perform likelihood ratio test to determine the final model. Based on the experience from Chapter 2, both methods should agree on the final selected rank; I plan to study this via extensive simulation and application.

### 3.4.5 Spatial Prediction

SGLMMs are also often used for spatial interpolation/prediction at unsampled locations. Here, we describe spatial prediction using the projection-based models with a focus on the continuous domain, since prediction for discrete spatial domain is often less of interest in practice. Let $S^* = \{s_1^*, \ldots, s_n^*\}$ be a set of unsampled
locations where we want to make predictions. In SGLMMs, the observation \( Z(s^*) \) for any \( s^* \in S^* \) has dependence with the observed \( Z \) induced by the latent Gaussian process \( W \). In the projection-based model, the covariance between \( W^* \) at \( S^* \) and the reparameterized random effects \( \tilde{W} = U_\phi D_\phi^{1/2} \delta \) is

\[
\text{cov} \left\{ \begin{pmatrix} \tilde{W} \\ W^* \end{pmatrix} \right\} = \begin{pmatrix} (U_\phi D_\phi^{-1/2})^T \Sigma_\theta (U_\phi D_\phi^{-1/2}) & (U_\phi D_\phi^{-1/2})^T \Sigma_{\theta, ss} \\ \Sigma_{\theta, ss} (U_\phi D_\phi^{-1/2}) & \Sigma_{\theta, **} \end{pmatrix},
\]

where \((U_\phi D_\phi^{-1/2})^T \Sigma_\theta (U_\phi D_\phi^{-1/2})\) is simply \( \sigma^2 I_{m \times m} \). The best linear unbiased predictor (BLUP) of \( W^* \) given is \( \tilde{W} \) therefore based on the conditional distribution \( W^* | \tilde{W}, \theta \sim \text{MVN}(\mu_{W^* | \tilde{W}}, \Sigma_{W^* | \tilde{W}}) \) (Stein, 1999), where

\[
\mu_{W^* | \tilde{W}} = \frac{1}{\sigma^2} \Sigma_{\theta, ss} (U_\phi D_\phi^{-1/2}) \tilde{W},
\]

\[
\Sigma_{W^* | \tilde{W}} = \Sigma_{\theta, **} - \frac{1}{\sigma^2} \Sigma_{\theta, ss} (U_\phi D_\phi^{-1/2} U_\phi^T) \Sigma_{\theta, ss}.
\]

Therefore, to make a prediction, we just need to sample from the above multivariate normal distribution using the Monte Carlo sample of the random effects and the parameter estimates from the last iteration of the EM algorithm.

### 3.4.6 Approximating Eigencomponents

Fitting projection-based models for the continuous case requires eigendecomposition on \( R_\phi \) for every \( \phi \) update. When the number of data points is large, exact eigendecomposition is infeasible; however, we can approximate the principal components using a probabilistic version of Nyström’s method. Probabilistic algorithms have been increasing in popularity for fast matrix decompositions (see [Halko et al., 2011](#) for a summary of algorithms). [Banerjee et al., 2012](#) proposed using them for approximating covariance matrices in the linear Gaussian process.
regression setting. This was extended in Guan and Haran (2017) to approximate eigencomponents for SGLMMs. Here, we provide details of the probabilistic algorithm used by Guan and Haran (2017) to approximate eigencomponents; this algorithm combines ideas from random-projection with Nyström’s method.

We first introduce the deterministic Nyström’s method (Williams and Seeger, 2001). Let $K$ denote an $n \times n$ positive semi-definite matrix and $\Phi$ an $n \times m$ truncation matrix by permuting the rows of $[I_{m \times m}, 0_{m \times (n-m)}]^T$. The Nyström’s method partitions $K$ into four blocks, \[
\begin{bmatrix}
K_{11} & K_{12} \\
K_{21} & K_{22}
\end{bmatrix}
\], by sub-sampling its columns and rows, and letting $K_{11} = \Phi^T K \Phi$. Then, it performs an exact decomposition on the $m \times m$ sub-matrix $K_{11}$ to obtain its eigenvectors $V_{11}$ and eigenvalues $\Lambda_{11} = \text{diag}(\lambda_{11,1}, \ldots, \lambda_{11,m})$. Finally, it maps the low-dimensional eigenvectors $V_{11}$ to high dimension via $\sqrt{\frac{m}{n}} [K \Phi] [V_{11} \Lambda_{11}^{-1}]$ (Drineas and Mahoney, 2005).

In the probabilistic algorithm, we replace the truncation matrix $\Phi$ with $K^a \Omega$, where $\Omega$ is an $n \times (m + l)$ random matrix with $\Omega_{ij} \sim N \left(0, 1/\sqrt{(m + l)}\right)$, and $a = 0, 1, \text{ or } 2$ is a small non-negative integer to improve approximation (see Guan and Haran, 2017, for a discussion on $a$); $l$ is an oversampling factor to reduce approximation error (Halko et al., 2011). Here, we take $l = m$ ($n > m + l$), and $a = 1$. The eigenvectors approximated using the Nyström’s method are not guaranteed to be orthogonal; therefore, we take a slight variant of the Nyström’ method and an additional step to orthogonalize the column vectors, i.e. we will use the first $m$ columns of the left singular vectors of $[K \Phi] [V_{11} \Lambda_{11}^{-1/2}]$ as the final approximation to the eigenvectors. The algorithm is summarized in Algorithm 2.
Algorithm 2 Probabilistic Nyström’s Approximation

This algorithm approximates the leading $m$ eigencomponents of an $n \times n$ positive semi-definite matrix $K$, combining random projection and the Nyström’s method.

1. Low dimensional projection from $R^{n \times n}$ to $R^{n \times (m+l)}$:
   Form $\Phi = K \Omega$, where $\Omega_{ij} \sim N(0, 1/\sqrt{(m+l)})$.

2. Nyström’s method to approximate eigencomponents:
   Form $K_1 = \Phi^T K \Phi$
   SVD for $K_1$: $V_{11} \Lambda_{11} V_{11}^T$
   Form Nyström extension $C = [K\Phi][V_{11} \Lambda_{11}^{-1/2}]$
   SVD for $C$: $UDV^T$

3. Take the first $m$ columns of $U$, and the first $m$ diagonal elements of $D^2$ as our approximation to the leading $m$ eigencomponents of $K$.

3.5 A Simulation Study

We study our method for spatial counts and binary observations for both the continuous and discrete domains. Results for count observations are presented below. Results for binary observations are similar to those for the count data and are therefore not presented here.

3.5.1 Count Data in Continuous Domain

We simulate $n=1400$ random effects $W$ in the unit domain $[0,1]^2$, using a covariance function from the Matérn class \cite{Stein1999} with smoothing parameter $\nu=2.5$, $C(h) = \sigma^2 \left( 1 + \frac{\sqrt{5} h}{\phi} + \frac{5 h^2}{3 \phi^2} \right) \exp \left( -\frac{\sqrt{5} |h|}{\phi} \right)$; here, we use $\sigma^2 = 1$ and $\phi = 0.2$. Conditional on $W$, we simulate count observations $Z_i$ from Poisson($\mu_i$) using the natural logarithm link function, $\log(\mu_i) = x_{i,1} + x_{i,2} + W_i$, where $x_{i,1}, x_{i,2}$ are the xy-coordinates of $W_i$. The first 1000 observations, randomly located in the unit domain, are used for model fitting; the last 400 observations, on a $20 \times 20$ grid, are used for testing.
We first fit a non-spatial generalized linear model (GLM) using the glm function in R ([R Core Team](https://www.r-project.org/)) (R Core Team, 2014), the resulting coefficient estimates and residual variance are used as starting values. It is typically difficult to obtain an estimate for the range parameter $\phi$ from the non-Gaussian observations; therefore, we take roughly half of the spatial domain as the initial value. For the simulated example, we have $\beta^{(0)} = (1.38, 1.40)^T$ and $\theta^{(0)} = (1.55, 0.5)^T$. We then fit the MCMC-EM algorithm described in Section 3.4 with $m=50$, $\alpha=0.15$, $\gamma=0.05$, and $\epsilon=0.001$. When the stopping threshold is reached, we predict the random effects on the $20 \times 20$ grid as described in Section 3.4.5.

To monitor convergence, we run the algorithm for a fixed number of MCMC-EM iterations at three different starting values. Figure 3.2 shows the parameter estimates at each iteration; we see that the parameter estimates converge to the same values, regardless of starting values, all within 30 EM iterations. Figure 3.3 shows the estimated and predicted random effects corresponding to running the algorithm with starting values obtained from GLM estimates; they have similar spatial pattern to the simulated true random effects. Results for different starting values are similar, therefore not shown. Figure 3.4 shows the Monte Carlo sample sizes at each EM iteration; we see that most of the simulation efforts are spent in the first and last 2-3 EM iterations. Typically, when the stopping threshold is reached (indicated by the vertical dashed line), the ascent-based MCMC-EM algorithm provides a large Monte Carlo sample. This is a desirable feature, since the last MC sample is used in subsequent analyses, for instance, for estimating the observed information matrix and spatial prediction. Finally, the integrated log-likelihood function corresponding to different starting values stabilizes as the EM iteration increases (Figure 3.4).

We repeat the simulation 100 times to study the distribution of the MCMC-
Figure 3.2: Parameter estimates with different starting values converge within a few iterations.

EM estimates. For each simulated data, we fit the projection-based model as above using only the GLM estimates as starting values. Figure 3.5 shows the boxplots of the estimates from the MCMC-EM algorithm; it seems that $\hat{\beta}$ are unbiased, and $\hat{\theta}$ have positive biases. Table 3.2 shows the coverages obtained from the approximated observed information matrix and from bootstrap; the coverages based on the observed information matrix are much lower than the nominal rate 95% whereas the coverages based on bootstrap are near the nominal rate. A
Figure 3.3: The left panel shows the simulated random effects at the observation (top) and prediction locations (bottom). The right panel shows the estimated (top) and predicted (bottom) random effects using the GLM starting value.

Figure 3.4: Monte Carlo sample size is adjusted automatically, with increasing simulation efforts as EM iteration increases (left). We typically obtain a large MC sample when the algorithm stops; vertical line indicates the stopping threshold is reached. Integrated log-likelihood function corresponding to different starting values stabilize as EM iteration increases (right).
potential reason for the low coverage can be seen in Table 3.1 where we show the analyses using several ranks for a simulated Poisson data. The parameter estimates are close to the simulated true values, however, the confidence intervals (CIs) provided by the observed information matrix is much narrower than the ones from bootstrap and therefore missed the true values. The computation time for fitting the projection-based model using \( m = 50 \) is typically less than 30 minutes. This is much faster than the Bayesian with MCMC approach proposed in Chapter 2, which took roughly 4 hours. A more thorough comparison is planned to see the exact savings in computation time for different ranks and data sizes.

Table 3.1: Results of our analyses for Poisson data simulated from the SGLMM in a continuous domain

<table>
<thead>
<tr>
<th>( m )</th>
<th>( \beta_1 )</th>
<th>CI(( \beta_1 ))</th>
<th>CI(( \beta_1 ))</th>
<th>( \beta_2 )</th>
<th>CI(( \beta_2 ))</th>
<th>CI(( \beta_1 ))</th>
<th>time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1.115</td>
<td>(1.014,1.217)</td>
<td>(-2.236,4.645)</td>
<td>1.446</td>
<td>(1.347,1.545)</td>
<td>(1.406,1.543)</td>
<td>13.72</td>
</tr>
<tr>
<td>50</td>
<td>1.133</td>
<td>(1.027,1.239)</td>
<td>(-0.714,2.612)</td>
<td>1.405</td>
<td>(1.301,1.509)</td>
<td>(-0.375,3.222)</td>
<td>24.68</td>
</tr>
<tr>
<td>75</td>
<td>1.121</td>
<td>(1.016,1.226)</td>
<td>(-0.161,2.41)</td>
<td>1.405</td>
<td>(1.302,1.508)</td>
<td>(-0.27,2.64)</td>
<td>40.35</td>
</tr>
<tr>
<td>100</td>
<td>1.117</td>
<td>(1.012,1.222)</td>
<td>(-0.256,2.668)</td>
<td>1.407</td>
<td>(1.303,1.512)</td>
<td>(-0.046,2.871)</td>
<td>47.58</td>
</tr>
</tbody>
</table>

Table 3.2: Coverages for Poisson data simulated from the SGLMM on a continuous domain

<table>
<thead>
<tr>
<th>Method based on</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \sigma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Information</td>
<td>0.14</td>
<td>0.11</td>
<td>0.60</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>0.96</td>
<td>0.94</td>
<td>1</td>
</tr>
</tbody>
</table>

### 3.5.2 Count Data on a Lattice

We simulate \( n=900 \) count data on a \( 30 \times 30 \) lattice grid by first simulating random effects from \( N(0,3Q_\delta^{-1}) \); here, we follow the simulation study in Hughes and Haran.
Figure 3.5: Distributions of the estimates from the MCMC-EM algorithm for spatial counts in continuous domain. It seems that $\beta$ are unbiased and $\theta$ has positive biases.

[2013] and use the first 400 eigenvectors to simulate the data, i.e. $\dim(\delta) = 400$ and $M$ is $900 \times 400$. We then simulate count observations from Poisson($\mu_i$), where $\log(\mu_i) = x_{i,1} + x_{i,2} + W_i$ and $x_{i,1}, x_{i,2}$ are the xy-coordinates of the vertices. Convergence of parameter estimates (Figure 3.6), allocation of computing resources and the behavior of the integrated log-likelihood function (Figure 3.7) using different starting values are similar to the cases of the continuous domain. However, the point estimates are distributed more tightly around the true value (Figure 3.6), because in our simulation of the data, we have restricted the random effects to be
orthogonal to the fixed effects (no spatial confounding). This is what we typically see for both of the continuous and lattice cases; when there is confounding, the distributions of point estimates have larger variability, and when confounding is not an issue, the point estimates center closely to the true values.

Figure 3.6: Parameter estimates with different starting values converge within a few iterations (top row). Distributions of \( \hat{\beta} \) from the MCMC-EM algorithm for count data on a lattice. It seems that \( \hat{\beta} \) are unbiased and tightly center around the true value and \( \hat{\tau} \) has a negative bias (bottom row).

Table 3.3: Results of our analyses and computation time (minutes) for Poisson data simulated from the SGLMM on a lattice.

<table>
<thead>
<tr>
<th>m</th>
<th>( \beta_1 )</th>
<th>CI(( \beta_1 ))</th>
<th>( \beta_2 )</th>
<th>CI(( \beta_2 ))</th>
<th>( \sigma_2 )</th>
<th>CI(( \sigma_2 ))</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1.041 (0.941, 1.141)</td>
<td>1.091 (0.992, 1.191)</td>
<td>2.638 (0.733, 4.543)</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.039 (0.937, 1.141)</td>
<td>1.071 (0.97, 1.172)</td>
<td>2.564 (1.194, 3.934)</td>
<td>11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>1.032 (0.928, 1.136)</td>
<td>1.053 (0.95, 1.157)</td>
<td>2.551 (1.356, 3.746)</td>
<td>29.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.025 (0.921, 1.13)</td>
<td>1.05 (0.946, 1.153)</td>
<td>2.715 (1.531, 3.898)</td>
<td>39.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.7: Monte Carlo sample size is adjusted automatically, with increasing simulation efforts as EM iteration increases (a). We typically obtain a large MC sample when the algorithm stops; vertical line indicates the stopping threshold is reached. Integrated log-likelihood function corresponding to different starting values stabilize as EM iteration increases (b).

Table 3.4: Coverages for Poisson data simulated from the SGLMM on a lattice

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Information</td>
<td>0.91</td>
<td>0.82</td>
<td>0.70</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>0.96</td>
<td>0.90</td>
<td>0.63</td>
</tr>
</tbody>
</table>

3.6 An Application on US Infant Mortality Data

We fit the projection-based model for the county-level US infant mortality from 2002 to 2004, a data set analyzed in Hughes and Haran (2013) under a Bayesian approach. The response variable is the 3-year average number of infant deaths before the first birthday, and the predictors are the rate of low birth weight (low), the percentage of black residents (black), the percentage of Hispanic residents (Hisp), a measure of income inequality (the Gini coefficient proposed by Gini, 1921), a composite score of social affluence (aff, proposed by Yang et al., 2009)
and residential stability (stab, an average z-score of two variables). Similar to
Hughes and Haran (2013), we use the 3-year average number of live births as an
offset to adjust for the population difference in these counties. Figure 3.8 shows
that the parameter estimates stabilize quickly. The point and interval estimates
from the maximum likelihood inference using MCMC-EM method are shown in
Table 3.5; the results are very comparable to the ones from the Bayesian inference
with MCMC method showed in Hughes and Haran (2013).

Figure 3.8: Parameter updates for infant mortality analysis
Table 3.5: Results of fitting model with rank 50 to the infant mortality data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter</th>
<th>Estimate</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$\beta_0$</td>
<td>-5.423</td>
<td>(-5.609,-5.238)</td>
</tr>
<tr>
<td>Low birth weight</td>
<td>$\beta_1$</td>
<td>8.779</td>
<td>(7.521,10.037)</td>
</tr>
<tr>
<td>Black</td>
<td>$\beta_2$</td>
<td>0.004</td>
<td>(0.003,0.006)</td>
</tr>
<tr>
<td>Hisp</td>
<td>$\beta_3$</td>
<td>-0.004</td>
<td>(-0.005,-0.003)</td>
</tr>
<tr>
<td>Gini</td>
<td>$\beta_4$</td>
<td>-0.568</td>
<td>(-1.001,-0.136)</td>
</tr>
<tr>
<td>aff</td>
<td>$\beta_5$</td>
<td>-0.077</td>
<td>(-0.089,-0.065)</td>
</tr>
<tr>
<td>stab</td>
<td>$\beta_6$</td>
<td>-0.029</td>
<td>(-0.044,-0.015)</td>
</tr>
<tr>
<td>–</td>
<td>$\tau$</td>
<td>7.939</td>
<td>(2.556,13.323)</td>
</tr>
</tbody>
</table>

3.7 Summary

We have described an approach for maximum likelihood inference for SGLMMs. Our MCMC-EM approach to maximizing the likelihood is computationally efficient; it appears to be faster than the corresponding Bayesian approach using MCMC in the cases of continuous domain, and comparable to that in the cases of discrete domain. For point estimation, parameter estimates seem to converge quickly under the current simulation settings; more simulation study using different covariance functions, parameter values and simulation schemes (spatial confounded or not) is planned to further investigate the proposed approach. For interval estimation, confidence intervals from parametric bootstrap seem to be more reliable than asymptotic confidence intervals based on the observed Fisher information. The conclusion here is based on our simulation study of 100 data sets, for each of which a bootstrap sample of 100 replicates was used to compute the confidence intervals; this is repeated for both count and binary observations on both continuous and discrete spatial domains. Bootstrap confidence interval is computationally expensive to obtain, but can be easily parallelized. I plan to further investigate combining the approximated observed information matrix with the a posteriori
adjustment proposed by Hanks et al. (2015) to obtain approximate confidence intervals without computationally intensive bootstrap sampling.

Maximum likelihood inference has not been as popular as Bayesian inference for SGLMMs, at least in part because of computational issues. We hope that the methodology we develop here, which addresses inference for a very large class of models, including both latent Gaussian process and Gaussian Markov random field models, will allow researchers to routinely fit SGLMMs using maximum likelihood inference. We do not believe that this will entirely replace Bayesian approaches as Bayesian models allow for a greater range of flexibility in terms of adding additional hierarchies, handling missing data, and combining information from multiple variables routinely. However, for a large range of problems, the class of SGLMMs for which we have developed a computationally efficient set of methods here, maximum likelihood inference may be a convenient and viable option. This fact is borne out by our simulation and real data examples where the ML approach results in estimates that are comparable to those obtained by using Bayesian methods.
Inferring Ice Thickness from a Glacier Dynamics Model and Multiple Surface Datasets

In this chapter we develop a Bayesian hierarchical model that is flexible and capable of integrating a simple glacier dynamics model, multiple data sets and uncertainty sources to estimate the key parameter in a glacier dynamics model. The data set we consider is from the Thwaites glacier in West Antarctica. The methods allow us to quantify the contribution of uncertainties from various sources to reconstruct ice sheet thickness using sparsely observed ice thickness data. We show that there are some advantages of our approach compared to a naive non-statistical approach: (1) we obtain uncertainty estimates for the parameters of the physical model, and (2) predicting ice sheet thickness with uncertainty provides more useful inference.
4.1 Introduction

A common challenge in glaciology is the estimation of unknown bed properties below glaciers and ice sheets using high-quality surface observations and some knowledge or models of the glacial physics. Here we focus on estimating or interpolating poorly known bedrock topography (or equivalently ice thickness) based on well understood ice sheet physics and higher quality information regarding surface quantities. The principal physics commonly applied is the conservation of ice mass; given high-quality observations of surface elevation, ice velocity and surface mass balance, we can then deduce the ice thickness assuming the ice sheet is in a steady state. Attempts based on this principle have been made over continental Antarctica with very coarse grid spacing (Warner and Budd, 2000) and for a major glacier in northeast and in other locations of Greenland using a much higher-resolution spacing (Morlighem et al., 2011, 2013, 2014). For simplicity, we focus on a 1-D flowline along the centerline of a glacier with two improvements to the physical model: (1) we add a new component to the dynamics model, the Shallow-Ice Approximation (SIA), which goes beyond mass conservation for shearing flow, and (2) we include the varying glacier width to account for tributaries, which contribute to mass flux in the downstream.

A purely statistical interpolation approach for deriving ice thickness will violate the underlying physics, resulting in a physically implausible reconstruction of the ice sheet. Furthermore, it would not contribute to our understanding of ice sheet dynamics which is intrinsically of scientific interest. On the other hand, a purely physics-based approach using the simple physical model may not capture the entire glacier process no matter how well we adjust the model parameters, because it ignores various errors and uncertainties resulting in estimates that are over-confident and unreliable. Therefore, we need a modeling approach that com-
bines both the physics as well as the observational data sets, and allows for errors and data-model discrepancies. We propose an approach that allows us to: (1) combine multiple data sources and simple physical laws for deducing ice thickness, (2) estimate the key parameter of the physical model along with their uncertainties, and (3) account for errors and for data-model discrepancies. Our main focus is on parameter estimation of the physical model and reconstructing the ice thickness using sparsely observed thickness data, high-quality surface observations and well-understood ice sheet physics.

Efforts have been made on modeling glacier using a Bayesian approach with a focus on smoothing resolutions for the North East Ice Stream in Greenland (cf. Berliner et al., 2008a,b). A mass conservation approach, similar to the one we use here, was employed by Berliner et al. (2008a,b) for a 1-dimensional glacier dynamics model. This was used to relate basal (bed) elevation to surface velocity speed and surface elevation. In their study, they use all available ice thickness, the difference between the surface and basal elevations to estimate the ice deformation coefficient and to select models of different smoothing resolutions. The problem we consider in this paper differs from their work by posing at least two significant additional challenges: (1) improvements to the glacier dynamics model require us to estimate the flow width, which is an unknown quantity that varies along the flowline, and (2) we use sparse thickness observations to reconstruct the ice profile. Problems that involve combining a physical model and observations for improving estimation and prediction have been an active research area. Some studies focus on combining multiple information from deterministic physical model outputs and observations for improving estimation of a spatial temporal field without explicitly solving the physical model (cf. Reich and Fuentes, 2007; Wikle et al., 2001); because running the physical model is computationally intensive. The glacier dynamics
model we propose has a simple form that can be solved explicitly and efficiently, allowing us to work with the model directly. Others solve the physical model explicitly with a focus on inferring model parameters, which are typically scalars, from noisy observations (cf. Li et al., 2005; Ramsay et al., 2007; Xun et al., 2013). However, it is not clear how to adapt these approaches to our problem. Our problem and methodology is different in the following ways. (1) Our variable of main interest, ice thickness, is highly nonlinearly related to other physical model inputs. Therefore, a basis function approximation to ice thickness using generalized smoothing approaches (Ramsay et al., 2007; Xun et al., 2013) does not simplify the form of the differential equation; the estimation problem in a statistical formulation therefore remains difficult. (2) The unknown flow width parameter is spatially varying. Therefore, it is impossible to approximate the Hessian matrix required by optimization approaches (cf. Li et al., 2005). (3) Moreover, the multiple data inputs to the physical model have different spatial resolutions and are subject to observational errors. These challenges motivate the Bayesian hierarchical approach proposed in Section 4.3.

We apply our method to the Thwaites Glacier in West Antarctica, because (1) a relatively high-quality dataset of all necessary surface observations is available along its centerline (see Section 4.2.2), (2) Thwaites Glacier is the center of considerable current attention and concern, as it may be the first major Antarctic outlet to undergo dramatic future retreat (Rignot et al., 2014; Joughin et al., 2014; DeConto and Pollard, 2016), and (3) it is not in steady state, but has experienced rapid thinning in recent decades (Pritchard et al., 2012; Rignot et al., 2014), and hence provides a good testbed for our method.

The outline of the remainder of the paper is as follows. We start by introducing in Section 4.2 the physical model and the available data sets. We then introduce
the statistical model and its inference in Section 4.3. We present the results of the Bayesian approach on simulated data sets and application to the Thwaites Glacier in Section 4.4. We discuss the extension of the proposed method to other glaciers in Section 4.5.

4.2 Glacier Dynamics and Data

4.2.1 Glacier Dynamics

The physical model we employ is based on conservation of ice mass along a flowline, bounded by streamlines separated by the width of the glacier, with an additional physics component, the Shallow-Ice Approximation (SIA), for shearing flow. The spatial domain of the model is 1-dimensional along the centerline, which represents the fastest flow path of a glacier, bounded by the divide (initial location) and upstream of the grounding line (Figure 4.1a). The physical model relates several processes, for instance, surface elevation, surface velocity of ice flow, and surface accumulation and thinning rates to ice thickness, allowing us to combine multiple data sets in a physically plausible manner.

The principal physics used is conservation of ice mass. This implies the change in mass flux downstream is equal to the net accumulation rate and thinning rate. That is, the glacial flow of ice out of a control volume in the downstream direction, minus the flow into the volume from upstream, must equal the net rate of ice externally added on the surface (snowfall minus melt, with no melt occurring here), plus any observed reduction in ice thickness (the thinning rate),

\[
\frac{\partial}{\partial x} (\bar{v}(x) h(x) \omega(x)) = (a(x) - \tau(x)) \omega(x),
\] (4.1)
where $x$ indicates location along the flowline (in meters), $\bar{v}(x)$ is the centerline depth-averaged and width-averaged velocity (m/year), $h(x)$ represents the width-average thickness (m), $\omega(x)$ is the flow width to account for downstream variations in glacier width, $a(x)$ is local surface net accumulation rate (m/year), and $\tau = \frac{\partial h(x)}{\partial t}$ is the ice-column thickness thinning rate (m/year). This equation states that the mass is balanced between the change in flux along downstream and the net change in the centerline depth, the left and right side of Equation 4.1 respectively.

The glacier width $\omega(x)$ does not necessarily represent the extreme edges of the glacier; it is the transverse distance between a pair of streamlines across which no ice flows out of the domain. If the flowline extends upstream beyond where the glacier splits into tributaries, in principle the glacier width includes the tributaries too, because they contribute to mass flux into the downstream portions of the domain. This is the case for the Thwaites application (see Figure 1). It is assumed that the observed centerline velocities are close to uniform across the width of the glacier; a correction factor could be applied in follow-up work to account for side drag and lateral shear.

This approach closely follows Morlighem et al. (2011), who used conservation of mass to deduce ice thicknesses on a major Greenland outlet glacier. They used observed surface velocity $v_s(x)$ for depth-averaged velocity $\bar{v}(x)$. Here, we apply a correction to $v_s(x)$ based on the Shallow Ice Approximation (e.g. Van der Veen, 2013) to account for the internal ice deformation due to gravitational stress and basal drag due to friction at the ice bed. Let $v_b(x)$ denote the basal sliding velocity, $v_s^{SIA}(x)$ and $\bar{v}^{SIA}(x)$ denote SIA-modeled surface velocity and depth averaged velocity respectively. Our correction to surface velocity is

$$\bar{v}(x) = v_s(x) - (v_s^{SIA}(x) - \bar{v}^{SIA}(x)),$$
where

\[ v_s^{\text{SIA}}(x) = \frac{A}{4} (\rho g s(x))^3 h(x)^4 + v_b(x), \]

\[ \bar{v}^{\text{SIA}}(x) = \frac{A}{5} (\rho g s(x))^3 h(x)^4 + v_b(x), \]

so that

\[ \bar{v}(x) = v_s(x) - \frac{A}{20} (\rho g s(x))^3 h(x)^4. \] (4.2)

\[ A \] is the ice rheologic coefficient (\( Pa^{-3}/\text{year} \)) in Glens Law with exponent 3, \( \rho \) is ice density (\( kgm^{-3} \)), \( g \) is gravitational acceleration (\( m/s^2 \)), and \( s(x) \) is the observed downstream surface slope derived from surface elevation. A scalar value of the rheological coefficient \( A \) is used, neglecting its dependence on ice temperature or fabric variations along the centerline. Note that the difference between \( v_s^{\text{SIA}}(x) \) and \( \bar{v}^{\text{SIA}}(x) \) is used as a correction to the observed velocity, so that the unknown basal velocity \( v_b(x) \) cancels. Thus, no explicit knowledge of basal sliding (and the effect of frozen versus thawed beds) is needed, because its contribution to the depth-averaged transport is included implicitly via the observed surface velocity.

We first describe a naive non-statistical approach for solving ice thickness using (4.1) and (4.2). This involves finite differences and solving fifth order polynomials. Typically, \( a(x) \) and \( \tau(x) \) are taken to be well observed surface quantities, as are the surface velocity and surface elevation. For a fixed value of \( A \) and a plug-in estimate of \( \omega \), we can obtain the flux \( \bar{v}(x) h(x) \omega(x) \) on a predetermined grid along the \( x \)-axis by solving the following ordinary differential equation derived from (4.1),

\[ \bar{v}(x) h(x) \omega(x) - \bar{v}(x_0) h(x_0) \omega(x_0) = \int_{x_0}^{x} (a(s) - \tau(s)) \omega(s) ds, \] (4.3)

and \( \bar{v}(x_0) h(x_0) \omega(x_0) = C_0 \) is a given initial value.
Plugging in (4.2) into (4.3), we obtain a fifth order polynomial of $h(x)$. Using a numerical solver, we can compute the ice thickness as a function of other variables; denote this as $M(v_s(x), s(x), a, \tau, \omega, A, h_0)$. Here, we write $h_0 = h(x_0)$ as the initial thickness value for simplicity. However, the solutions for the system may not be unique or do not exist for some locations. For locations where solutions exist, we select the ones that are closest to the observed ice thickness. The results from this non-statistical inversion method using plugged-in estimates of the flow width from Figure 4.1b are shown in Figure 4.2; we see that none of these solutions seem to reconstruct the ice thickness well, for instance, some sections along the flowline have large departures, up to 1 km, from the observed thickness. Also the bedrock topography is not recognized by the reconstructed ice thickness, for example, the thin ice near 130 km, 200 km and 250 km cannot be reproduced from the naive method. Therefore, we conclude that although this non-statistical approach provides fast and rough estimates for the unknown, there are a few drawbacks that make it undesirable. First, the plug-in estimates for $\omega$ are chosen informally; this choice is not based on a well-defined criterion. Figure 4.1a shows a few possible streamlines estimated by tracing the direction of surface velocity vector field using the MATLAB toolbox Ice Flowlines (Greene et al., 2017). For every pair of streamlines at each side of the centerline, we can take the transverse distance as the flow width. However, depending on different starting upstream locations, the widths can differ substantially. Second, measurement errors in observations have not been accounted for, which may lead to over-confident estimates. Third, without quantifying uncertainty associated with the estimates, any conclusions obtained would be unreliable.

These drawbacks motivate us to propose a statistical approach that can integrate multiple sources of information, multiple data sets, underlying physical law
and error sources, in a single framework to provide parameter estimates along with its uncertainty and probabilistic prediction for unobserved ice thickness.

4.2.2 Data

Several surface data sets are available for the WAIS. These include (1) surface elevation measured by using Airborne Topographic Mapper (ATM) instrumentation as part of Operation IceBridge funded aircraft survey campaigns (Krabill, 2016), (2) surface velocity with spatial resolution of 450 m as part of the NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Program (Rignot and Scheuchl, 2011) and (3) estimates of ice accumulation rate with spatial resolution $\sim 55$ km from Van de Berg et al. (2006). We first smooth the surface elevation, then derive the surface slope $s(x)$ by taking central differences. Surface elevation has been smoothed, because the glacier dynamics model is very sensitive to small variations in the surface slope. Based on the physical understanding of the glacier, the small variations in surface slope are not realistic and mostly due to observational errors. For the Thwaites Glacier, we also have ice thickness observations from airborne radar with $\sim 14$ m spacing along the flowline of total length 272.8 km (Leuschen et al., 2016). Note that surface velocity, surface slope, ice accumulation and thinning rates have different spatial resolutions; velocity and slope are interpolated to the locations where we have thickness data, and accumulation and thinning rate are interpolated to the same grid as the flowline width (details described in Section 4.3.4). Our methods would adapt in straightforward fashion to situations where the data are on different resolutions; the only additional burden is computational. All the data that can be observed or derived from observable quantities are summarized in Table 4.1.
(a) The red, green, and blue dashed lines represent three pairs of streamlines, and the centerline (white dashed) for the Thwaites Glacier.

(b) Flow widths obtained from taking the transverse distance perpendicular to the centerline.

Figure 4.1: Depending on different starting locations at the divide, the flow widths vary.
Figure 4.2: Thwaites Glacier thickness solutions corresponding to different flowline widths and rheologic coefficient values. The non-statistical approach provides deterministic ice thickness solutions without uncertainty estimates, and none of these solutions seem to reconstruct the ice thickness well.

4.3 Hierarchical Approach

In this section, we introduce a hierarchical approach that can integrate the multiple data sets summarized in Table 4.1 and the glacier dynamics model described by (4.1) and (4.2) to infer ice thickness along the flowline. This framework has
the flexibility of incorporating multiple error sources, including the discrepancy between the physical model and the true underlying process, multiple sources of observational errors and error due to discretization of the glacier dynamics model. We discuss the computational challenges of inference based on the full hierarchical approach. Computational considerations require us to approximate this approach. Our main goal is to (1) estimate the rheologic coefficient $A$ and quantify its uncertainty due to various sources of data and the physical model, and (2) predict the ice thickness incorporating multiple sources of information.

4.3.1 Model Overview

Given the need to combine multiple sources of data with a physical model, a hierarchical Bayesian approach is natural. Here we introduce the notation and begin with an overview of the statistical model. We then describe each component in details in the following subsections. Let $\Gamma \subseteq \mathbb{R}$ denote the continuous spatial domain along the flowline with total length of 272.8 km from the divide to upstream of the grounding line. At a location $x \in \Gamma$, we have the true processes $v_s(x)$ for surface velocity, $s(x)$ for surface slope, $a(x)$ for ice accumulation rate, $\tau(x)$ for thinning rate, and $h(x)$ for ice thickness. These processes obey the physical

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Spatial resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface velocity, m/year ($V_s$)</td>
<td>450 m</td>
</tr>
<tr>
<td>Surface elevation, m ($E$)</td>
<td>$\sim 14$ m</td>
</tr>
<tr>
<td>Net ice accumulation rate, m ice eq./year ($a$)</td>
<td>55 km</td>
</tr>
<tr>
<td>Thinning rate, m ice eq./year ($\tau$)</td>
<td>$\sim 1.5 - 5$ km</td>
</tr>
<tr>
<td>Ice thickness, m ($H$)²</td>
<td>$\sim 14$ m</td>
</tr>
</tbody>
</table>

1 Ice equivalent (ice eq.)
2 We use only 5, 10 and 25 observed thickness to fit our model to keep our method realistic for applying to other glaciers on WAIS.
relationship described by the glacier dynamics model (see Figure 4.3), in which rheologic coefficient $A$, and flowline width process $\omega$, are unknowns to be estimated. Let $x = (x_1, \ldots, x_n)^T \in \Gamma$ denote a vector of locations where we have ice thickness observations, $H^{\text{obs}} = (H_{x_1}^{\text{obs}}, \ldots, H_{x_n}^{\text{obs}})^T$ be a vector of observed thickness, and $h = (h(x_1), \ldots, h(x_n))^T$ be the true process evaluated at $x$. Similarly, observed inputs are denoted as $V^{\text{obs}}_s, S^{\text{obs}}, a^{\text{obs}},$ and $\tau^{\text{obs}}$. Note that the input observations may be from different spatial locations.

We formulate the statistical model in the following hierarchical form.

**Ice thickness Model:**

- observations model: $H^{\text{obs}} | h, \theta \sim N(h, \sigma^2_H I)$, \hspace{1cm} (4.4)
- physics (deterministic) model: $h | v_s, s, a, \tau, \omega, \theta = M (v_s, s, a, \tau, \omega, A, h_0)$. \hspace{1cm} (4.5)

**Flowline Width Model:** $\omega | \theta \sim \text{GP} (C(\theta_\omega))$. \hspace{1cm} (4.6)
Input Process Model: \( v_s, s, a, \tau \mid \theta, V_s^{\text{obs}}, S^{\text{obs}}, a^{\text{obs}}, \tau^{\text{obs}} \sim f(v_s \mid V_s^{\text{obs}}, \theta_v) f(s \mid S^{\text{obs}}, \theta_s) f(a \mid a^{\text{obs}}, \theta_a) f(\tau \mid \tau^{\text{obs}}, \theta_\tau), \)

velocity model: \( f(v_s \mid V_s^{\text{obs}}, \theta_v), \)
slope model: \( f(s \mid S^{\text{obs}}, \theta_s), \)
accumulation rate model: \( f(a \mid a^{\text{obs}}, \theta_a), \)
thinning rate model: \( f(\tau \mid \tau^{\text{obs}}, \theta_\tau). \)

Prior: \( p(\theta) = p(\sigma_H^2) p(h_0) p(A) p(\theta_\omega) p(\theta_v) p(\theta_s) p(\theta_a) p(\theta_\tau), \)

where \( \theta = (\sigma_H^2, h_0, A, \theta_\omega, \theta_v, \theta_s, \theta_a, \theta_\tau)^T \) includes all the unknown parameters in the model. The ice thickness observations model (4.4) implies conditional independence for the observations, with errors determined by \( \sigma_H^2 \). The physics model (4.5) imposes physical constraints for multiple processes: conditional on the initial thickness value \( h_0 \), rheologic coefficient \( A \), the flow width \( \omega \) and the input processes (surface velocity, surface slope, ice accumulation and thinning rates), the thickness process is related to the input processes through the deterministic glacier dynamics model \( M \). The flow width process \( \omega \) is modeled with a Gaussian process (4.6) with a Matén covariance function \( C(\theta_\omega) \) where \( \theta_\omega = (\nu, \sigma_\omega^2, \phi, \tau_\omega^2)^T \); we fix the smoothness parameter \( \nu = 1.5 \) to reflect that the width is smooth based on expert opinion. The input processes, however, are unknown and assumed to be independent, conditional on their respective model parameters \( \theta_v, \theta_s, \theta_a, \theta_\tau \) and observations. We now provide details for the components of the hierarchical model and comment on their extensions.
4.3.2 Ice Thickness Model

The \( n \)-dimensional thickness observations \( H_{\text{obs}} \), conditional on the underlying true thickness process \( h \), are modeled as independent normal variables centered at the true values with unknown variance \( \sigma_H^2 \). A simple error model is justified since its main source is the instrumental error in acquiring the thickness data. The true thickness at the observation locations have a deterministic relationship with the true surface velocity, surface slope, accumulation rate and thinning rate as described in the glacier dynamics model (4.1) and (4.2), which balances the change in flux \( \left( \frac{\partial}{\partial x} (v_s(x) - \frac{A}{20} \rho g |s(x)|^3 h(x)^4) h(x) \omega(x) \right) \) with the net change in depth \( (a(x) - \tau(x)) \omega(x) \) for mass conservation.

Conditional on the initial thickness value \( h_0 \), the rheologic coefficient \( A \), the flowline width \( \omega(x) \) and the input processes \( (v_s(x), s(x), a(x), \tau(x)) \), solving for thickness \( h(x) \) involves two steps: (i) integrating \( (a(x) - \tau(x)) \omega(x) \) from the divide (initial location used when solving the differential equation) \( x_0 \) to an observation location \( x_j \in \mathbf{x} \), and (ii) computing the flux, \( \bar{v}(x_j) h(x_j) \omega(x_j) \), at \( x_j \) and solving for ice thickness.

Here we describe the details for steps (i) and (ii) listed immediately above. Step (i) integration over the spatial domain can be approximated with a finite sum since, in practice, we only obtain the continuous processes flow width, accumulation and thinning rates at finite locations. Let \( \mathbf{x}^g = (x_{g0}^g, \ldots, x_{gm}^g) \) be a collection of locations; here, we take \( \mathbf{x}^g \) on a grid with \( x_{g0}^g = x_0 \) and \( x_{gm}^g > x_n \) to cover the study domain. We approximate the integral \( \int_{x_0}^{x_j} (a(s) - \tau(s)) \omega(s) ds \) by a sum of the products evaluated at \( \mathbf{x}^g \), \( \sum_{i=0}^{I} \left( a(x_{i,i+1/2}^g) - \tau(x_{i,i+1/2}^g) \right) \omega(x_{i,i+1/2}^g) \Delta_{i,i+1}, \) from \( x_0^g \) to \( x_j^g \), where the interval \( [x_{i,i}^g, x_{i+1}^g] \) contains the observation location \( x_j \), \( a(x_{i,i+1/2}^g) \) is the average of \( a(x_i^g) \) and \( a(x_{i+1}^g) \), \( \tau(x_{i,i+1/2}^g) \) is the average of \( \tau(x_i^g) \) and \( \tau(x_{i+1}^g) \), \( \omega(x_{i,i+1/2}^g) \) is the average of \( \omega(x_i^g) \) and \( \omega(x_{i+1}^g) \), and \( \Delta_{i,i+1} \) is the distance
between locations $x_i^g$ and $x_{i+1}^g$. Note that the grid $x^g$ can be different from the observation locations $x$, and there is a trade off between accuracy and computational efficiency in determining its dimension: high (low) resolution $x^g$ increase (decrease) approximation accuracy and reduce (increase) efficiency. An error term can be added to the process model to capture the discretization error, for instance, $\int_{x_0}^{x_x} (a(s) - \tau(s)) \omega(s) ds = \sum_{i=0}^{I} \left( a(x_{i,i+1/2}^g) - \tau(x_{i,i+1/2}^g) \right) \omega(x_{i,i+1/2}^g) \Delta_{i,i+1} + \epsilon(x)$. However, we neglect $\epsilon(x)$ here, because accumulation rate, thinning rate and width are smooth processes based on physical understanding; in this case, the discretization error is small. Moreover, the Gaussian process model for flow width is very flexible; it may adjust itself to reduce model discrepancy. We repeat step (i) for all observation locations $x_j, j = 1, \ldots, n$. In step (ii) we compute the flux $\bar{v}(x_j) h(x_j) \omega(x_j) \approx \bar{v}(x_0) h(x_0) \omega(x_0) + \sum_{i=0}^{I} \left( a(x_{i,i+1/2}^g) - \tau(x_{i,i+1/2}^g) \right) \omega(x_{i,i+1/2}^g) \Delta_{i,i+1}$ for each observation location $x_j$. Denote the flux quantities as $F_j, j = 1, \ldots, n$. Then, we solve the below fifth order polynomial for ice thickness, 

$$F_j - \left( v_s(x_j) - \frac{A}{20} (\rho gs(x_j))^3 h(x_j)^4 \right) h(x_j) \omega(x_j) = 0. \quad (4.7)$$

### 4.3.3 Flowline Width Model

The flowline width $\omega(x), x \in \Gamma$ is not observed, however, we know it is relatively smooth based on the estimates by tracing the direction of velocity vector field and our physical understanding of glaciers [Docquier et al. 2014]. We impose smoothness to flowline width by modeling it with a Gaussian process. Here, we choose a Matérn covariance function with smoothness $\nu = 1.5$, since it produces a relatively smooth process. Therefore, the flowline width $\omega = (\omega(x_0^g), \ldots, \omega(x_m^g))^T$ at the grid $x^g$ has a multivariate normal distribution with covariance matrix $\Sigma$. The $(i, j)$ element of the covariance matrix, $\Sigma_{ij}$, is the covariance between $\omega(x_i^g)$
and $\omega(x_i^q)$, and $\Sigma_{ij} = C(|| x_i^q - x_j^q ||) = \sigma^2_\omega \left( 1 + \frac{\sqrt{3}|| x_i^q - x_j^q ||}{\phi} \right) \exp \left( -\frac{\sqrt{3}|| x_i^q - x_j^q ||}{\phi} \right)$ for $i \neq j$, or $\Sigma_{ij} = \sigma^2_\omega + \tau^2$ for $i = j$ (Rasmussen and Williams 2005). Other models for flow width that preserve the smoothness of the process and can interpolate in between observations, for instance, B-spline basis function and Gaussian process with a smoother covariance function, are also reasonable for this application.

Based on the observed flowline width, we know it is wider upstream and narrower downstream; this knowledge is incorporated in the flowline width model by specifying a mean function for the Gaussian process. In the glacier dynamics model (4.1), we can assume a value for the rheologic coefficient $A = 0$ and plug in the narrowest observed flowline width (red dashed line in Figure 4.1b) to evaluate the integral from the divide to the thickness observation locations $x$ assuming there are no observational errors, we then solve for the glacier width at $x$ using the observed ice thickness $H^{obs}$ and (4.7). A linear interpolation of the glacier width at $x$ is then the mean function for the Gaussian process. Based on our simulated examples, the choice of $A$ value and the observed flowline width to deduce the mean function does not affect the results of the Bayesian model.

### 4.3.4 Input Process Model

The multiple data sets are implicitly integrated via their respective true processes through the glacier dynamics model in (4.1). We model the input processes (surface velocity, surface slope, ice accumulation and thinning rates) as independent processes conditional on their respective parameters ($\theta_v, \theta_s, \theta_a, \theta_\tau$). Our assumption does not imply that individual input processes are marginally independent, rather we assume the small-scale variation in the true processes, conditional on the smoothed observations, are independent. Let $V^{obs}_v, S^{obs}_s, A^{obs}_a$ and $\tau^{obs}_\tau$ be vectors of observed surface velocity, surface slope, ice accumulation and thinning rates,
respectively. Each individual process is related to their respective observations by the following.

velocity model: \[ v_s(x) \mid \theta_v, S_{\lambda v} v_{\text{obs}} = S_{\lambda v} V_s + \epsilon_v(x), \quad \epsilon_v(x) \sim N(0, \sigma_v^2), \]
slope model: \[ s(x) \mid \theta_s, S_{\lambda s} s_{\text{obs}} = S_{\lambda s} S_s \quad + \epsilon_s(x), \quad \epsilon_s(x) \sim N(0, \sigma_s^2), \]
accumulation rate model: \[ a(x) \mid \theta_v, a_{\text{obs}} = S_{\lambda a} a + \epsilon_a(x), \quad \epsilon_a(x) \sim N(0, \sigma_a^2), \]
thinning rate model: \[ \tau(x) \mid \theta_v, \tau_{\text{obs}} = S_{\lambda \tau} \tau + \epsilon_{\tau}(x), \quad \epsilon_{\tau}(x) \sim N(0, \sigma_{\tau}^2), \]

where \( S_{\lambda v}, S_{\lambda s}, S_{\lambda a}, \) and \( S_{\lambda \tau} \) are smoothing matrices, for instance smoothing splines \(^\text{(Nychka 1988)}\), and \( \epsilon_v(x), \epsilon_s(x), \epsilon_a(x) \) and \( \epsilon_{\tau}(x) \) are small-scale variation in the true processes not captured by the smoothed observations. In practice, surface observations are obtained from various sources and do not share the same locations; therefore, they are often interpolated to the same locations before entering physical models. The input process models allow for observations with different spatial resolutions.

In the glacier dynamics model, we evaluate the accumulation and thinning rates at location \( \mathbf{x}^g, \mathbf{a} = (a(x_{0}^g), \ldots, a(x_{m}^g))^T \) and \( \mathbf{\tau} = (\tau(x_{0}^g), \ldots, \tau(x_{m}^g))^T \), to approximate the integral using step (i) in Section 4.3.2, and we evaluate surface velocity and slope at observation locations \( \mathbf{x} \), denoted by \( \mathbf{v}_s = (v_s(x_1), \ldots, v_s(x_n))^T \) and \( \mathbf{s} = (s(x_1), \ldots, s(x_n))^T \), respectively, to solve for thickness using step (ii) in Section 4.3.2. Therefore, the dimension of the input process models \( \mathbf{v}_s, \mathbf{s}, \mathbf{a}, \mathbf{\tau} \mid \theta \) is about \( 2(n + m) \).

4.3.5 Model Inference and Computational Details

Maximum likelihood inference for the hierarchical model is difficult. The challenge is that the glacier dynamics model can not be solved analytically; its solutions
can only be obtained through numerical approximation. The likelihood function is hence not in closed form; therefore, its maximization requires sophisticated optimization over a high-dimensional parameter space. Bayesian inference based on the posterior distribution \( \pi(\theta, v_s, s, a, \tau, \omega | H_{\text{obs}}, V_{s_{\text{obs}}}, S_{\text{obs}}, a_{\text{obs}}, \tau_{\text{obs}}) \) is not difficult in principle.

However, the computational cost of inference for the hierarchical model poses a major challenge in practice. To handle the computational challenge, two compromises are made in the hierarchical model: (1) we pre-smooth the input processes using smoothing splines and treat them as the true processes for solving the ice thickness in the glacier dynamics model, and (2) instead of estimating the range parameter \( \phi \) for the flowline width, we estimate it from the observed flowline width. After these two modifications to the full hierarchical model, its posterior distribution reduces to \( \pi(A, \sigma^2_\omega, h_0, \tau^2, \omega | H_{\text{obs}}, V_{s_{\text{obs}}}, S_{\text{obs}}, a_{\text{obs}}, \tau_{\text{obs}}) \). We use Markov Chain Monte Carlo (MCMC) with Metropolis-Hasting updates to sample from the posterior distribution. We use a uniform prior with a range from 0 to \( 10^{-16} \) for the rheologic coefficient \( A \); the range is selected based on expert opinion. We use inverse gamma priors for variance \( \sigma^2_\omega \) (IVG(2,10^8)) and \( \tau^2 \) (IVG(2,10^4)), based on exploratory analysis as the follow: for fixed values of rheologic coefficient, if we assume that the errors in thickness observations are negligible, then we can derive flow width using (4.3) and (4.7); we explore the variation in the flowline width by solving it using several values of \( A \) and thickness observations with added errors.

Fixing the range parameter at a reasonable value reduces computational expenses and we find in practice that this approach does not affect the results significantly; also the main purpose of assuming a spatial process here is to interpolate assuming some smoothness. Based on arbitrarily chosen flow width as shown in Figure 4.1b, we estimated \( \hat{\phi} = 40,000 \) (m). In our implementation, we solve the
high-order polynomial \((4.7)\) using the R package rootSolve \cite{Soetaert2009}. Each MCMC is run from three dispersed starting values; based on the trace plots from several MCMC runs, we decide that 200,000 MCMC iterations suffice.

### 4.4 Thwaites Glacier

We first apply the hierarchical approach described in Section 4.3 to synthetic observations simulated from the physical model. This allows us to investigate the following: (1) inference and prediction performance, (2) the skill of recovering the local features of thickness by combining the physical model and surface observations, and (3) the number of thickness observations needed to provide useful predictions. Then, we apply the model to the Thwaites glacier with 5, 10 and 25 thickness observations, about 46 km, 30 km and 14 km apart in distance, respectively along the flowline.

#### 4.4.1 Simulated Example

To preserve the topographic features of the ice thickness, we treat the observed thickness on Thwaites glacier as the true underlying process and simulate surface velocity from the physical model \((4.1)\). We then fit the hierarchical model using a small number of synthetic thickness observations generated from the synthetic true thickness with different magnitude of \textit{i.i.d} normal error (Figure 4.4). This model fitting procedures are repeated for 5, 10, and 25 training locations. We then compare the performance in estimating the rheologic coefficient \(A\), and predicting the ice thickness. To make synthetic data match realistic observations, we also use the observed surface slope, ice accumulation rate, thinning rate and the observed flow width corresponding to the red dashed line in Figure 4.1b for data generation.
We let the true rheologic coefficient to be $A = 10^{-18}$.

The inference on the rheologic coefficient $A$ under the three error schemes are similar: posterior means are all around $5.07 \times 10^{-17}$, and 95% credible intervals decrease for increasing training locations, for instance, 95% CI’s are $[2.085 \times 10^{-18}, 9.683 \times 10^{-17}]$ and $[3.002 \times 10^{-18}, 9.632 \times 10^{-17}]$ for 5 and 25 training locations (under the medium error scheme), respectively. The predicted flow width for the entire flowline is shown in Figure 4.5. We see that as the number of observations increases, the uncertainties associated with the width predictions decrease dramatically. We also study the 95% credible intervals from the width predictions; the coverages, shown in Table 4.2, should be around 95%, however, are substantially lower than the nominal as the number of training locations increases. The thickness predictions from the physical model are shown in Figure 4.6. From the few model realizations, we see that the fine scale structure is preserved from the prediction, even when using only 5 training locations. As the number of training locations increases, the predicted thickness is tighter around the synthetic true thickness. However, the predicted thickness using 5 training locations has a larger uncertainty as its location gets further from the divide. The increasing uncertainty in thickness prediction is due to the ODE solution for the flux in (4.3); in approximating the integral from the divide to downstream, small departures from the width estimate to the synthetic true are accumulated along the distance. This reflects the sensitivity of ODE solution to errors, and we could reduce the uncertainty by solving the system using smaller segments along the glacier. The uncertainty also reduces as the number of training locations increases, results from better width estimation. It seems that estimating the observational error in thickness using our approach is very difficult; observational error $\sigma_H$ has posterior means around 5,000 m for all three error schemes. This is mostly likely a result from the discretiza-
tion in approximating the integral and simplification of observational error model. A practical implication of our results above is that thickness observational errors does not have substantial impact in thickness reconstruction, because much of the information about ice thickness comes from the physical model and well observed surface observations.

Table 4.2: The coverage of 95% CI’s for flow width prediction

<table>
<thead>
<tr>
<th>Number of thickness observations</th>
<th>Magnitude of Observational Errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small ($\sigma_H = 10$)</td>
<td>Medium ($\sigma_H = 50$)</td>
</tr>
<tr>
<td>5</td>
<td>0.962</td>
<td>0.960</td>
</tr>
<tr>
<td>10</td>
<td>0.948</td>
<td>0.855</td>
</tr>
<tr>
<td>25</td>
<td>0.628</td>
<td>0.475</td>
</tr>
</tbody>
</table>

Figure 4.4: Synthetic ice thickness data with different magnitude of observational errors. Small, medium and large errors have standard deviations of 10, 50 and 100 m, respectively.
### Figure 4.5: Predicted flow widths using 5, 10 and 25 synthetic ice thickness observations. Flow width predictions using small (a), medium (b) and large (c) magnitude of observational errors. The magnitude of observational errors does not change the shape of the width. Larger number of training locations produces tighter flow width that is more comparable to the synthetic true width.

#### 4.4.2 Application

We now apply the hierarchical approach to the Thwaites Glacier using 5, 10 and 25 training locations. Figure 4.7 shows the model’s thickness prediction and Figure 4.8 shows the flow width prediction. Our estimate for the rheologic coefficient is $5.085 \times 10^{-17} (2.259 \times 10^{-18}, 9.845 \times 10^{-17})$ using 25 training locations; the inference using 5 and 10 training locations are similar, therefore, not reported. As in the
(a) Use 5 synthetic observations with small error  
(b) Use 5 synthetic observations with medium error  
(c) Use 5 synthetic observations with large error  

(d) Use 10 synthetic observations with small error  
(e) Use 10 synthetic observations with medium error  
(f) Use 10 synthetic observations with large error  

(g) Use 25 synthetic observations with small error  
(h) Use 25 synthetic observations with medium error  
(i) Use 25 synthetic observations with large error

Figure 4.6: Predicted ice thickness (m) from flow model using 5, 10, and 25 synthetic thickness observations. Prediction results are similar under different error schemes, and a larger number of training locations provide more certain thickness prediction.
simulation study, the uncertainty in thickness prediction increases as its location is further from the divide, reflecting the sensitivity to error in solving ODE.

4.5 Conclusions and Future Work

We provide a Bayesian framework with the flexibility of integrating a physical model, multiple observations and uncertainty sources. We show via simulated example and observational data that there are some advantages of our Bayesian approach compared to a non-statistical approach: (1) we obtain uncertainty estimates for the parameters of the physical model, and (2) we provide uncertainties for our predictions of ice sheet thickness.

There are some important scientific implications of our study. Using 10 training locations along the glacier can reasonably reconstruct the ice structure in some regions. This implies that spatially sparse thickness observations, about 30 km apart can be used, in combination of good quality surface observations to deduce ice thickness. There is a potential to apply our approach to other regions of Antarctic Ice Sheet for estimating ice thickness.

Although we would like to use as little as 5 thickness observations for reconstructing the ice thickness along the entire glacier, our current result shows that more data is required; whether we can improve ice thickness prediction using less observations remains an open question. The physical model in our method is very sensitive to small variations in the input processes, especially the surface elevation. Therefore, it is pre-smoothed in the Bayesian model. This pre-smoothing procedure is often performed in practice (Pollard and DeConto 2012; Kamb and Echelmeyer 1986; Berliner et al. 2008b) to eliminate small-scale features that would otherwise result in unreasonable variations. However, a more rigorous statistical approach requires modeling and quantifying the uncertainties in the input processes. Due to
Figure 4.7: Predicted ice thickness from the dynamics model using 5 (a), 10 (c), and 25 (c) training locations. The left panel shows the 95% credible intervals for predicted ice thickness plus observational errors using 5 (b), 10 (d), and 25 (f) training locations.
Figure 4.8: Thwaites Glacier flow width estimates using different number of training locations. More training locations recreates the local feature of the width.

computation consideration, we have not include these uncertainties in the current approach.

Further interesting questions can be studied. For instance, we may be interested in finding out whether the proposed method works well in some regions along the glacier and works poorly in others. An equivalent statement is: to what extent can we use this simple physical law to model complex 3-dimensional ice dynamics? This may allow us to use the simple model for some regions and the complex model for others in combination to reconstruct ice thickness.
Chapter 5

Discussion

In this thesis, I have presented a projection-based model for high-dimensional spatial data. This approach is powerful as it reduces the number of random effects and makes them nearly independent, leading to easier construction of efficient samplers for random effects. It also provides researchers the flexibility to adjust for spatial confounding. The projection-based model can be extended easily to SGLMMs with other link functions and can be fitted under either a Bayesian or maximum likelihood inferential framework. I also presented a Bayesian hierarchical model that integrates a simple glacier dynamics model, multiple data sets and uncertainty sources to estimate key quantities in a glacier dynamics model. This allows us to quantify the contribution of uncertainties from various sources to reconstruct ice sheet thickness in the Thwaites glacier in West Antarctica using sparsely observed ice thickness data.
5.1 Projection-Based Spatial Models

In order for our methods to be conveniently used by non-experts, I am currently developing, with the assistance of undergraduate researchers, an R package (R Core Team, 2014) for fitting various projection-based SGLMMs. This includes the Poisson model with the log or probit link function and binomial model with the logit or probit link function. The final package will provide researchers a fast and flexible computational tool for modeling high-dimensional non-Gaussian spatial data. A goal will be to allow for users to fit these models under both the Bayesian and frequentist paradigms.

The current method relies on parallelization to handle large matrix computations. For Bayes inference, users are recommended to fit our model on machines with multi-core CPUs. If computational power is limited, users may consider discretizing the range parameter and pre-compute the eigencomponents. I find that maximum likelihood inference for the projection-based model is less computationally demanding, since the algorithm typically converges relatively quickly (it requires fewer eigendecompositions of the covariance matrix). We have successfully implemented this model for data sizes of 10,000 for both Bayesian and maximum likelihood approaches; it is unclear how to adapt our method to carry out inference for hundreds of thousands of data points.

There are several promising directions for future work:

(a) Eigencomponent approximation. In Chapter 2, I have shown that the random projection algorithm approximates the eigenvectors and eigenvalues well via a numerical experiment. A more rigorous approach to show the properties of these estimates will be very interesting and a major contribution to the area of probabilistic algorithms for constructing approximate matrix decompositions. To obtain such theoretical results, one would need to combine
well established approximation results based on random projections, with some existing important findings on convergence of eigenspaces \cite{zwald2006} and operation perturbation theory \cite{koltchinskii2000}.

(b) Projection-based model and INLA. The INLA approach \cite{rue2009} provides a fast approximate numerical method for carrying out inference for latent Gaussian random field models without expensive Monte Carlo sampling. An interesting avenue for future research is combining our reduced-dimensional reparameterization with INLA. \cite{eidsvik2012} considers an analogous approach that involves fitting a predictive process approach using INLA. We provide an outline of our approach below.

The projection-based models have the following form,

$$g\{E(Z | \beta, \delta, \theta)\} = X\beta + U\phi D^{1/2}\delta = Hv,$$

where \(H = [X, U\phi D^{1/2}]\), \(\theta = (\phi, \sigma^2)^T\) and \(v = (\beta^T, \delta^T)^T\). The prior for \(v\) is \(N(0, \Sigma_v)\), where

\[
\Sigma_v = \begin{bmatrix} \sigma^2 I_m \times m & 0 \\ 0 & \Sigma_\beta \end{bmatrix}.
\]

The marginal posterior distribution of \(\theta\) given observations \(Z\) is,

$$\pi(\theta | Z) \propto \frac{\pi(Z, v, \theta)}{\pi(v | Z, \theta)},$$

where the full conditional \(\pi(v | Z, \theta)\) is not available in closed-form when the likelihood model is non-Gaussian. The INLA approach consists of Gaussian approximation \(\tilde{\pi}_G(v | Z, \theta)\) to the full conditional, then approximate the posterior distribution in (5.1) denoted by \(\tilde{\pi}(\theta | Z)\), and lastly compute the posterior marginals for any regression effects or spatial effects by numerically
integrating over covariance parameters $\tilde{\pi}(v_j \mid Z) = \int \tilde{\pi}_G(v_j \mid Z, \theta) \tilde{\pi}(\theta \mid Z) d\theta$. Details of each step are described below.

1. Gaussian approximation to $\pi(v \mid Z, \theta)$, denoted by $\tilde{\pi}_G(v \mid Z, \theta)$.
   - Approximate $\pi(Z \mid v, \theta)$ by Taylor expansion around the mode $\tilde{v}$ to second order. For binary data,
     \[
     \log(\pi(Z \mid v, \theta)) \approx -\frac{1}{2} v^T H^T D_2 H v + v^T H^T (Z - d_1 + D_2 H \tilde{v}) + \text{const}(\tilde{v}),
     \]
     where $D_2 = \text{diag}(\frac{N \exp(H \tilde{v})}{1 + \exp(H \tilde{v})})$ is an $n \times n$ diagonal matrix, $d_1 = \frac{N \exp(H \tilde{v})}{1 + \exp(H \tilde{v})}$ is an $n$-dimensional vector and const($\tilde{v}$) is a constant involving $\tilde{v}$.
   - $\tilde{\pi}_G(v \mid Z, \theta) \propto \pi(G) \pi(v \mid \theta)$ is multivariate Normal $N(Q^{-1}b, Q^{-1})$, where $Q = H^T D_2 H + \Sigma_v^{-1}$, and $b = H^T (Z - d_1 + D_2 H \tilde{v})$
   - Find the mode $\tilde{v}_G^*$ using Newton-Raphson. We can solve $v^{(l)} = Q^{-1}b$ iteratively until convergence.

2. Laplace approximation to (5.1) using
   \[
   \tilde{\tau}(\theta \mid Z) \propto \frac{\pi(Z, v, \theta)}{\tilde{\pi}_G(v \mid Z, \theta)} \bigg|_{v=\tilde{v}_G}, \quad (5.2)
   \]
   and the posterior approximation is explored by numerical routines.
   - Choose a starting value $\theta$ and evaluate $\ln \tilde{\pi}(\theta \mid Z)$
   - Optimize $\log(\tilde{\pi}(\theta \mid Z))$ to obtain a mode $\theta^*$
   - Compute Hessian matrix at the mode by taking finite differences.
     Let $\Sigma_\theta = \text{Hess}^{-1}$, and let $\Sigma_\theta = V \Lambda V^T$ be its eigendecomposition.
     Transform parameters by standardizing $\theta(z) = \theta^* + V \Lambda^{1/2} z$.  

– Explore \( \log(\tilde{\pi}(\theta \mid Z)) \) by evaluating it on a grid with z-parameterization.

3. Approximate posterior marginal for \( \tilde{\pi}(v_j \mid Z) \)

INLA for SGLMMs is computationally expensive due to evaluations of \( Q^{-1} \), which has a dimension of \((n + p) \times (n + p)\) ([Rue et al., 2009]). The computation can be reduced by combining the predictive process to reduce the dimension of the random effects ([Eidsvik et al., 2012]). In Chapter 2, I compared the projection-based approach with the predictive process ([Banerjee et al., 2008]); for the same rank of random effects, our model produces better results both in terms of inference and prediction. These improvements are due to model specification, therefore, are independent to model-fitting algorithm. This indicates that the projection-based model combined with INLA will likely provides the same improvements over the predictive process, and computationally faster compared to using MCMC in Bayesian inference.

(c) Extension to non-stationary and spatio-temporal processes. In many applications, the processes of interest are non-stationary or spatio-temporal. Modeling such processes has been an active area of research. Developing projection-based methods in the context of these approaches would be useful. This is an interesting and challenging area of future research.

(d) Fast approximate methods for SGLMMs for semi-continuous spatial data. Ice sheet thickness data are actually semi-continuous in that many of the data points are zeroes (no ice sheet) while the rest are continuous and positive. It would be of interest to develop an extension of the methods I have developed here in order to obtain an efficient approach for modeling such data.
5.2 Statistical Methods for Ice Sheet Thickness

Computationally efficient statistical methods that combine complex physical models and multiple observations for modeling physical processes are essential to improve our understanding of the physical system and to provide better uncertainty quantification and prediction. The Bayesian approach I have shown for modeling glaciers on WAIS is an example of such a problem. In the methodology developed in Chapter 4, I have chosen a rather simplified flowline model for one-dimensional ice streams, and a simplified error model for the observed data. The reconstructed ice profile has large uncertainties when only 5 thickness observations are used in model fitting; these are reduced substantially as I increase the number of observations to 25. However, there are a few regions showing large departures, up to 200 meters, between predicted (from the physical model) and observed ice thickness. This might be because some of the input processes, such as ice velocity, have non-negligible errors that are currently not accounted for in the statistical model. Some potentially fruitful and challenging avenues of future research include:

- Investigate sources of errors and biases in data inputs required by the flowline model; this will likely provide more accurate ice thickness prediction.

- Incorporate more complex physical models that allow the extension from one-dimensional to three-dimensional reconstruction of ice thickness on WAIS.

- Address computational issues involved in more sophisticated statistical and physical models. Increasing the model complexity inevitably increases the computational cost. Therefore, the investigation of the applicability of integrating projection-based model, complex physical models and high-dimensional observations for modeling physical processes will potentially open avenues for interesting future research.
SGLMMs with Small-Scale (Nugget) Spatial Effect

The following is an extension of the method discussed in Chapter 2. For SGLMMs where inclusion of small scale, non-spatial heterogeneity is appropriate, the model becomes,

$$g \{ E(Z(s) \mid \beta, W(s)) \} = X(s)\beta + w(s) + \epsilon(s), \quad (A.1)$$

where $\epsilon(s) \overset{iid}{\sim} N(0, \tau^2)$. We provide implementations of our method for two cases: (1) when Gibbs sampling of the latent variables is available, and (2) when it is not. Examples for case (1) are the spatial binary model with probit link (considered by Berrett and Calder, 2016) and spatial probit model for correlated ordinal data (Schliep and Hoeting, 2015); examples for case (2) are already considered in this manuscript.

We begin by redefining some notation. Let $W = (W_1, \ldots, W_n)^T$ denote the latent variable, $Z = (Z_1, \ldots, Z_n)^T$ the observed spatial binary data and $X$ the $n \times p$ design matrix.

**Case (1):** We first consider the case where Gibbs sampling is available for the
latent variables, for example when using SGLMM with a probit link for binary data. The model is defined as

$$Z_i = \begin{cases} 
1, & Y_i \geq 0 \\
0, & Y_i < 0 
\end{cases}$$

(A.2)

where $$Y_i = X_i \beta + W_i + \epsilon_i$$. $$W \sim MVN(0, \sigma^2 R_\phi)$$ captures large-scale spatial variation and $$\epsilon_i \sim i.i.d. N(0, \tau^2)$$ captures small-scale variation. The conditional distribution for $$Y | \beta, \sigma^2, \phi, \tau^2$$ is therefore multivariate normal with mean $$X \beta$$ and variance $$\sigma^2 R_\phi + \tau^2 I$$. Our method can be used to facilitate model fitting in this case as follows: We approximate the eigen-components of $$R_\phi$$ using random projections and obtain its first $$m$$ eigenvectors $$U_\phi = [u_1, \ldots, u_m]$$ and eigenvalues $$D_\phi = \text{diag}(\lambda_1, \ldots, \lambda_m)$$. Let $$M_\phi = U_\phi D_\phi^{1/2}$$ be the projection matrix, then we reduce the dimension of the latent variables by approximating $$W$$ with $$M_\phi \delta$$. For a specific value of $$\phi$$, we can treat $$M_\phi$$ as fixed spatial covariates and $$\delta$$ the corresponding coefficients. Write $$X_\phi = [X, M_\phi]$$ and $$\beta_\phi = (\beta^T, \delta^T)^T$$ as the reparameterized design matrix and coefficients, respectively, then $$Y_i$$ is approximated by $$X_i \beta + M_{\phi,i} \delta + \epsilon_i$$ and can be rewritten as $$X_{\phi,i} \beta_\phi + \epsilon_i$$. We use a normal conjugate prior for $$\beta$$, inverse gamma conjugate priors for $$\sigma^2$$ and $$\tau^2$$, and a uniform prior for $$\phi$$. Then, fitting the reduced-rank Bayesian probit model involves the following steps.

At the $$t^{th}$$ iteration of the algorithm,

Step 1: Gibbs update for latent variables. Sample $$Y^{(t)}$$ from

$$Y|Z, \beta^{(t-1)}, \sigma^2(t-1), \phi^{(t-1)}, \tau^2(t-1)$$

(a) Compute projection matrix $$M_\phi$$ for $$\phi^{(t-1)}$$. Form $$X_\phi$$ and $$\beta_\phi$$. 
(b) For \( i = 1, \ldots, n \), draw \( Y_i \) from

\[
Y_i \mid Z_i, Y_{-i}, \beta, \sigma^2, \phi, \tau^2 \sim \begin{cases} 
\text{TN}(X_{\phi,i}\beta_{\phi}, \tau^2, 0, \infty), & \text{if } Z_i = 1 \\
\text{TN}(X_{\phi,i}\beta_{\phi}, \tau^2, -\infty, 0), & \text{if } Z_i = 0,
\end{cases}
\]

where \( \text{TN}(\mu_Y, \sigma_Y^2, 0, \infty) \) is a truncated normal distribution with lower bound 0, upper bound \( \infty \), mean \( X_{\phi,i}\beta_{\phi} \) and variance \( \tau^2 \).

Step 2: Gibbs update for \( \beta_{\phi} \).

Sample from

\[
\beta_{\phi} \mid Z, Y^{(t)}, \sigma^{2(t-1)}, \phi^{(t-1)}, \tau^{2(t-1)} \sim \text{MVN}
\left( \hat{\beta}_{\phi}, \left( \frac{1}{\tau^{2(t-1)}} - \Sigma^{-1}_\beta \right)^{-1} X^T_{\phi} Y^{(t)} \right),
\]

where

\[
\hat{\beta}_{\phi} = \left( \frac{1}{\tau^{2(t-1)}} X^T_{\phi} X_{\phi} + \Sigma^{-1}_\beta \right)^{-1} \frac{1}{\tau^{2(t-1)}} X^T_{\phi} Y^{(t)},
\]

and

\[
\Sigma^{-1} = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & \sigma^2(t-1)I_{m \times m} \end{bmatrix}
\]

with \( \Sigma_\phi \) denotes the normal prior variance.

Step 3: Gibbs update for \( \tau^2 \).

Step 4: Gibbs update for \( \sigma^2 \).

Step 5: Metropolis-Hastings update for \( \phi \).

We have not provided details for steps 3-5 since they remain the same as when fitting SGLMMs in general. Furthermore, techniques for dealing with non-identifiable
parameters (Berrett and Calder, 2012, 2016) can also be used.

Case (2): We now consider the case where Gibbs sampling from the latent variable is not available. We first explain why the reparameterization for Case (1) is not suitable here, and then provide an alternative strategy. In Case (1) above, $W$ is reparameterized with a low-rank representation, however, the dimension of latent variable $Y$ remains high; $Y$ is approximated by $X\beta + M_\phi \delta + \epsilon$, and has a normal distribution with mean $X\beta$ and covariance $\sigma^2 M_\phi M_\phi^T + \tau^2 I$. Constructing efficient MCMC to sample $Y$ from its full conditional distribution is not easy due to its high dimensions. Hence, we propose an alternative: reduce the dimension of $Y$ by approximating $W + \epsilon$ with $U_\theta D_\theta^{1/2} \delta$, where $U_\theta$ and $D_\theta$ are eigenvectors and eigenvalues of $\sigma^2 R_\phi + \tau^2 I$, respectively. Hence, the eigencomponents here depend on all parameters $\theta = (\sigma^2, \phi, \tau^2)^T$ of the covariance function. In fact $U_\theta$ is identical to $U_\phi$ from Case (1), and $D_\theta$ is identical to $\sigma^2 D_\phi + \tau^2 I_{m \times m}$. This alternative reparameterization provides some computational gains. The latent variable $Y$ is now approximated by $X\beta + M_\theta \delta = [X, M_\theta](\beta^T, \delta^T)^T$ whose full conditional distribution has $m + p$ dimensions. Reducing the dimension of the posterior distribution allows for easier construction of efficient MCMC.
Bibliography


Vita
Yawan Guan

Education

• Ph.D. Statistics, Pennsylvania State University, 2017
  Advisor: Dr. Murali Haran
  Committee: Ephraim M. Hanks, Francesca Chiaromonte, Chris E. Forest
  Thesis title: Reduced-dimensional Non-Gaussian Spatial Models and Statistical Methods for Studying the West Antarctic Ice Sheet

• B.S. Finance and Mathematics, Pennsylvania State University, 2012

Publications


Honors and Awards

• Recipient of the 2016 J. Keith Ord Scholarship in Statistics for research and mentoring in the area of spatial and environmental statistics.

• Recipient of the 2016 Jack and Eleanor Pettit Scholarship in Science.